Strategies for Securing Evidence through Model Criticism

Kent W. Staley
Saint Louis University

November 28, 2010

Abstract: Some accounts of evidence regard it as an objective relationship holding between data and hypotheses, perhaps mediated by a testing procedure. Mayo's error-statistical theory of evidence is an example of such an approach. Such a view leaves open the question of when an epistemic agent is *justified* in drawing in inference from such data to a hypothesis. Using Mayo's account as a launching point, I propose a framework for addressing the justification questions via a relativized notion, which I designate *security*, meant to conceptualize practices aimed at the justification of inferences from evidence. I then show how the notion of security can be put to use by showing how two quite different theoretical approaches to model criticism in statistics can both be viewed as strategies for securing (in my sense) claims about statistical evidence.

Contents

1 Introduction

Error-statistics (ES) proposes that evidence derives from testing procedures that constitute severe error probes. In statistical settings, ES employs a modified version of Neyman-Pearson Theory (NPT). Like NPT, the error-statistical approach uses probability distributions as models of the reliability of testing procedures, i.e., the rate at which they yield errors with regard to a family of competing hypotheses, which are themselves represented within the statistical model. Roughly, good tests in the ES view are those with appropriately low rates of error in indicating discrepancies from a family of competing hypotheses, and good evidence for a hypothesis results from the appropriate use of good tests. Mayo writes, "Data in accordance with hypothesis H indicate the correctness of H to the extent that the data result from a procedure that with high probability would have produced a result more discordant with H, were H incorrect" (Mayo 1996, 445n). Putting this idea in more schematic terms, the ES theory of evidence can be articulated in terms of Mayo's 'severe test' requirement: Supposing that hypothesis H is subjected to test procedure T employing test statistic x, resulting in data x_0,

> Data x_0 in test T provide good evidence for inferring H (just) to the extent that H passes severely with x_0, i.e., to the extent that H would (very

> probably) not have survived the test so well were H false. (Mayo and Spanos 2006, 328)

The idea of severity is elaborated according to the following schema: H passes a severe test T with data x_0 if

SR1 x_0 fits H, and

SR2 with very low probability, test T would have produced a result that fits H as well as (or better than) x_0 does, if H were false (and some alternative incompatible with H were true).

To a first approximation, one can say that the features of testing procedures (their error rates) that probability statements are meant to capture in this context are putatively objective features that obtain or not independently of what is known or believed by any individual.[1] These features can be thought of as characterizing a certain kind of reliability for the procedures employed in the inference from data to statistical generalizations.

Here I propose to press a question regarding the relationship between *evidence* as defined by the error-statistical approach and the *justification* of inferences based upon such evidential relations. Does the fact that data x_0 in test T provide good evidence for inferring H suffice for an individual to be justified in drawing that inference? ES certainly does propose a relationship between the error probabilities of the testing procedures used in drawing inferences from data and the justificatory status of those inferences. For example, a recent defense and elaboration of the ES account provides the following "inferential rationale" to articulate the basis for methodologies centered on error-probabilities:

> Error probabilities provide a way to determine the evidence a set of data x_0 supplies *for making warranted inferences* about the process giving rise to data x_0 (Mayo and Spanos 2006, 327, emphasis added)[2]

A careful reading, however, reveals this statement to concern how error probabilities are to be *used* in the ES account, and not a statement about the conditions under which inferences are in fact justified. More precisely, the severe test requirements quoted above articulate conditions under which data count as good evidence for a hypothesis. To say that (i) data x_0 are good evidence for H, however, is not the same as saying that (ii) a person in such-and-such an epistemic situation is justified in accepting H on the basis of x_0.

In this paper, I will present the ES account of evidence as an unrelativized account. I do not propose here to assess the ES view vis a vis other accounts of evidence. Rather, I will argue that satisfying the $SR1$ and $SR2$ requirements does not suffice for the justification of inference by a given epistemic agent. To help close this gap between evidence and justification, I will propose a relativized concept that is compatible with, though distinct from, the requirements of the ES account of evidence. This concept, which I call *security*, is defined in terms of truth across epistemically possible scenarios. Since epistemic possibility is a relative notion, so is security. I propose that the value of this concept lies chiefly in its heuristic use as way of thinking about and developing justificatory practices of *securing* inferences from data (i.e., increasing the relative range of epistemically possible scenarios across which those inferences are valid), in a manner that is independent of the statistical framework in which one works (whether error-statistical, Bayesian, or otherwise). To illustrate the value of the security framework, I discuss two general strategies of securing inferences: weakening and strengthening strategies. I then turn to theoretical statistics and discuss two approaches to model criticism in statistics — robust statistics and mis-specification testing — as examples of the weakening and strengthening strategies respectively.

2 From N-P to ES: Reliability, Evidence, and Methodology

The roots of the error-statistical approach lie in the frequentist tradition of mathematical statistics as that tradition has evolved from its origins in the work of R. A. Fisher and the joint efforts of Jerzy Neyman and Egon Pearson. To help clarify the error-statistical approach, then, it will be useful to consider first the orthodox Neyman-Pearson approach to statistical inference, and then to consider how ES departs from such orthodoxy.

Suppose that we seek answers to questions regarding the value of a "location parameter" μ_x for a distribution function f governing a series of random variables $\mathbf{X} \equiv X_1, X_2, \ldots, X_n$. This parameter might, for example, correspond to a physical quantity such as the mass of a newly discovered elementary particle, and the random variable might correspond to estimates of that quantity based on measurements made on its decay products. Suppose further that we know f to be normal, with unknown mean μ_x and known variance σ_0^2. We might start by asking whether μ_x exceeds a certain minimum value μ_0.

Consider first an "orthodox" Neyman-Pearson approach to specifying a test. The basic idea behind N-P testing is that, by specifying in advance the hypotheses among which a discrimination is to be made, and by specifying a statistical model that adequately represents data-generation as a stochastic process, one can exploit the probabilistic features of the statistical model, and use it as a basis for drawing inferences by using testing rules with error probabilities that are good or even optimal in a certain sense, to be explored below. Orthodox N-P thus provides a framework for making decisions in which the rate at which errors are committed can in principle be controlled. For this reason, many interpret N-P testing in *behaviorist* terms, according to which N-P tests serve the aim, not of evaluating the evidence for or against some proposition, but of deciding what to do in a way that will limit one's long-run losses from erroneous inferences. Neyman himself advocated such an approach (Neyman 1950).

An N-P test thus requires the prior specification of a statistical model. This model can be written as $\mathscr{M}_{\boldsymbol{\theta}}(\mathbf{x}) = \{f(\mathbf{x};\boldsymbol{\theta}), \boldsymbol{\theta} \in \Theta\}, \mathbf{x} \in \mathbb{R}^n_X$. Here $f(\mathbf{x};\boldsymbol{\theta}), \mathbf{x} \in \mathbb{R}^n_X$ is the joint distribution of $\mathbf{X}$ and the vector $\boldsymbol{\theta}$ gives the statistical parameters for that distribution, which are represented as lying somewhere in the parameter space Θ. The primary function of such a model is to represent "often in considerably idealized form, the data-generating process" and is thus a "model of physically generated variability" (Cox 2006). A statistical model can be described by reference to the assumptions it makes regarding particular statistical characteristics of the data-generating process. In the present example, these assumptions could be given with reference to a particular probability model defined as follows:

$$\Phi = \{f(x;\boldsymbol{\theta}) = \frac{1}{\sigma\sqrt{2\pi}}\exp\{-\frac{(x-\mu)^2}{2\sigma^2}\}, \boldsymbol{\theta} \equiv (\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}_+, x \in \mathbb{R}\}. \qquad (1)$$

Our assumptions are that: (1) $E(X_i) = \mu, i = 1, 2, \ldots$ (the expectation value of X_i, or distribution mean, is constant), (2) $Var(X_i) = \sigma^2$ (the variance, defined as $Var(X) \equiv E[(X - E(X))^2]$, is constant), (3) the random variables $\mathbf{X}$ are independent (i.e., $f(x_1, x_2, \ldots, x_n) = f_1(x_1) \cdot f_2(x_3) \cdots f_n(x_n)$, for all $(x_1, x_2, \ldots, x_n) \in \mathbb{R}^n$). Finally, we make the sampling assumption that $\mathbf{X} \equiv X_1, X_2, \ldots X_n$ is a random sample.

Next, the N-P approach requires us to make explicit both the *null* hypothesis and the *alternative* against which it is to be tested. The former might be thought of as that hypothesis, departures from which we are particularly interested to discover. Thus, the null here might state $H_0 : \mu_x \leq \mu_0$. The alternative would state $H_1 : \mu_x > \mu_0$. This can be thought of as a matter of demarcating two regions in the space Θ of possible values of the parameter μ. $H_0 : \mu \in \Theta_0$ is to be tested against $H_1 : \mu \in \Theta_1$.

To figure out the error probabilities for the test we use, we need to choose a feature of the data that will serve as a criterion for accepting or rejecting the null hypothesis. For this example the test statistic $\kappa(\mathbf{X}) \equiv \sigma^{-1}\sqrt{n}(\hat{\mu}_n - \mu_0)$, where

$\hat{\mu}_n \equiv n^{-1} \sum_{i=1}^{n} X_i$ is the sample mean, will allow us to employ a test of H_0 versus H_1 that is optimal in the following sense: We note first that under the assumption that the distribution F is normal, $\hat{\mu}_n$, itself a random variable, is the best estimator for μ_x, in the sense that it is *unbiased* (the mean of the sampling distribution for $\hat{\mu}_n$ equals the value of the parameter μ_x), *efficient* (its variance is minimized, both with regard to a finite sample and asymptotically as n goes to infinity), and *strongly consistent* (as n goes to infinity, the value of $\hat{\mu}_n$ is equal to the true value of μ_x with probability one). Moreover, as shown by the central limit theorem, $\sigma^{-1}\sqrt{n}(\hat{\mu}_n - \mu)$ is asymptotically Standard Normally distributed, with mean equal to zero and variance equal to one (thus, $\hat{\mu}_n$ is *asymptotically Normal*).

This last point allows us the convenience of using a Standard Normal table to follow the orthodox Neyman-Pearson procedure of first choosing a cut-off or critical value of the test statistic for rejecting the null hypothesis, such that the probability of rejecting the null hypothesis when true does not exceed a certain predetermined value, such as 0.05 (this would require setting the cutoff at $c = 1.96$). We would then consider the *power* of this test, defined to be one minus its probability of accepting the null when the alternative is true (the type II error). Since the alternative here is a compound hypothesis (it encompasses all hypotheses regarding the value of μ_x such that $\mu_x > \mu_0$), the power of this test does not take on a single value, but is instead a function, defined over the entire parameter space Θ, which can then be used to determine the type II error probability (and thus the power) for particular values of μ. An optimal test will be one that is *uniformly most powerful* (UMP), i.e., is such that, for any $\mu \in \Theta$, its power is at least as great as that of any test using another test statistic. Although UMP tests do not always exist, there is such a test for this example; it is the test just described, using $\kappa(\mathbf{X})$ as a test statistic.

Supposing, then, that we obtain data such that the value of $\kappa(\mathbf{x})$ lies in the critical region (e.g., $\kappa(\mathbf{x}) = 2.10$), the orthodox Neyman-Pearson inference would be to reject the null hypothesis, and the probability of doing so erroneously would be reliable insofar as it would be limited to not more than 0.05. By contrast, should the result fall short of the cutoff value (say, $\kappa(\mathbf{x}) = 1.20$), then the test would yield the result that H_0 is not rejected. We can then ask what the probability is of failing to reject H_0 given specific alternative values of μ_x. For example, if we suppose that the true value of μ_x is 10.8, then the probability that this test would fail to reject H_0 (i.e., that $\kappa(\mathbf{x}) < 1.96$) is 0.02, and the power of the test is 0.98. However, assuming that $\mu_x = 10.2$, the probability of a type II error is 0.83, and the power of the test is a mere 0.17. These results derive from the distribution of the test statistic, under the assumed underlying distribution f.

Thus far, however, we have been considering only the "orthodox" N-P approach. As developed by Mayo, however, the ES approach would have us go beyond the more "behavioristic" orthodox approach to N-P testing.[3] In the ES approach, reliability in the form of error-probabilities enters not merely as a way of limiting potential losses in a series of repeated decisions based on data, but as a tool for characterizing how well the data discriminate between various possible answers to the question being investigated. This is most apparent in considering how ES utilizes post-data severity analyses with regard to a range of hypotheses.

Mayo and Spanos (2006) note a number of difficulties with the orthodox approach, many of which turn on the fact that the error probabilities we have used thus far are concerned only with whether the observed results fall inside or outside the critical region. In effect this is to treat outcomes that fall just short of the cutoff the same as those that fall very close to the expectation value of the test statistic under the null, while treating a result just short of the cutoff entirely differently from a result that just barely exceeds it. Meanwhile, any outcome that exceeds the cutoff value is treated the same, regardless of the magnitude by which it exceeds it. Although such an approach may well be suitable for a behavioristic approach that is concerned only to limit the rate at which erroneous decisions are made, it is ill-suited for the purpose of determining just what inferences are warranted by the data in hand regarding the family of hypotheses under consideration.

To address such issues, the ES methodology proposes the use of a severity analysis based on the actual data that looks at the error probabilities with the cutoff set at the observed value of the test statistic, under a range of possible alternative hypotheses. Such analysis is guided by "meta-statistical principles" and aims to address the problem previously mentioned of determining what kinds of inferences from a given test can be justified. As this approach involves a certain kind of "re-use" of data that some hold to be problematic, it should be emphasized that ES draws a disinction between the *primary* inference (basically, the inference to an "accept/reject" conclusion, drawn using a prespecified testing procedure with specific error probabilities), and the post-data severity analysis that relies on assumptions about *counterfactual* error probabilities. The error-probabilities employed are counterfactual in the sense that they are the probabilities that the test statistic would take some value greater or less than the observed value under various hypotheses of interest.[4] These post-data severity analyses do not constitute *new* statistical tests based on the same data, but are rather a means of determining the epistemological import of the initial testing results – something for which the orthodox behaviorist approach is inadequate.

Suppose, following Mayo's standard notation, we use $SEV(T, d(\mathbf{x_0}), H)$ to mean "the severity with which hypothesis H passes test T with an observed value for the test statistic of $d(\mathbf{x_0})$." As I will next illustrate, such measures of severity depend on error probabilities in a manner that depends on the discrimination of interest, and on whether the original test T (defined pre-data in terms of a particular cut-off) led to an acceptance or rejection of the null hypothesis.

In our example above, suppose that the observed value of $\hat{\mu}_n = 10.35$. This corresponds to a value of $\kappa(\mathbf{x}) = 1.75$, which falls short of the cutoff value of 1.96, but not by a lot. Thus our original two-sided test T gives the output "accept H." With a severity analysis, we would next ask the probability of getting so large a value of $\kappa(\mathbf{x})$, supposing that the true value of μ exceeds $\mu_1 = \mu_0 + \gamma$, for some relevant values of γ. In other words, although our test accepts the null, we are interested to know whether the test would probably have yielded so large a value of $\kappa(\mathbf{x})$, even though the value of μ is actually greater than μ_0 by particular amounts.

So, for example, to determine whether $\mu \leq 10.2$ passes with high severity, we could evaluate the probability $P(\kappa(\mathbf{x}) > \kappa(\mathbf{x_0}); \mu = 10.2) = 0.22$. This gives the lower bound of the severity with which $\mu \leq 10.2$ passes against particular alternative values of μ. Mayo and Spanos articulate the relevant principle as follows (notation adapted to mine):

> If there is a very *high* probability that $\kappa(\mathbf{x_0})$ would have been larger than it is, were $\mu > \mu_1$, then $\mu \leq \mu_1$ passes the test with *high* severity, i.e. $SEV(\mu \leq \mu_1)$ is high. If there is a very *low* probability that $\kappa(\mathbf{x_0})$ would have been larger than it is, even if $\mu > \mu_1$, then $\mu \leq \mu_1$ passes the test with *low* severity, i.e. $SEV(\mu \leq \mu_1)$ is low. (Mayo and Spanos 2006, 337)

Clearly, in this example, the severity with which $\mu \leq 10.2$ passes is rather low. For comparison, given the same data, the hypothesis that $\mu \leq 10.4$ passes with a severity of 0.60, while $\mu \leq 10.8$ passes with severity 0.99. Of these three possible inferences, only the last is well supported by the evidence.

Notice that these rules for post-data severity analysis provide methods for determining what hypotheses have passed with what degree of severity. As methods, therefore, they constitute a means by which investigators can acquire the requisite knowledge for distinguishing between justified and unjustified inferences. The severity relationships themselves remain objective in the sense that they obtain independently of any individual's epistemic situation, thus underscoring that as a theory of evidence, the ES account relies on objective criteria.

When satisfied, these criteria – stated in terms of error probabilities – ensure that the investigator is *using a severe error probe.* This is the central notion of error-elimination that is at work in the ES account. What it means is that a testing procedure is being used that reliably discriminates between different possible answers to an investigator's question, and in that sense supports learning about the phenomenon which that question is about. *However, that a testing procedure can serve as the basis for learning does not entail that any particular individual is in a position to learn from that procedure.* I propose that we here encounter a gap that needs to be filled if we are to have an adequate epistemology of science built on the error statistical approach.

Significantly, the ES criteria for evidence as articulated in $SR1$ and $SR2$ are *not relativized* to an investigator's epistemic situation. They depend on objective features of the testing situation and on a particular class of alternative hypotheses. This latter dependence simply reflects the fact that a test that is a good probe for discriminating a hypothesis against one set of alternatives might not be good for discriminating against another, an issue that has been noted in Staley (2008). However, given a set of competing hypotheses, the error probabilities are determined by the test criteria, and do not depend in any way on the epistemic situation of the investigator, except insofar as it plays a causal role in leading her to specify the criteria that she does.

Thus, if data x_0, testing procedure T, and hypothesis H satisfy $SR1$ and $SR2$, then they do so *independently* of the knowledge, beliefs, or abilities of any epistemic agents, whether performing experiments, drawing inferences, or reading research reports. Of course, such factors may be of instrumental value in producing conditions that allow for $SR1$ and $SR2$ to be satisfied. The point is that facts about epistemic agents, real or hypothetical, play no *constitutive* role in evidential relations in the ES account.

This leaves a gap in the error-statistical philosophy of science, however, if we focus just on the account of evidence as stated in $SR1$ and $SR2$. As noted above, ES aims to provide resources for the investigator to determine which inferences are justified with regard to data produced as part of a given testing procedure. Unlike ES's unrelativized criteria for evidence, however, justification in the sciences *does* seem to be relative to an investigator's epistemic situation. So although Mayo and Spanos, in their advocacy of post-data use of severity analyses, with the accompanying rules of acceptance and rejection have articulated a methodology for justifying statistical inferences, their account falls short of an epistemology insofar as it lacks a concept of justification that links the results of applying these methods with the epistemic situations of investigators. To put it another way, the ES methodology provides a means for evaluating evidence; the ES theory of evidence relates evidence to features of the testing situation; but nothing in the ES account connects the obtaining of an evidential relationship with the question of what is required for an investigator in a particular epistemic situation to be justified in drawing a particular conclusion from the experimental data.

Before proceeding, it would be useful to explain what I mean by "epistemic situation," as this is meant to be a somewhat richer notion than simply a set of background beliefs. This term is borrowed from Achinstein (2001), who describes an epistemic situation as a situation in which "among other things, one knows or believes that certain propositions are true, one is not in a position to know or believe that others are, and one knows (or does not know) how to reason from the former to [a particular] hypothesis" (ibid., 20).

3 Security in the justification of evidence claims

In a nutshell, the problem is this: For an investigator to *justify* an inference from x_0 to H via test procedure T, or the claim that data x_0 from T are evidence for H, it is not sufficient that H pass a severe test with x_0, T. In addition, the investigator must be able to offer reasons in support of the claim that H does pass a severe test with x_0, T. Justification thus attaches not simply to the data, test, and hypothesis, but to the inference as an epistemic act of the investigator. Put differently, evidential relations depend on an epistemic agent only insofar as they depend on that agent's decisions to test hypotheses in certain ways. But justification depends on the agent's epistemic situation: What does she believe? What does she know?

For scientific claims, justification is directed at an audience of some sort.

Suppose that a researcher presents a conclusion from data gathered during research. The decision to present a conclusion indicates that the researcher and her collaborators are convinced that they are prepared to justify their inference in response to whatever challenges they might plausibly encounter. Their confidence will result from their having already posed many such challenges to themselves. New challenges will emerge from the community of researchers with which they communicate. Such challenges take many forms, depending on the nature of the experiment and of the conclusions: Are there biases in the sampling procedure? Have confounding variables been ruled out? To what extent have alternative explanations been considered? Are estimates of background reliable? Can the conclusion be reconciled with the results of other experiments? Have instruments been adequately shielded, calibrated, and maintained? Is the correct model being employed? Is the reference class used for determining probabilities appropriate? Is the test-statistic well-defined and appropriate for the inference drawn? What policy was followed in deciding to terminate the experiment?

To a large extent, such challenges can be thought of as presenting possible scenarios in which the experimenters have gone wrong in drawing the conclusions that they do. But such challenges are not posed arbitrarily. Being logically possible does not suffice, for example, to constitute a challenge that the experimenter is responsible for addressing. Rather, both experimenters in anticipating challenges and their audience in posing them draw upon a body of knowledge in determining the kinds of challenges that are significant (Staley 2008).

Here I propose a heuristic that might serve to systematize the strategies that experimenters use in responding to such challenges and allow for a clearer understanding of the epistemic function of such strategies (see also Staley and Cobb 2010).

Already we can identify certain features of the problem situation just described that can guide us in formulating the concept at which we aim. Responses to the kinds of challenges we have in mind are concerned with scenarios in which the claim to have found evidence supporting the conclusion turns out to be erroneous; they are posed as more than mere logical possibilities, but as scenarios judged significant by those in a certain kind of epistemic situation, incorporating relevant disciplinary knowledge; and an appropriate response needs to provide a basis for concluding that the scenario in question is not actual.

I propose that we think of the practices of justifying an evidence claim as the *securing* of that claim against scenarios under which it would be incorrect. Such a perspective introduces a second notion of error-elimination that is distinct from the use of a severe error probe. The latter is unrelativized: testing procedures have their error rates independently of our judgments about them. One eliminates error by using a procedure that *as a matter of fact* rarely leads to false conclusions, a matter that is independent of one's epistemic situation. The former is relativized: one eliminates error by showing that, given what one knows (more precisely, given one's epistemic situation), the ways in which one might go wrong can be ruled out, or else make no difference to the evidential conclusion one is drawing. That is to say, one secures the evidence.

To clarify how this heuristic works, let me offer the following definition:

Definition (security)**.** *Suppose that Ω_0 is the set of all epistemically possible scenarios relative to epistemic situation K, and $\Omega_1 \subseteq \Omega_0$. A proposition P is* secure throughout Ω_1 *relative to K iff for any scenario $\omega \in \Omega_1$, P is true. If P is secure throughout Ω_0 then it is* fully secure.

Some explanation of terminology is in order. This definition employs the notion of epistemic possibility, which can be thought of as the modality employed in such expressions as "For all I know, there might be a third-generation leptoquark with a rest of mass of 250 GeV/c2" and "For all I know, I might have left my sunglasses on the train." Hintikka, whose (Hintikka 1962) provides the origins for contemporary discussions, there takes expressions of the form "It is possible, for all that S knows, that P" to have the same meaning as "It does not follow from what S knows that not-P."[5] Borrowing Chalmers' notion of a scenario for heuristic purposes, we use that term to refer to what might be intuitively thought of as a maximally specific way things might be (Chalmers 2011).[6] In practice, no one ever considers scenarios as such, of course, but rather focuses on salient differences between one scenario and another.

Although security is defined so as to be applicable to any proposition, we are here concerned with the context of inductive reasoning from data and thus with the security of propositions expressing an *evidence claim*, i.e., a claim of the form 'Data E (resulting from test T) are evidence for the hypothesis that H.' Moreover, I will sometimes apply the term security to *inferences*. Such usage should always be understood to be informal and to refer to inductive inferences, such that an inductive *inference* from data E to hypothesis H is secure exactly to the extent that the *proposition* 'E is good evidence for H' is secure.

An evidence claim is thus secure for an agent to the extent that it holds true across a range of scenarios that are epistemically possible for that agent. Exactly which scenarios are epistemically possible for a given epistemic agent is opaque, and not all epistemically possible scenarios are equally relevant, so the methodologically significant concept turns out to be *relative security*: how do investigators make their evidential inferences more secure? And which scenarios are the ones against which they ought to secure such inferences?

At this point, some readers — particularly those with Bayesian leanings — might wonder why one does not simply introduce a prior distribution or some such measure across the class of possible scenarios. Introducing a measure across a range of possible statistical models is a natural thing to do in a Bayesian framework. Here I will not pursue such an approach. Primarily, my reason for not considering a Bayesian approach is that I am here concerned with an error-statistical approach. An important motivation for advocates of error-statistical approaches is to avoid assigning probabilities to anything that resists being modeled as the outcome of a stochastic process of some sort. (In this paper I do not enter into debate over the justification for this motivation.) If the measure over scenarios were to be a probability measure, then it would be conceptually at odds with the approach to evidence that is here assumed as a starting point. If the measure were not a probability, it would need to be provided with some other interpretation to be meaningful. Having no defensible interpretation at hand, I deem it advisable for now to eschew such a measure entirely.

Furthermore, and more fundamentally, insofar as one might wish to pursue a Bayesian approach, this should not be seen as a *substitute* for a security perspective as here described, but rather as a different evidence-theoretic context in which the need to address security remains, but the conceptual tools for doing so are different. An elaboration of this point would take us beyond the scope of this paper, however.

To return to my main argument, then, I contend that numerous scientific practices already aim at enhancing the security of evidence claims, and that these can be usefully viewed as instances of two types of strategy: *weakening* and *strengthening*. In weakening, the conclusion of an evidential inference is logically weakened in such a way as to remain true across a broader range of epistemically possible scenarios than the original conclusion. Strengthening strategies operate by adding to knowledge, reducing the overall space of epistemically possible scenarios so as to eliminate some in which the conclusion of the evidential inference would be false.

In what follows I survey the pursuit of these two strategies through two developments within theoretical statistics. The first of these is *robust statistics*, a branch of mathematical statistics that has received little attention from philosophers of science. The second is the program of misspecification testing (MST) and model respecification advocated by Spanos (1999) and by Mayo and Spanos (2004) from a standpoint firmly within the error-statistical approach. The first can be viewed as an example of a weakening strategy, while the latter operates by strengthening. Viewing both approaches as efforts to address the problem of securing evidence claims yields insight into the justification of claims regarding the evidential support of scientific hypotheses.

4 Security through robust statistics

Robust statistics originates in the insight that many classical statistical procedures depend upon parametric models that may hold only approximately. One might hope that when those models are approximately valid, so are the conclusions drawn. However, it is well established that small departures from such a model can dramatically affect the performance of statistical measures (Tukey 1960). In particular, theorists have been concerned with three reasons why a parametric model might fail to hold exactly (Hampel et al. 1986):

1. Rounding of observations
2. Occurrence of gross errors (bad data entry, instrument malfunction, etc.)
3. Idealization or approximation in the model

As Stephen Stigler notes, "Scientists have been concerned with what we would call 'robustness' – insensitivity of procedures to departures from assumptions . . . for as long as they have been employing well-defined procedures, perhaps longer" (Stigler 1973, 872).[7] Statisticians continue to use the term 'robustness' to refer broadly to this notion of insensitivity, and there are several theoretical approaches to the development of frameworks for robust statistical inference.

The theoretical interest of these approaches derives from their methodological significance: In practice, data analysis often uses estimators or test statistics[8] that do not behave at all like they are supposed to in the presence of even small violations of the parametric models on which they depend. Put another way, the reliability properties that are understood to hold for these estimators are an indicator of the evidential strength of the results of their application *only if* those properties really do hold. In many situations in which calculations based on a parametric model attribute such reliability properties to an estimator, the model does not in fact hold exactly, and in many of *those* situations, the result is that *the attributed reliability properties do not hold even approximately.*[9]

Here I will survey some influential robustness notions that originated in the 1960s in work by Peter Huber (1964) and Frank Hampel (1974, 1971, 1968).

Their approaches have been extended and applied to problems far beyond simple one-dimensional estimation problems to multi-dimensional and testing contexts, but I will here discuss some of the early developments on one-dimensional estimators. My aim is not to survey the state of robust statistical theory, but to argue that from the outset the theoretical work has been guided by a methodological concern with the security of statistical conclusions, and that the theory of robust statistics exemplifies systematic thinking about how to secure evidence via a weakening strategy.[10]

4.1 Huber's minimax approach

In his groundbreaking 1964 paper, Peter Huber introduced a class of estimators that he called "M-estimators."[11] Huber introduces these as a kind of generalization of least-squares estimators. To return to our previous example, recall that we selected a test statistic that employed the sample mean $T = \bar{x} = n^{-1}\Sigma_i x_i$ as a best estimator of μ, the "location parameter" of distribution F. This choice of estimator emerges as the solution to a problem of minimizing the sum of the *errors*, i.e., the squares of the differences between the observed values and those that would be predicted under the hypothesis chosen by that estimator. In other words, supposing T initially to be some unspecified function of random variables $x_1, x_2, \ldots x_n$, we seek to choose T so that $\Sigma_i(x_i - T)^2$ takes its minimum value. The solution to this particular minimization problem is in fact to define T to be the sample mean $T = \frac{1}{n}\Sigma_i x_i$.

The class of M-estimators is then introduced as those that solve the more general problem of minimizing some function of the errors, i.e., they minimize $\Sigma_i \rho(x_i - T)$, for some non-constant function ρ.[12] It was well-known that other statistics besides the mean performed better as location estimators when assumed exact parametric models failed. Since the choice of the mean as a location estimator could be defended on the grounds that it solves a particular minimization problem, perhaps more robust estimators would emerge as solutions to alternative minimization problems.

Of course, to determine whether this is the case, one needs some means of evaluating robustness. Huber's analysis assumes that the unknown underlying distribution F can be represented in the form of a mixture of a normal distribution Φ with another, possibly non-normal but symmetric distribution H: $F = (1-\epsilon)\Phi + \epsilon H$. This is sometimes called a "model of indeterminacy." (Here H is assumed unknown, but ϵ is assumed to be known.) In this setting, Huber opts to use the supremum of the *asymptotic variance* of an estimator as an indicator of its robustness.

More specifically: suppose that T is an estimator to be applied to observations $x_1, x_2, \ldots, x_n$ drawn from a family $\mathscr{P}_\epsilon$ of models that have the form of F just given, for some value of ϵ (call the resulting estimate T_n). Then the asymptotic variance of T at a distribution $G \in \mathscr{P}_\epsilon$ is understood to be the expected value of the squares of the differences between estimator values and the expected estimator values, evaluated at G, as $n \to \infty$, i.e., $V(T,G) = \lim_{n\to\infty} E_G[(T_n - E(T_n))^2]$. Then the most robust M-estimator for a given family F of distributions would be that which minimizes the maximal asymptotic variance across $\mathscr{P}_\epsilon$. In other words, the most robust M-estimator T_0 is the one that satisfies the condition:

$$\sup_{G\in\mathscr{P}_\epsilon} V(T_0, G) = \min_T \sup_{G\in\mathscr{P}_\epsilon} V(T,G) \tag{2}$$

Intuitively, this criterion selects the optimum choice for the "worst case scenario" compatible with the model of indeterminacy, in which the observed random variable is the least informative about the value of the parameter.

4.2 Hampel's infinitesimal approach

Beginning in his 1968 thesis and in a series of subsequent papers (Hampel 1974, 1971, 1968), Frank Hampel laid the foundations for the "infinitesimal" approach to robust statistics. Whereas Huber's approach begins by replacing the usual exact parametric model with a model of indeterminacy and then seeks to formulate a generalized minimization problem for that particular model, Hampel's approach begins with an exact parametric model and then considers the behavior of estimators in "neighborhoods" of that model.

First consider a qualitiative definition of robustness, as introduced in Hampel (1971).[13] Suppose that we consider a sequence of estimates $T_n = T_n(x_1, x_2, \ldots, x_n)$, where the x_i are independent and identically distributed observations, with common distribution F. Let $\mathscr{L}_F(T_n)$ denote the distribution of T_n under F. The sequence T_n is *robust at* $F = F_0$ iff, for a suitable distance function d,[14] for any $\epsilon > 0$, there is a $\delta > 0$, and an $n_0 > 0$, such that for all distributions F and all $n \geq n_0$,

$$d(F_0, F) \leq \delta \Rightarrow d(\mathscr{L}_{F_0}(T_n), \mathscr{L}_F(T_n)) \leq \epsilon \tag{3}$$

Intuitively, qualitative robustness requires that an estimator be such that closeness of the assumed distribution of the observations F_0 to their actual distribution F ensures that the assumed distribution of the estimator is close to its actual distribution.[15]

Alongside this qualitative criterion, Hampel introduced the notion of the influence function (IF) to quantify how much the value of an estimator would change with the addition of a single new data point with a particular value x.

Hampel described the IF as "essentially the first derivative of an estimator, viewed as a functional, at some distribution" (Hampel 1974, 383). More specifically, supposing an estimator functional T, a probability measure F on a subset of the real line R, and $x \in R$, the IF is defined as:

$$\mathrm{IF}_{T,F}(x) = \lim_{\epsilon \downarrow 0} \frac{T((1-\epsilon)F + \epsilon\delta_x) - T(F)}{\epsilon} \tag{4}$$

where δ_x denotes the pointmass 1 at x.

In practice, the importance of the influence function lies in derived quantities that serve as measures of different kinds of robustness. Three of these deserve mention here, as they are adapted to quite distinct worries involving robustness. The point I would like to emphasize about these quantities is that they all seek to capture behaviors of estimators in some kind of generic "worst-case scenario."

The first ("and most important," according to Hampel et al. (1986, 87)) of these derived concepts is the *gross-error senstivity* γ^*. Suppose that T is an estimator and F a distribution. Then the *gross-error sensitivity* for (T, F) is defined as:

$$\gamma^*(T, F) \equiv \sup_x |\mathrm{IF}_{T,F}(x)|, \tag{5}$$

and is described by Hampel as a measure of the "worst (approximate) influence which a small amount of contamination of fixed size can have on the value of the estimator" (ibid., 87). The gross-error sensitivity is thus useful for understanding how estimators react to outliers or other "contamination" (Hampel 1974, 387).

The *local-shift sensitivity* λ^*, defined as:

$$\lambda^*(T, F) \equiv \sup_{x \neq y} |\mathrm{IF}_{T,F}(y) - \mathrm{IF}_{T,F}(x)|/|y - x|, \tag{6}$$

measures the effects of small changes in the values of observations, such as might result from either rounding or grouping of observations, among other sources. Here one in effect removes an observation at point x and replaces it with an observation at a neighboring point y. Local-shift sensitivity is thus a "measure for the worst (approximate and standardized) effect of 'wiggling'" the data in this way (Hampel et al. 1986, 88)(Hampel 1974, 389).

Finally, the *rejection point* ρ^*, defined as:

$$\rho^*(T, F) \equiv \inf\{r > 0; \mathrm{IF}_{T,F}(x) = 0 \text{ when } |x| > r\}, \tag{7}$$

can be used to describe approaches to estimation that simply *reject* outliers – the most time-honored approach to robust estimation. The rejection point can be thought of as the smallest absolute value that an observation might have that would lead to its being rejected outright, thus having no influence on the value of the estimate. If data are never to be rejected, regardless of their value, then $\rho^* = \infty$.

We can now see how robust statistics responds to its motivating problem by giving investigators tools for evaluating how well statistical conclusions drawn with a particular claimed reliability hold up in the face of particular kinds of departures from a given model. Or, to put it in terms used in the definition of security: robustness notions in statistics aim to allow the investigator to determine and employ an estimator that would allow her evidence claims to remain valid for various ways in which, for all she knows, her initial assumptions might be wrong.

The general approach that the Huber/Hampel framework takes to enhancing security is a weakening strategy: the security of the inference is enhanced by weakening its conclusion. This is reflected in a comparison of the variance of different estimators. The variance of the sample mean, which has generally poor robustness characteristics, is easily shown to be smaller than that of other more robust estimators *at the Normal distribution*. The sample mean, thus, is a more *efficient* estimator than its more robust counterparts, allowing one to draw, ceteris paribus, a stronger conclusion from a given body of data (see Hampel 1974, esp. the table on p. 392 and accompanying text).

This last advantage is of course illusory if in fact the process generating data is not adequately modeled using the Normal distribution. A more robust estimator is thus a more secure choice for the inquirer who has assumed a statistical model based on the Normal distribution, although for all she knows the process might not be correctly described by a Normal distribution. The price paid is that the less narrowly distributed, but more robust estimators will in general lead to less precise estimates, making less efficient use of the information in the data than one would if the Normal model were valid and one used the mean as an estimator. The strategy is clearly a weakening one in the sense that one draws a weaker conclusion (an estimate that results in a larger interval for the same confidence level), relying on what is implicitly a "compound" or disjunctive premise: the conclusion is sound so long as either the assumed model or an alternative that is "close" to it (in a sense defined by the relevant robustness measure) is correct. The contrast between weakening and strengthening will emerge more clearly as we turn in the next section to an alternative strengthening strategy: rather than draw a weaker conclusion that remains sound across a range of models of epistemically possible scenarios, attempt to determine a statistically adequate model, and then choose the optimal inferential strategy for that model.

5 Security through misspecification testing

In this section I will discuss an alternative approach to model criticism. *Misspecification testing* (MST) constitutes a systematic method for probing assumed statistical models for errors of specification that would result in statistical inadequacy and respecifying as necessary. The method has been advocated by defenders of ES (Mayo and Spanos 2004, Spanos 1999). Here I will present just enough of an overview of the method to substantiate the contrast I wish to draw with robustness approaches to model criticism: Both robust statistics and MST address the issue of securing statistical evidence. The former does so through a weakening strategy while the latter does it through a strengthening strategy.

By its nature, MST calls for testing outside of the original parametric model. Indeed, because MST aims to consider *all* possible distributions as alternatives to that in the assumed model, it cannot proceed on a fully parametric basis at all. As Spanos notes, "the implicit maintained hypothesis [is] $\mathscr{P}$, the set *of all possible* *probability models*," including nonparametric models (ibid., 733, emphasis in original). This poses a difficulty, however. One might attempt to carry out a test of the assumed model by treating *it* as a null that can be specified parametrically, thus defining a subset $\mathfrak{B}_\theta \subset \mathscr{P}$, but given the impossibility of parameterizing the *alternative* $\mathscr{P} - \mathfrak{B}_\theta$, one seems to be forced into testing in an ad hoc and local manner, with no framework for evaluating the power of such tests. The situation seems to demand a Fisherian approach[16] to testing in which the aim is really to subject the null hypothesis to testing, but without the specification of an alternative hypothesis (apart from the implicit alternative that the true distribution lies within $\mathscr{P} - \mathfrak{B}_\theta$), thus leading one only to conclusions about how compatible the data are with the null. Yet one would also like to be able to systematize one's search for possible departures from the assumed model in a way that allows one to judge sensitivity of the test to such departures.

Spanos proposes to solve this difficulty by strategically employing a series of pseudo-Neyman-Pearson tests of the assumed model that situate that model within an "encompassing" statistical model, not as a true Neyman-Pearson test, but as a provisional setting for a kind of operationalization of testing unsupported in a Fisherian framework. In other words, rather than ad hoc scrutiny of single assumptions, Spanos's MST approach uses techniques of data analysis (largely graphical) to look for "specific directions of possible departures from the assumptions of the postulated model" (ibid., 763). Based on such information, one then postulates a new model that includes the original model as a special (null) case, and tests within the enlarged model for departures from that null. This allows for the full parametrization of the misspecification test, as required in Neyman-Pearson approaches. Nonetheless, Spanos insists, these are not true Neyman-Pearson tests because the context demands explicit openness to the possibility that the true model lies outside, not only the original postulated model, but also outside the encompassing model. Moreover, the "basic objective" of MST is that of Fisherian testing: "The significance level α, interpreted in terms of what happens in the long run when the experiment is repeated a large number of times, is irrelevant because the question the modeler poses concerns the particular sample realization" (ibid., 764).

Recall the simple Normal model of the example in section two of this paper. That model incorporated assumptions regarding distribution, dependence and heterogeneity. The aim of MST would be to use the data in hand to test these assumptions against their *alternatives*: that $X_1, \ldots, X_n$ are not Normally distributed, that some of them are probabilistically dependent on others, that they are not all identically distributed.

In the present case, then, the MST approach of specifying an encompassing statistical model that includes the original postulated model as a special case might lead one to replace the Independence assumption with an assumption that allows for Markov dependence. Suppose that we use notation $f(x; \boldsymbol{\theta})$ to denote a density function of random variable X with parameters $\boldsymbol{\theta}$, that $\mathbb{T}$ is the "index set" used to represent the dimension according to which the data are ordered, and that R is the Borel σ-field generated by the real numbers $\mathbb{R}$. Whereas the initial independence assumption regarding $\{X\}$ could be expressed in terms of the identity

$$f(x_1, x_2, \ldots, x_T; \boldsymbol{\phi}) = \prod_{i=1}^{T} f_t(x_t; \boldsymbol{\psi}_t) \text{ for all } t \in \mathbb{T}, \\ \text{and all } \mathbf{x} := (x_1, \ldots, x_T) \in R, \tag{8}$$

our new assumption would be that of Markov dependence:

$$f_k(x_k | x_{k-1}, x_{k-2}, \ldots, x_1; \boldsymbol{\phi}_k) = f_k(x_k | x_{k-1}; \boldsymbol{\psi}_k), k = 2, 3, \ldots. \tag{9}$$

Consistency then requires us also to replace the original heterogeneity assumption of identical distribution with that of second-order stationarity. We then have the following *statistical generating model*:

$$X_t = \alpha_0 + \alpha_1 X_{t-1} + u_t, t \in \mathbb{T} \tag{10}$$

(here u_t is the error term).

These modifications amount to the specification of an encompassing model that allows one to test the hypothesis H_0: that $(X_1, X_2, \ldots, X_T)$ are independent against the alternative H_1: that they are Markov dependent. In parametric terms this is a matter of testing $H_0 : \alpha_1 = 0$ against $H_1 : \alpha_1 \neq 0$.[17]

This brings us naturally to the question of what to do with the results of such tests. Although the mathematical apparatus is that of the Neyman-Pearson approach, the aims and interpretation of the tests are Fisherian, and some care is needed in the interpretation of test outcomes.

A chief distinction between MST and NP testing is the role played by the statistical model. For an NP test, the statistical model must be statistically adequate for it to guide the interpretation of test outcomes. It is this feature that allows one to draw *positive* evidential conclusions both in the case where the null hypothesis is accepted and in the case where it is rejected, with regard to those hypotheses that are tested with high severity (see Mayo and Spanos 2006). But the role of the statistical model in MST is different, as it serves only to allow for the development of tests that *potentially* have high power in testing the null model against alternatives in a particular direction. In our example, we may have a t-test that tests the null model postulating independence with potentially high power against alternatives postulating some degree of Markov dependence. This high power is potential in the sense that our determination of the power of the test relies on the encompassing model, which in Fisherian mode we allow may be false.

Suppose, then, that the null model *passes* this test. We then can say that, as far as the direction of departure from the null that is tested with high power is concerned, we have evidence that the null model is not in error by more than a magnitude to which the test is sensitive. This supports at least the provisional and approximate endorsement of the power assessments of the misspecification test. Our next step may be to consider other possible directions of departure, by turning to our assumptions regarding dependence or heterogeneity, for example, or by looking for higher order dependence. If the null model passes such a series of misspecification tests, then, insofar as we believe that we have ruled out all of the relevant ways in which that model fails, we may also believe our power calculations for the misspecification tests used, because the null model is contained by all of its encompassing models. We may in fact be in a position to say that we have evidence for the hypothesis for which we claimed evidence in the original inference *and* for the statistical model on which that evidence claim depended. In this way, we have secured our original evidence claim by *strengthening* the support for its premises.

Things look rather different if the null model *fails* this misspecification test. In an NP test, data that leads to the rejection of the null hypothesis can potentially be interpreted as evidence supporting an alternative. In misspecification testing, this is not the case. In the absence of support for the null model, the adequacy of the encompassing model is also called into question. Thus, rejecting the null in a misspecification test that was designed to have power against alternatives in a particular direction "simply points the direction one should search for a better model" (Spanos, personal communication). Such information is useful for purposes of respecifying the assumed model. The methodology of respecification goes beyond the scope of the present paper. For our purposes it suffices to note that any such respecified model will itself need to be tested before it can be securely employed.

6 Conclusion

Given that both robust statistics and mis-specification testing serve the same purpose, it is natural to ask which approach is to be preferred in the pursuit of that aim. Answering that question is not the aim of the present paper. Prima facie, MST, precisely because it is not a weakening strategy, enjoys the advantage of greater efficiency. No satisfactory comparison could be made in the absence of considerations of computational costs, however. The comparison here is meant only to draw attention to two points: first, that both approaches can be viewed as pursuing the same aim of securing evidence claims; second, that by examining *how* the approaches differ in their pursuit of security, we can see that they exemplify the two different general strategies of securing evidence here discussed.[18]

To see the constrast between the two strategies, consider the situation of the researcher who seeks to draw inferences from a body of data using some statistical model. Supposing an initial model to be postulated, perhaps on the basis of a combination of plausibility and convenience considerations, the researcher is then faced with the problem that, for all she knows, that model might well be wrong. The Huber/Hampel approach would have her consider a range of epistemically possible error scenarios in which the postulated model is wrong, and then seek an estimator or test statistic that would allow her to draw weaker evidential conclusions that would remain sound across that range, as opposed to the stronger (but possibly false) conclusions that could be drawn using a procedure that is optimal for the postulated model. The MST approach, by contrast, would advise the researcher to subject the postulated model to a series of tests against epistemically possible errors in particular directions. Such testing would lead either to the validation of the postulated model, or to the respecification of the posulated model, whereupon the MST procedure would be reiterated, until at length a model would be specified that would withstand and be validated by such testing. By thus strengthening the support for the model employed, one would be in a position to derive the strongest possible conclusion from the data compatible with one's own reliability standards. Of course, there is nothing to prevent the researcher from drawing upon *both* strategies, such as by applying robustness considerations to a model that has been subjected to misspecification testing.

However one views the relative merits of Huber/Hampel robustness theory vs. the MST tesing approach, it is clear that the context for both belongs to the stage of inquiry in which one is engaged, not in the *use* of a reliable inferential process, but in the scrutiny, relative to one's epistemic situation, of the possible modes of error for the assessment of such a process's reliability. For an advocate of the ES theory of evidence, which employs reliability as the core objective and unrelativized notion behind the evidential relationship, either approach could be used to enhance security as a mode of evidential assessment that is relativized to epistemic situation. Thus, both the application of robustness theory and the MST methodology belong to that stage of inquiry that is sometimes referred to as "model criticism," which can be described in terms of a shift of perspective on the part of the investigator from "tentative sponsor to tentative critic" (Box and Tiao 1973, 8). In neither approach discussed here is model criticism carried out blindly, but rather rests upon a prior reflection on what is and is not known about the possible sources and modes of error in an initial set of assumptions.

7 Funding

This work was supported by the National Science Foundation [grant number SES-0750691].

8 Acknowledgments

I am grateful to Deborah Mayo, Jan Sprenger, Aris Spanos, and two anonymous referees for their helpful suggestions for this paper, and also for stimulating comments from Teddy Seidenfeld, Michael Strevens, Paul Weirich, and audiences at the Missouri Philosophy of Science Workshop, the Second Valencia Meeting on Philosophy, Probability, and Methodology, and the 2009 EPSA meeting.

References

Achinstein, P. (2001). *The Book of Evidence.* Oxford University Press, New York.

Box, G. E. P. and Tiao, G. C. (1973). *Bayesian Inference in Statistical Analysis.* Addison-Wesley, Reading, Mass.

Chalmers, D. (2011). The nature of epistemic space. In Egan, A. and Weatherson, B., editors, *Epistemic Modality*. Oxford University Press, Oxford.

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

DeRose, K. (1991). Epistemic possibilities. *The Philosophical Review*, 100:581–605.

Fisher, R. A. (1949). *The Design of Experiments*. Hafner Publishing Co., New York, fifth edition.

Hampel, F. (1968). *Contributions to the Theory of Robust Estimation*. PhD thesis, University of California, Berkeley.

Hampel, F. (1971). A general qualitative definition of robustness. *The Annals of Mathematical Statistics*, 42:1887–96.

Hampel, F. (1974). The influence curve and its role in robust estimation. *Journal of the American Statistical Association*, 69:383–93.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., and Stahel, W. A. (1986). *Robust Statistics: The Approach Based on Influence Functions*. John Wiley and Sonss, New York.

Hintikka, J. (1962). *Knowledge and Belief: An Introduction to the Logic of the Two Notions*. Cornell University Press, Ithaca, NY.

Huber, P. (1964). Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35:73–101.

Huber, P. (1981). *Robust Statistics*. John Wiley and Sons, New York.

Kratzer, A. (1977). What 'must' and 'can' must and can mean. *Linguistics and Philosophy*, 1:337–55.

MacFarlane, J. (2011). Epistemic modals are assessment-sensitive. In Egan, A. and Weatherson, B., editors, *Epistemic Modality*. Oxford University Press, Oxford.

Magnus, J. R. (2007). Local sensitivity in econometrics. In Boumans, M., editor, *Measurement in Economics: A Handbook*, pages 295–319. Academic, Oxford.

Magnus, J. R. and Vasnev, A. L. (2007). Local sensitivity and diagnostic tests. *Econometrics Journal*, 10:166–92.

Mayo, D. G. (1992). Did Pearson reject the Neyman-Pearson philosophy of statistics? *Synthese*, 90:233–62.

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

Mayo, D. G. and Spanos, A. (2004). Methodology in practice: Statistical misspecification testing. *Philosophy of Science*, 71:1007–1025.

Mayo, D. G. and Spanos, A. (2006). Severe testing as a basic concept in a Neyman-Pearson philosophy of induction. *The British Journal for the Philosophy of Science*, 57(2):323–357.

Neyman, J. (1950). *First Course in Probability and Statistics.* Henry Holt, New York.

Neyman, J. (1955). The problem of inductive inference. *Communications on Pure and Applied Mathematics*, VIII:13–46.

Pearson, E. S. (1962). Some thoughts on statistical inference. *Annals of Mathematical Statistics*, 33:394–403.

Spanos, A. (1999). *Probability Theory and Statistical Inference.* Cambridge University Press, Cambridge.

Sprenger, J. (2010). Science without (parametric) models: the case of bootstrap resampling. *Synthese*, forthcoming.

Staley, K. (2008). Error-statistical elimination of alternative hypotheses. *Synthese*, 163:397–408.

Staley, K. and Cobb, A. (2010). Internalist and externalist aspects of justification in scientific inquiry. *Synthese*, pages 1–18. 10.1007/s11229-010-9754-y.

Stigler, S. (1973). Simon Newcomb, Percy Daniell, and the history of robust estimation 1885–1920. *Journal of the American Statistical Association*, 68:872–79.

Tukey, J. (1960). A survey of sampling from contaminated distributions. In Olkin, I., editor, *Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling*, pages 448–85. Stanford University Press, Stanford, CA.

Notes

[1] The precise sense, however, in which objectivity can be predicated of the ES account, however, is a subtle and disputed issue. Suffice it here to note two points. First, the error probabilities that figure into the *application* of criteria $SR1$ and $SR2$ are, as will be discussed, predicated on a statistical model on which the statistical inference is premised. Much of the present paper is concerned with statistical procedures for coping with the possible inadequacy of these premises for the kind of inference that the investigtor seeks to draw. Second, although the very fact that such premises are susceptible to the kind of criticism discussed here would seem to indicate that some notion of objectivity is applicable, the exact sense in which such a notion applies is not a simple matter of "correspondence to the facts." As will be seen, for example, the procedure to be here discussed for testing statistical models does not test for the truth of their defining assumptions but for their *statistical adequacy* (see section five). Here I simply note these issues, as the main concern of this paper is not objectivity *per se* (granting that the notion does play a role in the discussion), but rather the justification of statisitcal inferences.

[2] In epistemology, warrant is sometimes used to denote that which, in addition to truth, qualifies a belief as knowledge. For our purposes, it will suffice to regard the use of the term here as synonymous with justification.

[3] Mayo has in fact argued that Neyman and Pearson themselves should not be understood as having consistently advocated the orthodox N-P approach. Pearson distanced himself from the behavioristic interpretation typically associated with orthodox N-P (Mayo 1996, 1992, Pearson 1962), and Neyman advocated post-data power analyses similar to those

employed in ES (see, e.g., Mayo and Spanos 2006, Neyman 1955).

[4] Although degrees of severity as here deployed look mathematically like ordinary "p-values" as used in Fisherian significance testing, their methodological use in a post-data meta-statistical scrutiny distinguishes them from p-values.

[5] Just how to formulate the semantics of such statements is, however, contested (see, e.g., Kratzer 1977, DeRose 1991, Chalmers 2011). To note one difficulty for Hintikka's original understanding, consider the status of mathematical theorems. Arguably, if Goldbach's conjecture is true, then it does follow from what I know (though I do not realize this), if I know the axioms of number theory. Yet it also seems correct to say that it is possible, for all I know, that Goldbach's conjecture is false, even if I do know the axioms of number theory. More recently, contextualist and relativist approaches have been formulated (DeRose 1991, MacFarlane 2011). One reason for relativizing security to epistemic situations as characterized above, rather than to, e.g., sets of beliefs is that such approaches to modal semantics make factors beyond simply an agent's beliefs or knowledge relevant to the status of an epistemic modal proposition. Relativizing to epistemic situations makes the account sufficiently flexible to be compatible with such varied approaches to modal semantics.

[6] As suggested by an anonymous referee, one might wish to formulate the idea of security in terms of 'possible worlds' (as used by Lewis and many others) or 'state descriptions' (as used by Carnap). No difficulty seems to arise on either approach, provided one keeps firmly in view that the modality at issue is neither subjunctive nor logical, but epistemic in nature.

[7] In the history of statistics, Stigler traces the first mathematical contributions to robust estimation back to Laplace, but focuses on the work of Simon Newcomb and of P. J. Daniell as exemplars of early work on robust estimation that was both clear and rigorous.

[8] Henceforth, in making general points about robustness theory, I shall refer only to estimators. It must be borne in mind that robustness theory has been developed for testing as well as estimation and all the same general points obtain in that context, but with attention shifted from the properties of estimators to those of test statistics.

[9] Jan Magnus and his collaborators have pursued a distinct but related approach to this problem that is noteworthy but not explored in the present paper. Magnus advocates a model-perturbation approach to sensitivity analysis that studies "the effect of small changes in model assumptions on an estimator of a parameter of interest" (see, e.g., Magnus and Vasnev 2007, Magnus 2007).

[10] Here I discuss these developments in the context of frequentist statistics in the Neyman–Pearson tradition. However, robustness theory is also applicable in Bayesian settings and likelihood-based approaches (Hampel et al. 1986, 52–56). That this is so provides additional support to the argument above regarding the relevance of security for statistical approaches other than ES.

[11] Cf. Huber (1964). The discussion that follows also owes much to Hampel et al. (1986, esp. 36–39, 172–78).

[12] As Huber notes, this class turns out to include as special cases the sample mean ($\rho(t) = t^2$), the sample median ($\rho(t) =\mid t \mid$), and all maximum likelihood estimators ($\rho(t) = -\log f(t)$, where f is the assumed density of

the distribution).

[13] The following discussion owes much to (Huber 1981). Many technical details are omitted, as the aim is to convey an intuitive notion that only approximates the more rigorous mathematical approach taken by Hampel.

[14] Just what makes a function d "suitable" to be a distance function in this context, beyond some obvious but underdetermining constraints, is not perfectly clear. See Huber 1981, 25–34, for some functions that have received the attention of theorists.

[15] Note that this notion only serves to characterize robustness with respect to assumptions about the distribution, not about dependence or heterogeneity, since the definition assumes the data are distributed independently and identically.

[16] Fisher's approach contrasts with the N-P approach by eschewing teh specification of alternative hypotheses, emphasizing instead the effort to test the null hypothesis itself. Thus, when the null is rejected, no alternative is accepted, but rather the statistical signficance with which the null is rejected is reported. Fisherian testing also differs from a behaviorist construal of N-P insofar as its aim is not to control the cost of erroneous decisions, but to determine how well the observations agree with a particular hypothesis (see, e.g., Fisher 1949).

[17] The model in question is the Normal autoregressive model, and the optimal test is a t-test; see Spanos (1999, 757–60) for details.

[18] Another statistical methodology that can be put to use in securing inferences not discussed here is the use of nonparametric techniques, discussed recently by Jan Sprenger (2010).