Model-based Induction and the Frequentist Interpretation of Probability

Aris Spanos

Spanos, A. (2013), "A frequentist interpretation of probability for model-based inductive inference," *Synthese*, **190**:1555–1585. DOI 10.1007/s11229-011-9892-x

1 Introduction: the frequentist interpretation

The conventional wisdom in philosophy of science. The frequentist interpretation of probability, which relates $P(A)$ to the limit of the relative frequency $\frac{m}{n}$ of the occurrence of A as $n \to \infty$, does *not* meet the basic criteria of:

$$\boxed{\text{(a) Admissibility, (b) Ascertainability, (c) Applicability.}}$$

In particular (Salmon, 1967, Hajek, 2009), argue that:

- (i) its definition is 'circular' (invokes probability to define probability) [(a)],
- (ii) it relies on 'random samples' [(a), (b)],
- (iii) it cannot assign probabilities to 'single events', and
- (iv) frequencies must be defined relative to a 'reference class' [(b)-(c)].

Koop, Poirier and Tobias (2007), p. 2: "... frequentists, argue that situations not admitting repetition under essentially identical conditions are not within the realm of statistical enquiry, and hence 'probability' should not be used in such situations. Frequentists define the probability of an event as its long-run relative frequency. The frequentist interpretation cannot be applied to (i) unique, once and for all type of phenomena, (ii) hypotheses, or (iii) uncertain past events. Furthermore, this definition is nonoperational since only a finite number of trials can ever be conducted."

Howson and Urbach (2006): "... 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)

Since the 1950s, discussions in philosophy of science have concentrated primarily on a number of **defects** in frequentist reasoning that give rise to **fallacious and counter-intuitive results**, and highlighted the **limited scope and applicability** of the frequentist interpretation of probability; see Kyburg (1974), Giere (1984), Seidenfeld (1979), Gillies (2000), Sober (2008), inter alia.

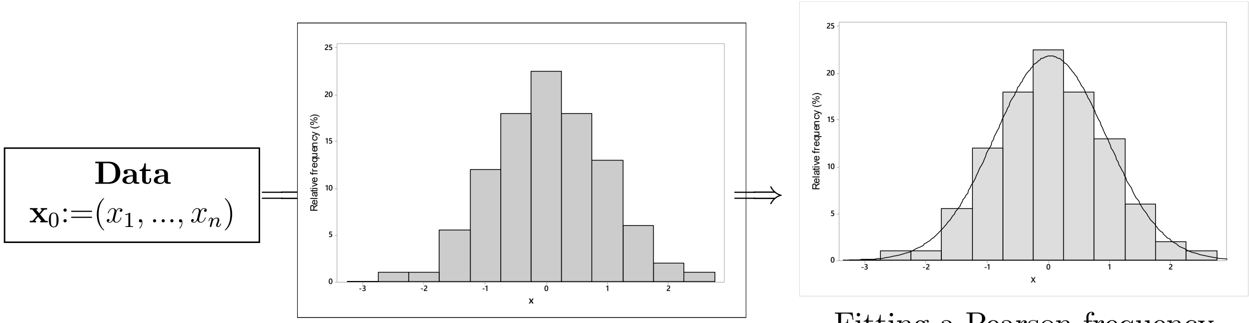
Proponents of the **Bayesian approach** to inductive inference *muddied the waters* further and hindered its proper understanding by introducing several **mis-interpretations** and **cannibalizations** of the frequentist approach to inference; see Berger and Wolper (1988), Howson (2000), Howson and Urbach (2005).

These discussions have discouraged philosophers of science to take frequentist inductive inference seriously and attempt to address some of its foundational problems; Mayo (1996) is the exception.

1.1 Frequentist approach: foundational 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, into a frequentist 'model-based induction', relying on *finite sampling distributions*.

Karl Pearson (1920) would commence with data $\mathbf{x}_0 := (x_1, ..., x_n)$ in search of a frequency curve to describe the resulting histogram.



Histogram of the data

Fitting a Pearson frequency curve $f(x; \widehat{\theta}_1, \widehat{\theta}_2, \widehat{\theta}_3, \widehat{\theta}_4)$

Fig. 1: The Karl Pearson approach to statistics

In contrast, Fisher (1922) proposed to begin with:

(a) a prespecified model (a hypothetical infinite population), say, ***the simple Normal model*:**

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

(b) view $\mathbf{x}_0$ as a typical realization of of the process $\{X_k, \; k \in \mathbb{N}\}$ underlying $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$. Indeed, he made specification (the initial choice) of the prespecified statistical model a response to the question:

"Of what population is this a random sample?" (p. 313),

emphasizing that:

'the adequacy of our choice may be tested *a posteriori*' (p. 314).

Since then, the notions (a)-(b) have been extended and formalized in purely *probabilistic* terms to define the concept of a **statistical model:**

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

where $f(\mathbf{x}; \boldsymbol{\theta})$ is the *distribution of the sample* $\mathbf{X} := (X_1, ..., X_n)$.

What is the key difference between the approach proposed by Fisher and that of K-Pearson?

In the K-Pearson approach the IID assumptions are made implicitly, but Fisher brought them out explicitly as the relevant statistical (inductive) premises $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$, i.e. it is assumed that $\{X_k, \; k \in \mathbb{N}\}$ is NIID, and, as a result, one can test them vis-a-vis data $\mathbf{x}_0$.

Does it make a difference in practice? A big difference!

Statistical misspecification renders the **nominal error probabilities** different from the **actual** ones.

Fisher (1925, 1935) constructed 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 Cox and Hinkley (1974).

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).

▶ Neyman argued for 'inductive behavior' based on Neyman-Pearson (N-P) testing (Neyman, 1952).

Unfortunately, several foundational problems remained unresolved.

Inference foundational problems:

■ [a] a sound frequentist interpretation of probability that offers a proper foundation for frequentist inference,

■ [b] the form and nature of inductive reasoning underlying frequentist inference,

■ [c] the initial vs. final precision (Hacking, 1965), i.e. the role of pre-data vs. post-data error probabilities,

■ [d] safeguarding frequentist inference against unwarranted interpretations, including:

(i) the fallacy of acceptance: interpreting accept H_0 [no evidence against H_0] as evidence for H_0; e.g. the test had low power to detect existing discrepancy,

(ii) the fallacy of rejection: interpreting reject H_0 [evidence against H_0] as evidence for a particular H_1; e.g. conflating statistical with substantive significance (Mayo, 1996).

Modeling foundational problems:

■ [e] the role of substantive subject matter information in statistical modeling (Lehmann, 1990, Cox, 1990),

■ [f] statistical model specification: how to narrow down a (possibly) infinite set $\mathcal{P}(\mathbf{x})$, of all possible models that could have given rise to data $\mathbf{x}_0$, to a single statistical model $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$.

■ [g] Mis-Specification (M-S) testing: assessing the adequacy a statistical model $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ *a posteriori.*

■ [h] statistical model respecification: how to respecify a statistical model $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ when found *misspecified.*

■ [i] Duhem's conundrum: are the substantive claims false or the inductive premises misspecified.

These issues created endless confusions in the minds of practitioners concerning the appropriate implementation and interpretation of frequentist inductive inference.

Error statistics (A) *extends* the Fisher-Neyman-Pearson (F-N-P) approach by supplementing it with a **post-data severity** assessment component, in an attempt to address problems [b]-[d] (Mayo, 1996, Mayo & Spanos, 2006, 2010, 2011).

(B) It *refines* the F-N-P approach by proposing a broader framework with a view to secure **statistical adequacy**, motivated by the aim to deal with the foundational problems [e]-[i]; Mayo and Spanos (2004), Spanos (1986, 1999, 2007, 2018).

★ **This paper** focuses on [a] by defending the *frequentist interpretation* of probability against several well-rehearsed charges, including:

(i) the circularity of its definition,

(ii) its reliance on 'random samples',

(iii) its inability to assign 'single event' probabilities, and

(iv) the 'reference class' problem.

▶ The argument in a nutshell is that, although charges (i)-(iv) might constitute legitimate criticisms of **enumerative induction** and the **von Mises (1928) rendering** of the frequentist interpretation of probability, they constitute misplaced indictments when directed against the model-based 'stable long-run frequencies' interpretation (Neyman, 1952), grounded on the Strong Law of Large Numbers (SLLN).

Key difference between enumerative and model-based induction.

Enumerative induction relies on simple (implicit) statistical models whose premises are **vaguely framed** in terms of a priori stipulations like the 'uniformity of nature' and the 'representativeness of the sample' (Skyrms, 2000).

Model-based induction revolves around $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ whose premises are specified in terms of probabilistic assumptions that are testable vis-à-vis data $\mathbf{x}_0$.

2 Frequentist interpretation of probability

This section articulates a frequentist interpretation, that revolves around the notion of a *statistical model*, as opposed to the 'collective' for the von Mises variant.

2.1 Kolmogorov's axiomatic formulation of probability

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

(a) S denotes the set of all possible distinct outcomes.

(b) $\mathfrak{F}$ denotes a set of subsets of S, called *events* of interest, endowed with the mathematical structure of a σ-field, i.e. it satisfies the following conditions:

(i) $S \in \mathfrak{F}$, (ii) if $A \in \mathfrak{F}$, then $\overline{A} \in \mathfrak{F}$,

(iii) if $A_i \in \mathfrak{F}$ for $i{=}1, 2, ..., n, ...$ then $\bigcup_{i=1}^{\infty} A_i \in \mathfrak{F}$.

(c) $\mathbb{P}(.)$: $\mathfrak{F} \rightarrow [0, 1]$ is a set function satisfying the axioms:

[A1] $\mathbb{P}(S) = 1$, for any outcomes set S,

[A2] $\mathbb{P}(A) \geq 0$, for any event $A \in \mathfrak{F}$,

[A3] *Countable Additivity.* For $A_i \in \mathfrak{F}$, $i{=}1, ..., n, ..$, s.t. $A_i \cap A_j{=}\emptyset$, for all $i{\neq}j$, $i, j{=}1, 2, ..., n, ...$, 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).

Can the above **Kolmogorov formalism** be given an *interpretation* by assigning a *meaning* to the primitive term probability?

"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." (Cramer, 1946)

Primary objective. Modeling observable stochastic phenomena of interest giving rise to data that exhibit chance regularity patterns (Spanos, 1999).

2.2 Random variables and statistical models

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

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

That is, $X(.)$ assigns numbers to the elementary events in S 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 $(S, \mathfrak{F}, \mathbb{P}(.))$ and the observable stochastic phenomena of interest, since observed data come usually in the form of *numbers*.

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

$$(S, \mathfrak{F}, \mathbb{P}(.)) \overset{X(.)}{\longrightarrow} \mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x}){=}\{f(\mathbf{x}; \boldsymbol{\theta}),\ \boldsymbol{\theta} \in \Theta\},\ \mathbf{x} \in \mathbb{R}_X^n.$$

Hence, the notion of probability associated with $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ 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:

$$\{X{=}1\}{=}A,\ \{X{=}0\}{=}\overline{A},\ \text{with } \mathbb{P}(A){=}p,\ \mathbb{P}(\overline{A}){=}1{-}p,$$

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) $\{X_k,\ k\in\mathbb{N}\}$ assumed to be IID, i.e. the underlying statistical model is ***the simple Bernoulli model***:

$$\boxed{\mathcal{M}_\theta(\mathbf{x}):\ X_k \backsim \mathsf{BerIID}(\theta,\theta(1-\theta)),\ \ k\in\mathbb{N}.} \tag{1}$$

■ In general, the statistical model $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ 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 $\mathbf{x}_0$:=$(x_1,...,x_n)$ a *truly typical realization* thereof.

Example 1. What would a truly typical realization from this model look like?

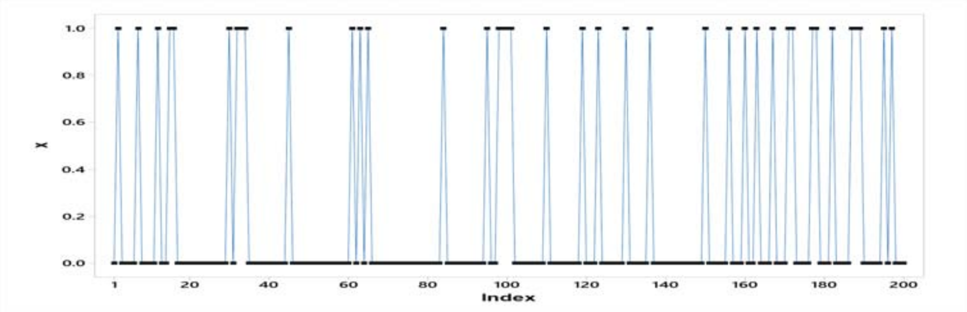


Fig. 3 - Typical realization from a BerIID process: $\theta = .2$

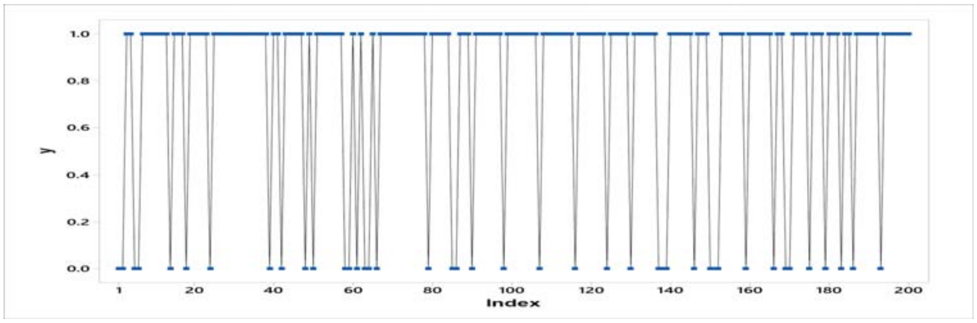


Fig. 4 - Typical realization from a BerIID process: $\theta = .8$

2.3 The frequentist interpretation anchored on the SLLN

The proposed *frequentist interpretation* identifies the probability of an event A with the *limit* of the relative frequency of its occurrence, $\overline{x}_n=\frac{1}{n}\sum_{k=1}^{n}x_k=\frac{m}{n}$, in the context of a well-defined stochastic mechanism $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$.

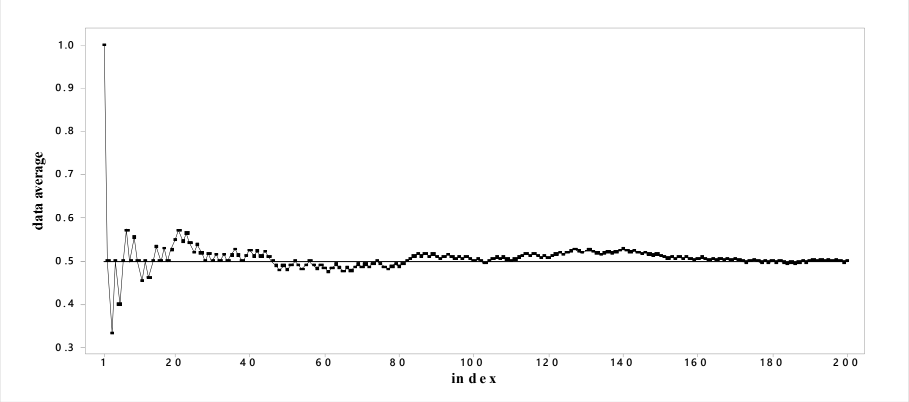
The SLLN gives precise probabilistic meaning to the **unwarranted claim**:

$$\boxed{\text{the sequence of relative frequencies } \{\overline{x}_n\}_{n=1}^{\infty} \text{ converges to } p \text{ as } n\to\infty.}$$

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

$$\boxed{\mathbb{P}(\lim_{n\to\infty}(\tfrac{1}{n}\textstyle\sum_{k=1}^{n}X_k)=p)=1.} \tag{2}$$

That is, as $n \to \infty$ the stochastic sequence $\{\overline{X}_n{=}\frac{1}{n}\sum_{k=1}^{n} X_k\}_{n=1}^{\infty}$, converges to a constant p *with probability one* or *almost surely* (*a.s.*) $[\overline{X}_n \overset{a.s.}{\to} p]$; see Billingsley (1995).



$\overline{x}_n$ for Bernoulli IID data with $n{=}200$

▲ Let us clarify the notion of convergence in (2) 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 (2) holds everywhere in a domain $D_1 \subset S$ except on $D_0 \subset S$, *a set of measure zero* ($\mathbb{P}(D_0){=}0$), i.e.

$$D_1{=}\{s\colon \lim_{n\to\infty} \overline{X}_n(s){=}p,\, s{\in}S\},\;\; D_0{=}\{s\colon \lim_{n\to\infty} \overline{X}_n(s){\neq}p,\, s{\in}S\}.$$

"Thus, D is the set of outcomes for which the 'long-term relative frequency' idea works. Then D is an event, and $\mathbb{P}(D){=}1$." (Williams, 2001, p. 111).

Second, the result in (2) 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 that one needs to invoke the *Law of Iterated Logarithm* (LIL), which quantifies the *rate* of convergence of the process $\{\overline{X}_n\}_{n=1}^{\infty}$. For an IID process $\{X_k,\ k{\in}\mathbb{N}\}$ with $E(X_k){=}\mu,\ Var(X_k){=}\sigma^2{<}\infty,\ k{\in}\mathbb{N}$:

$$\boxed{\textbf{Khinchin LIL}\text{:}\;\; \mathbb{P}\left(\lim_{n\to\infty}\sup\left[\frac{\left|\sum_{k=1}^{n}(X_k-\mu)\right|}{\sqrt{n\ln(\ln(n))}}\right]{=}\sqrt{2\sigma^2}\right){=}1.}$$

Third, the result in (2) holds when $\{X_k,\ k{\in}\mathbb{N}\}$ satisfies certain probabilistic assumptions, the most restrictive being IID, i.e. these assumptions are sufficient to secure the limit exists.

★ This suggests that from a modeling perspective, the SLLN is essentially an *existence* result for stable (constant) relative frequencies ($\overline{X}_n \overset{a.s.}{\to} p$), 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.

That is, the absence of stable relative frequencies implies that the phenomenon of interest is beyond the scope of statistical modeling, because it exhibits no k-invariant chance regularities.

Fourth, $\overline{X}_n \overset{a.s.}{\to} p$ does *not* involve any claims about the **mathematical convergence** of the sequence of numbers $\{\overline{x}_n\}_{n=1}^{\infty}$ to p in a purely mathematical sense: $\lim_{n\to\infty} \overline{x}_n{=}p$.

Unfortunately, the line between probabilistic (*a.s.*) and mathematical convergence was blurred by von Mises's (1928) **collective** which was defined in terms of infinite realizations $\{x_k\}_{k=1}^{\infty}$ whose partial sums $\{\overline{x}_n\}_{n=1}^{\infty}$ converge to p; Gillies (2000). However, any attempt to make rigorous the convergence $\lim_{n\to\infty} \overline{x}_n{=}p$ is ill-fated for 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)

The **long-run metaphor** associated with the frequentist interpretation, anchored on the SLLN, enables one to conceptualize the frequentist interpretation of probability by bringing out the connection between the stochastic generating mechanism (i.e. IID Bernoulli) and the probability of event(s) of interest (e.g.. $X{=}1$).

In conclusion, it is important to emphasize that, by themselves, mathematical results, such as the SLLN (2) and the LIL, do not suffice to provide an apposite frequentist interpretation that addresses the foundational problems pertaining to the *inductive reasoning* underlying frequentist inference.

▲ **Statistical induction** requires a pertinent link between such mathematical results and the actual data-generating mechanism. In error statistics 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 '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 M-S testing.

That is, the set of all typical realizations – they satisfy the invoked probabilistic assumptions (IID) – comprise the *uncountable* set $D_1{=}\{s{:}\ \lim\limits_{n\to\infty} \overline{X}_n(s){=}p,\, s{\in}S\}$ of measure one, but the non-typical realizations such as:

$$\begin{array}{l}\{x_k\}_{k=1}^{\infty} = \{0, 0, ..., 0, ...\},\\ \{x_k\}_{k=1}^{\infty} = \{1, 1, ..., 1, ...\},\\ \{x_k\}_{k=1}^{\infty} = \{1, 0, 1, 0..., 1, 0, ...\},\\ \{x_k\}_{k=1}^{\infty} = \{1, 1, 0, 0, 1..., 1, 0, 0, ...\},\\ \{x_k\}_{k=1}^{\infty} = \{1, 1, 1, 0, 0, 0, ..., 0, 1, 1, 1, ...\},\ \text{etc.}\end{array} \tag{3}$$

define a *countable* set $D_0{=}\{s{:}\ \lim\limits_{n\to\infty} \overline{X}_n(s){\neq}p,\, s{\in}S\}$ of measure zero; see Adams and Guillemin (1996). But how would one know that the particular realization $\mathbf{x}_0$ in hand is non-typical? They do not satisfy the probabilistic assumptions (IID). Hence, in practice the falsity of the IID assumptions can be detected using simple **Mis-Specification (M-S) tests**, like a *runs test*, which relies solely on mathematical combinatorics; see Spanos (2019).

2.4 Von Mises' frequentist interpretation

The early 20th century rendering of the frequentist interpretation of probability was put forward by von Mises (1928). In contrast to the model-based frequentist interpretation that revolves around the concept of a statistical model $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$, von Mises interpretation of probability is anchored on:

> 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 $(x_k{=}1)$ that satisfies two conditions:

Table 10.5: Conditions for von Mises collective

(C) Convergence:	$\lim_{n\to\infty}\left(\frac{1}{n}\sum_{k=1}^{n}x_k\right){=}p_A,$	(4)
(R) Randomness:	$\lim_{n\to\infty}\left(\frac{1}{n}\sum_{k=1}^{n}\varphi(x_k)\right){=}p_A,$	

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

Since the 1940s, the philosophy of science literature has called into question von Mises's frequentist interpretation of probability on several grounds by viewing it as providing the link between the empirical relative frequencies and the corresponding mathematical probabilities in conjunction with *induction by enumeration*; 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.

Model-based induction. In contrast to the use of enumerative induction in philosophical discussions, practitioners in most applied fields rely on frequentist *model-based* induction based on the notion of a statistical model $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$, assumed to represent an idealized generating mechanism that could have given rise to data $\mathbf{x}_0{:=}(x_1,...,x_n)$. The key difference between the two perspectives stems from the nature and justification of their inductive premises and the ensuing inferences; see Spanos (2013a).

Model-based induction relies on a statistical model $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ whose inductive premises are specified in terms of testable probabilistic assumptions pertaining to a general stochastic process $\{X_t,\ t{\in}\mathbb{N}{:=}(1,2,...,n,...)\}$ 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

$\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ 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, 1999).

Long-run metaphor. 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 these charges stem primarily from mis-attributing to the long-run metaphor a temporal and/or a physical dimension instead of **the 'repeatability' in principle** of the underlying stochastic mechanism described by $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$.

2.5 Revisiting the circularity charge

The common sense intuition underlying the SLLN in (2), that the relative frequency of occurrence of event A converges to $\mathbb{P}(A){=}p$ as n increases, is often the source of the charge that the frequentist interpretation of probability 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 Renyi (1970, p. 159):

"It may seem that there lurks some vicious circle here: probability was indeed defined by means of the stability of relative frequency, and yet in the definition of stability of relative frequency the concept of probability is hidden. In reality there is no logical fault. The "definition" of the probability stating that the probability is the numerical value around which the relative frequency is fluctuating at random is not a mathematical definition: it is an intuitive description of the realistic background concept of probability. 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."

▶ Elaborating on his last sentence, the SLLN is an **existence result** for 'stable relative frequencies' (converging to 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 $\{\overline{X}_n\}_{n=1}^{\infty}$ [not $\{\overline{x}_n\}_{n=1}^{\infty}$], which involves the purely measure-theoretic notion of a set of measure zero.

Anonymous referee: *To suggest, as he/she does, that Lindley lacks sufficient expertise in the measure-theoretic treatment of probability is insulting, false and rebounds back on him/her: what Lindley and those other authors possess besides unchallengeable mathematical competence is a sensitivity to philosophical problems and a realisation that appeals to convergence except on sets of measure zero, 'strong consistency' etc. do not solve them.*

Response: Lindley referring to $|\overline{x}_n - p| \leq \epsilon$ is invoking mathematical convergence of the form $\overline{x}_n \underset{n\to\infty}{\to} p$ has nothing to do with almost sure convergence of $\overline{X}_n \xrightarrow{a.s.} p$.

Adams and Guillemin (1996), in the introduction to a book entitled "Measure Theory and Probability", argue:

"What we hope to convey here is that had the Lebesgue theory of measure not existed, one would be forced to invent it to contend with the paradoxes of large numbers." (p. x)

Given that the SLLN and the LIL are purely measure-theoretic results, the circularity charge is clearly misplaced. Why do critics keep reiterating this charge?

▼ One possible explanation might be that these critics consider the 'long-run frequency' *itself* as providing a 'definitional link' between "statements of probability calculus" and "the physical reality" (Howson and Urbach, 2005, p. 48-49).

The pertinence of this link was challenged by Kolmogorov (1963), p. 369:

"[the long-run frequency] 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."

The model-based frequentist interpretation invokes **no** such link. The link 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 $\{X_k,\ k\in\mathbb{N}\}$.

[i] data $\mathbf{x}_0 := (x_1, x_2, \ldots, x_n)$ is viewed as a '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 M-S testing.

The same interpretive provisions [i]-[ii] are used by Kolmogorov's algorithmic information theory (Li and Vitanyi, 2008), whose notion of randomness is based on the effective computability and **incompressibility** of finite sequences.

This provides a purely *non-probabilistic* (algorithmic) rendering to the frequentist interpretation that operationalizes all the above measure-theoretic results:

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

★ This algorithmic rendering dispels any intimation of **circularity** stemming from the interpretive provisions [i]-[ii].

2.6 The frequentist interpretation and 'random samples'

Does the proposed frequentist interpretation of probability rely on the notion of a random sample $\mathbf{X}$ (IID random variables $(X_1, ..., X_n)$)?

It is fair to say that the IID assumptions appear to constitute an integral part of von Mises's (1928) frequentist interpretation, being reflected in his condition of 'invariance under place selection' for *admissible collectives* $\{x_k\}_{k=1}^{\infty}$.

However, the frequentist interpretation anchored on the SLLN does not require such restrictive 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 issue 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.

▼ This motivated the notion of 'typical realization', which can be easily extended to non-IID processes. The only restriction on the latter is that they retain a form of *t-invariance* encapsulating the *unvarying features* of the phenomenon being modeled in terms of the unknown parameter(s) $\boldsymbol{\theta}$.

Example. 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) is:

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

where $\boldsymbol{\theta}:=(\beta_0,\beta_1,...\beta_m,\alpha_1,\alpha_2,...,\alpha_p,\sigma^2)$ are t-invariant.

Indeed, the reason for defining $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ in terms of the *joint* distribution, $f(\mathbf{x};\boldsymbol{\theta})$ is to account for the dependence/heterogeneity in non-IID samples; a key result first established by Kolmogorov (1933).

Since Borel (1909) the sufficient probabilistic assumptions on the process $\{X_k,\ k\in\mathbb{N}\}$ giving rise to the SLLN result in (2) have been weakened considerably; Spanos (2019), ch. 9. 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 process or a martingale process.

Second the result has been extended from the linear function $\overline{X}_n=\frac{1}{n}\sum_{k=1}^{n}X_k$, to any Borel function of the sample, say $Y_n=h(X_1,X_2,...,X_n)$; Billingsley (1995).

For a general statistical model $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ based on a non-IID sample, the assignment of the probabilities using $f(\mathbf{x};\boldsymbol{\theta})$, $\mathbf{x}\in\mathbb{R}_X^n$ depends crucially on being able to estimate consistently the unknown parameter(s) $\boldsymbol{\theta}$. Indeed, the constancy of the parameters $\boldsymbol{\theta}$ renders possible the estimation of stable relative frequencies associated with $f(\mathbf{x};\boldsymbol{\theta})$.

Hence, in the context of $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$, the SLLN can be extended to secure the existence of a **strongly consistent** estimator $\widehat{\boldsymbol{\theta}}_n(\mathbf{X})$ of $\boldsymbol{\theta}$:

$$\mathbb{P}(\lim_{n\to\infty}\widehat{\boldsymbol{\theta}}_n(\mathbf{X})=\boldsymbol{\theta})=1. \tag{5}$$

The result in (5) 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, also founded on 'statistical regularities', was articulated even earlier by Cramer (1946), pp. 137-151.

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

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

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

The above discussion suggests that the various criticisms of the frequentist interpretation on admissibility and ascertainability grounds, stemming from the convergence/divergence of the sequence of relative frequencies $\{\overline{x}_n\}_{n=1}^{\infty}$ (Salmon, 1967, pp. 84-87), are simply misplaced.

To be fair, they constitute valid criticisms of the von Mises (1928) frequentist interpretation, but they are misdirected when leveled against the frequentist interpretation anchored on the SLLN.

3 Error statistics and model-based induction

The notion of a statistical model $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x}),\ \mathbf{x}\in\mathbb{R}^n_X$, describing an idealized stochastic mechanism that could have given rise to $\mathbf{x}_0$, provides the cornerstone of the proposed frequentist interpretation of probability.

In error statistics, the **statistical model** $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ plays a pivotal role because:

- (i) it specifies the inductive premises of inference,
- (ii) it delimits legitimate events in terms of an univocal *sample space* $\mathbb{R}^n_X$,
- (iii) it assigns probabilities to all legitimate events via $f(\mathbf{x};\boldsymbol{\theta})$,
- (iv) it defines what are legitimate hypotheses and/or inferential claims,
- (v) it determines the relevant error probabilities in terms of which the optimality and reliability of inference methods is assessed, and
- (vi) it lays out what constitute legitimate data $\mathbf{x}_0$ for inference purposes.

In relation to (v), $\mathcal{M}_\theta(\mathbf{x})$ also determines the optimality of inference procedures in terms of the relevant error probabilities. This is because for any statistic (estimator, test statistic), say $T_n{=}g(X_1,...,X_n)$, its sampling distribution is derived from $f(\mathbf{x};\boldsymbol{\theta})$ via:

$$F(t;\boldsymbol{\theta}){:=}\mathbb{P}(T_n\leq t;\boldsymbol{\theta})=\underbrace{\int\int\cdots\int}_{\{\mathbf{x}:\ g(x_1,...,x_n)\leq t;\ \mathbf{x}\in\mathbb{R}_X^n\}} f(\mathbf{x};\boldsymbol{\theta})dx_1dx_2\cdots dx_n.$$

3.1 Frequentist interpretation: an empirical justification

The statistical model underlying Borel's SLLN is the simple Bernoulli model $\mathcal{M}_\theta(\mathbf{x})$ in (1), which can be specified more explicitly as in table 1. The validity of assumptions [1]-[4] vis-à-vis data $\mathbf{x}_0$ is what secures the reliability of any inference concerning θ, including the SLLN.

Table 1 - Simple Bernoulli Model

Statistical GM:		$X_k=\theta+u_k,\ \ k\in\mathbb{N}.$	
[1]	Bernoulli:	$X_k\backsim\mathsf{Ber}(.,.),\ x_k{=}0,1,$	$k\in\mathbb{N}.$
[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	

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 $\mathbf{x}_0$.

Hence, the only relevant question 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 'typical realization' of this process. Such typical realizations of zeros and ones form the uncountable set $D_1{=}\{s{:}\ \lim_{n\to\infty}\overline{X}_n(s){=}p,\ s{\in}S\}$ of measure one ($\mathbb{P}(D_1){=}1$), and $D_0{=}\{s{:}\ \lim_{n\to\infty}\overline{X}_n(s){\neq}p,\ s{\in}S\}$ the set of non-typical realizations such as the ones in (3) of measure zero ($\mathbb{P}(D_0){=}0$).

In summary, the justification of the above frequentist interpretation of $\mathbb{P}(A)=p$, is *not* in terms of *a priori* stipulations, but stems from the adequacy of the statistical model $\mathcal{M}_\theta(\mathbf{x})$ (table 1) originating in the interpretive provisions [i]-[ii]. That is, statistical adequacy secures the meaningfulness of identifying the limit of the relative frequencies $\{\overline{x}_n\}_{n=1}^{\infty}$ with the probability p by invoking (2). Given that the probabilistic assumptions [1]-[4] are testable vis-à-vis data $\mathbf{x}_0$, the frequentist interpretation is justifiable on *empirical* grounds.

▲ One could go even further and make a case that frequentist model-based induction has provided the missing empirical cornerstone for **ampliative induction**. First, it has formalized the philosopher's vague a priori stipulations like the 'uniformity of nature' and the 'representativeness of the sample' into clear probabilistic assumptions (IID) that are testable vis-à-vis data $\mathbf{x}_0$. Second, it has extended the IID-based statistical models (implicitly used), to more general ones based on non-IID processes.

3.2 Kolmogorov complexity: an algorithmic perspective

A crucial feature of the above error-statistical stochastic perspective on randomness is that it can be viewed as a dual to an algorithmic perspective based on the notion of *Kolmogorov complexity*, associated with the work of Kolmogorov, Solomonoff, Martin-Löf and Chaitin (Li and Vitanyi, 2008). The duality stems from the fact that both perspectives rely on the same inductive interpretive stipulations [i]-[ii], but grounded on entirely different mathematical formulations.

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 (Li and Vitanyi, 2008):

"random infinite sequences as sequences all of whose initial finite segments pass all effective randomness tests"; see Kolmogorov (1963), p. 56.

Moreover, these tests rely on **algorithmic** notions of **partial recursive functions** and **incompressibility**.

The key to the duality between the stochastic and algorithmic perspectives is provided by:

"Martin-Löf's [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." (Li and Vitanyi, 2008, p. 146)

This duality can be used to dispel any lingering suspicions concerning the circularity of the frequentist interpretation of probability. This is because the Kolmogorov complexity framework provides an operational algorithmic (non-probabilistic) interpretation to all the above measure-theoretic results, including non-typical realizations defined on a set of measure zero, rendered as a countable set of recursively-enumerable sequences; see Nies (2009). That is, the notion of Kolmogorov complexity provides the first successful attempt to operationalize stochastic randomness, by ensuring the compliance of algorithmically incompressible sequences to the above measure theoretic results, including the SLLN (2) and the LIL; see chapter 9.

In a certain sense, 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 stochastic view.

▶ This link relies on the initial finite segment $\{x_k\}_{k=1}^{n}$ being 'typical', i.e. passing all effective randomness tests, and provides the first successful attempt to operationalize randomness, by ensuring the compliance of algorithmically incompressible sequences to the above measure theoretic results, including the SLLN (2) and the LIL.

▶ A key result for this elucidation is the notion of **pseudo-randomness**: sequences that exhibit statistical randomness while being generated by a deterministic recursive process.

In summary, the model-based frequentist and the algorithmic perspective based on Kolmogorov complexity, despite being grounded on entirely different mathematical formulations, share several features, including:

▶ the link between the measure-theoretic results and real-world phenomena is provided by viewing data $\mathbf{x}_0$ as a "typical realization" of the stochastic process $\{X_t,\ t\in\mathbb{N}\}$ underlying $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ and give rise to two **in sync complementary interpretations of frequentist probability**.

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 can be used to shed additional light on:

(a) Why **von Mises's (1928) 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; see Li and Vitanyi (2008), pp. 49–56.

(b) Dispelling certain confusions relating to charges leveled against the frequentist interpretation by **summoning infinite realizations** $\{x_k\}_{k=1}^{\infty}$, a well as any lingering doubts concerning the **circularity charge**.

(c) The algorithmic employs 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 $\{x_k\}_{k=1}^{\infty}$.

3.3 The Propensity Interpretation of Probability

The propensity interpretation is associated with the philosophers Charles Sanders Peirce (1839–1914) and Karl Raimund Popper (1902–1994) Popper; see Gillies (2000) and Gavalotti (2005). It interprets probability as a **propensity (disposition, or tendency)** of a real world stochastic mechanism to yield a certain **stable long-run relative frequency** of particular outcomes. The propensity interpretation is invoked to explain why such stochastic mechanisms will generate a given *outcome type* at a stable rate.

Example 10.2. It is well-known in physics that **a radioactive atom has a 'propensity to decay'** that gives rise to **stable relative frequencies**, despite the fact that the particular instant of the decay is unpredictable because it depends on an unobservable mechanism in the nucleus of the atom. Radioactive decay represents the process by which an atom with unstable atomic nucleus loses energy by emitting radiation in a variety of forms. Every **radioactive substance decays over time** in a law-like rate that can be accurately modeled using an exponential function:

$$X(t){=}X_0e^{-\lambda t},\ \ \lambda{>}0,$$

where $X(t)$ represents the amount of radioactive material present at time t, that is used for dating substances using their half-life period. For instance, the **half-life of radium-226** is 1590 years.

The propensity interpretation of probability has a clear affinity with the frequentist interpretation in so far as:

(i) [it] assumes the presence of a stochastic generating mechanism,
(ii) [it] is defined in terms of long-run stable relative frequencies, and
(iii) [it] views probability as a feature of the real world.

This affinity has generated confusion in the philosophy of science literature that classifies this interpretation as different from the frequentist interpretation; see Gillies (2000).

Causal asymmetry in probability. A particular example that is often used to contrast the two interpretations was proposed by Humphreys (1985) as a **paradox**. He argued above, the propensity interpretation associated with real world stochastic generating mechanisms carries with it a built-in **causal connection** between different events, say A and B, which renders reversing conditional probabilities such as $\mathbb{P}(A|B)$ to $\mathbb{P}(B|A)$ meaningless when A is the **effect** and B is the **cause**. This is viewed as indicating that the **propensity** interpretation **does not satisfy the basic rules of mathematical probability**.

Humphrey's paradox, however, can be easily explained away when one distinguishes between a **statistical model** $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$, and a **substantive model** $\mathcal{M}_{\boldsymbol{\varphi}}(\mathbf{x})$, where the two are related via certain parameter restrictions $\mathbf{G}(\boldsymbol{\theta},\boldsymbol{\varphi}){=}\mathbf{0}$; see Spanos (2006c). $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ is a purely probabilistic construal that comprises the probabilistic assumptions imposed on the data $\mathbf{x}_0$ and represents a particular parameterization of the stochastic process $\{X_k,\ k\in\mathbb{N}\}$ underlying $\mathbf{x}_0$. In the context of $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$, probabilities are **generic and consistent with the Kolmogorov axioms**. In contrast, $\mathcal{M}_{\boldsymbol{\varphi}}(\mathbf{x})$ is based on substantive subject matter information, including causal assumptions, and aims to approximate the real-world GM as faithfully as possible. In the context of $\mathcal{M}_{\boldsymbol{\varphi}}(\mathbf{x})$ probabilities could and often have causal interpretation assigned to them, including the case of a radioactive atom's decay. As argued in chapter 1, in empirical modeling one needs to separate the two models, ab initio, with a view to allow the substantive information in $\mathcal{M}_{\boldsymbol{\varphi}}(\mathbf{x})$ (including causality assumptions) to be tested against the data before being imposed. In this sense, there is no conflict between the frequentist and propensity interpretations of probability, as the former is germane to the statistical $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$, and the latter to the substantive model $\mathcal{M}_{\boldsymbol{\varphi}}(\mathbf{x})$.

4 Operationalizing the 'long-run' metaphor

The notion of pseudo-random sequences, exhibiting particular statistical regularities, can be used to operationalize the relevant 'long-run' metaphor of the frequentist interpretation.

Table 2 - Simple Normal Model

Statistical GM:	$X_k=\mu+u_k,\quad k\in\mathbb{N}:=\{1,2,...\}$	
[1] Normality:	$X_k\backsim\mathsf{N}(.,.),\ x_k\in\mathbb{R}:=(-\infty,\infty),$	$k\in\mathbb{N}.$
[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	

For this model, one can used the statistical GM:

$$X_k=\mu+\sigma\varepsilon_k,\ \varepsilon_k\backsim\mathsf{N}(0,1),\quad k=1,2,..., \tag{6}$$

to emulate the long-run metaphor by using the following algorithm.

Step 1: Specify values for (or estimate) the unknown parameters $\boldsymbol{\theta}\boldsymbol{:=}(\mu,\sigma^2)$.

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

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

Using the artificial data $\mathbf{X}_N$ one can construct the empirical counterparts to the sampling distribution of any statistic of interest, including the estimators $\widehat{\boldsymbol{\theta}}_n:=(\overline{X}_n,s^2)$.

This simulation algorithm operationalizes the model-based long-run metaphor and provides an 'empirical counterpart' to any relevant distribution of interest, including the evaluation of the empirical relative frequency corresponding to $\mathbb{P}(A)$, for any legitimate event A.

4.1 Error probabilities and relative frequencies

The above framing of the frequentist interpretation of probability for an event A is general enough to be extended in all kinds of different set-ups within frequentist inference, including **the error probabilities**. In the context of a statistical model $\mathcal{M}_{\theta}(\mathbf{x})$, $\mathbf{x}{\in}\mathbb{R}_X^n$, the sequence of data come in the form of N realizations $\mathbf{x}^1$, $\mathbf{x}^2, ..., \mathbf{x}^N$ from the same sample space $\mathbb{R}_X^n$.

Example. Consider the following hypotheses:

$$H_0\text{: } \mu \leq \mu_0, \text{ vs. } \quad H_1\text{: } \mu > \mu_0, \tag{7}$$

in the context of the *simple* (one parameter) *Normal model*:

$$\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})\text{: } X_t \backsim \mathsf{N}\left(\mu, \sigma^2\right), \ [\sigma^2 \text{ known}], \ t{=}1, 2, ..., n, ...,$$

for which the optimal test is $\mathcal{T}_{\alpha}{:=}\{\kappa(\mathbf{X}),\ C_1(\alpha)\}$:

$$\begin{array}{l} \text{test statistic: } \kappa(\mathbf{X}){=}\frac{\sqrt{n}(\overline{X}_n-\mu_0)}{\sigma},\ \overline{X}_n{=}\frac{1}{n}\sum_{k=1}^n X_k, \\ \text{rejection region: } C_1(\alpha){=}\{\mathbf{x}\text{: } \kappa(\mathbf{x}) > c_{\alpha}\}. \end{array} \tag{8}$$

To evaluate the error probabilities one needs the distribution of $\kappa(\mathbf{X})$ under H_0 and H_1:

[i] $\kappa(\mathbf{X}){=}\frac{\sqrt{n}(\overline{X}_n-\mu_0)}{\sigma} \overset{\mu=\mu_0}{\backsim} \mathsf{N}(0,1)$,

[ii] $\kappa(\mathbf{X}){=}\frac{\sqrt{n}(\overline{X}_n-\mu_0)}{\sigma} \overset{\mu=\mu_1}{\backsim} \mathsf{N}(\delta_1,1)$, $\delta_1{=}\frac{\sqrt{n}(\mu_1-\mu_0)}{\sigma}{>}0$ for all $\mu_1 > \mu_0$.

These hypothetical sampling distributions are then used to compare H_0 or H_1 via $\kappa(\mathbf{x}_0)$ to the true value $\mu{=}\mu^*$ represented by data $\mathbf{x}_0$ via $\overline{X}_n$, the best estimator of μ. The evaluation of the type I error probability and the p-value is based on [i] $\kappa(\mathbf{X}){=}\frac{\sqrt{n}(\overline{X}_n-\mu_0)}{\sigma} \overset{\mu=\mu_0}{\backsim} \mathsf{N}(0,1)$:

$$\begin{array}{l} \alpha{=}\mathbb{P}(\kappa(\mathbf{X}) > c_{\alpha}; \mu{=}\mu_0), \\ p(\mathbf{x}_0){=}\mathbb{P}(\kappa(\mathbf{X}) > \kappa(\mathbf{x}_0); \mu{=}\mu_0), \end{array}$$

and the evaluation of type II error probabilities and power is based on:

$$\text{[ii] } \kappa(\mathbf{X}){=}\frac{\sqrt{n}(\overline{X}_n-\mu_0)}{\sigma} \overset{\mu=\mu_1}{\backsim} \mathsf{N}(\delta_1,1), \text{for } \mu_1 > \mu_0.$$

$$\beta(\mu_1){=}\mathbb{P}(\kappa(\mathbf{X}) \leq c_{\alpha}; \mu{=}\mu_1) \text{ for all } \mu_1 > \mu_0.$$

$$\pi(\mu_1){=}\mathbb{P}(\kappa(\mathbf{X}) > c_{\alpha}; \mu{=}\mu_1) \text{ for all } \mu_1 > \mu_0.$$

How do these error probabilities fit into the above frequentist interpretation of probability that revolves around the long-run metaphor?

Type I error probability. The event of interest for the evaluation of α is:

$$(Z{=}1){:=}A{=}\{\mathbf{x}\text{: } \kappa(\mathbf{x}) > c_{\alpha}\},\ \forall\mathbf{x}{\in}\mathbb{R}^n,$$

and the distribution where the probabilities come from is:

$$[\text{i}]\ \kappa(\mathbf{X}){=}\tfrac{\sqrt{n}(\overline{X}_n-\mu_0)}{\sigma}\ \overset{\mu=\mu_0}{\backsim}\ \mathsf{N}(0,1).$$

One draws N IID samples of size n from $\mathsf{N}(0,1)$, that give rise to the realizations $\mathbf{x}^1$, $\mathbf{x}^2,...,\mathbf{x}^N$.For each sample realization one evaluates $\kappa(\mathbf{x}^i)$ and considers the relative frequency of event A occurring. That relative frequency is the sample equivalent to the significance level α.

The power of the test. The event of interest is:

$$(Z{=}1){:=}A{=}\{\mathbf{x}:\kappa(\mathbf{x})>c_\alpha\},\ \forall\mathbf{x}{\in}\mathbb{R}^n,$$

but the distribution from where the realizations $\mathbf{x}^1$, $\mathbf{x}^2,...,\mathbf{x}^N$ come from is:

$$[\text{ii}]\ \kappa(\mathbf{X}){=}\frac{\sqrt{n}(\overline{X}_n-\mu_0)}{\sigma}\ \overset{\mu=\mu_1}{\backsim}\ \mathsf{N}(\delta_1,1),\ \mu_1>\mu_0.$$

The same evaluation as that associated with α will now give rise to the relative frequency associated with power of the test at $\pi(\mu_1)$ for a specific μ_1.

The type II error probability. The event of interest is also

$$(Z{=}1){:=}A{=}\{\mathbf{x}:\kappa(\mathbf{x})\leq c_\alpha\},\ \forall\mathbf{x}{\in}\mathbb{R}^n,$$

and the distribution from where the realizations $\mathbf{x}^1$, $\mathbf{x}^2,...,\mathbf{x}^N$ come from is

$$[\text{ii}]\ \kappa(\mathbf{X}){=}\tfrac{\sqrt{n}(\overline{X}_n-\mu_0)}{\sigma}\ \overset{\mu=\mu_1}{\backsim}\ \mathsf{N}(\delta_1,1),\ \mu_1>\mu_0.$$

The p-value. The event of interest for the evaluation of the p-value is:

$$(Z{=}1){:=}A{=}\{\mathbf{x}{:}\ \kappa(\mathbf{x})>\kappa(\mathbf{x}_0)\},\ \forall\mathbf{x}{\in}\mathbb{R}^n,$$

and the distribution where the probabilities come from is [i] $\kappa(\mathbf{X}){=}\tfrac{\sqrt{n}(\overline{X}_n-\mu_0)}{\sigma}\ \overset{\mu=\mu_1}{\backsim}\ \mathsf{N}(0,1)$. The data specificity of the p-value does not matter in this case because:

$$p(\mathbf{X})\ \overset{\mu=\mu_0}{\backsim}\ \mathsf{U}(0,1),$$

which implies that $\mathbb{P}(\kappa(\mathbf{X})<c;\mu{=}\mu_0){=}c$.

Post-data severity. The event of interest will be either of the events

$$(Z{=}1){:=}A{=}\{\mathbf{x}:\kappa(\mathbf{x})\gtrless\kappa(\mathbf{x}_0)\},\ \forall\mathbf{x}{\in}\mathbb{R}^n,$$

depending on the inferential claim $\mu\gtrless\mu_0{+}\gamma$ evaluated. The distribution from where the realizations $\mathbf{x}^1$, $\mathbf{x}^2,...,\mathbf{x}^N$ come from is [ii] $\kappa(\mathbf{X}){=}\tfrac{\sqrt{n}(\overline{X}_n-\mu_0)}{\sigma}\ \overset{\mu=\mu_1}{\backsim}\ \mathsf{N}(\delta_1,1)$, with one **caveat**: the legitimate realizations $\mathbf{x}^1$, $\mathbf{x}^2,...,\mathbf{x}^N$ should take values $\kappa(\mathbf{x}_0)\pm\varepsilon$. This is necessary because under $\mu{=}\mu_1$ the distribution associated with events $\{\mathbf{x}{:}\ \kappa(\mathbf{x})\gtrless\kappa(\mathbf{x}_0)\}$ is *not* Uniformly distributed.

4.2 Enumerative vs. model-based induction

A closer look at the philosophy of science literature concerning the frequentist interpretation of probability reveals that the SLLN has been invoked, implicitly or explicitly, for two different, but related, tasks. The first has to do with the justification of the frequentist interpretation itself, but the second is concerned with the justification of the **straight rule** as a form of *inductive inference.*

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).

The above discussion has called into question the latter claim by bringing out the crucial differences between that and model-based induction.

In relation 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 $\overline{x}_n = \frac{m}{n}$, to infer that: $\lim_{n\to\infty} \overline{x}_n = \frac{m}{n}$." (p. 86)

The primary justification for this rule is that asymptotically (as $n \to \infty$) $\overline{x}_n$ converges to the true probability θ; NO such result can be mathematically established! Indeed, there is nothing in **model-based point estimation** that could justify the **inferential claim** $\overline{x}_n \simeq p$.

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, $\overline{x}_n \pm \varepsilon$, originally proposed by Reichenbach (1934), and argued for codifying the error ε in terms of de Finetti's subjective interpretation of probability.

▶ Viewing the straight rule $\mathbb{P}(A) = \frac{m}{n}$ in the context of the error statistical perspective, it becomes clear that none of these proposals provides an adequate justification for it as an inferential procedure.

▶ What has not been sufficiently appreciated in these discussions is how *model-based induction* focalizes the 'signal' by distilling the data into a parsimonious statistical model that enhances both the reliability and precision of inference.

▶ From the error statistical perspective the relevant statistical model, implicit in the discussion, is the simple Bernoulli $\mathcal{M}_\theta(\mathbf{x})$ (table 1) with $\mathbb{P}(A)=\theta$. Viewing $\overline{x}_n$ in the context of $\mathcal{M}_\theta(\mathbf{x})$ reveals that one knows much more about $\overline{x}_n$ as an *estimate* of θ than the straight rule suggests.

The SLLN asserts that $\overline{X}_n = \frac{1}{n}\sum_{k=1}^{n} X_k$ is a *strongly consistent* estimator of θ, which secures only minimal reliability because the result in (2) is necessary, but not sufficient for the reliability of inference for a given n; that calls for the relevant error probabilities.

▶ Any attempt to evaluate such **error probabilities relying exclusively on the SLLN** will give rise to very crude results because they are invariably based on inequality bounds.

For instance, when invoking Borel's SLLN, arguably the best inequality one can use is *Hoeffding's* (Wasserman, 2006):

$$\mathbb{P}\left(|\overline{X}_n - \theta| \geq \varepsilon\right) \leq 2\exp\left(-2n\varepsilon^2\right), \quad \text{for any } \varepsilon > 0. \tag{9}$$

In contrast, the **model-based frequentist approach** makes full use of the model assumptions [1]-[4] (table 2) to derive the **exact sampling distribution**:

$$\overline{X}_n \backsim \ \mathsf{Bin}\left(\theta, \tfrac{\theta(1-\theta)}{n}\right), \tag{10}$$

where 'Bin' denotes the Binomial distribution.

Contrasting (9) and (10) brings out the key difference between **enumerative** and **model-based induction** in so far as these bounds turn out to be very crude, giving rise to imprecise error probability evaluations; Spanos (1999).

To illustrate this let n=100, θ=.5 and ε=.1: (9) yields:

$$\mathbb{P}\left(|\overline{X}_n - \theta| \geq .1\right) \leq .271, \text{ compared to } \mathbb{P}\left(|\overline{X}_n - \theta| \geq .1\right) = .0455,$$

given by (10). Such a sixfold imprecision in error probabilities undermines completely the reliability of any inference!

Focusing on reliable and precise inferences, (10) can be used to construct a $(1-\alpha)$ Confidence Interval:

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

which, apropos, provides a proper frequentist interpretation to Reichenbach's approximate straight rule:

$$\overline{X}_n \pm \varepsilon_n,\ \varepsilon_n = c_{\frac{\alpha}{2}}\sqrt{Var(\overline{X}_n)}.$$

The difference is that, when the statistical adequacy of $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ (table 1) has been secured, one can assess the reliability, as well as the precision, of this rule, using the associated error probabilities. There is no reason to invoke as $n \to \infty$.

Taking stock: model-based frequentist interpretation

► In addition to demarcating explicitly the probabilistic premises of inference and rendering them testable vis-a-vis the data, frequentist model-based induction has extended the scope of induction beyond IID processes to include general statistical models with dependence and/or heterogeneity.

► It has enhanced the reliability and precision of inductive inferences by grounding them on finite sampling distributions rather than relying solely on asymptotic results like the SLLN.

5 The ‘single case’ and the ‘reference class’

5.1 Revisiting the problem of the ‘single case’ probability

A crucial criticism of the frequentist interpretation of probability raised in philosophy of science literature has been on (c) **applicability** grounds, in so far as it cannot be used to assign probabilities to **single case events**.

According to Salmon (1967):

“The frequency interpretation also encounters applicability problems in dealing with the use of probability as a guide to such practical action as betting. We bet on single occurrences: a horse race, a toss of the dice, a flip of a coin, a spin of the roulette wheel. The probability of a given outcome determines what constitutes a reasonable bet. 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.” (ibid. p. 90)

This passage raises two separate issues.

▶ The **first** concerns a notion of probability for ‘individual decision making (betting) under uncertainty’. This might call for a different interpretation of probability, but I leave that issue aside.

▶ The **second** issue concerns the charge that the frequentist interpretation cannot be used to assign a probability to events such as: ‘heads’ on a single flip of a coin, a ‘six’ on the next toss of a dice, or ‘red’ on a single spin of the roulette wheel. To a frequentist statistician this charge seems totally bizarre because there is no difficulty attaching a probability to the event:

$$A_{k+1}=\{X_{k+1}=1\} \text{ -‘heads’ on the next toss of the coin,}$$

since it is a generic event – an event within the intended scope of $\mathcal{M}_{\theta}(\mathbf{x})$. The probabilistic assignment is straightforward:

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

and presents no conceptual or technical difficulties.

In light of this, why do philosophers of science keep reiterating this charge?

Perhaps the only way to explain its persistence is in terms of misidentifying the frequentist interpretation of probability with von Mises's variant. Salmon's last sentence reads like a paraphrasing of von Mises's (1957) 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**’ in this quotation, the single event probability charge fades away in model-based induction. That is, when the single event of interest belongs to the intended scope of a particular model $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$, $f(\mathbf{x};\boldsymbol{\theta})$ assigns probabilities to all such *generic* events.

What does replacing a '**collective**' with a **statistical model** $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ accomplish for frequentist modeling and inference?

A. The notion of probability used in the context of $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ follows directly from **Kolmogorov's axioms** and nothing else.

B. There is nothing arbitrary about the choice of an appropriate $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ in the context of the model-based induction framework, because it depends crucially on being **statistically adequate** vis-à-vis data $\mathbf{x}_0$.

5.2 Assigning probabilities to 'singular events'

Sometimes, the single case probability is raised, not in terms of a generic event like:

A – a **randomly selected** individual from the population of 40-year old Englishmen, will die before his next birthday,

but in relation to a seemingly interchangeable *singular event* (Gillies, 2000:

B – Mr Smith, an Englishman who is 40 today, will die before his next birthday.

The charge is that the frequentist interpretation cannot be used to assign probabilities to events like B because the **long-run makes no sense** in this case.

The question is: are events A and B interchangeable when viewed in the context of model-based induction?

The implicit statistical model $\mathcal{M}_{\theta}(\mathbf{x})$ that includes A as a legitimate (generic) event is the simple Bernoulli (table 2) which aims to provide an idealized description of the survival of a target population (40-year old Englishmen), **treating each individual generically**: a randomly selected individual survives (X=1) or dies (X=0), before his next birthday with $\mathbb{P}(X{=}1){=}\theta$.

When Mr Smith is *randomly selected* from the target population, one can attach the same probability to A as to B because they are indistinguishable.

However, when Mr Smith is *not* randomly selected – he cannot be viewed as a generic individual – no probability can be attached to event B in the context of $\mathcal{M}_{\theta}(\mathbf{x})$ because the latter requires every individual X_k in the sample $\mathbf{X}{:=}(X_1, ..., X_n)$ to be generic (IID); purposeful selection precludes that.

Hence, assigning a probability to event B is problematic because it lies outside the intended scope of $\mathcal{M}_{\theta}(\mathbf{x})$, and that has **no bearing** on the long-run frequentist interpretation.

Common sense suggests that in the context of $\mathcal{M}_{\theta}(\mathbf{x})$ the only relevant information is whether Mr Smith – as a generic individual – will survive past his next birthday or not, because $\mathcal{M}_{\theta}(\mathbf{x})$ was meant to be an idealized description of the survival of **the population as a whole**.

On the other hand, if one is actually interested in **Mr Smith's survival per se**, a **very different statistical model** is called for.

▶ For instance, a **logit model** $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})$, based on the vector process:

$$\{\mathbf{Z}_t{:=}(y_k, \mathbf{W}_k),\ k{\in}\mathbb{N}\},$$

whose statistical Generating Mechanism (GM) is:

$$y_k = \tfrac{\exp(\boldsymbol{\theta}^\top \mathbf{w}_k)}{1+\exp(\boldsymbol{\theta}^\top \mathbf{w}_k)} + u_k,\ k{\in}\mathbb{N}, \tag{11}$$

$$\boldsymbol{\theta}^\top \mathbf{w}_k{=}\textstyle\sum_{j=1}^{m} \beta_j w_{jk},\ \mathbf{W}_k{:=}(W_{1k}, ..., W_{mk}),$$

$$E(y_k|\mathbf{W}_k{=}\mathbf{w}_k){=}\tfrac{\exp(\boldsymbol{\theta}^\top \mathbf{w}_k)}{1+\exp(\boldsymbol{\theta}^\top \mathbf{w}_k)}{=}\mathbb{P}(y_k{=}1|\mathbf{W}_k{=}\mathbf{w}_k){=}p(\mathbf{w}_k).$$

$\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})$ is envisaged as an idealized description of Mr Smith's survival y_k as it relates to potential contributing factors $\mathbf{W}_k$, such as **age**, **family medical history**, **smoking habits**, **nutritional habits**, **stress factors**, etc.; see Balakrishnan and Rao (2004).

Not surprisingly, in the context of $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})$ assigning a probability to B makes perfectly good sense, and so does the long-run frequentist interpretation. What is more, this repudiates the view expressed by von Mises (1957):

"We can say nothing about the probability of death of an individual even if we know his condition of life and health in details." (p. 11)

In summary, from the **model-based frequentist perspective**, probabilities of events of interest are defined in the context of a statistical model $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ whose structure renders events like A legitimate (generic), but events like B illegitimate, on statistical adequacy grounds. Event B calls for a very different statistical model.

5.3 Revisiting the 'reference class' problem

Related to the 'single case' probability is the **reference class problem** where it is argued that since Mr Smith's survival y_k can be related to **several different factors:**

$$\mathbf{W}_k{:=}(W_{1k}, W_{2k}, ..., W_{mk})$$

the frequentist probability of y_k will be different when the reference class is relative to each of these distinct potential factors; see Hajek (2007).

A closer look at this argument reveals that it stems from inadequate understanding of the role of a s**tatistical model** since the multiplicity of potential contributing factors in $\mathbf{W}_k$ does not render the frequentist interpretation problematic in any sense. On the contrary, a most reliable way to address the multiplicity problem is to combine all the potential factors into a single statistical model $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})$, like (11), specified in terms of the stochastic process

$$\{\mathbf{Z}_t{:=}(y_k, \mathbf{W}_k),\ k{\in}\mathbb{N}\},$$

aiming to describe how these factors (collectively and individually) are likely to influence Mr Smith's survival y_k.

Having said that, one might argue that a more sympathetic interpretation of the 'reference class' problem is that it concerns the selection of the **'correct' subset**, say

$\mathbf{W}_{1k}$, of the relevant contributing factors in $\mathbf{W}_k$ giving rise to an adequate explanation for y_k. Again, this suggests inadequate appreciation of the role of **substantive information vs. a statistical model**.

An encompassing statistically adequate model $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})$, like (11), offers an effective way to address the potential *confounding* problems in appraising the substantive significance of different potential factors. Delineating the role of these potential effects raises genuine *substantive adequacy* issues pertaining to whether a 'structural' model

$$\mathcal{M}_{\boldsymbol{\varphi}}(\mathbf{z}_1),\ \mathbf{z}_1{:=}(\mathbf{y},\mathbf{W}_1)$$

provides a veritable explanation for the phenomenon of interest (Spanos, 2006).

Securing substantive adequacy raises additional issues and often calls for further probing of (potential) errors in bridging the gap between $\mathcal{M}_{\boldsymbol{\varphi}}(\mathbf{z}_1)$ and the phenomenon of interest. This problem, however, has no bearing on the frequentist interpretation of probability per se.

Finally, a closer look at the examples used to articulate the reference class problem (Hajek, 2007), reveals that the **difficulties stem primarily from the restrictive and overly simplistic nature of statistical models**, like the Bernoulli model $\mathcal{M}_{\theta}(\mathbf{x})$ (table 1), implicitly invoked by enumerative induction.

In that sense, the discussions pertaining to the choice of the reference class, like '**the broadest homogeneous**' (Salmon, 1967, p. 91), beg the question

'**homogeneous with respect to what dimension (ordering)**?'

whose answer would invariably intimate **certain omitted variable(s)**; a substantive adequacy issue! These can be viewed as *ad hoc* attempts to extend these simple models to accommodate additional (potentially) relevant variables ($\mathbf{W}_k$), demarcating the relevant reference class.

Viewed in the context of **model-based modeling**, these attempts can be formalized using **logit-type models** like (11) for different sub-groups ($i{=}1,2,...,\ell$) of the original population (classified by gender, race, ethnicity etc.). The idea is that if there is homogeneity within but heterogeneity between these groups, the heterogeneity in the probability of survival p_i might be explainable by certain conditioning variables $\mathbf{W}_k$:

$$p_i(\mathbf{w}_k){:=}E(y_{ik}\mid\mathbf{W}_{ik}{=}\ \mathbf{w}_{ik}){=}\tfrac{\exp(\boldsymbol{\theta}_i^{\top}\mathbf{w}_{ik})}{1+\exp(\boldsymbol{\theta}_i^{\top}\mathbf{w}_{ik})},\quad i{=}1,2,...,\ell. \tag{12}$$

This transforms the original (nebulous) reference class problem into a (clear) modeling question that concerns the deliberate selection of the relevant variables $\mathbf{W}_k^*$ so that a model based on $f(y_{ik}\mid\mathbf{w}_{ik}^*;\boldsymbol{\theta})$ is both statistically and substantively adequate.

In summary, the difficulties associated with the reference class problem amount to **posing question(s) of interest in the context of an inappropriate model**; one that does not contain the information sought.

▶ The reasons for that might be practical (the right data are unavailable), or conceptual (one cannot think of a model), but neither of these deficiencies can be blamed on the model-based frequentist interpretation of probability.

▶ Indeed, one can make a case that error statistics has paved the way for addressing the issues raised by the reference class problem by transforming them into modeling questions in the context of general statistical models beyond the overly simplistic ones implicitly invoked by enumerative induction.

6 Summary and conclusions

The error statistical perspective identifies the probability of an event A with the *limit* of its relative frequency of occurrence – invoking the SLLN – in the context of a statistical model $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x}),\ \mathbf{x}\in\mathbb{R}^n_X$.

This frequentist interpretation is defended against the charges of:
(i) the circularity of its definition,
(ii) its reliance on 'random samples',
(iii) its inability to assign 'single event' probabilities, and
(iv) the 'reference class' problem,
by showing that the perceived target is unduly influenced by enumerative induction and the von Mises rendering of the frequentist interpretation.

An important feature of the error-statistical view of randomness is its duality to an algorithmic view based on the notion of *Kolmogorov complexity.* Both perspectives adopt the same *interpretive provisions:*

[i] data $\mathbf{x}_0{:=}(x_1, x_2, \ldots, x_n)$ is viewed as a '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 M-S testing.

This links mathematical results, such as the SLLN and LIL, to the actual data-generating mechanism – data $\mathbf{x}_0$ is viewed as a 'typical realization' of the process $\{X_k,\ k{\in}\mathbb{N}\}$ – but are grounded on entirely different mathematical formulations.

The Kolmogorov complexity provides a purely *non-probabilistic* (algorithmic) rendering that operationalizes all the measure-theoretic results associated with the probabilistic perspective.

In model-based induction there is no difficulty in assigning probabilities to **any legitimate (generic) event** A within the model's ($\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$) intended scope.

▲ It is argued that the difficulties raised by the '**singular event**' probability and the '**reference class**' problems stem from posing questions of interest in the context of **nebulous and incomplete inductive premises**.

▲ Error statistics paves the way for addressing these issues by transforming them into well-defined modeling questions in the context of statistical models beyond the simple ones (IID) invoked by enumerative induction.

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 $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$, 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 $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$,

[d] it provides an empirical justification for frequentist induction stemming from securing the statistical adequacy of $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ 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**.