Two Ways to Rule out Error: Severity and Security

Kent W. Staley
Saint Louis University

December 4, 2008

Abstract: I contrast two modes of error-elimination relevant to evaluating evidence in accounts that emphasize frequentist reliability. The contrast corresponds to that between the *use* of of a reliable inference procedure and the *critical scrutiny* of a procedure with regard to its reliability, in light of what is and is not known about the setting in which the procedure is used. I propose a notion of *security* as a category of evidential assessment for the latter. In statistical settings, robustness theory and misspecification testing exemplify two distinct strategies for securing statistical inferences.

Contents

1 Introduction

I highlight a distinction between two notions of error elimination at work in the error-statistical (ES) philosophy of scientific evidence advocated by Deborah Mayo, and discuss the methodological implications of that distinction. The shift between these two notions is connected with a shift from the context in which a model is *used* to that in which it is *criticized*. Correspondingly, they are distinguishable in that one notion is unrelatived, while the other relative to epistemic situation.

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 specific types of errors. Roughly, good tests in the ES view are those with appropriately low rates of error in indicating discrepancies from a family of competing hypotheses under consideration, 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, resulting in data E, E constitutes evidence

for H just in case:

SR1 E fits H, and

SR2 the probability of H passing T with an outcome such as E (i.e., one that fits H at least as well as E does), given that H is false, is very low (Mayo 1996, esp. 178–87).

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. That a particular probability distribution is an adequate model of such features is a *judgment* that the researcher must make from the perspective of her epistemic situation; it is held as a virtue of the approach that such judgments are themselves potentially erroneous (they may fail to correspond to the facts) but also corrigible (further testing, if carried out reliably, will probably bring judgment into closer correspondence to the facts).

Here I seek to articulate an additional dimension of epistemic appraisal, which I call *security*, to complement that embodied in the ES account of evidence, and to relate that dimension to efforts in theoretical statistics to systematize the critical assessment of model assumptions. In particular, I consider the development of robust statistics and the program of misspecification testing/model respecification as two modes of response, in the context of *model criticism*, to the problem of insecure evidence.

I proceed as follows: In section two I seek to clarify the epistemic character of ES evidence by situating it within what I call a "reliable indicator" view of evidence. A central characteristic of reliability in these accounts is that it is to be understood in terms of *frequencies* of various types of error. In section three I contrast two perspectives on claims about ES evidence, insofar as these involve reliability. Roughly put, it is the distinction between whether a reliability claim is true, a matter of fact that is unrel-

ativized to any epistemic situation, and whether a given agent is justified in making an evidence claim. I argue that the latter is not only relativized to epistemic situation, but that it involves a category of epistemic appraisal distinct from reliability. I propose a definition for this category of epistemic appraisal, which I call the *security* of inferences or evidence claims. Section four surveys some work in the field of *robust statistics*, which I present as a theoretical approach to securing inferences by means of a strategy of weakening one's conclusions. In section five I discuss the contrasting approach of misspecification testing, presented as an example of security through the strengthening of support for premises. I conclude in section six by considering the prospects for a systematic approach to security, focusing on the lessons that we might draw from work in robustness theory and misspecification testing.

2 Reliable indicators

In this section, I present Mayo's ES account of scientific evidence as a *reliable indicator* (RI) account. By this I mean that it incorporates two related yet distinct conceptual dimensions: *reliability* and *semiotic function*. Both of these can be understood as characteristics of the testing procedures from which ES evidence derives.

Consider a rather ordinary kind of reliable indicator: a magnetic compass. In describing a particular compass as a reliable indicator, I indicate that, used appropriately, one can take the direction of its needle as indicating the approximate direction of magnetic north (this is its semiotic function), and that in doing so, one would not at all often draw the wrong conclusion in this regard (this is reliability). Obviously, reliability is a matter of degree. Moreover, whether we choose to call a particular indicator reliable will rest on contextual matters: a degree of reliability sufficient for finding one's way back to the trail head may not suffice for guiding a nuclear submarine through a coral reef. That said, whether a particular indicator is

sufficiently reliable to meet a particular standard is an objective matter with a dual basis in the physical functioning of the device itself and the procedure employed in arriving at conclusions on the basis of that functioning.

The ES account of evidence can be thought of as a kind of generalization of the way in which we use such every-day devices. Although ES evidence is not restricted to contexts in which quantitative statistical measures are employed, an overtly statistical example may help to make the conceptual components most apparent, as well as pave the way for the discussion robust statistical estimation to come.

Suppose that we are interested in estimating the "location parameter" μ_x for a distribution function F governing a series of random variables $X_1, X_2, \ldots, X_n$. We know that F is normal, with unknown mean μ_x and unknown variance σ_0^2. We might follow the standard instructions for constructing a confidence interval as such: First, we use as an estimator for the population mean μ_x the sample mean $\bar{x} = n^{-1} \sum_1^n x_i$, and as an estimator for the population variance the sample variance $s_x^2 = (n-1)^{-1} \sum_i^n (x_i - \bar{x})^2$. Now, suppose that we already know what standard of reliability we have in mind. For example, we might want to make the estimate by means of an instrument with no greater than a 5% error rate. Using the quantities $\bar{x}$ and $s_{\bar{x}} = n^{-1/2} s_x$ (the standard deviation of the sample mean), and with the help of a statistical table giving t-variate probabilities, one can determine a confidence interval with just that characteristic. Specifically, it can be shown that, under these assumptions,

$$Pr(\bar{x} - t_{\alpha/2} s_{\bar{x}} < \mu_x < \bar{x} + t_{\alpha/2} s_{\bar{x}}) = 1 - \alpha. \tag{1}$$

Supposing then that we let $\alpha = 0.05$ and that $n = 20$, this becomes

$$Pr(\bar{x} - 2.093 s_{\bar{x}} < \mu_x < \bar{x} + 2.093 s_{\bar{x}}) = 0.95. \tag{2}$$

This result derives from the distribution of the estimator itself, under the assumed underlying distribution F, and can be understood as follows: The method of constructing an interval estimate just described is such that

95% of the intervals thus constructed will include the true value of the location parameter μ_x. Our conclusion might then be that the data give evidence that $\mu_x = \bar{x} \pm 2.093 s_{\bar{x}}$.

Here again, as in the example of the compass, we have a semiotic function (i.e., the interval constructed) which is an indicator that points to a range of possible values for μ_x. Moreover, the limiting of the probability of error can serve to ensure the reliability of this semiotic function. The test's semiotic function and its reliability are determined by the probabilistic dependence on the value of μ_x of the values recorded for $x_1, x_2, \ldots, x_n$ and by the appropriate choice of an estimator (itself in turn dependent on those same values) so as to achieve the requisite limits on error rates. As in the case of the compass, the appropriate standard of reliability may be contextual, but whether the standard is met by the test as constituted is an objective matter in that it is independent of what anyone knows or believes about the test or the hypotheses under consideration.

To put the point differently, the investigator employing a particular statistical testing procedure judges its reliability by means of the probability values calculated according to the statistical model employed (in the present example, this model includes the use of the normal distribution, for example). That the model used is an adequate representation of the relevant aspects of the testing procedure is a judgment made by the investigator. It is precisely because reliable semiotic function is an objective matter that there is always the possibility of such a judgment being in error. Moreover, as emphasized by Mayo, such errors themselves may be discovered and corrected. Therefore, those claims about the evidence relevant to a particular hypothesis that depend on such tests are also susceptible to error — and those errors, too, may be discovered and corrected. Mayo quotes approvingly Henry Kyburg's assertion that this possibility of error is "almost a touchstone of objectivity" (Kyburg 1993, 147, quoted in Mayo 1996, 83).

Thus, the first notion of error-elimination at work in the ES account is

the notion of *using a severe error probe*, i.e., a rule for drawing conclusions about hypotheses via the evidence given by experimentally-generated data that, when applied to hypotheses about a particular question, would only rarely lead one to infer a conclusion, supposing it to be false. It is this that allows the investigator to draw an inference with the assurance that she has, with high probability, ruled out an erroneous conclusion. Determining whether a given testing procedure meets that severity standard requires relying on some assumptions. The problem can be seen most clearly in the kind of quantitative statistical context exemplified in the estimation example given above, where the relevant assumptions serve to define a statistical model. In such cases, the distinction at issue is that between using the model and criticizing it. Two questions immediately arise concerning model criticism. First, what does it mean for a model to be adequate? Second, how can one establish the adequacy of a model?

I propose here to show how a notion of error-elimination distinct from the severe-test notion just mentioned plays a central role both in characterizing model adequacy and in the evaluation of statistical models (hence also in the evaluation of evidence claims that rely upon such models).

The kind of uncertainty that is addressed by error probabilities is of the following sort: Is the procedure presently used to draw conclusions one that is very often leads to errors? How probable is it, given the procedure at hand, that one will draw an incorrect conclusion? The researcher's situation is also characterized by uncertainty of a different sort, which model criticism seeks to address: What are the various ways in which error might arise, and which are ruled out by what is already known? Answers to these questions are necessarily relative to epistemic situation of the researcher in a way that error rates are not. Or, to put it in another light, ES evidential *relations*, being independent of epistemic situations, enjoy a more robust objectivity than evidential *claims*, the epistemic status of which is relative to some epistemic situation.[1] In the next section I will elaborate on this

[1]Note that this is a different distinction than that drawn by Peter Achinstein between

point, introducing a concept that will help to clarify the nature of the latter perspective on the epistemic status of evidence claims.

3 Security in the evaluation of evidence claims

The viewpoint of the researcher making an evidential judgment thus brings into perspective two distinct notions of error elimination not previously distinguished by ES advocates. The first 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 ones epistemic situation. The second is relativized: one eliminates error by showing that, given what one knows, the ways in which ones premises or underlying assumptions might be wrong can be ruled out, or else make no difference to the evidential conclusion one is drawing.

I propose a division of labor between these two notions. In accordance with ES, the first, unrelativized notion of frequency reliability is appropriate to the concept of evidence itself. The second notion is appropriate for the appraisal of claims *about* evidence (or reliability). I employ the term *security* for this latter concept.

The concept of security that I develop is meant to capture an intuitive notion regarding how investigators make claims about evidence. Let an *evidence claim* be a claim of the form 'Data E (resulting from test T) are evidence for the hypothesis that H.' At the time that such a claim is made, the claimant will believe or rely upon many propositions, some of which, for all she knows, may be wrong. Assuming she does not want to make an

relativized and unrelativized evidence concepts (Achinstein 2001). Achinstein's point is that evidence concepts of both kinds are invoked on different occasions by scientists, a point that I do not dispute. My point is that, even restricting ourselves to the kind of unrelativized evidence concept that is the subject of the ES theory, the epistemic status (as opposed to simply the truth) of evidence claims using that unrelativized concept is relative to epistemic situation.

evidence claim that might subsequently be refuted, she has reason therefore to consider the ways in which her claim might fail. More specifically, she should, I argue, wish for her evidence claims to be secure in the following sense:

Definition 1 (secure evidence) *Suppose that Ω_0 is the set of all epistemically possible scenarios relative to epistemic situation K, and $\Omega_1 \subseteq \Omega_0$. An evidence claim C is* secure throughout Ω_1 *relative to K iff for any scenario $\omega \in \Omega_1$, C is true. If C is secure throughout Ω_0 then it is* fully secure.[2]

Before proceeding, some explanation of terminology is in order. Following Chalmers (2008), I use the term *scenario* to refer to what might be intuitively thought of as a "maximally specific way things might be." In practice, no one ever considers scenarios as such, of course, but rather focuses on salient differences between scenario and another. Scenarios thus function rather like possible worlds, although here the relevant modality is distinct from the subjunctive use to which possible worlds are typically put.

The modality of interest here is *epistemic possibility*, which can be thought of as the modality invoked in such expressions as "For all I know, there might be a third-generation leptoquark with a rest of mass of 250 GeV/c^2" and " For all I know, I might have left my sunglasses on the train." Hintikka, whose (1962) provides the origins for contemporary discussions, took 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." Just how to formulate the semantics of such statements is, however, contested (see, e.g., DeRose 1991 and Chalmers 2008).[3] The central claims

[2]There may be reasons for the investigator to consider someone else's epistemic situation to be more relevant than her own. The choice of the relevant epistemic situation for a given evidence claim is an outstanding problem of the theory of secure reasoning.

[3]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

of the present proposal are independent of disputed issues regarding the semantics of epistemic possibility.

Finally, the notion of an *epistemic situation* 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 the hypothesis" (ibid., 20). To this I would add as components of the epistemic situation that one knows (or does not know) how to do things (such as the manipulation of data or instruments, or the performance of speech acts) that facilitate the inference from data and other propositions to the hypothesis of interest.

The basic idea is that an evidence claim is 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*: An investigator can make her evidence claim *more secure* either by decreasing the range of epistemically possible scenarios so as to exclude some in which her claim is false, or by expanding, across the range of possible scenarios, the scope of those in which the claim she makes is true.

I contend that numerous scientific practices already aim at enhancing the security of evidence claims.We can tentatively classify such practices within some broad categories, such as strategies for *weakening* an evidential conclusion, for *strengthening* the support for assumptions employed in evaluating evidence, and for arguing from *robustness*, in the sense of appealing to convergent results from independent tests (Staley 2004).[4]

is false, even if I do know the axioms of number theory.

[4]It appears to be an unfortunate coincidence that the term 'robust' and its cognates enters this discourse in two rather distinct roles, one coming from philosophy of science, and the other from statistics. Here the emphasis will be on the sense of the term as used

The present paper focuses on the strengthening and weakening strategies (see Staley 2004 and 2008b for a discussion of the robustness strategy). The security framework here proposed allows for a unified understanding of these strategies. 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 (M-S) and model respecification advocated by Aris 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. I argue that viewing both approaches as efforts to address the problem of securing evidence claims yields insight into the evaluation of scientific knowledge.

4 Security through robust statistics

In this section, I argue that the development of robust statistics serves as an example of how security considerations can guide (even if implicitly) the development of rigorous theoretical frameworks with epistemic advantages over their non-secure counterparts.

Robust statistics originates in the insight that many classical statistical procedures depend upon parametric models that may hold only approximately. Although one might hope that when those models are approximately valid, so are the conclusions drawn, this is often not the case. In an oft-cited

by statisticians.

and clever paper, Tukey (1960) considered the following situation, prompted by his work at the end of World War II analyzing data on the effectiveness of bomber machine-gun fire against attacking fighters. Suppose there are two normal populations with identical means, but where the standard deviation of one is three times larger than another, and suppose that a large sample of data is generated from a population that is a mixture of the narrower population with some small "contamination" from the wider population. (Then, of course, the population sampled is no longer described by a single normal distribution; the contamination data might be thought of as outliers relative to the original narrower distribution.) Suppose, further, that one wishes to estimate the scale parameter for the population sampled. Tukey notes that for a normal distribution, the relative efficiency of the mean deviation as compared to the standard deviation, as estimators of scale, is 88%. But he shows that, not only does the addition of contributions from the wider distribution to the narrower distribution render the mean deviation more efficient, but that the point at which the mean deviation just matches the standard deviation in efficiency is when a mere .008 of the population sampled comes from the wider distribution.

What Tukey's discussion shows is that, in spite of the fact that much of statistical practice rests on the use of statistical measures the properties of which are determined under the assumption that some parametric statistical model of the population holds exactly, small departures from such an exact model can have dramatic effects on the performance of such measures. 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

Awareness of these problems significantly pre-dates Tukey's work. 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).[5] 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. Here I will survey some influential robustness notions that originated in the 1960s in work by Peter Huber (1964) and Frank Hampel (1968; 1971; 1974).

Many theoretical advances have been made since that early work, and robustness has been extended beyond simple one-dimensional estimation problems to multi-dimensional and testing contexts, but I will here simply discuss some of the early developments on one-dimensional estimators. My aim is not to survey the state of robust statistical theory, but to instead 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 can serve as an exemplar for further systematic thinking about security.

4.1 Huber's minimax approach

In his groundbreaking 1964 paper, Peter Huber introduced a class of estimators that he called "M-estimators."[6] Huber introduces these as a kind of generalization of least-squares estimators. Consider, in our original example attempting to estimate the location parameter of the distribution F, our choice of test statistic $T = \frac{1}{n}\Sigma_i x_i$. This emerges as the solution to a

[5] 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.

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

problem of minimizing the sum of the squares of the differences between the observed values and those that would be predicted under the hypothesis chosen by that estimator (the "errors"). 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 ρ of the errors (possibly not the sum of their squares). I.e., M-estimators are those functions that minimize $\Sigma_i \rho(x_i - T)$, for some non-constant function ρ.[7] Huber's motivation here is initially just that "It is quite natural to ask whether one can obtain more robustness by minimizing another function of the errors" (Huber 1964, 74). Tukey and others had already noticed 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 of it's solving a particular minimization problem, perhaps alternative, more robust estimators would emerge as solutions to alternative, but related, minimization problems.

Of course, to determine whether this is the case, one needs some means of evaluating robustness. Here it should be noted that Huber's discussion is not perfectly general, but 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." (Note that although H is assumed unknown, 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.

[7] 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.

More specifically: suppose that ψ 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 ψ_n). Then the asymptotic variance of ψ 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 F_0, as $n \to \infty$, i.e., $V(\psi, G) = E_{n\to\infty}[(\psi_n - E(\psi_n))^2]$. Then the most robust M-estimator for a given family F of distributions would be that which minimizes the maximal asymptotic variance across $\mathscr{P}_\epsilon$. Huber's approach, in other words, is to select as most robust that M-estimator ψ_0 that satisfies the condition:

$$\sup_{G\in\mathscr{P}_\epsilon} V(\psi_0, G) = \min_{\psi} \sup_{G\in\mathscr{P}_\epsilon} V(\psi, G) \tag{3}$$

Huber then goes on to show, among many other important results, that the solution to this problem corresponds to determining first the "least favorable distribution" F_0, which is the distribution that minimizes the Fisher information over all $G \in \mathscr{P}_\epsilon$. The estimator that satisfies the robustness criterion above is then the maximum likelihood estimator for that least favorable distribution. Intuitively, the approach is to pick the approach that is the optimum choice for the "worst case scenario," i.e., the scenario in which the observed random variable is the least informative about the value of the estimated parameter.

4.2 Hampel's influence function/breakdown point approach

Beginning in his 1968 thesis and in a series of subsequent papers (Hampel 1968; 1971; 1974), Frank Hampel laid the foundations for the "infinitesimal" approach to robust statistics, beginning as Huber did with one-dimensional estimation problems. Huber's approach begins by replacing the usual exact parametric model with a model of indeterminacy (originally, a normal distribution with a specified degree of "contamination") and then seeks to formulate a generalized minimization problem for that particular model,

Hampel's approach begins with an exact parametric model (not necessarily normal) and then considers the behavior of estimators in "neighborhoods" of that model.

First consider a qualitiative definition of robustness, as introduced in Hampel (1971).[8] 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,[9] for any $\epsilon > 0$, there is a $\delta > 0$, and an $n_0 > 0$, such that for all distributions F and all $n \geq n_0$,

$$d(F_0, F) \leq \delta \Rightarrow d(\mathscr{L}_{F_0}(T_n), \mathscr{L}_F(T_n)) \leq \epsilon \tag{4}$$

To express qualitative robustness intuitively, Hampel's definition requires that an estimator be such that closeness of the assumed distribution of the observations to their actual distribution ensures that the assumed distribution of the estimator is close to its actual distribution.

Such a definition allows for the systematic use of the designation "robust," but one might also wish to know *how much* difference a particular error in one's assumptions will make to the behavior of an estimator or test statistic T. Hampel introduced the notion of the influence function (IF) to address specifically the question of how much the value of T would change with the addition of a single new data point with a particular value x. The motivation seems in particular to have been to deal with questions of how to handle gross errors that turn up as outliers in the data. (The sample mean, for example, as a location estimate, responds dramatically to the addition

[8] 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.

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

of a single observation with x large relative to the rest of the sample.)

In his first publication on what he was then calling the "influence curve," after having introduced the notion in his 1968 dissertation, Hampel described it as "essentially the first derivative of an estimator, viewed as a functional, at some distribution" (Hampel 1974, 383). More specifically, the following definition is the one Hampel gives for dealing with an estimator functional T, a probability measure F on a subset of the real line R, and $x \in R$:

$$\mathrm{IF}_{T,F}(x) = \lim_{\epsilon \downarrow 0} \frac{T((1-\epsilon)F + \epsilon\delta_x) - T(F)}{\epsilon} \tag{5}$$

where δ_x denotes the pointmass 1 at x.

In practice, the importance of the influence function lies in various 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." (Here I will only introduce them with their intuitive interpretations. Mathematical definitions are given in the appendix; all of their definitions involve the influence function.)

The first ("and most important," according to Hampel et al. 1986, 87) of these derived concepts is the *gross-error senstivity* γ^*, a measure of the "worst (approximate) influence which a small amount of contamination of fixed size can haver 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" – what Hampel calls the results of "throwing in" operations (Hampel 1974, 387).

A rather different concern motivates the use of the *local-shift sensitivity* λ^*. Here the concern is with the effects of small changes in the values of observations, such as might result from either rounding or grouping of observations, among other sources. Supposing that one thinks of such a change in terms of removing an observation at point x and replacing it with an

observation at a neighboring point y, one can think of this as asking about the change in the estimate brought about by such a change, standardized by dividing out the difference between y and x. Local-shift sensitivity is thus a a "measure for the worst (approximate and standardized) effect of 'wiggling"' (Hampel et al. 1986, 88; also Hampel 1974, 389).

Finally, the *rejection point* ρ^* can be used to describe approaches to estimation that simply *reject* outliers – the most time-honored approach to robust estimation, whether based on "objective" or "subjective" criteria. 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 $\rho^* = \infty$.

It should be noted that the influence function and its associated robustness measures are all *local* in the sense that they are evaluated at a particular distribution, with the effects of deviations from that distribution evaluated in a piecemeal manner. In order to arrive at a *global* characterization of an estimator, Hampel introduced the *breadown point*, a measure that "describes up to what distance from the model distribution the estimator still gives some relevant information" (Hampel et al. 1986, 96), in the sense of "excluding part of the parameter space" (Hampel 1971, 1894). Hampel has also stresed the usefulness of the breakdown point as guiding how far from the assumed model F the IF can be used (Hampel et al. 1986, 41).

The theoretical interest of robustness theory in statistics derives from its methodological significance: In practice, data analysis often uses estimators or test statistics[10] 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

[10]Henceforth, in making general points about robustness theory, I shall refer only to estimators. It must be born 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.

hold for these estimators are an indicator of the evidential strength provided by the results of their application – but only if those properties really to hold. In many situations in which calculations based on a parametric model attribute such reliability properties to an estimator (and hence the results of its application), the model does not in fact hold exactly, and in many of *those* situations, the result is that the attributed reliability properties do not even hold approximately.

Robust statistics responds to this 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 parametric 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 (parametric statistical) assumptions might be wrong.

Without invoking security explicitly under the terms I have used to characterize it, robustness theorists have shown how to treat problems of the security of statistical inference in a systematic way. Specifically, the mathematical frameworks above provide frameworks for building models of various ways in which a particular parametric model used to calculate the long-run error behavior of an estimator might fail, so as to permit understanding of how such failure influences the behavior of that estimator. The particular robustness concepts developed seem to depend both on their suitability for capturing the relevant aspects of prominent model-defeating scenarios (gross errors, contamination, etc.) and on their mathematical tractability: at least in principle, one can use these concepts for discussing the behavior of estimators within such scenarios.

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 can be seen very clearly by

considering Hampel's comparison of the robustness properties of the mean to those of others at the Standard Normal distribution (Table 1, based on a similar table in Hampel 1974). Apart from the local-shift sensitivity λ^* (typically used to evaluate sensitivity to rounding errors), the mean fares poorly in comparison to the robustness properties of some other common estimators. It is the only one of these to fail to be qualitatively robust, and has a strong susceptibility to gross errors. (Since none of these estimators is defined to reject values on the basis of their magnitude as such, they all have infinite rejection points.) However, the mean has one very strong advantage *at the Normal distribution*, which is that its variance is so much smaller, making it a much more efficient estimator than its more robust counterparts.

As the emphasis in that last sentence indicates, this last advantage is illusory if in fact the process generating data is not adequately modeled using the Normal distribution. The use of a more robust estimator is then 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 sharply distributed, but more 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), but relies 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 valid. 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

estimator	qr[a]	σ^2	γ^*	λ^*	ρ^*
mean	−	1.00	∞	1.00	∞
Hodges-Lehmann[b]	+	1.047	1.77	1.41	∞
median	+	1.571	1.25	∞	∞
5% trim[c]	+	1.026	1.83	1.11	∞
10% trim[d]	+	1.060	1.60	1.25	∞

Table 1: Robustness properties of some common estimators at the Normal distribution (based on Hampel 1974)

[a] qr = qualitative robustness
[b] median of pairwise means of observations
[c] mean after smallest/largest $[.05n]$ observations are removed
[d] mean after smallest/largest $[.10n]$ observations are removed

possible scenarios, focus on determining a statistically adequate model, and then choose the optimal inferential strategy for that model.

A final note regarding these robustness notions. The basic strategy employs models of error that incorporate their own assumptions. For example, recall that Huber's initial work on M-estimators used an error model that assumed the contaminating distribution was symmetric, an assumption unlikely to be exactly met in most applications. Of course, this was an early attempt, and subsequent work by Huber and others has extended the mathematical treatment of security to more general scenarios involving much weaker assumptions. Nonetheless, the point remains that more definite statements regarding security can be made when one has more resources for representing what one does not know.

5 Security through misspecification testing

As argued by Aris Spanos (2008), such robustness arguments suffer from two disadvantages. The first is that just noted: applying the mathematical tools of robustness theory typically requires considerable knowledge of the nature

of the error in the original model, in particular the "form and structure of potential misspecifications". In the case where we lack such knowledge, those tools are inapplicable and the tendency to invoke robustness nonetheless leads to a "false sense of security" (ibid., 22). In the case we are able to determine the nature of the problem, this will be precisely through some sort of testing of the original model, just as advocated by the misspecification testing (M-S) approach, and the natural next step would be, not to use the less efficient robust estimators, but to respecify the model and choose an optimal estimator based on the new, statistically adequate model.

The second problem noted by Spanos is that both Huber's minimax approach and Hampel's influence function approach are based on changes in or distance measures applied to *distributions as a whole* – i.e., the assumed vs. the actual distribution characterizing the asymptotic behavior of estimators – when what is relevant to the evaluation of evidence in the error-statistical setting is not the entire distribution, but rather the error probabilities. Thus the basis for robustness assessment regarding claims about error-statistical evidence should be the sensitivity of the error probabilities to epistemically possible flaws in the assumed model.[11]

Thus, Spanos (1999; Mayo and Spanos 2004) argues that the appropriate strategy for addressing possible departures from assumed parametric models is to carry out a systematic approach to testing those models (misspecification testing – M-S), replacing the model, if necessary, with one that is more statistically adequately (model respecficiation).

A full explanation of the M-S testing approach would go beyond the

[11]Indeed, this is the way in which robustness often is considered when evaluating the sensitivity of particular inferences to departures from model assumptions. Consider, for example, G.E.P. Box's (1953) demonstration that analysis of variance tests using Bartlett's modification of Neyman and Pearson's L_1 test that involve more than two variances are very *non-robust* with regard to departures from Normality. The first table in the paper shows, for various values of kurtosis, how the true probability of exceeding a nominal 0.05 significance level using Bartlett's test statistic can vastly exceed 0.05, and the more so the larger the number of variances being compared (Box 1953).

aims of the present paper. My procedure here will be to discuss M-S testing in general terms, with attention to its aims, and the theoretical apparatus it employs.[12] The point I wish to emphasize is that M-S testing, like the minimax and infinitesimal approaches to robustness, arises from the need to address the security of evidence claims and their associated inferences. Understanding the epistemological difficulty that M-S and robustness theory aim to address will facilitate the evaluation of their quite different approaches to the problem.

By its nature, M-S testing calls for testing outside of the original parametric model. Indeed, because M-S 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 absence of a parametrization of 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 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-

[12] The discussion here follows closely that of Spanos (1999, 729–65).

Pearson test, but as a kind of *ansatz* to allow for the kind of operationalization of testing that a strict Fisher-type test does not allow. In other words, rather than ad hoc scrutiny of single assumptions, Spanos's M-S testing 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 M-S test, as required in Neyman-Pearson approaches. Nonetheless, Spanos insists, these are not true Neyman-Pearson tests because the context is one in which one is explicitly open 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 M-S testing 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).

The statistical model in our example was the simple Normal model and comprises probabilistic assumptions falling into three categories. Regarding *distribution*, the model assumes that random variables $X_1, \ldots, X_n$ are all **N**ormally distributed. The *dependence* assumption is that $X_1, \ldots, X_n$ are probabilistically **I**ndependent. Finally, the model assumption regarding *heterogeneity* is that all random variables $X_1, \ldots, X_n$ are **I**dentically **D**istributed (hence the abbreviation **NIID**). The aim of M-S would then 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 M-S testing 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

$$f(x_1, x_2, \ldots, x_T; \boldsymbol{\phi}) = \prod_{i=1}^{T} f_t(x_t; \boldsymbol{\psi}_t) \text{ for all } t \in \mathbb{T}, \\ \text{and all } \mathbf{x} := (x_1, \ldots, x_T) \in R, \tag{6}$$

our new assumption would be that of Markov dependence:

$$f_k(x_k|x_{k-1}, x_{k-2}, \ldots, x_1; \boldsymbol{\phi}_k) = f_k(x_k|x_{k-1}; \boldsymbol{\psi}_k), k = 2, 3, \ldots. \tag{7}$$

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*:

$$X_t = \alpha_0 + \alpha_1 X_{t-1} + u_t, t \in \mathbb{T} \tag{8}$$

(here u_t is the error term).

These modifications amount to the specification of an encompassing model (the Normal autoregressive model) that allows one to carry out a test of 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 $H_1 : \alpha_1 \neq 0$. As Spanos explains, the optimal test here is a t-test using an appropriately defined test statistic (see appendix for details). Moreover, the Normal autoregressive model can easily be extended to capture higher order Markov dependence, thus allowing for an optimal test of the null against such alternatives by means of the F-test (ibid., 757–60).

This brings us naturally to the question of what to do with the results of such tests. Although the mathematical apparatus is precisely that of the Neyman-Pearson approach that is directed at underwriting the kind of reliable indicator evidence claims at the heart of the ES 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 M-S testing 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 M-S testing is different, as it serves only to allow for the development of tests that *potentially* have high power in testing the null model (the assumed model of the original inference) 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, at least 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 deficient. This supports at least the provisional endorsement of the power assessments of the M-S 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 M-S 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 M-S 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 reliable evidence, not only for the hypothesis for which we claimed evidence in the original inference, but also for the statistical model on which that original reliability assessment depended. In this way, we have secured our original evidence claim by *strengthening* the support for our original premises.

Things look rather different if the null model *fails* this M-S test. In an NP test, data that leads to the rejection of the null hypothesis can potentially be interpreted as evidence supporting the alternative, or some subset of the alternative in the case of a compound alternative. In M-S 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 an M-S 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.

The contrast with the Huber/Hampel approach can now be seen quite clearly, if we 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 M-S approach, by contrast, would advise the researcher to subject the posulated 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 M-S 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.

However one views the relative merits of Huber/Hampel robustness theory vs. the M-S 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 M-S testing 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.

6 What next?

It should be clear by now that the most pressing problem for any attempt to theorize systematically about security is the relevance problem: When making an evidence claim, an agent need not worry equally about *all* the ways in which she might err. There are possibilities of error, after all, that are quite remote, and as Peirce pointed out long ago, the mere possibility of error is not by itself grounds for genuine doubt.[13] A sensible approach to evaluating the security of evidence claims would seem to call for some sort of ranking of which scenarios most demand scrutiny.

Here it must be acknowledged that the Bayesian enjoys an advantage. The Bayesian apparatus comes equipped already with a measure over a set of propositions, and isn't that just what is needed?

Before proceeding to take up this question, let me hasten to note that I do not propose here to consider the merits of an overall Bayesian approach to scientific evidence. From the outset this discussion has been concerned with articulating a more complete epistemological understanding of ES evidence and how it is evaluated. Not only is the Bayesian account (whether subjective or objective) not compatible with ES account, it is not any kind of reliable indicator view of evidence, in the frequentist sense of reliability here considered.

So the question for the present discussion is really whether an advocate of an ES view of evidence should adopt a supplementary Bayesian framework for the evaluation of evidence claims from the perspective of the epistemic agent, as called for in section two?

Here there are two routes one might contemplate: the subjective or the objective understanding of Bayesian probability. The first does such

[13]Indeed, it is important to note that a policy of withholding evidence claims unless *all possible* ways of going wrong have been eliminated would introduce a new form of *unreliability*: one would, with probability one, *fail* to infer a correct hypothesis from any data, no matter how compelling (see Mayo 1996; Staley 2008).

conceptual violence to the ES framework that is our starting point, that an ES advocate should reject it outright. The latter, objective approach, might be made compatible with the ES approach, if confined to this secondary evaluative role, but brings along its own foundational difficulties.

The subjective Bayesian approach would presumably be to weight the error-scenarios according to the agent's personal probabilities for those scenarios. Suppose that investigator S makes an evidence claim on the basis that a particular hypothesis H has passed a severe test. On this view, such a claim would be justified for S, even in the absence of any testing of the assumed model of the test, and even though that model made rather strong and possibly false assumptions, simply on the grounds that S attaches only a negligible degree of belief to any of the scenarios in which the assumed model is false. This should have little appeal for the ES advocate who holds the truth conditions for evidence claims to be objective matters of fact independent of belief, since it juxtaposes a strongly objective view of the content of evidence claims with a strongly subjective view of their justification. By contrast, the security concept here advocated, although it is relativized to an epistemic situation, is nonetheless objective insofar as what is epistemically possible relative to an epistemic situation is an objective matter, at least in the sense of depending on what an epistemic agent knows. (Here I do assume that there is more to knowing than simply believing very strongly. At a minimum, to know that p requires that p is true.)

Turning to objective versions of Bayesianism, things look rather different. Indeed, beyond just compatibility, there might even seem to be a family resemblance between security and logical probability. Just as an evidence claim is secure to the extent that it is true over all the scenarios that are possible relative to an epistemic situation, logical probability has often been framed in terms of the satisfaction of a formula by a class of models consistent with a certain body of background knowledge. Thus one might

entertain a kind of two-probability approach like Carnap's (1962) in order to rank those scenarios that are "most compatible" with a given epistemic situation.

This challenge raises problems that cannot be satisfactorily addressed in a brief discussion. Let me for now simply observe that, whatever the advantages of such approach, it amounts to a multiplication of foundational challenges, insofar as one adds the notorious problems of fixing a prior distribution, determining a likelihood for the catchall, etc., to whatever conceptual problems might be raised for frequentists. It is hard enough to defend *one* interpretation of probability!

Better, then, to follow the examples of the model criticism approaches surveyed here. One can, like the robustness theorists, try to "break off" particular categories of error-scenarios that can be represented in way that allows them to be treated more or less rigorously. One can furthermore, as the M-S approach advises, extend frequentist modes of testing to those modes of insecurity that relate directly to the assumptions (involving distributions, independence, and homogeneity) that define statistical models.

However, let us not stop there. Robust statistics emphasizes the readily quantifiable aspects of security appraisal. As important as this is, investigators also must, and often do, reflect on possible errors that are not readily quantifiable in this way. Furthermore, possibilities of error that cannot be approached quantitatively may nevertheless be approached systematically. Returning to Deborah Mayo's *Error and the Growth of Knowledge*, she there called for the articulation of "canonical models of error" (Mayo 1996, e.g., 450–51). By now we have seen how, in addition the canonical parametric models so commonly used in statistical data analysis, robustness theorists introduced additional canonical models of how *those* models might be violated, to facilitate the investigation of the behavior (whether asymptotic or finite) of various estimators.

But just as there are qualitative, "informal" approaches to testing a

hypothesis reliably (as when we give our students a test that it would be hard for them to pass if they did not know the material), so there are ways to secure our conclusions from severe tests that are not readily modeled in a mathematical framework (such as when we space out their desks, thus securing our estimate of the severity of the test based on its difficulty against defeat due to cheating). To advance the cause of such informal, qualitative efforts at securing our evidence claims, it may be less important to develop sophisticated mathematical theories, and more important to reflect, as experimentalists have always done, on a kind of typology of causes of error in different kinds of experimental undertakings. This kind of enterprise has been joined by a handful of philosophers, pursuing various philosophical agendas. A concern with the security of evidence might provide a setting in which the work of various philosophers of science who have not embraced error-statistics can be seen as nonetheless contributing to it (see, e.g., Hon 1998; 2003; Franklin 1986; 2002; Schickore 2005, among others).

However, for such categorization to constitute a real advance, I propose that we not rest content with compiling a kind of catalogue of types of errors – rather the goal should be render such a catalogue useful for the planning of experiment and the appraisal of experimental evidence. As in the example of robustness theory, this requires that we not merely consider the causes of error, but also its *effects*, and that we seek to draw general conclusions about those.

7 Appendix

7.1 The weak topology and distance measures

Suppose that Ω denotes a topological space (i.e., a set on which a function d is defined such that for any two points in Ω the four conditions on a distance function given above are satisfied) that is complete (every Cauchy sequence in Ω is convergent) and separable (Ω has a dense countable subset), and

that $\mathscr{B}$ is the Borel-σ-algebra on Ω (i.e., the σ-algebra generated by the open subsets of Ω). Let $\mathscr{M}$ be the space of all probability measures on $(\Omega, \mathscr{B})$.

Supposing that F and G are distribution functions, the *Lévy distance* between F and G is defined to be

$$d_L(F,G) \equiv \inf\{\epsilon | \forall x F(x-\epsilon) - \epsilon \leq G(x) \leq F(x+\epsilon) + \epsilon\}. \tag{9}$$

Next, suppose that F and G are two probability measures in $\mathscr{M}$. The *Prohorov distance* between F and G is defined to be

$$d_P(F,G) = \inf\{\epsilon > 0 | F\{A\} \leq g\{A^\epsilon + \epsilon \text{forall} A \in \mathscr{B}\}. \tag{10}$$

Here A^ϵ is the closed ϵ-neighborhood of subset $A \subset \Omega$, defined as

$$A^\epsilon \equiv \{x \in \Omega | \inf_{y \in A} d(x,y) \leq \epsilon\}. \tag{11}$$

Finally, suppose that the distance function d in Ω is bounded by 1 (otherwise, replace d with $d' = d(x,y)/[1+d(x,y)]$ to obtain a distance function so bounded). Then consider the class Ψ of all functions ψ satisfying the Lipschitz condition: $|\psi(x) - \psi(y)| \leq d(x,y)$. The *bounded Lipschitz metric* is then defined as:

$$d_{BL}(F,G) = \sup_{\psi \in \Psi} |\int \psi dF - \int \psi dG|. \tag{12}$$

7.2 Definitions of IF-related concepts and breakdown point

Suppose that T is an estimator and F a distribution. Then the *gross-error sensitivity* for (T,F) is defined as:

$$\gamma^*(T,F) = \sup_x |\text{IF}_{T,F}(x)| \tag{13}$$

The *local-shift sensitivity* for (T,F) is defined as:

$$\lambda^*(T,F) = \sup_{x \neq y} |\text{IF}_{T,F}(y) - \text{IF}_{T,F}(x)|/|y-x| \tag{14}$$

Finally, the *rejection point* for (T, F) is defined as:

$$\rho^*(T, F) = \inf\{r > 0; \mathrm{IF}_{T,F}(x) = 0 \text{ when } |x| > r\} \tag{15}$$

Suppose that F and G are distributions in sample space $\mathscr{X}$, $\{T_n\}$ is a sequence of estimators, and Θ is a parameter space (e.g., the real number line R). Then the *breakdown point* ϵ^* *of* $\{T_n\}$ *at* F is defined as:

$$\begin{aligned}\epsilon^* \equiv \sup\{\epsilon \leq 1;\ &\text{there is a compact set } K_\epsilon \subsetneqq \Theta \text{ such that} \\ &d_P(F, G) < \epsilon \text{ implies } G(\{T_n \in K_\epsilon\}) \overset{n\to\infty}{\to} 1\}.\end{aligned} \tag{16}$$

Here $d_P(F, G)$ is the Prohorov distance between the distributions F and G. A finite-sample definition of the breakdown point involving no reference to probability distributions was introduced by Donoho and Huber (1983; see also Rousseeuw and Leroy 2003, 9-12). It is worth noting that an advantage of the breakdown point over the influence function is that it is defined for all estimators, whereas the IF is not.

7.3 The Normal autoregressive model and testing independence

As noted above, the optimal test of independence against Markov dependence within the Normal autoregressive model is a t-test. The test statistic for such a test is defined as $\tau(\mathbf{X}) = \frac{\sqrt{T}(\hat{\alpha}_1)}{s}$, where $\hat{\alpha}_1 = \frac{\sum_{t=1}^{T}(X_t - \bar{X})(X_{t-1} - \bar{X})}{\sum_{t=1}^{T}(X_{t-1} - \bar{X})^2}$, $s^2 = \frac{1}{T-2}\sum_{t=1}^{T}(X_t - \hat{\alpha}_0 - \hat{\alpha}_1 X_{t-1})^2$, $\hat{\alpha}_0 = \bar{X} - \hat{\alpha}_1 \bar{X} - 1$, $\bar{X} = \frac{1}{T}\sum_{t=1}^{T} X_t$, and $\bar{X}_{-1} = \frac{1}{T-1}\sum_{t=1}^{T-1} X_t$. Under the null hypothesis, the quantity $\tau(\mathbf{X})$ is asymptotically approximated by Student's t-distribution with $n - 2$ degrees of freedom.

REFERENCES

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

Chalmers, David. 2008. "The Nature of Epistemic Space," URL (cited on June 26, 2008): `http://consc.net/papers/espace.html`.

DeRose, Keith. 1991. "Epistemic Possibilities," *The Philosophical Review* 100: 581–605.

Donoho, D. L. and Peter Huber. 1983. "The Notion of Breakdown Point." In P. Bickel, K. Doksum, and J. L. Hodges, Jr. (eds.) *A Festschrift for Erich Lehmann*. Belmont, California: Wadsworth.

Franklin, Allan. 1986. *The Neglect of Experiment*. New York: Cambridge University Press.

Franklin, Allan. 2002. *Selectivity and Discord: Two Problems of Experiment*. Pittsburgh, Pennsylvania: University of Pittsburgh Press.

Hampel, Frank. 1968. "Contributions to the Theory of Robust Estimation." Ph.D. thesis. University of California, Berkeley.

Hampel, Frank. 1971. "A General Qualitative Definition of Robustness." *The Annals of Mathematical Statistics* 42: 1887–96.

Hampel, Frank. 1974. "The Influence Curve and Its Role in Robust Estimation." *Journal of the American Statistical Association* 69: 383–93.

Hampel, Frank, Elvezio Ronchetti, Peter Rousseeuw, and Werner Stahel. 1986. *Robust Statistics: The Approach Based on Influence Functions*. New York: John Wiley and Sons.

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

Hon, Giora. 1998. "Exploiting Errors." *Studies in History and Philosophy of Science* 29A: 465–79.

Hon, Giora. 2003. "The Idols of Experiment: Transcending the 'Etc. List'." In Hans Radder (ed.), *The Philosophy of Scientific Experimentation*. Pittsburgh, Pennsylvania: University of Pittsburgh Press, 174–197.

Huber, Peter. 1964. "Robust Estimation of a Location Parameter." *The Annals of Mathematical Statistics* 35: 73–101.

Huber, Peter. 1981. *Robust Statistics*. New York: John Wiley and Sons.

Kyburg, Henry. 1993. "The Scope of Bayesian Reasoning." In *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Assocation*, vol, 2: Symposia and Invited Papers, pp. 139–52.

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

Mayo, Deborah and Aris Spanos. 2004. "Methodology in Practice: Statistical Misspecification Testing," *Philosophy of Science* 71: 1007–1025.

Mayo, Deborah 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: 323–57.

Rousseeuw, Peter and Annick Leroy. 2003. *Robust Regression and Outlier Detection*. Hoboken, NJ: John Wiley and Sons.

Schickore, Jutta. 2005. "Through Thousands of Errors We Reach the Truth – But How? On the Epistemic Roles of Error in Scientific Practice." *Studies in History and Philosophy of Science* 36A: 539–56.

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

Spanos, Aris. 2008. "Misspecification, Robustness, and the Reliability of Inference: The Simple t-Test in the Presence of Markov Dependence." Unpublished ms, available at `http://www.econ.vt.edu/faculty/2008vitas_research/spanos_working_papers/2Spanos-reliability.pdf`. Accessed December 4, 2008.

Staley, Kent. 2004. "Robust Evidence and Secure Evidence Claims." *Philosophy of Science* 71: 467–88.

Staley, Kent. 2008a. "Error-statistical Elimination of Alternative Hypotheses." *Synthese* 163: 397–408.

Staley, Kent. 2008b. "Securing Scientific Evidence." Unpublished ms, available at `http://staleykw.googlepages.com/securingevidence.pdf`.

Stigler, Sephen. 1973. "Simon Newcomb, Percy Daniell, and the History of Robust Estimation 1885–1920." *Journal of the American Statistical Association* 68: 872–79.

Tukey, John. 1960. "A Survey of Sampling from Contaminated Distributions." In Ingram Olkin et al. (eds.), *Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling.* Stanford, CA: Stanford University Press, pp. 448–85.