Is Frequentist Testing Vulnerable to the Base-Rate Fallacy?
Author(s): Aris Spanos
Source: *Philosophy of Science*, Vol. 77, No. 4 (October 2010), pp. 565-583
Published by: The University of Chicago Press on behalf of the Philosophy of Science Association
Stable URL: http://www.jstor.org/stable/10.1086/656009
Accessed: 30/03/2011 15:36

Is Frequentist Testing Vulnerable to the Base-Rate Fallacy?*

Aris Spanos†‡

This article calls into question the charge that frequentist testing is susceptible to the base-rate fallacy. It is argued that the apparent similarity between examples like the Harvard Medical School test and frequentist testing is highly misleading. A closer scrutiny reveals that such examples have none of the basic features of a proper frequentist test, such as legitimate data, hypotheses, test statistics, and sampling distributions. Indeed, the relevant error probabilities are replaced with the false positive/negative rates that constitute deductive calculations based on known probabilities among events. As a result, the ampliative dimension of frequentist induction—learning from data about the underlying data-generating mechanism—is missing.

1. Introduction. In psychology, the base-rate fallacy refers to the error people commit by ignoring the relative sizes of population subgroups when assessing the likelihood of contingent events involving the subgroups; see Tversky and Kahneman (1982). Various recent experiments, however, have shown that when the same information is presented to people in terms of 'natural (relative) frequencies', the error in reasoning often disappears; see Krämer and Gigerenzer (2005).

In the context of Bayesian reasoning, the base-rate fallacy has been formalized as the error involved when the conditional probability of a hypothesis H given some evidence E, $P(H|E)$—known as the posterior probability—is assessed on the basis of $P(E|H)$ without taking account

*Received August 2009; revised January 2010.

†To contact the author, please write to: Department of Economics, Virginia Tech, Blacksburg, VA 24061; e-mail: aris@vt.edu.

‡I would like to thank Stathis Psillos for encouraging me to focus on this problem and two anonymous referees whose constructive comments and suggestions helped to improve the article.

Philosophy of Science, 77 (October 2010) pp. 565–583. 0031-8248/2010/7704-0008$10.00

of the prior probability (base rate) of H, $P(H)$. This fallacy stems from the fact that since probability calculus gives the relationship

$$P(H|E) = P(E|H)\left[\frac{P(H)}{P(E)}\right], \tag{1}$$

when evaluating $P(H|E)$, the base rates $[P(H), P(E)]$ can be ignored at one's peril.

What is not so apparent is whether frequentist testing, which attaches no probabilities to hypotheses ($P(H)$ is meaningless), is also vulnerable to such a fallacy. Howson (1997, 2000) and Achinstein (2001, 2010), inter alia, contend that frequentist testing, in general, and the severity assessment (Mayo 1996), in particular, are not just susceptible to this fallacy, they are totally undermined by it.

The main objective of this article is to call into question this claim by demonstrating that the circumstantial case made against frequentist testing and the severity assessment, on the basis of examples like the Harvard Medical School test, is highly misleading. A closer scrutiny of such examples reveals that the reasoning connecting a medical to a statistical test using the false positive and false negative rates as analogous to the type I and type II error probabilities, respectively, is fallacious because it grossly misrepresents frequentist testing. It is argued that such examples have none of the basic features of a proper frequentist test (legitimate hypotheses, test statistics, data, sampling distributions, etc.), and relevant error probabilities are misconstrued as conditional probabilities among events under the guise of false positive and false negative rates. In fact, proper error probabilities are not conditional; they depend crucially on the sample size n, and they are invariably assigned to inference procedures, never to events.

Section 2 presents a brief introduction to frequentist testing, including the severity assessment, with a view to bring out certain crucial features that are either misunderstood or misrepresented by the base-rate fallacy argument. Section 3 revisits this argument with a view to untangle the various confusions pervading this argument and to reveal the fallacious reasoning underlying the charge against error-statistical testing.

2. Frequentist Testing and Error Probabilities. Frequentist statistics, pioneered by Fisher (1922), is model-based inductive inference that commences with a prespecified statistical model $\mathcal{M}_\theta(\mathbf{x})$, in the context of which the observed data $\mathbf{x}_0 := (x_1, \ldots, x_n)$ are viewed as a truly typical realization. Fisher (1925, 1934), almost single-handedly, proposed a frequentist theory of optimal estimation, and Neyman and Pearson (1933) modified Fisher's significance testing to put forward an analogous theory for optimal testing; see Cox and Hinkley (1974).

2.1. The Notion of a Statistical Model. Consider the material experiment of randomly selecting n newborn babies in New York City during 2008, the aim being to pose the substantive question of interest:

Is the ratio of boys (B) to girls (G) equal to or greater than 1? (2)

Performing the experiment gives rise to a sequence of outcomes: (B, G, G, B, G, . . . , B). To specify an appropriate statistical model in the context of which this sequence of outcomes can be used to answer the substantive question of interest (2), one needs to embed the material experiment into a statistical model viewed as a purely probabilistic construct.

To that end, one begins by specifying the set of all possible distinct outcomes associated with the above experiment: $\Omega = \{B, G\}$. The mathematical formalism of probability theory requires one to define the event space $\mathfrak{F}$ (the set of events of interest and related events) to be a field—a set of subsets of Ω that is closed under the set theoretic operations of union, intersection, and complementation; see Billingsley (1995). In this example, $\mathfrak{F} = \{\Omega, \varnothing, B, G\}$, where $\varnothing = \{\,\}$ denotes the impossible event. The next step is to assign probabilities to all events in $\mathfrak{F}$:

$$\mathbb{P}(\Omega) = 1, \quad \mathbb{P}(\varnothing) = 0, \quad \mathbb{P}(B) = \theta, \quad \text{and } \mathbb{P}(G) = 1 - \theta,$$
$$\text{where } 0 \leq \theta \leq 1. \tag{3}$$

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

$$X(.): \Omega \to \mathbb{R}, \quad \text{such that } \{X = x\} \in \mathfrak{F} \quad \text{for all } x \in \mathbb{R}. \tag{4}$$

The relevant random variable for the above experiment is defined by

$$\{X = 1\} = B \text{ [or } X(B) = 1] \quad \text{and} \quad \{X = 0\} = G \text{ [or } X(G) = 0].$$

This function defines an r.v. relative to $\mathfrak{F}$ because it satisfies (4):

$$\{X = 1\} \in \mathfrak{F}, \{X = 0\} \in \mathfrak{F}, \text{ and } \{X = x\} = \varnothing \in \mathfrak{F},$$
$$\text{for all real numbers } x \neq 0 \text{ or } x \neq 1.$$

In this sense $X(.)$ assigns numbers to the elementary events in Ω in such a way so as to preserve the original event structure of interest ($\mathfrak{F}$). This extension is important for bridging the gap between the mathematical formalism $(\Omega, \mathfrak{F}, \mathbb{P}(.))$ and the observable stochastic phenomena of interest, since observed data usually come in the form of numbers. The r.v. $X(.)$ maps the original probability space $(\Omega, \mathfrak{F}, \mathbb{P}(.))$ onto the real line in the sense that all the relevant information in $(\Omega, \mathfrak{F}, \mathbb{P}(.))$ and (3) is now encapsulated by the Bernoulli density (nonnegative) function:

$$f(x; \theta) = \theta^x(1 - \theta)^{1-x}, \quad x = 0, 1, \quad 0 \leq \theta \leq 1, \tag{5}$$

where $E(X) = \theta$, $\text{Var}(X) = \theta(1 - \theta)$, and $\sum_{k=0}^{1} f(x; \theta) = (1 - \theta) + \theta = 1$.

The material experiment of "randomly selecting n newborns in New York City during 2008 and noting their gender" can now be framed in the form a stochastic process (an indexed sequence of random variables) $\{X_k, k \in \mathbb{N} := (1, 2, \ldots)\}$, assumed to be Bernoulli, independent, and identically distributed (BerIID).

Collecting all these pieces together defines the simple Bernoulli model:

$$\mathcal{M}_\theta(\mathbf{x})\text{: } X_k \sim \text{BerIID}[\theta, \theta(1-\theta)],$$

$$0 \leq \theta \leq 1, \quad x_k = 0, 1, \quad k = 1, 2, \ldots, \tag{6}$$

assumed to describe an idealized mechanism that gave rise to data $\mathbf{x}_0 := (1, 0, 0, 1, 0, \ldots, 1)$; $X(.)$ has transformed the original sequence $(B, G, G, B, G, \ldots, B)$ into $(1, 0, 0, 1, 0, \ldots, 1)$.

In general, a complete description of a statistical model for inference purposes is given in terms of $f(\mathbf{x}; \theta)$, the distribution of the sample $\mathbf{X} := (X_1, \ldots, X_n)$:

$$\mathcal{M}_\theta(\mathbf{x}) = \{f(\mathbf{x}; \theta), \theta \in \Theta\}, \quad \mathbf{x} \in \mathbb{R}^n_X, \tag{7}$$

where Θ denotes the parameter space (the set of all possible values of the unknown parameter θ) and $\mathbb{R}^n_X$, the sample space (the set of all possible values of the sample $\mathbf{X}$). The data, generically denoted by $\mathbf{x}_0 := (x_1, \ldots, x_n)$ are viewed as a typical realization of the process specified by (7). Note that r.v.'s like the sample $\mathbf{X}$ are specified in capital letters and realizations like the data $\mathbf{x}_0$ by the corresponding small letter.

In the case of the simple Bernoulli model, the parameter and sample spaces are $\Theta := [0, 1]$ and $\mathbb{R}^n_X := \{0, 1\}^n$, respectively, and $f(\mathbf{x}; \theta)$ simplifies to

$$f(\mathbf{x}; \theta) \overset{I}{=} \prod_{k=1}^{n} f_k(x_k; \theta_k) \overset{\text{IID}}{=} \prod_{k=1}^{n} f(x_k; \theta) \overset{\text{BerIID}}{=} \theta^{\Sigma_{k=1}^n x_k}(1-\theta)^{\Sigma_{k=1}^n (1-x_k)}, \tag{8}$$

where the first equality follows by imposing Independence, the second by imposing IID, and the last by imposing all three assumptions: Bernoulli, IID. In practice, the IID assumptions can and should be tested, vis-à-vis data $\mathbf{x}_0$, to secure the reliability of the inductive inferences based on $\mathcal{M}_\theta(\mathbf{x})$; see Spanos (1999).

Philosophers of science often prefer to hide the "ugliness" of expressions like (6)(7)–(8) behind the veil of "technical details" better swept under the carpet. The truth is that dodging this ugliness seriously impairs adequate understanding of frequentist inference. In particular, poor understanding of the pivotal role played by the notion of a statistical model could easily give rise to serious confusions because $\mathcal{M}_\theta(\mathbf{x})$

(*a*) specifies the inductive premises of inference,
(*b*) delimits legitimate events in terms of an univocal sample space $\mathbb{R}^n_X$,
(*c*) assigns probabilities to all legitimate events via $f(\mathbf{x}; \theta)$, $\mathbf{x} \in \mathbb{R}^n_Z$,
(*d*) delimitates what are legitimate hypotheses and/or inferential claims,

(*e*) identifies the relevant error probabilities in terms of which the optimality and reliability of inference methods are assessed, and
(*f*) designates what constitute legitimate data $\mathbf{x}_0$ for inference purposes

Legitimate events are defined in terms of well-behaved (Borel) functions of the sample $\mathbf{X}$; formally an event is legitimate when it belongs to the σ-field generated by $\mathbf{X}$ (Billingsley 1995). Legitimate data come in the form of randomly selected newborns in New York City during 2008. For instance, the information that at St. Jude's Hospital in New York City the ratio of boys to girls during 2008 is 1.07 does not constitute legitimate data in the context of $\mathcal{M}_\theta(\mathbf{x})$ in (6) because this information cannot be regarded as comprising generic (IID) observations from the target population; purposeful selection invalidates the IID assumptions.

2.2. Hypothesis Testing. To understand what constitutes legitimate frequentist hypotheses, consider framing the above substantive question of interest (2)—whether the ratio of boys (B) to girls (G) is equal or greater than one—in the context of the simple Bernoulli model. Since $\theta \leq .5$ and $\theta > .5$ imply that the ratio is equal to or less than one and bigger than one, respectively, the archetypal Neyman-Pearson (N-P) framing takes the form of the statistical hypotheses:

$$H_0\text{: } \theta \leq \theta_0 \quad \text{versus} \quad H_1\text{: } \theta > \theta_0, \quad \text{where } \theta_0 = .5. \tag{9}$$

What renders the hypotheses in (9) legitimate is that (i) they pose questions concerning the underlying statistical data-generating mechanism, (ii) they are framed in terms of the unknown parameter θ, and (iii) they are framed in a way that partitions $\mathcal{M}_\theta(\mathbf{x})$. This can be more clearly seen by recognizing that (9) can be equivalently specified as

$$H_0\text{: } f_*(\mathbf{x}) \in \mathcal{M}_0(\mathbf{x}) \quad \text{versus} \quad H_1\text{: } f_*(\mathbf{x}) \in \mathcal{M}_1(\mathbf{x}),$$

where $f_*(\mathbf{x})$ denotes the 'true' distribution of the sample, and $\mathcal{M}_0(\mathbf{x}) = \{f(\mathbf{x}; \theta), \theta \leq .5\}$, and $\mathcal{M}_1(\mathbf{x}) = \{f(\mathbf{x}; \theta), \theta > .5\}$. That is, the question posed by (9) is whether the data $\mathbf{x}_0$ were generated by $\mathcal{M}_0(\mathbf{x})$ or its complement $\mathcal{M}_1(\mathbf{x})$ relative to $\mathcal{M}_\theta(\mathbf{x})$.

An example of an N-P test for hypothesis (9) is $T_\alpha := \{d(\mathbf{X}), C_1(\alpha)\}$, where

$$\text{test statistic:} \quad d(\mathbf{X}) = \frac{\sqrt{n}(\overline{X}_n - \theta_0)}{\sqrt{\theta_0(1-\theta_0)}}, \quad \text{where } \overline{X}_n = \frac{1}{n}\sum_{k=1}^{n} X_k,$$

$$\text{rejection region:} \quad C_1(\alpha) = \{\mathbf{x} : d(\mathbf{x}) > c_\alpha\}. \tag{10}$$

In words, the N-P rule is this: reject H_0 when $d(\mathbf{x}_0)$ exceeds c_α, otherwise accept it. The optimality of T_α is defined in terms of the probability of

the two types of error associated with accepting/rejecting H_0:

$$\begin{aligned} &\text{type I:} \quad \text{rejecting } H_0 \text{ when true, } \mathbb{P}[\mathbf{x}: d(\mathbf{x}) > c_\alpha; H_0(\theta)], \quad \text{for } \theta \leq \theta_0. \\ &\text{type II:} \quad \text{accepting } H_0 \text{ when false, } \mathbb{P}[\mathbf{x}: d(\mathbf{x}) \leq c_\alpha; H_1(\theta_1)], \quad \text{for } \theta_1 > \theta_0. \end{aligned} \tag{11}$$

It is important to emphasize that these error probabilities are associated with all the different outcomes of $\mathbf{x}$ satisfying the above inequalities ($d(\mathbf{x}) > c_\alpha$ and $d(\mathbf{x}) \leq c_\alpha$), stemming from the sampling distribution of $d(\mathbf{X})$ evaluated under different hypothetical (single) values of θ, relating to H_0 and H_1, respectively.

An optimal N-P test, known as uniformly most powerful (UMP), minimizes the probability of type II error for all $\theta_1 > \theta_0$, for a given (small) type I error probability:

$$\alpha = \max_{\theta \leq \theta_0} \mathbb{P}[\mathbf{x}: d(\mathbf{X}) > c_\alpha; \ H_0] = \mathbb{P}[\mathbf{x}: d(\mathbf{X}) > c_\alpha; \theta = \theta_0].$$

Equivalently, test T_α is said to be UMP if its power (Lehmann 1986),

$$\pi(\theta_1) = \mathbb{P}[\mathbf{x}: d(\mathbf{X}) > c_\alpha; \theta = \theta_1], \tag{12}$$

is higher than that of any other α-level test for all $\theta_1 > \theta_0$ ($\theta_1 = \theta_0 + \gamma, \gamma \geq 0$). In this sense, a UMP test provides the most effective α-significance-level probing procedure for detecting any discrepancy (γ) of interest from the null.

To evaluate the error probabilities in (11) and (12), one needs to derive the sampling distribution of $d(\mathbf{X})$ under hypothetical values of θ. In particular, the evaluation of the type I error probability α is based on the sampling distribution

$$d(\mathbf{X}) = \frac{\sqrt{n}(\overline{X}_n - \theta_0)}{\sqrt{\theta_0(1-\theta_0)}} \overset{H_0(\theta_0)}{\sim} \text{Bin}(0, 1; n), \tag{13}$$

where "$\overset{H_0(\theta_0)}{\sim}$ Bin(0, 1; n)" stands for "distributed under $H_0(\theta_0)$: $\theta = \theta_0$ as a binomial with mean 0 and variance 1, based on a sample size n." The result in (13) stems from the fact that, in the context of the simple Bernoulli model (6), the sampling distribution of $\sum_{k=1}^{n} X_k$ is Bin$[n\theta, n\theta(1-\theta); n]$, which implies that

$$\overline{X}_n = \frac{1}{n}\sum_{k=1}^{n} X_k \sim \text{Bin}\left(\theta, \frac{\theta(1-\theta)}{n}; \ n\right),$$

but, when standardized,

$$\frac{\overline{X}_n - E(\overline{X}_n)}{\sqrt{\text{Var}(\overline{X}_n)}} = d(\mathbf{X})$$

(evaluated under $\theta = \theta_0$) has mean 0 and variance 1.

Similarly, the type II error probability (and the power) are evaluated

"under all point alternatives" $H_1(\theta_1)$: $\theta = \theta_1$:

$$d(\mathbf{X}) = \frac{\sqrt{n}(\overline{X}_n - \theta_0)}{\sqrt{\theta_0(1-\theta_0)}} \overset{H_1(\theta_1)}{\sim} \text{Bin}[\delta(\theta_1), V(\theta_1); n], \quad \text{for } \theta_1 > \theta_0, \tag{14}$$

where

$$\delta(\theta_1) = \frac{\sqrt{n}(\theta_1 - \theta_0)}{\sqrt{\theta_0(1-\theta_0)}},$$

and

$$V(\theta_1) = \frac{\theta_1(1-\theta_1)}{\theta_0(1-\theta_0)}.$$

The key difference between (13) and (14) is that, under $H_1(\theta_1)$, the sampling distribution of $d(\mathbf{X})$ has a nonzero mean $\delta(\theta_1)$. It is interesting to note that, for moderately large values of n (say, $n > 20$) and values of θ close to .5, both distributions (13) and (14) can be closely approximated using the normal distribution; see Spanos (1999). Two remarks are in order.

First, it is unfortunate that most statistics books use the vertical bar (|) instead of the semicolon (;) in formulas (11)–(12) to denote the evaluation under H_0 or H_1, as it relates to (13)–(14), encouraging practitioners to misinterpret error probabilities as being conditional on H_0 or H_1; see Cohen (1994). Alas, such conditioning makes no probabilistic sense in frequentist statistics because both hypotheses are framed in terms of the unknown parameter θ, which is an unknown constant, not an r.v. As mentioned above, legitimate events that render conditioning meaningful in frequentist inference are only the ones specified in terms of the sample $\mathbf{X}$, which is a set of r.v.'s.

Second, it is important to emphasize that the optimality of test T_α depends crucially on $\overline{X}_n$, the sample mean, being a minimal sufficient statistic for θ (see Lehmann 1986). Hence, any cannibalization of $\overline{X}_n$, like rendering some of its values indistinguishable (Howson and Urbach 2005, 136–37), will destroy its sufficiency, distort both sampling distributions (13)–(14), and ruin the optimality of the resulting test by blunting its capacity to detect certain discrepancies from the null.

Although the formal apparatus of the Fisher-Neyman-Pearson (F-N-P) frequentist testing was largely in place by the late 1930s, the nature of the underlying inductive reasoning was clouded in disagreements. Fisher (1935) argued for a purely falsificationist 'inductive inference', spearheaded by his significance testing. This relied exclusively on the notion of the observed significance level, or a p-value:

$$p(\mathbf{x}_0) = \mathbb{P}[\mathbf{x}: d(\mathbf{X}) > d(\mathbf{x}_0); \theta = \theta_0], \tag{15}$$

where a small enough $p(\mathbf{x}_0)$ is interpreted as indicating discordance with H_0 (Fisher 1955). Neyman argued for 'inductive behavior' based on N-P

testing and allowed both the acceptance and rejection of H_0 (Neyman 1956). However, neither account gave a satisfactory answer to the basic question: when do data $\mathbf{x}_0$ provide evidence for (or against) a hypothesis or a claim H? Indeed, several crucial foundational problems were left unanswered, including

(*a*) the role of predata versus postdata error probabilities (Hacking 1965),
(*b*) the fallacy of acceptance—no evidence against H_0 is misinterpreted as evidence for H_0 (Mayo 1996), and
(*c*) the fallacy of rejection—evidence against H_0 is misinterpreted as evidence for a particular H_1; see Mayo and Spanos (2006).

2.3. A Severity-Based Evidential Interpretation. Error statistics (Mayo and Spanos 2010) proposes a postdata evaluation of inference to supplement the F-N-P testing in order to address the foundational problems *a*–*c*. To simplify the exposition, let us consider a particular example.

Numerical example. For the simple Bernoulli model (6), consider applying test T_α for $n = 2{,}000$ and data $\mathbf{x}_0$, yielding $\bar{x}_n = 0.503$. Fixing the significance level at $\alpha = .025$, one can use the normal approximation to the binomial distribution in (13), then derive the threshold value $c_\alpha = 1.96$; this approximation is excellent for moderate values of n and θ close to .5.

Evaluating the test statistic in (10) yields

$$d(\mathbf{x}_0) = \frac{\sqrt{2000}(.503 - .5)}{\sqrt{.5(.5)}} = .268.$$

Since $d(\mathbf{x}_0) < c_\alpha$, the test leads to accepting H_0. The lack of discordance with the null is also affirmed by the p-value

$$\mathbb{P}[d(\mathbf{x}) > .268; H_0] = .394.$$

Does this mean that data $\mathbf{x}_0$ provide evidence for H_0, or, more accurately, does $\mathbf{x}_0$ provide evidence for no substantive discrepancy from H_0? Not necessarily! To answer that question, one needs to go beyond the coarse N-P "accept/reject" rules and apply the postdata severe-testing reasoning to evaluate the discrepancy $\gamma \geq 0$, for $\theta_1 = \theta_0 + \gamma$, warranted by data $\mathbf{x}_0$ (Mayo and Spanos 2006).

A hypothesis or a claim H passes a severe test T with data $\mathbf{x}_0$ if

(S-1) $\mathbf{x}_0$ accords with H (for a suitable notion of accordance) and

(S-2) with very high probability test T would have produced a result that accords less well with H than $\mathbf{x}_0$ does, if H were false.

Severity can be viewed as a feature of a test T as it relates to a particular

TABLE 1. Severity Evaluation in the Case of "Accept H."

Relevant Claim	$\theta \leq \theta_1 = \theta_0 + \gamma$									
γ	.001	.005	.01	.015	.017	.0173	.018	.02	.025	.03
SEV(T_α; $\mathbf{x}_0$; $\theta \leq \theta_1$)	.429	.571	.734	.858	.895	.900	.910	.936	.975	.992

data $\mathbf{x}_0$ and a specific claim H. Hence, the severity function has three arguments, SEV(T, $\mathbf{x}_0$, H), denoting the severity with which H passes T with $\mathbf{x}_0$.

In the case of the above numerical example, the test $T_\alpha := \{d(\mathbf{X}), C_1(\alpha)\}$ in (10) passes H_0, and the idea is to establish the smallest discrepancy $\gamma \geq 0$ from H_0 warranted by data $\mathbf{x}_0$ by evaluating (postdata) the claim $\theta \leq \theta_1$ for $\theta_1 = \theta_0 + \gamma$. Condition S-1 of severity is satisfied since $\mathbf{x}_0$ accords with H_0 because $d(\mathbf{x}_0) < c_\alpha$, but, in addition, severity requires one to evaluate the probability of "all outcomes $\mathbf{x}$ for which test T_α accords less well with H_0 than $\mathbf{x}_0$ does"; that is, [$\mathbf{x}$: $d(\mathbf{x}) > d(\mathbf{x}_0)$], under the hypothetical scenario that "$\theta \leq \theta_1$ is false" or equivalently "$\theta > \theta_1$ is true":

$$\text{SEV}(T_\alpha; \mathbf{x}_0; \theta \leq \theta_1) = \mathbb{P}[\mathbf{x}: d(\mathbf{X}) > d(\mathbf{x}_0); \theta > \theta_1], \quad \text{for } \gamma \geq 0. \quad (16)$$

The evaluation of (16) relies on the sampling distribution (14) and for different discrepancies ($\gamma \geq 0$) yields the results in table 1. Note that SEV(T_α; $\mathbf{x}_0$; $\theta \leq \theta_1$) is evaluated at $\theta = \theta_1$ because the probability increases with γ. Assuming a severity threshold of, say, .90 is high enough, the above results indicate that the 'smallest' discrepancy warranted by $\mathbf{x}_0$ is

$$\gamma \geq .0173, \quad \text{since SEV}(T_\alpha; \mathbf{x}_0; \theta \leq .5173) = .9.$$

That is, data $\mathbf{x}_0$, in conjunction with test T_α, provide evidence (with severity at least .9) for the presence of a discrepancy as large as $\gamma \geq .0173$. Is this discrepancy substantively insignificant?

In general, to answer this question one needs to appeal to substantive subject matter information to assess the warranted discrepancy on substantive grounds. In human biology it is commonly accepted that the sex ratio at birth is approximately 105 boys to 100 girls; see Hardy (2002). Translating this ratio in terms of θ yields $\theta^* = 105/205 = .512$, which suggests that the warranted discrepancy $\gamma \geq .0173$ is, indeed, substantively significant, since this outputs $\theta \geq .5173$, which exceeds θ^*—the substantive value of the human sex ratio at birth.

This is an example where the postdata severity evaluation addresses the fallacy of acceptance. The severity assessment indicates that the statistically insignificant result $d(\mathbf{x}_0) = .268$, at $\alpha = .025$, actually provides evidence against H_0: $\theta \leq .5$ and for $\theta \geq .5173$ with severity at least .90.

Mayo and Spanos (2006) show how the same conditions S-1–S-2 can be applied to the case of 'reject H_0' to derive the corresponding severity

evaluation function

$$\text{SEV}(T_\alpha; \mathbf{x}_0; \theta > \theta_1) = \mathbb{P}[\mathbf{x}: d(\mathbf{X}) \leq d(\mathbf{x}_0); \theta \leq \theta_1], \quad \text{for } \gamma \geq 0. \quad (17)$$

The severe testing reasoning underlying (16) and (17) can be used to circumvent the fallacy of acceptance and rejection, respectively, by establishing the smallest (largest) discrepancy $\gamma \geq 0$ from H_0 warranted by data $\mathbf{x}_0$, associated with the N-P decision to accept (reject) H_0.

A direct comparison between, on the one hand, the type I and II error probabilities (11) and the power function (12) and the severity functions (16)–(17), on the other hand, reveals that they are all evaluated in terms of the sampling distributions (13)–(14) with one crucial difference: the former are predata (the threshold c_α is derived on the basis of a prespecified α), but the latter are postdata (the threshold $d(\mathbf{x}_0)$ becomes available only after $\mathbf{x}_0$ is observed). Severity shares this postdata feature with the p-value, which is a postdata error probability, but unlike the p-value whose probability stems from (13), the probabilities associated with severity are based on the sampling distribution (14), resembling, in this respect, the type II error probability and power. Indeed, severity can be used to explain why the use of the p-value is also susceptible to the same fallacies as the accept/reject decision rules (Mayo and Spanos 2006).

One can go a step further and make a case that the error-statistical framework anchored on the postdata severity component provides a harmonious blending of the Fisherian and Neyman-Pearsonian perspectives by weaving a coherent frequentist inductive reasoning anchored firmly on error probabilities. Predata, error probabilities are used to appraise the generic capacity (for any $\mathbf{x} \in \mathbb{R}_X^n$) of different testing procedures. Postdata, however, the particular realization $\mathbf{x}_0$ can be used to transcribe the generic capacity, as it relates to the test outcome $d(\mathbf{x}_0)$, to customize an evidential interpretation—beyond the crude true/false dichotomy—that pertains to whether data $\mathbf{x}_0$ provide evidence for or against relevant claims concerning specific discrepancies (γ) from H_0. Indeed, the error-statistical perspective provides a unifying inductive reasoning for frequentist testing that addresses all three foundational problems *a–c* mentioned above; see Mayo and Spanos (2006) for further discussion.

3. The Base-Rate Fallacy Revisited.

3.1. Summarizing the Fallacy. Psillos (2007, 17–18) offers the following succinct summary:

> Base-rate fallacy.—Best introduced by the Harvard Medical School test. A test for the presence of a disease has two outcomes, 'positive' and 'negative' (call them + and −). Let a subject (Joan) take the test. Let H be the hypothesis that Joan has the disease and $-H$ the

> hypothesis that Joan doesn't have the disease. The test is highly reliable: it has zero false negative rate. That is, the **likelihood** that the subject tested negative given that she has the disease is zero (i.e., $prob(-|H) = 0$). The test has a small *false positive rate*: the likelihood that Joan is tested positive though she doesn't have the disease is, say, 5 per cent ($prob(+| -H) = 0.05$). Joan is tested positive. What is the **probability** that Joan has the disease given that she tested positive? When this problem was posed to experimental subjects, they tended to answer that the probability that Joan has the disease given that she tested positive was very high—very close to 95 per cent. However, given only information about the likelihoods $prob(+|H)$ and $prob(+| -H)$, the question above—what is the posterior probability $prob(H|+)$?—is indeterminate. There is some crucial information missing: the incidence rate (base-rate) of the disease in the population. If this incidence rate is very low, for example, only 1 person in 1,000 has the disease, it is very unlikely that Joan has the disease even though she tested positive: $prob(H|+)$ would be very small. For $prob(H|+)$ to be high, it must be the case that the prior probability that Joan has the disease (i.e. $prob(H)$) is not too small. The lesson that many have drawn from cases such as this is that it is a fallacy to ignore the base rates because it yields wrong results in probabilistic reasoning.

It is clear from the above summary that any procedure relying on the posterior probability $\Pr(H|+)$ will give rise to fallacious inferences when base rates are ignored because, as can be seen in (1), going from $\Pr(+|H)$ to $\Pr(H|+)$ depends crucially on the ratio of the prior probabilities $\Pr(+)$ and $\Pr(H)$.

The question is whether the frequentist approach to testing is also vulnerable to the base-rate fallacy. Howson (2000, 54) contends that this fallacy seriously undermines frequentist testing, in general, and the severity assessment, in particular:

> The central methodological claims of a recent book on methodology (Mayo 1996) are based on committing it. In this book a version of the argument very similar to Fisher's is proposed. As in the tea-tasting, there is a notion of outcomes agreeing better or worse with some hypothesis H. . . . Similarly, spelling out explicitly what is implicit in Fisher's discussion, if E "fits" H, while there is a very small chance that the test procedure "would yield so good a fit if H is false", then, "E should be taken as *good grounds for* H to the extent that H has passed a severe test with E" (Mayo 1996:177; my italics). In the Harvard Medical School test we have Mayo's formal criteria for H "passing a severe test with E" satisfied.

TABLE 2. JOINT DISTRIBUTION OF (X, Y).

$X\backslash Y$	0	1	$f(x)$
0	$p_{00} = .99997998$	$p_{01} = .00001992$	$(1-\theta_1) = .9999999$
1	$p_{10} = .00000002$	$p_{11} = .00000008$	$\theta_1 = .0000001$
$f(y)$	$(1-\theta_2) = .99998$	$\theta_2 = .00002$	1

A similar view is expressed in Howson (1997), Achinstein (2001, 2010), and Howson and Urbach (2005).

3.2. Unraveling the Base-Rate Fallacy Argument. In order to simplify the untangling of the various confusions vitiating the base-rate fallacy argument, consider a numerical example due to Achinstein (2010), specified in terms of the following probabilities:

$$\Pr(+|H) = .8, \quad \Pr(+|-H) = .00002, \quad \Pr(H) = .0000001. \tag{18}$$

3.2.1. *What Is the Underlying Statistical Model?* Since frequentist statistical inference is model based, the first thing one needs to uncover is the implicit statistical model that can then be used to delineate what constitute legitimate events, hypotheses, data, test statistics, error probabilities, and so forth.

Since there are only two outcomes in the Harvard Medical School test example, it is easy to recognize that the relevant random variables are Bernoulli distributed:

> X denotes having the disease ($X = 1$) or not ($X = 0$), and Y denotes the medical test result, positive ($Y = 1$) or negative ($Y = 0$).

Thus, the probabilities in (18) can be written more transparently in terms of the joint probabilities $\{p_{ij} := \mathbb{P}(X = i, Y = j), i, j = 0,1\}$:

$$\mathbb{P}(Y = 1|X = 1) = \frac{p_{11}}{p_{10} + p_{11}} = .8, \quad \mathbb{P}(Y = 1|X = 0) = \frac{p_{01}}{p_{00} + p_{01}} = .00002,$$

$$\mathbb{P}(X = 1) = p_{10} + p_{11} = .0000001, \quad p_{00} + p_{01} + p_{10} + p_{11} = 1. \tag{19}$$

Solving (19) for $\{p_{ij} i, j = 0, 1\}$ gives rise to the joint distribution in table 2.

This suggests that the underlying statistical model—assumed to describe the incidence of a disease in a particular population as it relates to the result of a medical test for that disease—is a simple (bivariate) Bernoulli model:

$$\mathcal{M}_\theta(\mathbf{z})\text{: } \mathbf{Z}_k \sim \text{BerIID}[E(\mathbf{Z}_k), \text{Cov}(\mathbf{Z}_k)], \quad k = 1, 2, \ldots, n, \ldots, \tag{20}$$

where

$$\mathbf{Z}_k := \begin{pmatrix} X_k \\ Y_k \end{pmatrix}, \quad E(\mathbf{Z}_k) = \begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix},$$

$$\mathrm{Cov}(\mathbf{Z}_k) = \begin{bmatrix} \theta_1(1-\theta_1) & \theta_3 - \theta_1\theta_2 \\ \theta_3 - \theta_1\theta_2 & \theta_2(1-\theta_2) \end{bmatrix}.$$

In direct analogy to (6), the distribution of the sample

$$\mathbf{Z} := ([X_1, Y_1], [X_2, Y_2], \ldots, [X_n, Y_n])$$

simplifies to

$$f(\mathbf{z}; \theta) \overset{1}{=} \prod_{k=1}^{n} f_k(\mathbf{z}_k; \theta) \overset{\text{IID}}{=} \prod_{k=1}^{n} f(x_k, y_k; \theta),$$

where

$$f(x_k, y_k; \theta) = p_{11}^{x_k y_k} p_{00}^{(1-x_k)(1-y_k)} p_{01}^{(1-x_k)y_k} p_{10}^{(1-y_k)x_k}, \quad x = 0, 1, \quad y = 0, 1; \tag{21}$$

$$\theta := (\theta_1, \theta_2, \theta_3) \in [0, 1]^3,$$

$$\theta_1 = p_{10} + p_{11}, \quad \theta_2 = p_{01} + p_{11}, \quad \theta_3 = p_{11} \tag{22}$$

denote the unknown parameters. The additional "ugliness" notwithstanding, $\mathcal{M}_\theta(\mathbf{z})$ in (20) is identical to the simple Bernoulli in (6), apart from the underlying process $\{\mathbf{Z}_k, k \in \mathbb{N}\}$ being bivariate!

The first thing one notices about $\mathcal{M}_\theta(\mathbf{z})$ in (20) and table 2 is that the latter is an instantiation of the former that involves no unknown parameters. If frequentist inductive inference is all about learning from data about the data-generating mechanism, the question that naturally arises is, what could learning based on table 2 possibly mean?

3.2.2. *Frequentist Hypotheses versus Events.* As argued in section 2.2, the statistical hypotheses in frequentist testing are always framed in terms of the unknown parameters θ, and they invariably pertain to the stochastic mechanism—assumed to be described by $\mathcal{M}_\theta(\mathbf{z})$—that gave rise to data $\mathbf{z}_0 := [(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)]$. An example of such N-P hypotheses in the context of the bivariate Bernoulli model might be

$$H_0: \phi \leq \phi_0 \quad \text{versus} \quad H_1: \phi > \phi_0, \quad \text{for } \phi_0 = .05, \tag{23}$$

where $\phi = \mathbb{P}(X = 1|Y = 1) = (\theta_3/\theta_2)$ (unknown) denotes the proportion of the population that has the disease, given that they tested positive. The N-P hypotheses in (23) pose the sharp question whether or not $\phi \leq .05$, with a view to learn from data by narrowing the original model

$\mathcal{M}_\theta(\mathbf{z})$ down to a subset:

$$\mathcal{M}_0(\mathbf{z}) = \{f(\mathbf{z};\, \phi), \phi \leq \phi_0\} \quad \text{or} \quad \mathcal{M}_1(\mathbf{z}) = \{f(\mathbf{z};\, \phi), \phi > \phi_0\}.$$

A frequentist test for properly defined hypotheses such as (23) will involve constructing a test similar to T_α in (10), say,

$$\tau(\mathbf{Z}) = \frac{\hat{\phi} - \phi_0}{\sqrt{\text{Var}(\hat{\phi})}}, \quad C_1(\alpha) = \{\mathbf{z}\colon \tau(\mathbf{z}) > c_\alpha\}, \tag{24}$$

for an appropriate estimator $\hat{\phi}$ of ϕ. Using the sampling distribution of $\tau(\mathbf{Z})$ under both the null and alternative hypotheses, one can evaluate the predata and postdata error probabilities as shown in sections 2.2–2.3. One feature of the hypotheses in (23) worth bringing out is that because $\phi_0 = .05$ is rather small, one can guesstimate that the sample size n needed to give adequate power to the test to detect small discrepancies (third decimal) from ϕ_0 is likely to be huge.

How does this frequentist set up relate to the "hypothesis"

$$h\text{-Joan has the disease,} \tag{25}$$

around which the base-rate fallacy example revolves? The assignment of probability to h is a giveaway that it is not a legitimate frequentist hypothesis, since in that context hypotheses are framed in terms of θ, which is assumed to be an unknown constant, not a random variable. The difficulty, however, is to disentangle the connections between h, the data $\mathbf{z}_0$, and $\mathcal{M}_\theta(\mathbf{z})$.

In contrast to (23), h in (25) assumes that ϕ is known ($\phi = .004$; table 2) and poses the question whether Joan—a particular member of the target population—has the disease or not. Due to the ambiguity as to the status of h vis-à-vis $\mathcal{M}_\theta(\mathbf{z})$, one needs to distinguish between two different cases.

Case 1.—Joan is randomly selected from the target population. In this case, h concerns the event (say $X_{13} = 1$) and represents an element of the random sample $Z\colon [X_{13}, Y_{13}]$. In view of the fact that h does not pertain to the statistical data-generating mechanism, it is not a legitimate frequentist hypothesis; see Mayo (1997a). Indeed, in this case $[x_{13}, y_{13}] = [1, 1]$ constitutes a single observation from a size n sample $\mathbf{Z}$.

Case 2.—Joan is not randomly selected from the target population; for example, she requested the test. In this case the single observation $[x_{13}, y_{13}] = [1, 1]$ is no longer legitimate for $\mathcal{M}_\theta(\mathbf{z})$. Indeed, the event $h(x_{13} = 1)$ lies outside the intended scope of the statistical model $\mathcal{M}_\theta(\mathbf{z})$, which aims to provide an idealized description of the disease's incidence in the target population, and not the affliction of a particular individual; for the latter, one needs to specify a different statistical model (see Spanos 2009). The variable h is illegitimate as an event in the context of $\mathcal{M}_\theta(\mathbf{z})$ because $[x_{13}, y_{13}] = [1, 1]$ does not constitute a "typical" reali-

zation of $[X_{13}, Y_{13}]$; purposeful selection invalidates the IID assumptions underlying $\mathcal{M}_\theta(\mathbf{z})$. Any attempt to interpret $[x_{13}, y_{13}] = [1, 1]$ as an instantiation of the generic event $(X = 1, Y = 1)$ will introduce statistical misspecifications that often lead to unreliable inferences.

Statistical adequacy—the validity (vis-à-vis data $\mathbf{z}_0$) of the IID assumptions underlying $\mathcal{M}_\theta(\mathbf{z})$—ensures that the actual error probabilities are approximately equal to the nominal ones, rendering the reliability of inference ascertainable; see Mayo and Spanos (2004). Applying a .025 significance level test when its actual type I error probability is closer to .99 will lead the inference astray!

In summary, *h*-Joan has the disease does not constitute a legitimate frequentist hypothesis because, at best, it pertains to an event (not to the data-generating mechanism) and, at worst, it lies outside the intended scope of $\mathcal{M}_\theta(\mathbf{z})$.

3.2.3. *Error Probabilities versus Conditional Probabilities of Events.* Focusing on the best-case scenario for the proponents of the base-rate argument, it is clear that when Joan is randomly selected, the event $[x_{13}, y_{13}] = [1, 1]$ constitutes a single observation from $\mathbf{Z}$. It is well known, however, that with $n = 1$ no consistent (minimally reliable) estimator or test concerning θ is possible; see Cox and Hinkley (1974).

In light of this, how do the proponents of the base-rate argument make their case that frequentist testing is vulnerable to the fallacy? Their argument is that the hypothesis $h\colon (X = 1)$ has "passed a severe test" on the basis of the following two conditional probabilities (table 2):

$$\mathbb{P}(Y = 1|X = 1) = .8, \quad \mathbb{P}(Y = 1|X = 0) = .00002, \tag{26}$$

known as the false positive and the false negative probabilities, respectively. As the base-rate argument goes, since $\mathbb{P}(X = 1) = .0000001$ (the prior probability of h) is low, the conditional (posterior or epistemic) probability,

$$\mathbb{P}(X = 1|Y = 1) = \frac{.00000008}{.00002} = 0.004, \tag{27}$$

is also low, and, thus, on the basis of (27), Joan's positive result gives very little reason to believe $h := (X = 1)$, despite h's passing a severe test; see Achinstein (2010, 182).

Mayo (1997a, 1997b, 2005) has called repeatedly into question the basic claim that "h has passed a severe test" on the basis of the probabilities in (26)–(27) on several grounds. Howson (2000, 54), however, maintains, "Mayo discusses the example in Mayo 1997, but I cannot see that she mitigates in any way its force"; see also Howson and Urbach (2005, 25).

Focusing on the most telling of such grounds, the conditional proba-

bilities (26)–(27) have nothing to do with any proper predata or postdata error probabilities. In particular, the severity functions in (16)–(17) have three arguments (a test T, data $\mathbf{z}_0$, and a frequentist hypothesis or a claim H), and none of them is present in the base-rate argument. This argument replaces frequentist hypotheses with events; the test and its error probabilities with conditional probabilities among events; and the data, at best, amount to a single observation ($n = 1$).

One can go further and call into question the base-rate argument on other grounds, even when h is the legitimate event ($X = 1$) in the context of $\mathcal{M}_\theta(\mathbf{z})$. First, the claim that $\mathbb{P}(X = 1)$ and $\mathbb{P}(X = 1|Y = 1)$ represent the prior and posterior probabilities of h is unwarranted on Bayesian statistical grounds. This is because Bayesian inference is deemed legitimate when grounded on a posterior distribution:

$$\pi(\theta|\mathbf{z}_0) \propto \pi(\theta) \times f(\mathbf{z}_0|\theta), \theta := (\theta_1, \theta_2, \theta_3) \in [0, 1]^3,$$

where θ is defined in (22), $\pi(\theta)$ denotes the prior distribution of θ, and $L(\mathbf{z}_0|\theta)$ denotes the likelihood function. The latter is defined as being proportional to the distribution of the sample:

$$f(\mathbf{z}; \theta) = \theta_3^{\sum_{k=1}^{n} x_k y_k}[1 + \theta_3 - (\theta_1 + \theta_2)]^{\sum_{k=1}^{n}(1-x_k)(1-y_k)}$$
$$\times [\theta_2 - \theta_3]^{\sum_{k=1}^{n}(1-x_k)y_k}[\theta_1 - \theta_3]^{\sum_{k=1}^{n}(1-y_k)x_k},$$

evaluated at $\mathbf{z}_0$ and viewed as a function of θ. However, any attempt to relate the assignments $\mathbb{P}(X = 1)$ and $\mathbb{P}(X = 1|Y = 1)$ to $\pi(\theta)$, $\pi(\theta|\mathbf{z}_0)$, or/and $L(\mathbf{z}_0|\theta)$ is a hopeless task because all these crucial components of Bayesian inference are missing. Indeed, $\mathbb{P}(X = 1|Y = 1)$ in (27) is nothing more than a deductive calculation within the context of a known statistical model (table 2). Second, any attempt to associate $\mathbb{P}(X = 1)$ and $\mathbb{P}(X = 1|Y = 1)$ with subgroups of the original target population in a desperate attempt to provide some basis for any form of make-believe inductive inference is misguided; see Spanos (2009). Finally, one can test legitimate hypotheses about the unknown parameters $\theta_1 = \mathbb{P}(X = 1)$ and $\mathbb{P}(X = 1|Y = 1) = \phi$ in the context of $\mathcal{M}_\theta(\mathbf{z})$ in (20), but that requires a proper data set $\mathbf{z}_0(n > 1)$ associated with randomly selected individuals from the target population.

3.2.4. *What about the False Positive and False Negative Probabilities?* The medical test is conflated with a proper frequentist test using a beguiling analogical argument concerning the former's false positive and false negative conditional probabilities as being equivalent to the type I and type II error probabilities (Howson 2000). The analogy misrepresents frequentist testing because there is no such a thing as a generic type I and II error probability for a frequentist test.

A glance at the sampling distributions of $d(\mathbf{X})$ in (13) and (14) reveals

that all error probabilities (see [11]–[12] and [16]–[17]) depend crucially on the sample size n. For example, for $n = 10$ and $\alpha = .025$, a power value of .8 at some discrepancy, say $\gamma = .3$, from the null might be considered adequate, but for $n = 100{,}000$, it is not! Moreover, any attempt to identify the false positive/negative with asymptotic error probabilities, as being more generic, will not work because the power of any half-decent (consistent) N-P test goes to one as the sample size n goes to infinity; see Lehmann (1986).

In addition, a medical test aims to prognosticate the occurrence of an event (Joan has the disease), but a frequentist test aims to guide the quest for the data-generating mechanism that gave rise to the data. Although very important, distinguishing between the statistical and substantive data-generating mechanism is beyond the scope of this article; see Spanos (2007). Hence, the use of conditional probabilities associated with particular events by the base-rate argument is a far cry from proper error probabilities pertaining to inference procedures. Why?

The type I and II error probabilities and the power of the test in (24) would take an almost identical form to (11) and (12). What is crucially important for this discussion is that these error probabilities are always associated with the test procedure (not the hypotheses); they are never conditional on a hypothesis—they are evaluated under different hypothetical values of θ to yield (13)–(14)—and they invariably involve an infinity of events (tail areas). For example, the type I error probability in (11) involves all outcomes $\mathbf{x} \in \mathbb{R}_X^n$ such that $d(\mathbf{x}) > c_\alpha$, not to mention that they are meaningless for $n = 1$.

3.3. The Ampliative Dimension of Frequentist Induction. Summarizing the discussion so far, the base-rate argument relying on medical-test type examples is only superficially related to frequentist testing, 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.

At best, the data amount to a single observation ($n = 1$); the hypothesis of interest is illegitimate; and the test statistic, its sampling distribution, and error probabilities are replaced by conditional probabilities among events. As a result, the ampliative dimension of frequentist induction, in the sense of "learning from data" about the underlying statistical data-generating mechanism, is absent. When viewed in the context of frequentist induction, $\mathbb{P}(X = 1|Y = 1)$ amounts to a deductive calculation of conditional probabilities associated with particular events within the context of a known model (table 2). This is the reason why from that perspective the base-rate fallacy arguments appear to rely on $n = 1$ and

there are no unknown parameters in table 2; they are totally at odds with the implicit bivariate Bernoulli model $\mathcal{M}_\theta(\mathbf{z})$ in (20).

Mayo (2010) elaborates further on how misleading the claim is that h passes a severe test with data $[x_{13}, y_{13}] = [1, 1]$ on the basis of the conditional probabilities in (19). For our purposes it suffices to reiterate that examples like the Harvard Medical School test demonstrate inadequate understanding of frequentist testing, insofar as (i) frequentist testing takes place within the boundaries of a clearly defined statistical model $\mathcal{M}_\theta(\mathbf{z})$ chosen so as to account for the regularities in data $\mathbf{z}_0$ and not on the basis of a cluster of probabilities like (19) attached to particular events; (ii) frequentist hypotheses are framed in terms of the unknown parameter(s) θ (assumed to be constant) and always pertain to the statistical data-generating mechanism; and (iii) the ampliative dimension of frequentist testing revolves around the relevant error probabilities, which are (*a*) never conditional, (*b*) always assigned to inference procedures (never to hypotheses), and (*c*) invariably depend on the sample size $n > 1$.

4. Summary and Conclusion. The above discussion has demonstrated that the base-rate fallacy argument is riddled with obfuscations and false analogies that stem from inadequate understanding of frequentist testing anchored on the notion of a statistical model $\mathcal{M}_\theta(\mathbf{z})$.

Despite the apparent similarities between the Harvard Medical School test example and a frequentist test, it was pointed out that the two are fundamentally different. The medical test and its false positive and false negative probabilities have only superficial resemblance to a proper frequentist test and its type I and II error probabilities, because the latter are crucially dependent on the sample size n but the former are built-in.

In particular, learning from data about the underlying data-generating mechanism is absent from the base-rate fallacy example because the relevant error probabilities—around which the ampliative dimension of frequentist inference revolves—are replaced by conditional probabilities among events. These probabilities represent simple deductive calculations within the context of a known data-generating mechanism.

REFERENCES

Achinstein, Peter. 2001. *The Book of Evidence*. Oxford: Oxford University Press.

———. 2010. "Mill's Sins or Mayo's Errors?" In *Error and Inference: Recent Exchanges on Experimental Reasoning, Reliability, and the Objectivity and Rationality of Science*, ed. D. G. Mayo and Aris Spanos, 170–88. Cambridge: Cambridge University Press.

Billingsley, Patrick. 1995. *Probability and Measure*. 3rd ed. New York: Wiley.

Cohen, Jacob. 1994. "The Earth Is Round ($p < .05$)." *American Psychologist* 49:997–1003.

Cox, D. R., and D. V. Hinkley. 1974. *Theoretical Statistics*. London: Chapman & Hall.

Fisher, R. A. 1922. "On the Mathematical Foundations of Theoretical Statistics." *Philosophical Transactions of the Royal Society* A 222:309–68.

———. 1925. "Theory of Statistical Estimation." *Proceedings of the Cambridge Philosophical Society* 22:700–725.
———. 1934. "Two New Properties of Maximum Likelihood." *Proceedings of the Royal Statistical Society* A 144:285–307.
———. 1935. *The Design of Experiments*. Edinburgh: Oliver & Boyd.
———. 1955. "Statistical Methods and Scientific Induction." *Journal of the Royal Statistical Society* B 17:69–78.
Hacking, Ian. 1965. *Logic of Statistical Inference*. Cambridge: Cambridge University Press.
Hardy, I. C. W., ed. 2002. *Sex Ratios: Concepts and Research Methods*. Cambridge: Cambridge University Press.
Howson, Colin. 1997. "A Logic of Induction." *Philosophy of Science* 64:268–90.
———. 2000. *Hume's Problem*. Oxford: Oxford University Press.
Howson, Colin, and Peter Urbach. 2005. *Scientific Reasoning: The Bayesian Approach*. 3rd ed. Chicago: Open Court.
Krämer, Walter, and Gerd Gigerenzer. 2005. "How to Confuse with Statistics; or, The Use and Misuse of Conditional Probabilities." *Statistical Science* 20:223–30.
Lehmann, E. L. 1986. *Testing Statistical Hypotheses*. 2nd ed. New York: Wiley.
Mayo, D. G. 1996. *Error and the Growth of Experimental Knowledge*. Chicago: University of Chicago Press.
———. 1997a. "Error Statistics and Learning from Error." *Philosophy of Science* 64:195–212.
———. 1997b. "Response to Howson and Laudan." *Philosophy of Science* 64:323–33.
———. 2005. "Evidence as Passing Severe Tests: Highly Probable versus Highly Probed Hypotheses." In *Scientific Evidence: Philosophical Theories and Applications*, ed. Peter Achinstein, 95–127. Baltimore: Johns Hopkins University Press.
———. 2010. "Sins of the Epistemic Probabilist: Exchanges with Achinstein." In *Error and Inference: Recent Exchanges on Experimental Reasoning, Reliability, and the Objectivity and Rationality of Science*, ed. D. G. Mayo and Aris Spanos, 189–201. Cambridge: Cambridge University Press.
Mayo, D. G., and Aris Spanos. 2004. "Methodology in Practice: Statistical Misspecification Testing." *Philosophy of Science* 71:1007–25.
———. 2006. "Severe Testing as a Basic Concept in a Neyman-Pearson Philosophy of Induction." *British Journal for the Philosophy of Science* 57:323–57.
———. 2010. "Error Statistics." In *Philosophy of Statistics*, vol. 7, *The Handbook of Philosophy of Science*, ed. Dov Gabbay, Paul Thagard, and John Woods, 151–96. North-Holland: Elsevier.
Neyman, J. 1956. "Note on an Article by Sir Ronald Fisher." *Journal of the Royal Statistical Society* B 18:288–94.
Psillos, Stathis. 2007. *Philosophy of Science A–Z*. Edinburgh: Edinburgh University Press.
Spanos, Aris. 1999. *Probability Theory and Statistical Inference: Econometric Modeling with Observational Data*. Cambridge: Cambridge University Press.
———. 2007. "Curve-Fitting, the Reliability of Inductive Inference, and the Error-Statistical Approach." *Philosophy of Science* 74:1046–66.
———. 2009. "Model-Based Inference and the Frequentist Interpretation of Probability." Working paper, Department of Economics, Virginia Tech.
Tversky, Amos, and Daniel Kahneman. 1982. "Evidential Impact of Base Rates." In *Judgment under Uncertainty: Heuristics and Biases*, ed. Daniel Kahneman, Paul Slovic, and Amos Tversky, 153–60. Cambridge: Cambridge University Press.