Œconomia
History, Methodology, Philosophy

9-4 | 2019
Varia / Public Debt

What Foundations for Statistical Modeling and Inference?

About Ian Hacking, Logic of Statistical Inference, and Deborah G. Mayo, Statistical Inference as Severe Testing: How to Get Beyond the Statistics Wars

Aris Spanos

Electronic version
URL: http://journals.openedition.org/oeconomia/7521
DOI: 10.4000/oeconomia.7521
ISSN: 2269-8450

Publisher
Association Œconomia

Printed version
Date of publication: 1 December 2019
Number of pages: 833-860
ISSN: 2113-5207

Electronic reference
Aris Spanos, « What Foundations for Statistical Modeling and Inference? », *Œconomia* [Online], 9-4 | 2019, Online since 01 December 2019, connection on 28 December 2020. URL : http://journals.openedition.org/oeconomia/7521 ; DOI : https://doi.org/10.4000/oeconomia.7521

Essais critiques | *Review essays*

What Foundations for Statistical Modeling and Inference?

Ian Hacking, *Logic of Statistical Inference*
Cambridge: Cambridge University Press, 1965, Reprinted in 2016, 226 pages, ISBN 978-131650814-5

Deborah G. Mayo *Statistical Inference as Severe Testing: How to Get Beyond the Statistics Wars*
Cambridge: Cambridge University Press, 2018, 500 pages, ISBN 978-110766464-7

Aris Spanos*

The primary aim of this article is to review the above books in a comparative way from the standpoint of my perspective on empirical modeling and inference.

These two books pertaining to the nature and justification of statistical inference were written by two philosophers of science more than 50 years apart. What they have in common is the critical eye of a philosopher of science scrutinizing the statistical reasoning employed by statisticians and practitioners in different fields, and endeavoring to provide answers to certain

*Department of Economics, Virginia Tech. aris@vt.edu.

foundational issues that emerge from such a philosophical investigation. They differ in so far as Hacking does not venture too far away from his philosophical ground and focuses primarily on the early pioneers of statistics, staying above the fray of the conflicts and disputes among statisticians in the 1960s. In contrast, Mayo dares to get into the current disputes pertaining to the replication crisis and the trustworthiness of empirical evidence with a view to delineate the issues and make progress by using a philosopher's perspective to redress the balance between heat and light.

1 Hacking (1965). Logic of Statistical Inference

Hacking's (1965) perspective consists in scrutinizing statistical inference using *formal logic* based on *axiomatic foundations* supplemented with probabilities. In chapter 1, Hacking discusses the notion of a 'long run frequency', which he considers the quintessential concept of frequentist inference. In summarizing chapter 1 Hacking argues: "Statistical inference is chiefly concerned with a physical property, which may be indicated by the name of *long run frequency*. The property has never been well defined. Because there are some reasons for denying that it is a physical property at all, its definition is one of the hardest of conceptual problems about statistical inference—and it is taken as the central problem in this book."(v)

After tracing the development of the notion of the 'long run frequency' to Cournot, Ellis and Venn, he adopts the notion as framed by von Mises in the form of a *collective*: an infinite sequence $\{x_k\}_{k=1}^{\infty}$ of Bernoulli *outcomes* of 0's and 1's, representing the occurrence of event $A{=}(X_k{=}1),\ P(A){=}p_A$, where X_k denotes the random variable and x_k its observed value, that satisfies two conditions:

$$\text{Convergence}: \lim_{n\to\infty}[\frac{1}{n}\sum\nolimits_{k=1}^{n} x_k]{=}p_A$$

$$\text{Randomness}: \lim_{n\to\infty}[\frac{1}{n}\sum\nolimits_{k=1}^{n} \varphi(x_k)]{=}p_A$$

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

As argued below his adopting of von Mises notion of the long run frequency undermines Hacking's perspective on the nature and justification of frequentist inference.

In his attempt to avoid the notion of 'sampling from populations' as the hypothetical engine that generates long run frequencies, Hacking introduces in chapter 2 his notion of a *chance set-up*: "A chance set-up is a device or part of the world on which might be conducted one or more trials, experiments, or observations; each trial must have a unique result which is a member of a class of possible results."(12)

His supporting argument is that the notion of a chance set-up also avoids the problems associated with Fisher's (1922) notion of a 'hypothetical infinite population: "the object of statistical methods is the reduction of data... This object is accomplished by constructing a hypothetical infinite population, of which the actual data are regarded as constituting a sample. ... The postulate of randomness thus resolves itself into the question, 'Of what population is this a random sample?'"(313).

Explaining the potential vicious circle Hacking (1965, 14) argues: "Fisher's remark recalls that when frequency is taken as a characteristic of populations investigated by sampling, the sampling is generally supposed to be random. But randomness in this context can only be explained by frequency in the long run. So there is some danger of a silly circle. The danger is skirted by a study of chance set-ups."

This was an insightful change of concepts by Hacking (1965), but it turns out that his chance set-up did not go far enough to formalize the notion of a 'chance process' giving rise to data $\mathbf{x}_0\mathbf{:}=(x_1, ..., x_n)$ mentioned below: "Chance set-ups at least seem a natural and general introduction to the study of frequency. Especially since they lend themselves naturally to the idea of a chance process; to describe a chance process in terms of sampling from populations, you probably need an hypothetical infinite array of hypothetical populations. Chimaeras are bad enough, but a chimaera riding on the back of a unicorn cannot be tolerated." (24)

In chapter 3, Hacking argues that Kolmogorov's axiomatic approach to probability, grounded on the abstract probability space $(S, \mathfrak{F}, \mathbb{P}(.))$ with $\mathbb{P}(.)$: $\mathfrak{F} \rightarrow [0,1]$ satisfying the three well-known axioms, is inadequate for the purpose of establishing support for statistical hypotheses by data $\mathbf{x}_0$. With that in mind, he adopts the axioms in Koopman's (1940) 'logic of intuitive probability' and appends to those axioms the *Law of Likelihood*: Data $\mathbf{x}_0$ support hypothesis H_0 over hypothesis H_1 if and only if $L(H_0; \mathbf{x}_0) > L(H_1; \mathbf{x}_0)$, where $L(.;.)$ denotes the likelihood function.

For Hacking the law of likelihood provides a logic of *comparative support* stemming from evidence e for pairs of propositions h and g that belong to a Boolean algebra of propositions; a set that is closed with respect to negation and disjunction (chapter 5). In chapter 4 he considers the question of relating his notion of *comparative support* to the frequentist long run interpretation of probability ('chance' in his terminology), but after discussing the various problems of this interpretation concludes that no such definite link can be established. Hence, he argues that his law of likelihood "will serve as a foundation not only for guessing by frequency, but also for what is more commonly called statistical inference." (47).

Despite his spirited defence of the Law of Likelihood (LL) throughout this book, soon afterwards Hacking backed away from it in no uncertain terms. In his review of Edward's (1972) book entitled "Likelihood", conveys serious doubts about the LL: "... Birnbaum has given it up and I have become pretty dubious." (137)

In Hacking (1980) he rejects outright his perspective on 'logicism' and his Law of Likelihood: "It may be tempting to sum up this opinion in the words, 'There is no such thing as a logic of statistical inference.' But to say that is to grant too much to the logicist, for it is to suppose that (1) The probability of H on A is p, and (2) $H$is more probable on A than H^* on A^* and the like are the province of inductive logic. On the contrary they are grounded on a false analogy with deductive logic. ... I do not mean that there is no role for likelihood or significance levels etc., but only that these fundamental concepts should not

be understood in a logicist way." (145).

Hacking's (1980) rejection of the Law of Likelihood (LL) and his 'logic of statistical inference' puts his discussion of frequentist inference in the 1965 book in a very different light altogether. In particular, his change of mind renders void his criticisms of the Neyman-Pearson (N-P) testing and Confidence Intervals (CIs) in chapters 6-7, as well as his reinterpretation of Fisher's 'fiducial probability' to render it consistent with his LL in chapter 9.

Does this render his 1965 discussion of frequentist inference in general, and N-P testing in particular, irrelevant? No. He raises several issues that are as relevant today as they were in 1965.

One of his most perceptive comments in chapter 7 pertains to the role of *pre-data* error probabilities (significance level and power) for the *post-data* evaluation of testing results: "the whole point of testing is usually to evaluate the hypotheses after a trial on the set-up has been made. And although size and power may be the criteria for before-trial betting, they need not be the criteria for after-trial evaluation."(88)

This is especially relevant for this review because the concept of a *post-data severity evaluation,* discussed in Mayo (2018) aims to addresses this very issue.

Another important point made by Hacking in explaining the notion of support for a hypothesis (chapter 3) is the difference between inference and decision:

> to conclude that an hypothesis is best supported is, apparently, to decide that the hypothesis in question is best supported. Hence it is a decision like any other. But this inference is fallacious. Deciding that something *is* the case differs from deciding to *do* something. ... Hence deciding to do something falls squarely in the province of decision theory, but deciding that something is the case does not. (29)

This gem of an argument calls into question any claims that decision theory provides all the answers to problems of inference; see Spanos (2017).

Another highly insightful argument that Hacking articulates in chapter 11 is on point estimation. He argues that estimator $\widehat{\theta}(\mathbf{X})$ of θ, however optimal, does not justify the inferential claim that $\widehat{\theta}(\mathbf{x}_0) \simeq \theta^*$, where '$\simeq$' denotes 'approximately equal to' and θ^* denotes the true value of θ; the value that could have generated data $\mathbf{x}_0$. His discussion of why, however, does not explain adequately the fact that $\widehat{\theta}(\mathbf{x}_0)$ represents just a *single value* from an infinite range of possible values in the parameter space Θ $(\theta \in \Theta)$ as it relates to the sampling distribution of $\widehat{\theta}(\mathbf{X})$. What is more, interval estimation and hypothesis testing rectify this problem by taking into account the sampling distribution of an estimator $\widehat{\theta}(\mathbf{X})$ when deriving the relevant error probabilities; type I, II and coverage. Having said that, optimal interval estimation and hypothesis testing begin with an optimal estimator to ensure their own effectiveness (optimality) in learning from data about θ^*; a point not emphasized by Hacking (1965). Instead, he focuses on 'uniform bestness' for point estimates and discusses *admissibility* of estimators which runs afoul the primary aim of frequentist inference: learning from data about the 'true' value θ^* that could have generated data $\mathbf{x}_0$. This is because the idea an estimator which is 'best' for all values $\theta \in \Theta$ is at odds with how well an estimator $\widehat{\theta}(\mathbf{X})$ pinpoints θ^*; see Spanos (2017).

Where Hacking is led astray by his comparative support notion is when he argues: "An hypothesis should be rejected if and only if there is some rival hypothesis much better supported than it is"(81). But as pointed out by Barnard (1972) "there *always* is such a rival hypothesis, *viz.* that things just had to turn out the way they actually did."(129), i.e. the Maximum Likelihood (ML) estimate $\widehat{\theta}_{ML}(\mathbf{x}_0)$. This is relevant for the discussion that follows because Mayo (2018) uses this to question both the likelihoodist and the Bayesian approach to testing based on Bayes factors.

The biggest weakness of the case pertaining to the foundations of statistical induction articulated in Hacking (1965) is his view of 'long run frequency' (chapters 1 and 4). The discussion relies primarily on the philosophical literature based on

von Mises' frequentist interpretation of probability anchored on his notion of a 'collective' and the associated idea of *enumerative induction*: if m/n observed A's are B's, infer (inductively) that approximately m/n of all A's are B's; see Salmon (1967).

Although informative, the discussion ignores the probabilistic formulation associated with *model-based* frequentist interpretation of probability, grounded on the Strong Law of Large Numbers (SLLN), that is clearly articulated in Cramer (1946, 332) and Neyman (1952):

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

Model-based induction revolves around the concept of a statistical model:

$$\mathcal{M}_\theta(\mathbf{x})=\{f(\mathbf{x};\theta),\ \theta\in\Theta\subset\mathbb{R}^m\},\ \mathbf{x}\in\mathbb{R}^n_X,\ n>m, \tag{1}$$

where $f(\mathbf{x};\theta),\ \mathbf{x}\in\mathbb{R}^n_X$ denotes the joint distribution of the sample $\mathbf{X}:=(X_1,\ldots,X_n)$ and Θ and $\mathbb{R}^n_X$ denote the parameter and sample spaces, respectively. This formalization stems directly from the Kolmogorov axiomatization of probability based on an abstract probability space $(S,\mathfrak{F},\ \mathbb{P}(.))$ with $\mathbb{P}(.)$ satisfying the three axioms. The link between $(S,\mathfrak{F},\ \mathbb{P}(.))$ and $\mathcal{M}_\theta(\mathbf{x})$ is provided by the concept of a random variable X:

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

$\mathcal{M}_\theta(\mathbf{x}),\mathbf{x}\in\mathbb{R}^n_X$ comprises the probabilistic assumptions imposed on stochastic process $\{X_t,\ t\in\mathbb{N}\}$ underlying the data $\mathbf{x}_0$. That is,

$\mathcal{M}_\theta(\mathbf{x})$ represents the stochastic mechanism assumed to have given rise to $\mathbf{x}_0$.

Example 1. Consider the simple Bernoulli model:

$$\mathcal{M}_\theta(\mathbf{x})\text{: } X_k \backsim \mathsf{BerIID}(\theta, \theta(1-\theta)),\ x_k=0,1,\ \theta\in[0,1],\quad k=1,2,\dots n,\dots, \tag{2}$$

where '$\mathsf{BerIID}(\theta, \theta(1-\theta))$' stands for Bernoulli, Independent and Identically Distributed (IID), with mean θ and variance $\theta(1-\theta)$; k is an index that denotes the ordering of the sample, say the first trial, second trial, etc. In this case the distribution of the sample $f(\mathbf{x};\theta),\ \mathbf{x}\in\mathbb{R}^n_X$, takes the form:

$$\begin{aligned} f(\mathbf{x};\theta) &\overset{\text{IID}}{=} \textstyle\prod_{k=1}^n f(x_k;\theta)=\theta^{\sum x_k}(1-\theta)^{\sum(1-x_k)} \\ &=\theta^y(1-\theta)^{n-y},\ \mathbf{x}\in\{0,1\}^n, \end{aligned} \tag{3}$$

where $Y=\sum_{k=1}^n X_k$ is Binomially (Bin) distributed:

$$Y=\textstyle\sum_{k=1}^n X_k \backsim \mathsf{Bin}\left(n\theta,\ n\theta(1-\theta);n)\right),\ y=0,1,2,...,n, \tag{4}$$

whose density function is: $f(y;\theta,n)=\binom{n}{y}\theta^y(1-\theta)^{n-y}$, $y=0,1,2,...,n$; $\binom{n}{y}=\frac{n!}{(n-y)!y!}$.

The probabilistic assumptions of the model, encapsulated in $f(\mathbf{x};\theta),\ \mathbf{x}\in\mathbb{R}^n_X$, determine the likelihood function via:

$$L(\theta;\mathbf{x}_0)\propto f(\mathbf{x}_0;\theta),\ \theta\in\Theta, \tag{5}$$

For the simple Bernoulli model in (3), the likelihood is:

$$L(\theta;\mathbf{x}_0)\propto \theta^y(1-\theta)^{n-y},\ \theta\in[0,1]. \tag{6}$$

The model-based frequentist interpretation of probability has certain distinct features that render it very different from the von Mises interpretation (Spanos, 2013):

(a) It revolves around the concept of a statistical model $\mathcal{M}_\theta(\mathbf{x})$, broadly viewed to accommodate non-IID samples.

(b) It is firmly anchored on the SLLN. In the case of (3):

$$\mathbb{P}(\lim_{n\to\infty}(\tfrac{1}{n}\textstyle\sum_{k=1}^n X_k)=\theta)=1. \tag{7}$$

That is, as $n\to\infty$ the stochastic sequence $\{\overline{X}_n=\frac{1}{n}\sum_{k=1}^n X_k\}_{n=1}^\infty$, converges to a constant θ *with probability one* or *almost surely* (*a.s.*) $[\overline{X}_n \overset{a.s.}{\to} \theta]$. It is a measure-theoretic result (Williams,

2001, 111) that does not involve any form of convergence of $\{\overline{x}_n=\frac{1}{n}\sum_{k=1}^{n} x_k\}_{n=1}^{\infty}$ to θ, where $\mathbf{x}_0=\{x_k\}_{k=1}^{n}$; recall that X_k denotes the random variable and x_k a particular value of X_k.

Unfortunately, the line between probabilistic (*a.s.*) and mathematical convergence was blurred by von Mises's (1928) *collective* which was defined in terms of infinite realizations $\{x_k\}_{k=1}^{\infty}$ whose partial sums $\{\overline{x}_n\}_{n=1}^{\infty}$ converge to θ. It turns out that any attempt to make rigorous the convergence $\lim_{n\to\infty}\overline{x}_n=\theta$ is ill-fated for mathematical reasons; see Williams (2001, 25).

(c) The 'long-run' is just a metaphor relating probabilities to relative frequencies and can be rendered operational by validating $\mathcal{M}_\theta(\mathbf{x})$ and using it to simulate faithful replicas of the original data $\mathbf{x}_0$. This model-based framing draws a clear distinction between 'probability' in terms of which the inductive premises of inference ($\mathcal{M}_\theta(\mathbf{x})$) is specified and 'relative frequencies' associated with the long-run metaphor. The latter represents an intuitive way to visualize and understand 'probabilities' in terms of 'relative frequencies' that relate to data $\mathbf{x}_0$.

(d) The link between the mathematical (measure-theoretic) results and real-world phenomena is provided by viewing data $\mathbf{x}_0:=(x_1,...,x_n)$ as a 'typical realization' of the stochastic process $\{X_t,\ t\in\mathbb{N}\}$ underlying $\mathcal{M}_\theta(\mathbf{x})$. Hence, it is justified on empirical grounds, establishing the validity of the probabilistic assumptions of $\mathcal{M}_\theta(\mathbf{x})$.

Admittedly, the concept of a statistical model $\mathcal{M}_\theta(\mathbf{x})$ is a form of a *chance set-up,* as envisioned by Hacking, but the latter is too vague to serve as the cornerstone of statistical modeling and inference. What is needed is a more precisely defined concept, such as $\mathcal{M}_\theta(\mathbf{x})$, that describes 'a stochastic generating mechanism' in terms of the stochastic process $\{X_t,\ t\in\mathbb{N}\}$ underlying data $\mathbf{x}_0$.

The vague generality of a chance set-up in conjunction with von Mises' notion of a collective create a number of problems because they replace the precisely defined and easily testable set of probabilistic assumptions with notions like 'random samples' stemming from (a) the "uniformity" of nature (population) and (b) the "representativeness" of the sample. These, in turn raise additional issues, including the *reference class* and

the *single case* problems; see Salmon (1967). The key difference between enumerative and model-based induction is that latter revolves around $\mathcal{M}_\theta(\mathbf{x})$ whose premises are specified in terms of probabilistic assumptions that are testable vis-à-vis data $\mathbf{x}_0$. The model-based perspective sheds very different light on the single case and reference class problems; see Spanos (2013).

Worse, in defining the most crucial concept in his book, the *likelihood function*, Hacking ends up with a very cumbersome notation for a function with six arguments: "a joint proposition is one which states, 'the distribution of chances on trials of kind K on set-up X is a member of the class Δ; outcome E occurs on trial T of kind K'. Here K' might be K, but need not be. A joint proposition is represented: $< X, K, \Delta; T, K', E >$"(52)

Moreover, his framing of a chance set-up requires one to distinguish between *discrete* and *continuous* probability distributions, which is a completely unnecessary distinction because $L(\theta; \mathbf{x}_0)$ is a function of θ and not $\mathbf{x}$, and Θ is usually an uncountable subset of the real line $\mathbb{R}$. Instead of (6), Hacking defines the likelihood as:

$$L(\theta; \mathbf{x}_0) \propto \binom{n}{y}\theta^y(1-\theta)^{n-y},\ \theta\in[0,1], \tag{8}$$

which coincides with the sampling distribution of Y as defined in (4); see also Royal (1997, 19). That is, for evaluating a continuous function of θ ($L(\theta; \mathbf{x}_0),\ \theta\in\Theta$) he uses a discrete distribution in conjunction with specific values of θ. This works in practice because both (6) and (8) comply with (5) and the proportionality factor $\binom{n}{y}$ is free of θ.

In relation to Hacking's (1965) doubts on whether the notion of relative frequency in the context of model-based induction is "a physical property at all," Cramer (1946, 332), gave an affirmative answer almost 20 years earlier. After his change of mind, Hacking (1980) adopts the same answer: "Probability in this sense [objective] does not mean 'relative frequency', but probabilities are typically manifested by stable frequencies." (150). His revised view all but echoes Neyman (1952, 27) quoted above.

The last two chapters of Hacking (1965) provide a philosophical scrutiny of *Bayesian statistics* when examined from the viewpoint of formal logic and relates that to his proposed logic

of *statistical support*. Chapter 12 brings out explicitly several assumptions underlying Bayes' theory in an attempt to place Bayes' original problem in a formal framework with a view to relate it to Hacking's logic of statistical support. The proposed Bayesian framework is extended to include Jeffreys' (1938) *The Theory of Probability* by adding an additional assumption pertaining to the uniqueness of prior probabilities. Although both Bayesian formulations are consistent with Hacking's logic of statistical support, he rejects them as formal ways to learn from data primarily because of the lack of any persuasive arguments to render logical the choice of different types of prior distributions, including the uniform and Jeffrey's parameterization invariant priors. These two chapters include a number of insightful and discerning arguments that have not been pursued further by Bayesians or their critics.

In light of Hacking's change of mind on the merits of his approach to *comparative support* of hypotheses (propositions) anchored on his Law of the Likelihood, one might conclude that his 1965 book is mostly of historical interest. That will be a rushed judgment because the book includes several insightful comments and suggestions pertaining to both frequentist and Bayesian statistical inference alluded to above.

On the question posed by the title of this review, Hacking (1965) proposes founding statistical induction on formal logic, but Hacking (1980) reverses course: "My *Logic of Statistical Inference* took vigorous issue with Neyman. This essay is a retraction. I now believe that Neyman, Peirce and Braithwaite were on the right lines to follow in the analysis of inductive arguments." (141).

2 Mayo (2018). Statistical Inference as Severe Testing: How to Get Beyond the Statistics Wars

The sub-title of Mayo's (2018) book provides an apt description of the primary aim of the book in the sense that its focus is

on the current discussions pertaining to replicability and trustworthy empirical evidence that revolve around the main fault line in statistical inference: the nature, interpretation and uses of probability in statistical modeling and inference. This underlies not only the form and structure of inductive inference, but also the nature of the underlying statistical reasonings as well as the nature of the evidence it gives rise to.

A crucial theme in Mayo's book pertains to the current confusing and confused discussions on reproducibility and replicability of empirical evidence. The book cuts through the enormous level of confusion we see today about basic statistical terms, and in so doing explains why the experts so often disagree about reforms intended to improve statistical science.

Mayo makes a concerted effort to delineate the issues and clear up these confusions by defining the basic concepts accurately and placing many widely held methodological views in the best possible light before scrutinizing them. In particular, the book discusses at length the merits and demerits of the proposed reforms which include: (a) replacing p-values with Confidence Intervals (CIs), (b) using estimation-based effect sizes and (c) redefining statistical significance.

The key philosophical concept employed by Mayo to distinguish between a *sound* empirical evidential claim for a hypothesis H and an unsound one is the notion of a *severe test*: if little has been done to *rule out flaws (errors and omissions)* in pronouncing that data $\mathbf{x}_0$ provide evidence for a hypothesis H, then that inferential claim *has not passed* a severe test, rendering the claim untrustworthy. One has trustworthy evidence for a claim C only to the extent that C passes a severe test; see Mayo (1983; 1996). A distinct advantage of the concept of severe testing is that it is sufficiently general to apply to both frequentist and Bayesian inferential methods.

Mayo makes a case that there is a two-way link between philosophy and statistics. On one hand, philosophy helps in resolving conceptual, logical, and methodological problems of statistical inference. On the other hand, viewing statistical inference as severe testing gives rise to novel solutions to crucial philosophical problems including induction, falsification

and the demarcation of science from pseudoscience. In addition, it serves as the foundation for understanding and getting beyond the statistics wars that currently revolves around the replication crises; hence the title of the book, *Statistical Inference as Severe Testing*.

Chapter (excursion) 1 of Mayo's (2018) book sets the scene by scrutinizing the different role of probability in *statistical inference*, distinguishing between:

(i) Probabilism. Probability is used to assign a *degree of confirmation, support or belief* in a hypothesis H, given data $\mathbf{x}_0$ (Bayesian, likelihoodist, Fisher (fiducial). An inferential claim H is warranted when it is assigned a *high* probability, support, or degree of belief (absolute or comparative).

(ii) Performance. Probability is used to ensure the *long-run reliability* of inference procedures; type I, II, coverage probabilities (frequentist, behavioristic Neyman-Pearson). An inferential claim H is warranted when it stems from a procedure with a low long-run error.

(iii) Probativism. Probability is used to assess the *probing capacity* of inference procedures, *pre-data* (type I, II, coverage probabilities), as well as *post-data* (p-value, severity evaluation). An inferential claim H is warranted when the different ways it can be false have been adequately probed and averted.

Mayo argues that probativism based on the severe testing account uses error probabilities to output an evidential interpretation based on assessing how severely an inferential claim H has passed a test with data $\mathbf{x}_0$. Error control and long-run reliability is necessary but not sufficient for probativism. This perspective is contrasted to probabilism (Law of Likelihood (LL) and Bayesian posterior) that focuses on the relationships between data $\mathbf{x}_0$ and hypothesis H, and ignores outcomes $\mathbf{x}\in\mathbb{R}^n_X$ other than $\mathbf{x}_0$ by adhering to the *Likelihood Principle (LP)*: given a statistical model $\mathcal{M}_\theta(\mathbf{x})$ and data $\mathbf{x}_0$, all relevant sample information for inference purposes is contained in $L(\theta;\mathbf{x}_0)$, $\forall\theta\in\Theta$. Such a perspective can produce unwarranted results with high probability, by failing to pick up on optional stopping, data dredging and other biasing selection effects. It is at odds with what is widely accepted as the most effective way to improve

replication: predesignation, and transparency about how hypotheses and data were generated and selected.

Chapter (excursion) 2 entitled 'Taboos of Induction and Falsification' relates the various uses of probability to draw certain parallels between probabilism, Bayesian statistics and Carnapian logics of confirmation on one side, and performance, frequentist statistics and Popperian falsification on the other. The discussion in this chapter covers a variety of issues in philosophy of science, including, the problem of induction, the asymmetry of induction and falsification, sound vs. valid arguments, enumerative induction (straight rule), confirmation theory (and formal epistemology), statistical affirming the consequent, the old evidence problem, corroboration, demarcation of science and pseudoscience, Duhem's problem and novelty of evidence. These philosophical issues are also related to statistical conundrums as they relate to significance testing, fallacies of rejection, the cannibalization of frequentist testing known as Null Hypothesis Significance Testing (NHST) in psychology, and the issues raised by the reproducibility and replicability of evidence.

Chapter (excursion) 3 on 'Statistical Tests and Scientific Inference' provides a basic introduction to frequentist testing paying particular attention to crucial details, such as specifying explicitly the assumed statistical model $\mathcal{M}_\theta(\mathbf{x})$ and the proper framing of hypotheses in terms of its parameter space Θ, with a view to provide a coherent account by avoiding undue formalism. The Neyman-Pearson (N-P) formulation of hypothesis testing is explained using a simple example, and then related to Fisher's significance testing. What is different from previous treatments is that the claimed 'inconsistent hybrid' associated with the NHST caricature of frequentist testing is circumvented. The crucial difference often drawn is based on the N-P emphasis on pre-data long-run error probabilities, and the behavioristic interpretation of tests as accept/reject rules. By contrast, the post-data p-value associated with Fisher's significance tests is thought to provide a more evidential interpretation. In this chapter, the two approaches are reconciled in the context of the error statistical framework. The N-P formulation

provides the formal framework in the context of which an optimal theory of frequentist testing can be articulated, but in its current expositions lack a proper evidential interpretation.

Example 2. Consider the case of the simple Normal model:

$$\begin{array}{c}\mathcal{M}_\theta(\mathbf{x})\text{: } X_t \backsim \mathsf{NIID}(\mu,\sigma^2),\ x_t\in\mathbb{R},\ \mu\in\mathbb{R},\ \sigma^2>0,\\ t\in\mathbb{N}:=(1,2,\dots n,\dots),\end{array} \tag{9}$$

where σ^2 is assumed known. An optimal N-P test for the hypotheses:

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

is defined by $\mathcal{T}_\alpha:=\{d(\mathbf{X}), C_1(\alpha)\}$, where $d(\mathbf{X})=\frac{\sqrt{n}(\overline{X}_n-\mu_0)}{\sigma}$, $\overline{X}_n=\frac{1}{n}\sum_{t=1}^n X_t$ is the test statistics and $C_1(\alpha)=\{\mathbf{x}\text{: } d(\mathbf{x})>c_\alpha\}$, the rejection region with the sampling distributions for evaluating the relevant error probabilities taking the form:

$$\begin{array}{c}\text{(i) } d(\mathbf{X}) \overset{\mu=\mu_0}{\backsim} \mathsf{N}(0,1),\ \text{(ii) } d(\mathbf{X}) \overset{\mu=\mu_1}{\backsim} \mathsf{N}(\delta_1,1),\ \delta_1=\frac{\sqrt{n}(\mu_1-\mu_0)}{\sigma},\\ \text{for all } \mu_1>\mu_0.\end{array} \tag{10}$$

The pre-data power function at each μ_1 is defined by:

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

Fisher's p-value aspired to provided an evidential interpretation: "The actual value of p ... indicates the strength of evidence against the hypothesis"(Fisher, 1925, 80). It fell short of that goal because of two crucial weakness:
(a) The p-value $p(\mathbf{x}_0)=\mathbb{P}(d(\mathbf{X})>d(\mathbf{x}_0);H_0)$ depends crucially on the particular *statistical context* that comprises four components:

$$\begin{array}{l}\text{(i) } \mathcal{M}_\theta(\mathbf{x}),\\ \text{(ii) } H_0\text{: } \theta\in\Theta_0 \text{ vs. } H_1\text{: } \theta\in\Theta_1,\\ \text{(iii) } \mathcal{T}_\alpha:=\{d(\mathbf{X}), C_1(\alpha)\},\\ \text{(iv) data } \mathbf{x}_0.\end{array} \tag{11}$$

(b) By itself the p-value cannot provide sufficient information for or against H_0 since the sampling distribution of $d(\mathbf{X})$ is evaluated *only* under H_0. One would need additional information pertaining to discrepancies from H_0 and the test's power to detect them. That is, a p-value $p(\mathbf{x}_0)<\alpha$ indicates 'some' discrepancy from $\theta=\theta_0$, but contains no information about its

magnitude. Indeed, since the 1930s it is known that there is always a large enough n to reject H_0, however small $(\theta^*-\theta_0)\neq 0$ and a threshold $c>0$, such that $p(\mathbf{x}_0) \underset{n\to\infty}{\to} 0$; *the large* n *problem*. This happens because the power of a consistent test increases with n. In example 2, the power of the test increases with δ_1 in (10)(ii), which increases monotonically with $\sqrt{n}$.

As argued in Mayo (2018), the *post-data severity evaluation* of the testing results can be used to shed light on (a)-(b) as well as provide the missing evidential interpretation. This is achieved by particularizing the notion of severity to frequentist testing within $\mathcal{M}_\theta(\mathbf{x})$. If a hypothesis H_0 passes a test $\mathcal{T}_\alpha$ that was highly capable of finding discrepancies from it, were they to be present, then the passing result indicates some evidence for their absence. The resulting evidential result comes in the form of the magnitude of the discrepancy γ from H_0 warranted with test $\mathcal{T}_\alpha$ and data $\mathbf{x}_0$ at different levels of severity. The intuition underlying the post-data severity is that a small p-value or a rejection of H_0 based on a test with low power (e.g. a small n) for detecting a particular discrepancy γ provides *stronger* evidence for the presence of γ than if the test had much higher power (e.g. a large n).

The *post-data severity evaluation* outputs the discrepancy γ stemming from the testing results and takes the probabilistic form:

$$SEV(\theta \lesseqgtr \theta_1;\mathbf{x}_0)=\mathbb{P}(d(\mathbf{X}) \gtrsim d(\mathbf{x}_0);\ \theta_1=\theta_0+\gamma),\ \text{for all } \theta_1\in\Theta_1,$$

where the inequalities are determined by the testing result and the sign of $d(\mathbf{x}_0)$. When the relevant N-P test result is 'accept (reject) H_0' one is seeking the smallest (largest) discrepancy γ, in the form of an inferential claim $\theta \lesseqgtr \theta_1=\theta_0+\gamma$, warranted by $\mathcal{T}_\alpha$ and $\mathbf{x}_0$ at a high enough probability, say .8 or .9. The severity evaluations are introduced by connecting them to more familiar calculations relating to observed confidence intervals and p-value calculations. A more formal treatment to the post-data severity evaluation is given in chapter (excursion) 5.

Mayo uses the post-data severity perspective to scorch several misinterpretations of the p-value, including the claim that the p-value is not a legitimate error probability. She also calls

into question any comparisons of the tail areas of $d(\mathbf{X})$ under H_0 that vary with $\mathbf{x}\in\mathbb{R}^n$, with posterior distribution tail areas that vary with $\theta\in\Theta$, pointing out that this is tantamount to comparing apples and oranges!

The real life examples of the 1919 eclipse data for testing the General Theory of Relativity, as well as the 2012 discovery of the Higgs particle are used to illustrate some of the concepts in this chapter.

The discussion in this chapter sheds light on several important problems in statistical inference, including several howlers of statistical testing, Jeffreys' tail area criticism, weak conditionality principle and the likelihood principle.

Chapter (excursion) 4 entitled 'Objectivity and Auditing' discusses 'error statistics', which can be viewed as a refinement/extension of the Fisher-Neyman-Pearson (F-N-P) framing of frequentist modeling and inference. Error statistics *extends* the F-N-P approach by supplementing it with a post-data severity evaluation that goes beyond the testing results to provide an evidential interpretation in the form of the warranted discrepancy from H_0; see Mayo and Spanos (2006; 2011). It *refines* the F-N-P framing by proposing a broader framework wherein the *modeling* and *inference facets* are separated. The modeling facet includes *specification* (initial choice of $\mathcal{M}_\theta(\mathbf{x})$), *Mis-Specification (M-S) testing* and *respecification* with a view to secure statistical adequacy: the validity of the probabilistic assumptions comprising the statistical model $\mathcal{M}_\theta(\mathbf{x})$ (Mayo and Spanos, 2004; Spanos, 2018). Only when the statistical adequacy of $\mathcal{M}_\theta(\mathbf{x})$ is established, should one proceed to the *inference facet* since the latter assumes the validity of $\mathcal{M}_\theta(\mathbf{x})$.

In error statistics particular emphasis is placed on specifying a statistical model in terms of a complete, internally consistent and testable set of probabilistic assumptions in terms of the observable process $\{X_t,\ t\in\mathbb{N}\}$ underlying data $\mathbf{x}_0$, and not an unobservable error term process. A quintessential example of such a specification is that of the Linear Regression model in table 1 comprising a statistical Generating Mechanism (GM), assumptions [1]-[5] and the statistical parametrization.

When $\mathcal{M}_\theta(\mathbf{x})$ is *statistically misspecified*–certain assumptions

are invalid for data $\mathbf{x}_0$–$\mathcal{M}_{\hat{\theta}}(\mathbf{x})$ will give rise to *untrustworthy evidence* since both $f(\mathbf{x};\theta),\ \mathbf{x}\in\mathbb{R}^n_X$ and the likelihood function $L(\theta;\mathbf{x}_0)\propto f(\mathbf{x}_0;\theta),\ \theta\in\Theta$ are erroneous. That, in turn *distorts* the sampling distribution $f(y_n;\theta)$ of any statistic (estimator, test statistic) $Y_n=g(X_1,X_2,...,X_n)$, rendering it erroneous since its derivation is based on $f(\mathbf{x};\theta),\ \mathbf{x}\in\mathbb{R}^n_X$. Hence, statistical misspecification could easily give rise to*inconsistent* estimators and or sizeable discrepancies between the *actual* error probabilities (type I, II, p-values, coverage) and the *nominal* (assumed) ones, rendering any inferences based on such statistics unreliable; see Spanos (2019).

Table 1: Normal, Linear Regression model

Statistical GM: $Y_t=\beta_0+\beta_1 x_t+u_t,\ \ t\in\mathbb{N}:=(1,2,...,n,...)$

[1]	Normality:	$(Y_t\|X_t=x_t)\backsim \mathsf{N}(.,.)$,	$\left.\begin{array}{l}\\\\\\\\\end{array}\right\}$ $t\in\mathbb{N}$.
[2]	Linearity:	$E\,(Y_t\|X_t=x_t))=\beta_0+\beta_1 x_t$,	
[3]	Homoskedasticity:	$Var\,(Y_t\|X_t=x_t)=\sigma^2$,	
[4]	Independence:	$\{(Y_t\|X_t=x_t)\,,\ t\in\mathbb{N}\}$ indep. process,	
[5]	t-invariance:	$(\beta_0,\beta_1,\sigma^2)$ are *not* changing with t,	

$\beta_0=E(Y_t)-\beta_1 E(X_t)\in\mathbb{R},\ \beta_1=\left(\frac{Cov(X_t,Y_t)}{Var(X_t)}\right)\in\mathbb{R}$,

$\sigma^2=Var(X_t)-\frac{[Cov(X_t,Y_t)]^2}{Var(X_t)}\in\mathbb{R}_+$.

These refinements/extensions render the error statistical approach more transparent and allows for third parties to reproduce/replicate the results with a view to independently affirm/deny the reported inferential claims. While knowledge gaps leave room for biases and inappropriate choices in empirical modeling, the error statistical approach brings out such choices and demands that they should be checked against the data before drawing any inferences. In error statistics a statistical method's objectivity requires the ability to audit an inference: check assumptions, pinpoint blame for anomalies, falsify, and directly register how biasing selection effects (p-hunting, multiple testing and cherry-picking) undermine its error probing capacities. This calls into question Bayesian claims, such as "likelihoods are as subjective as priors" or "statistical infer-

ence is just a matter of subjective choices and beliefs", in their attempt to deflect criticisms away from their favored approach.

M-S testing constitutes a form of significance testing that differs from N-P testing primarily because the latter is testing within $\mathcal{M}_\theta(\mathbf{x})$ and the former is testing outside it. The default H_0 in M-S testing pertains to the validity of $\mathcal{M}_\theta(\mathbf{x})$ as a whole, i.e. H_0: $f(\mathbf{x};\theta^*)\in\mathcal{M}_\theta(\mathbf{x})$, but the default H_1: $f(\mathbf{x};\theta^*)\notin\mathcal{M}_\theta(\mathbf{x})$ is non-operational and there are many different ways to operationalize it based on different ways to parameterize departures from the model assumptions; see Spanos (2018).

A particularly misleading slogan that is widely invoked as an alibi for ignoring the validation of the inductive premises defined by $\mathcal{M}_\theta(\mathbf{x})$ is 'all models are wrong, but some are useful' attributed to George Box (1979). Glancing at Box (1979), however, reveals that the slogan is referring to the 'realisticness' of the estimated model: "Now it would be very remarkable if any system existing in the real world could be exactly represented by any simple model."(202), and thus to substantive and not statistical misspecification. Indeed, Box proceeds to emphasize the importance of viewing modeling as an iterative process where testing the validity of the probabilistic assumptions is an invaluable guide to more reliable models!

This chapter also addresses a number of foundational problems in empirical modeling, including conciliating *substantive subject matter* with *statistical information* as well as *Duhem's problem*: a scientific hypothesis cannot be empirically tested in isolation because such a test requires additional auxiliary assumptions whose validity cannot be evaluated separately; it can in the context of error statistics.

The error statistical approach is then used to shed light on several confusions and problems relating to the current literature on replication, including the large n problem, the fallacy of rejection, claims that the p-values exaggerate evidence, the Jeffreys-Lindley paradox, Bayes factors and the use of spiked priors, randomized control trials (RCTs), bootstrapping and nonsense correlations/regressions.

Chapter (excursion) 5, entitled 'Power and Severity', provides an in-depth discussion of power and its abuses or mis-

interpretations, as well as scotch several confusions permeating the current discussions on the replicability of empirical evidence.

Confusion 1: The power of a N-P test $\mathcal{T}_\alpha{:=}\{d(\mathbf{X}), C_1(\alpha)\}$ is a *pre-data* error probability that calibrates the generic (for any sample realization $\mathbf{x}{\in}\mathbb{R}^n_X$) capacity of the test in detecting different discrepancies from H_0, for a given type I error probability α. As such, the power is *not* a point function one can evaluate arbitrarily at a particular value θ_1. It is defined for all values in the alternative space $\theta_1{\in}\Theta_1$.

Confusion 2: The power function is properly defined for all $\theta_1{\in}\Theta_1$ only when (Θ_0, Θ_1) constitute a partition of Θ. This is to ensure that θ^* is not in a subset of Θ ignored by the comparisons since the *main objective* is to *narrow down* the unknown parameter space Θ using *hypothetical* values of θ. Ideally, the narrowing reduces $\mathcal{M}_\theta(\mathbf{x})$ to a single point $\mathcal{M}^*(\mathbf{x}){=}\{f(\mathbf{x};\theta^*)\}$, $\mathbf{x}{\in}\mathbb{R}^n_X$, where θ^* denotes the 'true' value of θ in Θ; shorthand for saying $f(\mathbf{x};\theta^*), \mathbf{x}{\in}\mathbb{R}^n_X$, could have generated data $\mathbf{x}_0$. In practice, this 'ideal' situation is unlikely to be reached, except by happenstance. This, however, does not prevent 'learning from data' $\mathbf{x}_0$. Hypothesis testing poses questions as to whether a hypothetical value θ_0 is close enough to θ^* in the sense that the difference $(\theta^*{-}\theta_0)$ is 'statistically negligible'; a notion defined using error probabilities.

Confusion 3: Hacking (1965) raised the problem of using pre-data error probabilities, such as the significance level α and power, to evaluate the testing results post-data. As mentioned above, the post-data severity aims to address that very problem, and is extensively discussed in Mayo (2018), excursion 5.

Confusion 4: Mayo and Spanos (2006) define "attained power" by replacing c_α with the observed $d(\mathbf{x}_0)$. But this should not be confused with replacing θ_1 with its observed estimate $\widehat{\theta}(\mathbf{x}_0)$, as in what is often called "observed" or "retrospective" power. To compare the two in example 2, contrast:

> Attained power: $\mathcal{P}(\mu_1){=}\mathbb{P}(d(\mathbf{X}) > d(\mathbf{x}_0);\ \mu{=}\mu_1)$,
> for all $\mu_1{>}\mu_0$,

with what Mayo calls Shpower which is defined at $\mu{=}\overline{x}_n$:

$$\text{Shpower:} \quad \mathcal{P}(\overline{x}_n){=}\mathbb{P}(d(\mathbf{X}) > d(\mathbf{x}_0);\ \mu{=}\overline{x}_n).$$

Shpower makes very little statistical sense, unless point estimation justifies the inferential claim $\overline{x}_n \simeq \mu^*$, which it does not, as argued above. Unfortunately, the statistical literature in psychology is permeated with (implicitly) invoking such a claim when touting the merits of estimation-based effect sizes. The estimate $\overline{x}_n$ represents just a single value of $\overline{X}_n \backsim \mathsf{N}(\mu, \frac{\sigma^2}{n})$, and any inference pertaining to μ needs to take into account the uncertainty described by this sampling distribution; hence, the call for using interval estimation and hypothesis testing to account for that sampling uncertainty. The post-data severity evaluation addresses this problem using hypothetical reasoning and taking into account the relevant statistical context (11). It outputs the discrepancy from H_0 warranted by test $\mathcal{T}_\alpha$ and data $\mathbf{x}_0$, with high enough severity, say bigger than .85. Invariably, inferential claims of the form $\mu \gtrless \mu_1{=}\overline{x}_n$ are assigned low severity of .5.

Confusion 5: Frequentist error probabilities (type I, II, coverage, p-value) are *not conditional* on H (H_0 or H_1) since $\theta{=}\theta_0$ or $\theta{=}\theta_1$ being 'true or false' do not constitute *legitimate events* in the context of $\mathcal{M}_\theta(\mathbf{x})$; θ is an *unknown constant*. The clause 'given H is true' refers to *hypothetical scenarios* under which the sampling distribution of the test statistic $d(\mathbf{X})$ is evaluated as in (10).

This confusion undermines the credibility of Positive Predictive Value (PPV):

$$PPV{=}\Pr(F|R){=}\frac{\Pr(R|F)\Pr(F)}{\Pr(R|F)P(F){+}\Pr(R|\overline{F})P(\overline{F})},$$

where (i) $F{=}H_0$ is false, (ii) R=test rejects H_0, and (iii) H_0: no disease, used by Ioannidis (2005) to make his case that 'most published research findings are false' when $PPV{=}\Pr(F|R){<}.5$. His case is based on 'guessing' probabilities at a discipline wide level, such as $\Pr(F){=}.1$, $\Pr(R|F){=}.8$ and $\Pr(R|\overline{F}){=}.15$, and presuming that the last two relate to the power and significance

level of a N-P test. He then proceeds to blame the wide-spread abuse of significance testing (p-hacking, multiple testing, cherry-picking, low power) for the high de facto type I error (.15). Granted, such abuses do contribute to untrustworthy evidence, but not via false positive/negative rates since (i) and (iii) are not legitimate events in the context of $\mathcal{M}_\theta(\mathbf{x})$, and thus $\Pr(R|F)$ and $\Pr(R|\overline{F})$ have nothing to do with the significance level and the power of a N-P test. Hence, the analogical reasoning relating the *false positive and false negative rates* in medical detecting devices to the type I and II error probabilities in frequentist testing is totally misplaced. These rates are established by the manufacturers of medical devices after running a very large number (say, 10000) of medical 'tests' with specimens that are *known* to be positive or negative; they are prepared in a lab. Known 'positive' and 'negative' specimens constitute legitimate observable events one can condition upon. In contrast, frequentist error probabilities (i) are framed in terms of θ (which are not observable events in $\mathcal{M}_\theta(\mathbf{x})$) and (ii) depend crucially on the particular statistical context (11); there is no statistical context for the *false positive and false negative rates*.

A stronger case can be made that abuses and misinterpretations of frequentist testing are only symptomatic of a more extensive problem: the *recipe-like/uninformed implementation of statistical methods*. This contributes in many different ways to untrustworthy evidence, including: (i) statistical misspecification (imposing invalid assumptions on one's data), (ii) poor implementation of inference methods (insufficient understanding of their assumptions and limitations), and (iii) unwarranted evidential interpretations of their inferential results (misinterpreting p-values and CIs, etc.).

Mayo uses the concept of a post-data severity evaluation to illuminate the above mentioned issues and explain how it can also provide the missing evidential interpretation of testing results. The same concept is also used to clarify numerous misinterpretations of the p-value throughout this book, as well as the fallacies:

(a) **Fallacy of acceptance (non-rejection)**. No evidence against H_0 is misinterpreted as evidence for it. This fallacy can easily

arise when the power of a test is low (e.g. small n problem) in detecting sizeable discrepancies.
(b) **Fallacy of rejection**. Evidence against H_0 is misinterpreted as evidence for a particular H_1. This fallacy can easily arise when the power of a test is very high (e.g. large n problem) and it detects trivial discrepancies; see Mayo and Spanos (2006).

In chapter 5 Mayo returns to a recurring theme throughout the book, the mathematical duality between Confidence Intervals (CIs) and hypothesis testing, with a view to call into question certain claims about the superiority of CIs over p-values. This mathematical duality derails any claims that observed CIs are less vulnerable to the large n problem and more informative than p-values. Where they differ is in terms of their inferential claims stemming from their different forms of reasoning, factual vs. hypothetical. That is, the mathematical duality does not imply inferential duality. This is demonstrated by contrasting CIs with the post-data severity evaluation.

Indeed, a case can be made that the post-data severity evaluation addresses several long-standing problems associated with frequentist testing, including the large n problem, the apparent arbitrariness of the N-P framing that allows for simple vs. simple hypotheses, say H_0: $\mu = -1$ vs. H_1: $\mu=1$, the arbitrariness of the rejection thresholds, the problem of the sharp dichotomy (e.g. reject H_0 at .05 but accept H_0 at .0499), and distinguishing between statistical and substantive significance. It also provides a natural framework for evaluating reproducibility/replicability issues and brings out the problems associated with observed CIs and estimation-based effect sizes; see Spanos (2019).

Chapter 5 also includes a retrospective view of the disputes between Neyman and Fisher in the context of the error statistical perspective on frequentist inference, bringing out their common framing and their differences in emphasis and interpretation. The discussion also includes an interesting summary of their personal conflicts, not always motivated by statistical issues; who said the history of statistics is boring?

Chapter (excursion) 6 of Mayo (2018) raises several important foundational issues and problems pertaining to Bayesian

inference, including its primary aim, subjective vs. default Bayesian priors and their interpretations, default Bayesian inference vs. the Likelihood Principle, the role of the catchall factor, the role of Bayes factors in Bayesian testing, and the relationship between Bayesian inference and error probabilities. There is also discussion about attempts by 'default prior' Bayesians to unify or reconcile frequentist and Bayesian accounts.

A point emphasized in this chapter pertains to model validation. Despite the fact that Bayesian statistics shares the same concept of a statistical model $\mathcal{M}_\theta(\mathbf{x})$ with frequentist statistics, there is hardly any discussion on validating $\mathcal{M}_\theta(\mathbf{x})$ to secure the reliability of the posterior distribution:
$\pi(\theta|\mathbf{x}_0) \propto \pi(\theta)\cdot L(\theta|\mathbf{x}_0),\ \forall\theta\in\Theta,$ upon which all Bayesian inferences are based. The exception is the indirect approach to model validation in Gelman et al (2013) based on the *posterior predictive distribution:*

$$m(\mathbf{x})=\int_{\theta\in\mathbf{\Theta}} f(\mathbf{x};\theta)\pi(\theta|\mathbf{x}_0)d\theta,\ \forall\mathbf{x}\in\mathbb{R}^n_X. \qquad (12)$$

Since $m(\mathbf{x})$ is parameter free, one can use it as a basis for simulating a number of replications $\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N$ to be used as *indirect* evidence for potential departures from the model assumptions vis-à-vis data $\mathbf{x}_0$, which is clearly different from frequentist M-S testing of the $\mathcal{M}_\theta(\mathbf{x})$ assumptions. The reason is that $m(\mathbf{x})$ is a smoothed mixture of $f(\mathbf{x};\theta)$ and $\pi(\theta|\mathbf{x}_0)$ and one has no way of attributing blame to one or the other when any departures are detected. For instance, in the case of the simple Normal model in (9), a highly skewed prior might contribute (indirectly) to departures from the Normality assumption when tested using simulated data using (12). Moreover, the 'smoothing' with respect to the parameters in deriving $m(\mathbf{x})$ is likely to render testing departures from the IID assumptions a lot more unwieldy.

On the question posed by the title of this review, Mayo's answer is that the error statistical framework, a refinement or extension of the original Fisher-Neyman-Pearson framing in the spirit of Peirce, provides a pertinent foundation for frequentist modeling and inference.

3 Conclusions

A retrospective view of Hacking (1965) reveals that its main weakness is that its perspective on statistical induction adheres too closely to the philosophy of science framing of that period, and largely ignores the formalism based on the theory of stochastic processes $\{X_t,\ t\in\mathbb{N}\}$ that revolves around the concept of a statistical model $\mathcal{M}_\theta(\mathbf{x})$. Retrospectively, its value stems primarily from a number of very insightful arguments and comments that survived the test of time. The three that stand out are: (i) an optimal point estimator $\widehat{\theta}(\mathbf{X})$ of θ does not warrant the inferential claim $\widehat{\theta}(\mathbf{x}_0)\ \simeq\ \theta^*$, (ii) a statistical inference is very different from a decision, and (iii) the distinction between the pre-data error probabilities and the post-data evaluation of the evidence stemming from testing results; a distinction that permeates Mayo's (2018) book. Hacking's change of mind on the aptness of logicism and the problems with the long run frequency is also particularly interesting. Hacking's (1980) view of the long run frequency is almost indistinguishable from that of Cramer (1946, 332) and Neyman (1952, 27) mentioned above, or Mayo (2018), when he argues: "Probabilities conform to the usual probability axioms which have among their consequences the essential connection between individual and repeated trials, the weak law of large numbers proved by Bernoulli. Probabilities are to be thought of as theoretical properties, with a certain looseness of fit to the observed world. Part of this fit is judged by rules for testing statistical hypotheses along the lines described by Neyman and Pearson. It is a "frequency view of probability" in which probability is a dispositional property..." (Hacking, 1980, 150-151).

Probability as a dispositional property' of a chance set-up alludes to the *propensity interpretation* of probability associated with Peirce and Popper, which is in complete agreement with the model-based frequentist interpretation; see Spanos (2019).

The main contribution of Mayo's (2018) book is to put forward a framework and a strategy to evaluate the trustworthiness of evidence resulting from different statistical accounts. Viewing statistical inference as a form of severe testing eluci-

dates the most widely employed arguments surrounding commonly used (and abused) statistical methods. In the severe testing account, probability arises in inference, not to measure degrees of plausibility or belief in hypotheses, but to evaluate and control how severely tested different inferential claims are. Without assuming that other statistical accounts aim for severe tests, Mayo proposes the following strategy for evaluating the trustworthiness of evidence: begin with a minimal requirement that if a test has little or no chance to detect flaws in a claim H, then H's passing result constitutes untrustworthy evidence. Then, apply this minimal severity requirement to the various statistical accounts as well as to the proposed reforms, including estimation-based effect sizes, observed CIs and redefining statistical significance. Finding that they fail even the minimal severity requirement provides grounds to question the trustworthiness of their evidential claims. One need not reject some of these methods just because they have different aims, but because they give rise to evidence that fail the minimal severity requirement. Mayo challenges practitioners to be much clearer about their aims in particular contexts and different stages of inquiry. It is in this way that the book ingeniously links philosophical questions about the roles of probability in inference to the concerns of practitioners about coming up with trustworthy evidence across the landscape of the natural and the social sciences.

References

Barnard, George. 1972. Review article: Logic of Statistical Inference. *The British Journal of the Philosophy of Science, 23*: 123-190.

Cramer, Harald. 1946. *Mathematical Methods of Statistics*, Princeton: Princeton University Press.

Fisher, Ronald A. 1922. On the Mathematical Foundations of Theoretical Statistics. *Philosophical Transactions of the Royal Society A, 222*(602): 309-368.

Fisher, Ronald A. 1925. *Statistical Methods for Research Workers*. Edinburgh: Oliver & Boyd.

Gelman, Andrew. John B. Carlin, Hal S. Stern, Donald B. Rubin. 2013. *Bayesian Data Analysis,* 3rd ed. London: Chapman & Hall/CRC.

Hacking, Ian. 1972. Review: Likelihood. *The British Journal for the Philosophy of Science, 23*(2): 132-137.

Hacking, Ian. 1980. The Theory of Probable Inference: Neyman, Peirce and Braithwaite. In D. Mellor (ed.), *Science, Belief and Behavior: Essays in Honour of R. B. Braithwaite*. Cambridge: Cambridge University Press, 141-160.

Ioannidis, John P. A. 2005. Why Most Published Research Findings Are False. *PLoS medicine, 2*(8): 696-701.

Koopman, Bernard O. 1940. The Axioms and Algebra of Intuitive Probability. *Annals of Mathematics, 41*(2): 269-292.

Mayo, Deborah G. 1983. An Objective Theory of Statistical Testing. *Synthese, 57*(3): 297-340.

Mayo, Deborah G. 1996. *Error and the Growth of Experimental Knowledge*. Chicago: The University of Chicago Press.

Mayo, Deborah G. 2018. *Statistical Inference as Severe Testing: How to Get Beyond the Statistical Wars*. Cambridge: Cambridge University Press.

Mayo, Deborah G. and Aris Spanos. 2004. Methodology in Practice: Statistical Misspecification Testing. *Philosophy of Science, 71*(5): 1007-1025.

Mayo, Deborah G. and Aris Spanos. 2006. Severe Testing as a Basic Concept in a Neyman–Pearson Philosophy of Induction. *British Journal for the Philosophy of Science, 57*(2): 323-357.

Mayo, Deborah G. and Aris Spanos. 2011. Error Statistics. In D. Gabbay, P. Thagard, and J. Woods (eds), *Philosophy of Statistics, Handbook of Philosophy of Science*. New York: Elsevier, 151-196.

Neyman, Jerzy. 1952. *Lectures and Conferences on Mathematical Statistics and Probability,* 2nd ed. Washington: U.S. Department of Agriculture.

Royall, Richard. 1997. *Statistical Evidence: A Likelihood Paradigm*. London: Chapman & Hall.

Salmon, Wesley C. 1967. *The Foundations of Scientific Inference*. Pittsburgh: University of Pittsburgh Press.

Spanos, Aris. 2013. A Frequentist Interpretation of Probability for Model-Based Inductive Inference. *Synthese, 190*(9):1555-1585.

Spanos, Aris. 2017. Why the Decision-Theoretic Perspective Misrepresents Frequentist Inference. In *Advances in Statistical Methodologies and Their Applications to Real Problems*. `http://dx.doi.org/10.5772/65720`, 3-28.

Spanos, Aris. 2018. Mis-Specification Testing in Retrospect. *Journal of Economic Surveys, 32*(2): 541-577.

Spanos, Aris. 2019. *Probability Theory and Statistical Inference: Empirical Modeling with Observational Data,* 2nd ed. Cambridge: Cambridge University Press.

Von Mises, Richard. 1928. *Probability, Statistics and Truth*, 2nd ed. New York: Dover.

Williams, David. 2001. *Weighing the Odds: A Course in Probability and Statistics*. Cambridge: Cambridge University Press.