Error Statistics and the Frequentist Interpretation of Probability[1]

Aris Spanos

Department of Economics

Virginia Tech

June 2009

Outline:

1. **Introduction: Enumerative vs. Model-based Induction**
2. **Error statistics: a refinement/extension of the Fisher-Neyman–Pearson (F-N-P) framework**
3. **The frequentist interpretation of probability**
4. **Enumerative vs. Model-based Induction**
5. **The Single Case Probability and the Reference Class Problems**
6. **Frequentist vs. Epistemic Probability in Statistical Inference**
7. **Summary and conclusions**

---

[1] **Presentation at the conference on the "Multiplicity and Unification in Statistics and Probability", 25-26 June, 2009, The University of Kent, Canterbury.**

1 Introduction: enumerative vs. model-based induction

The frequentist approach to statistical inference has been largely ignored by the recent philosophy of science literature due primarily to various seemingly crucial inadequacies of the frequentist interpretation of probability and various serious flaws in its underlying inductive reasoning that gave rise to several fallacious and counter-intuitive results; see Savage (1962), Hacking (1965), Salmon (1967), Kyburg (1974), Gillies (2000), Howson (2000), Howson and Urbach (2005).

This paper focuses on charges against the frequentist interpretation of probability:
(i) the circularity of its definition, (ii) its reliance on 'random samples',
(iii) the 'single event' probability issue, (iv) the 'reference class' problem,
(v) the incompleteness of the evidential interpretation provided by severity.

▶ Charges (i)-(ii) are misplaced because the frequentist interpretation of probability is often misleadingly identified with von Mises's (1928) rendering, instead of the 'stable long-run frequencies' variant relying on the SLLN.

▶ Charges (iii)-(iv) is due to insufficient appreciation of the differences between induction by enumeration and the practitioners' model-based induction.

▶ Any attempt to use epistemic probability to supplement the evidential interpretation of frequentist inference raises serious incompatibility issues.

2 Error statistics: a refinement/extension of the Fisher-Neyman-Pearson (F-N-P) framework

The **refinement** is primarily concerned with securing the reliability of inference, and the **extension** with addressing chronic foundational problems, including the fallacies of acceptance and rejection, the large *n* problem, statistical vs. substantive significance, and several Bayesian-instigated confusions; Mayo & Spanos (2006).

2.1 The traditional frequentist approach

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 (see Pearson, 1920), into a frequentist 'model-based induction', relying on *finite sampling distributions*.

► Unlike Karl Pearson, who would commence with data $\mathbf{x}_0$:=$(x_1, ..., x_n)$ in search of a frequency curve $f(x;\boldsymbol{\theta})$ to describe $\mathbf{x}_0$, Fisher proposed to begin with:
(a) a prespecified model $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ ("a hypothetical infinite population"), and
(b) view $\mathbf{x}_0$ as a typical realization thereof, by rendering the specification of $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ 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 *a posteriori*" (ibid., p. 314). In sharp contrast to *a priori* justifications. Fisher went on to formalize notions (a)-(b) in purely *probabilistic* terms by defining the concept of a (*parametric*) *statistical model*:

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

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

Example - The simple Normal model: $X_k \backsim \mathsf{NIID}(\mu, \sigma^2),\ k=1, 2, ..., n, ...$

The problem of ***statistical model specification*** [selecting the prespecified statistical model $\mathcal{M}_\theta(\mathbf{x})$ from the set of all possible models $\mathcal{P}(\mathbf{x})$] has *not* been adequately addressed in frequentist statistics; see Lehmann (1990), Cox (1990).

2.2 Error Statistics and the reliability of inference

Error statistics provides a reconciliation of the Fisherian and N-P perspectives that weaves a coherent frequentist inductive reasoning anchored firmly on error probabilities. The role of a statistical model $\mathcal{M}_\theta(\mathbf{x})$ in error statistics is manifold: (i) to specify the premises of inference, (ii) to assign probabilities to all events of interest and related events, and (iii) to provide ascertainable error probabilities in terms of which one can assess the optimality and reliability of inference methods.

2.3 Reliability of inference

▶ Statistical adequacy refers to the validity–vis-à-vis data $\mathbf{x}_0$–of the probabilistic assumptions comprising the statistical model $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ in question, and provides the sole criterion for 'when $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ accounts for the regularities in data $\mathbf{x}_0$.'
In error statistics the reliability of inference is evaluated in terms of the relevant *error probabilities* associated with different forms of inference.
Error-reliability. Statistical adequacy renders the relevant error probabilities ascertainable by ensuring that the *nominal* error probabilities for assessing substantive claims are approximately equal to the *actual* ones. The surest way to draw invalid inferences is to apply a .05 significance level test when its actual type I error probability is closer to .99 due to misspecification; see Spanos (2006a).
A. A purely probabilistic construal of a statistical model.
The most difficult hurdle in addressing the question, 'how does one assess the adequacy of $\mathcal{M}_\theta(\mathbf{x})$ *a posteriori*?' has been to determine the role of *substantive* subject matter information; see Lehmann (1990).
★ The key to dealing with this difficulty is to find a way to distinguish, *ab initio*, between *statistical* and *substantive* information. Such a distinction was first proposed in Spanos (1986), where *statistical information* pertains to the

chance regularity patterns exhibited by data $\mathbf{x}_0$ when viewed as a realization of a *generic* stochastic process $\{X_k,\ k\in\mathbb{N}\}$, irrespective of what the data quantify.

► This provides a purely probabilistic construal of a statistical model $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$, by viewing it as a particular *parameterization* of the probabilistic structure of a process $\{X_k,\ k\in\mathbb{N}\}$ that would have rendered $\mathbf{x}_0$ a typical realization thereof.

► *Substantive* subject matter information often comes in an estimable form – in view of data $\mathbf{x}_0$ – of the theory, often called a structural model $\mathcal{M}_{\boldsymbol{\varphi}}(\mathbf{x})$.

B. Reconciling the substantive and statistical information

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

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

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

Statistical adequacy is necessary for appraising **substantive adequacy**.

Error statistics provides the framework for securing these objectives by:
▶ (i) specifying $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ in terms of a complete list of (internally consistent) probabilistic assumptions, in a form that is testable vis-à-vis data $\mathbf{x}_0$, and
▶ (ii) supplementing that with a statistical Generating Mechanism (GM) to provide a bridge between the statistical and substantive information (Cox, 1990).
Given a stochastic process $\{X_k,\ k\in\mathbb{N}\}$ defined on a probability space $(\Omega, \mathfrak{F},\ \mathbb{P}(.))$, the statistical GM can be specified in the form of an orthogonal decomposition:

$$X_k = E(X_k \mid \mathcal{D}_k) + u_k, \quad k\in\mathbb{N}, \tag{2.3.1}$$

where $E(X_k \mid \mathcal{D}_k)$ and u_k denote the *systematic* and *non-systematic* components, respectively. the relevant conditioning information set $\mathcal{D}_k \subset \mathfrak{F}$ is chosen to render $\{(u_k|\mathcal{D}_k),\ k\in\mathbb{N}\}$ a martingale difference process; see Spanos (1999).
Example. The simple Normal model, viewed from the error statistical perspective, is given in table 1 in terms of the statistical GM:

$$X_k = \mu + \sigma\varepsilon_k,\ \varepsilon_k \backsim \mathsf{N}(0,1), \quad k=1,2,..., \tag{2.3.2}$$

specifying explicitly how one can generate typical realizations of the NIID process $\{X_k,\ k\in\mathbb{N}\}$, using pseudo-random numbers, and assumptions [1]-[4].

Table 1 - Simple Normal Model

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

That is, the statistical GM (iii) **operationalizes** the hypothetical 'long-run'.

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

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

Step 3: Feed sequentially each $\boldsymbol{\varepsilon}^{(k)}$ into: $\boxed{\mathbf{x}^{(k)} = \mathbf{1}\mu + \sigma\boldsymbol{\varepsilon}^{(k)},}$

$\mathbf{1}$:=$(1,...,1)^\top$, to generate the artificial data: $\mathbf{X}_N$:=$(\mathbf{x}^{(1)}, \cdots, \mathbf{x}^{(N)})$, $\mathbf{x}^{(k)}$:=$(x_1, ..., x_n)^\top$.

The artificial data $\mathbf{X}_N$ can be used to construct the empirical counterparts to the sampling distributions of of the estimators $\widehat{\boldsymbol{\theta}}_n$:=$(\overline{X}_n, s^2)$.

■ For every statistical model there is a statistical GM which operationalizes the 'long-run' metaphor using pseudo-random numbers.

3 The frequentist interpretation of probability

The frequentist interpretation of probability has been called into question in the philosophy of science literature since the 1930s, by focusing on the weaknesses of the von Mises (1928) rendering; Salmon (1967), Gillies (2000), Hajek (2007).

3.1 Mathematical 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 $\boxed{\sigma\text{-field}}$ 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 $i = 1, 2, ..., n, ...$ 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 $A_i \in \mathfrak{F}$, $i=1, ..., n, ..$, such that $A_i \cap A_j = \emptyset$, 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 *measure theory* concerned more broadly with assigning size, length, content, area, volume, etc. to sets.
The notion of an event in probability plays an analogous role to the notion of a point in geometry. The notion of a σ-field [the events of interest and related events] $\mathfrak{F}$ is crucial for the single case probability and the reference class problems.
▶ A crucial question that arises is whether the above Kolmogorov formulation can be given an interpretation by assigning a meaning to the primitive term probability.
■ The relationship between mathematics and empirical modeling adopted in this paper has been first articulated by **Harald Cramer** (1946), p. 332:
"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."
▶ The interpretation discussed here relates to the particular objective of modeling observable stochastic phenomena of interest; they exhibit chance regularity

patterns. The primary aims of this interpretation are to: (a) facilitate the task of bridging the gap between such phenomena and the mathematical set up, and (b) shed additional light on empirical modeling and inference.

▶ It is argued that the frequentist interpretation, not only achieves the objectives stated above, but contrary to the conventional wisdom, it does meet Salmon's (1967) criteria of adequacy: Admissibility, Ascertainability and Applicability.

▼ The above empirical modeling objective should be contrasted to the use of probability to model individual decision making under uncertainty.

3.2 Relative frequencies and probabilities

Frequentist is the interpretation that identifies the probability of an event A with the limit of the relative frequency of its occurrence in a large number of trials. An important extension of the original Kolmogorov formulation $(\Omega, \mathfrak{F}, \mathbb{P}(.))$ is the notion of a *random variable* (r.v.), a real-valued function:

$$X(.)\colon \Omega \to \mathbb{R}{:=}(-\infty, \infty),$$

such that its pre-image defines events in $\mathfrak{F}$: $X^{-1}(-\infty, x] \in \mathfrak{F}$ for all $x \in \mathbb{R}$, i.e. $X(.)$ attaches numbers to the elementary events in Ω in such a way so as to

preserve the original event structure of interest ($\mathfrak{F}$).
The relevant random variable for the frequentist interpretation is defined by:

$$\{X{=}1\}{=}\{A \text{ occurs}\},\ \{X{=}0\}{=}\{\overline{A} \text{ occurs}\}, \text{ where } A{\in}\mathfrak{F} \text{ and } \mathbb{P}(A){=}p.$$

Repeating this experiment under identical conditions generates the stochastic process $\{X_k,\ k{\in}\mathbb{N}\}$ where $s_n{=}\frac{1}{n}\sum_{k=1}^{n} x_k{=}\frac{m}{n}$ is the relative frequency of occurrence of event A, which converges to p as $n \to \infty$.
■ This *'long-run' metaphor* enables one to conceptualize the notion of probability using the relative frequency s_n, but it is not intended to replace the notion of probability as a 'measure' attached to events in $\mathfrak{F}$. This metaphor is intended to help elucidate certain aspects of statistical modeling and inference.

3.3 The Strong Law of Large Numbers

A formal justification for the frequentist interpretation stems from the Strong Law of Large Numbers (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 *Strong Law of Large Numbers (SLLN)* asserts that for an *Independent and Identically Distributed (IID) Bernoulli* process $\{X_k,\ k{\in}\mathbb{N}\}$:

$$\boxed{\mathbb{P}(\lim_{n\to\infty}(\tfrac{1}{n}\sum_{k=1}^{n}X_k)=p)=1.} \qquad (3.3.3)$$

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 (2003).

"The relative frequency of occurrence of A converges to $\mathbb{P}(A){=}p$" needs to be qualified by clarifying the notion of convergence in (3.3.3) and delineating what the result does and does *not* mean.

▶ *First*, the result $S_n \overset{a.s.}{\to} p$ does *not* involve any claims that the sequence of numbers $\{s_n\}_{n=1}^{\infty}$ converges to p in a purely mathematical sense: $\lim_{n\to\infty} s_n = p$. The confusion between the probabilistic convergence of $\{S_n\}_{n=1}^{\infty}$ and the mathematical convergence of $\{s_n\}_{n=1}^{\infty}$ can be traced back to the notion of a *collective*, defined in terms of a realization $\{x_k,\ k{\in}\mathbb{N}\}$, proposed by von Mises (1928).

■ Now we know that any attempt to make rigorous the mathematical convergence of relative frequencies $\{s_n\}_{n=1}^{\infty}$ 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 ..." (Williams, 2001, p. 25)

▶ *Second*, the SLLN asserts that $\mathbb{P}\left(\lim_{n\to\infty} S_n(\omega)=p\right)=1$, where the result (3.3.3) holds everywhere in a domain $D=\{\omega\colon \lim_{n\to\infty} S_n(\omega)=p,\ \omega\in\Omega\}$ except on **a set of measure zero:** $D_0=\{\omega\colon \lim_{n\to\infty} S_n(\omega)\neq p,\ \omega\in\Omega\}$, i.e. $\mathbb{P}(D)=1$ and $\mathbb{P}(D_0)=0$.
▶ *Third*, the result in (3.3.3) holds only when $\{X_k,\ k\in\mathbb{N}\}$ satisfies certain probabilistic assumptions, such as IID. The *converse* result: for any IID process $\{X_k,\ k\in\mathbb{N}\}$, if $\{S_n\}_{n=1}^{\infty}$ converges a.s. to p, then $E(X_k)$ exists and is equal to p.
▶ *Fourth*, the result in (3.3.3) 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 the Law of Iterated Logarithm (LIL); Billingsley (2003).

3.4 The circularity charge

One of the issues often raised is that the justification of the frequentist interpretation of probability invoking (3.3.3) is *circular*! Lindley (1965), p. 5:
"... 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 some notable probabilists including Borel. Renyi (1970, p. 159) argues that the concept of probability in the above SLLN is purely a mathematical concept, and as such it cannot suffer from circularity.

▶ The SLLN is a theorem whose assertions revolve around the Kolmogorov mathematical formalism in measure-theoretic language. Indeed, a closer look at the word 'unlikely' that Lindley argues renders the argument circular, shows that it refers to the convergence of $\{S_n\}_{n=1}^{\infty}$ [not $\{s_n\}_{n=1}^{\infty}$], which holds everywhere in a domain D except on a set D_0 of measure zero, i.e. $\mathbb{P}(D)=1$ and $\mathbb{P}(D_0)=0$.

▶ The SLLN provides a bridge to what Cramer (1946) calls the 'empirical counterparts' of the cumulative distribution (cdf) via the Glivenko-Cantelli theorem, which states that under the same IID assumptions, the empirical cdf $\widehat{F}_n(x)$:

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

$\widehat{F}_n(x)$ is directly related to the *empirical* density function $\widehat{f}_n(x)$ via:

$$\widehat{f}_n(x)=\tfrac{1}{2c_n}\left[\widehat{F}_n(x+c_n)-\widehat{F}_n(x-c_n)\right],\ c_n\underset{n\to\infty}{\to}0 \text{ and } nc_n\underset{n\to\infty}{\to}\infty.$$

■ **In light of the fact that these are purely mathematical results, the circularity charge seems completely misplaced.**

4 Error statistics and model-based induction

The stochastic process $\{X_k,\ k\in\mathbb{N}\}$ underlying the SLLN is assumed to be Bernoulli, IID, giving rise to the simple Bernoulli model specified in table 2. The validity of these assumptions vis-a-vis data $\mathbf{x}_0$ secures the reliability and precision of any inference concerning θ, including the SLLN.

Table 2 - Simple Bernoulli Model

Statistical GM: $X_k = \theta + u_k,\quad t\in\mathbb{N}.$

$$\left.\begin{array}{ll} [1] \text{ Bernoulli:} & X_k \backsim \mathsf{Ber}(.,.),\ x_k{=}0,1, \\ [2] \text{ constant mean:} & E(X_k)=\theta, \\ [3] \text{ constant variance:} & Var(X_k)=\theta(1{-}\theta), \end{array}\right\} t\in\mathbb{N}.$$

[4] Independence: $\{X_k,\ k\in\mathbb{N}\}$ is an independent process

4.1 IID renders certain realizations 'almost' impossible

There is nothing stochastic about a particular data $\mathbf{x}_0{:=}\{x_k\}_{k=1}^n$ because it denotes a set of numbers which exhibit certain chance regularity patterns *reflecting* the probabilistic structure of the underlying process $\{X_k,\ k\in\mathbb{N}\}$.

▶ From this perspective the 'randomness' is firmly attached to $\{X_k,\ k\in\mathbb{N}\}$ and is only reflected in data $\mathbf{x}_0$. The only relevant question for $\mathbf{x}_0$ is 'do the chance regularity patterns exhibited by $\{x_k\}_{k=1}^n$ reflect 'faithfully enough' the probabilistic structure presumed for $\{X_k,\ k\in\mathbb{N}\}$? Moreover, what renders D_0 **a set of measure zero** [the set of realizations for which $S_n \overset{a.s.}{\to} p$ does *not* hold], is the fact that the IID structure renders 'almost' impossible realizations such as:

$$\{x_k\}_{k=1}^n = \{0,0,...,0\},$$
$$\{x_k\}_{k=1}^n = \{1,1,...,1\},$$
$$\{x_k\}_{k=1}^n = \{1,0,1,0...,1,0\},$$
$$\{x_k\}_{k=1}^n = \{1,1,0,0,1,1,0,...,0,1,1,0,0\},\ \text{etc.}$$

Such sample realizations would be rejected on statistical adequacy grounds!

▶ It is interesting to note that this empirical view of 'randomness' can be used to shed some additional light on von Mises (1928) randomness condition.

■ To conclude, the justification of the above frequentist interpretation of $\mathbb{P}(A)=p$, does not stem from any *a priori* presuppositions, but is anchored on the empirical adequacy of the statistical model in question; the validation of the model assumptions secures the meaningfulness of $\boxed{\mathbb{P}(\lim_{n\to\infty}(\frac{1}{n}\sum_{k=1}^n X_k)=p)=1.}$

4.2 The frequentist interpretation and 'random samples'

Is a random sample necessary for the frequentist interpretation of probability?

No! The development of stochastic processes since the 1930s has extended statistical modeling from IID to non-IID processes $\{X_k,\ k\in\mathbb{N}\}$ that possess certain *invariant* features which are captured by k-invariant parameter(s) $\boldsymbol{\theta}$.

■ The SLLN, as it relates to the frequentist interpretation of probability, has been extended in two different, but interrelated, directions. The *first* concerns the weakening of the IID assumptions, and the *second* concerns the extension from the arithmetic average $S_n=\frac{1}{n}\sum_{k=1}^{n} X_k$, to any well-behaved (Borel) function of the sample, say $Y_n = h(X_1, X_2, ..., X_n)$; see Billingsley (2003).

In the context of $\mathcal{M}_\theta(\mathbf{x})$, the SLLN can be extended to secure the existence of a *strongly consistent* estimator $\widehat{\boldsymbol{\theta}}_n= \mathbf{H}(\mathbf{X})$ of $\boldsymbol{\theta}$: $\boxed{\mathbb{P}(\lim_{n\to\infty} \widehat{\boldsymbol{\theta}}_n = \boldsymbol{\theta}) = 1.}$

▶ The strong consistency of $\widehat{\boldsymbol{\theta}}_n$ gives rise to 'stable long-run relative frequencies' (Neyman, 1952), whose existence is sufficient for the phenomenon of interest to be amenable to statistical modeling and inference. This bestows a clear frequentist interpretation upon the estimated model: $\boxed{\mathcal{M}_{\widehat{\boldsymbol{\theta}}_n}(\mathbf{x})=\{f(\mathbf{x};\widehat{\boldsymbol{\theta}}_n)\},\ \mathbf{x}\in\mathbb{R}^n_X,}$ which can be used to evaluate (estimate) the probability (relative frequency) of any event in $\sigma(\mathbf{X})\subset\mathfrak{F}$, as well as the reliability of any inference concerning $\boldsymbol{\theta}$.

5 Enumerative vs. model-based induction

Discussions of the frequentist interpretation of probability reveal 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 interpretation itself, but the *second* with the justification of the *straight rule* as a form of induction by enumeration. 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 questioned the argument that all forms of induction can be reduced to enumerative induction; the latter involves only averaging!

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

The primary justification for this rule is the SLLN; asymptotically (as $n \to \infty$) s_n converges to the true probability θ.

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

■ ***What has* not *been appreciated enough in these discussions is the way* model-based induction *enhances the 'signal' by drastically distilling the data into a* statistically adequate model *to render the inference more reliable and precise.***

The relevant statistical model, implicit in the discussion, is the simple Bernoulli with $\mathbb{P}(X{=}1){=}\mathbb{P}(A){=}\theta$, specified in table 2. Viewing $s_n{=}\frac{1}{n}\sum_{k=1}^{n} x_k$ in the context of this statistical model reveals that one knows much more about s_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, but that secures only minimal reliability for $\overline{X}_n$; it is necessary but not sufficient for the reliability of inference for a given n. In contrast, model-based induction offers a way to assess both the reliability and precision by making full use of the model assumptions [1]-[4] (table 2) to derive the sampling distribution:

$$\boxed{\overline{X}_n \backsim \ \mathsf{Bin}(\theta, \tfrac{\theta(1-\theta)}{n}).} \qquad (5.0.4)$$

▶ Both forms of induction rely on the same probabilistic assumptions [1]-[4], but only model-based induction makes effective use of them to derive the sampling distribution of $\overline{X}_n$ in (5.0.4), to be used as a basis of inference for any $n > 1$.

6 The single case and the reference class problems

Salmon's (1967) criteria for the adequacy of an interpretation.

(a) Admissibility. The above frequentist interpretation, anchored on the SLLN, satisfies *admissibility* because relative frequencies can be viewed as an instantiation of the Kolmogorov formalism.

(b) Ascertainability. Viewing the *ascertainability* criterion from the error statistical perspective: $\mathcal{M}_{\widehat{\boldsymbol{\theta}}_n}(\mathbf{x})=\{f(\mathbf{x};\widehat{\boldsymbol{\theta}}_n)\},\ \mathbf{x}\in\mathbb{R}^n_X,$ where $\widehat{\boldsymbol{\theta}}_n$ is a strongly consistent estimator of $\boldsymbol{\theta}$, provides a general way to evaluate reliably the probability of any event of interest [within the model's intended scope] in $\sigma(\mathbf{X})\subset\mathfrak{F}$.

■ In summary, the criticisms about (a)-(b) of the frequentist interpretation – anchored on the SLLN – (Salmon, 1967, pp. 84-87), are rather misplaced.

6.1 The problem of the 'single case' probability

What about **(c) Applicability**? 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 the interpretation of mathematical probability for a particular *objective* which might be very different from modeling observable stochastic phenomena of interest. Salmon goes on to allude to individual decision making (betting) under uncertainty. That might require a different interpretation of probability, but I will leave that question aside.

▶ The *second* issue concerns the claim that the frequentist interpretation cannot be used to assign a probability to single events, like $A_{n+1}=\{X_{n+1}=1\}$-'heads' in the next toss. This is false since one can use the prediction rule:

$$\boxed{\widehat{\mathbb{P}}(A_{n+1}) = \overline{X}_n,}$$

as well as evaluate its reliability using $\overline{X}_n \backsim \mathsf{Bin}(\theta, \frac{\theta(1-\theta)}{n})$.

Hence, the non-applicability charge does not apply to **generic events** like A_{n+1}. It is just a matter of *choosing the appropriate statistical model* $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ (that includes A_{n+1}) – validating it, and then using it to predict $\mathbb{P}(A_{n+1})$.

6.2 Assigning a probability to a 'singular event'

Sometimes the problem of the 'single event' probability is **not** discussed in terms of **generic events**:

> 'A – an Englishman *randomly selected* from the population of 40-year olds today, will die before his next birthday.'

but in terms of **singular events** such as (Gillies, 2000, p. 114):

> 'B – Mr Smith, and 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 B because the 'long-run' metaphor makes no sense in this case.

▶ A moments reflection suggests that the difficulty in assigning a probability to event B has nothing to do with the 'long-run' of the frequentist interpretation. The difficulty stems from the fact that the original statistical model $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ aims to provide an idealized description of the survival of a particular population, [the 40-year olds in England] treating each individual generically. Hence, the σ- field of events of interest ($\mathfrak{F}$) includes event A, but excludes B as a legitimate event; $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ was never meant to be an idealized description of Mr Smith's survival.

▶ That would require a different statistical model, say $\mathcal{M}_{\boldsymbol{\psi}}(\mathbf{z})$, $\mathbf{Z}_k{:=}\ (y_k, \mathbf{X}_k)$, aiming to describe the mechanism for the survival of Mr Smith y_k as it relates the potential contributing factors $\mathbf{X}_k{:=}\ (X_{1k}, X_{2k}, ..., X_{km})$, such as his age, family history, smoking habits, nutritional habits, stress factors, etc., etc.; there are numerous such models in the survival analysis literature; see Balakrishnan and Rao (2004). Hence, what is a legitimate event depends crucially on the choice of the underlying model. This brings us conveniently to the next charge.

6.3 The 'reference class' problem

The *reference class problem* is said to arise since Mr Smith's survival y_k can be related to several different factors $\mathbf{X}_k$ and the frequentist probability of y_k is different relative to each of these distinct potential factors; see Hajek (2007).

▶ A closer look at this argument reveals a serious misunderstanding of model-based inference. The multiplicity of potential contributing factors in $\mathbf{X}_k$ does not render the frequentist interpretation problematic in any sense. To the contrary, there exist numerous statistical models specified in terms of a vector stochastic process $\{\mathbf{Z}_k{:=}(y_k, \mathbf{X}_k),\ k{\in}\mathbb{N}\}$, where the potential factors are combined into a single statistical model $\mathcal{M}_{\boldsymbol{\psi}}(\mathbf{z})$ aiming to describe how these factors collectively

might influence Mr Smith's survival y_k; see Martinussen and Scheike (2006).

■ To conclude, the single case probability and the reference class problem have **nothing to do** with the frequentist interpretation of probability as such.

They essentially constitute admissions that one cannot specify a statistical model for the phenomenon of interest, because of:

(a) practical reasons (the right data are not available), or

(b) conceptual reasons (one cannot think of an appropriate statistical model).

▶ The critics miss the point that frequentist inference depends crucially on being able to specify an **appropriate statistical model**:

(a) $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ accounts for the chance (recurring) regularities in data $\mathbf{x}_0$, and

(b) $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ represents an idealized description of the stochastic phenomenon of interest that gives rise to the events of interest and related events by varying $\mathbf{x} \in \mathbb{R}^n_X$.

Indeed, (b) reflects faithfully the Kolmogorov formalism where the probabilities are defined over $\mathfrak{F}$ – the set of events of interest and related events. That is, probabilities are defined within a statistical model $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$, and they are reliably ascertainable via $f(\mathbf{x};\widehat{\boldsymbol{\theta}}_n)\}$, $\mathbf{x} \in \mathbb{R}^n_X$, when $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ is statistically adequate.

Caution: be wary of approaches that forge probabilities out of thin air!

7 Probability in Frequentist Inference

7.1 The 'coherence' of frequentist probabilities

Armed with a statistical model $\mathcal{M}_\theta(\mathbf{x}) = \{f(\mathbf{x};\boldsymbol{\theta}),\ \boldsymbol{\theta} \in \Theta\},\ \mathbf{x} \in \mathbb{R}^n_X$, frequentist statistical induction takes a number of different forms, including estimation (point and interval), hypothesis testing, prediction and simulation.
The inference is based on a certain statistic $T_n{=}g(X_1, ..., X_n)$ and its sampling distribution can be derived from $f(\mathbf{x};\boldsymbol{\theta})$ via:

$$F(T_n{\leq}t;\boldsymbol{\theta}) = \underbrace{\int\int\cdots\int}_{\{\mathbf{x}:\ g(x_1,x_2,...,x_n){\leq}t;\ \mathbf{x}\in\mathbb{R}^n_X\}} f(\mathbf{x};\boldsymbol{\theta})dx_1dx_2\cdots dx_n. \tag{7.1.5}$$

This sampling distribution furnishes the relevant error probabilities in terms of which the 'effectiveness' (optimality) of any frequentist inference is evaluated.
In this sense, the statistical model $\mathcal{M}_\theta(\mathbf{x})$ furnishes, not only the probabilities of all events of interest, but also all the relevant error probabilities associated with inductive inferences; they all revolve around $f(\mathbf{x};\boldsymbol{\theta}),\ \mathbf{x} \in \mathbb{R}^n_X$. These error probabilities stem from the sampling distribution of $T_n(\mathbf{X})$, say $f(t_n(\mathbf{x});\boldsymbol{\theta})$, $\mathbf{x} \in \mathbb{R}^n_X$, evaluated under different scenarios [factual or hypothetical] relating to $\boldsymbol{\theta}$.

Example. Consider testing the hypotheses: H_0:μ=μ_0 vs. H_1:$\mu > \mu_0$, in the context of the simple Normal model (table 1). The Neyman-Pearson (N-P) optimal testing theory shows that the test T_α:= $\{\tau(\mathbf{X}),\ C_1(\alpha)\}$, where:

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

is *Uniformly Most Powerful* (UMP); see Cox and Hinkley (1974).
That is, $\{\tau(\mathbf{X}),\ C_1(\alpha)\}$ provides the most effective (powerful) α-significance level test for detecting any discrepancy (γ) of interest: μ_1=$\mu_0 + \gamma$, $\gamma \geq 0$.
As mentioned in the introduction, however, the N-P accept/reject rules and the Fisherian p-value do not provide an adequate answer to the question:

'when do data $\mathbf{x}_0$ provide evidence for or against a hypothesis or a claim?'

because they are both vulnerable to the fallacies of acceptance and rejection as well as the initial vs. final precision problem (see Savage, 1962, Hacking, 1965).
▶ Does the the post-data evaluation of inference, based on severe testing, provide an adequate evidential interpretation of frequentist testing?
Certain critics say no! Indeed, they argue for framing any evidential interpretation in terms of epistemic probability and they propose examples like the base-rate fallacy to make their case against severe testing.

7.2 Epistemic probability and frequentist inference

Achinstein's (2001) view of evidence requires one to evaluate the objective epistemic probability of a hypothesis h, given data $\mathbf{x}_0$ and test T, which reflects the 'reasonableness' of believing h. How could one evaluate such an epistemic probability? According to Achinstein (2009) this will require "that a posterior probability, however vaguely characterized, can be attributed to the hypothesis".

▶ When viewed from the error-statistical perspective, any probability attached to h belie the very foundation of frequentist inference.

In frequentist testing hypotheses are framed in terms of the unknown (but constant) parameter(s) $\boldsymbol{\theta}$ of: $\mathcal{M}_\theta(\mathbf{z}) = \{f(\mathbf{z};\boldsymbol{\theta}),\ \boldsymbol{\theta} \in \Theta\},\ \mathbf{z} \in \mathbb{R}^n_Z$.

The archetypal N-P hypotheses formulation is based on partitioning Θ into (Θ_0, Θ_1) to specify :

$$H_0 : \boldsymbol{\theta} \in \Theta_0 \ \text{ vs. } \ H_1 : \boldsymbol{\theta} \in \Theta_1. \tag{7.2.6}$$

What is often insufficiently appreciated is that these hypotheses always concern the actual data-generating mechanism because (7.2.6) is equivalent to:

$$H_0\text{: } f_*(\mathbf{z}) \in \{f(\mathbf{z};\boldsymbol{\theta}),\ \boldsymbol{\theta} \in \Theta_0\} \text{ vs. } H_1\text{: } f_*(\mathbf{z}) \in \{f(\mathbf{z};\boldsymbol{\theta}),\ \boldsymbol{\theta} \in \Theta_1\},$$

where $f_*(\mathbf{z})$ denotes the 'true' distribution of the sample.

That is, the question posed is whether one can narrow down the original model to one of its subsets. It is *not* a question that concerns a certain event or a family of events associated with $\mathbf{X}$.

▶ Hence, attaching probabilities to $\boldsymbol{\theta}$ undermines one of the crucial pillars of frequentist inference: there is a 'true' but unknown parameter, say $\boldsymbol{\theta}^*$, that defines the actual data-generating mechanism, $\mathcal{M}_{\boldsymbol{\theta}^*}(\mathbf{z})=\{f(\mathbf{z};\boldsymbol{\theta}^*)\}$, $\mathbf{z}\in\mathbb{R}^n_Z$, and the primary aim is to learn from $\mathbf{x}_0$ as much as possible about $\mathcal{M}_{\boldsymbol{\theta}^*}(\mathbf{z})$.

■ These crucial features of frequentist testing are often insufficiently appreciated by philosophers of science who argue that the above severity-based evidential interpretation is vulnerable to the so-called *base-rate fallacy* when the prior probability of h is very small; see Howson (2000) and Sober (2008).

Example – base-rate fallacy. To see how misleading this argument is, consider a typical example used by Achinstein (2009). When properly stated, this example assumes a statistical model $\mathcal{M}_\theta(\mathbf{z})$ describing the incidence of a disease in a particular population as it relates to the result of a medical test for that disease. The underlying joint distribution is defined in terms of two Bernoulli random variables: X denotes having the disease (X=1) or not (X=1), and Y denotes the test outcome, positive (Y=1) or negative (Y=0).

$Y \backslash X$	0	1	$f(y)$
0	$p_{11} = .9999799$	$p_{12} = .00000002$	.99998
1	p_{21}=.000019999998	p_{22}=.00000008	.00002
$f(x)$	.9999999	.0000001	1

(7.2.7)

Achinstein selects a particular member of the target population, Irving, and considers the hypothesis of interest to be:

$$h : \text{Irving has the disease.} \tag{7.2.8}$$

He goes on to claim that on the basis of the two conditional probabilities:

$$\begin{array}{l} \mathsf{Prob}(\text{Irving testing positive} \mid \text{he has the disease}) = \frac{.00000008}{.0000001} = 0.8, \\ \mathsf{Prob}(\text{Irving testing positive} \mid \text{he is disease-free}) = \frac{.000019999998}{.9999999} = .00002, \end{array} \tag{7.2.9}$$

the hypothesis h has 'passed' a **severe test** with Irving testing positive.

◀ However, as the argument goes, since Prob(h)= .0000001 is very low, the epistemic probability: $\mathsf{Prob}(h \mid \text{Irving testing positive}) = \frac{.00000008}{.00002} = 0.004$, is also low, and thus, on such grounds Irving's positive result gives very little reason to believe h, despite h's passing a severe test. This is supposed to show how misleading the severity evaluation can be in certain situations.

■ A closer look at Achistein's argument reveals that it is only superficially and misleadingly related to frequentist testing (or severity), because it involves *none* of the components that define a proper frequentist test:

(a) appropriate data, (b) legitimate hypotheses, (c) a test statistic,
(d) sampling distributions and (e) the relevant error probabilities.

To avoid a long digression let me summarize what a proper frequentist test would look like, and compare that with Achinstein's example.

A proper frequentist test. Consider the N-P hypotheses:

$$H_0\textbf{:}\ \phi \leq \phi_0 \text{ vs. } H_1\text{: } \phi > \phi_0, \text{ for } \phi_0 = .00001, \tag{7.2.10}$$

where $\phi{=}\mathbb{P}(X{=}1){=}p_{12}{+}p_{22}$ (assumed to be *unknown*), in the context of a simple (bivariate) Bernoulli model $\mathcal{M}_\theta(\mathbf{z})$ with a density function:

$$f(x,y;\boldsymbol{\theta}){=}p_{11}^{(1-x)(1-y)}p_{22}^{xy}\ p_{21}^{(1-x)y}p_{12}^{(1-y)x},\ \boldsymbol{\theta}\textbf{:}{=}(p_{11},p_{21},p_{12},p_{22})\quad x{=}0,1,\ y{=}0,1.$$

Assuming an IID sample of the form: $\mathbf{Z}\textbf{:}{=}([X_1,Y_1],[X_2,Y_2],...,[X_n,Y_n])$,

yields: $f(\mathbf{z};\boldsymbol{\theta}) = p_{11}^{\sum_{k=1}^n(1-x_k)(1-y_k)}p_{22}^{\sum_{k=1}^n x_k y_k}p_{21}^{\sum_{k=1}^n(1-x_k)y_k}p_{12}^{\sum_{k=1}^n(1-y_k)x_k}$,

and the relevant data take the form: $\mathbf{z}_0\textbf{:}{=}([x_1,y_1],[x_2,y_2],...,[x_n,y_n])$.

The test will involve constructing a test statistic, say $d(\mathbf{Z})=(\widehat{\phi}-\phi_0)/\sqrt{Var(\widehat{\phi})}$, for some appropriate estimator $\widehat{\phi}$ of ϕ, whose sampling distribution is known under both the null and alternative hypotheses. For a given n, these distributions, in conjunction with $C_1(\alpha):=\{\mathbf{z}: d(\mathbf{z})>c_\alpha\}$ of significance level α, will then be used to evaluate the relevant error probabilities, i.e. type I and II (or power) and select an optimal test. For a given $d(\mathbf{z}_0)$, the same sampling distributions will be used to evaluate the relevant *post-data* error probabilities, including the severity of claims concerning discrepancies of interest.

In contrast, Achinstein's base-rate fallacy example is based on:

(i) *Insufficient (inappropriate) data* in the form of *one* illegitimate observation relating to a particular individual (Irving), say $[x_{13}, y_{13}]=[1, 1]$, and not n observations of the form $\mathbf{z}_0$ relating to the sample $\mathbf{Z}$ from the prespecified population.

► Even if one were to ignore the illegitimacy and assume that this single observation was randomly selected from the prespecified population, no *consistent* (minimally reliable) estimator or test concerning $\boldsymbol{\theta}$ is possible with n=1.

(ii) An *improper 'hypothesis'* of interest (h) because: (a) h concerns a *singular event* $[x_{13}=1]$ beyond the intended scope of the assumed statistical model $\mathcal{M}_\theta(\mathbf{z})$, (b) h is erroneously misidentified with the proper generic event $\{X=1\}$-an indi-

vidual *randomly selected* from the prespecified population has the disease,
(c) h, being an event, it is *not* a proper frequentist hypothesis; the latter could not even be defined because all the parameters in (7.2.7) are known!
(iii) A medical (*not* a frequentist) test whose outcome, a singular event $[y_{13}{=}1]$, is misidentified with the proper generic event $\{Y{=}1\}$-a test result *randomly selected* from the prespecified population is positive. This medical test is then conveniently conflated with a *proper frequentist test* using a beguiling analogical argument concerning the former's false negative and false positive probabilities as equivalent to the type I (α) and II (β) error probabilities (Howson, 2000); for a given α, the power depends crucially on the sample size n, e.g. for $n{=}10000$, power equal to .8 is not something you should write home about!
(iv) *An improper final 'inference'* pertaining to the occurrence of a singular event $[x_{13}, y_{13}] = [1, 1]$ (outside the intended scope of $\mathcal{M}_\theta(\mathbf{z})$) and not about the actual generating mechanism that gave rise to $\mathbf{z}_0$.
(v) In place of the relevant error probabilities, the above example uses *illegitimate conditional probabilities* assigned to singular events.
▶ Even if one were to pretend they refer to the generic events, they still denote probabilities of *events,* which are a far cry from proper error probabilities;

the latter are *never* conditional (they are evaluated under different hypothetical scenarios) and they invariably denote tail areas.

■ One can elaborate further on how misleading Achinstein's claim is that h passes a severe test with data $[x_{13}, y_{13}]$=[1, 1] on the basis of a medical test (see Mayo, 2009), but it suffices here to say that such examples demonstrate the lack of basic understanding of frequentist inference in general.

Incompatibility of frequentist and epistemic probabilities

Attempts to apply epistemic probabilities from a Bayesian or non-Bayesian perspectives to frequentist inference would inevitably entail crucial distortions and serious misinterpretations because they, necessarily, involve assigning probabilities to $\boldsymbol{\theta}$. The very notion of a probability distribution for $\boldsymbol{\theta}$ contravenes the fundamental frequentist premise that there exists a true $\boldsymbol{\theta}^*$ which characterizes the actual data-generating mechanism**.** Assigning probabilities to $\boldsymbol{\theta}$ brings specification uncertainty into the inference stage and ruins the very notion of error probability because, now, there is *no* single $f(\mathbf{z};\boldsymbol{\theta})$, $\mathbf{z}\in\mathbb{R}^n_Z$ [actual or hypothetical] in terms of which the relevant sampling distributions can be derived;

$$F(T_n{\leq}t;\boldsymbol{\theta}) = \underbrace{\int\int\cdots\int}_{\{\mathbf{z}:\ g(\mathbf{z}){\leq}t;\ \mathbf{z}\in\mathbb{R}^n_Z\}} f(\mathbf{z};\boldsymbol{\theta})dz_1dz_2\cdots dz_n.$$

8 Summary and conclusions

The error statistical perspective was used to shed light on a number of charges and criticisms leveled against the frequentist interpretation of probability.

▶ The **circularity charge** stems from misidentifying the frequentist interpretation of probability with von Mises's rendering, instead of invoking the SLLN to secure 'stable long-run relative frequencies'. The SLLN is a theorem whose assertions revolve around the Kolmogorov mathematical formalism in measure-theoretic language. Hence, the justification for the frequentist interpretation stems from the statistical adequacy of the prespecified model $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ that renders the SLLN valid, and there is no circularity.

▶ The charge that the frequentist interpretation necessitates **random samples** was also called into question. The original SLLN can been extended to hold for non-IID samples, giving rise to strongly consistent estimators, $\widehat{\boldsymbol{\theta}}_n \overset{a.s.}{\to} \boldsymbol{\theta}$, for general statistical models $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ indexed by certain t-invariant parameters $\boldsymbol{\theta}$.

▶ Model-based induction has no difficulty assigning probabilities to **generic events** within the relevant field of events of interest as demarcated by $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$, $\mathbf{x} \in \mathbb{R}^n_X$. But it cannot assign probabilities to events outside its intended scope.

The charges of 'single case' probabilities and the 'reference class' problem constitute admisions of inability to construct an appropriate statistical model $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ for practical or conceptual reasons; neither deficiency can be blamed on the frequentist interpretation of probability as such.

▶ Is there a role for **epistemic probability**, interpreted as reflecting the degree of "reasonableness of belief", in supplementing the severity-based evidential interpretation of inference? The short answer proposed is **No**, because there is a *serious incompatibility* between the epistemic and frequentist interpretations of probability, stemming from assigning probabilities to $\boldsymbol{\theta}$.

▶ The base-rate fallacy illustrates the lack of basic understanding of frequentist testing on behalf of the critics.

CRUCIAL POINTS:

(i) frequentist testing always pertains to the actual data-generating mechanism and never to generic events within the intended scope of $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$,

(ii) genuine error probabilities are never conditional and they always depend crucially on the sample size n; inference with n=1 is always a bad idea!

(iii) attaching probabilities to $\boldsymbol{\theta}$ belies the very foundation of frequentist inference; it mires the notion of relevant error probabilities.

9 Appendix: A severity-based evidential interpretation

Example. Consider the above t-test with μ_0=0, significance level α=.05, c_α=1.796, n=12, and data $\mathbf{x}_0$ yielded $\overline{x}_n$=2.583, s=2.778, giving rise to $\tau(\mathbf{x}_0)$=$\frac{\sqrt{12}(2.583-0)}{2.778}$=3.221. Since $\tau(\mathbf{x}_0)$ is greater than c_α, the null hypothesis H_0: μ=0 is rejected. This is confirmed by the p-value: $\mathbb{P}(\tau(\mathbf{X}) > 3.221;\ H_0)$=.004.
Does this mean that the data provide evidence for a substantive discrepancy from the null? Not necessarily! To establish that, one needs to use a post-data evaluation of the decision to reject H_0 based on the severity evaluation for different discrepancies $\gamma \geq 0$, which for μ_1=$\mu_0 + \gamma$ takes the form (Mayo and Spanos, 2006):

$$SEV(T_\alpha; \mu > \mu_1) = \mathbb{P}(\mathbf{x} : \tau(\mathbf{X}) \leq \tau(\mathbf{x}_0); \mu \leq \mu_1).$$

Assuming a severity threshold of, say .90, one can evaluate the 'largest' discrepancy warranted by $\mathbf{x}_0$, which in this case is $0 < \gamma \leq 1.5$, since
$SEV(T_\alpha; \mu > 1.5) = .899$. That is, data $\mathbf{x}_0$ in conjunction with test T_α provide evidence (with severity .9) for the presence of a discrepancy $\gamma \leq 1.5$.
This threshold provides a way to gradate the *trustworthiness* of the inductive inference as it pertains to the claim $\gamma \leq 1.5$.

▶ In the above example, the data measure changes in yield due to the application of a fertilizer, and the discrepancy γ quantifies the increase in yield per acre. Hence, the warranted discrepancy $\gamma \leq 1.5$ with data $\mathbf{x}_0$ can be used, in conjunction with other substantive information, including economic costs, to decide if such a discrepancy is also substantively significant or not.

■ This **severity-based evidential interpretation** can be used to address:
(i) the fallacies of acceptance and rejection,
(ii) the large n problem,
(iii) statistical vs. substantive significance,
(iv) the prosecutor's fallacy,
(v) several confusions pertaining to conditional vs. error probabilities,
(vi) several confusions pertaining to observed confidence intervals,
(vii) the arbitrariness of the prespecified significance level α,
(viii) the asymmetry between the null and alternative hypotheses,
(ix) a number of confusions perpetrated by Bayesian statisticians and philosophers using several "classic" examples; see Berger and Wolpert (1988).