The Logic of Severe Testing

Samuel C. Fletcher*

Department of Philosophy
University of Minnesota, Twin Cities
&
Munich Center for Mathematical Philosophy
Ludwig Maximilian University of Munich

August 7, 2017

1 Introduction

In the last decades of philosophy of science, the dominant approach to understanding confirmation—how do data support hypotheses?—and induction—what are rational rules for defeasible reasoning?—has been Bayesianism. Although I. J. Good famously quipped that there are more distinct varieties of Bayesianism than Bayesians, for present purposes the doctrine treats both confirmation and induction with degrees of belief, modeled as probabilities, changing through conditionalization. This is all well, as far as it goes, but a thoroughgoing naturalistic attitude about the philosophy of scientific methodology yields some pause: The Bayesian project, for all its heroic reconstructions, does not describe how most scientific studies, from data, extract support for conclusions and test the viability of hypotheses.[1] The tools they use, developed from the ideas of Fisher, Neyman, Pearson, and others, fall within classical statistics.

Philosophers have (characteristically?) not been kind to classical (i.e., so-called "frequentist") statistics. Its foundations have been roundly criticized, its apparent central concepts subject to withering counterexamples. One of its few steadfast philosophical defenders has been Deborah G. Mayo, who, over the past 20 years or so, has developed her own approach to solving these problems, dubbed *Error Statistics*. Yet, despite a monograph, a healthy blog following, and numerous articles—often co-authored with econometrician Aris Spanos—there remains confusion as to what, exactly, Error Statistics amounts to. For instance, George Casella has written that "Mayo outlines a 'framework of inquiry' that is compelling in its ideas, frustrating in its vagueness, and

*Thanks to Charlie Geyer and the Young Turks' Philosophy of Statistics Group, especially Conor Mayo-Wilson, for their comments on a previous draft.

[1] I do not wish to dismiss the considerable support that Bayesian statistical methods rightly enjoy, but rather show how it is possible to agitate for viable and productive reform without revolution.

unworkable in its rigidity" (Taper and Lele, 2004, 100). Mayo and Spanos seem unwilling or uninterested both in expounding Error Statistics in a way accessible to statisticians and in considering applications of Error Statistics beyond the one-sided z-test. This has made Mayo's claims to have solved many of the main conceptual problems of classical statistics difficult to evaluate.

The purpose of this note is to begin to sketch how to construct a real, honest-to-goodness mathematical theory of *severe testing*, which one finds in Error Statistics. In a sense, though, my goal is more akin to Carnapian explication than exegesis, for my theory modifies and expands upon Mayo's while, I contend, preserving its central insights. In doing so, not only ought one be able to gain a perspective to better evaluate Mayo's claims and the foundations of classical statistics, but perhaps as well illuminate new corners of logic, traditional problems in philosophy of science, and connections with broader debates in epistemology.

Severe testing, as I shall try to reconstruct, provides a confirmation theory that, while steeped in probability, does not make confirmation itself probabilistic. This aspect of the theory expands upon the Fisherian evidential use of p-values through further evidential criteria partly inspired from Neyman-Pearson theory. On the back of this confirmation theory it erects a method for inductive reasoning according to which the "acceptance" of a proposition amounts to its confirmation criteria being sufficiently high (not unlike how the Lockean account of full belief can be built on Bayesian degress of belief).[2] Conversely, these criteria being sufficiently low for a proposition leads to its "rejection." By including the aforementioned evidential criteria in its inductive inference rules, it expands upon the Neyman-Pearson theory for statistical decisions to genuinely epistemic contexts.

The titular connection with logic arises because of this treatment of confirmation and induction. Reasoning within the language describing the hypotheses is governed by a particular sort of mathematical fuzzy logic, whose semantic valuations in $[0, 1]^n$ (for positive integers $n \leq 4$, depending on the version,) represent neither degrees of partial truth nor probabilities, but n sorts of degrees of compatibility of data with the hypothesis in question. Data sets are the semantic models for this logic. The inductive part "coarse-grains" the semantical values to just three: accept, reject, and neither.

2 The Formalism of Hypothesis Testing

Classical statistics describes support for hypotheses not through degrees of uncertain belief, but by their reliability, quantified by their (probabilistic) compatibility with data. The process of making this comparison, called hypothesis testing, depends not just on the data collected, but on data that could have been collected, were the hypothesis true. In this sense, classical statistics has affinities with reliabilist and externalist epistemology.

I shall return to some of these connection with epistemology below. First, here are some of the ingredients of hypothesis testing:

Data The set of possible data sets, $\mathcal{X}$, consists in all the possible recorded outcomes of the scientific study in question. For example, if the study consists in an n-fold sequential experiment, then $\mathcal{X}$ will consist in n-fold Cartesian product of sets $\mathcal{X}_i$ for $i \in \{1, \ldots, n\}$.

[2] Mayo accedes to the goal of induction but not confirmation, so on this aspect of interpretation I part ways with her.

Hypotheses Informally, the space of hypotheses Θ consists in possible states in which one is considering the world to be. They are mutually exclusive but not necessarily exhaustive. Formally, each $\theta \in \Theta$ determines a probability distribution over $\mathcal{X}$.

Tests Each element of the set of testing criteria, τ, measures in a particular way how well a data set coheres with a certain hypothesis $\theta \in \Theta$. Formally, each $T \in \tau$ is a random variable, i.e., a map from $\mathcal{X} \times \Theta$ into a measurable space $\mathcal{Y}$, such that for each $(x, \theta) \in (X, \Theta)$, the upper set $\{x' \in X : T(x', \theta) >_\theta T(x, \theta)\}$ is measurable, where data sets whose images under some T are smaller in the order $\leq_\theta$ would be more expected if the world were in state θ. (Often $\leq_\theta$ is the ordering on $\mathcal{Y}$ inverse to the ordering given by the probability density (or mass) function on $\mathcal{Y}$.)

In many applications, such generality is not needed. Here's an example: let $\mathcal{X} = \mathbb{R}^n$ and Θ consist in the hypotheses that each of the n components of $\mathcal{X}$ is independently and identically normally distributed, with arbitrary mean and variance. In other words, we can parameterize Θ as $\mathbb{R} \times (0, \infty)$ so that each $(\mu, \sigma^2) \in \Theta$ determines a collection of n random variables

$$X_i \sim_{iid} N(\mu, \sigma^2) \tag{1}$$

with probability density functions

$$P_{(\mu,\sigma^2)}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right). \tag{2}$$

Standard test criteria when one is interested in hypotheses about the mean of the normal distribution are the z- and t-statistics,[3]

$$z(x, \mu, \sigma^2) = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}, \tag{3}$$

$$t(x, \mu, \sigma^2) = \frac{\bar{x} - \mu}{s/\sqrt{n}}, \tag{4}$$

and their absolute values, $|z|$ and $|t|$, where

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i,$$

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(\bar{x} - x_i)^2,$$

are the sample mean and variance, respectively. These test criteria takes on values in $\mathcal{Y} = \mathbb{R}$, equipped with its standard ordering, which is independent of the value of (μ, σ^2) and measures how much a data set departs positively from the mean μ. The z-statistic is used when the variance σ^2 is known, while the t-statistic is used when it is not.

[3] A serious question, which as far as I know Mayo never considered, concerns the determination of the test statistics τ. This is because the sorts of conclusions drawn from severe testing, in the sense considered here, depend quite a bit on the choice. I hope to have more to say about this in future drafts.

When (μ, σ^2) is true, $z(x, \mu, \sigma^2) \sim N(0, 1)$ and $t(x, \mu, \sigma^2) \sim t(n-1)$, Student's t-distribution with $n-1$ degrees of freedom. That these distributions are independent of the (μ, σ^2) just means that z and t are pivotal quantities for these parameters. However, if some $(\mu', \sigma'^2) \neq (\mu, \sigma^2)$ is true instead, then $z(x, \mu, \sigma^2) \sim N(\mu' - \mu, 1)$ and $t(x, \mu, \sigma^2) \sim t(n-1, \delta)$, Student's non-central t-distribution with $n-1$ degrees of freedom and non-centrality parameter $\delta = (\mu' - \mu)/(\sigma'/\sqrt{n})$.

3 Measures of Evidence

3.1 Fit

Since $T(X, \theta)$ is a random variable, one can quantify the T-fit of data x with an hypothesis θ in terms of the probability of getting results as least as ill-fitting:

$$\text{FIT}(T, x, \theta) = P_\theta(T(X, \theta) >_\theta T(x, \theta)). \tag{5}$$

Describing this criterion in words brings out its connection with the *adherence* criterion sometimes used in epistemology for evidence:

> Data x is FIT T-evidence *for* θ to the extent that, if θ were true, then data less T-fitting to θ would likely be observed.

Usually the consequent of the adherence condition is formulated in terms of the probability of x, which is problematic in cases of continuous data, but since the T-fit to θ tracks the value of the probability density at x under θ, there is a sense in which the data is probable to the extent that data less T-fitting to θ are probable. This is the sense in which FIT is a quantitative measure of the adherence of a data set to an hypothesis. Moreover, smaller values of FIT indicate evidence against an hypothesis. In other words:

> Data x is FIT T-evidence *against* θ to the extent that, if θ were true, then data less T-fitting to θ would unlikely be observed.

Finally, middling values of FIT indicate equivocal evidence. It will be the work of the inductive component, yet to be described, to partition FIT values for an hypothesis into those necessary for acceptance, rejection, or neither.

The generalization of equation 5 to composite hypotheses, i.e., some $H \subseteq \Theta$ that is not just a singleton, is important and straightforward. To assert a composite hypothesis is just to assert the disjunction of its atomic (or "simple") constituents; the degree to which a data set fits with a composite hypothesis is just the highest degree to which it fits any of those atomic constituents. So in the general case:

$$\text{FIT}(T, x, H) = \sup_{\theta \in H} P_\theta(T(X, \theta) >_\theta T(x, \theta)). \tag{6}$$

Suppose, for example, one has data sets x_1 and x_2 with $\bar{x}_1 = 2$, $\bar{x}_2 = 6$, $s_1 = s_2 = 4$, and $n_1 = n_2 = 4$. Then, rounding to three decimal places,

$$\begin{aligned}
\text{FIT}(z, x_1, \mu = 0) &= \text{FIT}(z, x_1, \mu \leq 0) = 0.159 & \text{FIT}(z, x_2, \mu = 0) &= \text{FIT}(z, x_2, \mu \leq 0) = 0.001 \\
\text{FIT}(t, x_1, \mu = 0) &= \text{FIT}(t, x_1, \mu \leq 0) = 0.196 & \text{FIT}(t, x_2, \mu = 0) &= \text{FIT}(t, x_2, \mu \leq 0) = 0.029 \\
\text{FIT}(z, x_1, \mu = 2) &= \text{FIT}(z, x_1, \mu \leq 2) = 0.500 & \text{FIT}(z, x_2, \mu = 2) &= \text{FIT}(z, x_2, \mu \leq 2) = 0.023 \\
\text{FIT}(t, x_1, \mu = 2) &= \text{FIT}(t, x_1, \mu \leq 2) = 0.500 & \text{FIT}(t, x_2, \mu = 2) &= \text{FIT}(t, x_2, \mu \leq 2) = 0.070
\end{aligned}$$

Readers familiar with classical statistics may recognize that the T-FIT of an hypothesis θ with a data set x is just the p-value associated with the test statistic $T_\theta(x) = T(x, \theta)$. In this sense, severe testing incorporates the Fisherian perspective on p-values as measures of evidence. That perspective, however, has always warned that while sufficiently low p-values for a statistic (i.e., low FIT) may be sufficient to provide evidence *against* a hypothesis, sufficiently high p-values (i.e., high FIT) are only *necessary* (but *not* sufficient) to provide evidence *for* an hypothesis. In the above examples, for instance, a simple hypothesis may have a high FIT score for a reasonable test criterion. Yet it seems unjustified to say that that alone is strong evidence for it being true in particular: its truth would rule out all other simple hypotheses, even though some of these may have high FIT as well. A further criterion for evidence that addresses this concern, to which I turn in the next subsection, is the concept of severity.

3.2 Severity

The trouble with using FIT as the close criterion of evidential support for a hypothesis is that it ignores whether and to what extent the negation of the hypothesis is also supported. This is problematic if one wants to use sufficiently evidential support as grounds for inferring the truth of an hypothesis. The concept of severity is introduced exactly to take into account how well different hypotheses fit with the data comparatively.

For the sake of simplicity, first consider the case in which $\Theta = \{\theta, \theta'\}$. In order for the data x to truly support (say) θ, one must have that "with high probability, the test would have produced an outcome that fits θ less well than x does if θ were not the case (hence θ' were the case)" (Taper and Lele, 2004, 110):

$$\text{SEV}(T, x, \theta) = P_{\theta'}(T(X, \theta) >_\theta T(x, \theta)). \tag{7}$$

Describing this criterion in words brings out its connection with the *sensitivity* criterion sometimes used in epistemology for evidence:

> Data x is SEV T-evidence *for* θ to the extent that, if θ were not true, then data less T-fitting to θ would likely be observed.

In light of the identity of the consequents of FIT and SEV, one might wonder whether the latter makes sense as a criteria of evidence. The important difference between them is that while data less T-fitting to θ is less "probable" if θ is true, it is not necessarily so for $\theta' \neq \theta$. So one cannot in general interpret the event $T(X, \theta) >_\theta T(x, \theta)$ as one which is probable. Rather, data provide T-severe evidence for a hypothesis when it is likely that less expected data would have been likely if the hypothesis were false. Also unlike FIT, low values of SEV do not necessarily provide T-evidence against an hypothesis: it's perfectly compatible with the truth of an hypothesis that other mutually exclusive hypotheses entail a high probability of similar data.

Readers familiar with classical statistics may recognize the similarity between severity and the power of a test T of hypothesis θ of size α at θ',

$$\text{POW}(T, \theta, \alpha, \theta') = P_{\theta'}(T(X, \theta) >_\theta c_{\theta,\alpha}), \tag{8}$$

where $c_{\theta,\alpha}$ is the value of any $T(\theta, x)$ such that $\alpha = P_\theta(T(X, \theta) >_\theta T(x, \theta))$. The T-severity of x for θ is just the power of the test T of θ of size $T(\theta, x)$ at θ': $\text{SEV}(T, x, \theta) = \text{POW}(T, \theta, T(x, \theta), \theta')$. In this sense, severity is to power as p-values are to the size of a test: the latter are features of testing

procedures, while the former are their data-dependent versions. Severity is thus distinct from "post hoc" power, in which a data-based estimate of θ is substituted in for θ', and I see this as the central conceptual innovation of Error Statistics.

The generalization of equation 7 to composite hypotheses is also important but less straightforward. The first step is the case of the T-severity for a single θ when there are many distinct elements of Θ. In this case, there are many ways in which θ can be false, so in order for x to provide T-severe data for θ, less T-fitting data to θ would have to be likely observed for all these ways. Thus one takes the worst severity score when calculated amongst all the alternatives:

$$\text{SEV}(T, x, \theta) = \inf_{\theta' \in \Theta \setminus \{\theta\}} P_{\theta'}(T(X, \theta) >_\theta T(x, \theta)). \tag{9}$$

In this case, the T-severity of x for θ is the minimum power against all alternatives: $\text{SEV}(T, x, \theta) = \inf_{\theta' \in \Theta \setminus \{\theta\}} \text{POW}(T, \theta, T(x, \theta), \theta')$. Readers familiar with classical power analysis will thus see that, for many continuous spaces of hypotheses, a single $\theta \in \Theta$ will not be severely tested (have a sufficiently high T-SEV score) for a reasonable data set: all the hypotheses "close" to θ will yield similar results and SEV takes on the minimum value of the power curve, which is just the FIT of the hypothesis.

So, it is important to generalize equation 9 to composite hypotheses $H \subseteq \Theta$, which can be considered as a disjunction of its constituent elements. The most appropriate generalization seems at this point to be

$$\text{SEV}(T, x, H) = \inf_{\theta' \in \Theta \setminus H} P_{\theta'}(T(X, \hat{\theta}) >_{\hat{\theta}} T(x, \hat{\theta})), \tag{10}$$

where $\hat{\theta} \in H$ is such that $\text{FIT}(T, x, \hat{\theta}) = \text{FIT}(T, x, H)$, i.e., it is the best-fitting hypothesis amongst those in H.[4]

When the test criterion is z, the event $z(X, \mu, \sigma^2) >_{(\mu,\sigma^2)} z(X, \mu, \sigma^2)$ can be written as the event $\bar{X} > \bar{x}$, which does not depend on the values of (μ, σ^2). Mayo and Spanos (2011, p. 178) calculate SEV for a variety of data sets with $\sigma = 2$ and $n = 100$:

$$\begin{aligned}
&\text{SEV}(z, \bar{x} = 0.39, \mu \leq 0.2) = 0.171 && \text{SEV}(z, \bar{x} = 0.39, \mu \leq 0.6) = 0.853 \\
&\text{SEV}(z, \bar{x} = 0.30, \mu \leq 0.2) = 0.309 && \text{SEV}(z, \bar{x} = 0.30, \mu \leq 0.6) = 0.933 \\
&\text{SEV}(z, \bar{x} = 0.10, \mu \leq 0.2) = 0.691 && \text{SEV}(z, \bar{x} = 0.10, \mu \leq 0.6) = 0.995
\end{aligned}$$

However, when the test criterion is $|z|$, $z(X, \mu, \sigma^2) >_{(\mu,\sigma^2)} z(X, \mu, \sigma^2)$ can be written as the event $|\bar{X} - \mu| > |\bar{x} - \mu|$, which *does* depend on the value of μ.

This sort of case helps illustrate why certain initially plausible alternative generalizations are not satisfactory. For example, consider

$$\text{SEV}'(T, X, H) = \sup_{\theta \in H} \inf_{\theta' \in \Theta \setminus H} P_{\theta'}(T(X, \theta) >_\theta T(x, \theta)). \tag{11}$$

This alternative definition calculates the pointwise severity (equation 9) for each $\theta \in H$ and then returns their least upper bound. But then given any data with $\bar{x} < 0$, $\text{SEV}'(|z|, \bar{x}, \mu \leq 0) = 1$, while intuitively lower values of $\bar{x}$ should provide more severe evidence for $\mu \leq 0$. (Similar reasoning can be used to reject a version of severity with $\sup_{\theta \in H}$ replaced by $\inf_{\theta \in H}$.)

[4]Existence and uniqueness of such a best fit is a substantive assumption, but I believe it can be dropped for a more complicated definition involving limits of level sets (with respect to fir) of hypotheses.

However, SEV can still behave counterintuitively when one uses the test criterion t or $|t|$. Suppose that one is interested in the t-severity attributed to the hypothesis $\mu \leq \mu_0$ for some $\mu_0 \in \mathbb{R}$ by data x. As I stated before, for any values of μ and σ^2, $t \sim (n-1, \delta)$ for n data points, so the probability of data at least as t-extreme as that observed will be minimized when δ is minimized. Since in this case $\delta = (\mu' - \mu)/(\sigma'/\sqrt{n})$, this occurs as $\sigma' \to \infty$, for $\mu' - \mu > 0$. Consequently, the resulting (central) t-distribution yields the same probability as $P_{(\hat{\mu},\hat{\sigma})}(t(X,(\hat{\mu},\hat{\sigma})) >_{(\hat{\mu},\hat{\sigma})} t(x,(\hat{\mu},\hat{\sigma})))$, i.e., $\mathrm{SEV}(t, x, \mu \leq \mu_0) = \mathrm{FIT}(t, x, \mu \leq \mu_0)$. Similarly, if one is interested in the $|t|$-severity of an hypothesis $\mu_- \leq \mu \leq \mu_+$ given by data x, the probability of data at least as t-extreme as that observed will be minimized when $|\delta|$ is minimized, which is again when $\delta \to 0$. So the equality of FIT and SEV holds in these cases as well. Obviously, SEV is no more informative of the confirmation of an hypothesis than FIT in these cases, contrary to its stated ambitions.

Intuitively, the problem with equation 10 revealed in the above examples is that the infimum ranges over *all* simple hypotheses outside of the one being tested. One should not be interested in just any hypothesis that could be true if the one tested were false, but only ones that themselves are sufficiently plausible—i.e., have themselves sufficient FIT with the data. Indeed, guidelines for pedestrian applied classical statistics indicate that in power calculations for the one-sample t-test, one should perform the calculation at a variety of values of the true variance *close to the sample variance*, i.e., values that fit sufficiently closely with the data. Here it is therefore important to ensure that the test criteria used for FIT incorporate not just the parameter of interest, but other "nuisance" parameters needed to determine the probability distribution as well.

This gives rise to the final (?) modification of the definition of severity. Since we must set a threshold, α for those alternative hypotheses with sufficiently high FIT, we have now a four-place function:

$$\mathrm{SEV}(T, x, H, \alpha) = \inf_{\{\theta' \in \Theta \setminus H : \mathrm{FIT}(T,x,\theta') > \alpha\}} P_{\theta'}(T(X,\hat{\theta}) >_{\hat{\theta}} T(x,\hat{\theta})) \tag{12}$$

In practice, the value of α should be low enough to make the set over which the infimum ranges non-empty, but should not be so large so as to

In the examples with the z-test, the value of α did not matter because the alternative simple hypotheses contributing the lower values over which severity minimizes were already those which best fit the data. But one cannot presume this in general.

3.3 Co-Fit

Another consideration following from the foregoing $|t|$-test example comes from the case in which $\mu_- = \mu_+ := \mu_0$, i.e., the testing of a simple hypothesis. Intuitively, one should not expect such an arbitrarily precise hypothesis to be well supported for any data set. In this case, the best fit is just $\hat{\mu} = \mu_0$, so regardless of what the best fit is for σ, the best-fitting alternatives also have $\mu \approx \mu_0$. Consequently, even in this case severity ends up being equal to fit. So if μ_0 ends up quite close to $\bar{x}$, both the fit and severity will be high.

One response to this problem is to draw attention to the bounds for SEV needed for good evidence, e.g., requiring that they be (perhaps much) higher than FIT. This may require a more detailed investigation of the structure of FIT values, i.e., whether they are best characterized as ordinal, interval, or ratio quantities.

Another, which I shall pursue here, is to employ a third dimension of confirmation. The intuition behind it is simple. To accept an hypothesis H, whether simple or composite, is to *reject* the

hypotheses $\theta \in \Theta\backslash H$. This would only be justified if the criterion for rejection of $\Theta\backslash H$ is satisfied, i.e., if its FIT is sufficiently low. This leads to the obvious definition of what I shall call co-fit:

$$\begin{aligned}\text{coFIT}(T, x, H) &= 1 - \text{FIT}(T, x, \Theta\backslash H) \\ &= 1 - \sup_{\theta' \in \Theta\backslash H} P_{\theta'}(T(X, \theta') >_{\theta'} T(x, \theta')) = \inf_{\theta' \in \Theta\backslash H} P_{\theta'}(T(X, \theta') \leq_{\theta'} T(x, \theta')). \quad (13)\end{aligned}$$

This rules out the simple hypothesis $\mu = \mu_0$ from being highly confirmed in the $|t|$-test because data more T-fitting to $\mu \neq \mu_0$ would *not* likely be observed; in fact, $\text{coFIT}(|t|, x, \mu = \mu_0) = 0$.

Describing this criterion in words brings out its connection with the *safety* criterion sometimes used in epistemology for evidence:

> Data x is coFIT T-evidence for H to the extent that, if H were false, then data more T-fitting to $\neg H$ would likely be observed.

Usually safety is expressed as the condition that, if the data x were likely to be observed, then H would be true. In other words, H contains all the hypotheses that make the data likely. This is logically equivalent to the condition that $\neg H$ contains only hypotheses that make the data unlikely. Hence, in the context of a testing criterion T, data more T-fitting to $\neg H$ would likely be observed.

There is also a connection with the severity criterion of the previous section. In their discussions of severity, Mayo and Spanos often gives two different severity criteria, one for accepting an hypothesis and another for rejecting it. My own discussion has followed theirs for acceptance.[5] An hypothesis H_0 can be justifiably rejected, Mayo and Spanos (2011, p. 168) claim, when it fits $\Theta\backslash H_0$ and "there is a high probability ([e.g.,] .99) that a less statistically significant difference [from what H_0 predicts] would have resulted, were H_0 true." Once one understands "less statistically significant difference" as "more T-fitting," this just becomes the definition of co-fit. Perhaps it is only a matter of the vocabulary one uses to express a concept, but I find that severity and co-fit, unfamiliar as they are to many, are best seen as distinct concepts, both of which are needed in order to assess evidence, rather than only one in situations of acceptance and another only in situations of rejection.

3.4 Co-Severity

Since in the previous section I have defined the coFIT criterion for evidence, it would perhaps seem natural to define by analogy co-severity:

$$\begin{aligned}\text{coSEV}(T, x, H, \alpha) &= 1 - \text{SEV}(T, x, \Theta\backslash H, \alpha) \\ &= 1 - \inf_{\{\theta \in H : \text{FIT}(T, x, \theta) > \alpha\}} P_\theta(T(X, \hat{\theta}') >_{\hat{\theta}'} T(X, \hat{\theta}')) \\ &= \sup_{\{\theta \in H : \text{FIT}(T, x, \theta) > \alpha\}} P_\theta(T(X, \hat{\theta}') \leq_{\hat{\theta}'} T(X, \hat{\theta}')), \quad (14)\end{aligned}$$

where $\hat{\theta}'$ is the best fitting $\theta' \in \Theta\backslash H$.

[5] There are some differences, though: they don't register the (in general) hypothesis-dependence of the event whose probability is being calculated, nor the need for the α parameter to restrict the calculation to sufficiently fit alternatives. One finds in Spanos (2006) implicitly the use of the best fit but without acknowledgment of the subtleties that nuisance parameters introduce (as a plug-in estimate for the variance is used in the t-test).

That said, it is less clear that co-severity should play an important role as a criterion of evidence. This is because, unlike fit, a low severity score for an hypothesis is not evidence to reject the hypothesis, but only weak or equivocal evidence for it. So a pattern of reasoning similar to that which led to the adoption of co-fit does not hold for co-severity.

Perhaps there is some other reason to adopt it as an evidential criterion. But until such a reason is forthcoming it is reasonable to decline including co-severity.

4 Induction

As a theory of confirmation, severe testing takes on all hypotheses on a par: there is no asymmetric preference for "null" hypotheses over "alternative" hypotheses. The FIT, SEV, and coFIT scores assigned to atomic or composite hypotheses act as a kind of three-dimensional confirmation score in $[0, 1]^3$ that, while based in probabilistic concepts, does not itself represent in general any kind of uncertainty. These confirmation scores themselves do not license any particular inferences about the state of the world (i.e., the truth of some hypotheses $H \subset \Theta$). In this sense severe testing is Fisherian in spirit, although expanded with additional evidential criteria to allow for positive, rather than just negative, evidence.

To apply these scores to inductive inference, however, one can introduce *threshold values* for FIT, SEV, and coFIT. The logic itself does not determine these threshold values; they will in general depend on the values of and risks perceived by the user. (This connection with decision theory, though worth pursuing, is beyond the scope of the present article.) But once these thresholds are set, the resulting inductive logic is three-valued, in the sense that a data set, as a model, assigns to each $H \subseteq \Theta$ one of "accept," "reject," or neither:

Accept. The necessary and jointly sufficient criteria for acceptance are values of FIT, SEV, and coFIT above the indicated acceptance threshold. Acceptance of an hypothesis H entail the rejection of its negation, $\Theta \backslash H$.

Reject. The necessary and sufficient criterion for rejection is a FIT value below the indicated rejection threshold (which is in general not the acceptance threshold). Rejection of an hypothesis H does not entail the acceptance of its negation, $\Theta \backslash H$, as the hypotheses in Θ are not in general exhaustive.

Neither. All hypotheses neither accepted nor rejected are held in abeyance. These are hypotheses with either middling FIT, or too low SEV or coFIT.

The sense in which this is an inductive logic is that, as more data accumulates, previously accepted or rejected hypotheses may be no longer so judged, and hypotheses held in abeyance may become accepted or rejected.

The spirit of this procedure is similar to Neyman-Pearson testing, except that that procedure requires an "accept" or "reject" decision for every test, yielding in the process an asymmetry between null and alternative hypotheses. By contrast, on the present approach, weak evidence accrued may not be enough to reject or accept any hypothesis, thus avoiding the behaviorism with which classical statistics is sometimes charged. Indeed, it is possible to accrue evidence warranting the *rejection* of all hypotheses in Θ—such cases can commonly arise in the process of

model verification. This reveals the important role that the delimited set of working hypotheses Θ plays in the theory of severe testing.

Why can't one include *all* statistical hypotheses in Θ? While there is nothing mathematically wrong with doing so, it will typically entail extremely difficulty in calculate SEV and coFIT for hypotheses of interest, and when this is possible those values will likely be too low to warrant acceptance. (This may be a point at which measures of fit that also depend on simplicity considerations play a role.) However, the Fisherian logic of rejection only requires calculating FIT, which can be done while taking Θ to include all statistical hypotheses. This Fisherian fragment of severe testing thus projects the more complicated confirmation values in $[0, 1]^3$ to one dimension, but in doing so limits the kinds of inductive inferences one can make.

(More to be said about the non-classical features of this logic.)

References

Mark L. Taper and Subhash R. Lele. (2004). *The Nature of Scientific Evidence*. Chicago: University of Chicago Press.

Deborah G. Mayo 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: 323–357.

Deborah G. Mayo and Aris Spanos. (2011). "Error Statistics" in *Philosophy of Statistics*, Vol. 7 of Handbook of Philosophy of Science. Eds. Prasanta S. Bandyopadhyay and Malcolm R. Forster. Elsevier: 153–198.

Aris Spanos. (2006). "Revisiting the omitted variables argument: Substantive vs. statistical adequacy," *Journal of Economic Methodology* 13.2: 179–218.