IV On a New Philosophy of Frequentist Inference Exchanges with David Cox and Deborah G. Mayo

Aris Spanos

1. *Experimental Reasoning and Reliability*: How can methods for controlling long-run error probabilities be relevant for inductive inference in science? How does one secure error reliability in statistical inference? How do statistical inferences relate to substantive claims and theories?
2. *Objectivity and Rationality*: What is objectivity in statistical inference? What is the relationship between controlling error probabilities and objectivity? Can frequentist statistics provide an account of inductive inference? Is a genuine philosophy for frequentist inference possible? Does model validation constitute illegitimate double use of data?

1 Introduction

The twin papers by Cox and Mayo in this chapter constitute a breath of fresh air in an area that has long suffered from chronic inattention. A renowned statistician and a well-known philosopher of science and statistics have joined forces to grapple with some of the most inveterate foundational problems that have bedeviled frequentist statistics since the 1950s. The end result is more than a few convincing answers pertaining to some of these chronic problems, which are rarely discussed explicitly in either statistics or philosophy of science. Although here I give my own conception of what this amounts to, it is my hope that others recognize that the twin papers put forward some crucial steps toward providing a genuine philosophy for frequentist inference.

I need to profess at the outset that I am *not* an uninterested outsider to these philosophical issues and discussions. I learned my statistics as an undergraduate at the London School of Economics (LSE) from Cox and Hinkley (1974); and over the years, both as a graduate student at LSE and later, as a practicing econometrician, I have been greatly influenced

by Cox's books and papers (Hand and Herzberg, 2005). What appealed to me most in his writings was a dauntless inclination to raise as well as grapple with fundamental issues in statistical modeling and inference as they arise at the level of a practitioner, offering what seemed to me right-headed suggestions guided by a discerning intuition. From his writings, I learned to appreciate several subtle issues and problems in statistical modeling, including the value of preliminary data analysis and graphical techniques in learning from data, the importance of model adequacy, the difficulties in fusing statistical and substantive information, the distinctness of Fisher's inductive reasoning, and the perplexing variety of statistical testing. The challenge to seek more systematic accounts for some of his right-headed suggestions in a unifying framework fascinated me, and that objective greatly influenced my research agenda over the years.

Over the past eight years or so, I have collaborated with Mayo on several projects pertaining to various foundational issues. Our discussions often overlapped with some of the issues Mayo and Cox were grappling with, and I was aware of the exchanges that eventually gave rise to these twin papers. Most intriguing for me was the sense of discovery at how the frequentist procedures and principles Cox had long developed primarily on intuitive grounds obtained their most authentic and meaningful justification within a unified philosophy of learning from data. In what follows, I try to articulate that sense of discovery.

What is particularly interesting (and unique) about these twin papers is that they show us the weaving together of threads from Cox's perspective on frequentist inference, which has a distinct Fisherian undertone, and Mayo's error-statistical perspective, which appears to enjoy more affinity with the Neyman-Pearson (N-P) framework. The end result is a harmonious blend of the Fisherian and N-P perspectives to weave a coherent frequentist inductive reasoning anchored firmly on error probabilities. The key to this reconciliation is provided by recognizing that Fisher's p-value reasoning is based on a *post-data* error probability, and Neyman and Pearson's type I and II error reasoning is based on *pre-data* error probabilities. In the coalescing, both predata and post-data error probabilities fulfill crucial complementary roles, contrary to a prevailing view of critics, e.g. Savage (1962a).

Some of the particular problems and issues discussed in 7(I), (II), (III) include: frequentist inductive reasoning as it relates to the p-value and the accept/reject rules, the relevant error probabilities – especially when selection effects are involved, the different roles of conditioning in frequentist inference, model adequacy and its relationship to sufficiency, the use and abuse of the likelihood principle, objectivity in frequentist and Bayesian

statistics, and Bayesian criticisms of frequentist inference. In addition, these papers raise interesting issues pertaining to the Popperian philosophy of science as it relates to frequentist inductive reasoning.

2 Statistics and Philosophy of Science since the 1950s: A Bird's-Eye View

To do even partial justice to the joint papers by Mayo and Cox, I need to place them in the proper context, which includes both statistics and philosophy of science, and, by necessity, I have to use very broad brushstrokes to avoid a long digression; I apologize for that at the outset.

The modern approach to frequentist (classical) statistics was pioneered by Fisher (1922) as model-based statistical induction, anchored on the notion of a *statistical model.* Fisher (1925, 1934), almost single-handedly, erected the current theory of "optimal" point estimation and formalized *p*-value significance testing. Neyman and Pearson (1933) proposed an "optimal" theory for hypothesis testing by modifying and extending Fisher's significance testing. Neyman (1937) proposed an "optimal" theory for interval estimation analogous to N-P testing.

Broadly speaking, the probabilistic foundations of frequentist statistics and the technical apparatus associated with statistical inference methods were largely in place by the late 1930s, but the philosophical foundations associated with the proper form of the underlying *inductive reasoning* were rather befuddled. Fisher was arguing for "inductive inference," spearheaded by his significance testing in conjunction with *p*-values and his fiducial probability for interval estimation. Neyman was arguing for "inductive behavior" based on N-P testing and confidence interval estimation in conjunction with predata error probabilities (see Mayo, 2006).

The last exchange between these pioneers of frequentist statistics took place in the mid 1950s (see Fisher, 1955; Neyman, 1956; Pearson, 1955) and left the philosophical foundations of the field in a state of confusion with many more questions than answers: What are the differences between a Fisher significance test and an N-P test? Does a proper test require the specification of an alternative hypothesis? What about goodness-of-fit tests like Pearson's? Are the notions of type II error probability and power applicable to Fisher-type tests? What about the use of error probabilities postdata? Is the *p*-value a legitimate error probability? What is the relationship between *p*-values and posterior probabilities? Does Fisher's fiducial distribution give rise to legitimate error probabilities? Can one distinguish between different values of the unknown parameter within an observed confidence interval

(CI)? Can one infer substantive significance from an observed CI? In what sense does conditioning on an ancillary statistic enhance the precision and data specificity of inference?

In addition to these questions, it was not at all obvious under what circumstances an N-P tester could safeguard the coarse *accept/reject decisions* against:

1. the *fallacy of acceptance*: interpreting *accept* H_0 [*no evidence against* H_0] as *evidence for* H_0, or
2. the *fallacy of rejection*: interpreting *reject* H_0 [*evidence against* H_0] as *evidence for* H_1.

A well-known example of the latter is the conflation of *statistical* with *substantive significance.*

Fisher's use of the p-value to reflect the "strength of evidence" against the null was equally susceptible to the fallacy of rejection because the p-value often goes to zero as the sample size $n \to \infty$.

Moreover, interpreting a p-value that is *not* "small enough" as evidence for H_0 would render it susceptible to the fallacy of acceptance.

The subsequent literature on frequentist statistics sheds very little additional light on these philosophical/foundational issues. The literature in philosophy of science overlooked apparent connections between the Popperian and Fisherian versions of falsification and more or less ignored the important developments in frequentist statistics.[1] In direct contrast to the extensive use of statistics in almost all scientific fields, by the early 1950s, logical empiricism had adopted combinations of hypothetico-deductive and Bayesian perspectives on inductive inference, with Carnap's confirmatory logics (logical relations between statements and evidence) dominating the evidential accounts in philosophy of science (see Neyman's [1957] reply to Carnap).

Not surprisingly, because of the absence of genuine guidance from statistics or philosophy of science, the practitioners in several disciplines, such as epidemiology, psychology, sociology, economics, and political science, came up with their own "pragmatic" ways to deal with the philosophical puzzles bedeviling the frequentist approach. This resulted in a hybrid of the Fisher and N-P accounts, criticized as "inconsistent from both perspectives and burdened with conceptual confusion" (Gigerenzer, 1993, p. 323). That these methods were open to unthinking use and abuse made them a

[1] Exceptions include Giere (1969), Hacking (1965), Seidenfeld (1979), philosophical contributors to Harper and Hooker (1976), and Godambe and Sprott (1971).

convenient scapegoat for the limits and shortcomings of several research areas, a practice that continues unabated to this day (see references). In my own field of economics, Ziliak and McCloskey (2008) propose their own "economic" way to deal with the statistical versus substantive significance problem; see Spanos (2008) for a critical review.

By the early 1960s, disagreements about the philosophical underpinnings of frequentist inductive reasoning were increasingly taken as prima facie evidence that it failed to provide a genuine account for inference or evidence. This encouraged the supposition of the philosophical superiority of Bayesian inference, which does away with error probabilities altogether and upholds foundational *principles* like the *likelihood* and *coherency*. Despite the vast disagreements between different Bayesian schools, this impression continues to be reiterated, largely unchallenged in both statistics (see Ghosh et al., 2006) and philosophy of science (see Howson and Urbach, 1993).

3 Inductive Reasoning in Frequentist Statistics

A pivotal contribution of Mayo and Cox is a general *frequentist principle for inductive reasoning*, which they motivate as a modification and extension of the *p*-value reasoning:

FEV (i): data $\mathbf{z}_0$ provides (strong) evidence against the null H_0 (*for* a discrepancy from H_0), if and only if (iff) the *p*-value, $\mathrm{P}(d(\boldsymbol{Z}) > d(\mathbf{z}_0); H_0) = p(\mathbf{z}_0)$ is very low or, equivalently, $\mathrm{P}(d(\boldsymbol{Z}) \leq d(\mathbf{z}_0); H_0) = (1 - p(\mathbf{z}_0))$ is very high.

Corollary. Data $\mathbf{z}_0$ do *not* provide (strong) evidence against H_0, if $\mathrm{P}(d(\boldsymbol{Z}) > d(\mathbf{z}_0); H_0) = p(\mathbf{z}_0)$ is *not* very low.

This is a formal version of our "minimal scientific principle for evidence" (p. 3). The question that naturally arises is whether the aforementioned conditions relating to the *p*-value can be strengthened enough to avoid the fallacies of acceptance and rejection. The answer provided by Mayo and Cox is that, in cases where one can quantify departures from H_0 using a discrepancy parameter $\gamma \geq 0$, one can strengthen the corollary to guard against the fallacy of acceptance in the following form:

FEV(ii): A moderate $p(\mathbf{z}_0)$-value is evidence of the absence of a discrepancy γ from H_0, only if $\mathrm{P}(d(\boldsymbol{Z}) > d(\mathbf{z}_0); \mu_0 + \gamma)$ is very high.

What about the fallacy of rejection? It is well known that a very low *p*-value establishes the existence of some discrepancy $\gamma \geq 0$ from H_0 but provides

no information concerning the magnitude of γ licensed by data $\mathbf{z}_0$. This magnitude can be established using an obvious modification of FEV(ii) to strengthen FEV(i) in safeguarding it against the fallacy of rejection:

FEV(iii): A very low $p(\mathbf{z}_0)$-value is evidence *for* a discrepancy $\gamma \geq 0$ from H_0, only if $\mathrm{P}(d(\boldsymbol{Z}) \leq d(\mathbf{z}_0); \mu_0 + \gamma)$ is very high.

The preceding principles constitute crucial extensions of post-data frequentist inductive reasoning in cases where one can quantify departures from H_0 using a *discrepancy parameter* γ. Under such circumstances, the FEV(ii) and FEV (iii) rules can be seen as special cases of the *severity evaluations* associated with N-P accept/reject decisions (Mayo 1996, Mayo and Spanos, 2006):

$\mathrm{SEV}(T_\alpha; \mathbf{z}_0; \mu \leq \mu_1) = \mathrm{P}(d(\boldsymbol{Z}) > d(\mathbf{z}_0); \mu > \mu_1)$, for $\mu_1 = \mu_0 + \gamma, \gamma \geq 0$,
$\mathrm{SEV}(T_\alpha; \mathbf{z}_0; \mu > \mu_1) = \mathrm{P}(d(\boldsymbol{Z}) \leq d(\mathbf{z}_0); \mu \leq \mu_1)$, respectively.

At first sight these evaluations give the impression that they stem exclusively from the Neyman-Pearson (N-P) testing perspective because they remind one of the evaluation of power and the probability of type II error. This first impression is misleading, however, because on closer examination the severity evaluations draw from both the N-P and Fisherian perspectives on testing; they constitute a harmonious reconciliation of the two that can be used to address several of the questions mentioned in Section 2. Like the p-value, but unlike the type II error probability and power, the severity evaluations constitute *post-data error probabilities.* They involve events in the sample space denoting *lesser* (or *greater*) accordance with H_0 than $\mathbf{z}_0$ is. Like the type II error probability and power, the severity evaluations involve scenarios with specific discrepancies from the null, $\mu_1 = \mu_0 + \gamma$ (for some $\gamma \geq 0$), but, unlike them, the emphasis here is on evaluating the *post-data capacity* of the test in question. Notwithstanding Fisher's rhetoric against type II errors and power (see Fisher, 1955), enough evidence exists to suggest that he also viewed the optimality of tests in terms of their capacity (sensitivity) to detect discrepancies from the null hypothesis: "By increasing the size of the experiment, we can render it more sensitive, meaning by this that it will allow of the detection of... a quantitatively smaller departure from the null hypothesis" (Fisher, 1935, pp. 21–22).

This emphasis on capacity to detect discrepancies is exactly what is needed to provide an evidential construal of frequentist tests when combined with the following principle:

Severity Principle (SP): Data $\mathbf{z}_0$ do *not* provide good evidence for hypothesis H (H_0 or H_1) if $\mathbf{z}_0$ is used in conjunction with a test procedure that

Table 7.1. *Simple Normal Model*

$Z_t = \mu + u_t, \quad t \in \mathbb{N},$
[1] Normality: $Z_t \sim N(.,.)$,
[2] Constant mean: $E(Z_t) = \mu$,
[3] Constant variance: $\text{Var}(Z_t) = \sigma^2$,
[4] Independence: $\{Z_t, \ t \in \mathbb{N}\}$ is an independent process.

has a very low capacity to uncover discrepancies from H when present (see Mayo, 1996).

4 Model Adequacy

Another largely neglected area of frequentist statistics that Cox and Mayo discuss is that of *model adequacy*. The issues of statistical model specification and adequacy can be traced back to Cox's (1958) paper and constitutes a recurring theme in many of his writings, including Cox and Hinkley (1974) and Cox (1990, 2006).

4.1 Model Adequacy and Error Statistics

Since Fisher (1922) it has been known, but not widely appreciated, that the reliability of inductive inference depends crucially on the validity of the prespecified statistical model, which is generically denoted by $\mathcal{M}_\theta(\mathbf{z}) = \{f(\mathbf{z};\boldsymbol{\theta}), \boldsymbol{\theta}\in\Theta\}$, $\mathbf{z} \in \mathbb{R}^n_z$, where $f(\mathbf{z};\boldsymbol{\theta})$ denotes the distribution of the sample $\mathbf{Z}:=(Z_1, Z_2, \ldots, Z_n)$, and Θ and $\mathbb{R}^n_z$ denote the parameter and sample spaces, respectively. The statistical model is chosen to render the data $\mathbf{z}_0:=(z_1, z_2, \ldots, z_n)$ a "typical realization" of the stochastic process $\{Z_t, \ t \in \mathbb{N}:=(1, 2, \ldots, n \ldots)\}$, whose probabilistic structure is parameterized by $\mathcal{M}_\theta(\mathbf{z})$. A crucial first step in assessing model adequacy is to be able to specify a statistical model in terms of a complete set of probabilistic assumptions that are testable vis-à-vis data $\mathbf{z}_0$ (see Spanos, 1999).

The quintessential statistical model is given in Table 7.1.

Statistical model adequacy – that the assumptions underlying a statistical model $\mathcal{M}_\theta(\mathbf{z})$ (e.g., conditions [1]–[4] in Table 7.1) are valid for data $\mathbf{z}_0$ – is crucially important for frequentist inference because it secures *error reliability* – the nominal and actual error probabilities coincide (approximately) – which is fundamental in learning from data and is an important component of objectivity. Departures from model assumptions give rise to *error unreliability* – divergences between nominal and actual error probabilities – leading inductive inferences astray. Spanos (2005) showed that

even seemingly minor departures (e.g., the presence of correlation, say, $\rho = .1$, instead of assumption [4]), can give rise to sizeable divergences between nominal (.05) and actual (.25) error probabilities.

4.2 Misspecification (M-S) testing

The idea underlying model validation is to construct M-S tests using "distance" functions (test statistics) whose distribution under:

H_0: the probabilistic assumptions constituting $\mathcal{M}_\theta(\mathbf{z})$ hold for data $\mathbf{z}_0$,

is known, and at the same time they have adequate power against potential departures from the model assumptions. The logic of M-S testing is the one underlying Fisher's significance testing (Mayo and Cox, p. 259), where one identifies a test statistic $d(\boldsymbol{Z})$ to measure the distance between what is expected under H_0 with $d(\mathbf{z}_0)$. When the relevant p-value, $\mathrm{P}(d(\boldsymbol{Z}) > d(\mathbf{z}_0); H_0 \text{ true}) = p(\mathbf{z}_0)$, is very small, then there is evidence of violations of the model assumption(s) entailed by H_0. If the p-value is not small enough, one is entitled to rule out only departures the test had enough capacity to detect. Mayo and Cox bring out the importance of combining *omnibus* and *directional* M-S tests because they shed different light on possible departures from the model assumptions.

4.3 Model Adequacy and Sufficiency

An important impediment to model validation using M-S testing has been the argument that it involves *illegitimate double-use of data*, a topic discussed in Chapter 4. The same data are used to draw inferences concerning $\boldsymbol{\theta}$ as well as test the validity of assumptions [1]–[4]. Although nobody can deny that M-S testing involves double use of data, the charge of illegitimacy can be challenged on several grounds (see Mayo and Spanos, 2004). Cox and Mayo offer a purely statistical argument based on sufficiency. This argument arises in cases where there exists a sufficient statistic $\boldsymbol{S}(\boldsymbol{Z})$ for $\boldsymbol{\theta}$, which gives rise to the following reduction:

$$\mathrm{f}(\mathbf{z};\boldsymbol{\theta}) \propto \mathrm{f}(\mathbf{z}|\mathbf{s}) \cdot \mathrm{f}(\mathbf{s};\boldsymbol{\theta}) \quad \text{for all } \mathbf{z} \in \mathbb{R}^n_z \tag{1}$$

In this reduction, the information in the data about the model is split into two parts, one $[f(\mathbf{s};\boldsymbol{\theta})]$ captures all the information assuming the model to be correct, and the other $[f(\mathbf{z}|\mathbf{s})]$ allows checking the adequacy of the model. By a simple *modus tollens* argument, $f(\mathbf{z};\boldsymbol{\theta})$ implies $f(\mathbf{z}|\mathbf{s})$; thus, any departure from $f(\mathbf{z}|\mathbf{s})$ implies that $f(\mathbf{z};\boldsymbol{\theta})$ is false. In the case of a simple Bernoulli

model, $f(\mathbf{z}|\mathbf{s})$ is a discrete uniform distribution, and in the case of the simple Poisson model, $f(\mathbf{z}|\mathbf{s})$ is a multinomial distribution with identical cell probabilities. In both of these cases the form of $f(\mathbf{z}|\mathbf{s})$ gives rise to testable restrictions, which can be used to assess the validity of the original model, $\mathcal{M}_\theta(\mathbf{z})$.

This result can be extended to other situations where the structure of $f(\mathbf{z};\boldsymbol{\theta})$ gives rise to reductions that admit statistics whose sampling distributions are free of $\boldsymbol{\theta}$.

4.4 Model Validation and the Role of Sufficiency and Ancillarity

Motivated by informal arguments of Mayo (1981) and Hendry (1995), Spanos (2007a) showed that under certain assumptions $f(\mathbf{z};\boldsymbol{\theta})$ can be reduced into a product of two components as in reduction (1), but now it involves a *minimal sufficient* $\boldsymbol{S}(\boldsymbol{Z})$ and a *maximal ancillary* statistic $\boldsymbol{R}(\boldsymbol{Z})$ for $\boldsymbol{\theta}$:

$$\mathrm{f}(\mathbf{z};\boldsymbol{\theta}) \propto \mathrm{f}(\mathbf{s}|\boldsymbol{\theta}) \cdot \mathrm{f}(\mathbf{r}) \quad \text{for all } (\mathbf{r},\mathbf{s}) \in \mathbb{R}^n_z. \tag{2}$$

This reduction is analogous to reduction (1), but here $\mathbf{s}$ is also independent of $\mathbf{r}$. The crucial argument for relying on $f(\mathbf{r})$ is that the probing for departures from $\mathcal{M}_\theta(\mathbf{z})$ is based on error probabilities that do not depend on the true $\boldsymbol{\theta}$.

Example. In the case of the simple Normal model with $\boldsymbol{\theta} = (\mu, \sigma^2)$ (Table 7.1), reduction (2) holds with

$$\mathbf{s} = \left(\bar{z} = (1/n)\sum_{t=1}^{n} z_t,\ \ s^2 = [1/(n-1)]\sum_{t=1}^{n}(z_t - \bar{z})^2\right),\ \ \mathbf{r} = (r_3, \ldots, r_n)$$

$$r_t = \frac{\sqrt{n}(z_t - \bar{z})}{s}, \qquad t = 3, 4, \ldots, n,$$

where $\mathbf{r}$ represents the *studentized residuals*. These results extend to the linear regression and hold approximately in many cases where asymptotic Normality is invoked.

From a methodological perspective, the separation in reductions (1) and (2) reflects the drastically different questions posed for inference and model adequacy purposes. The question posed for model adequacy purposes is as follows: "Do data $\mathbf{z}_0$ represent a truly typical realization of the stochastic mechanism specified by $\mathcal{M}_\theta(\mathbf{z})$?" The generic form of the model validation hypotheses is:

$$H_0: f^*(\mathbf{z}) \in \mathcal{M}_0(\mathbf{z}) \quad \text{vs.} \quad H_1: f^*(\mathbf{z}) \in [\mathcal{P}(\mathbf{z}) - \mathcal{M}_0(\mathbf{z})],$$

where $f^*(\mathbf{z})$ denotes the true (but unknown) distribution of the sample, and $\mathcal{P}(\mathbf{z})$ the set of all possible models that could have given rise to data $\mathbf{z}_0$. This form clearly indicates that M-S probing takes place *outside the boundaries* of $\mathcal{M}_\theta(\mathbf{z})$. Note that the generic alternative $[\mathcal{P}(\mathbf{z})-\mathcal{M}_\theta(\mathbf{z})]$ is intractable as it stands; thus, in practice one must consider different forms of departures from H_0, which can be as vague as a direction of departure or as specific as an encompassing model $\mathcal{M}_\varphi(\mathbf{z})$; $\mathcal{M}_\theta(\mathbf{z}) \subset \mathcal{M}_\varphi(\mathbf{z})$ (see Spanos, 1999).

In contrast, the questions posed by inferences concerning $\boldsymbol{\theta}$ take the model $\mathcal{M}_\theta(\mathbf{z})$ as given (adequate for data $\mathbf{z}_0$) and probe their validity *within the model's boundaries* (see Spanos, 1999).

On a personal note, I ascertained the crucial differences between testing *within* (N-P) and testing *outside the boundaries* (M-S) of a statistical model and ramifications thereof, after many years of puzzling over what renders Fisher's significance testing different from N-P testing. What eventually guided me to that realization was the "unique" discussion of testing in Cox's writings, in particular the exposition in chapters 3–6 of Cox and Hinkley (1974). I also came to appreciate the value of preliminary data analysis and graphical techniques in guiding and enhancing the assessment of model adequacy from Cox's writings. After years of grappling with these issues, the result is a unifying modeling framework wherein these techniques become indispensable facets of statistical modeling and inference (see Spanos, 1986).

5 Revisiting Bayesian Criticisms of Frequentist Statistics

Of special importance in Cox and Mayo (7(II)) is their use of the new philosophy of frequentist inference to shed light on earlier philosophical debates concerning frequentist versus Bayesian inference and more recent developments in objective (O) Bayesianism.

5.1 Revisiting the Likelihood Principle (LP)

A crucial philosophical debate concerning frequentist versus Bayesian inference began with Birnbaum's (1962) result claiming to show that the (strong) LP follows from Sufficiency Principle (SP) and the Weak Conditionality Principle (WCP); if frequentists wish to condition on $\mathbf{z}_0$, they are faced with either renouncing sufficiency or renouncing error probabilities altogether. Cox and Mayo counter this argument as follows:

> It is not uncommon to see statistics texts argue that in frequentist theory one is faced with the following dilemma, either to deny the appropriateness of conditioning on the precision of the tool chosen by the toss of a coin, or else to embrace the strong likelihood principle which entails that frequentist sampling distributions are

irrelevant to inference once the data are obtained. This is a false dilemma: Conditioning is warranted in achieving objective frequentist goals, and the conditionality principle coupled with sufficiency does not entail the strong likelihood principle. The "dilemma" argument is therefore an illusion. (Cox and Mayo, p. 298)

Indeed, in 7(III) Mayo goes much further than simply raising questions about the cogency of the LP for frequentist inference. She subjects Birnbaum's "proof" to a careful logical scrutiny. On logical grounds alone, she brings out the *fallacy* that shows that those who greeted Birnbaum's paper as a "landmark in statistics" (see Savage, 1962b) with skepticism had good cause to withhold assent. By and large, however, Birnbaum's result is taken at face value (see Berger and Wolpert, 1988). As such, arguments for conditioning are taken as arguments for LP, which is just a short step away from Bayesianism; Ghosh et al. (2006) argue:

Suppose you are a classical statistician and faced with this example [Welch's uniform] you are ready to make conditional inference as recommended by Fisher. Unfortunately, there is a catch. Classical statistics also recommends that inference be based on minimal sufficient statistics. These two principles, namely the conditionality principle (CP) and sufficiency principle (SP) together have a far reaching implication. Birnbaum (1962) proved that they imply one must then follow the likelihood principle (LP), which requires inference be based on the likelihood alone, ignoring the sample.... Bayesian analysis satisfies the LP since the posterior depends on the data only through the likelihood. Most classical inference procedures violate the LP. (p. 38)

Even frequentist statisticians who treated the WCP + S = LP equation with skepticism are legitimately challenged to give a principled ground for conditioning rather than using the unconditional error probabilities. When we consider the kinds of canonical examples that argue for conditioning, especially in light of FEV, it is not difficult to find a principled argument, and that alone attests to its value as a frequentist principle. According to Cox and Mayo, "[M]any criticisms of frequentist significance tests (and related methods) are based on arguments that overlook the avenues open in frequentist theory for ensuring relevancy of the sampling distribution on which p-values are to be based" (Chapter 7, p. 295). Indeed, one may explain the inappropriateness of the unconditional inference by appealing to the notion of *relevant error probabilities*, as elaborated in the twin Cox and Mayo papers, where "relevance" includes *error reliability* (stemming from model adequacy) and *inference specificity* (relating to the inference at hand).

5.2 Revisiting the Welch Example for Conditional Inference

I argue that the various examples Bayesians employ to make their case involve some kind of "rigging" of the statistical model so that it appears

as though embracing the conditionality principle (CP) is the only way out, when in fact other frequentist principles invariably allow extrication. To illustrate the general point, I consider the case of the Welch uniform, which seems the most realistic among these examples; Cox and Mayo discuss two other examples (pp. 295–7). In the Welch (1939) uniform case where $Z_k \sim U(\theta - .5, \theta + .5)$, the rigging stems from the fact that this distribution is *irregular* in that its support depends on the unknown parameter θ (see Cox and Hinkley, 1974, p. 112). This irregularity creates a constraint between θ and the data $\mathbf{z}_0$ in the sense that, whatever the data, $\theta \in A(\mathbf{z}_0) = [z_{[n]} - .5, z_{[1]} + .5]$, where $z_{[n]} = \max(z_1, \ldots, z_n)$ and $z_{[1]} = \min(z_1, \ldots, z_n)$. Hence, post-data, the unconditional sampling distribution $f(\hat{\theta}; \theta)$, $[\theta - .5 \leq \hat{\theta} \leq \theta + .5]$, where $\hat{\theta} = [(Z_{[n]} + Z_{[1]})/2]$ is an estimator of θ, ignores the support information $\theta \in A(\mathbf{z}_0)$, and thus a confidence interval may include *infeasible values* of θ. The CP argument suggests that the conditional distribution $f(\hat{\theta} \mid R; \theta)$, where $R = (Z_{[n]} - Z_{[1]})$ is an ancillary statistic, gives rise to much better inference results (see Berger and Wolpert, 1988; Ghosh et al., 2006; Young and Smith, 2005). Notwithstanding such claims, Cox and Hinkley (1974, pp. 220–1) showed that this conditional inference is also *highly problematic* because $f(\hat{\theta} \mid R; \theta)$ has *no* discriminatory capacity because it is uniform over $[-.5(1 - R), .5(1 - R)]$.

Using the notion of relevant error probabilities, Spanos (2007b) showed that, *post-data*, the truncated sampling distribution $f(\hat{\theta} \mid A(\mathbf{z}_0); \theta)$, for $\theta \in A(\mathbf{z}_0)$, provides the relevant basis for inference because it accounts for the deterministic support information $\theta \in A(\mathbf{z}_0)$ – created by the irregularity of the model – without sacrificing the discriminatory capacity of $f(\hat{\theta}; \theta)$.

Learning from data can only result from using *relevant* error probabilities in the sense that they reflect faithfully (in an error-reliable sense) the mechanism that actually generated the data $\mathbf{z}_0$.

5.3 Objective Bayesian Perspective

It is refreshing to see Cox and Mayo give a hard-nosed statement of what scientific objectivity demands of an account of statistics, show how it relates to frequentist statistics, and contrast that with the notion of "objectivity" used by O-Bayesians (see Berger, 2004). They proceed to bring out several weaknesses of this perspective. The question that naturally arises from their discussion is "if one renounces the likelihood, the stopping rule, and the coherence principles, marginalizes the use of prior information as largely untrustworthy, and seeks procedures with 'good' error probabilistic properties (whatever that means), what is left to render the inference Bayesian,

apart from a *belief* (misguided in my view) that the only way to provide an evidential account of inference is to attach probabilities to hypotheses?"

6 Error Statistics and Popperian Falsification

The philosophy of frequentist inference growing out of 7(I) bears much fruit in 7(II). Among the most noteworthy is the stage it sets for philosophical debates concerning Popperian falsification. Although recognizing the importance of a "new experimentalist" focus on the details of obtaining, modeling, and making inferences about data, a reluctance exists among Popperian philosophers of science to see that full-fledged ampliative or inductive inferences are involved. Part of the reason is the supposition that inductive inference is a matter of assigning post-data degrees of probability or belief to hypotheses (Musgrave, Achinstein, this volume), whereas a true Popperian ("progressive" Popperian, as Mayo calls him) would insist that what matters is not *highly probable*, but rather *highly probed* hypotheses resulting from severe tests – genuine attempts at refutation using tests with enough capacity to detect departures. In a sense, the error-statistical interpretation of tests in these joint papers offers solutions to Popper's problems about deductive "falsification" and "corroboration," wherein, by essentially the same logic as that of statistical falsification, one may warrant claims that pass severe tests.

This suggestion is importantly different from using *p*-values and other error probabilities as post-data degrees of confirmation, support, or the like – in contrast with the most familiar "evidential" interpretations of frequentist methods, or the well-known O-Bayesian "reconciliations" (see Berger, 2003). Where in philosophy of science we might say this enables us to move away from what Musgrave (in this volume) calls "justificationist" approaches, in statistics, it allows us to answer those skeptical of regarding "frequentist statistics as a theory of inductive inference."

7 Conclusion

My hope is that I was able to convey to the reader the sense of discovery that I felt in coming to see how several of the frequentist procedures and principles developed by Cox over many years obtained their philosophical justification within a coherent statistical framework, the central goal of which is "to extract what can be *learned from data* that can also be vouched for" (p. 278).

The series of exchanges between Cox and Mayo exemplifies the central goals of the "two-way street," wherein statistical ideas inform philosophical

debates and problems while philosophical and foundational analysis offer unification and justification for statistical methodology. In the back-and-forth exchanges that gave rise to 7(I) and 7(II), it is apparent that both scholars have moved away from their usual comfort zones to some degree; yet by taking the two papers together, the shared principles of the error-statistical philosophy begin to crystallize and the essential pieces of the puzzle fall into place.

Aside from their specific contributions, the Cox and Mayo papers are noteworthy for taking some crucial steps toward legitimating the philosophy of frequentist statistics as an important research program in its own right. I hope that others, statisticians and philosophers of science, will join them in what may properly be described as an expansion of contemporary work in the foundations of statistics.

References

Berger, J. (2003), "Could Fisher, Jeffreys and Neyman Have Agreed on Testing?" *Statistical Science*, 18: 1–12.

Berger, J. (2004), "The Case for Objective Bayesian Analysis," *Bayesian Analysis*, 1: 1–17.

Berger, J., and Wolpert, R. (1988), *The Likelihood Principle*, 2nd ed., Institute of Mathematical Statistics, Hayward, CA.

Birnbaum, A. (1962), "On the Foundations of Statistical Inference" (with discussion), *Journal of the American Statistical Association*, 57: 269–306.

Cox, D.R. (1958), "Some Problems Connected with Statistical Inference," *Annals of Mathematical Statistics*, 29: 357–72.

Cox, D.R. (1990), "Role of Models in Statistical Analysis," *Statistical Science*, 5: 169–74.

Cox, D.R. (2006), *Principles of Statistical Inference*, Cambridge University Press, Cambridge.

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

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

Fisher, R.A. (1925), "Theory of Statistical Estimation," *Proceedings of the Cambridge Philosophical Society*, 22: 700–25.

Fisher, R.A. (1934), "Two New Properties of Maximum Likelihood," *Proceedings of the Royal Statistical Society A*, 144: 285–307.

Fisher, R.A. (1935), *The Design of Experiments*, Oliver and Boyd, Edinburgh.

Fisher, R.A. (1955) "Statistical Methods and Scientific Induction," *Journal of the Royal Statistical Society B*, 17: 69–78.

Ghosh, J.K., Delampady, M., and Samanta, T. (2006), *An Introduction to Bayesian Analysis: Theory and Methods*, Springer, New York.

Giere, R.N. (1969), "Bayesian Statistics and Biased Procedures," *Synthese*, 20: 371–87.

Gigerenzer, G. (1993), "The Superego, the Ego, and the Id in Statistical Reasoning," pp. 311–39 in G. Keren and C. Lewis (eds.), *A Handbook of Data Analysis in the Behavioral Sciences: Methodological Issues*, Lawrence Erlbaum Associates, Hillsdale, NJ.

Godambe, V.P., and D.A. Sprott (eds.) (1971), *Foundations of Statistical Inference: a Symposium*, Holt, Rinehart and Winston, Toronto.

Hacking, I. (1965), *Logic of Statistical Inference*, Cambridge University Press, Cambridge.

Hand, D.J., and Herzberg, A.M. (2005), *Selected Statistical Papers of Sir David Cox*, vols. 1–2, Cambridge University Press, Cambridge.

Harper, W.L., and C.A. Hooker (eds.) (1976), *Foundations of Probability Theory, Statistical Inference, and Statistical Theories of Science, Vol. II: Foundations and Philosophy of Statistical Inference*, Reidel, Dordrecht, The Netherlands.

Hendry, D.F. (1995), *Dynamic Econometrics*, Oxford University Press, Oxford.

Howson, C., and Urbach, P. (1993), *Scientific Reasoning: The Bayesian Approach*, 2nd ed., Open Court, Chicago.

Mayo, D.G. (1981), "Testing Statistical Testing," pp. 175–230 in J. Pitt (ed.), *Philosophy in Economics*, D. Reidel, Dordrecht.

Mayo, D.G. (1996), *Error and the Growth of Experimental Knowledge*, University of Chicago Press, Chicago.

Mayo, D.G. (2005), "Philosophy of Statistics," pp. 802–15 in S. Sarkar and J. Pfeifer (eds.), *Philosophy of Science: An Encyclopedia*, Routledge, London.

Mayo, D.G., and Spanos, A. (2004), "Methodology in Practice: Statistical Misspecification Testing," *Philosophy of Science*, 71: 1007–25.

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

Neyman, J. (1937), "Outline of a Theory of Statistical Estimation Based on the Classical Theory of Probability," *Philosophical Transactions of the Royal Statistical Society of London A*, 236: 333–80.

Neyman, J. (1956), "Note on an Article by Sir Ronald Fisher," *Journal of the Royal Statistical Society B*, 18: 288–94.

Neyman, J. (1957), "Inductive Behavior as a Basic Concept of Philosophy of Science," *Revue de L'Institut International de Statistique*, 25: 7–22.

Neyman, J., and Pearson, E.S. (1933), "On the Problem of the Most Efficient Tests of Statistical Hypotheses," *Philosophical Transactions of the Royal Society A*, 231: 289–337.

Pearson, E.S. (1955), "Statistical Concepts in Their Relation to Reality," *Journal of the Royal Statistical Society B*, 17: 204–7.

Savage, L., ed. (1962a), *The Foundations of Statistical Inference: A Discussion*. Methuen, London.

Savage, L. (1962b), "Discussion" on Birnbaum (1962), *Journal of the American Statistical Association*, 57: 307–8.

Seidenfeld, T. (1979), *Philosophical Problems of Statistical Inference: Learning from R.A. Fisher*, Reidel, Dordrecht, The Netherlands.

Spanos, A. (1986), *Statistical Foundations of Econometric Modelling*, Cambridge University Press, Cambridge.

Spanos, A. (1999), *Probability Theory and Statistical Inference: Econometric Modeling with Observational Data*, Cambridge University Press, Cambridge.

Spanos, A. (2005), "Misspecification, Robustness and the Reliability of Inference: The Simple t-Test in the Presence of Markov Dependence," Working Paper, Virginia Tech.

Spanos, A. (2007a), "Sufficiency and Ancillarity Revisited: Testing the Validity of a Statistical Model," Working Paper, Virginia Tech.

Spanos, A. (2007b), "Revisiting the Welch Uniform Model: A Case for Conditional Inference?" Working Paper, Virginia Tech.

Spanos, A. (2008), "Stephen Ziliak and Deirdre McCloskey's The Cult of Statistical Significance: How the Standard Error Costs Us Jobs, Justice, and Lives," *Erasmus Journal for Philosophy and Economics*, 1: 154–64.

Welch, B.L. (1939), "On Confidence Limits and Sufficiency, and Particular Reference to Parameters of Location," *Annals of Mathematical Statistics*, 10: 58–69.

Young, G.A., and Smith, R.L. (2005), *Essentials of Statistical Inference*, Cambridge University Press, Cambridge.

Ziliak, S.T., and McCloskey, D.N. (2008), *The Cult of Significance: How the Standard Error Costs Us Jobs, Justice and Lives*, University of Michigan Press, Ann Arbor, MI.

Related Exchanges

Bartz-Beielstein, T. (2008), "How Experimental Algorithmics Can Benefit From Mayo's Extensions To Neyman-Pearson Theory Of Testing," *Synthese (Error and Methodology in Practice: Selected Papers from ERROR 2006)*, vol. 163(3): 385–96.

Casella, G. (2004), "Commentary on Mayo," pp. 99–101 in M.L. Taper and S.R. Lele (eds.), *The Nature of Scientific Evidence: Statistical, Philosophical, and Empirical Considerations*, University of Chicago Press, Chicago.

Mayo, D.G. (2003), "Severe Testing as a Guide for Inductive Learning," pp. 89–118 in H. Kyburg, Jr. and M. Thalos (eds.), *Probability Is the Very Guide of Life: The Philosophical Uses of Chance*, Open Court, Chicago and La Salle, Illinois.

Mayo, D.G. (2004), "An Error-Statistical Philosophy of Evidence," pp. 79–97 and "Rejoinder" pp. 101–15, in M.L. Taper and S.R. Lele (eds.), *The Nature of Scientific Evidence: Statistical, Philosophical, and Empirical Considerations*, University of Chicago Press, Chicago.

Taper, M.L. and Lele, S.R. (eds.) (2004), *The Nature of Scientific Evidence: Statistical, Philosophical, and Empirical Considerations*, University of Chicago Press, Chicago.

Additional References on the Significance Test Debates (see webpage for updates)

Altman, D.G., Machin D., Bryant T.N., and Gardner M.J. (2000), *Statistics with Confidence*, (eds.), British Medical Journal Books, Bristol.

Barnett, V. (1982), *Comparative Statistical Inference*, 2nd ed., John Wiley & Sons, New York.

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

Harlow L.L., Mulaik, S.A., and Steiger, J.H. (1997) *What If There Were No Significance Tests?* Erlbaum, Mahwah, NJ.

Lieberman, B. (1971), *Contemporary Problems in Statistics: a Book of Readings for the Behavioral Sciences*, Oxford University Press, Oxford.

Morrison, D.E., and Henkel, R.E. (1970), *The Significance Test Controversy: A Reader*, Aldine, Chicago.