Synthese (2021) 199:13949–13976
https://doi.org/10.1007/s11229-021-03405-z

ORIGINAL RESEARCH

Bernoulli's golden theorem in retrospect: error probabilities and trustworthy evidence

Aris Spanos[1]

Received: 5 October 2020 / Accepted: 3 September 2021 / Published online: 1 November 2021

Abstract

Bernoulli's 1713 golden theorem is viewed retrospectively in the context of modern model-based frequentist inference that revolves around the concept of a prespecified statistical model $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$, defining the inductive premises of inference. It is argued that several widely-accepted claims relating to the golden theorem and frequentist inference are either misleading or erroneous: (a) Bernoulli solved the problem of inference 'from probability to frequency', and thus (b) the golden theorem cannot justify an approximate Confidence Interval (CI) for the unknown parameter θ, (c) Bernoulli identified the probability $P(A)$ with the relative frequency $\frac{1}{n}\sum_{k=1}^{n} x_k$ of event A as a result of conflating $f(\mathbf{x}_0|\theta)$ with $f(\theta|\mathbf{x}_0)$, where $\mathbf{x}_0$ denotes the observed data, and (d) the same 'swindle' is currently perpetrated by the p value testers. In interrogating the claims (a)–(d), the paper raises several foundational issues that are particularly relevant for statistical induction as it relates to the current discussions on the replication crises and the trustworthiness of empirical evidence, arguing that: [i] The alleged Bernoulli swindle is grounded in the unwarranted claim $\hat{\theta}_n(\mathbf{x}_0) \simeq \theta^*$, for a large enough n, where $\hat{\theta}_n(\mathbf{X})$ is an optimal estimator of the true value θ^* of θ. [ii] Frequentist error probabilities are *not* conditional on hypotheses (H_0 and H_1) framed in terms of an unknown parameter θ since θ is neither a random variable nor an event. [iii] The direct versus inverse inference problem is a contrived and misplaced charge since neither conditional distribution $f(\mathbf{x}_0|\theta)$ and $f(\theta|\mathbf{x}_0)$ exists (formally or logically) in model-based ($\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$) frequentist inference.

Keywords Bernoulli's golden theorem · Direct versus inverse inference · Factual versus hypothetical reasoning · Law of large numbers · p Value · Effect sizes · Bayes' rule · Straight rule · Error probabilities · Bernoulli's swindle

This article belongs to the topical collection "Recent Issues in Philosophy of Statistics: Evidence, Testing, and Applications", edited by Sorin Bangu, Emiliano Ippoliti, and Marianna Antonutti.

✉ Aris Spanos
aris@vt.edu

[1] Department of Economics, Virginia Tech, Blacksburg, VA 24061, USA

1 Introduction

James (Jacob) Bernoulli (1713), in Part IV of his book entitled "The Art of Conjecturing" derived what he called the 'golden theorem' (*theorema aureum*). This theorem was particularly influential for subsequent developments in both probability theory (especially limit theorems) and statistical inference (frequentist vs. Bayesian inference); see Hald (1998), Gorroochurn (2012). Since then, the golden theorem has become a topic of recurring disputes relating to its importance, interpretation and implications for inference, which are motivated by several of its unique features, including (i) Bernoulli's own motivation and interpretation, (ii) its direct link to his numerical example aiming to illustrate it, (iii) its inferential interpretation in terms of the inverse versus direct inference, and (iv) its interpretation, and implications for a finite sample ($n < \infty$) and its asymptotic ($n \to \infty$) renderings.

In an attempt to narrow the scope of the discussion, the paper focuses on Diaconis and Skyrms (2018) that summarizes a widely-held perspective on the golden theorem as follows:

"Bernoulli's motivation for his golden theorem was the determination of chance from empirical data."(p. 64).

"What does it mean to determine chances a posteriori from frequencies? The question is, given the data—the number of trials and the relative frequencies of success in those trials—what is the probability that the chances fall within a certain interval? It is evident that this is *not* the problem that Bernoulli solved. He solved an inference from chances to frequencies, not the inverse problem from frequencies to chances. The inverse problem had to wait for Thomas Bayes." (p. 65).

"Bernoulli argued that he had shown that with a large enough number of trials, it will be *morally certain* that relative frequency would be (approximately) equal to chance. But if frequency equals chance, then chance equals frequency. So, the argument goes, we have solved the problem of inference from frequency to chance. This is *Bernoulli's swindle*. Try to make it precise and it falls apart." (p. 65).

"To be explicit, Bernoulli's conditional probabilities are probabilities about frequencies given chances, rather than probabilities about chances given frequencies." (p. 66).

It is important to note at the outset that Bernoulli (1713) viewed $\theta = \mathbb{P}(X = 1)$ as probability a priori (chances) and $\overline{x}_n = \frac{1}{n}\sum_{k=1}^{n} x_k$, based on binary data $\mathbf{x}_0 := (x_1, x_2, \ldots, x_n)$, as probability a posteriori (relative frequencies), which should not be conflated with the modern Bayesian interpretation of these terms.

Diaconis and Skyrms (2018) also argue that current p value testers routinely perpetrate Bernoulli's swindle by conflating $P(H_0|D)$ with $P(D|H_0)$: "The untutored think they are getting the probability of effectiveness given the data, while they are being given conditional probabilities going in the opposite direction." (p. 67).

The above quotations include several different but interrelated claims:

(a) Bernoulli (1713) solved the problem of inference 'from probability θ to frequency $\overline{x}_n$', but the inverse problem was addressed by Bayes (1764), because:

(b) Bernoulli committed a swindle by identifying the probability (θ) with relative frequency ($\overline{x}_n$) as a result of conflating 'direct' inference based on $f(\mathbf{x}_0|\theta)$ with 'inverse' inference based on $f(\theta|\mathbf{x}_0)$, and thus:
(c) the golden theorem does not justify an approximate confidence interval for θ, and
(d) the same swindle permeates current frequentist testing whose error probabilities fail to distinguish between $P(H_0|\mathbf{x})$ and $P(\mathbf{x}|H_0)$.

Claims and criticisms similar to (a)–(d) are repeated by most Bayesian statistics textbooks (O'Hagan, 1994, and Robert, 2007), as well as philosophy of science books on 'probability and evidence' (Howson & Urbach, 2006; Sober, 2008).

Viewing Bernoulli's (1713) golden theorem retrospectively in the context of modern model-based [$\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$] frequentist inference, the claims in (a)–(d) are called into question as grounded in misconceptions. Their interrogation brings out several broader foundational problems that are particularly relevant for the current discussions on the replication crisis and the trustworthiness of empirical evidence, including:

[i] misapplying/misconstruing limit theorems (as $n \to \infty$) in inference,
[ii] misinterpreting the p value, type I and II error probabilities and the power as conditional on H_0 or H_1,
[iii] the alleged 'swindle' is a special case of a well-known unwarranted claim, $\hat{\theta}(\mathbf{x}_0) \simeq \theta^*$ for $n < \infty$, where θ^* denotes the true value of θ, $\hat{\theta}(\mathbf{x}_0)$ is the estimate corresponding to an optimal estimator $\hat{\theta}(\mathbf{X})$ of θ, which is routinely committed by effect size users, and not by frequentist testers, and
[iv] the direct versus inverse inference criticism is not just misplaced, it is motivated by a misguided attempt to justify a dubious crosscut in vindicating Bayes' formula by reimagining the distribution of the sample $f(\mathbf{x};\theta)$, $\mathbf{x} \in \mathbb{R}^n_X$, as *conditional* on θ, i.e. $f(\mathbf{x}|\theta)$, $\mathbf{x} \in \mathbb{R}^n_X$, which is meaningless in frequentist statistics; see Spanos (2010).

2 Statistical induction

2.1 Induction by enumeration

The *problem of induction* boils down to justifying an inference from particular instances to potential realizations (generalizations), or from past to future instances. Hume (1748) argued that no rational justification of induction based on experience can be invoked since the argument that 'a regularity that has held in the past will or must continue to hold in the future' is circular and question-begging in the sense that it presupposes a belief in the 'uniformity of nature' that has no rational defence in reason. Instead, it reflects custom of the mind or habit. Hume's stance has been bedeviling philosophy of science since then; see Henderson (2020).

Induction by enumeration: if (m/n) is the relative frequency of event A from a sample of n realizations, infer that:

$$P(A) \simeq \frac{\mathrm{m}}{n}, \tag{1}$$

i.e. the 'long-run' relative frequency is (m/n); see Salmon (1967), p. 50.

This is widely viewed in philosophy of science as the quintessential form of statistical induction, and von Mises's (1928) frequentist interpretation of probability as providing the link between the empirical relative frequencies $(m/n) = \frac{1}{n}\sum_{k=1}^{n} x_k$ and the corresponding mathematical probability $P(A)$ using the notion of *a collective*: an infinite sequence of outcomes $\{x_k\}_{k=1}^{\infty}$, $x_k = \begin{cases} 0 \text{ not } A \\ 1 \ A \end{cases}$, via $\lim_{n\to\infty}\left(\frac{1}{n}\sum_{k=1}^{n} x_k\right) = P(A)$, with this limit being *invariant to place selections*, i.e. $\lim_{n\to\infty}\left(\frac{1}{n}\sum_{k=1}^{n} \varphi(x_k)\right) = P(A)$, where $\varphi(.)$ is a mapping of admissible place-selection sub-sequences $\{\varphi(x_k)\}_{k=1}^{\infty}$.

Hacking (1965), p. 261, questions Salmon's claim: "Reichenbach equated induction with acceptance of a certain estimator, the *straight rule*: If m of the n observed A are B, estimate the long-run frequency of B among A as m/n. Salmon and Reichenbach maintain that *if long-run frequencies exist, the straight rule for estimating long-run frequencies is to be preferred to any rival estimator*. Other propositions are needed to complete their vindication of induction, but only this one concerns us. Salmon claims to have proved it. This is more interesting than mere academic vindications of induction; practical statisticians need good criteria for choosing among estimators, and, if Salmon were right, he would have very largely solved their problems, which are much more pressing than Hume's."

The key feature of inductive inference is that it is *ampliative* in the sense that it goes beyond the observed data (m/n) to the unknown $\theta = \mathbb{P}(A)$, enhancing our knowledge about the underlying set-up that gave rise to the observed data. As argued in the sequel, when this claim is viewed in the context of model-based induction where $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ provides the inductive premises of inference, Hacking is right to question Salmon's claim since (1) is a special case of a more general unwarranted claim:

$$\hat{\theta}_n(\mathbf{x}_0) \simeq \theta^*, \quad \text{for a large enough } n < \infty, \tag{2}$$

when $\hat{\theta}_n(\mathbf{X})$ is an 'optimal' estimator of the unknown true parameter θ^*; (1) assumes the simple Bernoulli model in (5). Viewing Hacking's "Other propositions needed to complete their vindication of induction" in the context of $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ in (5), they include (i) the validity of the inductive premises [Independent and Identically Distributed (IID)] for data $\mathbf{x}_0$, which ensures the reliability of inference, as well as (ii) the optimality of the estimator $\hat{\theta}_n(\mathbf{X}) = \frac{1}{n}\sum_{k=1}^{n} X_k$, which secures the effectiveness of the inference. The reliability and effectiveness of inference lie at the core of *inductive* (statistical) *inference*: how we learn from data about phenomena of interest.

2.2 Model-based frequentist inference

Fisher (1922) recast Pearson's descriptive statistics into model-based induction that revolves around the concept of a prespecified parametric *statistical model*, generically defined by:

$$\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x}) = \left\{f(\mathbf{x};\boldsymbol{\theta}),\ \boldsymbol{\theta} \in \Theta \subset \mathbb{R}^m\right\}, \quad \mathbf{x} \in \mathbb{R}_X^n, \quad n > m, \tag{3}$$

where $f(\mathbf{x};\boldsymbol{\theta}), \mathbf{x} \in \mathbb{R}_X^n$ denotes the joint distribution of the sample $\mathbf{X} := (X_1, \ldots, X_n)$, $\mathbb{R}_X^n$ denotes the sample space and Θ the parameter space, specifying (explicitly) the inductive premises of inference. The revolutionary nature of Fisher's recasting stems from the fact that $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ aims to describe the stochastic mechanism that gave rise to data $\mathbf{x}_0$, and not to summarize/describe $\mathbf{x}_0$, and thus transforming descriptive statistics into statistical induction.

Example 1 Consider the simple Normal model:

$$X_t \sim \text{NIID}\left(\mu, \sigma^2\right), \ x_t \in \mathbb{R}, \ E(X_t) = \mu \in \mathbb{R}, \ Var(X_t) = \sigma^2 > 0, \quad t \in \mathbb{N} := (1, 2, \ldots, n, \ldots), \tag{4}$$

where 'NIID' stands for Normal, Independent, and Identically Distributed (IID), and for simplicity we assume that σ^2 is *known*.

Example 2 Consider the *simple Bernoulli model*, specified by:

$$X_k \backsim \ \text{BerIID}(\theta, \theta(1-\theta)), \ x_k = 0, 1, \ E(X_k) = \theta, \quad 0 < \theta < 1, \ k \in \mathbb{N}, \tag{5}$$

where 'Ber' denotes the 'Bernoulli distribution' with $\theta = \mathbb{P}(X_k = 1)$.

The *primary objective* of frequentist inference is to use the statistical information, as summarized by $f(\mathbf{x};\boldsymbol{\theta}), \mathbf{x} \in \mathbb{R}_X^n$, in conjunction with data $\mathbf{x}_0$, to *narrow down* Θ as much as possible, ideally, to a single point $\boldsymbol{\theta}^*$—the 'true' value of $\boldsymbol{\theta}$ in Θ—which is shorthand for saying that the generating mechanism $\mathcal{M}^*(\mathbf{x}) = \left\{f\left(\mathbf{x};\boldsymbol{\theta}^*\right)\right\}, \mathbf{x} \in \mathbb{R}_X^n$, could have generated data $\mathbf{x}_0$; see Spanos and Mayo (2015).

The evaluation of the effectiveness (optimality) of an inference procedure is calibrated in terms of the relevant error probabilities that revolve around the sampling distribution, $f(y_n;\boldsymbol{\theta}), \ \forall y_n \in \mathbb{R}$, of a statistic (estimator, test, predictor) $Y_n = h(X_1, X_2, \ldots, X_n)$ derived via:

$$F_n(y) = \mathbb{P}(Y_n \leq y) = \underbrace{\int \int \cdots \int}_{\{\mathbf{x}:\, h(\mathbf{x}) \leq y\}} f(\mathbf{x};\, \boldsymbol{\theta}) d\mathbf{x}, \quad \forall y \in \mathbb{R} \tag{6}$$

The parameter $\boldsymbol{\theta}$ is viewed as an *unknown constant* whose values in (6) in deriving the sampling distribution, $f(y_n;\boldsymbol{\theta}), \ \forall y_n \in \mathbb{R}$, are always *prespecified* and based on two different forms of reasoning:

(i) **factual** (estimation and prediction): presuming that $\boldsymbol{\theta} = \boldsymbol{\theta}^*$, whatever that value happens to be in $\boldsymbol{\Theta}$, and

(ii) **hypothetical** (hypothesis testing): various hypothetical scenarios based on $\boldsymbol{\theta}$ taking different prespecified values under H_0: $\boldsymbol{\theta} \in \boldsymbol{\Theta}_0$ (presuming that $\boldsymbol{\theta} \in \boldsymbol{\Theta}_0$) versus H_1: $\boldsymbol{\theta} \in \boldsymbol{\Theta}_1$ (presuming that $\boldsymbol{\theta} \in \boldsymbol{\Theta}_1$), where $\boldsymbol{\Theta}_0 \cup \boldsymbol{\Theta}_1 = \boldsymbol{\Theta}$, $\boldsymbol{\Theta}_0 \cap \boldsymbol{\Theta}_1 = \emptyset$; see Spanos (2019), p. 576. Note that neither form of reasoning involves conditioning on $\boldsymbol{\theta}$, since the latter makes no mathematical or logical sense; see Sect. 2.5 for further discussion.

It is important to emphasize that the reliability and effectiveness of statistical inference depend crucially on *statistical adequacy*: the validity of the probabilistic assumptions comprising the prespecified $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ For example 1, the invoked assumptions are NIID and their validity should be evaluated using mis-specification (M-S) testing before any inference is drawn; see Spanos (2018). When any of these assumptions are invalid for data $\mathbf{x}_0$, the actual error probabilities associated with the invoked inference procedures are likely to be very different from the nominal (assumed based on $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$) ones. Applying a .05 significance level test when the actual type I error (due to statistical misspecification) is closer to .9, will lead that inference astray; see Spanos and McGuirk (2001).

Example 1 (continued). For the simple Normal model in (4):

$$\begin{aligned} &\text{(i)}\ \ \overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \sim \mathrm{N}\left(\mu, \frac{\sigma^2}{n}\right), \\ &\text{(ii)}\ \ s^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}_n\right)^2 \sim \left(\frac{n-1}{\sigma^2}\right)\chi^2(n-1), \end{aligned} \tag{7}$$

and (iii) $\overline{X}_n$ is independent of s^2, implies that (Lehmann & Romano, 2005, p. 156):

$$\tau(\mathbf{X};\mu) = \frac{\sqrt{n}\left(\overline{X}_n - \mu\right)}{s} \sim \ \mathrm{St}(n-1), \tag{8}$$

where $\mathrm{St}(n-1)$ denotes the Student's t distribution with $(n-1)$ degrees of freedom. What is not obvious is how to interpret (8), since it is not apparent why $E(\tau(\mathbf{X};\mu)) = 0$. A simple answer is that it follows from the fact that $\overline{X}_n$ is an unbiased estimator of μ, i.e. $E\left(\overline{X}_n\right) = \mu^*$. Using this unbiasedness in conjunction with the independence in (iii), one can show (Williams, 2001, p. 101) that under *factual* reasoning:

$$E\left(\frac{\sqrt{n}\left(\overline{X}_n - \mu\right)}{s}\right) \overset{\mu=\mu^*}{=} E\left(\overline{X}_n - \mu^*\right)\cdot E\left(\sqrt{n}/s\right) = 0, \ \text{for any } E\left(\sqrt{n}/s\right) > 0.$$

Hence, a more transparent way to specify (8) is:

$$\tau(\mathbf{X};\mu^*) = \frac{\sqrt{n}\left(\overline{X}_n - \mu^*\right)}{s} \overset{\mu=\mu^*}{\sim} \mathrm{St}(n-1), \tag{9}$$

despite the cumbersome notation that overuses '*' to elucidate it.

It is interesting to note that when the von Mises 'collective' $\{x_k\}_{k=1}^{\infty}$ is viewed from the model-based ($\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})$) perspective, it becomes clear that an infinite realization of an IID Bernoulli process $\{X_t, t \in \mathbb{N}\}$ is a non-operational concept. What operationalizes the idea behind the collective is to view the data $\mathbf{x}_0 = \{x_k\}_{k=1}^{n}$ its initial segment that constitutes a realization of the sample $\mathbf{X}$; see Spanos (2013a).

2.3 Estimation (point and interval)

For estimation and prediction purposes the underlying reasoning is factual.

Example 1 (continued). For the simple Normal model in (4) with σ^2 known, the Maximum Likelihood (ML) estimator of μ is $\hat{\theta}_{ML}(\mathbf{X}) = \frac{1}{n}\sum_{i=1}^{n} X_i$. Its optimality revolves around its sampling distribution evaluated using *factual reasoning*:

$$\hat{\theta}_{ML}(\mathbf{X}) \overset{\mu=\mu^*}{\sim} \mathrm{N}\left(\mu^*, \frac{\sigma^2}{n}\right). \tag{10}$$

where $\hat{\theta}_{ML}(\mathbf{X})$ is unbiased, sufficient, fully efficient, and strongly consistent; note that these properties hold only when the model assumptions 'NIID' are valid!

As Fisher (1922) points out, the statistics literature until the 1920s conflated the sample $\mathbf{X} := (X_1, X_2, \ldots, X_n)$ with the sample realization $\mathbf{x}_0$ (the observed data), as well as the estimator $\hat{\theta}(\mathbf{X})$, the estimate $\hat{\theta}(\mathbf{x}_0)$ and the unknown parameter θ.

What is often insufficiently appreciated by the effect size literature (Cohen, 1988) is that an optimal (consistent, unbiased, fully efficient, sufficient) estimator $\hat{\theta}_n(\mathbf{X})$ of θ does *not* justify the inferential claim in (2).

Example 1 (continued). The ML estimator $\hat{\theta}_{ML}(\mathbf{X}) = \frac{1}{n}\sum_{i=1}^{n} X_i$ of μ enjoys all optimal properties, but that does not underwrite the claim $\hat{\theta}_{ML}(\mathbf{x}_0) \simeq \mu^*$, since $\hat{\theta}_{ML}(\mathbf{x}_0)$ represents a *single value* from the range of possible values of $\hat{\theta}_{ML}(\mathbf{x})$ associated with its sampling distribution $f\left(\hat{\theta}_{ML}(\mathbf{x}); \theta^*\right)$, $\mathbf{x} \in \mathbb{R}^n$, as in (10). What (10) implies is that $Var\left(\hat{\theta}_{ML}(\mathbf{X})\right) = \frac{\sigma^2}{n}$ decrease to zero as $n \to \infty$. Therefore, invoking the strong consistency of $\hat{\theta}_{ML}(\mathbf{X})$ does not address the problem since $\mathbb{P}\left(\lim_{n\to\infty}\hat{\theta}_{ML}(\mathbf{X}) = \theta^*\right) = 1$ pertains to what happens at the limit ($n = \infty$), and not at any $n < \infty$; see Spanos (2013a). That is, as n increases $f\left(\hat{\theta}_{ML}(\mathbf{x}); \theta^*\right)$ concentrates around θ^*, but it is defined over an unknown interval for any $n < \infty$. As shown in Sect. 5.1, this interval can be approximated using bounds provided by the Law of Iterated Logarithm; see Billingsley (1995).

The unwarranted inferential claim in (2) was a primary motivation for Neyman (1937) to go beyond point estimation to propose the method of Confidence Intervals (CIs) that takes into consideration the uncertainty that relates to the point estimate as described by its sampling distribution $f\left(\hat{\theta}_{ML}(\mathbf{x}); \theta^*\right)$, $\mathbf{x} \in \mathbb{R}^n_X$.

Example 1 (continued). For (4), the $(1-\alpha)$ CI takes the form:

$$\mathbb{P}\left(\overline{X}_n - c_{\frac{\alpha}{2}}\left(\frac{s}{\sqrt{n}}\right) \leq \mu < \overline{X}_n + c_{\frac{\alpha}{2}}\left(\frac{s}{\sqrt{n}}\right);\ \mu = \mu^*\right) = 1-\alpha, \tag{11}$$

where $c_{\frac{\alpha}{2}}$ is derived from the distribution of $\tau(\mathbf{X}; \mu^*)$ in (8). Having said that, it should be emphasized that the observed CI, $\left(\overline{x}_n - c_{\frac{\alpha}{2}}\left(\frac{s}{\sqrt{n}}\right) \leq \mu < \overline{x}_n + c_{\frac{\alpha}{2}}\left(\frac{s}{\sqrt{n}}\right)\right)$, where

$\overline{x}_n$ is the estimate of μ, cannot be assigned the probability $(1-\alpha)$ post-data; it either includes or excludes μ^*, but it is invariably unknown which one holds. The length of the observed CI does, however, provide some additional information about the uncertainty relating to the estimate $\overline{x}_n$.

2.4 Neyman–Pearson (N–P) testing

Example 1 (continued). Consider testing the hypotheses:

$$H_0: \mu \leq \mu_0 \text{ vs. } H_1: \mu > \mu_0, \tag{12}$$

where the framing of H_0 and H_1 constitutes a partition of $\mathbb{R}$. For statistical inference purposes, all values of μ are of interest, irrespective of whether only a few values are of substantive interest. Using *hypothetical reasoning* one can evaluate the sampling distribution of $\tau(\mathbf{X}) = \frac{\sqrt{n}(\overline{X}_n-\mu_0)}{s}$ under H_0 and H_1 yielding:

$$\text{(i)}\tau(\mathbf{X}) = \frac{\sqrt{n}\big(\overline{X}_n-\mu_0\big)}{\sigma} \overset{\mu=\mu_0}{\sim} \text{St}(n-1), \quad \text{(ii)}\tau(\mathbf{X}) = \frac{\sqrt{n}\big(\overline{X}_n-\mu_0\big)}{\sigma} \overset{\mu=\mu_1}{\sim} \text{St}(\delta_1; n-1), \tag{13}$$

where $\delta_1 = \frac{\sqrt{n}(\mu_1-\mu_0)}{\sigma}$, for $\mu_1 > \mu_0$, is the noncentrality parameter.

More generally, N–P testing is based on hypothetical reasoning using prespecified values of μ that could 'approximate closely' μ^*, in the sense that the difference $||\mu^*-\mu_0||$, where $||.||$ denotes a distance function (norm), is statistically insignificant/significant (negligible/substantial). The primary role of the error probabilities is to operationalize the concepts of 'statistically significant/insignificant' as it relates to $||\mu^*-\mu_0||$. The test statistic $\tau(\mathbf{X})$ reflects the difference $||\mu^*-\mu_0||$, in the sense that (i) μ^* is replaced by its best estimator, and (ii) $\tau(\mathbf{X})$ increases monotonically with this distance. For instance, the test T_α in (14) uses $\tau(\mathbf{X}) = \big[\sqrt{n}\big(\overline{X}_n-\mu_0\big)/s\big]$, a standardized distance between $\overline{X}_n$ (best estimator of μ^*) and μ_0.

For the hypotheses in (12), an α-significance level Uniformly Most Powerful (UMP) test is defined by:

$$T_\alpha := \{\tau(\mathbf{X}) = \frac{\sqrt{n}\big(\overline{X}_n-\mu_0\big)}{\sigma},\ C_1(\alpha) = \{\mathbf{x}: \tau(\mathbf{x}) > c_\alpha\}\}, \tag{14}$$

Lehmann and Romano (2005, p. 58). The type I error probability and the p value are evaluated using (i) in (13):

$$\mathbb{P}(\tau(\mathbf{X}) > c_\alpha; \mu=\mu_0) = \alpha,\ \mathbb{P}(\tau(\mathbf{X}) > \tau(\mathbf{x}_0); \mu=\mu_0) = p(\mathbf{x}_0). \tag{15}$$

The power of T_α is evaluated using (ii) in (13):

$$\mathcal{P}(\mu_1) = \mathbb{P}(\tau(\mathbf{X}) > c_\alpha; \mu=\mu_1), \text{ for all } \mu_1 > \mu_0. \tag{16}$$

The power of a test measures its *generic* (for any $\mathbf{x} \in \mathbb{R}^n$) *capacity* to detect discrepancies from H_0. As argued next, none of the above error probabilities (type I, II, power, p value) are conditional on values of μ. Hence the use of the notation ';' instead of 'l' to separate the observable random variable $\tau(\mathbf{X})$ from the unknown (and unobservable) constant θ to avoid confusion.

Particularly important for the current discussions on replicability are two crucial preconditions proposed by Neyman and Pearson (1933) which relate to the framing of H_0 and H_1 to secure the effectiveness of N–P testing: [i] H_0 and H_1 should constitute a partition of Θ, in a way that renders [ii] the type I error probability as the more serious of the two to ensure that the framing of H_1 includes the potential range of values around θ^*. Precondition [i] is needed to eliminate the scenario where θ^* lies outside $\mathbf{\Theta}_0 \cup \mathbf{\Theta}_1$, and [ii] to ensure that the test has power where is needed for effective learning from data.

Example 2 (continued). For the simple Bernoulli model let the framing be:

$$H_0 : \theta = \theta_0 \text{ vs. } H_1 : \theta > \theta_0, \tag{17}$$

and consider the case where $\theta_0 = .5, n = 20, \overline{x}_n = .2$. This framing ensures that a UMP N–P test for the hypotheses in (17) based on $\mathrm{T}_\alpha := \{d(\mathbf{X}) = \frac{\sqrt{n}(\overline{X}_n - \theta_0)}{\sqrt{\theta_0(1-\theta_0)}}, C_1(\alpha) = \{\mathbf{x}: d(\mathbf{x}) > c_\alpha\}\}$ for $\alpha = .05$ yields $d(\mathbf{x}_0) = -2.683$, which indicates that the relevant range of values for θ^* lies outside $\mathbf{\Theta}_0 \cup \mathbf{\Theta}_1$. $d(\mathbf{x}_0) = -2.683$ gives rise to 'accept H_0' with a p value $p(\mathbf{x}_0) = .996$! This absurd result stems from the ill-chosen framing in (17) that disregards both N–P preconditions [i]–[ii] and ensures that the (implicit) power of this test in detecting all relevant discrepancies $(\theta - \theta_0) < 0$ is less than α. Such absurd testing results are easily preventable by adhering to the N–P preconditions.

Hence, when no reliable information about the potential range of values for θ^* is available, the N–P test will be more appropriate with a two-sided partition of Θ:

$$H_0 : \theta = \theta_0 \text{ vs. } H_1 : \theta \neq \theta_0. \tag{18}$$

When such information is available, the appropriate framing is one-sided (directional) with H_1 framed to include the relevant range of value for θ^*. In the case of the above example, the framing would be H_0: $\theta \geq \theta_0$ versus H_1: $\theta < \theta_0$, which would have rejected H_0 with a p value $p(\mathbf{x}_0) = .004$!

Regrettably, such ill-chosen framings of H_0 and H_1 are routinely used to (misleadingly) criticize N–P testing as inherently problematic when in fact the framing in (17) runs afoul one or both preconditions [i]–[ii]!

2.5 Error probabilities *cannot* be conditional on θ

To shed light on why conditioning on θ makes no formal or logical sense in frequentist inference, one needs to return to the basic axiomatic approach (Kolmogorov, 1933) where probability theory is erected on a probability space $(S, \Im, \mathbb{P}(.))$, with S denoting

the set of all (logically) possible distinct outcomes, $\Im$ the set of all events ($A \subset S$) of interest and related events that enjoys the mathematical structure of a sigma (σ)-field-$\Im$ is closed under the set-theoretic operations of union, intersection, and complementation –, and $\mathbb{P}(.)$: $\Im \to 0, 1]$ assigns probabilities to events (elements) in $\Im$. Kolmogorov (1933, p. v) points out that the concept of a σ-field played a key role in the axiomatization of probability through Lebesgue's measure theory (Shiryaev, 2016, p. 187). Random variables are defined relative to $\Im$ in the sense that a function $X(.)$: $S \to \mathbb{R}$ is said to be a random variable if its pre-image $((X(s) \leq x) = X^{-}(x)$, for all $s \in S$ and $x \in \mathbb{R}$) defines events in $\Im$ ensuring that X defines a subset of events $\sigma(X)$ of $\Im$ known as the minimal σ-field generated by X.

To make the case that error probabilities are conditional on θ, one needs to demonstrate the mathematical meaning of $f(h(\mathbf{x})|\theta)$, for any statistic $h(\mathbf{X})$, and defined by (Williams, 2001, p. 258):

$$f(h(\mathbf{x})|\theta = \vartheta) = \frac{f(h(\mathbf{x}), \theta = \vartheta)}{f(\vartheta)}, \quad \forall \mathbf{x} \in \mathbb{R}^n_X \tag{19}$$

for a particular value ϑ in Θ. Given that, in frequentist inference, θ is not an event or a random variable defined relative to σ-field $\Im$ of the probability space $(S, \Im, \mathbb{P}(.))$ underlying $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$, (19) makes no mathematical sense. That is, (19) does not exist as a probabilistic concept since there is no well-defined joint distribution $f(\mathbf{x}, \theta)$ to determine the numerator $f(\mathbf{x}, \theta = \vartheta)$, or the denominator $f(\vartheta) = \int_{\mathbf{x} \in \mathbb{R}^n_X} f(\mathbf{x}, \vartheta) d\mathbf{x}$. This is not just a matter of 'inept' terminology, but a crucial issue that concerns the *non-existence* of the two concepts $f(\mathbf{x}|\theta = \vartheta)$, $\forall \mathbf{x} \in \mathbb{R}^n_X$ and $f(\theta|\mathbf{X} = \mathbf{x}_0)$, $\forall \theta \in \Theta$, in the context of frequentist inference. Even when viewed at a more intuitive level, factual (presuming that $\theta = \theta^*$) and hypothetical (presuming that $\theta = \theta_0$) reasoning do not entail probabilistic conditioning since the latter pertains to 'information that an event A in $\Im$ has occurred'. Hence, invoking the misleading set phrase 'given H_0' as bespeaking mathematical conditioning is ridiculous. What makes mathematical and logical sense is to define $f(h(\mathbf{x}); \theta = \vartheta)$, $\forall \mathbf{x} \in \mathbb{R}^n_X$, for prespecified values of θ and derived it via (6) using factual or hypothetical reasoning.

As a counter-argument to the above case, one might hazard the counter-claim that θ can be transformed into a special random variable that relates to two events $A = \{\theta$: $\theta = \theta_0\}$ and $\overline{A} = \{\theta$: $\theta \neq \theta_0\}$, with the relevant σ-field of interest being $\mathcal{F} = \left\{S, \emptyset, A, \overline{A}\right\}$, and $\mathbb{P}(A) = 0$, or $\mathbb{P}(A) = 1$. Regrettably, this idea crumbles instantly since the two random variables X and θ can only be related as in (19) when they are both defined on the *same* probability space, $(S, \Im, \mathbb{P}(.))$, whose σ-field $\Im$ is required to include all possible unions, intersections, and complementations of all the events relating to both! Worse, the mapping $\theta(s) = \theta_0$ for all $s \in S$ defines a *degenerate* (constant) random variable which, by construction, is independent of every other random variable X defined on $(S, \Im, \mathbb{P}(.))$ (Renyi, 1970, p. 201), i.e., there is no joint $\mathbb{P}(x, \theta)$ or conditional $\mathbb{P}(x|\theta)$ probability to be had.

More astounding is the impossibility of constructing a σ-field $\Im$ that includes all the joint events associated with θ and X even when θ is a *proper* random variable with its own prior distribution $\pi(\theta)$, $\forall \theta \in \Theta$. That is, this problem lies abeyant at

the very foundation of Bayesian statistics. The traditional derivation of Bayes theorem circumnavigates this problem by *reimagining* the frequentist distribution of the sample $f(\mathbf{x};\theta)$ as (somehow) conditional on θ, i.e. $f(\mathbf{x}|\theta)$, $\forall \mathbf{x} \in \mathbb{R}^n_X$. This finessing enables Bayesians to define—without any intellectual effort—the (contrived) joint distribution by $f(\mathbf{x},\theta) = f(\mathbf{x}|\theta)\cdot\pi(\theta)$, $\forall\theta\in\Theta$, $\forall \mathbf{x} \in \mathbb{R}^n_X$; see Sect. 6.2.

3 Bernoulli's golden theorem in retrospect

Assuming the simple Bernoulli model in (5), Bernoulli's *golden theorem* asserts:

$$\mathbb{P}\big(\big|\overline{X}_n-\theta\big|<\varepsilon\big)\geq 1-\delta,\ \text{for } \varepsilon>0,\ \delta>0,\ \text{and all } n\geq N. \tag{20}$$

The retrospective view of this theorem is guided by Le Cam's (1986) perspective on limit theorems encapsulated by the following quotation: "… limit theorems 'as n tends to infinity' are logically devoid of content about what happens at any particular n. All they can do is suggest certain approaches whose performance must then be checked on the case at hand. Unfortunately, the approximation bounds we could get were too often too crude and cumbersome to be of any practical use." (p. xiv).

3.1 Bernoulli's law of large numbers

The most pivotal way Bernoulli's golden theorem influenced probability and statistical inference arose from its implications as $n\to\infty$ (Hald, 1998, 2007). When placed in the context of model-based frequentist inference, the statistical model underlying the result is the simple Bernoulli model.

3.1.1 Bernoulli's WLLN

For a Bernoulli IID process $\{X_k, k\in\mathbb{N}\}$ in (5):

$$\lim_{n\to\infty}\mathbb{P}\big(\big|\overline{X}_n-\theta\big|<\varepsilon\big)=1,\ \text{for } \varepsilon>0, \tag{21}$$

where $\overline{X}_n=\frac{1}{n}\sum_{k=1}^n X_k$. (21) follows from (20) and (25) since $\delta=\frac{\theta(1-\theta)}{\varepsilon^2 n}\to_{n\to\infty}0$; see Billingsley (1995), p. 5.

The result in (21) provided the first formal justification for the *frequentist interpretation* of probability of an event A as the *limit* of the 'stable long-run relative frequency' $\overline{x}_n=\frac{1}{n}\sum_{k=1}^n x_k$, in the context of the statistical model $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ in (5). This was the first limit theorem known as the Weak Law of Large Numbers (WLLN). Almost two centuries later, Bernoulli's WLLN was strengthened by Borel in 1909 in the form of a Strong Law of Large Number (SLLN).

3.1.2 Borel's SLLN

For an IID Bernoulli process $\{X_k, k \in \mathbb{N}\}$ in (5):

$$\mathbb{P}\Big(\lim_{n\to\infty} \overline{X}_n = \theta\Big) = 1. \tag{22}$$

That is, as $n \to \infty$ the process $\{\overline{X}_n\}_{n=1}^{\infty}$, converges to $\theta = E(X_k)$ *with probability one*, or *almost surely* (*a.s.*); see Billingsley (1995), p. 8.

3.1.3 Probabilistic versus mathematical convergence

It is important to distinguish between the above forms of probabilistic convergence (21)–(22) from the *mathematical convergence* invoked by von Mises (1928):

$$\lim_{n\to\infty} \overline{x}_n = \theta, \tag{23}$$

where $\overline{x}_n$ denotes the values of $\overline{X}_n$, since neither (21) nor (22) entails (23). As argued by Williams (2001), p. 25, any attempt to make rigorous the mathematical convergence $\lim_{n\to\infty}\overline{x}_n = \theta$ is ill-fated for purely mathematical reasons which can only be circumvented using measure theory. Historically, the line between probabilistic and mathematical convergence $\lim_{n\to\infty}\overline{x}_n = \theta$ was blurred by von Mises's (1928) notion of a *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 θ. This has led to the widespread confusion that lingers on to today between probabilities and relative frequencies by misidentifying the frequentist interpretation of probability with the long-run metaphor; see Spanos (2013a).

3.2 Bernoulli's golden theorem versus his numerical example

From today's perspective, Bernoulli's golden theorem amounts to a finite sample approximation to the WLLN in (21). Bernoulli (1713), derived the Binomial (Bin) distribution for $\sum_{k=1}^{n} X_k$ using the homonymous expansion in his discussion of proposition 12 of Part I. He used this result in Part IV, to derive the first finite 'sampling distribution' of the sum:

$$Y := n\overline{X}_n = \sum_{k=1}^{n} X_k \backsim \text{Bin}(n\theta, n\theta(1-\theta); n). \tag{24}$$

In retrospect, his derivation of (20) was based on approximating the Binomial tail areas, which today is better approximated using Chebyshev's inequality:

$$\mathbb{P}\big(\big|\overline{X}_n - \theta\big| \geq \varepsilon\big) \leq \frac{\theta(1-\theta)}{\varepsilon^2 n}, \tag{25}$$

which implies that since $\theta(1-\theta) \leq \frac{1}{4}$, for $\delta = \mathbb{P}\left(\left|\overline{X}_n - \theta\right| \geq \varepsilon\right)$ (20) holds for any:

$$n \geq N = \frac{\theta(1-\theta)}{\varepsilon^2 \delta} \leq \frac{1}{4\varepsilon^2 \delta}. \tag{26}$$

3.2.1 Bernoulli's example

In discussing the golden theorem's interpretation and implications for inference, it is important to distinguish between the above generic results in (24)–(26) and Bernoulli's numerical example based on $\theta^* = .6$, $\varepsilon = .2$ and $\delta = .001$ since the example has often been misinterpreted.

Using Bernoulli's numerical example (26) implies that: $N = [4(.2)^2(.001)]^{-1} = 6250$. That is, for any $n \geq 6250$ the lower and upper bounds, $\left(\overline{X}_n - \varepsilon\right)$ and $\left(\overline{X}_n + \varepsilon\right)$, respectively, will include (overlay) the true value of θ, say θ^*, with probability $(1-\delta) = .999$.

It is worth noting that Bernoulli's (1713) bound for δ was much less accurate than (26), yielding $N = 25550$, "… because of two crude approximations. First, he requires that the basic inequality holds for each tail separately, instead of their sum. … Second, he uses the arithmetic approximation for the tail probability instead of the geometric one." (Hald, 2007, p. 14). It is also important to bring out the fact that the lower bound for (20) yielding $N = 6250$ does not use the information that $\theta^* = .6$. Naturally, when this information is used, $\theta(1-\theta) = .24$, yielding a smaller $N = (.24)[(.2)^2(.001)]^{-1} = 6000$.

In light of the above comments, one should separate the golden theorem from Bernoulli's numerical example to illustrate it. His illustration is no different in substance from *demonstrating* the golden theorem today using simulation or an analytical calculation for particular values of $\theta, \varepsilon, \delta$, and N. The simulation can be used to establish the relevant tail areas empirically based on a large number (say $N = 10{,}000$) of sample realizations $\mathbf{x}_i$, $i = 1, 2, \ldots, N$, of size n. Hence, it will be a mistake from today's perspective to view Bernoulli's theorem as (somehow) tainted by his use of the information $\theta^* = .6$ to illustrate it since such information is irrelevant for the theorem in (20) to hold.

3.3 Revisiting Bernoulli's alleged swindle

Influenced by the legal tradition of his time, Bernoulli (1713) understood the magnitude of probability $\mathbb{P}(A)$ as degrees of certainty along a graduated spectrum of belief ranging from total ignorance ($\mathbb{P}(A) \simeq 0$) to firm conviction ($\mathbb{P}(A) \simeq 1$) or moral certainty: "something is morally certain if its probability comes so close to complete certainty that the difference cannot be perceived." (p. 315). In his numerical example, an event (conjecture) A is morally certain when $\mathbb{P}(A) = .999$.

3.3.1 Bernoulli's alleged swindle

Diaconis and Skyrms (2018) argue that Bernoulli committed a 'swindle' by viewing his golden theorem in (20), in conjunction with his notion of *moral certainty,* to infer:

$$\overline{x}_n \simeq \theta^*, \text{ for } n \geq N, \tag{27}$$

where $\overline{x}_n$ denotes the observed value of $\overline{X}_n = \frac{1}{n}\sum_{k=1}^{n} X_k$.

A retrospective view suggests that (27) is just a special case of the unwarranted claim in (2), with $\hat{\theta}(\mathbf{x}_0) = \overline{x}_n$, potentially stemming from misinterpreting (21) as entailing (23); see Spanos (2013b). The claim in (27), to the extent it persists today, stems primarily from misconstruing the *long-run* metaphor that aims to conceptualize the link between relative frequencies and probabilities. In the context of model-based frequentist inference, probabilities are *not identified* with relative frequencies, but rather probabilities are evidenced by stable relative frequencies based on a statistically adequate $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$; see Spanos (2013a). As argued by Hacking (1980): "Probability in this sense [frequentist] does not mean 'relative frequency', but probabilities are typically manifested by stable frequencies." (p. 150). 'Typically' refers to 'the particular data $\mathbf{x}_0$ being a *typical realization* of the prespecified $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ or equivalently, the probabilistic assumptions comprising $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ are valid for $\mathbf{x}_0$. Hence, (21) is justified on empirical and not on a priori – rational defense in reason – grounds.

The *intuition* underlying Bernoulli's golden theorem could be illustrated in terms of relative frequencies (proportions) as follows: assuming the IID assumptions of $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ in (5) are valid, for large enough n, say $n \geq 6250$, a proportion $\delta = .001$ of the $N = 10{,}000$ sample realizations $\mathbf{x}_i := (x_{1i}, x_{2i}, \ldots, x_{ni})$, $i = 1, 2, \ldots, N$, is likely to exhibit errors (fluctuations around θ^*) outside the band $|\overline{x}_n - \theta^*| < \varepsilon$. Borel's SLLN states that under the same conditions, for $n \geq 6250$ *no* sample realization is likely to exhibit errors outside $|\overline{x}_n - \theta^*| < \varepsilon$. It is important to view this as a heuristic explanation of the theorems (21)–(22) where probabilities are manifested by the relative frequencies; see Spanos (2013a).

Is Bernoulli (1713) guilty of the swindle alleged by Diaconis and Skyrms (2018)? A retrospective case can be made that the combination of his numerical example and his notion of 'moral certainty', are likely to have misled modern readers into conflating the heuristic illustration with the theorem in (21).

3.4 The golden theorem and approximate CIs

As argued above, the statistical adequacy of $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ in (5) is critical for the golden theorem in (20), as well as (21)–(22), to hold. A crucial difference between Bernoulli's and Borel's Law of Large Numbers (LLN) and subsequent variants is that the inductive premises underlying (21) and (22), $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ include an explicit distribution assumption that can be used to simulate the underlying sampling distribution of $\sum_{k=1}^{n} X_k$ in in (24), as shown in Fig. 1, where the Binomial is approximated closely by the Normal distribution.

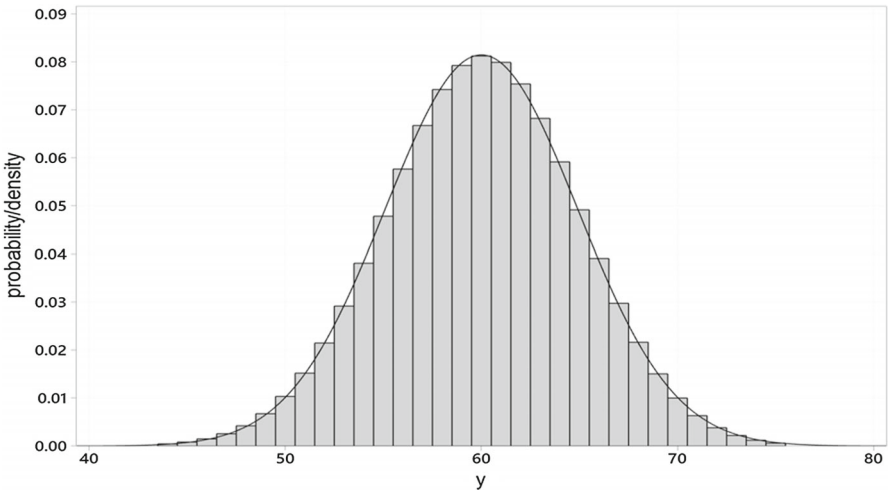

Fig. 1 Bin($n\boldsymbol{\theta}^*, n\boldsymbol{\theta}^*(1-\boldsymbol{\theta}^*); n$) versus N($n\boldsymbol{\theta}^*, n\boldsymbol{\theta}^*(1-\boldsymbol{\theta}^*)$), $\boldsymbol{\theta}^* = .6, n = 100$

Historically, almost all subsequent extensions (generalizations) of the original limit theorems (LLN, CLT) replaced that with indirect distribution assumptions (e.g. existence of certain moments); see Billingsley (1995).

In light of that, the golden theorem in (20) can be used in conjunction with the sampling distribution in (24) to derive an *approximate* frequentist CI:

$$\mathbb{P}\big(\overline{X}_n - \varepsilon < \theta \leq \overline{X}_n + \varepsilon;\ \theta = \theta^*\big) = (1-\alpha), \tag{28}$$

where $\varepsilon = c_{\frac{\alpha}{2}}\sqrt{\overline{X}_n\big(1-\overline{X}_n\big)/n}$, and $c_{\frac{\alpha}{2}}$ relates to the Normal approximation in Fig. 1. Hence, contrary to the Diaconis and Skyrms (2018) claim, Bernoulli did answer the question: "what is the probability that the chances [i.e. $\theta^* = \mathbb{P}(X=1)$] fall within a certain interval?", in the sense that the CI in (28) overlays θ^* with probability $(1-\alpha)$, and not the inverse probability interval $\mathbb{P}(\theta - \varepsilon < \overline{x}_n \leq \theta + \varepsilon|\theta)$.

That is, the legitimacy of this approximate CI in (28) stems from (24) and the fact that δ does not depend on θ^*. Indeed, Laplace (1812) was the first to put forward a similar interval based on *direct probabilities*; see Hald (2007), p. 5. Dempster (1966) argues that the golden theorem can be viewed as a forerunner of Neyman-type CIs. What is even more interesting is that (28) can be sharpened considerably by replacing the $(1-\delta)$ bound with the tails areas of (24).

Example 2 (continued). Using Chebyshev's inequality for $n = 2500$ and $\varepsilon = .1$, implies $\delta = [4(2500)(.1)^2]^{-1} = .01$, one can deduce that the approximate .99 CI:

$$\mathbb{P}\big(\big|\overline{X}_n - \theta\big| \leq \varepsilon\big) \geq .99. \tag{29}$$

On the other hand, when $Z = \frac{\sum_{k=1}^{n}(X_k - n\theta)}{\sqrt{n\theta(1-\theta)}}$ is used to approximate the Binomial with the Normal distribution (de Moivre, 1738), shown in Fig. 1 for $n = 100,\ \theta^* = .6$,

the finite sample .99 CI in (28) requires only $n = 166$ since $\sqrt{n}(.1)/\sqrt{.25} = 2.576 \rightarrow n = 166$, and thus:

$$\mathbb{P}\big(\big|\overline{X}_n - \theta\big| < .1\big) \geq \int_{-2.576}^{2.576} \left(\frac{\exp\big(-.5z^2\big)}{\sqrt{2\pi}}\right) dz = .99. \tag{30}$$

The sizeable reduction of the required sample size n from 2500 to 166 illustrates Le Cam's quotation about asymptotic approximations being "too crude and cumbersome to be of any practical use", and the reduction from 2500 to 166 is typical of asymptotic approximations versus finite sample results; see Spanos (2019).

3.5 Bernoulli and direct versus inverse inference

As argued above, the alleged Bernoulli's swindle in (27) is a special case of the more general unwarranted claim in (2). This calls into question the traditional argument articulated by Diaconis and Skyrms (2018) that the source of the swindle stems from conflating $f(\mathbf{x}|\theta)$ with $f(\theta|\mathbf{x})$. Let us unpack this claim.

Regrettably, Bernoulli's use of the true $\theta^* = .6$ in his numerical example has generated confusion in the literature about legitimate and illegitimate interpretations of the golden theorem, as well as whether the probability in (20) is direct (frequentist) or inverse (Bayesian). As shown above, the lower bound $(1 - \delta)$ in (20) need not rely on knowing θ^* since $\delta = \frac{\theta(1-\theta)}{\varepsilon^2 n} \leq \frac{1}{4\varepsilon^2 n}$. Also, it is not obvious what the claim by Diaconis and Skyrms (2018): "He solved an inference from chances to frequencies" (p. 65) refers to. Why?

To begin with, the probabilistic assignment $P(\theta - \varepsilon < \overline{x}_n \leq \theta + \varepsilon|\theta) \simeq 1$ is meaningless in frequentist inference since there is no random variable involved to justify the assignment $P(.)$; $\overline{x}_n$, θ and ε are known constants.

Second, it is not obvious what the inferential claim: 'assuming θ is known, for a given $\varepsilon > 0$ there is a large enough n such that $\mathbb{P}(\theta - \varepsilon < \overline{X}_n \leq \theta + \varepsilon|\theta) \simeq 1$' could (possibly) mean in frequentist statistics since the golden theorem pertains to a particular value of θ, i.e. θ^*. When $\theta = \theta^*$ is known, the underlying generating mechanism $\mathcal{M}^*(\mathbf{x}) = \{f(\mathbf{x};\theta^*),\ \mathbf{x} \in \{0, 1\}^n$, is fully known for any n; see Fig. 1 for $n = 100,\ \theta^* = .6$. That is, one can just use:

$$f\big(\mathbf{x};\theta^* = .6\big) = (.6)^{\sum_{k=1}^n x_k}(.4)^{\sum_{k=1}^n (1-x_k)} = (.6)^y(.4)^{(n-y)}, \quad y = 0, 1, \ldots, n, \tag{31}$$

where $y = \sum_{k=1}^n x_k$, to evaluate the exact probabilities for different $Y = y$ as in Table 1.

Given that the primary objective of frequentist inference is to learn from data $\mathbf{x}_0$ about θ^*, when θ^* is known no statistical inference is called for or warranted. The notion that one can use $\theta = \theta^*$ to infer something about $\mathbf{x}_0$ is nonsensical since there is no statistical inference to be had; there is no uncertainty about θ^*. Indeed, one can use $\mathcal{M}^*(\mathbf{x})$ to evaluate the probabilities associated with any values of $Y =$

Table 1 Probabilities relating to Fig. 1

y	…	58	59	60	61	62	…
$\mathbb{P}(Y=y)$	…	.0742	.0792	.0812	.0799	.0754	…

$\sum_{k=1}^{n} X_k$ of substantive interest beyond $\mathbf{x}_0$, including predicting future values of X_t. Moreover, since neither $f(\theta|\mathbf{x}_0)$, nor $f(\mathbf{x}_0|\theta)$, exist in frequentist inference (Sect. 2.5), (20) cannot (possibly) be susceptible to the charge of conflating *direct* with *inverse* inference.

One the other hand, when θ is assumed to be a random variable, as in Bayesian statistics, the probabilistic statement $\Pr(\overline{x}_n - \varepsilon < \theta \leq \overline{x}_n + \varepsilon|\mathbf{x}_0) \simeq 1$ stems from the posterior distribution, $\pi(\theta|\mathbf{x}_0) \propto f(\mathbf{x}_0;\theta) \cdot \pi(\theta)$, $\theta \in (0,1)$. This, however, does not render the frequentist interpretation of (20) problematic in the context of $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ in (5) in any logical or mathematical sense.

4 Revisiting the direct versus inverse inference

4.1 Bayesian deformation of the *p* value?

The question that naturally arises at this stage is: what is the merit of the Bayesian charge that frequentists often confuse $f(\theta|\mathbf{x}_0)$ with $f(\mathbf{x}_0|\theta)$ when neither exists in that context and what does that imply for frequentist testing in particular?

In a section entitled "Bernoulli swindle and hypothesis testing" Diaconis and Skyrms (2018, p. 67), argue: “Suppose a drug company runs randomized trials on a new drug. The drug is either effective or not. *You would like to know the probability that it is effective given the data*. The drug company computes the probability that one would get the result in the data or better, given that the drug is ineffective, and gets a very small number. … To those who do not understand statistics, this is an invitation to Bernoulli’s swindle. It is "morally impossible" to get this value if the drug is ineffective. Therefore the drug is effective.”

A Bayesian practitioner would wholeheartedly agree with the sentence in italics since probability refers to his/her degrees of belief, but why do the authors presume that this claim has any meaning in frequentist testing where the drug does not have a "probability of being effective", whether or not given the data. As argued below, N–P testing results can provide reliable evidence ‘whether the drug is effective or not’ when appropriately interpreted using the post-data severity evaluation to establish the warranted discrepancy γ from the null value; see also Mayo and Spanos (2011).

The above quotation echoes Cohen’s (1994) more direct calumny: “When one tests H_0, one is finding the probability that the data (D) could have arisen if H_0 were true, $P(D|H_0)$. If that probability is small, then it can be concluded that if H_0 is true, then D is unlikely. Now, what really is at issue, what is always the real issue, is the probability that H_0 is true, given the data, $P(H_0|D)$, the inverse probability.” (p. 998).

Numerous papers in the replication literature (Wasserstein et al., 2019) declare:

$$P(H_0|D) \neq P(D|H_0), \tag{32}$$

self-evident, and proceed to admonish frequentist testing. As argued in Sect. 2.5, when (32) is properly defined takes the form in (19), which does not exist in frequentist inference. Why the confusion? The unwarranted claim in (32) pertains to any two events A and B, where the relevant *formula:*

$$\mathbb{P}(A \cap B) = \mathbb{P}(A|B) \cdot \mathbb{P}(B) = \mathbb{P}(B|A) \cdot \mathbb{P}(A) \tag{33}$$

implies that $\mathbb{P}(B|A) \neq \mathbb{P}(A|B)$ unless $\mathbb{P}(A) = \mathbb{P}(B)$. What is insufficiently appreciated is that (33) involves observable events A and B, in $\Im$; see Spanos (2010). Calling B a hypothesis (H_0) and A data (D) does not render (32) a legitimate claim in the context of $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ since H_0: $\theta = \theta_0$ cannot be an event in $\Im$; see Sect. 2.5.

4.2 From accept/reject H_0 to an evidential interpretation

After a tongue-in-cheek 'praise' for Fisher for avoiding 'Bernoulli's swindle' by proposing "… a methodology and a story about why that is what you want", Diaconis and Skyrms (2018) take the praise back by claiming: "But it is not what you want, is it? You want the probability of effectiveness given the data." (p. 68). Instead of allowing frequentists to articulate what they really want, and try to understand their underlying reasoning, they pronounce "what you really want is a posterior probability from $f(\theta|\mathbf{x}_0)$, $\forall \theta \in \Theta$".

Fisher's (1925) significance testing driven by the p value was recast into an optimal theory of hypothesis testing by Neyman and Pearson (1933), where the type I and II (or power) are used to calibrate the *pre-data capacity* of the test to detect different discrepancies from H_0; see Spanos (2006). Unfortunately, neither account has provided a cogent evidential interpretation of the testing results. Mayo and Spanos (2006) proposed such an evidential interpretation based on a *post-data* evaluation of the testing results that outputs the discrepancy γ from H_0 warranted with high probability by test T_α and data $\mathbf{x}_0$. What is different from previous attempts at providing an evidential interpretation is that error probabilities are viewed and interpreted in the context of the particular *statistical set-up*:

$$[a]\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x}),\ [b]H_0 : \theta \in \Theta_0\ vs.\ H_1 : \theta \in \Theta_1,\ [c]T_\alpha := \{d(\mathbf{X}), C_1(\alpha)\},\ [d]data\,\mathbf{x}_0, \tag{34}$$

which includes the validity of the assumptions comprising $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ vis-à-vis data $\mathbf{x}_0$, the framing of H_0 and H_1 as a partition of Θ, and the sample size n. What is important to emphasize is that the discrepancy γ from H_0 warranted by T_α and $\mathbf{x}_0$, with high probability, provides a more reliable testing-based effect size, which is not vulnerable to the alleged Bernoulli swindle since it does not invoke the unwarranted claim in (2); see Spanos (2013b, 2021).

Contrary to the claim by Diaconis and Skyrms (2018), a frequentist tester agrees with their comment that: "… the p values are only part of the story. There is the *power* of the test …." (p. 116). Indeed, from the post-data severity perspective (Mayo & Spanos, 2011) $p(\mathbf{x}_0) < \alpha$ indicates the presence of 'some' discrepancy γ, but provides no information about its magnitude since (i) the underlying distribution for $p(\mathbf{x}_0)$ is evaluated only under H_0, and (ii) $p(\mathbf{x}_0)$ is vulnerable to the large n problem (e.g. high power). Both problems are addressed using the severity evaluation that takes into account the statistical context in (34), including the power, or equivalently the 'sensitivity' of the test: "By increasing the size of the experiment, we can render it more sensitive, meaning by this that it will allow the detection of … a quantitative smaller departure from the null hypothesis." (Fisher, 1925, pp. 21–22).

Regrettably, 'untutored' practitioners accept the misleading claims (a)–(d) by Diaconis and Skyrms (2018) in the introduction at face value, in concert with similarly erroneous testimonials from Bayesian textbooks, which include:

(e) Ignore the statistical context in (34) because only $\mathbf{x}_0$ has any bearing on the evidence for or against H_0 since Bayesian inference is *data specific*. A feature that has been lionized by Bayesians in the form of the *likelihood principle,* which asserts that for inference purposes $\mathbf{x}_0$ is the only relevant value of $\mathbf{X}$; see Berger and Wolpert (1988).

(f) Accept the unwarranted claim that the p value conflates $P(H_0|D)$ with $P(D|H_0)$ and disparage frequentist testers for conflating the two; see Nickerson (2000).

(g) Keep reminding practitioners that 'what they really want' in terms of inference is the conditional probability of different values of θ given $\mathbf{x}_0$, i.e. the posterior probability based on $f(\theta|\mathbf{x}_0)$, $\forall\theta \in \Theta$.

Arguably, the erroneous referrals and recommendations (a)–(g) have contributed a great deal to the misuse/abuse and misinterpretation of the p value in particular, and frequentist inference results more generally. Adding to this list:

(h) the confusion between the false positive/negative rates in medical diagnostic screening and the type I/II error probabilities that permeates the discussion in the replication crisis (Spanos, 2021), and

(i) a statistically misspecified $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$—its assumptions are invalid for data $\mathbf{x}_0$,

Taken together (a)–(i) provide a much better explanation of why a sizeable percentage of the empirical evidence published in scientific journals is untrustworthy.

5 Bernoulli's alleged swindle and effect sizes

Bernoulli's distinction between chances, referring to $\theta = \mathbb{P}(X = 1)$, and $\overline{x}_n$ as probability a posteriori, referring to relative frequencies, is important because θ is rarely a probability in the context of a statistical model $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$; the Bernoulli distribution is an exception. As argued below, the inferential claim in (27) is unwarranted, not because Bernoulli conflated $f(\mathbf{x}_0|\theta)$ with $f(\theta|\mathbf{x}_0)$, but since (27) is an instance of (2).

5.1 An unwarranted claim: $\hat{\theta}_n(\mathbf{x_0}) \simeq \theta^*$, for a large enough n

As argued in Sect. 2.3, the Law of Large Numbers (LLN) (weak or strong) does not justify the claim $\hat{\theta}_n(\mathbf{x}_0) \simeq \theta^*$, for a large enough n, since the LLN pertains only to what happens at the limit ($n = \infty$). What would it take to find a statistic, say $h(\mathbf{X})$, that would justify the claim $h(\mathbf{x}_0) \simeq \theta^*$? For that one needs to invoke another limit theorem, known as the Law of Iterated Logarithm (LIL) that quantifies the LLN fluctuations of $\hat{\theta}_n(\mathbf{X})$ around θ^*, as described by its sampling distribution $f\left(\hat{\theta}_n(\mathbf{x});\theta\right)$, $\mathbf{x} \in \mathbb{R}^n$, using upper and lower bounds.

As an aside, it is important to note that limit theorems, such as LLN and the LIL revolve around a specific statistic, $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$, but their results can be easily extended to more general statistics $h(\mathbf{X})$; see Spanos (2019), ch. 9.

To implement the LIL, however, one would need to generate additional sample information in the form of N faithful replicas—ones that exhibit the same chance regularity patterns—as the original data $\mathbf{x}_0$, say $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N$, using simulation or bootstrapping (resampling). These replicas are used to evaluate N estimates $\hat{\theta}_n(\mathbf{x}_i)$, $i = 1, 2, \ldots, N$, of θ whose (smoothed) histogram approximates the empirical distribution, say $\hat{f}_N\left(\hat{\theta}(\mathbf{x}_1, , \mathbf{x}_2, \ldots, \mathbf{x}_N);\theta\right)$; the empirical counterpart of the sampling distribution $f\left(\hat{\theta}(\mathbf{x});\theta\right)$, $\mathbf{x} \in \mathbb{R}^n$. Although no single $\hat{\theta}_n(\mathbf{x}_i)$ approximates θ^* unless by happenstance, the overall average of these N estimates provides a 'close enough' approximation:

$$\overline{\hat{\theta}}_N(\mathbf{x}_1, , \mathbf{x}_2, \ldots, \mathbf{x}_N) = \frac{1}{N}\sum_{i=1}^{N} \hat{\theta}_n(\mathbf{x}_i) \simeq \theta^*, \text{ for a large enough } N. \tag{35}$$

The LIL quantifies 'close enough' by providing bounds for the approximation error $\left|\frac{1}{N}\sum_{i=1}^{N} \hat{\theta}_{ML}(\mathbf{x}_i) - \theta^*\right| < \varepsilon$ (Billingsley, 1995, p. 153):

$$(1 \pm \varepsilon)\left(\frac{1}{N}\sqrt{2N\ln\ln(N)}\right) \; for \; any \; \varepsilon > 0. \tag{36}$$

For instance, when $N = 20{,}000$ (36) yields $(1 \pm \varepsilon)(.015)$, ensuring first decimal approximation accuracy, but for $N = 100$ the bounds are not as accurate $(1 \pm \varepsilon)(.175)$.

In practice, the histogram in Fig. 1 can be replicated using simple *bootstrapping* (Efron & Tibshirani, 1993), when the validity of the IID assumptions for data $\mathbf{x}_0$ has been established using comprehensive misspecification testing; see Spanos (2018). This qualification is particularly crucial because any departures from the IID assumptions will render the bootstrap replications unfaithful replicas—they will exhibit different chance regularities than $\mathbf{x}_0$—and the ensuing empirical sampling distribution and its summary statistics will be unreliable; see Spanos (2019), p. 463.

It is important to emphasize that the approximation in (35) is not equivalent to using an enlarged data set $\mathbf{X}_0$ with sample size nN to estimate θ and invoke consistency to claim $\hat{\theta}_{nN}(\mathbf{X}_0) \simeq \theta^*$. What is different in (35) is that the LIL bounds in (36) depend

crucially on the averaging of the N estimates which shortens the range of values of the sampling distribution $\hat{f}_N\Big(\hat{\theta}(\mathbf{x}_1, , \mathbf{x}_2, ..., \mathbf{x}_N); \theta\Big)$ as opposed to that of $\hat{f}\Big(\hat{\theta}_{nN}(\mathbf{x}); \theta\Big)$. That is, the LLN does justify $\hat{\theta}_{nN}(\mathbf{X}_0) \to \theta^*$ (in probability or almost surely), as $nN \to \infty$, but it cannot provide bounds for the approximation error $\left|\hat{\theta}_{nN}(\mathbf{X}_0) - \theta^*\right|$, otherwise the LIL would have been redundant!

It should also be noted that Bernoulli's LLN in the context of (5) can be somewhat misleading for the general case of an arbitrary consistent estimator $\hat{\theta}_n(\mathbf{X}) \to \theta^*$ as $n \to \infty$. As mentioned in Sect. 3.4, it constitutes a special case where the invoked probabilistic assumptions include a direct (explicit) distribution assumption, Bernoulli, ensuring that (a) the finite sampling distribution of $\overline{X}_n = \frac{1}{n}\sum_{k=1}^{n} X_k$, is known, (24), and (b) $Var\big(\overline{X}_n\big) = (\theta(1-\theta)/n)$ is bounded above by $(1/4n)$ since $\theta(1-\theta) \leq (1/4)$. This is not the case with more general limit theorems since they usually rely on indirect distribution assumptions, such as the existence of the first few moments; see Spanos (2019), ch. 9.

5.2 Estimation-based effect sizes

This approximation in (35) has important implications for the replication crisis as they relate to the estimation-based effect sizes. Usually, effect sizes are point estimates of a function of one or more parameters of $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$; see Cohen (1988), Ellis (2010). For instance, in the case of testing the difference between two means, the estimation-based effect size, known as Cohen's $d = \big[(\overline{x}_n - \overline{y}_n)/s\big]$, is nothing more than a point estimate $\hat{\theta}_n(\mathbf{z}_0) = \big[(\overline{x}_n - \overline{y}_n)/s\big]$ of the unknown parameter $\theta = [(\mu_1 - \mu_2)/\sigma]$. This suggests that such estimation-based effect sizes constitute instances of the unwarranted claim (2).

This is important for the current discussions on replicability since numerous recent papers (Nosek & Lakens, 2014) replicate published results to compare the point estimates $\hat{\theta}_n(\mathbf{z}_0)$ of two or more studies to draw inferences relating to the replicability and the trustworthiness of their evidence. Given that $\hat{\theta}_n(\mathbf{x}_0) \simeq \theta^*$ is unwarranted, this strategy is likely to give rise to highly misleading results by the replicators. The above discussion questions the reliability of conclusions of the form 'for a particular published study (i) the statistical significance is replicated based on observed CIs, but (ii) the effect size $\theta = [(\mu_1 - \mu_2)/\sigma]$, measured by Cohen's d is smaller/bigger than the original'. Since particular point estimates depend crucially on the sample size n, as well as the statistical adequacy of $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$, estimates based on different sample sizes or statistically misspecified models ($\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$) will give rise to highly misleading replications results.

A case can be made that a more reliable way to evaluate the replicability of studies is to compare the discrepancies from a null value warranted by an optimal test and data $\mathbf{z}_0$ stemming from the post-data severity evaluation of the testing results that takes fully into account the statistical context in (34); see Spanos (2021).

6 Bayes' theorem and direct versus inverse inference

The traditional interpretation of Bernoulli's golden theorem, as summarized by Diaconis and Skyrms (2018) in the introduction, has been that his inferential claim $\overline{x}_n \simeq \theta^*$, for $n \geq N$, is not just unwarranted, but the problem he posed did not have a legitimate frequentist answer. His answer is based on conflating two different conditional densities $f(\mathbf{x}_0|\boldsymbol{\theta})$ and $f(\boldsymbol{\theta}|\mathbf{x}_0)$. Instead, his inferential problem was solved by Bayes (1764) who introduced the distinction between the two densities. As argued in Sect. 2.5, conditioning on the *unknown* and *unobservable* constant $\boldsymbol{\theta}$ is both mathematically and logically meaningless in model-based frequentist inference; neither density exists. Despite this obvious mathematical fact, Bayesians have convinced many frequentists that the distribution of the sample, $f(\mathbf{x};\boldsymbol{\theta})$, $\mathbf{x} \in \mathbb{R}^n_X$, can be (legitimately) reimagined as $f(\mathbf{x}|\boldsymbol{\theta})$, $\mathbf{x} \in \mathbb{R}^n_X$, giving rise to a reinterpreted likelihood function $L(\boldsymbol{\theta}|\mathbf{x}_0) \propto f(\mathbf{x}_0|\boldsymbol{\theta})$, $\forall \boldsymbol{\theta} \in \Theta$, as well as a transposed conditioning, to define $f(\theta|\mathbf{x}_0)$, $\forall \theta \in \Theta$, when neither makes sense in frequentist statistics. Why? The short answer is that it allows Bayesians to use a dubious crosscut to render Bayes' rule easier to define, justify and apply. Let us unpack this claim in finer detail.

6.1 Revisiting the traditional Bayes' rule

According to Ghosh et al., (2006), Bayes' rule takes the form:

$$\pi(\boldsymbol{\theta}|\mathbf{x}) = \frac{f(\mathbf{x}|\boldsymbol{\theta}) \cdot \pi(\boldsymbol{\theta})}{\int_{\boldsymbol{\theta}\in\Theta} f(\mathbf{x}|\boldsymbol{\theta}) \cdot \pi(\boldsymbol{\theta}) \mathbf{d}\boldsymbol{\theta}}, \tag{37}$$

"where $\pi(\boldsymbol{\theta})$ is the prior density function and $f(\mathbf{x}|\boldsymbol{\theta})$ is the density of $\mathbf{X}$, interpreted as the conditional density of $\mathbf{X}$ given $\boldsymbol{\theta}$. The numerator is the joint density of $\boldsymbol{\theta}$ and $\mathbf{X}$ and the denominator is the marginal density of $\mathbf{X}$." (p. 31).

The formula in (37) and the Ghosh et al. (2006) description of its components are both misleading. To reveal the flaws, compare a more accurate definition of Bayes' rule that includes a needed quantifier:

$$\pi(\boldsymbol{\theta}|\mathbf{x}_0) = \frac{f(\mathbf{x}_0|\boldsymbol{\theta}) \cdot \pi(\boldsymbol{\theta})}{\int_{\boldsymbol{\theta}\in\Theta} f(\mathbf{x}_0|\boldsymbol{\theta}) \cdot \pi(\boldsymbol{\theta}) \mathbf{d}\boldsymbol{\theta}}, \quad \forall \boldsymbol{\theta} \in \Theta, \tag{38}$$

for $f(\mathbf{x}_0) = \int\limits_{\boldsymbol{\theta}\in\Theta} f(\mathbf{x}_0|\boldsymbol{\theta}) \cdot \pi(\boldsymbol{\theta}) \mathbf{d}\boldsymbol{\theta} > 0$, where data $\mathbf{x}_0$ represents a point in the sample space $\mathbb{R}^n_X$. When (37) is compared to (38), the obvious differences are that the subscript 0 for $\mathbf{x}_0$, and the quantifier $\forall \boldsymbol{\theta} \in \Theta$ are missing, rendering the description of its components problematic in so far as:

[i] $f(\mathbf{x}_0|\boldsymbol{\theta})$ is *not* the conditional density of $\mathbf{X}$ given $\boldsymbol{\theta}$; it is an amalgam from different conditional distributions with a fixed $\mathbf{x}_0$ and varying values of $\boldsymbol{\theta}$ in $\Theta \subset \mathbb{R}^m$, $n > m$. Besides, the conditional density of $\mathbf{X}$ given $\boldsymbol{\theta}$ requires the quantifier $\forall \mathbf{x} \in \mathbb{R}^n_X$, and not $\forall \boldsymbol{\theta} \in \Theta$.

[ii] The product $f(\mathbf{x}_0|\boldsymbol{\theta}) \cdot \pi(\boldsymbol{\theta})$, $\forall \boldsymbol{\theta} \in \Theta$, is not the joint density of $\boldsymbol{\theta}$ and $\mathbf{X}$, because $f(\mathbf{x}, \boldsymbol{\theta})$ would require a double quantifier $\forall \boldsymbol{\theta} \in \Theta$, $\forall \mathbf{x} \in \mathbb{R}^n_X$ with a generic $\mathbf{x} \in \mathbb{R}^n_X$.

[iii] $\int_{\boldsymbol{\theta}\in\Theta} f(\mathbf{x}_0|\boldsymbol{\theta})\cdot\pi(\boldsymbol{\theta})\mathbf{d}\grave{\boldsymbol{\theta}} = f(\mathbf{x}_0)$ is a scaling factor and not the marginal density of $\mathbf{X}$, which is defined by $f(\mathbf{x}), \quad \forall \mathbf{x}\in\mathbb{R}^n_X$.

When one points out the flaws [i]–[iii] in the above quotation from Ghosh et al., (2006), the reply is often framed in terms of 'sloppy language and clumsy notation'. The problem is that this interpretation is typical of Bayesian textbooks more generally; see Lindley (1965), p. 118, O'Hagan (1994), p. 4, and Robert (2007), pp. 8–9, inter alia. It will be equally misplaced to dismiss [i]–[iii] as restating the obvious that 'we all know that …' type of exculpation because the problem is more fundamental and has to do with Bayesians (purposely) reimagining the distribution of the sample $f(\mathbf{x};\boldsymbol{\theta}),\ \mathbf{x}\in\mathbb{R}^n_X$ as conditional on $\boldsymbol{\theta},\ f(\mathbf{x}|\boldsymbol{\theta}),\ \mathbf{x}\in\mathbb{R}^n_X$. Why?

6.1.1 Bayes' foundational problem

Given that in Bayesian inference, $\mathbf{X}$ and $\boldsymbol{\theta}$ are viewed as random variables (vectors), they are both functions defined on the same probability space $(S,\Im,\mathbb{P}(.))$ underlying the relevant $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ based on events:

$$A_x := \mathbf{X}^{-}(-\infty,\mathbf{x}]\in\Im,\ \forall\mathbf{x}\in\mathbb{R}^n,\ B_\vartheta := \boldsymbol{\theta}^{-}(-\infty,\vartheta]\in\Im,\quad \forall\vartheta\in\Theta,$$

where $Z^{-}(.)$ denotes the pre-image of $Z(.)$. Since $\mathbf{X}$ is *observable* and represents real-world events (data), but $\boldsymbol{\theta}$ is *unobservable* and denotes degrees of belief, the foundational problem that arises is how one is supposed to conceptualize and construct the joint density:

$$f(\mathbf{x},\vartheta),\ \ \forall\mathbf{x}\in\mathbb{R}^n_X,\ \ \forall\vartheta\in\Theta, \tag{39}$$

by assigning probabilities to the overlapping events $A_x\cap B_\vartheta\neq\emptyset$ aiming to blend coherently the observable ($\mathbf{X}$) with the unobservable ($\boldsymbol{\theta}$) worlds. If one were to imagine that such a task is (somehow) achievable, then Bayesian inference would be reduced to a simple deductive formula:

$$f(\vartheta|\mathbf{x}_0)=\frac{f(\mathbf{x}_0,\vartheta)}{f(\mathbf{x}_0)},\quad \forall\vartheta\in\Theta. \tag{40}$$

The key difference between (40) with (38) is that $f(\mathbf{x}_0,\boldsymbol{\theta})$ is replaced by $f(\mathbf{x}_0|\boldsymbol{\theta})\cdot\pi(\boldsymbol{\theta})$, where $\pi(\boldsymbol{\theta})$ is chosen independently of $f(\mathbf{x},\vartheta)$ instead of using $f(\vartheta)=\int_{\mathbf{x}\in\mathbb{R}^n_X} f(\mathbf{x},\vartheta)\mathbf{dx}$.

The traditional perspective on Bayesian statistics, however, ignores the above foundational conundrum and defines Bayes' rule using a dubious crosscut to evade the intellectually taxing task in defining (39). Instead of choosing $f(\mathbf{x},\boldsymbol{\theta})$, which will determine both $f(\mathbf{x}|\boldsymbol{\theta})$ and $f(\boldsymbol{\theta})$, Bayesian statistics selects $f(\mathbf{x}|\boldsymbol{\theta})$ and $\pi(\boldsymbol{\theta})$ separately and defines a (contrived) joint distribution via (Gelman, et al., 2004, p. 7):

$$f(\mathbf{x},\boldsymbol{\theta})=f(\mathbf{x}|\boldsymbol{\theta})\cdot\pi(\boldsymbol{\theta}),\quad \forall\mathbf{x}\in\mathbb{R}^n_X,\quad \forall\boldsymbol{\theta}\in\Theta. \tag{41}$$

This conveniently evades the mammoth conundrum of bridging the gap between the real world of data and the mathematical world of prior probabilities pointed out by Le Cam (1977):

"(2) It [Bayesian statistics] confuses 'theories' about nature with 'facts', and makes no provision for the construction of models. (3) It applies brutally to propositions about theories or models of physical phenomena the same simplified logic which every one of us uses ordinarily for 'events'. … (5) The theory blends in the same barrel all forms of uncertainty and treats them all alike." (p. 134).

To be more specific, after reimagining $f(\mathbf{x};\boldsymbol{\theta})$ as $f(\mathbf{x}|\boldsymbol{\theta})$, the second step involves invoking the multiplication rule for density functions which takes the form:

$$f(\mathbf{x},\boldsymbol{\theta}) = f(\boldsymbol{\theta}|\mathbf{x})\cdot f(\mathbf{x}) = f(\mathbf{x}|\boldsymbol{\theta})\cdot f(\boldsymbol{\theta}), \quad \forall \mathbf{x}\in\mathbb{R}^n_X, \quad \forall \boldsymbol{\theta}\in\Theta. \tag{42}$$

The third step mistakenly evaluates (42) at $\mathbf{X}=\mathbf{x}_0$:

$$f(\mathbf{x}_0,\boldsymbol{\theta}) = f(\boldsymbol{\theta}|\mathbf{x}_0)\cdot f(\mathbf{x}_0) = f(\mathbf{x}_0|\boldsymbol{\theta})\cdot f(\boldsymbol{\theta}), \quad \forall \boldsymbol{\theta}\in\Theta, \tag{43}$$

by ignoring the fact that $f(\boldsymbol{\theta}|\mathbf{x}_0)\cdot f(\mathbf{x}_0) \neq f(\mathbf{x}_0|\boldsymbol{\theta})\cdot f(\boldsymbol{\theta})$, since the multiplication rule in (42) holds only when both quantifiers are attached, unlike the one for simple events in (33), since random variables always define more than one simple event in $\Im$. To derive (38), the erroneously derived (43) is then solved for $f(\boldsymbol{\theta}|\mathbf{x}_0)$, thus eliminating $f(\mathbf{x}_0,\boldsymbol{\theta})$ as a result of the sleight of hand in step three hiding the misapplication of (42) as if it were (33).

This sleight of hand suggests that one way to render the above Ghosh et. al (2006) interpretation of Bayes rule's components formally correct is to add both quantifiers:

$$f(\boldsymbol{\theta}|\mathbf{x}) = \frac{f(\mathbf{x}|\boldsymbol{\theta})\cdot\pi(\boldsymbol{\theta})}{\int_{\boldsymbol{\theta}\in\Theta} f(\mathbf{x}|\boldsymbol{\theta})\cdot\pi(\boldsymbol{\theta})\mathbf{d}\grave{}}, \quad \forall\boldsymbol{\theta}\in\Theta,\ \forall\mathbf{x}\in\mathbb{R}^n_X. \tag{44}$$

This describes accurately the above quotation by Ghosh et al. (2006), but has two unusual features:

(i) $f(\boldsymbol{\theta}|\mathbf{x})$ is essentially a simple reparametrization of the contrived $f(\mathbf{x},\boldsymbol{\theta})$, and

(ii) The presence of the quantifier $\forall\mathbf{x}\in\mathbb{R}^n_X$ belies the *Likelihood Principle*: for inference purposes, the only relevant sample information pertaining to $\boldsymbol{\theta}$ is contained in $\mathbf{x}_0$ via the likelihood function $L(\mathbf{x}_0|\boldsymbol{\theta})\propto f(\mathbf{x}_0|\boldsymbol{\theta})$, $\forall\boldsymbol{\theta}\in\Theta$. Moreover, if two sample realizations are proportional, $\mathbf{x}_0 = c\mathbf{y}_0$, for some $c>0$, they contain the same information about $\boldsymbol{\theta}$ (Berger & Wolpert, 1988, p. 19).

In light of the above discussion, Bayesian statistics need to choose between a formally correct Bayes' rule as in (38) and forsake the misleading interpretation associated with (37), or adopt the formula in (44) and give up the likelihood principle. A third, and more practical choice is to do away with the interpretation of the various components in (37), as per Ghosh et al., (2006), and view it as an updating formula whose interpretation is deemed irrelevant. All three choices, however, leave unresolved the key conundrum of bridging the gap between the real world of data and the mathematical world of *prior* probability. The questionable crosscut in (41) can be defended on

pragmatic grounds, but using it to admonish frequentists for conflating $f(\boldsymbol{\theta}|\mathbf{x}_0)$ with $f(\mathbf{x}_0|\boldsymbol{\theta})$ is absurd, since neither density makes sense in that context.

6.2 Alternative ways to 'learn from data'

In frequentist inference, bridging the gap between the real world of data $\mathbf{x}_0$ and the mathematical world of probability constitutes the essence of *statistical induction*: learning from data $\mathbf{x}_0$ about $\mathcal{M}^*(\mathbf{x}) = \left\{f\left(\mathbf{x};\boldsymbol{\theta}^*\right),\ \mathbf{x} \in \mathbb{R}_X^n\right\}$ that gave rise to data $\mathbf{x}_0$. This is achieved by first securing the statistical adequacy of $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ vis-a-vis data $\mathbf{x}_0$, including the constancy of its parameters $\boldsymbol{\theta}$, and then proceed to use effective (optimal) procedures at the inference facet, estimation (point and interval), testing and prediction, knowing that this ensures both the reliability and effectiveness of inference as well as the trustworthiness of the ensuing evidence; see Spanos (2013a).

In contrast, learning from data in Bayesian inference takes the form of revising the prior probability $\pi(\boldsymbol{\theta})$, $\forall \boldsymbol{\theta} \in \Theta$ in light of data $\mathbf{x}_0$, to yield the *posterior* probability $\pi(\boldsymbol{\theta}|\mathbf{x}_0) \propto f(\mathbf{x}_0;\boldsymbol{\theta}) \cdot \pi(\boldsymbol{\theta})$, $\forall \boldsymbol{\theta} \in \Theta$. What is not so obvious is the nature and meaning of the inferential claim that accompanies the revised rankings of $\boldsymbol{\theta}$ by $\pi(\boldsymbol{\theta}|\mathbf{x}_0)$. Is the highest-ranked $\boldsymbol{\theta}$ value, say $\boldsymbol{\theta}^\dagger$, the one approximating $\boldsymbol{\theta}^*$ the best in some sense? There is no decipherable answer to that question in the Bayesian literature since the notion of a 'true value' $\boldsymbol{\theta}^*$ is not well-defined when $\boldsymbol{\theta}$ is a random variable (vector); no single number can characterize a non-degenerate random variable. Looking at Bayesian statistics textbooks, the pragmatic answer seems to be: it depends on the choice of 'a loss function'; see Ghosh et al., (2006), Robert (2007) inter alia. What does expected loss, based on information other than data $\mathbf{x}_0$ and statistical model $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ have to do with learning from data $\mathbf{x}_0$ about $\mathcal{M}^*(\mathbf{x})$? The answer to this question highlights incisively a key difference between the frequentist and Bayesian approaches to inference, as well as 'what learning from data' amounts to in the context of two approaches; see Spanos (2017). In that sense, Bernoulli's (1713) inferential problem was not addressed by Bayes (1764) as often claimed. Bayes recast Bernoulli's inference problem by viewing $\boldsymbol{\theta}$ as a latent random vector and offered an alternative way to learn from data $\mathbf{x}_0$ about observable phenomena of interest.

7 Conclusions

Viewing Bernoulli's (1713) golden theorem retrospectively in the context of model-based frequentist inference that revolves around a parametric statistical model, $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$, the following claims were called into question: (a) Bernoulli solved the problem of "an inference from chances to frequencies", and thus (b) the golden theorem does not justify an approximate CI for θ, (c) Bernoulli's 'swindle' in identifying probability with relative frequency stems from his conflating $f(\mathbf{x}_0|\theta)$ with $f(\theta|\mathbf{x}_0)$, and (d) the same swindle is routinely perpetrated by p value significance testers. In interrogating these claims, the paper argued that they are grounded in misconceptions that raise several broader foundational problems relating to the current replication crisis.

The main conclusions are: (i) Frequentist error probabilities are *not* conditional on hypotheses framed in terms of θ. They are attached to the inference procedure itself to calibrate its effectiveness and grounded in the relevant sampling distribution of a statistic (estimator, test, predictor), which is evaluated via (6) under *factual* (presuming that $\theta = \theta^*$, in estimation and prediction), or *hypothetical* (presuming that $\theta = \theta_0$ or $\theta = \theta_1$ in testing) reasoning. (ii) The alleged Bernoulli's swindle is an instance of the unwarranted claim $\hat{\theta}_n(\mathbf{x}_0) \simeq \theta^*$, for a large enough n, that also undermines point-estimation measures, such as the straight rule and the estimation-based effect sizes. (iii) More reliable measures for the 'scientific effect' can be evaluated using testing-based discrepancies warranted by data $\mathbf{x}_0$; see Spanos (2021). (iv) The direct versus inverse inference problem is a contrived issue that gives rise to misplaced criticisms of model-based frequentist inference since neither distribution $f(\mathbf{x}_0|\theta)$ or $f(\theta|\mathbf{x}_0)$ exists in that context. A case is made that (v) this criticism is motivated by a misguided attempt to justify a dubious crosscut in deriving Bayes' rule by reimagining the distribution of the sample $f(\mathbf{x};\theta)$, $\mathbf{x} \in \mathbb{R}^n_X$, as *conditional* on θ. (vi) The reliability and precision of inferences depend solely on the approximate validity of the probabilistic assumptions comprising $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ for the particular data $\mathbf{x}_0$, and nothing else. Any attempt to invoke limit theorems (as $n \to \infty$) is misplaced. (vii) Bayes (1764) did not address Bernoulli's (1713) inference problem. Instead, he recast the original problem by viewing $\boldsymbol{\theta}$ as a latent random vector and proposed a very different way to learn from data $\mathbf{x}_0$. (viii) Bayesians should consider the dormant foundational problems arising from the choice of a prior as it relates to the dubious crosscut and the erroneous use of the multiplication rule for random variables in (42)–(43) in defining the contrived joint distribution in (41).

Acknowledgements Thanks are due to Prakash Gorroochurn and two anonymous reviewers for several useful comments and suggestions that helped improve the discussion in the paper appreciably.

References

Bayes, T. (1764). An essay towards solving a problem in the doctrine of chances. *Philospical Transactions of the Royal Society of London, 53*, 370–402.

Berger, J. O., & Wolpert, R. W. (1988). *The likelihood principle. Lecture notes—Monograph series* (2nd ed., Vol. 6). Institute of Mathematical Statistics.

Bernoulli, J. (1713/2006). *The art of conjecturing*. JHU Press.

Billingsley, P. (1995). *Probability and measure* (4th ed.). Wiley.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum.

Cohen, J. (1994). The earth is round (p <. 05). *American Psychologist, 49*, 997–1003.

De Moivre, A. (1738). *The doctrine of chances: Or a method of calculating the probability of events in play*. W. Pearson.

Dempster, A. P. (1966). New methods for reasoning towards posterior distributions based on sample data. *The Annals of Mathematical Statistics, 37*(2), 355–374.

Diaconis, P., & Skyrms, B. (2018). *Ten great ideas about chance*. Princeton University Press.

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. Chapman & Hall.

Ellis, P. D. (2010). *The essential guide to effect sizes: Statistical power, meta-analysis, and the interpretation of research results*. Cambridge University Press.

Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society A, 222*, 309–368.

Fisher, R. A. (1925). *Statistical methods for research workers*. Oliver and Boyd.

Gelman, A., Carlin, J. B., & Rubin, D. B. (2004). *Bayesian data analysis* (2nd ed.). Chapman & Hall.
Ghosh, K. J., Delampady, M., & Samata, T. (2006). *Introduction to Bayesian analysis*. Springer.
Gorroochurn, P. (2012). *Classic problems of probability*. Wiley.
Hacking, I. (1965). Salmon's vindication of induction. *The Journal of Philosophy, 62*(10), 260–266.
Hacking, I. (1980). The theory of probable inference: Neyman, Peirce and Braithwaite. In D. Mellor (Ed.), *Science, belief and behavior: Essays in honour of Richard B* (pp. 141–160). Cambridge University Press, Cambridge.
Hald, A. (1998). *A history of mathematical statistics from 1750 to 1930*. Wiley.
Hald, A. (2007). *A history of parametric statistical inference from Bernoulli to Fisher, 1713–1935*. Springer.
Henderson, L. (2020). The problem of induction. *The Stanford Encyclopedia of Philosophy*. Edward N. Zalta (ed.). https://plato.stanford.edu/archives/spr2020/entries/induction-problem/.
Howson, C., & Urbach, P. (2006). *Scientific rasoning: The Bayesian approach* (3rd ed.). Open Court.
Hume, D. (1748). *An Enquiry Concerning Human Understanding*. Oxford: Oxford University Press.
Kolmogorov, A. N. (1933). *Foundations of the theory of Probability*, 2nd English edition, NY: Chelsea Publishing Co.
Laplace, P. S. (1812). *Théorie analytique des Probabilités* (Vol. 2). Courcier Imprimeur.
Le Cam, L. (1977). A note on metastatistics or 'an essay toward stating a problem in the doctrine of chances". *Synthese, 36*, 133–160.
Le Cam, L. (1986). *Asymptotic methods in statistical decision theory*. Springer.
Lehmann, E. L., & Romano, J. P. (2005). *Testing statistical hypotheses*. Springer.
Lindley, D. V. (1965). *Introduction to probability and statistics from the bayesian viewpoint* (Vol. 1). Cambridge University Press.
Mayo, D. G., & Spanos, A. (2006). Severe testing as a basic concept in a Neyman–Pearson philosophy of induction. *British Journal for the Philosophy of Science, 57*, 323–357.
Mayo, D. G., & Spanos, A. (2011). Error statistics. Philosophy of statisticsIn D. Gabbay, P. Thagard, & J. Woods (Eds.), *The handbook of philosophy of science* (Vol. 7, pp. 151–196). Amsterdam: Elsevier.
Neyman, J. (1937). Outline of a theory of statistical estimation based on the classical theory of probability. *Philosophical Transactions of the Royal Statistical Society of London, A, 236*, 333–380.
Neyman, J., & Pearson, E. S. (1933). On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Statistical Society of London, A, 231*, 289–337.
Nickerson, R. S. (2000). Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods, 5*(2), 241–301.
Nosek, B. A., & Lakens, D. E. (2014). A method to increase the credibility of published results. *Social Psychology, 45*, 137–141.
O'Hagan, A. (1994). *Bayesian inference*. Edward Arnold.
Renyi, A. (1970). *Foundations of probability*. Holden-Day.
Robert, C. P. (2007). *The Bayesian choice: From decision-theoretic foundations to computational implementation* (2nd ed.). Springer.
Salmon, W. C. (1967). *The foundations of scientific inference*. University of Pittsburgh Press.
Shiryaev, A. N. (2016). *Probability-1* (2nd ed.). Springer.
Sober, E. (2008). *Evidence and evolution: The logic behind the science*. Cambridge University Press.
Spanos, A. (2006). Where do statistical models come from? Revisiting the problem of specification. In J. Rojo (Ed.), *Optimality: The Second Erich L. Lehmann Symposium*. Lecture Notes-Monograph Series, (Vol. 49, pp. 98–119). Institute of Mathematical Statistics.
Spanos, A. (2010). Is frequentist testing vulnerable to the base-rate fallacy? *Philosophy of Science, 77*, 565–583.
Spanos, A. (2013a). A frequentist interpretation of probability for model-based inductive inference. *Synthese, 190*, 1555–1585.
Spanos, A. (2013b). Who should be afraid of the Jeffreys–Lindley paradox? *Philosophy of Science, 80*, 73–93.
Spanos, A. (2017). Why the decision-theoretic perspective misrepresents frequentist inference. In: *Advances in statistical methodologies and their applications to real problems* (pp. 3–28). ISBN 978-953-51-4962-0.
Spanos, A. (2018). Mis-specification testing in retrospect. *Journal of Economic Surveys, 32*(2), 541–577.
Spanos, A. (2019). *Probability theory and statistical inference: Empirical modeling with observational data*. Cambridge University Press.

Spanos, A. (2021). Revisiting noncentrality-based confidence intervals, error probabilities and estimation-based effect sizes. *Journal of Mathematical Psychology, 104*, 102580.
Spanos, A., & McGuirk, A. (2001). The model specification problem from a probabilistic reduction perspective. *Journal of the American Agricultural Association, 83*, 1168–1176.
Spanos, A., & Mayo, D. G. (2015). Error statistical modeling and inference: Where methodology meets ontology. *Synthese, 192*, 3533–3555.
Von Mises, R. (1928). *Probability, statistics and truth* (2nd ed.). Dover.
Wasserstein, R. L., Schirm, A. L., & Lazar, N. A. (2019). Moving to a world beyond p < .05. *American Statistian, 73*(Suppl. 1), 1–19.
Williams, D. (2001). *Weighing the odds: A course in probability and statistics*. Cambridge University Press.