Synthese (2013) 190:1555–1585
DOI 10.1007/s11229-011-9892-x

A frequentist interpretation of probability for model-based inductive inference

Aris Spanos

Received: 5 October 2010 / Accepted: 9 February 2011 / Published online: 26 February 2011

Abstract The main objective of the paper is to propose a *frequentist interpretation of probability* in the context of model-based induction, anchored on the Strong Law of Large Numbers (SLLN) and justifiable on empirical grounds. It is argued that the prevailing views in philosophy of science concerning induction and the frequentist interpretation of probability are unduly influenced by enumerative induction, and the von Mises rendering, both of which are at odds with frequentist model-based induction that dominates current practice. The differences between the two perspectives are brought out with a view to defend the model-based frequentist interpretation of probability against certain well-known charges, including [i] the circularity of its definition, [ii] its inability to assign 'single event' probabilities, and [iii] its reliance on 'random samples'. It is argued that charges [i]–[ii] stem from misidentifying the frequentist 'long-run' with the von Mises collective. In contrast, the defining characteristic of the long-run metaphor associated with model-based induction is neither its temporal nor its physical dimension, but its repeatability (in principle); an attribute that renders it operational in practice. It is also argued that the notion of a statistical model can easily accommodate non-IID samples, rendering charge [iii] simply misinformed.

Keywords Frequentist interpretation of probability · Circularity · Random samples · Single event probability · Randomness · Long-run metaphor · Strong law of large numbers · Error statistics · Duhem-Quine problem · Model-based induction · Post-data severity evaluation

A. Spanos (✉)
Department of Economics, Virginia Tech, Blacksburg, VA 24061, USA
e-mail: aris@vt.edu

1 Introduction

The frequentist approach to statistical modeling and inference (Cox and Hinkley 1974) has dominated empirical modeling in most applied fields since the 1940s, but a number of foundational problems relating to its underlying *inductive reasoning* have bedeviled and hindered its proper implementation in practice; see Godambe and Sprott (1971), Harper and Hooker (1976), Morrison and Henkel (1970) inter alia. The philosophy of science literature has focused primarily on articulating some of these foundational problems, and highlighting the limited scope and applicability of the frequentist interpretation of probability; see Salmon (1967), Kyburg (1974), Giere (1984), Seidenfeld (1979) and Gillies (2000). For example, Howson and Urbach (2006) present a grim picture of frequentist probability:

> While the use of the limit definition allows us to regard objective probabilities as probabilities in the purely formal sense of the probability calculus, it has nevertheless elicited from positivistically-minded philosophers and scientists the objection that we can never in principle, not just in practice, observe the infinite n-limits. Indeed, we know that in fact (given certain plausible assumptions about the physical universe) *these limits do not exist*. For any physical apparatus would wear out or disappear long before n got to even moderately large values. So it would seem that no empirical sense can be given to the idea of a limit of relative frequencies. (p. 47)

The primary objective of this paper is to defend the *frequentist interpretation* of probability for model-based induction against well-known charges, including: [i] the circularity of its definition, [ii] its inability to assign 'single event' probabilities, and [iii] its reliance on 'random samples' (Salmon 1967; Hajek 2007).

The main argument of this paper is that, although charges [i]–[iii] constitute legitimate criticisms of enumerative induction anchored on the von Mises 'collective', they are unfounded when leveled against the model-based 'stable long-run frequencies' interpretation (Neyman 1952), grounded on the Strong Law of Large Numbers (SLLN). When enumerative induction and von Mises's collective (1928) are viewed from a model-based perspective, several weaknesses become apparent: (i) they are both based on a highly restrictive statistical model, (ii) identifying the frequentist 'long-run' with von Mises's collective is highly misleading, and (iii) any attempt to provide a frequentist interpretation of probability using the notion of a collective is ill-fated for purely mathematical reasons; Williams (2001).

Section 2 presents a brief summary of enumerative induction and the von Mises frequentist interpretation. Section 3 traces the grounding of the notion of a statistical model $\mathcal{M}_\theta(\mathbf{x})$ to axiomatic probability. Section 4 summarizes the Fisher–Neyman–Pearson (F–N–P) approach to frequentist inference, briefly mentioning some of its long-standing foundational problems. The error-statistical perspective is presented in Sect. 5 as a model-based refinement/extension of the F–N–P framework aspiring to address some of these problems. This perspective is used in Sect. 6 to articulate a frequentist interpretation of probability and address charges [i]–[ii]. This interpretation is then used in Sect. 7 to bring out the differences between enumerative and model-based induction with a view to elucidate the nature and justification of model-based

inference. The same perspective is used in Sect. 8 to revisit charge [iii]. Section 9 attempts to place model-based induction in the context of the broader philosophical discussions pertaining to the interpretation of probability and the nature of inductive inference.

2 Enumerative induction and the von Mises frequentist interpretation

Since the 1940s, the philosophy of science literature has called into question the frequentist interpretation of probability on several grounds by focusing on *induction by enumeration* and the Von Mises (1928) frequentist rendering based on the notion of a *collective*; see Salmon (1967), Gillies (2000).

Induction by enumeration: if m/n observed A's are B's, infer (inductively) that approximately m/n of all A's are B's.

Enumerative induction is widely viewed in philosophy of science as the quintessential form of statistical induction, and Von Mises (1928) frequentist interpretation of probability as providing the link between the empirical relative frequencies and the corresponding mathematical probabilities.

The cornerstone of this link is von Mises notion of **a collective:** an infinite sequence of outcomes in the context of which each relevant event has a limiting relative frequency that is invariant to place selections. More formally, a collective is an infinite sequence $\{x_k\}_{k=1}^{\infty}$ of outcomes of 0's and 1's, representing the occurrence of event A [when $x_k = 1$] that satisfies two conditions:

$$\begin{aligned} &\text{(C)} \quad \text{Convergence:} \quad \lim_{n\to\infty}\left(\frac{1}{n}\sum_{k=1}^{n} x_k\right) = p_A, \\ &\text{(R)} \quad \text{Randomness:} \quad \lim_{n\to\infty}\left(\frac{1}{n}\sum_{k=1}^{n} \varphi(x_k)\right) = p_A, \end{aligned} \tag{1}$$

where $\varphi(.)$ is a mapping of admissible *place-selection* sub-sequences $\{\varphi(x_k)\}_{k=1}^{\infty}$.

In contrast to the use of enumerative induction in philosophical discussions, practitioners in most applied fields rely on frequentist *model-based* induction grounded on the notion of a statistical model $\mathcal{M}_\theta(\mathbf{x})$, assumed to represent an idealized generating mechanism that could have given rise to data $\mathbf{x}_0 := (x_1, \ldots, x_n)$.

The key difference between the two perspectives stems from the nature and justification of their inductive premises and the ensuing inferences. Although these differences will be discussed extensively in what follows, it is important to highlight some of the crucial points at the outset.

Model-based induction relies on a statistical model $\mathcal{M}_\theta(\mathbf{x})$ whose inductive premises are specified in terms of testable probabilistic assumptions pertaining to a general stochastic process $\{X_t,\ t\in\mathbb{N} := (1, 2, \ldots, n, \ldots)\}$ underlying data $\mathbf{x}_0$. In particular, data $\mathbf{x}_0$ is viewed as a 'truly typical' realization of $\{X_t,\ t\in\mathbb{N}\}$, and the appropriateness of $\mathcal{M}_\theta(\mathbf{x})$ is empirically justified by testing the 'typicality' of $\mathbf{x}_0$.

Viewed from this model-based perspective, 'enumerative induction' relies on a simple (implicit) statistical model whose premises are framed in terms of a *priori* stipulations like the 'uniformity of nature' and the 'representativeness of the

sample' (Skyrms 2000). The von Mises 'collective' $\{x_k\}_{k=1}^{\infty}$ represents an infinite realization of a 'random' process $\{X_t,\ t\in\mathbb{N}\}$ that is often identified by the critics of the frequentist interpretation with the 'long-run' metaphor. Such an interpretation is shown to be inapposite for model-based induction which relies on the 'typicality' of the finite realization $\mathbf{x}_0 := \{x_k\}_{k=1}^{n}$. It is argued that charges [i]–[ii] stem primarily from misattributing to the long-run metaphor a temporal and/or a physical dimension instead of the 'repeatability' of the underlying stochastic mechanism described by $\mathcal{M}_\theta(\mathbf{x})$.

3 Probabilistic foundations of frequentist model-based inference

In this section an attempt is made to bring out the mathematical foundations of model-based induction in an effort to contrast it to the von Mises collective and thus set the stage for revisiting some of the well-known criticisms. A key result is that the notion of a statistical model in frequentist inference is a purely mathematical construct rooted in measure theory and needs no interpretation of probability.

3.1 Kolmogorov's axiomatic formulation of probability

Mathematical probability, as formalized by Kolmogorov (1933), takes the form of a probability space $(\Omega, \mathfrak{F},\ \mathbb{P}(.))$, where:

(a) Ω denotes the set of all possible distinct outcomes.
(b) $\mathfrak{F}$ denotes a set of subsets of Ω, called *events* of interest, endowed with the mathematical structure of a σ-field, i.e. it satisfies the following conditions: (i) $\Omega\in\mathfrak{F}$, (ii) if $A\in\mathfrak{F}$, then $\overline{A}\in\mathfrak{F}$, (iii) if $A_i\in\mathfrak{F}$ for $i = 1, 2, \ldots, n, \ldots$ then $\bigcup_{i=1}^{\infty} A_i\in\mathfrak{F}$.
(c) $\mathbb{P}(.)$: $\mathfrak{F}\to[0, 1]$ denotes a set function which satisfies the axioms:

[A1] $\mathbb{P}(\Omega) = 1$, for any outcomes set Ω,
[A2] $\mathbb{P}(A)\geq 0$, for any event $A\in\mathfrak{F}$,
[A3] *Countable Additivity*. For $A_i\in\mathfrak{F},\ i = 1, \ldots, n, \ldots$, such that $A_i\cap A_j = \emptyset$, for all $i\neq j,\ i, j = 1, 2, \ldots, n, \ldots$, then $\mathbb{P}(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty}\mathbb{P}(A_i)$.

This formalization places probability squarely into the mathematical field of *measure theory* concerned more broadly with assigning size, length, content, area, volume, etc. to sets; see Billingsley (1995).

The probability space $(\Omega, \mathfrak{F},\ \mathbb{P}(.))$ provides an idealized description of the stochastic mechanism that gives rise to the events of interest and related events $\mathfrak{F}$ [which is closed under the set theoretic operations of union ($\cup$), intersection ($\cap$) and complementation ($^-$)], with $\mathbb{P}(.)$ assigning probabilities to events in $\mathfrak{F}$.

Can the above Kolmogorov formalism be given an *interpretation* by assigning a *meaning* to the primitive term probability? The general thesis concerning the relationship between mathematics and empirical modeling adopted in this paper has been articulated by Cramer (1946):

> The mathematical theory belongs entirely to the conceptual sphere, and deals with purely abstract objects. The theory is, however, designed to form a model of a certain group of phenomena in the physical world, and the abstract objects and

> propositions of the theory have their counterparts in certain observable things, and relations between things. If the model is to be practically useful, there must be some kind of general agreement between the theoretical propositions and their empirical counterparts. (p. 332)

The proposed frequentist interpretation relates to a specific *objective*: modeling observable phenomena of interest exhibiting chance regularity patterns, referred to as *stochastic* (Spanos 1999). Its primary *aim* is to facilitate the task of bridging the gap between stochastic phenomena and the mathematical set up, as well as shed additional light on model-based inference, including the nature and role of error probabilities in frequentist induction. It is argued that the proposed frequentist interpretation, not only achieves this objective, but contrary to the conventional wisdom, the charges of 'circularity', its inability to assign probabilities to 'single events', and its reliance on 'random samples' are shown to be unfounded.

3.2 Random variables and statistical models

An important extension of the initial Kolmogorov formalism based on $(\Omega, \mathfrak{F}, \mathbb{P}(.))$ is the notion of a *random variable* (r.v.): a real-valued function:

$$X(.): \Omega \to \mathbb{R}, \text{ such that } \{X \leq x\} \in \mathfrak{F} \text{ for all } x \in \mathbb{R}.$$

That is, $X(.)$ assigns numbers to the elementary events in Ω in such a way so as to preserve the original event structure of interest ($\mathfrak{F}$). This extension is important for bridging the gap between the mathematical model $(\Omega, \mathfrak{F}, \mathbb{P}(.))$ and the observable stochastic phenomena of interest, because observed data usually come in the form of *numbers* on the real line.

The most crucial role of the r.v. $X(.)$ is to transform the original abstract probability space $(\Omega, \mathfrak{F}, \mathbb{P}(.))$ into a statistical model $\mathcal{M}_\theta(\mathbf{x})$ defined on the real line:

$$(\Omega, \mathfrak{F}, \mathbb{P}(.)) \xrightarrow{X(.)} \mathcal{M}_\theta(\mathbf{x}) = \{f(\mathbf{x}; \theta), \ \theta \in \Theta\}, \ \mathbf{x} \in \mathbb{R}^n_X.$$

Hence, the notion of probability associated with $\mathcal{M}_\theta(\mathbf{x})$ in (3) is purely measure-theoretic and follows directly from the axioms A1–A3 above; see Spanos (1999).

The relevant random variable underlying the traditional frequentist interpretation is defined by: $\{X = 1\} = A$, $\{X = 0\} = \overline{A}$, where $A \in \mathfrak{F}$, $\mathbb{P}(A) = p$ and $\mathbb{P}(\overline{A}) = 1 - p$, which is a Bernoulli (Ber) distributed r.v. The limiting process associated with the relative frequency interpretation requires 'repeating the experiment under identical conditions', which is framed in the form of an indexed sequence of random variables (a stochastic process) $\{X_k, \ k \in \mathbb{N} := (1, 2, \ldots n, \ldots)\}$ assumed to be IID, i.e. the statistical model is:

$$\boxed{\textbf{\textit{The simple Bernoulli model:}} \ \mathcal{M}_\theta(\mathbf{x}) : X_k \backsim \mathsf{BerIID}(\theta, \theta(1-\theta)), \ k \in \mathbb{N}.} \quad (2)$$

In general, the statistical model $\mathcal{M}_\theta(\mathbf{x})$ is viewed as a parameterization of the stochastic process $\{X_k, k\in\mathbb{N}\}$ whose probabilistic structure is chosen so as to render data $\mathbf{x}_0 := (x_1, \ldots, x_n)$ a *truly typical realization* thereof.

4 The frequentist approach and its problems

Fisher (1922) initiated a change of paradigms in statistics by recasting the then dominating Bayesian-oriented *induction by enumeration*, relying on large sample size (n) approximations (Pearson 1920), into a frequentist 'model-based induction', relying on *finite sampling distributions*, inspired by Gossett (1908) derivation of the Student's t distribution for any sample size $n > 1$. Unlike Karl Pearson who would commence with data $\mathbf{x}_0$ in search of a frequency curve to describe their histogram, he proposed to begin with (a) a prespecified model (a hypothetical infinite population), and (b) view $\mathbf{x}_0$ as a realization thereof. Indeed, he made the initial choice (specification) of the prespecified statistical model a response to the question:

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

Indeed, Fisher proposed an empirical justification for frequentist probability by emphasizing the importance of developing goodness-of-fit tests to evaluate the statistical model a posteriori:

> The possibility of developing complete and self-contained tests of goodness of fit deserves very careful consideration, since *therein lies our justification* for the free use which is made of empirical frequency formulae. (Fisher 1922, p. 314)

It should be noted that in the early 1920s the only Mis-Specification (M-S) test available was Pearson's chi-square test, which was duly employed by Fisher (1925a) to test, not only the Normality, but also the IID assumptions for discrete data. Since then numerous M-S tests have been added in the statistics literature.

The notions (a)–(b) have been formalized in purely *probabilistic* terms to define the concept of a (*parametric*) *statistical model*:

$$\boxed{\mathcal{M}_\theta(\mathbf{x}) = \{f(\mathbf{x};\theta),\ \theta\in\Theta\},\ \mathbf{x}\in\mathbb{R}^n_X,\ \text{for } \theta\in\Theta\subset\mathbb{R}^m,\ m<n,} \tag{3}$$

where $f(\mathbf{x};\theta)$, denotes the (joint) *distribution of the sample* $\mathbf{X} := (X_1, \ldots, X_n)$, whose probabilistic structure is specified with a view to render $\mathbf{x}_0$ a typical realization of the process $\{X_k,\ k\in\mathbb{N}\}$ underlying $\mathcal{M}_\theta(\mathbf{x})$.

The quintessential example of a statistical model is:

$$\boxed{\textbf{\textit{The simple Normal model:}}\ X_k \backsim \mathsf{NIID}(\mu,\sigma^2),\ k\in\mathbb{N}.} \tag{4}$$

Fisher (1925b, 1934) put forward a (frequentist) theory of optimal *estimation* almost single-handedly. Neyman and Pearson (N–P) (1933) extended/modified Fisher's significance testing framework to propose an optimal *hypothesis testing*; see Lehmann

(1986). Although the formal apparatus of the Fisher–Neyman–Pearson (F–N–P) statistical inference was largely in place by the late 1930s, the nature of the underlying *inductive reasoning* was clouded in disagreements. Fisher argued for 'inductive inference' spearheaded by his significance testing (Fisher 1955), and Neyman argued for 'inductive behavior' based on Neyman–Pearson (N–P) testing (Neyman 1956; Pearson 1955).

Indeed, several crucial **foundational problems** were left largely unanswered:

[a] the initial vs. final precision (Hacking 1965),
[b] the fallacies of acceptance and rejection (Mayo 1996),
[c] testing the adequacy of $\mathcal{M}_\theta(\mathbf{x})$ a posteriori (Fisher 1922),
[d] delineating the role of a statistical model $\mathcal{M}_\theta(\mathbf{x})$ in inductive inference,
[e] delineating the role of substantive information in statistical modeling and inference (Lehmann 1990),
[f] formulating a pertinent frequentist interpretation of probability that provides an adequate foundation for frequentist inference (Salmon 1967).

This paper focuses primarily on addressing problems [d]–[f]. For extensive discussions pertaining to problems [a]–[c] see Mayo (1996); Mayo and Spanos (2006) and Mayo and Spanos (2004). These papers are relevant for the discussion that follows because, when taken together, they demarcate what Mayo (1996) called the 'error statistical approach' which can be viewed as an refinement/extension of the F–N–P approach that offers a unifying inductive reasoning for frequentist inference. Of particular importance for the discussion that follows is the error statistical perspective on the role of frequentist probability as it relates to both the pre-data and post-data *error probabilities* associated with inductive procedures. For the error statistician probability arises, post-data, not to measure degrees of confirmation or belief in hypotheses, but to quantify how well a statistical hypothesis has passed a test. There is evidence for the statistical hypothesis or claim just to the extent that the test it passes with the data is *severe*: that with high probability the hypothesis would not have passed so well as it did if it were false, or specific flaws were present; see Mayo and Spanos (2011).

5 Error statistics and model-based inference

Empirical models in most applied fields are usually constructed using a blend of statistical and substantive information. They range from a solely data-based formulation like an ARIMA(p,d,q) model, to a entirely theory-based formulation like a structural multi-equation model in econometrics; see Spanos (2006). The majority of empirical models lie between these two extremes; see Lehmann (1990), Cox (1990).

5.1 The Duhem-Quine problem in statistical inference

There is a long tradition in the social sciences where a statistical model is specified by attaching white-noise error terms to structural (substantive) models; see Spanos (2010a). Unfortunately, foisting one's favorite theory on the data often gives rise to an estimated model that is both statistically and substantively inadequate. Worse, one has no way to delineate the two sources of error:

are the substantive claims false? or are the inductive premises misspecified?

This an instance of the *Duhem-Quine problem*: it is impossible to reliably test a substantive hypothesis (or claim) in isolation, since any statistical test used to assess this hypothesis invokes auxiliary assumptions whose validity is unknown (Mayo 1997). The widely held view concerning the theory-ladeness of observation (Skyrms 2000) made this difficult conundrum even more challenging.

Attempts to address this Duhmen-Quine problem had to overcome two difficult conundrums. The *first* had to do with formalizing the notion of the inductive premises $\mathcal{M}_\theta(\mathbf{x})$ invoked by the statistical procedures in question. The *second*, and more difficult, had to do with separating the respective roles of *statistical* vs. *substantive* information in specifying $\mathcal{M}_\theta(\mathbf{x})$; see Lehmann (1990).

5.2 A probabilistic construal of statistical models

The key to dealing with the above Duhemian ambiguity is to distinguish, *ab initio*, between *statistical* and *substantive* information and specify the statistical premises solely in terms of the former. Spanos (1999) proposed a notion of *statistical information* that relates directly to the chance regularity patterns (distribution, dependence and heterogeneity) exhibited by data $\mathbf{x}_0$. Irrespective of the substantive information (vague conjectures, bold hypotheses or claims, well-informed structural models) that led to the choice of the particular data $\mathbf{x}_0$, the later takes on 'a life of its own' when $\mathbf{x}_0$ is viewed as a realization of a *generic*—free from any substantive information—stochastic process $\{X_k,\ k\in\mathbb{N}\}$; the latter is often multivariate ($\mathbf{X}_k := (X_{1k}, X_{2k}, \ldots, X_{mk})$) but for simplicity the discussion focuses primarily on univariate processes.

Statistical model specification. The delineation/formalization of the relevant *inductive (statistical) premises* begins with a given data $\mathbf{x}_0$ and poses the question: what kind of probabilistic structure pertaining to $\{X_k,\ k\in\mathbb{N}\}$ would render data $\mathbf{x}_0$ a *truly typical realization* thereof? The notion of a truly typical realization can be understood at two different but interrelated levels. At an intuitive level this means that if one were to simulate the prespecified process $\{X_k,\ k\in\mathbb{N}\}$ on a computer it would yield several realizations, say $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N$, that exhibit the same chance regularities as that exhibited by data $\mathbf{x}_0$; the actual numbers will be different but the chance regularity patterns will be broadly the same. At a more formal empirical level, data $\mathbf{x}_0$ constitute a truly typical realization of a prespecified process $\{X_k,\ k\in\mathbb{N}\}$ only to the extent that its probabilistic assumptions can be shown—using thorough Mis-Specification (M-S) testing—to be valid for data $\mathbf{x}_0$. This defines the notion of **statistical adequacy** which provides *the* sole criterion for 'when $\mathcal{M}_\theta(\mathbf{x})$ accounts for the (recurring) regularities in data $\mathbf{x}_0$. In this sense the statistical premises underlying $\mathcal{M}_\theta(\mathbf{x})$ pertain solely to the probabilistic structure of $\{X_k,\ k\in\mathbb{N}\}$ assumed to have generated $\mathbf{x}_0$. From this purely probabilistic construal a statistical model $\mathcal{M}_\theta(\mathbf{x})$ constitutes a particular *parameterization* of the process $\{X_k,\ k\in\mathbb{N}\}$ chosen to enable one to pose the substantive questions of interest.

Example Several well-known statistical models, including (a) the multivariate Normal, (b) the Normal/Linear Regression, (c) Factor Analysis and (d) Principal

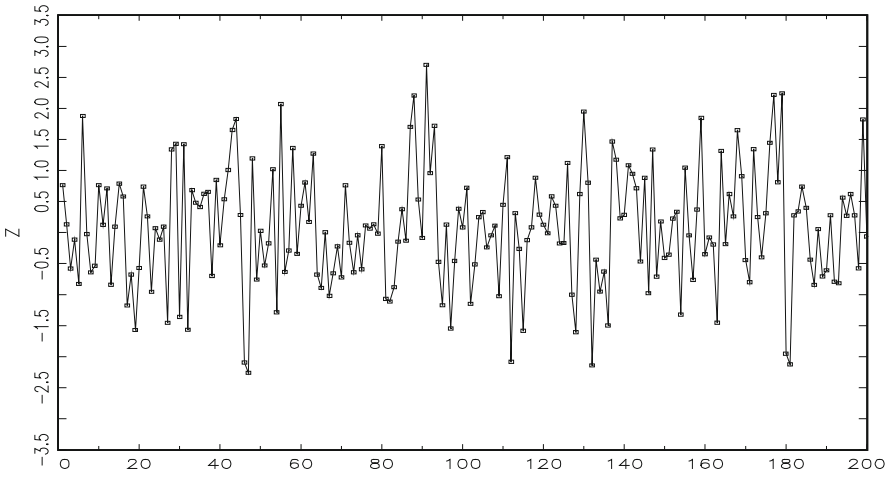

Fig. 1 A typical realization of a NIID process

Component Analysis, constitute different parameterizations of the same vector process $\{\mathbf{Z}_k,\ k\in\mathbb{N}\}$ assumed to be NIID. These different parameterizations enable one to pose very different substantive questions.

The key role of this purely probabilistic construal of statistical models stems from the fact that the validation of the statistical premises of $\mathcal{M}_\theta(\mathbf{x})$ is free of any substantive subject matter information pertaining to $\{X_k,\ k\in\mathbb{N}\}$. Both, the initial specification of $\mathcal{M}_\theta(\mathbf{x})$, as well as its validation (using M-S testing), do *not* invoke any form of such substantive information relating to $\{X_k,\ k\in\mathbb{N}\}$. They are solely based on the interplay between the chance regularities exhibited by the data in question and the formal probabilistic structure of $\mathcal{M}_\theta(\mathbf{x})$; see Spanos (2006).

To illustrate the notions of *statistical information*, *chance regularities*, *truly typical realizations* and how they can be used to select $\mathcal{M}_\theta(\mathbf{x})$ in practice, let us consider the t-plots of different data in Figs. 1, 2, 3, 4. When the data exhibit the chance regularities of Fig. 1, the simple Normal model in (4) will be appropriate; the data can be realistically viewed as a typical realization of a NIID process, and this can be formally confirmed by testing these assumptions.

In contrast, the data in Figs. 2, 3, 4, exhibit chance regularities that indicate a number of different departures from (4). Hence, if one adopts the simple Normal model for any of the data in Figs. 2, 3, 4, the estimated model will be *statistically misspecified*; this can easily confirmed using simple M-S tests; see Mayo and Spanos (2004). In particular, the data in Fig. 2 exhibit a distinct departure from Normality since the distribution chance regularity indicates a skewed distribution. The data in Fig. 3 exhibit a trending mean; a clear departure from the ID assumption. The data in Fig. 4 exhibit irregular cycles which indicate positive dependence; a clear departure from the Independence assumption.

Model-based inference. The notion of a statistical model $\mathcal{M}_\theta(\mathbf{x})$ plays a pivotal role in frequentist model-based inference because, among other things, it specifies the inductive premises of inference, delimits *legitimate* events [i.e. all well-behaved

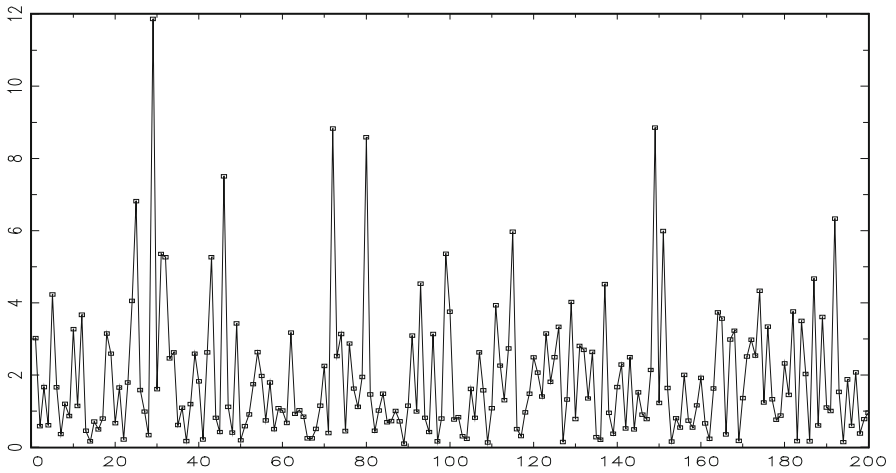

Fig. 2 A typical realization of an Exponential IID process

(Borel) functions of **X**], legitimate data, and assigns probabilities to all such events via $f(\mathbf{x};\theta)$, $\mathbf{x}\in\mathbb{R}^n_X$; see Spanos (2010c). In light of the fact that a statistic (estimator, test statistic, predictor), say $T_n = g(X_1,\ldots,X_n)$, is a Borel function of **X**, legitimate events include those generated by T_n. As a result, $\mathcal{M}_\theta(\mathbf{x})$ also determines the optimality of inference procedures because the sampling distribution of any statistic T_n is derived from $f(\mathbf{x};\theta)$ via:

$$F(t;\theta) := \mathbb{P}(T_n \leq t;\theta) = \underbrace{\int\int\cdots\int}_{\{\mathbf{x}:\, g(\mathbf{x})\leq t;\ \mathbf{x}\in\mathbb{R}^n_X\}} f(\mathbf{x};\theta)dx_1dx_2\cdots dx_n. \tag{5}$$

That is, in error statistics probability plays two interrelated roles. In addition to assigning probabilities to all legitimate events, it furnishes all relevant error probabilities associated with any statistic $T_n = g(\mathbf{Z})$ via (5). Pre-data these error probabilities quantify the generic capacity of any inference procedures to discriminate among alternative hypotheses. That is, error probabilities provide the basis for determining whether and how well a statistical hypothesis—a claim about the underlying data generating mechanism, framed in terms of an unknown parameter θ—is warranted by data $\mathbf{x}_0$ at hand. Post-data error probabilities are used to establish the warranted discrepancies from particular values of θ, using a post-data severity assessment; see Mayo and Spanos (2006).

Example In the case of the simple Normal model (4), one can use (5) to derive the following sampling distributions (Cox and Hinkley 1974):

$$\left(\overline{X}_n = \frac{1}{n}\sum_{k=1}^{n} X_k\right) \backsim \mathsf{N}\left(\mu, \frac{\sigma^2}{n}\right),\ (n-1)\,s^2 = \sum_{k=1}^{n}\left(X_k - \overline{X}_n\right)^2 \backsim \sigma^2\chi^2(n-1),$$

where '$\chi^2(n-1)$' denotes the chi-square distribution with $(n-1)$ degrees of freedom.

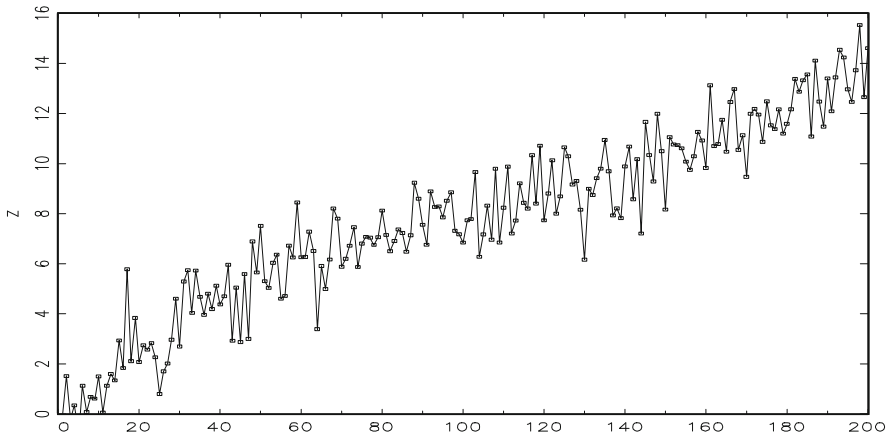

Fig. 3 A typical realization of a NI, but mean-heterogeneous process

These sampling distributions provide the relevant frequentist error probabilities used to determine the optimality of frequentist inference procedures. Such error probabilities include the Neyman–Pearson (N–P) type I [rejecting the null hypothesis when true] and II [accepting the null hypothesis when false] and Fisher's p-value [the smallest type I error probability at which the null would have been rejected with data $\mathbf{x}_0$]. Hence, in frequentist inference $f(\mathbf{x};\theta)$ provides the sole source of relevant probabilities used to calibrate the reliability of frequentist inference procedures.

The crucial role of statistical adequacy stems from the fact that it renders the relevant error probabilities ascertainable by ensuring that the *nominal* error probabilities for assessing substantive claims are approximately equal to the *actual* ones. Returning to the data in Figs. 2, 3, 4, any inference based the simple Normal model in (4) using such data will be unreliable in the sense that the actual error probabilities will *not* approximate the nominal ones; see Spanos (2009).

This is crucial for the error statistical account because a warranted inference requires the inference procedure to have low error probabilities. In light of the fact that these stem from the prespecified statistical model $\mathcal{M}_\theta(\mathbf{x})$ via (5), invalid model assumptions call into question such a warrant by inducing discrepancies between actual and nominal error probabilities. The surest way to draw an invalid inference is to apply a .05 significance level test when the actual one—due to misspecification—is closer to .99. This arises because invalid model assumptions render the distribution of the sample $f(\mathbf{x};\theta)$ erroneous, and that in turn gives rise to the wrong sampling distribution $F_n(t)$ of the relevant statistic $T_n = g(\mathbf{X})$ via (5).

5.3 **Reconciling substantive and statistical information**

Substantive subject matter information usually enters empirical modeling in a variety of forms varying in specificity from vague conjectures concerning the behavior of $\{X_k,\ k\in\mathbb{N}\}$ to well-defined *structural models,* say $\mathcal{M}_\varphi(\mathbf{x})$. In the discussion that follows we focus on the latter case where $\mathcal{M}_\varphi(\mathbf{x})$ constitutes an *estimable form* of a

substantive theory in view of the data $\mathbf{x}_0$; see Spanos (2010a). What is the connection between a statistical $\mathcal{M}_\theta(\mathbf{x})$ and a structural model $\mathcal{M}_\varphi(\mathbf{x})$?

The statistical model $\mathcal{M}_\theta(\mathbf{x})$ is built exclusively in terms of the *statistical systematic information* in data $\mathbf{x}_0$ with a view to meet two interrelated aims:

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

(II) to parameterize the probabilistic structure of $\{X_k,\ k\in\mathbb{N}\}$ in the form of an adequate statistical model $\mathcal{M}_\theta(\mathbf{x})$ in such a way so as to *embed* $\mathcal{M}_\varphi(\mathbf{x})$ in its context, via *reparametrization/restriction* $\mathbf{G}(\theta,\varphi)=\mathbf{0}$. That is, in frequentist inference substantive information is usually framed in terms of restrictions on well-defined statistical parameters; see Spanos (2006).

Error statistics provides the framework for securing these objectives by:

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

(ii) supplementing that with a *statistical Generating Mechanism* (GM) to provide a bridge between the statistical and substantive information; information that pertains to the actual data-generating mechanism.

The statistical GM is a crucial component because, as argued by Cox (1990):

> For empirical and indirect purposes, it may be enough that a model defines the joint distribution of the random variables concerned, but for substantive purposes it is usually desirable that the model can be used fairly directly to simulate data. The essential idea is that if the investigator cannot use the model directly to simulate artificial data, how can "Nature" have used anything like that method to generate real data? (p. 172)

The idea of a statistical model as a repeatable 'chance mechanism' is due to Neyman (1952) whose experience in modeling in a number of applied fields, including genetics, agriculture, epidemiology and astronomy, suggested that the Fisher's static 'infinite populations' metaphor was inadequate for the task.

Although the technical details associated with formalizing the notion of a statistical GM are beyond the scope of the present paper, it is important to note that, given a stochastic process $\{X_k,\ k\in\mathbb{N}\}$ defined on a probability space $(\Omega,\mathfrak{F},\ \mathbb{P}(.))$, it can be specified for any statistical model, under very general conditions, in the form of a decomposition:

$$X_k = E(X_k \mid \mathcal{D}_k) + u_k,\ \ k\in\mathbb{N}, \tag{6}$$

where the *systematic* $\mu_k = E(X_k|\mathcal{D}_k)$ and *non-systematic* u_k components are orthogonal, i.e. $E(\mu_k u_k)=0$. The relevant conditioning information set $\mathcal{D}_k\subset\mathfrak{F}$ is chosen so as to render $\{(u_k|\mathcal{D}_k),\ k\in\mathbb{N}\}$ a *martingale difference* process (Billingsley 1995). Intuitively, a martingale difference process represents a modern, and less restrictive, form of the older white-noise process. That is, $\mathcal{D}_k$ is chosen in such a way so as to

Table 1 Simple Normal Model

Statistical GM:		$X_k = \mu + u_k, \quad k \in \mathbb{N} := \{1, 2, \ldots\}$	
[1]	Normality:	$X_k \backsim \mathsf{N}(.,.),\ x_k \in \mathbb{R} := (-\infty, \infty),$	$\left.\begin{array}{l}\\\\\\\end{array}\right\} k \in \mathbb{N}.$
[2]	Constant mean:	$E(X_k) = \mu,$	
[3]	Constant variance:	$Var(X_k) = \sigma^2,$	
[4]	Independence:	$\{X_k,\ k \in \mathbb{N}\}$ independent process	

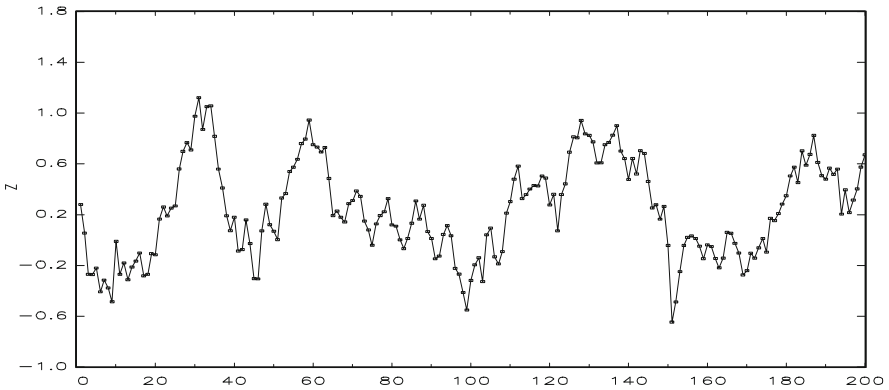

Fig. 4 A typical realization of a Normal, Markov, Stationary process

account for all the (probabilistic) systematic information in X_k, in the sense that what is left, $u_k = X_k - E(X_k \mid \mathcal{D}_k)$, is a non-systematic error process. In other words, the probabilistic concept of a martingale difference process is used to render operational the generic notion of 'non-systematicity' in delineating when an estimated model accounts for all the regularities in data $\mathbf{x}_0$.

For instance, the relevant conditioning information set for the simple Normal model (Table 1) is the trivial field $\mathcal{D}_0 = \{\Omega, \varnothing\}$, where Ω and $\varnothing$ denote the sure and impossible events, respectively, stemming from the IID assumptions. Given that $E(X_k|\mathcal{D}_0) = E(X_k)$, the decomposition in (6) yields the GM in Table 1.

To illustrate how the statistical GM will change when the probabilistic structure of $\{X_k,\ k \in \mathbb{N}\}$ changes, let us consider the case where $\{X_k,\ k \in \mathbb{N}\}$ is Normal, Markov and stationary (Spanos 1999); a typical realization of such a process in shown in Fig. 4, where the Markov dependence assumption is reflected in the cyclical patterns and the stationarity in the similarity of different portions of the graph representing sub-sets of consecutive observations $(x_{i_1}, x_{i_2}, \ldots, x_{i_m}), m < n$, of the data.

A particular parameterization of this process gives rise to the Normal, Autoregressive [AR(1)] model whose statistical GM takes the form:

$$X_k = E(X_k|\sigma(X_{k-1})) + u_k = \alpha_0 + \alpha_1 X_{k-1} + u_k, \quad k \in \mathbb{N}, \tag{7}$$

where $\sigma(X_{k-1})$ denotes the relevant 'past history' of the process $\{X_k,\ k \in \mathbb{N}\}$. The complete specification of the AR(1) model is given in Table 2; it is an appropriate statistical model for the data in Fig. 4.

Table 2 Normal, AutoRegressive [AR(1)] Model

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

Specifying a statistical model $\mathcal{M}_\theta(\mathbf{x})$ in terms of a complete set of probabilistic assumptions pertaining to the stochastic process(es) $\{X_k,\ k\in\mathbb{N}\}$ underlying data $\mathbf{x}_0$ is a necessary first step in securing *statistical adequacy* by applying thorough Mis-Specification (M-S) testing; Mayo and Spanos (2004).

In turn, statistical adequacy is a precondition for assessing *substantive adequacy*: establishing that the structural model $\mathcal{M}_\varphi(\mathbf{x})$ constitutes a *veritable explanation* of the phenomenon of interest. In the case where $\mathcal{M}_\varphi(\mathbf{x})$ is parametrically nested within $\mathcal{M}_\theta(\mathbf{x})$, statistical adequacy ensures the error-reliability of the procedure for testing the *restrictions* $\mathbf{G}(\theta,\varphi) = \mathbf{0}$ relating the two sources of information. Without it the reliability of any inference procedures used to assess the substantive information is at best unknown; see Spanos (2010d).

In summary, the key to circumventing the Duhem-Quine problem is to separate, ab initio, the statistical and substantive premises, specify $\mathcal{M}_\theta(\mathbf{x})$ exclusively in terms of the former, and secure the statistical adequacy of $\mathcal{M}_\theta(\mathbf{x})$ before assessing substantive adequacy; see Spanos (2010b).

6 A model-based frequentist interpretation

The *frequentist interpretation* articulated in this paper identifies the probability of an event A with the (probabilistic) *limit* of the relative frequency of its occurrence, $s_n = \frac{1}{n}\sum_{k=1}^{n} x_k$, in the context of a well-defined stochastic mechanism represented by the statistical model $\mathcal{M}_\theta(\mathbf{x})$. This aims to formalize the relationship between mathematical probability and 'stable long-run frequencies' that has been instinctively perceived by humans since the dawn of history.

6.1 What the SLLN does and does *not* entail

A formal justification for the frequentist interpretation as a *limit* is grounded on the SLLN which gives precise meaning to the claim 'the sequence of relative frequencies $\{s_n\}_{n=1}^{\infty}$ converges to p as $n\to\infty$'.

Borel (1909). The original SLLN asserts that for an *Independent and Identically Distributed (IID) Bernoulli* process $\{X_k,\ k\in\mathbb{N}\}$:

$$\boxed{\mathbb{P}\left(\lim_{n\to\infty}\left(\tfrac{1}{n}\textstyle\sum_{k=1}^{n} X_k\right) = p\right) = 1.} \tag{8}$$

That is, as $n \to \infty$ the stochastic sequence $\{S_n\}_{n=1}^{\infty}$, where $S_n = \frac{1}{n}\sum_{k=1}^{n} X_k$, converges to a constant p *with probability one*; this is also known as *convergence almost surely* (*a.s.*) and denoted by $S_n \overset{a.s.}{\to} p$; see Billingsley (1995). Let us clarify the notion of convergence in (8) and delineate what the result *does* and does *not* mean.

First, the SLLN is a measure-theoretic result which asserts that the probabilistic convergence in (8) holds everywhere in a domain D except on a subset D_0, the latter being *a set of measure zero* ($\mathbb{P}(D_0) = 0$). (Williams 2001, p. 111).

Second, it is well-known that $S_n \overset{a.s.}{\to} p$ does *not* involve any claims about the mathematical convergence of the sequence of numbers $\{s_n\}_{n=1}^{\infty}$ to p in a purely mathematical sense: $\lim_{n\to\infty} s_n = p$. Unfortunately, the line between probabilistic (*a.s.*) and mathematical convergence was blurred by Von Mises (1928) notion of a *collective* which was defined in terms of an infinite realization $\{x_k\}_{k=1}^{\infty}$ whose partial sums $\{s_n\}_{n=1}^{\infty}$ converge to p. It is clear that the SLLN *cannot* be invoked as a justification of the convergence condition (C) in (1). The truth of the matter is that any attempt to make rigorous the mathematical convergence $\lim_{n\to\infty} s_n = p$ is ill-fated for purely mathematical reasons:

> Trying to be 'precise' by making a *definition* out of the 'long-term frequency' idea lands us in real trouble. Measure theory gets us out of the difficulty in a very subtle way discussed in Chapter 4. (Williams 2001, p. 25)

Third, the result in (8) is essentially *qualitative*, asserting that convergence holds in the limit, but provides *no* quantitative information pertaining to the accuracy of $\frac{1}{n}\sum_{k=1}^{n} x_k$ as an approximation of $\mathbb{P}(A)$ for a given $n < \infty$. For the accuracy of the approximation one needs to use the *Law of Iterated Logarithm* (LIL):

$$\mathbb{P}\left(\lim_{n\to\infty} \sup\left[\frac{(S_n - p)/\sqrt{p(1-p)}}{\sqrt{2n\ln(\ln(n))}}\right] = 1\right) = 1, \tag{9}$$

which quantifies the *rate* of convergence of the process $\{S_n\}_{n=1}^{\infty}$; Billingsley (1995).

Fourth, the result in (8) holds when $\{X_k,\ k\in\mathbb{N}\}$ satisfies certain probabilistic assumptions, the most restrictive being IID. Their role is clearly brought out by the *converse* result, which states that:

For any IID process $\{X_k,\ k\in\mathbb{N}\}$, if $S_n \overset{a.s.}{\to} p$ then $E(X_k)$ exists and is equal to p.

NOTE that for the Bernoulli distribution $p = \mathbb{P}(X_k = 1)$ but also $p = E(X_k)$.

This suggests that from a modeling perspective, the SLLN is essentially an *existence result* for 'stable (constant) relative frequencies' ($S_n \overset{a.s.}{\to} p$), in the sense that it specifies *sufficient* conditions for the process $\{X_k,\ k\in\mathbb{N}\}$ to be amenable to statistical modeling and inference; see Cramer (1946); Neyman (1952). That is, when no stable relative frequencies exist, the phenomenon of interest is beyond the scope of statistical modeling. Put another way, when the (sufficient) probabilistic assumptions for the process $\{X_k,\ k\in\mathbb{N}\}$ do not hold for a particular data $\mathbf{x}_0$, they cannot be used to provide the basis for inductive inference about the underlying phenomenon of interest because $\mathbf{x}_0$ do not exhibit sufficient statistical regularities to be amenable to statistical modeling and inference.

Table 3 Simple Bernoulli Model

Statistical GM:	$X_k = \theta + u_k, \quad k \in \mathbb{N}.$	
[1]	Bernoulli:	$X_k \backsim \mathsf{Ber}(.,.), x_k = 0, 1,$
[2]	constant mean:	$E(X_k) = \theta,$
[3]	constant variance:	$Var(X_k) = \theta(1-\theta),$
[4]	Independence:	$\{X_k, \ k \in \mathbb{N}\}$is an independent process

6.2 **Revisiting the 'long-run' metaphor**

The *long-run* metaphor associated with the frequentist interpretation, anchored on the SLLN in (8), envisions repeating the stochastic mechanism represented by an IID Bernoulli process in (2), so that the relative frequency process $\{S_n\}_{n=1}^{\infty}$ (a.s.)-approximates p as $n \to \infty$. This metaphor enables one to conceptualize the frequentist interpretation of probability by elucidating the connection between the stochastic data-generating mechanism associated with $\mathcal{M}_\theta(\mathbf{x})$ and the assignment of probabilities to all legitimate (generic) events of interest, including 'single events', as well as events associated with sampling distributions as given in (5), like the type I and II errors. In this sense, the metaphor was never intended to replace $f(\mathbf{x};\theta)$ as *the* relevant tool for assigning probabilities (measure) to any legitimate event of interest associated with $\mathcal{M}_\theta(\mathbf{x})$. The metaphor's primary objective is to enhance our intuitive understanding of the connection between mathematical and empirical probabilities. This is achieved by conceptualizing the bridging of this gap via repeating $\mathcal{M}_\theta(\mathbf{x})$ to generate a large enough, but finite, number of sample realizations.

A. The single event probability charge

> Despite claims to the contrary: According to the frequency interpretation's official definition, however, the probability concept is meaningful only in relation to infinite sequences of events, not in relation to single events. (Salmon 1967, p. 90)

the assigning of probabilities to legitimate (generic) events in model-based induction does *not* invoke the 'long-run' metaphor. Indeed, there are no conceptual or technical difficulties in assigning a probability to single events such as: $A_{k+1} = \{X_{k+1} = 1\}$—'heads' on the next toss of the coin, in the context of the Bernoulli model $\mathcal{M}_\theta(\mathbf{x})$ (Table 3):

$$\mathbb{P}(A_{k+1}) = \mathbb{P}(X_{k+1} = 1) = \theta, \quad \text{for any } k = 1, 2, \ldots$$

In light of this, the only way to explain the persistence of this charge in the philosophy of science literature is in terms of misidentifying the frequentist interpretation of probability with von Mises's variant and its allusion to infinite realizations $\{x_k\}_{k=1}^{\infty}$. Indeed, the above quotation from Salmon reads like a paraphrasing of Von Mises (1928) original claim:

> It is possible to speak about probabilities only in reference to a properly defined collective. (p. 28)

If one replaces the word 'collective' with 'statistical model $\mathcal{M}_\theta(\mathbf{x})$' in this quotation, the single event probability problem vanishes in model-based induction.

B. Rendering the long-run metaphor operational

The misidentification of the frequentist interpretation with von Mises's variant is also examplified in the quotation by Howson and Urbach (2006) in Sect. 1, who go even further by invoking the 'physical apparatus' in generating the collective $\{x_k\}_{k=1}^{\infty}$.

When viewed in the context of model-based induction, however, the defining characteristic of the long-run metaphor is *neither* the *temporal* nor the *physical* dimension, but the *repeatability* (in principle) of the data-generating process represented by $\mathcal{M}_\theta(\mathbf{x})$. This idealized model-based mechanism is, by definition, hypothetical, but justifiable on empirical grounds, as clearly articulated by Neyman (1977):

> Guessing and then verifying the 'chance mechanism', the repeated operations of which produces the observed frequencies. ... Naturally, the guessed chance mechanism is hypothetical. (p. 99)

As argued above, the question of how $\mathcal{M}_\theta(\mathbf{x})$ relates to the 'physical data-generating mechanism' pertains to the issue of *substantive* not statistical *adequacy*. Moreover, the repeatability (in principle) renders the long-run metaphor operational for every statistical model. In practice, one can always simulate (using a computer) the statistical GM of a particular $\mathcal{M}_\theta(\mathbf{x})$ to generate numerous finite realizations (however large) of the process $\{X_k,\ k\in\mathbb{N}\}$ in nanoseconds.

Example In the case of the simple Normal model (Table 1), the statistical GM:

$$\boxed{X_k = \mu + \sigma\varepsilon_k,\ \varepsilon_k \backsim \mathsf{N}(0,1),\ \ k = 1, 2, \ldots, n,} \tag{10}$$

where $\varepsilon_k \backsim \mathsf{N}(0, 1)$ denotes *pseudo-random* numbers from the standard Normal distribution, can be used to simulate the long-run metaphor for a given set of values of (μ, σ^2), by applying the following algorithm.

Step 1: Specify values for (or estimates of) the unknown parameters $\theta := (\mu, \sigma^2)$.

Step 2: Generate, say $N = 10000$, realizations of sample size, say $n = 100$, of the process $\{\varepsilon_k, k = 1, 2, \ldots, N\}$ to create $\left(\varepsilon^{(1)}, \varepsilon^{(2)}, \ldots, \varepsilon^{(N)}\right)$, where $\varepsilon^{(k)} := (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)^\top$ represents a draw of n pseudo-random numbers from $\mathsf{N}(0, 1)$.

Step 3: Substitute sequentially each $\varepsilon^{(k)}$ into the GM: $\mathbf{x}^{(k)} = \mathbf{1}\mu + \sigma\varepsilon^{(k)}$, for $\mathbf{1} := (1, \ldots, 1)^\top$, to generate the simulated data: $\mathbf{X}_N := (\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(N)})$, $\mathbf{x}^{(k)} := (x_1, \ldots, x_n)^\top$.

Using the simulated data $\mathbf{X}_N$ one can construct the empirical counterpart to any probabilistic assignment $\mathbb{P}(A)$, for any legitimate event A: any well-behaved (Borel)

function of $\mathbf{X}$. It is important to emphasize that these legitimate events include those defined by any statistic $T_n = g(X_1, \ldots, X_n)$, and thus the probabilistic assignments include the sampling distribution of any statistic of interest like $\widehat{\theta}_n := (\overline{X}_n, s^2)$ and $\tau(\mathbf{X}) = \sqrt{n}(\overline{X}_n - \mu_0)/s$ derived via (5).

These empirical sampling distributions can be used to render operational, not only the pre-data error probabilities like the type I–II as well as the power of a test for different discepancies from the null, but also the post-data probabilities associated with the severity evaluation that provides an evidential interpretation of frequentist inference; see Mayo and Spanos (2006).

6.3 Interpretative provisions for the proposed interpretation

It is important to emphasize that, by themselves, mathematical results, such as the SLLN (8) and the LIL (9), do not suffice to provide an apposite frequentist interpretation that addresses the foundational problems pertaining to frequentist *inductive reasoning*. Statistical induction requires a pertinent link between the mathematical framework and the data-generating mechanism giving rise to $\mathbf{x}_0$. In model-based inference this link takes the form of the *interpretive provisions*:

[i] data $\mathbf{x}_0 := (x_1, x_2, \ldots, x_n)$ is viewed as a 'truly typical' realization of the process $\{X_k,\ k \in \mathbb{N}\}$ specified by the statistical model $\mathcal{M}_\theta(\mathbf{x})$, and

[ii] the 'typicality' of $\mathbf{x}_0$ (e.g. IID) can be assessed using trenchant Mis-Specification (M-S) testing.

This link can be used to refute another charge leveled against the anchoring of the frequentist interpretation on the SLLN. Howson and Urbach (2006), p. 48, claim:

> ... it is not difficult to see that in itself it [the SLLN] explains nothing at all, let alone sanctions the identification of probabilities with observed relative frequencies, since no empirical meaning has yet been given to the probability function P.

The truth is that the SLLN in conjunction with [i]–[ii] *does* provide a direct *link* to what Cramer (1946) calls the 'empirical counterpart' of $F(x) = \mathbb{P}(X \leq x)$ - the cumulative distribution function (cdf). The link comes in the form of the Glivenko-Cantelli (G-C) lemma, which states that under the same assumptions as the SLLN, the *empirical cdf* defined by $\widehat{F}_n(x) = \frac{[\text{ no. of } x_k s \leq x]}{n} x \in \mathbb{R}$, *converges uniformly* to the 'true' $F(x)$ (Williams 2001, p. 116):

$$\mathbb{P}\Big(\sup_{-\infty < x < \infty} |\widehat{F}_n(x) - F(x)| \to 0\Big) = 1, \quad \text{for every } x \in \mathbb{R}. \tag{11}$$

This is a fundamental result in statistical modeling and inference that provides the foundation for several highly-prized techniques such as resampling (including the bootstrap) and nonparametric procedures; see Wasserman (2006).

6.4 Revisiting the circularity charge

In explicating the SLLN one might invoke common sense intuition to describe the result in (8) as saying that the relative frequency of occurrence of event A converges to $\mathbb{P}(A) = p$, or fluctuates closer and closer around p as n increases. This intuition is often the source of the charge that the justification of the frequentist interpretation of probability invoking (8) is *circular*. For example, Lindley (1965), p. 5, argues:

> ... there is nothing impossible in $\frac{m}{n}$ differing from p by as much as 2ϵ, 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.

This charge of circularity is denied by notable probabilists like Renyi (1970), who argues that the charge stems from conflating the intuitive description of the convergence of relative frequencies with the mathematical result itself:

> Bernoulli's law of large numbers, on the other hand, is a theorem deduced from the mathematical concept of probability; there is thus no vicious circle. (p. 159)

Although Renyi is referring to the Weak Law of Large Numbers (Williams 2001), his comment is even more pertinent to the SLLN in (8). Elaborating on his last sentence, the SLLN is an *existence* result for 'stable relative frequencies' (convergence to a constant p), whose assertions rely exclusively on the Kolmogorov mathematical formalism. Indeed, a closer look at the word 'unlikely' that Lindley argues renders the argument circular, shows that the SLLN refers to the convergence of $\{S_n\}_{n=1}^{\infty}$ [not $\{s_n\}_{n=1}^{\infty}$], which involves the purely measure-theoretic notion of a set of measure zero; no hidden invocation of frequentist probability lurks behind (8).

Given that (8–9) are purely measure-theoretic results, the circularity charge is clearly misplaced. Why, then, do critics keep reiterating this charge?

One possible explanation might be that these critics consider the 'long-run frequency' *itself* as providing the link between mathematical and empirical probabilities. Howson and Urbach (2006), pp. 48–49, claim that:

> Some additional assumption linking statements of the probability calculus to physical reality is, ..., *indispensable*. We have made long-run frequency a definitional link, because that is the simplest procedure that avoids the objections.

It might be the simplest, but not the most appropriate. Kolmogorov (1963) challenged the pertinence of such a link by arguing that the notion of the long-run:

> ... does not contribute anything to substantiate the application of the results of probability theory to real practical problems where we always have to deal with a finite number of trials. (p. 369)

As argued above, the link between mathematical results, such as (8)–(9) and (11), and the data-generating mechanism comes in the form of the interpretive provisions [i]–[ii], focusing on the initial segment $\mathbf{x}_0$ by viewing it as a 'truly typical' realization of the process $\{X_k,\ k\in\mathbb{N}\}$.

What is particularly interesting from this interpretative perspective is that the frequentist interpretation proposed above shares the provisions [i]–[ii] with a completely different *algorithmic perspective* based on *Kolmogorov complexity*. This algorithmic perspective is relevant to the issues discussed in this paper for several reasons.

First, the algorithmic perspective can be used to shed additional light on why von Mises's frequentist interpretation based on the notion of a collective was ill-fated by clarifying the Wald (1937) and Church (1940) attempts to define admissible subsequences, and demonstrated by Ville (1939) to violate the LIL (9) (Li and Vitanyi 2008, pp. 49–56).

Second, the algorithmic perspective, being non-probabilistic in nature, can be used to elucidate certain confusions relating to charges leveled against the frequentist interpretation by summoning infinite realizations $\{x_k\}_{k=1}^{\infty}$, a well as dispel any lingering doubts concerning the circularity charge.

Third, the algorithmic shares with the error-statistical perspective the same notion of 'randomness' relating to the presence of 'chance regularities' exhibited by finite realizations $\mathbf{x}_0 := \{x_k\}_{k=1}^{n}$ of the processes $\{X_k,\ k\in\mathbb{N}\}$. This is in contrast to the von Mises notion relating to the absence of *predictability* in the context of infinite realizations $\{x_k\}_{k=1}^{\infty}$.

6.5 Kolmogorov complexity: an algorithmic perspective

Kolmogorov complexity provides a purely *non-probabilistic* rendering to the frequentist interpretation that instantiates all the above measure-theoretic results:

> ... algorithmic information theory is really a constructive version of measure (probability) theory. (Chaitin 2001, p. vi)

The algorithmic complexity perspective provides a *non-probabilistic* interpretation to infinite realizations of IID processes $\{x_k\}_{k=1}^{\infty}$ by focusing on the *effective computability* and *incompressibility* of its finite initial segment $\mathbf{x}_0 := \{x_k\}_{k=1}^{n}$. A particular finite sequence $\{x_k\}_{k=1}^{n}$ is 'algorithmically incompressible' iff the shortest program which will output $\mathbf{x}_0$ and halt is about as long as $\mathbf{x}_0$ itself. Incompressible sequences (strings) turn out to be indistinguishable, by any computable and measurable test, from typical realizations of IID Bernoulli processes, and vice versa. Hence, incompressible sequences provide a model of the most basic sort of probabilistic process which can be defined without any reference to probability theory; see Salmon (1984). Indeed, the complexity framework can be used to characterize:

> random infinite sequences as sequences all of whose initial finite segments pass all effective randomness tests (Li and Vitanyi 2008, p. 56).

Moreover, these tests rely on non-probabilistic (algorithmic) notions of partial recursive functions and incompressibility.

The key to the duality between the stochastic and algorithmic perspectives is provided by (Li and Vitanyi 2008, p. 146):

> Martin-Löf (1969) important insight that to justify any proposed definition of randomness one has to show that the sequences that are random in the stated

sense satisfy the several properties of stochasticity we know from the theory of probability.

The Kolmogorov complexity framework provides an operational algorithmic interpretation to all the above measure-theoretic results, including non-typical realizations defined on a set of measure zero; see Nies (2009).

In addition, the notion of Kolmogorov complexity provided the missing link between von Mises notion of randomness relying on infinite realizations $\{x_k\}_{k=1}^{\infty}$, and the above (error-statistical) stochastic view. This link relies on the initial finite segment $\{x_k\}_{k=1}^{n}$ being 'typical', i.e. passing all effective randomness tests, giving rise to the notion of *pseudo-randomness*: sequences that exhibit statistical randomness while being generated by a deterministic recursive process. This provided the first successful attempt to operationalize randomness, by ensuring the compliance of algorithmically incompressible sequences to the above measure theoretic results, including the SLLN (8), the LIL (9) and the G-C lemma (11).

In summary, the error-statistical stochastic view and the algorithmic perspective based on Kolmogorov complexity, despite being grounded on entirely different mathematical formulations, share several features and give rise to complementary interpretations of probability that do *not* contravene the measure-theoretic results; in contrast to the von Mises frequentist interpretation.

6.6 Frequentist interpretation: an empirical justification

As mentioned above, the statistical model underlying Borel's SLLN is the simple Bernoulli model $\mathcal{M}_\theta(\mathbf{x})$ in (2). A more explicit way to specify this model is given in Table 3, in term of a statistical GM and assumptions [1]–[4]. The validity of these assumptions vis-à-vis data $\mathbf{x}_0$ is what secures the reliability of any inference concerning θ, including the SLLN.

Viewing the 'stable long-run frequency' idea in the context of the error statistical perspective, it becomes apparent that there is *nothing stochastic* about a particular data $\mathbf{x}_0 := \{x_k\}_{k=1}^{n}$ when viewed as a realization of the process $\{X_k,\ k\in\mathbb{N}\}$. Data $\mathbf{x}_0$ denotes a set of numbers that exhibit certain chance regularity patterns *reflecting* the probabilistic structure of the underlying process $\{X_k,\ k\in\mathbb{N}\}$. From this perspective 'randomness' is firmly attached to $\{X_k,\ k\in\mathbb{N}\}$ and is only reflected in data $\mathbf{x}_0$. In light of that, the only relevant question for the data is whether the chance regularity patterns exhibited by $\mathbf{x}_0$ reflect 'faithfully enough' the probabilistic structure presumed for $\{X_k,\ k\in\mathbb{N}\}$, i.e. whether $\mathbf{x}_0$ constitutes a 'truly typical realization' of this process. If the process $\{X_k,\ k\in\mathbb{N}\}$ is IID, almost all its realizations for a large enough n will be 'non-systematic' in the sense of being 'truly typical' of an IID process. Equivalently, the set of 'systematic' (non-typical) realizations of $\{X_k,\ k\in\mathbb{N}\}$, say D_0, will be a set of measure zero. That is, for a large enough n the IID structure of $\{X_k,\ k\in\mathbb{N}\}$ renders 'almost impossible' non-typical (systematic) realizations such as:

$$\begin{aligned}\{x_k\}_{k=1}^{n} &= \{0, 0, \ldots, 0\},\\ \{x_k\}_{k=1}^{n} &= \{1, 1, \ldots, 1\},\\ \{x_k\}_{k=1}^{n} &= \{1, 0, 1, 0 \ldots, 1, 0\},\ \text{etc.}\end{aligned}$$

But how would one know that the particular realization $\mathbf{x}_0$ in hand is *not* 'truly typical'? The 'non-typicality' of a particular realization can be easily detected, in the form of departures from the IID assumptions, using simple *Mis-Specification (M-S) tests,* like a runs test, which rely solely on mathematical combinatorics; see Spanos (1999). As argued above, such M-S tests can be grounded exclusively on non-probabilistic notions of (pseudo) randomness associated with algorithmic incompressibility; see Li and Vitanyi (2008). That is, statistical adequacy ensures the meaningfulness of identifying the limit of the relative frequencies $\{s_n\}_{n=1}^{\infty}$ with the probability p by invoking (8). Given that the probabilistic assumptions [1]–[4] are testable vis-à-vis data $\mathbf{x}_0$, the frequentist interpretation is justifiable on *empirical*, not a priori, grounds.

7 Enumerative vs. model-based induction

In relation to induction, Salmon (1967) credits Reichenbach (1934) with two important contributions:

> a theory on inferring long run frequencies from very meagre statistical data, and a theory for reducing all inductions to just such inferences. (Hacking 1968, p. 44).

With respect to 'inferring long-run frequencies' Salmon argues that Reichenbach was the first to supplement the frequentist interpretation with a 'Rule of Induction by Enumeration':

> Given that $s_n = \frac{m}{n}$, to infer that: $\lim_{n\to\infty} s_n = \frac{m}{n}$. (p. 86)

The primary justification for this rule is that asymptotically (as $n \to \infty$) s_n converges to the true probability p. However, this claim is misleading because it does not follow from the SLLN. In addition to being framed in terms of probabilistic convergence, the SLLN implies that for any value $(\frac{m}{n}) \in [0, 1]$, $\mathbb{P}\left(\lim_{n\to\infty} S_n(\omega) = \frac{m}{n}\right) = 0$, unless $\frac{m}{n} = p$, which can only happen on a set of measure zero. To address that problem, the limiting process should read: $\lim_{n\to\infty} S_n(\omega) = p$ with probability 1.

Hacking (1968) questioned the justification of the straight rule on asymptotic grounds, and proposed an axiomatic justification in terms of properties like additivity, invariance and symmetry. He went as far as to suggest a return to the approximate form of the rule, $s_n \pm \varepsilon$, originally proposed by Reichenbach (1934), and argued for codifying the error ε in terms of de Finetti's subjective interpretation of probability.

A closer look at this literature reveals that the SLLN has been invoked, implicitly or explicitly, for two different, but related, tasks:

[a] in justifying the frequentist interpretation by providing a link between relative frequencies and mathematical probabilities, and
[b] in justifying the use of the straight rule $\mathbb{P}(A) = \frac{m}{n}$ as an inductive procedure for learning from data.

The model-based frequentist interpretation of probability proposed in this paper draws a clear distinction between [a] and [b] because their respective justifications are different. Focusing on the latter, none of the above proposals provides an adequate

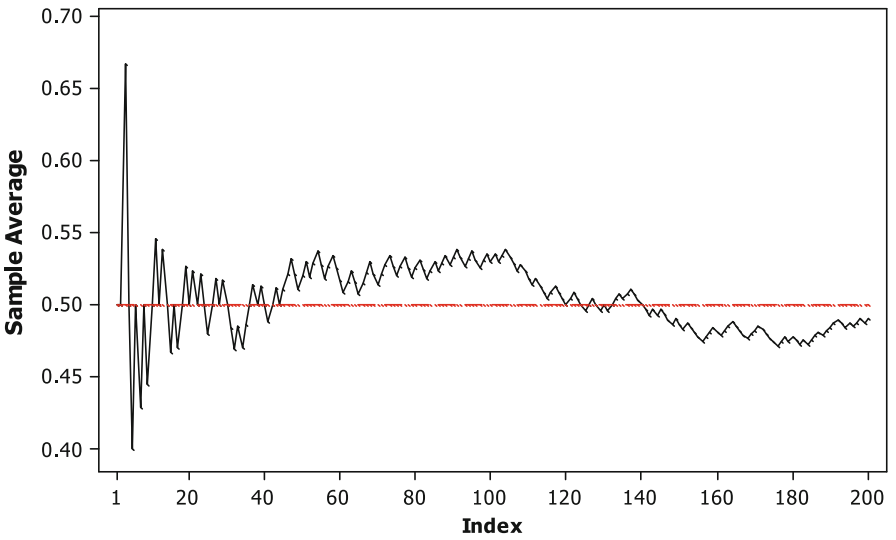

Fig. 5 t-plot of $\overline{x}_n$ for a Bernoulli IID realization with $n = 200$

justification for the straight rule $\mathbb{P}(A) = \frac{m}{n}$ as an inferential procedure. What has not been sufficiently appreciated in these discussions is that the SLLN holding for the stochastic process $\{X_k,\ k\in\mathbb{N}\}$ underlying data $\mathbf{x}_0$ ensures [a], but it does *not* necessarily secure [b] that concerns the nature and justification of model-based inductive inference. This key difference brings out the role of a statistical model $\mathcal{M}_\theta(\mathbf{x})$ in securing the *reliability* and *precision* of inference.

From the error statistical perspective the relevant statistical model, implicit in the discussions of Borel's SLLN is the simple Bernoulli $\mathcal{M}_\theta(\mathbf{x})$ (Table 3) with $\mathbb{P}(A) = \theta$. Viewing the straight rule $\mathbb{P}(A) = \frac{m}{n}$ in the context of $\mathcal{M}_\theta(\mathbf{x})$ reveals that one knows much more about $s_n = \frac{m}{n}$ as an *estimate* of θ than the SLLN indicates. The SLLN asserts that $S_n = \overline{X}_n = \frac{1}{n}\sum_{k=1}^{n} X_k$ is a *strongly consistent* estimator of θ. Does that secure the reliability of using $\overline{X}_n$ to draw inferences about θ? Not necessarily!

The SLLN secures only *minimal reliability* in the sense that the result in (8) is necessary, but not sufficient for the reliability of inference for a given n. This is because the SLLN holding secures potential (as $n \to \infty$) *learning from data* about the unknown parameter θ, for a sufficiently large n, in the sense that s_n will pinpoint the true θ. That learning, however, is not guaranteed for a particular n associated with data $\mathbf{x}_0$. This is demonstrated in Figs. 5, 6 where $n = 200$ is not sufficiently large for $\overline{x}_n$ to zero-in on the true θ, but $n = 1000$ comes very close. Since what is large enough depends on both n and the true θ, one cannot answer univocally the question how large is 'large enough'.

To secure the reliability and precision of any frequentist inference procedure, including estimation, testing and prediction, one needs to render the SLLN's *potential* into *actual learning* based on the given n associated with data $\mathbf{x}_0$, by evaluating the relevant error probabilities. That usually requires strengthening the probabilistic premises because the assumptions needed for the SLLN to hold are usually *weaker* than those

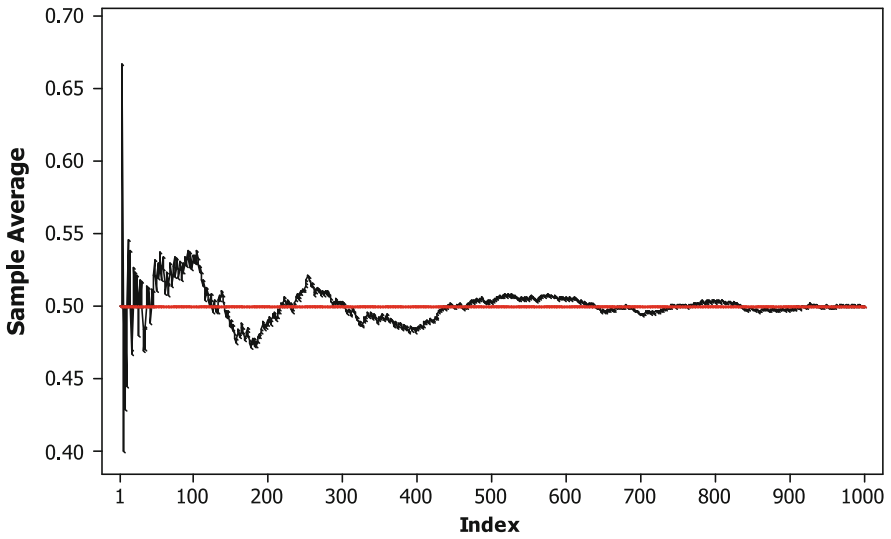

Fig. 6 t-plot of $\bar{x}_n$ for a Bernoulli IID realization with $n = 1000$

needed to secure the reliability and precision of inductive inferences based on a particular data $\mathbf{x}_0$. The weaker premises, however, come at a price when the SLLN is used as a basis of inference because any attempt to evaluate the relevant error probabilities will give rise to very crude and often inaccurate estimates. This stems from the asymptotic nature of the result which necessitates the use of inequality bounds such as *Hoeffding's* (Wasserman 2006):

$$\mathbb{P}\left(|\overline{X}_n - \theta| \geq \varepsilon\right) \leq 2\exp\left(-2n\varepsilon^2\right), \quad \text{for any } \varepsilon > 0. \tag{12}$$

For instance, if one were to use Borel's SLLN to evaluate the tail area of the sampling distribution of $\overline{X}_n$ beyond two standard deviations in the case where $\theta = .5, \varepsilon = .1$ and $n = 100$ (note that $\varepsilon = 2\sqrt{Var(\overline{X}_n)}$), the answer will be:

$$\mathbb{P}\left(|\overline{X}_n - \theta| \geq .1\right) \leq .271. \tag{13}$$

How accurate is this upper bound? The inaccuracy of (13) when used to evaluate error probabilities can be brought out by making full use of the model assumptions [1]–[4] (Table 3) to derive the finite sampling distribution:

$$\overline{X}_n \backsim \mathsf{Bin}(\theta,\ [\theta(1-\theta)/n]), \tag{14}$$

where 'Bin' denotes the Binomial distribution. Evaluation of the same tail area using (14) yields:

$$\mathbb{P}\left(|\overline{X}_n - \theta| \geq .1\right) = .0455, \tag{15}$$

which is considerably smaller than the upper bound; the sixfold difference between the *actual* error probability in (15) and the *evaluated* upper bound in (13) can easily lead an inference astray; see Spanos (1999) for several such examples. Note that the inaccuracy of these upper bounds worsens as n decreases.

It is very important to emphasize that although (in principle) both Borel's SLLN in (8) and the finite sampling distribution in (14) depend on exactly the same set of probabilistic assumptions ([1]–[4] in Table 3), the SLLN essentially *ignores* the Bernoulli distribution assumption [1]. However, this assumption is totally innocuous because in cases where there are only two outcomes it holds by definition. Hence, in this particular case the statistical adequacy of the simple Bernoulli model justifies both results. More generally, however, the probabilistic premises invoked by the SLLN and other limit theorems are weaker than those of a well-defined statistical model $\mathcal{M}_\theta(\mathbf{x})$ primarily because they make no direct distributional assumptions; see Spanos (1999). What has not been sufficiently appreciated in these discussions is how valuable a fully-specified statistical model $\mathcal{M}_\theta(\mathbf{x})$ can be in enhancing both the reliability and precision of inference when its statistical adequacy is secured. Inductive learning is assured by reliable and precise model-based inference.

The above discussion suggests that a better way to understand the straight rule as an inference procedure in the context of model-based inference is *not* in terms $s_n = \frac{m}{n}$ being a good *estimate* of θ by invoking the SLLN, but in terms of $\overline{X}_n$ being a good *estimator* of θ, with the relevant error probabilities provided by the finite sampling distribution (14). The latter distribution suggests that $\overline{X}_n$ is a good estimator, not just because it is a strongly consistent estimator of θ (stemming from the SLLN), but because it is also unbiased and fully efficient; see Cox and Hinkley (1974). Hence, when viewed in this context the key problem with induction by enumeration based on the straight rule is that, even in this revamped form, it does not extend beyond the simple Bernoulli model. Moreover, the proposed interpretation precludes the stance known as *finite frequentism* which attaches probabilities to events in a finite reference class using the straight rule (Hajek 2007).

Returning to model-based inference, (14) can be used to construct a $(1-\alpha)$ Confidence Interval:

$$\mathbb{P}\left(\overline{X}_n - c_{\frac{\alpha}{2}} s_{\overline{x}} \leq \theta \leq \overline{X}_n + c_{\frac{\alpha}{2}} s_{\overline{x}}\right) = 1-\alpha,$$

where $s^2_{\overline{x}} = \left(\overline{X}_n(1-\overline{X}_n)\right)/n$, which provides a proper frequentist interpretation to Reichenbach's approximate straight rule:

$$\overline{X}_n \pm \varepsilon_n, \quad \text{where } \varepsilon_n = c_{\frac{\alpha}{2}} s_{\overline{x}}.$$

The difference is that one can assess the reliability as well as the precision of this rule using the relevant error probabilities based on (14).

In summary, the SLLN specifies *sufficient* probabilistic assumptions pertaining to the process $\{X_k,\ k\in\mathbb{N}\}$ underlying data $\mathbf{x}_0$ to be amenable to statistical modeling and inference by ensuring the existence of stable relative frequencies. To ensure the reliability and precision of model-based inductive inference, however, one often needs to

go the extra mile and strengthen these premises to a fully-specified statistical model $\mathcal{M}_\theta(\mathbf{x})$ whose statistical adequacy for data $\mathbf{x}_0$ has been secured.

8 The 'random sample' charge revisited

Some critics claim that the frequentist interpretation of probability can only be applied to cases where the data $\mathbf{x}_0$ constitute a realization of a random sample $\mathbf{X}$, and thus it is inextricably bound up with the IID assumptions; see Fine (1973). It is fair to say that, historically speaking, the IID assumptions were implicitly imposed on the overwhelming majority of statistical modeling before the 1930s. Moreover, the IID assumptions appear to constitute an integral part of Von Mises (1928) frequentist interpretation, being reflected in his condition of 'invariance under place selection' for *admissible collectives* $\{x_k\}_{k=1}^{\infty}$; see (1). In contrast, the proposed frequentist interpretation anchored on the SLLN does *not* require such restrictive probabilistic assumptions imposed on the underlying process $\{X_k,\ k\in\mathbb{N}\}$.

Beginning in the 1930s, the literature on *stochastic processes* has greatly extended the intended scope of statistical modeling by a gradual weakening of the IID assumptions and the introduction of probabilistic notions of dependence and heterogeneity; see Doob (1953). This broadening brought about a shift away from the original von Mises notion of randomness. Kolmogorov (1983) reflecting on this argued:

> ... we should have distinguished between randomness proper (as absence of any regularity) and stochastic randomness (which is the subject of probability theory). There emerges the problem of finding reasons for the applicability of the mathematical theory of probability to the phenomena of the real world. (p. 1)

Von Mises randomness and the accompanying *unpredictability* of infinite sequences (impossibility of a gambling system), has been replaced by 'stochastic' randomness, reflected by the 'chance regularities' exhibited by finite realizations of processes that can be used to enhance statistical predictability; see Spanos (1999). This motivated the notion of 'truly typical' realization, which can be easily extended to non-IID processes. The only restriction on the latter is that they retain a form of *k-invariance* encapsulating the *unvarying features* of the phenomenon being modeled in terms of the unknown parameter(s) θ. For instance, assuming that the process $\{y_t,\ t\in\mathbb{N}\}$ is Normal, Markov and mean-heterogeneous, but covariance stationary, gives rise to an Autoregressive statistical model whose Generating Mechanism (GM) takes the form:

$$y_t = \beta_0 + \overbrace{\sum_{k=1}^{m}\beta_k t^k}^{\text{heterogeneity}} + \overbrace{\sum_{i=1}^{p}\alpha_i y_{t-i}}^{\text{dependence}} + u_t,\ \ t\in\mathbb{N},$$

where $\theta := (\beta_0, \beta_1, \ldots \beta_m, \alpha_1, \alpha_2, \ldots, \alpha_p, \sigma^2)$ are t-invariant. Indeed, the primary reason for defining $\mathcal{M}_\theta(\mathbf{x})$ in terms of the *joint* distribution, $f(\mathbf{x};\theta)$—and not the marginal $f(x_k;\theta)$—is exactly to be able to account for the dependence/heterogeneity in non-IID samples; a key result first established by Kolmogorov (1933). In relation to this, it will be rather anachronistic to criticize Von Mises (1928) for ignoring the

developments in stochastic processes in the 1930s, but the same courtesy cannot be extended to his advocates and critics in modern philosophy of science.

Since Borel (1909) the sufficient probabilistic assumptions on the process $\{X_k,\ k\in\mathbb{N}\}$ giving rise to the SLLN result in (8) have been weakened considerably. In particular, the SLLN, as it relates to the frequentist interpretation of probability, has been extended in two different, but interrelated, directions. *First* the result was proved to hold for processes considerably more sophisticated than BerIID, dropping the distributional assumption altogether and allowing for certain forms of non-IID structures such as $\{X_k,\ k\in\mathbb{N}\}$ being a heterogeneous Markov or a martingale process. *Second* the result has been extended from the linear function $S_n=\frac{1}{n}\sum_{k=1}^{n}X_k$, to any Borel function of the sample, say $Y_n=h(X_1,X_2,\ldots,X_n)$; Billingsley (1995).

For a general statistical model $\mathcal{M}_\theta(\mathbf{x})$ based on a non-IID sample, the assignment of the probabilities using $f(\mathbf{x};\theta)$, $\mathbf{x}\in\mathbb{R}^n_X$ depends crucially on being able to estimate consistently the unknown parameter(s) θ. Indeed, the constancy of the parameters θ renders possible the estimation of 'stable relative frequencies' associated with $f(\mathbf{x};\theta)$. Hence, in the context of $\mathcal{M}_\theta(\mathbf{x})$, the SLLN can be extended to secure the existence of a *strongly consistent* estimator $\widehat{\theta}_n(\mathbf{X})$ of θ :

$$\mathbb{P}(\lim_{n\to\infty}\widehat{\theta}_n(\mathbf{X})=\theta)=1. \tag{16}$$

The result in (16) underwrites what Neyman (1952) called 'stable long-run relative frequencies', whose existence is necessary for the phenomenon of interest to be amenable to statistical modeling and inference. A similar view, using the term 'statistical regularities', was articulated even earlier by Cramer (1946), pp. 137–151.

The strong consistency of $\widehat{\theta}_n$, in conjunction with the statistical adequacy of:

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

bestows an objective frequentist interpretation upon the probabilities assigned by $f(\mathbf{x};\widehat{\theta}_n)$, $\mathbf{x}\in\mathbb{R}^n_X$ which can be used to evaluate (estimate) the probability of any legitimate event in $\sigma(\mathbf{X})$, satisfying the *ascertainability* criterion. Similarly, such probabilistic assignments satisfy the *admissibility* criterion (Salmon 1967) because relative frequencies can be viewed as an instantiation of the mathematical probability.

9 A 'pragmatic' justification of induction?

The inductive premises of the SLLN provide *sufficient* conditions (probabilistic assumptions) pertaining to the process $\{X_k,\ k\in\mathbb{N}\}$ underlying data $\mathbf{x}_0$ to render it amenable to statistical modeling and inference by ensuring the existence of stable relative frequencies. This ensures 'potential' learning from data, but transforming that into 'actual' learning from data $\mathbf{x}_0$ one often needs to strengthen the SLLN premises to a fully-specified statistical model $\mathcal{M}_\theta(\mathbf{x})$ whose statistical adequacy has been secured. The proposed frequentist interpretation of probability addresses the charges [i]–[iii] by an apposite bridging of the gap between what the mathematical results of probability

(measure) theory provide and what can be empirically justified. That is, the justification for model-based induction is empirical; it stems from the appropriateness and pertinence of using a statistical model $\mathcal{M}_\theta(\mathbf{x})$ to learn about real-world phenomena of interest.

In summary, the key features of the proposed frequentist model-based inference are:

[a] it demarcates the inductive premises of inference by formalizing vague a priori stipulations like the 'uniformity of nature' and the 'representativeness of the sample' into formal probabilistic assumptions (IID) [revealing their restrictiveness],
[b] it extends the scope of inductive inference beyond IID samples by including statistical models $\mathcal{M}_\theta(\mathbf{x})$, that account for both dependence and heterogeneity,
[c] it provides a link between the mathematical set-up and the physical reality by viewing data $\mathbf{x}_0$ as a typical realization of the process $\{X_k,\ k\in\mathbb{N}\}$ underlying $\mathcal{M}_\theta(\mathbf{x})$,
[d] it provides an empirical justification for frequentist induction stemming from securing the statistical adequacy of $\mathcal{M}_\theta(\mathbf{x})$ using trenchant Mis-Specification (M-S) testing that relies solely on mathematical probability,
[e] it enhances the reliability and precision of inductive inferences by grounding them on finite sampling distributions rather than relying solely on asymptotic results like the SLLN and the Central Limit Theorem (CLT), and
[f] it renders the 'long-run' metaphor operational by bringing out its key attribute of repeatability in principle.

From the statistical perspective, the above model-based inductive inference can be viewed as an implementation of Fisher's (1922) original vision that was later elaborated upon by Neyman (1952):

> (i) If we wish to treat certain phenomena by means of the theory of probability we must find some element of these phenomena that could be considered as random, following the *law of large numbers*. This involves a construction of a mathematical model of the phenomena involving one of more probability sets. (ii) The mathematical model is found satisfactory, or not. This must be checked by observation. (iii) If the mathematical model is found satisfactory, then it may be used for deductions concerning phenomena to be observed in the future. (p. 27).

From the philosophical perspective, the proposed model-based frequentist interpretation suggests that there was something right-headed about the logical empiricists's denying any a priori foundation for probabilistic inference; see Glymour (1981). In this sense, the proposed interpretation is in the spirit of Reichenbach (1951) view:

> It is true that for the frequency interpretation the degree of probability is a matter of experience and not of reason. (p. 236)

Moreover, the proposed empirical justification of model-based induction can be viewed as making good on earlier (unsuccessful) attempts to provide a 'pragmatic' justification of induction as envisioned by Reichenbach; see Salmon (1967). It is

achieved by strengthening the inductive premises of the SLLN to a fully specified statistical model $\mathcal{M}_\theta(\mathbf{x})$, beyond the simple Bernoulli, that addresses the non-uniqueness of the inductive rule problem by grounding inductive inference on finite sampling distributions associated with the particular data $\mathbf{x}_0$.

Finally, it is important to re-iterate that the same finite sampling distributions will be used in error statistics to furnish, not only the pre-data error probabilities, but also the post-data severity evaluations that will provide an *evidential interpretation* of inference in the form of the discrepancy from the null hypothesis warranted by data $\mathbf{x}_0$. The underlying inductive reasoning differs from both the N–P and Fisherian thinking, but it borrows elements from both in the sense that it can be seen as a harmonious reconciliation of the two; see Mayo and Spanos (2006).

10 Summary and conclusions

The error statistical perspective identifies the probability of an event A—viewed in the context of a statistical model $\mathcal{M}_\theta(\mathbf{x})$, $\mathbf{x}\in\mathbb{R}^n_X$—with the *limit* of its relative frequency of occurrence by invoking the SLLN. This frequentist interpretation is defended against the charges of [i] 'circularity' and [ii] inability to assign 'single event' probabilities, by showing that in model-based induction the defining characteristic of the long-run metaphor is neither its temporal nor its physical dimension, but its repeatability (in principle) which renders it operational in practice. It is also shown that the notion of a statistical model $\mathcal{M}_\theta(\mathbf{x})$ can easily accomodate non-IID samples, rendering [iii] the 'random sample' charge simply misinformed.

The proposed frequentist interpretation of probability replaces enumerative induction with model-based induction and the von Mises rendering with the SLLN in conjunction with viewing data $\mathbf{x}_0 := (x_1, \ldots, x_n)$ as a 'truly typical' realization of the stochastic process $\{X_t,\ t\in\mathbb{N}\}$ specified by $\mathcal{M}_\theta(\mathbf{x})$. The SLLN provides *sufficient* conditions (inductive premises), relating to the process $\{X_k,\ k\in\mathbb{N}\}$ underlying data $\mathbf{x}_0$, to ensure the existence of stable relative frequencies. This demarcates the intended scope of frequentist model-based induction by rendering $\mathbf{x}_0$ amenable to statistical modeling and inference. To ensure 'learning from data', however, one often needs to go the extra mile and strengthen the SLLN premises to a fully-specified statistical model $\mathcal{M}_\theta(\mathbf{x})$ whose statistical adequacy for data $\mathbf{x}_0$ needs to be secured.

Acknowledgments I would like to thank Clark Glymour and Deborah Mayo for encouraging my interest in the issues discussed in this paper. Thanks are also due to Paul Humphreys, Cosma Shalizi and two anonymous referees for many valuable comments and suggestions that helped to improve the paper substantially.

References

Billingsley, P. (1995). *Probability and measure* (3rd ed.). NY: Wiley.

Borel, E. (1909). Sur les probabilites et leurs applications arithmetiques. *Rend. Circ. Mat. Palermo, 26*, 247–271.

Chaitin, G. J. (2001). *Exploring randomness*. NY: Springer.

Church, A. (1940). On the concept of a random sequence. *Bulletin of the American Mathematical Society, 46*, 130–135.

Cox, D. R. (1990). Role of models in statistical analysis. *Statistical Science, 5*, 169–174.

Cox, D. R., & Hinkley, D. V. (1974). *Theoretical statistics*. London: Chapman & Hall.
Cramer, H. (1946). *Mathematical methods of statistics*. NJ: Princeton Unversity Press.
Doob, J. L. (1953). *Stochastic processes*. New York: Wiley.
Fine, T. L. (1973). *Theories of probability—an examination of foundations*. NY: Academic Press.
Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society A, 222*, 309–368.
Fisher, R. A. (1925a). *Statistical methods for research workers*. Edinburgh: Oliver and Boyd.
Fisher, R. A. (1925b). Theory of statistical estimation. *Proceedings of the Cambridge Philosophical Society, 22*, 700–725.
Fisher, R. A. (1934). Two new properties of maximum likelihood. *Proceedings of the Royal Statistical Society A, 144*, 285–307.
Fisher, R. A. (1955). Statistical methods and scientific induction. *Journal of The Royal Statistical Society B, 17*, 69–78.
Giere, R. N. (1984). *Understanding scientific reasoning* (2nd ed.). NY: Holt. Rinehart and Winston.
Gillies, D. (2000). *Philosophical theories of probability*. London: Routledge.
Glymour, C. (1981). *Theory and evidence*. NJ: Princeton University Press.
Godambe, V., & Sprott, D. (Eds.). (1971). *Foundations of statistical inference holt*. Toronto: Rinehart and Winston of Canada.
Gossett, W. S. (aka Student) (1908). The probable error of the mean. *Biometrika, 6*, 1–25.
Hacking, I. (1965). *Logic of statistical inference*. Cambridge: Cambridge University Press.
Hacking, I. (1968). One problem about induction. In I. Lakatos (Eds.), *The problem of inductive logic* (pp. 44–59). Amsterdam: North-Holland.
Hajek, A. (2007). Interpretations of probability. In the Stanford Encyclopedia of Philosophy http://plato.stanford.edu/entries/probability-interpret/.
Harper, W., & Hooker, C. (Eds.). (1976). *Foundations of probability theory statistical inference and statistical theories of science* (Vol. 2). Dordrecht, The Netherlands: D. Reidel.
Howson, C., & Urbach, P. (2006). *Scientific reasoning: The Bayesian approach* (3rd ed.). Chicago, IL: Open Court.
Kolmogorov, A. N. (1933). *Foundations of the theory of probability* (2nd English edition). NY: Chelsea Publishing Co.
Kolmogorov, A. N. (1963). On tables of random numbers. *Sankhya, Indian Journal of Statistics A, 25*, 369–376.
Kolmogorov, A. N. (1983). On logical foundations of probability theory. In K. Ito., & J. V. Prokhorov (Eds.), *Probability theory and mathematical statistics* (pp. 1–5). NY: Springer-Verlag.
Kyburg, H. E. (1974). *The logical foundations of statistical inference*. Dorbrecht-Holland: Reidel.
Lehmann, E. L. (1986). *Testing statistical hypotheses* (2nd ed.). NY: Wiley.
Lehmann, E. L. (1990). Model specification: The views of Fisher and Neyman, and later developments. *Statistical Science, 5*, 160–168.
Li, M., & Vitanyi, P. (2008). *An introduction to Kolmogorov complexity and its applications* (3rd ed.). NY: Springer.
Lindley, D. V. (1965). *Introduction to probability and statistics from the bayesian viewpoint* (Vol. 1). Cambridge: Cambridge University Press.
Martin-Löf, P. (1969). The literature on von Mises' Kollectives revisited. *Theoria, 35*, 12–37.
Mayo, D. G. (1996). *Error and the growth of experimental knowledge*. Chicago: The University of Chicago Press.
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
Mayo, D. G., & Spanos, A. (2004). Methodology in practice: Statistical misspecification testing. *Philosophy of Science, 71*, 1007–1025.
Mayo, D. G., & Spanos, A. (2006). Severe testing as a basic concept in a Neyman–Pearson philosophy of induction. *British Journal for the Philosophy of Science, 57*, 323–357.
Mayo, D. G., & Spanos, A. (2011). Error statistics. In D. Gabbay, P. Thagard., & J. Woods, *Philosophy of statistics, handbook of philosophy of science*. Amsterdam: Elsevier.
Morrison, D. E., & Henkel, R. E. (1970). *The significance test controversy: A reader*. Chicago: Aldine.
Neyman, J. (1952). *Lectures and conferences on mathematical statistics and probability* (2nd ed.). Washington: U.S. Department of Agriculture.

Neyman, J. (1956). Note on an article by Sir Ronald Fisher. *Journal of the Royal Statistical Society B, 18*, 288–294.
Neyman, J. (1977). Frequentist probability and frequentist statistics. *Synthese, 36*, 97–131.
Neyman, J., & Pearson, E. S. (1933). On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society A, 231*, 289–337.
Nies, A. (2009). *Computability and randomness*. Oxford: Oxford University Press.
Pearson, E. S. (1955). Statistical concepts in their relation to reality. *Journal of the Royal Statistical Society B, 17*, 204–207.
Pearson, K. (1920). The fundamental problem of practical statistics. *Biometrika, XIII*, 1–16.
Reichenbach, H. (1934/1949). *The Theory of probability*. Berkeley, CA: University of California Press.
Reichenbach, H. (1951). *The rise of scientific philosophy*. Berkeley, CA: University of California Press.
Renyi, A. (1970). *Probability theory*. Amsterdam: North-Holland.
Salmon, W. C. (1967). *The foundations of scientific inference*. Pittsburgh: University of Pittsburgh Press.
Salmon, W. C. (1984). *Scientific explanation and the causal structure of the world*. Princeton, NJ: Princeton University Press.
Seidenfeld, T. (1979). *Philosophical problems of statistical inference*. Dorbrecht-Holland: Reidel.
Skyrms, B. (2000). *Choice and chance: An introduction to inductive logic* (4th ed.). US: Wadsworth.
Spanos, A. (1999). *Probability theory and statistical inference: Econometric modeling with observational data*. Cambridge: Cambridge University Press.
Spanos, A. (2006). Where do statistical models come from? Revisiting the problem of specification. In J. Rojo (Ed.) Optimality: The second Erich L. Lehmann symposium lecture notes-monograph series (Vol. 49, pp. 98–119), Institute of Mathematical Statistics, Beachwood, OH.
Spanos, A. (2007). Curve-fitting, the reliability of inductive inference and the error-statistical approach. *Philosophy of Science, 74*, 1046–1066.
Spanos, A. (2009). Statistical misspecification and the reliability of inference: The simple t-test in the presence of Markov dependence. *The Korean Economic Review, 25*, 165–213.
Spanos, A. (2010). Theory testing in economics and the error statistical perspective. In D.G. Mayo & A. Spanos (Eds.), *Error and inference* (pp. 202–246). Cambridge: Cambridge University Press.
Spanos, A. (2010). The discovery of Argon: A case for learning from data? *Philosophy of Science, 77*, 359–380.
Spanos, A. (2010). Is frequentist testing vulnerable to the base-rate fallacy? *Philosophy of Science, 77*, 565–583.
Spanos, A. (2010). Statistical adequacy and the trustworthiness of empirical evidence: Statistical vs. substantive information. *Economic Modelling, 27*, 1436–1452.
Ville, J. (1939). *Etude Critique de la Notion de Collectif*. Paris: Gauthier-Villars.
Von Mises, R. (1928). *Probability, statistics and truth* (2nd ed.). NY: Dover.
Williams, D. (2001). *Weighing the odds: A course in probability and statistics*. Cambridge: Cambridge University Press.
Wald, A. (1937). Die Widerspruchsfreiht des Kollektivbergriffes in der Wahrscheinlicheitsrechnung. *Ergebnisse Eines Mathematischen Kolloquiums, 8*, 38–72.
Wasserman, L. (2006). *All of nonparametric statistics*. NY: Springer.