ERROR STATISTICS

Deborah G. Mayo and Aris Spanos

1 WHAT IS ERROR STATISTICS?

Error statistics, as we are using that term, has a dual dimension involving philosophy and methodology. It refers to a standpoint regarding both:

1. a cluster of statistical tools, their interpretation and justification,

2. a general philosophy of science, and the roles probability plays in inductive inference.

To adequately appraise the error statistical approach, and compare it to other philosophies of statistics, requires understanding the complex interconnections between the methodological and philosophical dimensions in (1) and (2) respectively. To make this entry useful while keeping to a manageable length, we restrict our main focus to (1) the error statistical philosophy. We will however aim to bring out enough of the interplay between the philosophical, methodological, and statistical issues, to elucidate long-standing conceptual, technical, and epistemological debates surrounding both these dimensions. Even with this restriction, we are identifying a huge territory marked by generations of recurring controversy about how to specify and interpret statistical methods. Understandably, standard explications of statistical methods focus on the formal mathematical tools without considering a general philosophy of science or of induction in which these tools best fit. This is understandable for two main reasons: first, it is not the job of statisticians to delve into philosophical foundations, at least explicitly. The second and deeper reason is that the philosophy of science and induction for which these tools are most apt—we may call it ***the error statistical philosophy*** —will differ in central ways from traditional conceptions of the scientific method. These differences may be located in contrasting answers to a fundamental pair of questions that are of interest to statisticians and philosophers of science:

- *How do we obtain reliable knowledge about the world despite uncertainty and threats of error?*

- *What is the role of probability in making reliable inferences?*

Handbook of the Philosophy of Science. Volume 7: Philosophy of Statistics.
Volume editors: Prasanta S. Bandyopadhyay and Malcolm R. Forster. General Editors: Dov M. Gabbay, Paul Thagard and John Woods.

To zero in on the main issues, we will adopt the following, somewhat unusual strategy: We shall first set out some of the main ideas within the broader error statistical philosophy of induction and inquiry, identifying the goals and requirements that serve both to direct and justify the use of formal statistical tools. Next we turn to explicating statistical methods, of testing and estimation, while simultaneously highlighting classic misunderstandings and fallacies. The error statistical account we advocate builds on Fisherian and Neyman-Pearsonian methods; see [Fisher, 1925; 1956; Neyman, 1952; Pearson, 1962]. While we wish to set out for the reader the basic elements of these tools, as more usually formulated, we will gradually transform them into something rather different from both Fisherian and Neyman-Pearsonian paradigms.

Our goals are twofold: to set the stage for developing a more adequate philosophy of inductive inquiry, and to illuminate crucial issues for making progress on the "statistical wars", now in their seventh decade[1].

1.1 The Error Statistical Philosophy

Under the umbrella of error-statistical methods, one may include all standard methods using error probabilities based on the relative frequencies of errors in repeated sampling – often called *sampling theory* or *frequentist statistics*. Frequentist statistical methods are sometimes erroneously equated to other accounts that employ frequentist probability, for example, the "frequentism" in the logic of confirmation. The latter has to do with using relative frequencies of occurrences to infer probabilities of events, often by the straight rule, e.g., from an observed proportion of As that are Bs to infer the proportion of As that are Bs in a population.

1. One central difference, as Neyman [1957] chided Carnap, is that, unlike frequentist logics of confirmation, frequentist statistics always addresses questions or problems within a *statistical model*[2] (or family of models) $\mathcal{M}$ intended to provide an approximate (and idealized) representation of the process generating data $\mathbf{x}_0$:=$(x_1, x_2, \ldots, x_n)$. $\mathcal{M}$ is defined in terms of $f(\mathbf{x};\theta)$, the probability distribution of the sample $\mathbf{X}$:=$(X_1, ..., X_n)$, that assigns probabilities to all events of interest belonging to the sample space $\mathbb{R}^n_X$. Formal error statistical methods encompass the deductive assignments of probabilities to outcomes, given a statistical model $\mathcal{M}$ of the experiment, and inductive methods from the sample to claims about the model. Statistical inference focuses on the latter step: moving from the data to statistical hypotheses, typically couched in terms of unknown parameter(s) θ, which governs $f(\mathbf{x};\theta)$.

[1] The 'first act' might be traced to the papers by Fisher [1955], Pearson [1955], Neyman [1956].

[2] Overlooking the necessity of clearly specifying the statistical model — in terms of a complete set of probabilistic assumptions — is one of the cardinal sins still committed, especially by non-statisticians, in expositions of frequentist statistics.

2. The second key difference is how probability arises in induction. For the error statistician probability arises not to measure degrees of confirmation or belief (actual or rational) in hypotheses, but to quantify how frequently methods are capable of discriminating between alternative hypotheses and how reliably they facilitate the detection of error. These probabilistic properties of inductive procedures are *error frequencies* or *error probabilities*.

The statistical methods of significance tests and confidence-interval estimation are examples of formal error-statistical methods. A statistical inference might be an assertion about the value of θ, say that $\theta > 0$. Error probabilities attach, not directly to $\theta > 0$, but to the inference tools themselves, whether tests or estimators. The claims concerning θ are either correct or incorrect as regards the mechanism generating the data. Insofar as we are interested in using data to make inferences about this mechanism, in this world, it would make no sense to speak of the relative frequency of $\theta > 0$, as 'if universes were as plenty as blackberries from which we randomly selected this one universe', as Peirce would say (2.684). Nevertheless, error probabilities are the basis for determining whether and how well a statistical hypothesis such as $\theta > 0$ is warranted by data $\mathbf{x}_0$ at hand, and for setting bounds on how far off parameter values can be from 0 or other hypothetical values. Since it is the use of frequentist error probabilities, and not merely the use of frequentist probability that is central to this account, the term *error statistics* (an abbreviation of error probability statistics) seems an apt characterization.

Statistical Significance Test

Formally speaking, the inductive move in error statistics occurs by linking special functions of the data, $d(\mathbf{X})$, known as *statistics*, to hypotheses about parameter(s), θ of interest. For example, a test might be given as a rule: whenever $d(\mathbf{X})$ exceeds some constant c, infer $\theta > 0$, thereby rejecting $\theta = 0$:

Test Rule: whenever $\{d(\mathbf{x}_0) > c\}$, infer $\theta > 0$.

Any particular application of an inductive rule can be 'in error' in what it infers about the data generation mechanism, so long as data are limited. If we could calculate the probability of the event $\{d(\mathbf{X}) > c\}$ under the assumption that θ=0, we could calculate the probability of erroneously inferring $\theta > 0$. Error probabilities are computed from the distribution of $d(\mathbf{X})$, the *sampling distribution*, evaluated under various hypothesized values of θ. The genius of formal error statistics is its ability to provide inferential tools where the error probabilities of interest may be calculated, despite unknowns.

Consider a simple and canonical example of a statistical test, often called a statistical significance test. Such a test, in the context of a statistical model $\mathcal{M}$, is a procedure with the following components:

1. a *null hypothesis* H_0, couched in terms of unknown parameter θ, and

2. a function of the sample, $d(\mathbf{X})$, the *test statistic*, which reflects how well or poorly the data $\mathbf{x}_0$ accord with the null hypothesis H_0 — the larger the value of $d(\mathbf{x}_0)$ the further the outcome is from what is expected under H_0 — with respect to the particular question being asked. A crucial aspect of an error statistical test is its ability to ensure the sampling distribution of the test statistic can be computed under H_0 and under hypotheses discrepant from H_0. In particular, this allows computing:

3. the *significance level* associated with $d(\mathbf{x}_0)$: the probability of a worse fit with H_0 than the observed $d(\mathbf{x}_0)$, under the assumption that H_0 is true:

$$p(\mathbf{x}_0) = P(d(\mathbf{X}) > d(\mathbf{x}_0);\ H_0).$$

This is known either as the *observed* significance level or the *p-value*. The larger the value of the test statistic the smaller the p-value. Identifying a relevant test statistic, together with the need to calculate the error probabilities, restricts the choices of test statistic so as to lead to a uniquely appropriate test, whether one begins with Neyman-Pearson (N-P) or Fisherian objectives [Cox, 1958].

Consider for example the case of a random sample $\mathbf{X}$ of size n from a Normal distribution with *unknown* mean μ and, for simplicity, *known* variance σ^2 (denoted by $\mathsf{N}(\mu, \sigma^2)$). We want to test the hypotheses:

(1) $H_0 : \mu = \mu_0$ vs. $H_1 : \mu > \mu_0$.

The test statistic of this one-sided test is $d(\mathbf{X})=\frac{(\overline{X}-\mu_0)}{\sigma_x}$, where $\overline{X}=\frac{1}{n}\sum_{k=1}^{n} X_k$ denotes the sample mean and $\sigma_x=(\sigma/\sqrt{n})$. Given a particular outcome $\mathbf{x}_0$, we compute $d(\mathbf{x}_0)$. An inductive or ampliative inference only arises in moving from $d(\mathbf{x}_0)$ — a particular outcome — to a hypothesis about parameter μ. Consider the test rule: whenever $\overline{X}$ exceeds μ_0 by $1.96\sigma_x$ or more, infer H_1: $\mu > \mu_0$. Use of the statistic $d(\mathbf{X})$ lets us write this test rule more simply:

Test Rule T: whenever $\{d(\mathbf{x}_0) > 1.96\}$, infer H_1: $\mu > \mu_0$.

We deductively arrive at the probability of the event $\{d(\mathbf{X}) > 1.96\}$ under the assumption that H_0 correctly describes the data generating process, namely $P(d(\mathbf{X}) > 1.96;\ H_0)$=.025, giving the statistical significance level .025.

Tests, strictly speaking, are formal mapping rules. To construe them as inference tools requires an interpretation beyond the pure formalism, and this can be done in various ways. For instance, a Fisherian may simply output "the observed significance level is .025"; a Neyman-Pearson tester, might report "reject H_0" having decided in advance that any outcome reaching the .025 significance level will lead to this output. But these reports themselves require fleshing out, and a good deal of controversy revolves around some of the familiar ways of doing so.

Behavioristic and evidential philosophies

By a "statistical philosophy" we understand a general conception of the aims and epistemological foundations of a statistical methodology. Thus an important task for an error statistical philosophy is to articulate the interpretation and the rationale for outputs of inductive tests. For instance, continuing with our example, a Fisherian might declare that whenever $\{d(\mathbf{x}_0) > 1.96\}$:

"infer that H_0 is falsified at level .025".

A strict Neyman-Pearsonian might identify 'rejectH_0' with a specific *action*, such as:

"publish the result" or "reject the shipment of bolts".

Consider a third construal: whenever $\{d(\mathbf{x}_0) > 1.96\}$ infer that the data $\mathbf{x}_0$ is evidence for a (positive) discrepancy γ from μ_0 :

"infer $\mathbf{x}_0$ indicates $\mu > \mu_0 + \gamma$".

The weakest claim is to infer *some* discrepancy — as with the typical non-null hypothesis. More informatively, as we propose, one might specify a particular positive value for γ. For each we can obtain error probabilistic assertions:

Assuming that H_0 is true, the probability is .025 that:

- H_0 is falsified (at this level),
- the shipment of bolts is rejected,
- some positive discrepancy γ from μ_0 is inferred.

Each of these interpretations demands a justification or rationale, and it is a crucial part of the corresponding statistical philosophy to supply it. Error statistical methods are typically associated with two distinct justifications:

Behavioristic rationale. The first stresses the ability of tests to control error probabilities at some low level in the long run. This goal accords well with what is generally regarded as Neyman's statistical philosophy wherein tests are interpreted as tools for deciding "how to behave" in relation to the phenomena under test, and are justified in terms of their ability to ensure low long-run errors. Neyman [1971] even called his tests tools for *inductive behavior*, to underscore the idea that the test output was an action, as well as to draw the contrast with Bayesian inductive inference in terms of degrees of belief.

Inferential rationale A non-behavioristic or inferential justification stresses the relevance of error probabilities for achieving inferential and learning goals. To succeed it must show how error probabilities can be used to characterize warranted inference in some sense. The key difference between behavioristic and inferential construals of tests is not whether one views an inference as a kind of decision (which one is free to do), but rather in the justificatory role given to error probabilities.

On the behavioristic philosophy, the goal is to adjust our behavior so that in the long-run we will not act erroneously too often: it regards low long-run error rates (ideally, optimal ones) alone as what justifies a method. This does not yield a satisfactory error statistical philosophy in the context of scientific inference. How to provide an inferential philosophy for error statistics has been the locus of the most philosophically interesting controversies. Although our main focus will be on developing an adequate inferential construal, there are contexts wherein the more behavioristic construal is entirely appropriate, and we propose to retain it within the error statistical umbrella[3]. When we speak of "the context of scientific inference" we refer to a setting where the goal is an inference about what is the case regarding a particular phenomenon.

Objectivity in error statistics

Underlying the error statistical philosophy, as we see it, is a conception of the objective underpinnings for uncertain inference: although knowledge gaps leave plenty of room for biases, arbitrariness, and wishful thinking, in fact we regularly come up against experiences that thwart our expectations, disagree with the predictions and theories we try to foist upon the world, and this affords objective constraints on which our critical capacity is built. Getting it (at least approximately) right, and not merely ensuring internal consistency or agreed-upon convention, is at the heart of objectively orienting ourselves toward the world. Our ability to recognize when data fail to match anticipations is what affords us the opportunity to systematically improve our orientation in direct response to such disharmony. Much as Popper [1959] takes the ability to falsify as the foundation of objective knowledge, R.A. Fisher [1935, p. 16] developed statistical significance tests based on his view that "every experiment may be said to exist only in order to give the facts the chance of disproving the null hypothesis". Such failures could always be avoided by "immunizing" a hypothesis against criticism, but to do so would prevent learning what is the case about the phenomenon in question, and thus flies in the face of the aim of objective science. Such a strategy would have a very high probability of saving false claims. However, what we are calling the error-statistical philosophy goes beyond falsificationism, of both the Popperian and Fisherian varieties, most notably in its consideration of what positive inferences are licensed when data do not falsify but rather accord with a hypothesis or claim.

Failing to falsify hypotheses, while rarely allowing their acceptance as precisely true, may warrant excluding various discrepancies, errors or rivals. Which ones?

[3]Even in science there are tasks whose goal is avoiding too much noise in the network.

Those which, with high probability, would have led the test to issue a more discordant outcome, or a more statistically significant result. In those cases we may infer that the discrepancies, rivals, or errors are ruled out with severity.

Philosophy should direct methodology (not the other way around). To implement the error statistical philosophy requires methods that can accomplish the goals it sets for uncertain inference in science. This requires tools that pay explicit attention to the need to communicate results so as to set the stage for others to check, debate, scrutinize and extend the inferences reached. Thus, any adequate statistical methodology must provide the means to address legitimate critical questions, to give information as to which conclusions are likely to stand up to further probing, and where weaknesses remain. The much maligned, automatic, recipe-like uses of N-P tests wherein one accepts and rejects claims according to whether they fall into prespecified 'rejections regions' are uses we would also condemn. Rather than spend time legislating against such tests, we set out principles of interpretation that automatically scrutinize any inferences based on them. (Even silly tests can warrant certain claims.) This is an important source of objectivity that is open to the error statistician: choice of test may be a product of subjective whims, but the ability to critically evaluate which inferences are and are not warranted is not.

Background knowledge in the error statistical framework of 'active' inquiry

The error statistical philosophy conceives of statistics very broadly to include the conglomeration of systematic tools for collecting, modeling and drawing inferences from data, including purely 'data analytic' methods that are normally not deemed 'inferential'. In order for formal error statistical tools to link data, or data models, to primary scientific hypotheses, several different statistical hypotheses may be called upon, each permitting an aspect of the primary problem to be expressed and probed. An auxiliary or 'secondary' set of hypotheses are needed to check the assumptions of other models in the complex network; see section 4. Its ability to check its own assumptions is another important ingredient to the objectivity of this approach.

There is often a peculiar allegation (criticism) that:

(#1)	**error statistical tools forbid using any background knowledge,**

as if one must start each inquiry with a blank slate. This allegation overlooks the huge methodology of experimental design, data analysis, model specification and work directed at linking substantive and statistical parameters. A main reason for this charge is that prior probability assignments in hypotheses do not enter into the calculations (except in very special cases). But there is no reason to suppose the kind of background information we need in order to specify and interpret statistical methods can or should be captured by prior probabilities in the hypotheses being studied. (We return to this in section 3). But background knowledge must enter

in designing, interpreting, and combining statistical inferences in both informal and semi-formal ways. Far from wishing to inject our background opinions in the hypotheses being studied, we seek designs that help us avoid being misled or biased by initial beliefs. Although we cannot fully formalize, we can systematize the manifold steps and interrelated checks that, taken together, constitute a full-bodied experimental inquiry that is realistic.

The error statistician is concerned with the critical control of scientific inferences by means of stringent probes of conjectured flaws and sources of unreliability. Standard statistical hypotheses, while seeming oversimple in and of themselves, are highly flexible and effective for the piece-meal probes the error statistician seeks. Statistical hypotheses offer ways to couch conjectured flaws in inference, such as:

mistaking spurious for genuine correlations,

mistaken directions of effects,

mistaken values of parameters,

mistakes about causal factors,

mistakes about assumptions of statistical models.

The qualities we look for to express and test hypotheses about such inference errors are generally quite distinct from those required of the substantive scientific claims about which we use statistical tests to learn. Unless the overall error statistical philosophy is recognized, the applicability and relevance of the formal methodology will be misunderstood, as it often is. Although the overarching goal of inquiry is to find out what is (truly) the case about aspects of phenomena, the hypotheses erected in the actual processes of finding things out are generally approximations (idealizations) and may even be deliberately false.

The picture corresponding to error statistics is one of an activist learner in the midst of an inquiry with the goal of finding something out. We want hypotheses that will allow for stringent testing so that if they pass we have evidence of a genuine experimental effect. The goal of attaining such well-probed hypotheses differs crucially from seeking highly probable ones (however probability is interpreted). We will say more about this in section 3.

1.2 An Error Statistical Philosophy of Science

The error statistical philosophy just sketched alludes to the general methodological principles and foundations associated with frequentist error statistical methods. By an error statistical philosophy of science, on the other hand, we have in mind the application of those tools and their interpretation to problems of philosophy of science: to model scientific inference (actual or rational), to scrutinize principles of inference (e.g., prefer novel results, varying data), and to frame and tackle philosophical problems about evidence and inference (how to warrant data, pinpoint blame for anomalies, test models and theories). Nevertheless, each of the points

in 1.1 about statistical methodology has direct outgrowths for the philosophy of science dimension. The outgrowths yield:

(i) requirements for an adequate philosophy of evidence and inference, but also

(ii) payoffs for using statistical science to make progress on philosophical problems.

(i) In order to obtain a philosophical account of inference from the error statistical perspective, one would require forward-looking tools for finding things out, not for reconstructing inferences as 'rational' (in accordance with one or another view of rationality). An adequate philosophy of evidence would have to engage statistical methods for obtaining, debating, rejecting, and affirming data. From this perspective, an account of scientific method that begins its work only once well-defined evidence claims are available forfeits the ability to be relevant to understanding the actual processes behind the success of science.

(ii) Conversely, it is precisely because the contexts in which statistical methods are most needed are ones that compel us to be most aware of strategies scientists use to cope with threats to reliability, that considering the nature of statistical method in the collection, modeling, and analysis of data is so effective a way to articulate and warrant principles of evidence.

In addition to paving the way for richer and more realistic philosophies of science, we claim, examining error statistical methods sets the stage for solving or making progress on long-standing philosophical problems about evidence and inductive inference.

- Where the recognition that data are always fallible presents a challenge to traditional empiricist foundations, the cornerstone of statistical induction is the ability to move from less to more accurate data.

- Where the best often thought feasible is getting it right in some asymptotic long-run, error statistical methods ensure specific precision in finite samples, and supply ways to calculate how large a sample size n needs to be.

- Where pinpointing blame for anomalies is thought to present insoluble *Duhemian problems* and *underdetermination*, a central feature of error statistical tests is their capacity to evaluate error probabilities that hold regardless of unknown background or nuisance parameters.

- Where appeals to statistics in conducting a meta-methodology too often boil down to reconstructing one's intuition in probabilistic terms, statistical principles of inference do real work for us — in distinguishing when and why violations of novelty matter, when and why irrelevant conjuncts are poorly supported, and so on.

Although the extended discussion of an error statistical philosophy of science goes beyond the scope of this paper (but see [Mayo, 1996; Mayo and Spanos, 2010]), our discussion should show the relevance of problems in statistical philosophy for addressing the issues in philosophy of science — which is why philosophy of statistics is so rich a resource for epistemologists. In the next section we turn to the central error statistical principle that links (1.1) the error statistical philosophy and (1.2) an error statistical philosophy of science.

1.3 The Severity Principle

A method's error probabilities describe its performance characteristics in a hypothetical sequence of repetitions. How are we to use error probabilities in making particular inferences? This leads to the general question:

> When do data $\mathbf{x}_0$ provide good evidence for, or a good test of, hypothesis H?

Our standpoint begins with the situation in which we would intuitively *deny* $\mathbf{x}_0$ is evidence for H. Data $\mathbf{x}_0$ fail to provide good evidence for the truth of H if the inferential procedure had very little chance of providing evidence against H, even if H is false.

> **Severity Principle (weak)**. Data $\mathbf{x}_0$ (produced by process G) do *not* provide good evidence for hypothesis H if $\mathbf{x}_0$ results from a test procedure with a very low probability or capacity of having uncovered the falsity of H, even if H is incorrect.

Such a test we would say is insufficiently stringent or severe. The onus is on the person claiming to have evidence for H to show that they are *not* guilty of at least so egregious a lack of severity. Formal error statistical tools are regarded as providing systematic ways to foster this goal, as well as to determine how well it has been met in any specific case. Although one might stop with this negative conception (as perhaps Fisher and Popper did), we will go on to the further, positive one, which will comprise the full severity principle:

> **Severity Principle (full)**. Data $\mathbf{x}_0$ (produced by process G) provides good evidence for hypothesis H (just) to the extent that test T severely passes H with $\mathbf{x}_0$.

Severity rationale vs. low long-run error-rate rationale (evidential vs. behavioral rationale)

Let us begin with a very informal example. Suppose we are testing whether and how much weight George has gained between now and the time he left for Paris, and do so by checking if any difference shows up on a series of well-calibrated and stable weighing methods, both before his leaving and upon his return. If no change

on any of these scales is registered, even though, say, they easily detect a difference when he lifts a .1-pound potato, then this may be regarded as grounds for inferring that George's weight gain is negligible within limits set by the sensitivity of the scales. The hypothesis H here might be:

H: George's weight gain is no greater than δ,

where δ is an amount easily detected by these scales. H, we would say, has passed a severe test: were George to have gained δ pounds or more (i.e., were H false), then this method would almost certainly have detected this.

A *behavioristic rationale* might go as follows: If one always follows the rule going from failure to detect a weight gain after stringent probing to inferring weight gain no greater than δ, then one would rarely be wrong in the long run of repetitions. While true, this is not the rationale we give in making inferences about George. It is rather that this particular weighing experiment indicates something about George's weight. The long run properties — at least when they are relevant for particular inferences — utilize error probabilities to characterize the capacity of our inferential tool for finding things out in the particular case. This is the *severity rationale*.

We wish to distinguish the severity rationale from a more prevalent idea for how procedures with low error probabilities become relevant to a particular application; namely, the procedure is rarely wrong, therefore, the probability it is wrong in this case is low. In this view we are justified in inferring H because it was the output of a method that rarely errs. This justification might be seen as intermediate between full-blown behavioristic justifications, and a genuine inferential justification. We may describe this as the notion that the long run error probability 'rubs off' on each application. This still does not get at the reasoning for the particular case at hand. The reliability of the rule used to infer H is at most a necessary and not a sufficient condition to warrant inferring H. What we wish to sustain is this kind of counterfactual statistical claim: that were George to have gained more than δ pounds, at least one of the scales would have registered an increase. This is an example of what philosophers often call an *argument from coincidence*: it would be a preposterous coincidence if all the scales easily registered even slight weight shifts when weighing objects of known weight, and yet were systematically misleading us when applied to an object of unknown weight. Are we to allow that tools read our minds just when we do not know the weight? To deny the warrant for H, in other words, is to follow a highly unreliable method: it would erroneously reject correct inferences with high or maximal probability (minimal severity), and thus would thwart learning. The stronger, positive side of the severity principle is tantamount to espousing the legitimacy of strong arguments from coincidence. What statistical tests enable us to do is determine when such arguments from coincidence are sustainable (e.g., by setting up null hypotheses). It requires being very specific about which inference is thereby warranted—we may, for example, argue from coincidence for a genuine, non-spurious, effect, but not be able to sustain an argument to the truth of a theory or even the reality of an entity.

Passing a Severe Test.

We can encapsulate this as follows:

A hypothesis H passes a severe test T with data $\mathbf{x}_0$ if,

(S-1) $\mathbf{x}_0$ accords with H, (for a suitable notion of accordance) and

(S-2) with very high probability, test T would have produced a result that accords less well with H than $\mathbf{x}_0$ does, if H were false or incorrect.

Equivalently, (S-2) can be stated:

(S-2)*: with very low probability, test T would have produced a result that accords as well as or better with H than $\mathbf{x}_0$ does, if H were false or incorrect.

Severity, in our conception, somewhat in contrast to how it is often used, is not a characteristic of a test in and of itself, but rather of the test T, a specific test result $\mathbf{x}_0$, and a specific inference H (not necessarily predesignated) being entertained. That is, the severity function has three arguments. We use the notation: $SEV(T, \mathbf{x}_0, H)$, or even $SEV(H)$, to abbreviate:

"The severity with which claim H passes test T with outcome $\mathbf{x}_0$".

As we will see, the analyses may take different forms: one may provide a series of inferences that pass with high and low severity, serving essentially as benchmarks for interpretation, or one may fix the inference of interest and report the severity attained.

The formal statistical testing apparatus does not include severity assessments, but there are ways to use the error statistical properties of tests, together with the outcome $\mathbf{x}_0$, to evaluate a test's severity in relation to an inference of interest. This is the key for the inferential interpretation of error statistical tests. While, at first blush, a test's severity resembles the notion of a test's power, the two notions are importantly different; see section 2.

The severity principle, we hold, makes sense of the underlying reasoning of tests, and addresses chronic problems and fallacies associated with frequentist testing. In developing this account, we draw upon other attempts to supply frequentist foundations, in particular by Bartlett, Barnard, Birnbaum, Cox, Efron, Fisher, Lehmann, Neyman , E. Pearson; the severity notion, or something like it, affords a rationale and unification of several threads that we have extracted and woven together. Although mixing aspects from N-P and Fisherian tests is often charged as being guilty of an inconsistent hybrid [Gigerenzer, 1993], the error statistical umbrella, linked by the notion of severity, allows for a coherent blending of elements from both approaches. The different methods can be understood as relevant for one or another type of question along the stages of a full-bodied inquiry. Within the error statistical umbrella, the different methods are part of the panoply of methods that may be used in the service of severely probing hypotheses.

A principle for interpreting statistical inference vs. the goal of science

We should emphasize at the outset that while severity is the principle on which interpretations of statistical inferences are based, we are *not* claiming it is the goal of science. While scientists seek to have hypotheses and theories pass severe tests, severity must be balanced with informativeness. So for example, trivially true claims would pass with maximal severity, but they would not yield informative inferences[4]. Moreover, one learns quite a lot from ascertaining which aspects of theories have *not* yet passed severely. It is the basis for constructing rival theories which existing tests cannot distinguish, and is the impetus for developing more probative tests to discriminate them (see [Mayo, 2010a]).

2 A PHILOSOPHY FOR ERROR STATISTICS

We review the key components of error statistical tests, set out the core ingredients of both Fisherian and N-P tests, and then consider how the severity principle directs the interpretation of frequentist tests. We are then in a position to swiftly deal with the specific criticisms lodged at tests.

2.1 Key Components of Error-Statistical Tests

While we focus, for simplicity, on inferences relating to the simple normal model defined in section 1.1, the discussion applies to any well-defined frequentist test.

A One-Sided Test T_α. We elaborate on the earlier example in order to make the severity interpretation of tests concrete.

EXAMPLE 1. Test T_α. Consider a sample $\mathbf{X} := (X_1, ..., X_n)$ of size n, where each X_n is assumed to be Normal ($\mathsf{N}(\mu, \sigma^2)$), Independent and Identically Distributed (NIID), denoted by:

$$\mathcal{M} : X_k \backsim \mathsf{NIID}(\mu, \sigma^2),\ -\infty<\mu<\infty, k=1, 2, \ldots, n, \ldots$$

Let μ be the parameter of interest and assume at first that σ^2 is known; this will be relaxed later. Consider the following null and alternative hypotheses of interest:

(2) $H_0 : \mu = \mu_0$ vs. $H_1 : \mu > \mu_0$.

The *test statistic* or distance function is: $d(\mathbf{X})=\frac{(\overline{X}-\mu_0)}{\sigma_x}$, where $\sigma_x=(\sigma/\sqrt{n})$. Note that $d(\mathbf{X})$ follows a common pattern for forming such a measure of discordance:

> [estimated — expected (underH_0)], measured in standard deviation units.

[4]For an extreme case, demonstrating that not-H yields a logical contradiction shows, with maximal severity, that H is tautologous.

Under the null, $d(\mathbf{X})$ is distributed as standard Normal, denoted by:

$$d(\mathbf{X})=\frac{(\overline{X}-\mu_0)}{\sigma_x} \backsim \mathsf{N}(0,1).$$

This is a very important result because it enables one to ensure that:

$$P(d(\mathbf{X}) > c_\alpha;\ H_0) = \alpha.$$

For instance, c_α for $\alpha = .025$ is 1.96; see figures 1(a)-(d). We also have:

$$P(\text{observing a } p\text{-value } \leq \alpha) \leq \alpha.$$

In the simple Fisherian test, the p-value indicates the level of inconsistency between what is expected and what is observed in the sense that the smaller the p-value the larger the discordance between $\mathbf{x}_0$ and H_0 [Cox, 1958]. If the p-value is not small, if it is larger than some threshold α (e.g., .01) then the disagreement is not considered strong enough to indicate evidence of departures from H_0. Such a result is commonly said to be insignificantly different from H_0, but, as we will see, it is fallacious to automatically view it as evidence *for* H_0. If the p-value is small enough, the data are regarded as grounds to reject or find a discrepancy from the null. Evidence against the null suggests evidence for *some* discrepancy from the null, although it is not made explicit in a simple Fisherian test.

Reference to 'discrepancies from the null hypothesis' leads naturally into Neyman-Pearson [1933] territory. Here, the falsity of H_0 is defined as H_1 the complement of H_0 with respect to the parameter space Θ. In terms of the p-value, the Neyman-Pearson (N-P) test may be given as a rule:

> if $p(\mathbf{x}_0) \leq \alpha$, reject H_0 (infer H_1); if $p(\mathbf{x}_0) > \alpha$, do not reject H_0.

Equivalently, the test fixes c_α at the start as the cut-off point such that any outcome smaller than c_α is taken to "accept" H_0. Critics often lampoon an automatic-recipe version of these tests. Here the tester is envisioned as simply declaring whether or not the result was statistically significant at a fixed level α, or equivalently, whether the data fell in the rejection region.

Attention to the manner in which tests are used, even by Neyman, however, reveals a much more nuanced and inferential interpretation to which these formal test rules are open. These uses (especially in the work of E. Pearson) provide a half-way house toward an adequate inferential interpretation of tests:

> Accept H_0: statistically *insignificant result* — "decide" (on the basis of the observed p value) that there is insufficient evidence to infer departure from H_0, and
>
> Reject H_0: statistically *significant result* — "decide" (on the basis of the observed p-value) that there is some evidence of the falsity of H_0 in the direction of the alternative H_1.

Although one could view these as decisions, we wish to interpret them as inferences. All the N-P results would continue to hold with either construal.

The N-P test rule: Reject H_0 iff $d(\mathbf{x}_0) > c_\alpha$,

ensures the probability of rejecting (i.e., declaring there is evidence against) H_0 when H_0 is true — a type I error — is α. Having fixed α, the key idea behind N-P tests is to minimize the probability of a *type II error* (failing to reject H_0 when H_1 is true), written as $\beta(\mu_1)$:

$$P(d(\mathbf{X}) \leq c_\alpha; \mu{=}\mu_1) = \beta(\mu_1), \text{ for any } \mu_1 \text{ greater than } \mu_0.$$

That is, the test *minimizes* the probability of finding "statistical agreement" with H_0 when some alternative hypothesis H_1 is true. Note that the set of alternatives in this case includes all $\mu_1 > \mu_0$, i.e. H_1 is a *composite* hypothesis, hence the notation $\beta(\mu_1)$. Equivalently, the goal is to maximize the *power of the test*, for a fixed c_α :

$$POW(T_\alpha; \mu_1) = P(d(\mathbf{X}) > c_\alpha; \mu_1), \text{ for any } \mu_1 \text{ greater than } \mu_0.$$

In the behavioristic construal of N-P tests, these goals are put in terms of wishing to avoid behaving erroneously too often. But, the tests that grow out of the requirement to satisfy the N-P pre-data, long-run desiderata often lead to a uniquely appropriate test, whose error probabilities simultaneously can be shown to satisfy severity desiderata. The severity construal of N-P tests underscores the role of error probabilities as measuring the 'capacity' of the test to detect different discrepancies $\gamma \geq 0$ from the null, where $\mu_1{=}(\mu_0{+}\gamma)$. The power of a 'good' test is expected to increase with the value of γ.

Pre-data, these desiderata allow us to ensure two things:

(i) a rejection indicates with severity *some* discrepancy from the null, and

(ii) failing to reject the null rules out with severity those alternatives against which the test has high power.

Post-data, one can go much further in determining the magnitude γ of discrepancies from the null warranted by the actual data in hand. That will be the linchpin of our error statistical construal. Still, even N-P practitioners often prefer to report the observed p-value rather than merely whether the predesignated cut-off for rejection has been reached, because it "enables others to reach a verdict based on the significance level of their choice" [Lehmann, 1993, p. 62]. What will be new in the severity construal is considering sensitivity in terms of the probability of $\{d(\mathbf{X}) > d(\mathbf{x}_0)\}$, under various alternatives to the null rather than the N-P focus on $\{d(\mathbf{X}) > c_\alpha\}$. That is, the error statistical construal of tests will require evaluating this 'sensitivity' post-data (relative to $d(\mathbf{x}_0)$, not c_α); see [Cox, 2006].

We now turn to the task of articulating the error-statistical construal of tests by considering, and responding to, classic misunderstandings and fallacies.

2.2 How severity gives an inferential interpretation while scotching familiar fallacies

Suppose the observed p-value is .01. This report might be taken to reject the null hypothesis H_0 and conclude H_1. Why? An N-P *behavioristic rationale* might note that deciding to interpret the data this way would rarely be wrong. Were H_0 true, so large a $d(\mathbf{x}_0)$ would occur only 1% of the time. In our inferential interpretation, the fact that the p-value is small ($p(\mathbf{x}_0) = .01$) supplies evidence for H_1 because H_1 has passed a severe test: with high probability $(1 - p(\mathbf{x}_0))$ such an impressive departure from H_0 would not have occurred if H_0 correctly described the data generating procedure. The severity definition is instantiated because:

> (S-1): $\mathbf{x}_0$ accords with H_1, and (S-2): there is a high probability (.99) that a less statistically significant difference would have resulted, were H_0 true.

This is entirely analogous to the way we reasoned informally about George's weight. Granted, evidence from any one test might at most be taken as some evidence that the effect is genuine. But after frequent rejections of H_0, H_1 passes a genuinely severe test because, were H_1 false and the null hypothesis H_0 true, we would very probably have obtained results that accord less well with H_1 than the ones we actually observed. So the p-value gives the kind of data-dependency that is missing from the coarse N-P tests, and it also lends itself to a *severity construal* — at least with respect to inferring the existence of *some* discrepancy from the null. We have an inferential interpretation, but there are still weaknesses we need to get around. A pair of criticisms relating to statistically significant results, are associated with what we may call "fallacies of rejection".

2.3 Fallacies of rejection (errors in interpreting statistically significant results)

First there is the weakness that, at least on an oversimple construal of tests:

(#2)	**All statistically significant results are treated the same,**

and second, that:

(#3)	**The p-value does not tell us how large a discrepancy is found.**

We could avoid these criticisms if the construal of a statistically significant result were in terms of evidence for a particular discrepancy from H_0 (an effect size), that is, for inferring: $H{:}\mu > \mu_1 = (\mu_0 + \gamma)$, (there is evidence of a discrepancy γ).

The severity reasoning can be used to underwrite such inferences about particular discrepancies $\gamma \geq 0$ from the null hypothesis, i.e., $\mu > (\mu_0 + \gamma)$. For each

result we need to show: (a) the discrepancies that are not warranted, and (b) those which are well warranted. The basis for doing so is summarized in (a) and (b):

(a) If there is a very high probability of obtaining so large a $d(\mathbf{x}_0)$ (even) if $\mu \leq \mu_1$, then $SEV(\mu > \mu_1)$ is low. By contrast:

(b) If there is a very low probability of obtaining so large a $d(\mathbf{x}_0)$ if $\mu \leq \mu_1$, then $SEV(\mu > \mu_1)$ is high.

There are two key consequences. First, two different statistically significant results are distinguished by the inferences they severely warrant (criticism #2). Second, for any particular statistically significant result, the severity associated with $\mu > \mu_2$ will differ from (be less than) that associated with $\mu > \mu_1$, for any μ_2 greater than μ_1 (criticism #3).

Let us illustrate in detail with reference to our test T_α of hypotheses:

$$H_0 : \mu = 0 \text{ vs. } H_1 : \mu > 0.$$

For simplicity, let it be known that σ=2, and suppose n=100, i.e. σ_x= .2. Let us call a result "statistically significant" if it is statistically significant at the .025 level, i.e., $d(\mathbf{x}_0) > 1.96$. To address criticism #2, consider three different significant results: $d(\mathbf{x}_0)$=2.0 ($\overline{x}$=0.4), $d(\mathbf{x}_0)$=3.0 ($\overline{x}$=0.6), $d(\mathbf{x}_0)$=5.0 ($\overline{x}$=1.0).

Each statistically significant result "accords with" the alternative ($\mu > 0$). So (S-1) is satisfied. Condition (S-2) requires the evaluation of the probability that test T_α would have produced a result that accords *less well* with H_1 than $\mathbf{x}_0$ does (i.e. $d(\mathbf{X}) \leq d(\mathbf{x}_0)$), calculated under various discrepancies from 0. For illustration, imagine that we are interested in the inference $\mu > .2$. The three different statistically significant outcomes result in different severity assignments for the same inference $\mu > .2$.

Begin with $d(\mathbf{x}_0)$=2.0 ($\overline{x}$=0.4). We have:

$$SEV(\mu > .2) = P(\overline{X} < 0.4; \mu > .2 \text{ is false}) = P(\overline{X} < 0.4; \mu \leq .2 \text{ is true}).$$

Remember, we are calculating the probability of the event, $\{\overline{X} < .4\}$, and the claim to the right of the ";" should be read "calculated under the assumption that" one or another values of μ is correct. How do we calculate $P(\overline{X} < .4; \mu \leq .2$ is true) when $\mu \leq .2$ is a composite claim? We need only to calculate it for the point μ=.2 because μ values less than .2 would yield an even higher SEV value. The severity for inferring $\mu > .2$, when $\overline{x}$=.4 is $SEV(\mu > .2) = .841$. This follows from the fact that the observation $\overline{x}$=.4 is one standard deviation (σ_x=.2) in excess of .2. The probability of the event ($\overline{X} > .4$) under the assumption that μ=.2 is .16, so the corresponding SEV is .841. By standardizing the difference ($\overline{x}-\mu$), i.e. define a *standardized Normal* random variable $Z = \frac{(\overline{x}-\mu)}{\sigma_x} \backsim \mathsf{N}(0,1)$, one can read off the needed probabilities from the standard Normal tables. Figures 1(a)-(d) show the probabilities beyond 1 and 2 standard deviation, as well as the .025 and .05 thresholds, i.e. 1.645 and 1.96, respectively.

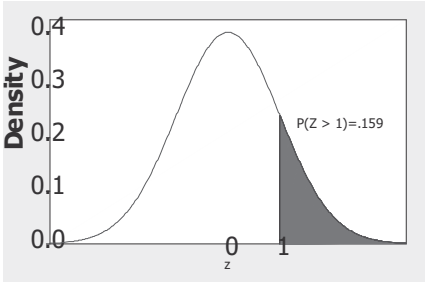


Fig 1a. $N(0,1)$: Right tail probability beyond 1 (one) standard deviation (SD).

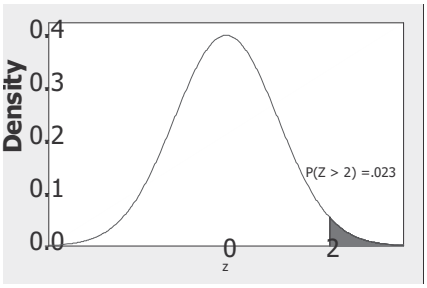


Fig 1b. $N(0,1)$: Right tail probability beyond 2 (two) standard deviations.

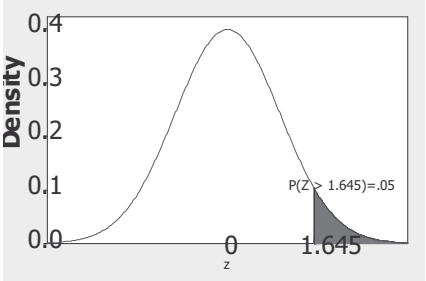


Fig 1c. $N(0,1)$: 5% right tail probability.

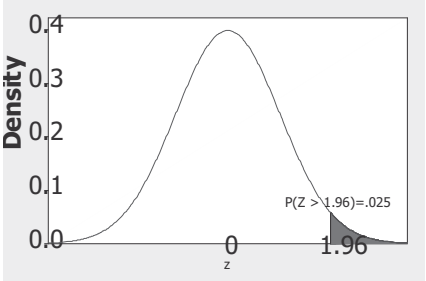


Fig 1d. $N(0,1)$: 2.5% right tail probability.

Figure 1. Tail area probabilities of the standard normal $N(0,1)$) distribution

Now, let us consider the two other statistically significant outcomes, retaining this same inference of interest. When $\overline{x}$=.6, we have $SEV(\mu > .2)$=.977, since $\overline{x}$=.6 is 2 standard deviation in excess of the μ=.2. When $\overline{x}$=1, $SEV(\mu > .2)$=.999, since $\overline{x}$=1 is 4 standard deviation in excess of μ=.2. So inferring the discrepancy $\mu > .2$ is increasingly warranted, for increasingly significant observed values. Hence, criticisms #2 and #3 are scotched by employing the severity evaluation.

If pressed, critics often concede that one can avoid the fallacies of rejection, but seem to argue that the tests are illegitimate because they are open to fallacious construals. This seems to us an absurd and illogical critique of the foundations of tests. We agree that tests should be accompanied by interpretive tools that avoid fallacies by highlighting the correct logic of tests. That is what the error statistical philosophy supplies. We do not envision computing these assessments each time, nor is this necessary. The idea would be to report severity values corresponding to the inferences of interest in the given problem; several benchmarks for well warranted and poorly warranted inferences would suffice.

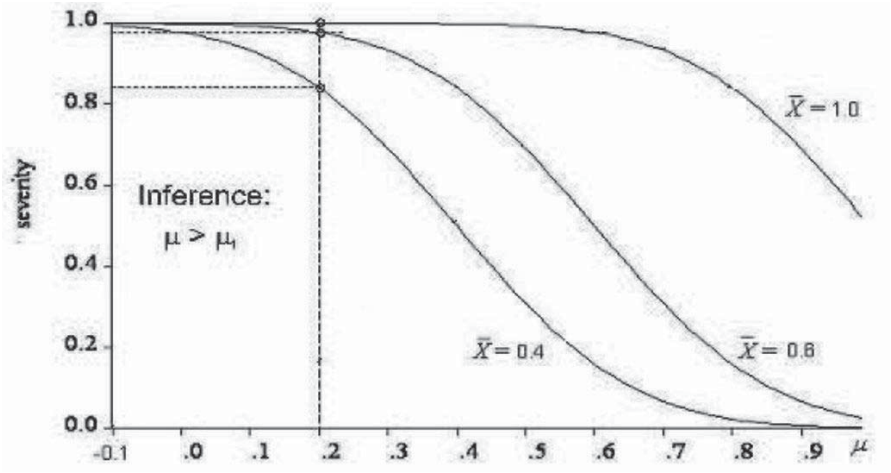


Figure 2. Significant result. Severity associated with inference $\mu > 0.2$, with different outcomes x_0.

Figure 2 shows three *severity curves* for test T_α, associated with different outcomes $\mathbf{x}_0$, where, as before, $\sigma = 2, n = 100, d(\mathbf{x}_0) = \frac{(\overline{x}-\mu_0)}{\sigma_x}, \mu_0 = 0$:

$$\begin{array}{ll} \text{for } d(\mathbf{x}_0)\text{=2.0 } (\overline{x}\text{=0.4}): & SEV(\mu > 0.2)\text{=.841,} \\ \text{for } d(\mathbf{x}_0)\text{=3.0 } (\overline{x}\text{=0.6}): & SEV(\mu > 0.2)\text{=.977,} \\ \text{for } d(\mathbf{x}_0)\text{=5.0 } (\overline{x}\text{=1.0}): & SEV(\mu > 0.2)\text{=.999.} \end{array}$$

The vertical line at μ=.2 pinpoints the inference in our illustration, but sliding it along the μ axis one sees how the same can be done for different inferences, e.g., μ>.3, μ>.4, ...

Criticism (#3) is often phrased as "statistical significance is not substantive significance". What counts as substantive significance is a matter of the context. What the severity construal of tests will do is tell us which discrepancies are and are not indicated, thereby avoiding the confusion between statistical and substantive significance.

To illustrate the notion of warranted discrepancies with data $\mathbf{x}_0$, consider figure 3 where we focus on just one particular statistically significant outcome, say $\overline{x}$=0.4, and consider different discrepancies γ from 0 one might wish to infer, each represented by a vertical line. To begin with, observe that $SEV(\mu > 0)$=.977, i.e. 1 minus the p-value corresponding to this test. On the other hand, as the discrepancy increases from 0 to .2 the $SEV(\mu > .2)$ is a bit lower, but not too bad: .841. We see that the SEV decreases as larger discrepancies from 0 are entertained (remembering the outcome is fixed at $\overline{x}$=0.4). An extremely useful benchmark is $\mu > .4$, since that is the inference which receives severity .5. So we know immediately that $SEV(\mu > .5)$ is less than .5, and in particular it is .3. So

$\overline{x}$=0.4 provides a very poor warrant for $\mu > .5$. More than half the time such a significant outcome would occur even if $\mu \leq .5$.

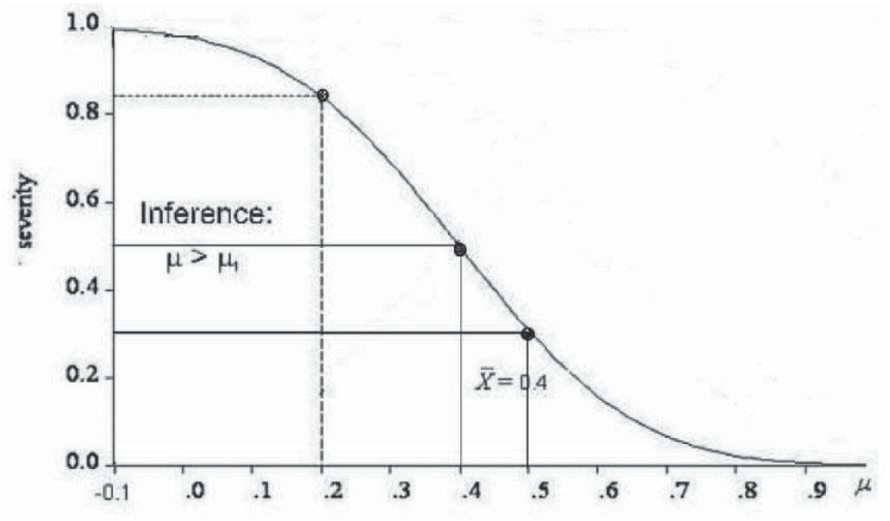


Figure 3. Significant result. The severity for inferring different discrepancies $\mu > \gamma$ with the same outcome $\overline{x}$=0.4

Many general relationships can be deduced. For example, since the assertions $\mu > \mu_1$ and $\mu \leq \mu_1$ constitute a partition of the parameter space of μ we have:

$$SEV(\mu > \mu_1) = 1 - SEV(\mu \leq \mu_1).$$

As before, severity is evaluated at a point μ_1, i.e.

$$SEV(\mu > \mu_1) = P(d(\mathbf{X}) \leq d(\mathbf{x}_0); \mu{=}\mu_1).$$

Severity and Power with Significant Results: two key ponts

(i) It is important to note the relationship between our data-specific assessment of an α-level statistically significant result and the usual assessment of the power of test T_α at the alternative: $\mu_1{=}(\mu_0 + \gamma)$. Power, remember, is always defined in terms of a rejection rule indicating the threshold (c_α) beyond which the result is taken as statistically significant enough to reject the null; see section 2.1.

If $d(\mathbf{x}_0)$ is then significant at the α-level, $d(\mathbf{x}_0) > c_\alpha$, the severity with which the test has passed $\mu > \mu_1$ is:

$$P(d(\mathbf{X}) \leq c_\alpha; \mu{=}\mu_1) = 1 - POW(T_\alpha; \mu_1).$$

But the observed statistically significant $d(\mathbf{x}_0)$ could exceed the mere cut-off value for significance c_α. Should we take a result that barely makes it to the cut-off

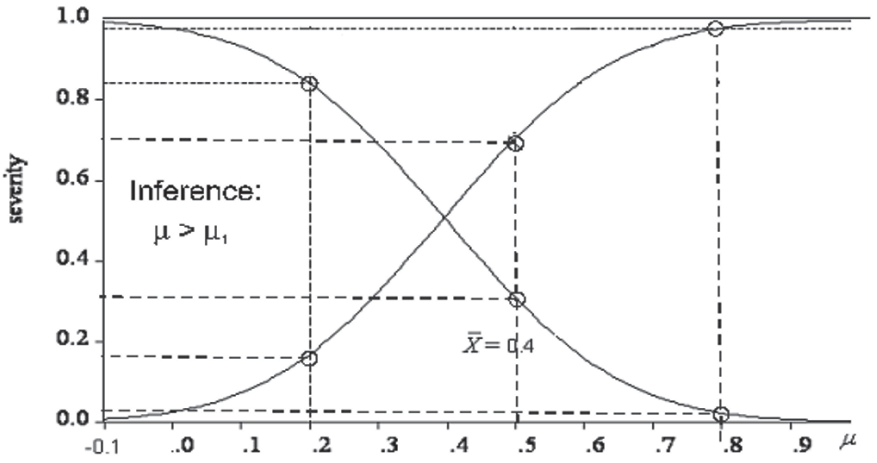


Figure 4. Juxtaposing the power curve with the severity curve for $\overline{x} = .4$.

just the same as one farther out into the rejection region? We think not, and the assessment of severity reflects this. As is plausible, the more significant result yields a higher the severity for the same inference $\mu > \mu_1$:

$$P(d(\mathbf{X}) \leq d(\mathbf{x}_0); \mu > \mu_1) \text{ exceeds } P(d(\mathbf{X}) \leq c_\alpha; \mu > \mu_1).$$

That is, one minus the power of the test at μ_1 provides a lower bound for the severity associated with the inference $\mu > \mu_1$.

The higher the power of the test to detect discrepancy γ, the lower the severity associated with the inference: $\mu > (\mu_0 + \gamma)$ when the test rejects H_0.

Hence, the severity with which alternative $\mu > (\mu_0 + \gamma)$ passes a test is not given by, and is in fact inversely related to, the test's power at: $\mu_1{=}(\mu_0 + \gamma)$.

This can be seen in figure 4 which juxtaposes the power curve with the severity curve for $\overline{x}{=}0.4$. It is seen that the power curve slopes in the opposite direction from the severity curve. As we just saw, the statistically significant result, $\overline{x}{=}0.4$, is good evidence for $\mu > .2$ (the severity was .841), but poor evidence for the discrepancy $\mu > .5$ (the severity was .3). If the result does not severely pass the hypothesis $\mu > .5$, it would be even less warranted to take it as evidence for a larger discrepancy, say $\mu > .8$. The relevant severity evaluation yields: $P(d(\mathbf{X}) \leq 2.0; \mu{=}.8) = .023$, which is very low, but the power of the test at $\mu{=}.8$ is very high, .977.

Putting numbers aside, an intuitive example makes the point clear. The smaller the mesh of a fishnet, the more capable it is of catching even small fish. So being given the report that (i) a fish is caught, and (ii) the net is highly capable of catching even 1 inch guppies, we would deny the report is good evidence of, say, a 9 inch fish! This takes us to our next concern.

2.4 Fallacies arising from overly sensitive tests

A common complaint concerning a statistically significant result is that for any discrepancy from the null, say $\gamma \geq 0$, however small, one can find a large enough sample size n such that a test, with high probability, will yield a statistically significant result (for any p-value one wishes).

(#4) With large enough sample size even a trivially small discrepancy from the null can be detected.

A test can be so sensitive that a statistically significant difference from H_0 only warrants inferring the presence of a relatively small discrepancy γ; a large enough sample size n will render the power $POW(T_\alpha; \mu_1{=}\mu_0 + \gamma)$ very high. To make things worse, many assume, fallaciously, that reaching statistical significance at a given level α is more evidence against the null the larger the sample size (n). (Early reports of this fallacy among psychology researchers are in Rosenthal and Gaito, 1963). Few fallacies more vividly show confusion about significance test reasoning. A correct understanding of testing logic would have nipped this fallacy in the bud 60 years ago. Utilizing the severity assessment one sees that an α-significant difference with n_1 passes $\mu > \mu_1$ less severely than with n_2 where $n_1 > n_2$.

For a fixed type I error probability α, increasing the sample size decreases the type II error probability (power increases). Some argue that to balance the two error probabilities, the required α level for rejection should be decreased as n increases. Such rules of thumb are too tied to the idea that tests are to be specified and then put on automatic pilot without a reflective interpretation. The error statistical philosophy recommends moving away from all such recipes. The reflective interpretation that is needed drops out from the severity requirement: increasing the sample size does increase the test's sensitivity and this shows up in the "effect size" γ that one is entitled to infer at an adequate severity level. To quickly see this, consider figure 5.

It portrays the severity curves for test T_α, $\sigma{=}2$, $n{=}100$, with the same outcome $d(\mathbf{x}_0){=}1.96$, but based on different sample sizes ($n{=}50, n{=}100, n{=}1000$), indicating that: the *severity for inferring* $\mu > .2$ decreases as n increases:

$$\begin{aligned} &\text{for } n{=}50: \quad SEV(\mu > 0.2){=}.895,\\ &\text{for } n{=}100: \quad SEV(\mu > 0.2){=}.831,\\ &\text{for } n{=}1000: \quad SEV(\mu > 0.2){=}.115. \end{aligned}$$

The facts underlying criticism #4 are also erroneously taken as grounding the claim:

"All nulls are false."

This confuses the true claim that with large enough sample size, a test has power to detect any discrepancy from the null however small, with the false claim that all nulls are false.

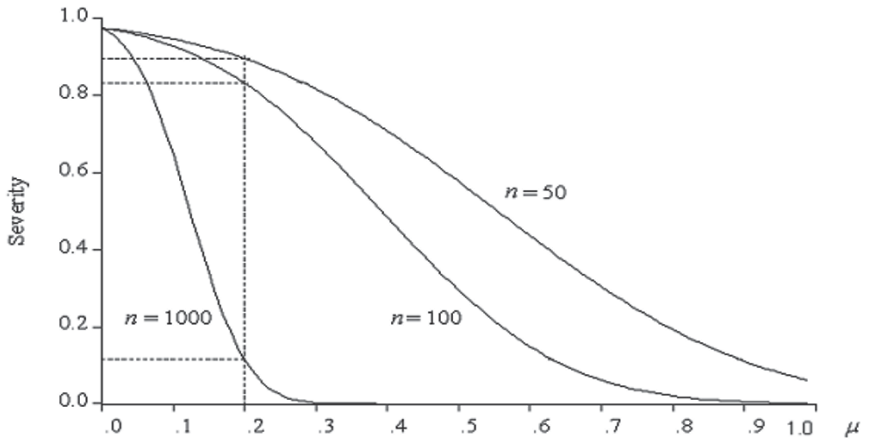


Figure 5. Severity associated with inference μ>0.2, $d(x_0)$=1.96, and different sample sizes n.

The tendency to view tests as automatic recipes for rejection gives rise to another well-known canard:

> **(#5) Whether there is a statistically significant difference from the null depends on which is the null and which is the alternative.**

The charge is made by considering the highly artificial case of two point hypotheses such as: μ=0 vs. μ=.8. If the null is μ=0 and the alternative is μ=.8 then $\overline{x}$=0.4 (being $2\sigma_x$ from 0) "rejects" the null and declares there is evidence for .8. On the other hand if the null is μ=.8 and the alternative is μ=0, then observing $\overline{x}$=0.4 now rejects .8 and finds evidence for 0. It appears that we get a different inference depending on how we label our hypothesis! Now the hypotheses in a N-P test must exhaust the space of parameter values, but even entertaining the two point hypotheses, the fallacy is easily exposed. Let us label the two cases:

Case 1: H_0: μ=0 vs. H_1: μ=.8, **Case 2**: H_0: μ=.8 vs. H_1: μ=0.

In case 1, $\overline{x}$=0.4 is indeed evidence of *some* discrepancy from 0 in the positive direction, but it is exceedingly poor evidence for a discrepancy as large as .8 (see figure 2). Even without the calculation that shows $SEV(\mu > .8)$=.023, we know that $SEV(\mu > .4)$ is only .5, and so there are far less grounds for inferring an even larger discrepancy[5].

[5]We obtain the standardized value by considering the sample mean ($\overline{x}$=.4) minus the hypothesize μ (.8), in standard deviation units (σ_x=.2), yielding $z = -2$, and thus $P(Z < -2)$=.023.

In case 2, the test is looking for discrepancies from the null (which is .8) in the negative direction. The outcome $\overline{x}$=0.4 ($d(\mathbf{x}_0)$=-2.0) is evidence that $\mu \leq .8$ (since $SEV(\mu \leq .8)$=.977), but there are terrible grounds for inferring the alternative μ=0!

In short, case 1 asks if the true μ exceeds 0, and $\overline{x}$=.4 is good evidence of some such positive discrepancy (though poor evidence it is as large as .8); while case 2 asks if the true μ is less than .8, and again $\overline{x}$=.4 is good evidence that it is. Both these claims are true. In neither case does the outcome provide evidence for the point alternative, .8 and 0 respectively. So it does not matter which is the null and which is the alternative, and criticism #5 is completely scotched.

Note further that in a proper test, the null and alternative hypotheses must exhaust the parameter space, and thus, "point-against-point" hypotheses are at best highly artificial, at worst, illegitimate. What matters for the current issue is that the error statistical tester never falls into the alleged inconsistency of inferences depending on which is the null and which is the alternative.

We now turn our attention to cases of statistically insignificant results. Overly high power is problematic in dealing with significant results, but with insignificant results, the concern is the test is not powerful enough.

2.5 Fallacies of acceptance: errors in interpreting statistically insignificant results

(#6) Statistically insignificant results are taken as evidence that the null hypothesis is true.

We may call this the fallacy of interpreting insignificant results (or the fallacy of "acceptance"). The issue relates to a classic problem facing general hypothetical deductive accounts of confirmation: positive instances "confirm" or in some sense count for generalizations. Unlike logics of confirmation or hypothetico-deductive accounts, the significance test reasoning, and error statistical tests more generally, have a very clear basis for denying this. An observed accordance between data and a null hypothesis "passes" the null hypothesis, i.e., condition (S-1) is satisfied. But such a passing result is *not* automatically evidence *for* the null hypothesis, since the test might not have had much chance of detecting departures even if they existed. So what is called for to avoid the problem is precisely the second requirement for severity (S-2). This demands considering error probabilities, the distinguishing feature of an error statistical account.

Now the simple Fisherian significance test, where the result is either to falsify the null or not, leaves failure to reject in some kind of limbo. That is why Neyman and Pearson introduce the alternative hypothesis and the corresponding notion of power. Consider our familiar test T_α. Affirming the null is to rule out a discrepancy $\gamma > 0$. It is unwarranted to claim to have evidence for the null if the test had little capacity (probability) of producing a worse fit with the null even though the null is false, i.e. $\mu > 0$. In the same paper addressing Carnap, Neyman makes this

point (p. 41)[6], although it must be conceded that it is absent from his expositions of tests. The severity account makes it an explicit part of interpreting tests (note that $d(x_0) = \frac{(\overline{x}-\mu_0)}{\sigma_x}$):

> (a) If there is a very low probability that $d(\mathbf{x}_0)$ would have been larger than it is, even if μ exceeds μ_1, then $\mu \leq \mu_1$ passes the test with low severity, i.e. $SEV(\mu \leq \mu_1)$ is low.

By contrast:

> (b) If there is a very high probability that $d(\mathbf{x}_0)$ would have been larger than it is, were μ to exceed μ_1, then $\mu \leq \mu_1$ passes the test with high severity, i.e. $SEV(\mu \leq \mu_1)$ is high.

To see how formal significance tests can encapsulate this, consider testing H_0: μ=0 vs. H_1: $\mu > 0$, and obtaining a *statistically insignificant* result: $d(\mathbf{x}_0) \leq 1.96$. We have (S-1): $\mathbf{x}_0$ agrees with H_0 since $d(\mathbf{x}_0) \leq 1.96$.

(S-1): $\mathbf{x}_0$ agrees with H_0 since $d(\mathbf{x}_0) \leq 1.96$. We wish to determine if it is good evidence for $\mu \leq \mu_1$, where μ_1=$\mu_0 + \gamma$, by evaluating the probability that test T_α would have produced a more significant result (i.e. $d(\mathbf{X}) > d(\mathbf{x}_0)$), if $\mu > \mu_1$:

$$SEV(T_\alpha, \mathbf{x}_0, \mu \leq \mu_1) = P(d(\mathbf{X}) > d(\mathbf{x}_0); \mu > \mu_1).$$

It suffices to evaluate this at μ_1=μ_0+γ because the probability increases for $\mu > \mu_1$. So, if we have good evidence that $\mu \leq \mu_1$ we have even better evidence that $\mu \leq \mu_2$ where μ_2 exceeds μ_1 (since the former entails the latter).

Rather than work through calculations, it is revealing to report several appraisals graphically. Figure 6 shows severity curves for test T_α, where σ=2, n=100, based on three different insignificant results:

$$d(\mathbf{x}_0)=1.95(\overline{x}=.392), d(\mathbf{x}_0)=1.5(\overline{x}=.3), d(\mathbf{x}_0)=.50(\overline{x}=.1).$$

As before, let a statistically significant result require $d(\mathbf{x}_0)$>1.96. None of the three insignificant outcomes provide strong evidence that the null is precisely true, but what we want to do is find the smallest discrepancy that each rules out with severity.

For illustration, we consider a particular fixed inference of the form ($\mu \leq \mu_1$), and compare severity assessments for different outcomes. The low probabilities

[6]In the context where H_0 had not been "rejected", Neyman insists, it would be "dangerous" to regard this as confirmation of H_0 if the test in fact had little chance of detecting an important discrepancy from H_0, even if such a discrepancy were present. On the other hand if the test had appreciable power to detect the discrepancy, the situation would be "radically different." Severity logic for insignificant results has the same pattern except that we consider the actual insignificant result, rather than the case where data just misses the cut-off for rejection.

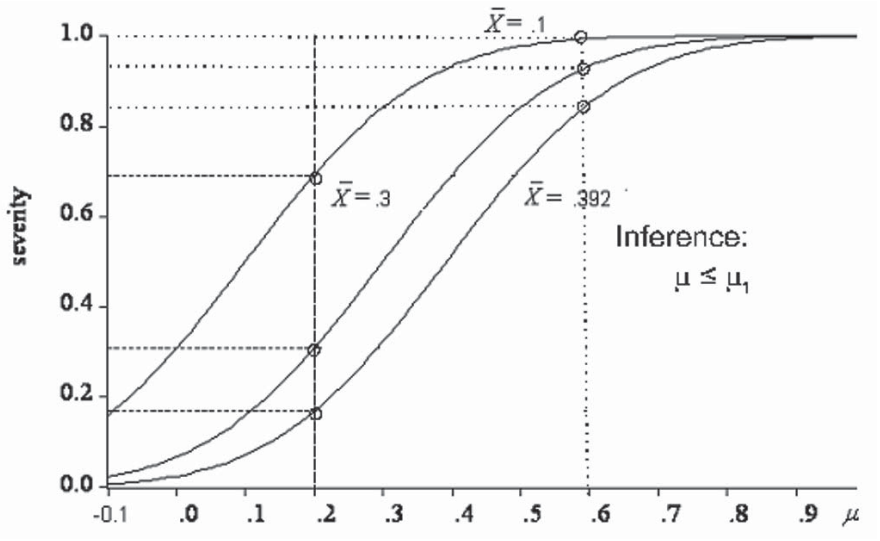


Figure 6. Insignificant result. Severity associated with inference $\mu \leq .2$ with different outcomes x_0.

associated with the severity assessment of $\mu \leq .2$ indicates that, in all three cases, the claim that the discrepancy is $\mu \leq .2$ is unwarranted (to the degrees indicated):

$$\begin{array}{ll} \text{for } d(\mathbf{x}_0)\text{=1.95 } (\overline{x}\text{=.39}): & SEV(\mu \leq 0.2) = .171, \\ \text{for } d(\mathbf{x}_0)\text{=1.5 } (\overline{x}\text{=0.3}): & SEV(\mu \leq 0.2) = .309, \\ \text{for } d(\mathbf{x}_0)\text{=0.5 } (\overline{x}\text{=0.1}): & SEV(\mu \leq 0.2) = .691. \end{array}$$

So it would be fallacious (to different degrees) to regard these as warranting $\mu \leq 0.2$. To have a contrast, observe that inferring $\mu \leq .6$ is fairly warranted (to different degrees) for all three outcomes:

$$\begin{array}{ll} \text{for } d(\mathbf{x}_0)\text{=1.95 } (\overline{x}\text{=.39}): & SEV(\mu \leq 0.6) = .853, \\ \text{for } d(\mathbf{x}_0)\text{=1.5 } (\overline{x}\text{=0.3}): & SEV(\mu \leq 0.6) = .933, \\ \text{for } d(\mathbf{x}_0)\text{=0.5 } (\overline{x}\text{=0.1}): & SEV(\mu \leq 0.6) = .995. \end{array}$$

Working in the reverse direction, it is instructive to fix a high severity value, say, .95, and ascertain, for different outcomes, the discrepancy that may be ruled out with severity .95. For $\overline{x}$=0.39, $SEV(\mu \leq .72)$=.95, for $\overline{x}$=0.3, $SEV(\mu \leq .62)$=.95, and $\overline{x}$=0.1, $SEV(\mu \leq .43)$=.95. Although none of these outcomes warrants ruling out all positive discrepancies at severity level .95, we see that the smaller the observed outcome $\overline{x}$, the smaller is the μ_1 value such that $SEV(\mu \leq \mu_1) = .95$.

It is interesting to note that the severity curve associated with $d(\mathbf{x}_0)$ =1.95 virtually coincides with the power curve since c_α=1.96 for α=.025. The power of

the test to detect μ_1 gives the lower bound for the severity assessment for $(\mu \leq \mu_1)$; this is the lowest it could be when an insignificant result occurs. High power at μ_1 ensures that insignificance permits inferring with high severity that $\mu \leq \mu_1$. Thus severity gives an inferential justification for the predesignated power, but it goes further. Once the result is available, it directs us to give a more informative inference based on $d(\mathbf{x}_0)$.

P-values are Not *Posterior Probabilities of* H_0

The most well-known fallacy in interpreting significance tests is to equate the p-value with a posterior probability on the null hypothesis. In legal settings this is often called the *prosecutor's fallacy*. Clearly, however:

(i) $P(d(X) \geq d(x_0);\ H_0)$ is not equal to (ii) $P(H_0 \mid d(X) \geq d(x_0))$.

The p-value assessment in (i) refers only to the sampling distribution of the test statistic $d(\mathbf{X})$; and there is no use of *prior probabilities*, as would be necessitated in (ii). In the frequentist context, $\{d(\mathbf{X}) > 1.96\}$ is an event and *not* a statistical hypothesis. The latter must assign probabilities to outcomes of an experiment of interest.

Could we not regard such events as types of hypotheses or at least predictions? Sure. But scientific hypotheses of the sort statistical methods have been developed to test are not like that. Moreover, no prior probabilities are involved in (i): it is just the usual computation of probabilities of events "calculated under the assumption" of a given statistical hypothesis and model. (It is not even correct to regard this as a conditional probability.) We are prepared to go further: it seems to us an odd way of talking to regard the null hypothesis as evidence for the event $\{d(\mathbf{X}) \leq 1.96\}$, or for its high probability. It is simply to state what is deductively entailed by the probability model and hypothesis. Most importantly, the statistical hypotheses we wish to make inferences about are not events; trying to construe them as such involves fallacies and inconsistencies (we return to this in Section 3).

Some critics go so far as to argue that despite it being fallacious (to construe error probabilities as posterior probabilities of hypotheses):

(#7) Error probabilities are invariably misinterpreted as posterior probabilities.

Our discussion challenges this allegation that significance tests (and confidence intervals) are invariably used in "bad-faith". We have put forward a rival theory as to the meaning and rationale for the use of these methods in science: properly interpreted, they serve to control and provide assessments of the severity with which hypotheses pass tests based on evidence. The quantitative aspects arise in the form of degrees of severity and sizes of discrepancies detected or not. This rival theory seems to offer a better explanation of inductive reasoning in science.

2.6 Relevance for finite samples

Some critics charge that because of their reliance on frequentist probability:

> **(#8) Error statistical tests are justified only in cases where there is a very long (if not infinite) series of repetitions of the same experiment.**

Ironically, while virtually all statistical accounts appeal to good asymptotic properties in their justifications, the major asset of error statistical methods is in being able to give assurances that we will not be too far off with specifiable finite samples. This is a crucial basis both for planning tests and in critically evaluating inferences post-data. Pre-data the power has an important role to play in ensuring that test T_α has enough 'capacity', say $(1-\alpha)$, to detect a discrepancy of interest γ for $\mu_1=\mu_0+\gamma$. To ensure the needed power one often has no other option but to have a large enough sample size n. How large is 'large enough' is given by solving the probabilistic equation:

$$\text{to ensure: } P(d(\mathbf{X}) > c_\alpha; \mu_1)=(1-\beta), \text{ set: } n=\{[(c_\alpha - c_\beta)\sigma]/\gamma\}^2,$$

where c_β is the threshold such that $P(Z \leq c_\beta)=\beta$ for $Z \backsim \mathsf{N}(0,1)$.

Numerical example. Consider test T_α with $\mu_0=0$, $\alpha=.025$, $\sigma=2$, and let the substantive discrepancy of interest be $\gamma=.4$. Applying the above formula one can determine that the sample size needed to ensure high enough power, say $(1-\beta)=.90$, to detect such a discrepancy is: $n=\{[(1.96+1.28)(2)]/(.4)\}^2 \approx 262$, i.e., the test needs 262 observations to have .9 power to detect discrepancies $\gamma \geq .4$.

If the sample size needed for informative testing is not feasible, then there are grounds for questioning the value of the inquiry, but not for questioning the foundational principles of tests.

This points to a central advantage of the error statistical approach in avoiding the limitations of those accounts whose reliability guarantees stem merely from asymptotic results, that is for n going to infinity. In particular, a test is consistent against some alternative μ_1 when its power $P(d(\mathbf{X}) > c_\alpha; \mu_1)$ goes to one as n goes to infinity. This result, however, is of no help in assessing, much less ensuring, the reliability of the test in question for a given n.

Considering discrepancies of interest restricts the latitude for test specification, not only in choosing sample sizes, but in selecting test statistics that permit error probabilities to be 'controlled' despite unknowns. We now turn to this.

2.7 Dealing with "Nuisance" Parameters

In practice, a more realistic situation arises when, in the above simple Normal model $(\mathcal{M})$, both parameters, μ and σ^2, are *unknown*. Since the primary inference concerns μ and yet σ^2 is needed to complete the distribution, it is often called a "nuisance" parameter. Given the way nuisance parameters are handled in this

approach, changes to the testing procedure are rather minor, and the reasoning is unchanged. That is why we were able to keep to the simpler case for exposition. To illustrate,

EXAMPLE 2. Test T^*_α, when σ^2 is *unknown.*

First, the test statistic now takes the form: $d^*(\mathbf{X})=\frac{\sqrt{n}(\overline{X}-\mu_0)}{s}$,
where $s^2=\frac{1}{n-2}\sum_{k=1}^{n}(X_k-\overline{X})^2$ is the sample variance, that provides an unbiased estimator of σ^2.

Second, the sampling distributions of $d^*(\mathbf{X})$, under both the null (H_0) and the alternative (H_1) hypotheses, are no longer Normal, but *Student's t*; see [Lehmann, 1986]. What matters is that the distribution of $d^*(\mathbf{X})$ under H_0 does not involve the nuisance parameter σ^2; the only difference is that one needs to use the Student's t instead of the Normal tables to evaluate c_α corresponding to a particular α. The distribution of $d^*(\mathbf{X})$ under H_1 *does* involve σ^2 but only through the noncentrality parameter affecting the power: $\delta=\frac{\sqrt{n}(\mu_1-\mu_0)}{\sigma}$. In practice one replaces σ with its estimate s when evaluating the power of the test at μ_1 — likewise for severity. All the other elements of test T_α remain the same for T^*_α, including the form of the test rule.

Ensuring error statistical calculations free of a nuisance parameter is essential for attaining objectivity: the resulting inferences are not threatened by unknowns. This important desideratum is typically overlooked in foundational discussions and yet the error statistical way of satisfying it goes a long way toward answering the common charge that:

(#9) Specifying statistical tests is too arbitrary.

In a wide class of problems, the error statistician attains freedom from a nuisance parameter by conditioning on a sufficient statistic for it; see [Cox and Hinkley, 1974], leading to a uniquely appropriate test. This ingenious way of dealing with nuisance parameters stands in contrast with Bayesian accounts that require prior probability distributions for each unknown quantity. (Nuisance parameters also pose serious problems for pure likelihood accounts; see [Cox, 2006]). Once this is coupled with the requirement that the test statistics provide plausible measures of "agreement", the uniquely appropriate test is typically overdetermined: one can take one's pick for the rationale (appealing to Fisherian, Neyman-Pearsonian, or severity principles).

2.8 Severe Testing and Confidence Interval (CI) Estimation

In CI estimation procedures, a statistic is used to set upper or lower (1-sided) or both (2-sided) bounds. For a parameter, say μ, a $(1-\alpha)$ CI estimation procedure leads to estimates of form: $\mu = \overline{x} \pm e$. Critics of significance tests often allege:

(#10) We should be doing confidence interval estimation rather than significance tests.

Although critics of significance tests often favor CI's, it is important to realize that CI's are still squarely within the error statistical paradigm. In fact there is a precise duality relationship between $(1-\alpha)$ CI's and significance tests: the CI contains the parameter values that would not be rejected by the given test at the specified level of significance [Neyman, 1935]. It follows that the $(1-\alpha)$ one sided interval corresponding to test T_α is:

$$\mu > \quad \overline{X} - c_\alpha(\sigma/\sqrt{n}).$$

In particular, the 97.5% CI estimator corresponding to test T_α is:

$$\mu > \overline{X} - 1.96(\sigma/\sqrt{n}).$$

Now it is true that the confidence interval gives a data-dependent result, but so does our post-data interpretation of tests based on severity. Moreover, confidence intervals have their own misinterpretations and shortcomings that we need to get around.

A well known fallacy is to construe $(1-\alpha)$ as the degree of probability to be assigned the particular interval estimate formed, once $\overline{X}$ is instantiated with $\overline{x}$. Once the estimate is formed, either the parameter is or is not contained in it. One can say only that the particular estimate arose from a procedure which, with high probability, $(1-\alpha)$, would contain the true value of the parameter, whatever it is. This affords an analogous "behavioristic" rationale for confidence intervals as we saw with tests: Different sample realizations $\mathbf{x}$ lead to different estimates, but one can ensure that $(1-\alpha)100\%$ of the time the true parameter value μ, whatever it may be, will be included in the interval formed. Just as we replace the behavioristic rationale of tests with the inferential one based on severity, we do the same with confidence intervals.

The assertion $\mu > \overline{x} - c_\alpha(\sigma/\sqrt{n})$ is the one-sided $(1-\alpha)$ interval corresponding to the test T_α and indeed, for the particular value $\mu_1 = \overline{x} - c_\alpha(\sigma/\sqrt{n})$, the severity with which the inference $\mu > \mu_1$ passes test T_α is $(1-\alpha)$. The severity rationale for applying the rule and inferring $\mu > \overline{x} - c_\alpha(\sigma/\sqrt{n})$ might go as follows:

Suppose this assertion is false, e.g., suppose $\mu_1 = \overline{x} - 1.96(\sigma/\sqrt{n})$. Then the observed mean is 1.96 standard deviations in excess of μ_1. Were μ_1 the mean of the mechanism generating the observed mean, then with high probability (.975) a result less discordant from μ_1 would have occurred. (For even smaller values of μ_1 this probability is increased.)

However, our severity construal also demands breaking out of the limitations of confidence interval estimation. In particular, in the theory of confidence intervals, a single confidence level is prespecified and the one interval estimate corresponding to this level is formed as the inferential report. The resulting interval is sometimes used to perform statistical tests: Hypothesized values of the parameter are accepted (or rejected) according to whether they are contained within (or outside) the resulting interval estimate. The same problems with automatic uses of tests with a single prespecified choice of significance level α reappear in

the corresponding confidence interval treatment. Notably, predesignating a single choice of confidence level, $(1-\alpha)$, is not enough.

Here is why: A $(1-\alpha)$ CI corresponds to the set of null hypotheses that an observed outcome would not be able to reject with the corresponding α-level test. In our illustrative example of tests, the null value is fixed (we chose 0), and then the sample mean is observed. But we could start with the observed sample mean, and consider the values of μ that would not be rejected, were they (rather than 0) the null value. This would yield the corresponding CI. That is, the observed mean is not sufficiently greater than any of the values in the CI to reject them at the α-level. But as we saw in discussing severity for insignificant results, this does not imply that there is good evidence for each of the values in the interval: many values in the interval pass test T_α with very low severity with $\mathbf{x}_0$. Yet a report of the CI estimate is tantamount to treating each of the values of the parameter in the CI on a par, as it were. That some values are well, and others poorly, warranted is not expressed. By contrast, for each value of μ in the CI, there would be a different answer to the question: How severely does $\mu > \mu_1$ pass with $\mathbf{x}_0$? The severity analysis, therefore, naturally leads to a sequence of inferences, or series of CIs, that are and are not warranted at different severity levels[7].

3 ERROR STATISTICS VS. THE LIKELIHOOD PRINCIPLE

A cluster of debates surrounding error statistical methods, both in philosophy and in statistical practice, reflects contrasting answers to the question: what information is relevant for evidence and inference? Answers, in turn, depend on assumptions about the nature of inductive inference and the roles of probabilistic concepts in inductive inference. We now turn to this.

Consider a conception of evidence based just on likelihoods: data $\mathbf{x}_0$ is evidence for H so long as:

(a) $P(\mathbf{x}_0; H) =$ high (or maximal)

(b) $P(\mathbf{x}_0; H$ is false$) =$ low.

Although at first blush these may look like the two conditions for severity ((S-1) and (S-2)), conditions (a) and (b) together are tantamount to a much weaker requirement: $\mathbf{x}_0$ is evidence for H so long as H is more likely on data $\mathbf{x}_0$ than on the denial of H — referring to the mathematical notion of likelihood. To see that this is scarcely sufficient for severity, consider a familiar example.

[7] A discussion of various attempts to consider a series of CI's at different levels, confidence curves [Birnbaum, 1961], p-value functions [Poole, 1987], consonance intervals [Kempthorne and Folks, 1971] and their relation to the severity evaluation is beyond the scope of this paper; see [Mayo and Cox, 2006].

Maximally Likely alternatives. H_0 might be that a coin is fair, and $\mathbf{x}_0$ the result of n flips of the coin. For each of the 2^n possible outcomes there is a hypothesis H_i^* that makes the data $\mathbf{x}_i$ maximally likely. For an extreme case, H_i^* can assert that the probability of heads is 1 just on those tosses that yield heads, 0 otherwise. For any $\mathbf{x}_i$, $P(\mathbf{x}_i;\ H_0)$ is very low and $P(\mathbf{x}_i; H_i^*)$ is high — one need only choose for (a) the statistical hypothesis that renders the data maximally likely, i.e., H_i^*. So the fair coin hypothesis is always rejected in favor of H_i^*, even when the coin is fair. This violates the severity requirement since it is guaranteed to infer evidence of discrepancy from the null hypothesis even if it is true. The severity of 'passing' H_i^* is minimal or 0. (Analogous examples are the "point hypotheses" in [Cox and Hinkley, 1974, p. 51], Hacking's [1965] "tram car" example, examples in [Mayo, 1996; 2008].)

This takes us to the key difference between the error statistical perspective and contrasting statistical philosophies; namely that to evaluate and control error probabilities requires going beyond relative likelihoods.

3.1 There is Often Confusion About Likelihoods

The distribution of the sample $\mathbf{X}$ assigns probability (or density) to each possible realization $\mathbf{x}$, under some fixed value of the parameter θ, i.e. $f(\mathbf{x};\theta)$. In contrast, the likelihood assigns probability (or density) to a particular realization $\mathbf{x}$, under different values of the unknown parameter θ. Since the data $\mathbf{x}$ are fixed at $\mathbf{x}_0$ and the parameter varies, the likelihood is defined as proportional to $f(\mathbf{x};\theta)$ but viewed as a function of the parameter θ:

$$L(\theta;\mathbf{x}_0) \propto f(\mathbf{x}_0;\theta) \text{ for all } \theta \in \Theta.$$

Likelihoods do not obey the probability axioms, for example, the sum of the likelihoods of a hypothesis and its denial is not one.

Hacking [1965] is known for having championed an account of comparative support based on what he called the "law of likelihood": data $\mathbf{x}_0$ support hypothesis H_1 less than H_2 if the latter is more likely than the former, i.e., $P(\mathbf{x}_0; H_2) > P(\mathbf{x}_0; H_1)$; when H_2 is composite, one takes the maximum of the likelihood over the different values of θ admitted by H_2. From a *theory of support* Hacking gets his *theory of testing* whereby, "an hypothesis should be rejected if and only if there is some rival hypothesis much better supported than it is..." [Hacking, 1965, p. 89].

Hacking [1980] distanced himself from this account because examples such as the one above illustrate that "there always is such a rival hypothesis, *viz.* that things just had to turn out the way they actually did" [Barnard, 1972, p. 129]. Few philosophers or statisticians still advocate a pure likelihood account of evidence (exceptions might be [Rosenkrantz, 1977; Sober, 2008], among philosophers, and [Royall, 1997] among statisticians). However, many who would deny that relative likelihoods are all that is needed for inference still regard likelihoods as all that is needed to capture the import of the data. For example, a Bayesian may hold

that inference requires likelihoods plus prior probabilities while still maintaining that the evidential import of the data is exhausted by the likelihoods. This is the gist of a general principle of evidence known as the *Likelihood Principle* (LP). Disagreement about the LP is a pivot point around which much of the philosophical debate between error statisticians and Bayesians has long turned. Holding the LP runs counter to distinguishing data on grounds of error probabilities of procedures.[8]

> "According to Bayes's theorem, $P(x|\mu)$...constitutes the entire evidence of the experiment, that is, it tells all that the experiment has to tell. More fully and more precisely, *if y is the datum of some other experiment, and if it happens that $P(x|\mu)$ and $P(y|\mu)$are proportional functions of μ (that is, constant multiples of each other), then each of the two data x and y have exactly the same thing to say about the values of μ*. . . ". [Savage, 1962, p. 17]

The italicized portion defines the LP. If a methodology allows data to enter only through the likelihoods, then clearly likelihoods contain all the import of the data — for that methodology. The philosophical question is whether relevant information is not thereby being overlooked. The holder of the LP considers the likelihood of the actual outcome, i.e., just $d(\mathbf{x}_0)$, whereas the error statistician needs to consider, in addition, the sampling distribution of $d(\mathbf{X})$ or other statistic being used in inference. In other words, an error statistician could use likelihoods in arriving at (S-1) the condition of accordance or fit with the data, but (S-2) additionally requires considering the probability of outcomes $\mathbf{x}$ that accord less well with a hypotheses of interest H, were H false. In the error statistical account, drawing valid inferences from the data $\mathbf{x}_0$ that happened to be observed is crucially dependent on the relative frequency of outcomes other than the one observed, as given by the appropriate sampling distribution of the test statistic.

3.2 Paradox of Optional Stopping

The conflict we are after is often illustrated by a two-sided version of our test T. We have a random sample from a Normal distribution with mean μ and standard deviation 1, i.e.,

$$X_k \backsim \mathsf{N}(\mu, 1),\ k = 1, 2, ..., n,$$

and wish to test the hypotheses:

$$H_0\colon\ \mu = 0, \text{vs. } H_1\colon\ \mu \neq 0.$$

To ensure an overall significance level of .05, one rejects the null whenever $|\overline{x}| > (1.96/\sqrt{n})$. However, instead of fixing the sample size in advance, we are to let n

[8] A weaker variation on the LP holds that likelihoods contain all the information within a given experiment, whereas the "strong" LP refers to distinct experiments. Here LP will always allude to the strong likelihood principle.

be determined by a *stopping rule*:

$$\text{keep sampling until } |\overline{x}| > (1.96/\sqrt{n}).$$

The probability that this rule will stop in a finite number of trials is 1, regardless of the true value of μ; it is a *proper* stopping rule. Whereas with n fixed in advance, such a test has a type 1 error probability of .05, with this stopping rule the test would lead to an *actual* significance level that would differ from, and be greater than .05. This is captured by saying that significance levels are sensitive to the stopping rule; and there is considerable literature as to how to adjust the error probabilities in the case of 'optional stopping', also called *sequential tests* [e.g., Armitage, 1975]. By contrast, since likelihoods are unaffected by this stopping rule, the proponent of the LP denies there really is an evidential difference between the cases where n is fixed and where n is determined by the stopping rule[9]. To someone who holds a statistical methodology that satisfies the LP, it appears that:

(#11) Error statistical methods take into account the intentions of the scientists analyzing the data.

In particular, the inference depends on whether or not the scientist intended to stop at n or intended to keep going until a statistically significant differnce from the null was found. The charge in #11 would seem to beg the question against the error statistical methodology which has perfectly objective ways to pick up on the effect of stopping rules: far from intentions "locked up in the scientist's head" (as critics allege), the manner of generating the data alter error probabilities, and hence severity assessments. As is famously remarked in [Edwards *et al.*, 1963]: "The likelihood principle emphasized in Bayesian statistics implies, ... that the rules governing when data collection stops are irrelevant to data interpretation. This irrelevance of stopping rules to statistical inference restores a simplicity and freedom to experimental design that had been lost by classical emphasis on significance levels — in the sense of [Neyman and Pearson, 1933, p. 239]. While it may restore "simplicity and freedom" it does so at the cost of being unable to adequately control error probabilities [Berger and Wolpert, 1988; Cox and Hinkley, 1974; Kadane *et al.*, 1999; Mayo and Kruse, 2001; Cox and Mayo, 2010].

3.3 The Reference Bayesians and the Renunciation of the LP

All error probabilistic notions are based on the sampling distribution of a statistic, and thus for an error statistician reasoning from data $\mathbf{x}_0$ always depends on considering how a procedure would handle outcomes other than $\mathbf{x}_0$; it is necessary to consider how often the result would occur in hypothetical repetitions. This conflicts with the likelihood principle (LP). Therefore, objectivity for the error

[9] Birnbaum [1962] argued that the LP follows from apparently plausible principles of conditionality and sufficiency. A considerable literature exists, see [Barnett, 1999]. Mayo [2010b] has recently argued that this "proof" is fallacious.

statistician entails violating the LP — long held as the core of Bayesianism [Mayo, 1983; 1985; 1996].

In fact Bayesians have long argued for foundational superiority over frequentist error statisticians on grounds that they uphold, while frequentists violate, the likelihood principle (LP), leading the latter into Bayesian *incoherency*. Frequentists have long responded that having a good chance of getting close to the truth and avoiding error is what matters [Cox and Hinkley, 1974]. However, many Bayesian statisticians these days, seem to favor the use of conventionally chosen or "reference" Bayesian priors, both because of the difficulty of eliciting subjective priors, and the reluctance of many scientists to allow subjective beliefs to overshadow the information provided by data. These reference priors however, violate the LP.

Over the past few years, leading developers of reference Bayesian methods [Bernardo, 2005; Berger, 2004] concede that desirable reference priors force them to consider the statistical model leading to violations of basic principles, such as the likelihood principle and the stopping rule principle; see [Berger and Wolpert, 1988]. Remarkably, they are now ready to admit that "violation of principles such as the likelihood principle is the price that has to be paid for objectivity" [Berger, 2004]. Now that the reference-Bayesian concedes that violating the LP is necessary for objectivity there may seem to be an odd sort of agreement between the reference Bayesian and the error statistician.

Do the concessions of reference Bayesians bring them closer to the error statistical philosophy? To even consider this possibility one would need to deal with a crucial point of conflict as to the basic role of probability in induction. Although Bayesians disagree among themselves about both the interpretation of posterior probabilities, and their numerical values, they concur that:

"what you 'really' want are posterior probabilities for different hypotheses."

It is well known that error probabilities differ from posteriors. In a variation on the charge of misinterpretation in (#6), critics seek examples where:

"p-values conflict with Bayesian posteriors,"

leading to results apparently counterintuitive even from the frequentist perspective. We consider the classic example from statistics.

(Two-sided) Test of a Mean of a Normal Distribution

The conflict between p-values and Bayesian posteriors often considers the familiar example of the two sided $T_{2\alpha}$ test for the hypotheses:

$$H_0{:}\mu = 0, \text{ vs. } H_1{:}\mu \neq 0.$$

The difference between p-values and posteriors are far less marked with one-sided tests , e.g., [Pratt, 1977; Cassella and Berger, 1987]. Critics observe:

> "If $n = 50$ one can classically 'reject H_0 at significance level $p = .05$,' although $P(H_0|x) = .52$ (which would actually indicate that the evidence favors H_0)." ([Berger and Sellke, 1987, p. 113], we replace Pr with P for consistency.)

Starting with a high enough prior probability to the point null (or, more correctly, to a small region around it), they show that an α significant difference can correspond to a posterior probability in H_0 that is not small. Where Bayesians take this as problematic for significance testers, the significance testers balk at the fact that use of the recommended priors can result in highly significant results being construed as no evidence against the null — or even "that the evidence favors H_0." If n=1000, a result statistically significant at the .05 level leads to a posterior to the null of .82! [Berger and Sellke, 1987]. Here, statistically significant results — results that we would regard as passing the non-null hypothesis severely — correspond to an *increase* in probability from the prior (.5) to the posterior.

What justifies this prior? The Bayesian prior probability assignment of .5 to H_0, the remaining .5 probability being spread out over the alternative parameter space, (e.g., recommended by Jeffreys [1939]) is claimed to offer an "objective" assessment of priors: the priors are to be read off from a catalogue of favored "reference" priors, no subjective beliefs are to enter. It is not clear how this negative notion of objectivity secures the assurance we would want of being somewhere close to the truth. The Bayesians do not want too small a prior for the null since then evidence against the null is merely to announce that an improbable hypothesis has become more improbable. Yet the spiked concentration of belief ("weight") in the null is at odds with the prevailing use of null hypotheses as simply a standard from which one seeks discrepancies. Finally, these examples where p-values differ from posteriors create a tension between the posterior probability in a testing context and the corresponding (highest probability) Bayesian confidence interval: the low posterior indicates scarce evidence against the null even though the null value is outside the corresponding Bayesian confidence interval [Mayo, 2005].

Some examples strive to keep within the frequentist camp: to construe a hypothesis as a random variable, it is imagined that we sample randomly from a population of hypotheses, some proportion of which are assumed to be true. The percentage "initially true" serves as the prior probability for H_0. This gambit commits what for a frequentist would be a fallacious instantiation of probabilities:

> 50% of the null hypotheses in a given pool of nulls are true.
>
> This particular null hypothesis H_0 was randomly selected from this pool.
>
> Therefore $P(H_0 \text{ is true}) = .5$.

Even allowing that the probability of a randomly selected hypothesis taken from an "urn" of hypotheses, 50% of which are true, is .5, it does not follow that

this particular hypothesis, the one we happened to select, has a probability of .5, however probability is construed [Mayo, 1997; 2005; 2010b].[10] Besides, it is far from clear which urn of null hypotheses we are to sample from. The answer will greatly alter whether or not there is evidence. Finally, it is unlikely that we would ever know the proportion of true nulls, rather than merely the proportion that have thus far not been rejected by other statistical tests! Whether the priors come from frequencies or from "objective" Bayesian priors, there are claims that we would want to say had passed severely that do not get a high posterior.

These brief remarks put the spotlight on the foundations of current-day reference Bayesians — arguably the predominant form of Bayesianism advocated for science. They are sometimes put forward as a kind of half-way house offering a "reconciliation" between Bayesian and frequentist accounts. Granted, there are cases where it is possible to identify priors that result in posteriors that "match" error probabilities, but they appear to mean different things. Impersonal or reference priors are not be seen as measuring beliefs or even probabilities — they are often improper.[11] Subjective Bayesians often quesiton whether the reference Bayesian is not here giving up on the central Bayesian tenets (e.g., [Dawid, 1997; Lindley, 1997]).

4 ERROR STATISTICS IS SELF-CORRECTING: TESTING STATISTICAL MODEL ASSUMPTIONS

The severity assessment of the primary statistical inference depends on the assumptions of the statistical model $\mathcal{M}$ being approximately true. Indeed, all model-based statistical methods depend, for their validity, on satisfying the model assumptions, at least approximately; a crucial part of the objectivity of error statistical methods is their ability to be used for this self-correcting goal.

Some critics would dismiss the whole endeavor of checking model assumptions on the grounds that:

#12 All models are false anyway.

This charge overlooks the key function in using statistical models, as argued by Cox [1995, p. 456]:

> "... it does not seem helpful just to say that all models are wrong. The very word model implies simplification and idealization. ... The construction of idealized representations that capture important stable

[10] The parallel issue is raised by Bayesian epistemologists; see [Achinstein, 2010; Mayo, 2005; 2010c, 2010d].

[11] Interestingly, some frequentist error statisticians are prepared to allow that reference Bayesian techniques might be regarded as technical devices for arriving at procedures that may be reinterpreted and used by error statisticians, but for different ends (see [Cox, 2006; Cox and Mayo, 2009; Kass and Wasserman, 1996]).

aspects of such systems is, however, a vital part of general scientific analysis."

In order to obtain reliable knowledge of "important stable aspects" of phenomena, tests framed within approximately correct models will do — so long as their relevant error probabilities are close to those calculated. Statistical misspecifications often create sizeable deviations between the calculated or *nominal* error probabilities and the *actual error probabilities* associated with an inference, thereby vitiating error statistical inferences. Since even Bayesian results depend on approximate validity of their statistical models, this might be an area for the Bayesian to employ non-Bayesian methods.

The error statistician pursues testing assumptions using three different types of tools: informal analyses of data plots, non-parametric and parametric tests, and simulation-based methods, including resampling.

Philosophers of science tend to speak of "the data" in a way that does not distinguish the different ways in which a given set of data are modeled, and yet such distinctions are crucial for understanding how reliable tests of assumptions are obtained. In using data to test model assumptions one looks, not at the reduced data in the test statistic (for primary testing), but rather the full data set $\mathbf{x}_0{:=}(x_1, x_2, \ldots, x_n)$. For example, in test T_α, $\overline{X}{=}\frac{1}{n}\sum_{k=1}^{n} X_k$ is a sufficient statistic for parameter μ. That means $\overline{X}$, together with its sampling distribution, contains all the information needed for those inferences. However, $\overline{X}$, by itself, does not provide sufficient information to assess the validity of the model assumptions underlying test T_α above. There are actually four distinct assumptions (table 1).

Table 1 - Simple Normal Model

$$X_k = \mu + u_k,\ k{\in}\mathbb{N},$$

[1]	Normality:	$X_k \backsim \mathsf{N}(.,.)$,	$\left.\begin{array}{l}\\\\\\\end{array}\right\}$ $k{\in}\mathbb{N}$.
[2]	constant mean:	$E(X_k){:=}\,\mu$,	
[3]	constant variance:	$Var(X_k){:=}\,\sigma^2$,	
[4]	Independence:	$\{X_k,\ k{\in}\mathbb{N}\}$ is an independent process.	

The inferences about μ depend on the assumptions, but the tests of those assumptions should not depend on the unknowns. The idea underlying model validation is to construct Mis-Specification (M-S) tests using 'distance' functions whose distribution under the null (the model is valid) is known, and at the same time they have power against potential departures from the model assumptions. M-S tests can be regarded as posing 'secondary' questions to the data as opposed to the primary ones. Whereas primary statistical inferences take place within a specified (or assumed) model $\mathcal{M}$, the secondary inference has to put $\mathcal{M}$'s assumptions to the test; so to test $\mathcal{M}$'s assumptions, we stand outside $\mathcal{M}$, as it were. The generic form of the hypothesis of interest in M-S tests is:

H_0: the assumption(s) of statistical model $\mathcal{M}$ hold for data $\mathbf{x}_0$,

as against the alternative not-H_0, which, in general, consists of all of the ways one or more of $\mathcal{M}$'s assumptions can founder. However, this alternative $[\mathcal{P}-\mathcal{M}]$, where $\mathcal{P}$ denotes the set of all possible models that could have given rise to data $\mathbf{x}_0$, is too unwieldy. In practice one needs to consider a specific form of departure from H_0, say H_r, in order to apply a statistical significance test to H_0 and test results must be interpreted accordingly. Since with this restricted alternative H_r, the null and alternative do not exhaust the possibilities (unlike in the N-P test), a statistically significant departure from the null would not warrant inferring the particular alternative in a M-S test H_r, at least not without further testing; see [Spanos, 2000].

In M-S testing, the logic of significance tests is this: We identify a test statistic $\tau(\mathbf{X})$ to measure the distance between what is observed $\mathbf{x}_0$ and what is expected assuming the null hypothesis H_0 holds, so as to derive the distribution of $\tau(\mathbf{X})$ under H_0. Now the relevant p-value would be:

$$P(\tau(\mathbf{X}) > \tau(\mathbf{x}_0);\ H_0 \text{ true}) = p,$$

and if it is very small, then there is evidence of violations of the assumption(s) in H_0. We leave to one side here the particular levels counting as 'small'. A central asset of the error statistical approach to model validation, is its ability to compute the p-value, and other relevant error probabilities, now dealing with erroneous inferences regarding the assumptions.

Although the alternative may not be explicit in this simple (Fisherian) test, the interest in determining what violations have been ruled out with severity leads one to make them explicit. This may be done by considering the particular violations from H_0 that the given test is capable of probing. This goes beyond what, strictly speaking, is found in standard M-S tests; so once again the severity requirement is directing supplements. The upshot for interpreting M-S test results is this: If the p-value is not small, we are entitled only to rule out those departures that the test had enough capacity to detect. In practice, the alternatives may be left vague or made specific. We consider an example of each, the former with a non-parametric test, the latter with a parametric test.

4.1 Runs Test for IID

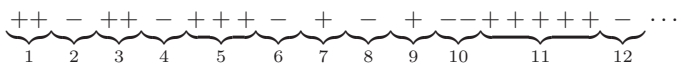
An example of a non-parametric M-S test for IID (assumptions [2]–[4]) is the well-known runs test. The basic idea is that if the sample $\mathbf{X}:=(X_1, X_2, \ldots, X_n)$ is random (IID), then one can compare the number of runs expected $E(R)$ in a typical realization of an IID sample with the number of runs observed $R = r$, giving rise to a test statistic:

$$\tau(\mathbf{X}) = \left|[R-E(R)]/\sqrt{Var(R)}\right|,$$

whose distribution under IID for $n \geq 20$ can be approximated by $\mathsf{N}(0,1)$. The number of runs R is evaluated in terms the residuals:

$$\widehat{u}_k = (X_k - \overline{X}),\ k = 1, 2, ..., n,$$

where instead of the particular value of each observed residual one records its sign, a "+", or negative, a "−", giving rise to patterns of the form:



The patterns we are interested in are called *runs*: a sub-sequence of one type (pluses only or minuses only) immediately preceded and succeeded by an element of the other type.

The appeal of such a non-parametric test is that its own validity does not depend on the assumptions of the (primary) model under scrutiny: we can calculate the probability of different numbers of runs just from the hypothesis that the assumption of randomness holds. As is plausible, then, the test based on R takes the form: Reject H_0 iff the observed R differs sufficiently (in either direction) from $E(R)$ — the expected R under the assumption of IID. The p-value is:

$$P(\tau(\mathbf{X}) > \tau(\mathbf{x}_0); \text{ IID true}) = p.$$

However, since the test is sensitive to any form of departures from the IID assumptions, rejecting the null only warrants inferring a denial of IID. The test itself does not indicate whether the fault lies with one or the other or both assumptions. Combining this test with other misspecification analyses, however, can; [Mayo and Spanos, 2004].

4.2 A Parameteric Test of Independence

Let us now compare this to a parametric M-S test. We begin by finding a way to formally express the denial of the assumption in question by means of a parameter value in an encompassing model. In particular, the dependence among the X_k's may be formally expressed as an assertion that the correlation between any X_i and X_j for $i \neq j$ is non-zero, which in turn may be parameterized by the following **AutoRegressive** (AR(1)) model:

$$X_k = \beta_0 + \beta_1 X_{k-1} + \varepsilon_t, \ k = 1, 2, ..., n.$$

In the context of this encompassing model the independence assumption in [4] can be tested using the parametric hypotheses:

$$H_0 : \beta_1 = 0, \text{vs. } H_1 : \beta_1 \neq 0.$$

Notice that under H_0 the AR(1) model reduces to $X_k = \mu + u_t$; see table 1. Rejection of the null based on a small enough p-value provides evidence for a violation of independence. Failing to reject entitles us to claim that the departures against which the test was capable of detecting are not present; see [Spanos, 1999] for further discussion.

4.3 Testing Model Assumptions Severely

In practice one wants to perform a variety of M-S tests assessing different subsets of assumptions [1]–[4]; using tests which themselves rely on dissimilar assumptions. The secret lies in shrewd testing strategies: following a logical order of parametric and non-parametric tests, and combining tests that jointly probe several violations with deliberately varied assumptions. This enables one to argue, with severity, that when no departures from the model assumptions are detected, despite all of these distinct probes, the model is adequate for the primary inferences. The argument is entirely analogous to the argument from coincidence that let us rule out values of George's weight gain earlier on.

To render the probing more effective, error statisticians employ data analytic techniques and data plots to get hints of the presence of potential violations, indicating the most fruitful analytic tests to try. In relation to this some critics charge:

#13 Testing assumptions involves illicit data-mining.

The truth of the matter is that data plots provide the best source of information pertaining to potential violations that can be used to guide a more informative and effective probing. Far from being *illicit data mining*, it is a powerful way to get ideas concerning the type of M-S tests to apply to check assumptions most severely. It provides an effective way to probe what is responsible for the observed pattern, much as a forensic clue is used to pinpoint the culprit; see [Spanos, 2000]. The same logic is at the heart of non-parametric tests of assumptions, such as the runs test.

4.4 Residuals provide the Key to M-S testing

A key difference between testing the primary hypotheses of interest and M-S testing is that they pose very different questions to data in a way that renders the the tests largely independent of each other. This can be justified on formal grounds using the properties of *sufficiency* and *ancillarity*; see [Cox and Hinkley, 1974].

It can be shown [Spanos, 2007] that, in many cases, including the above example of the simple Normal model, the information used for M-S testing purposes is independent of the information used in drawing primary inferences. In particular, the distribution of the sample for the statistical model in question simplifies as follows:

$$(3) \qquad f(\mathbf{y};\theta) \propto f(\mathbf{s};\theta)\cdot f(\mathbf{r}), \qquad \forall\,(\mathbf{s},\mathbf{r})\in\mathbb{R}_s^m\times\mathbb{R}_r^{n-m}.$$

where the statistics $\mathbf{R}$ and $\mathbf{S}$, are not only independent, but $\mathbf{S}$ is a *sufficient* statistic for $\theta{:=}(\mu,\sigma^2)$ (the unknown parameters of the statistical model) and $\mathbf{R}$ is *ancillary* for θ, i.e. $f(\mathbf{r})$ does *not* depend on θ. Due to these properties, the primary inference about θ can be based solely on the distribution of the sufficient

statistic $f(\mathbf{s};\theta)$, and $f(\mathbf{r})$ can be used to assess the validity of the statistical model in question.

In the case of the simple Normal model (table 1), the statistics $\mathbf{R}$ and $\mathbf{S}$ take the form:

$$\mathbf{S}{:=}(\overline{X},s),\ \text{where}\ \overline{X}{=}\tfrac{1}{n}\textstyle\sum_{k=1}^{n}X_k,\ s^2=\tfrac{1}{n-1}\sum_{k=1}^{n}(X_k-\overline{X})^2,$$

$$\mathbf{R}{:=}(\widehat{v}_3,..,\widehat{v}_n),\ \widehat{v}_k=\tfrac{\sqrt{n}\widehat{u}_k}{s}=\tfrac{\sqrt{n}(X_k-\overline{X})}{s}\backsim \mathsf{St}(n{-}1),\ k=3,4,..,n,$$

where $(\widehat{v}_1,..,\widehat{v}_n)$ are known as *studentized residuals*; see [Spanos, 2007]. Note that the runs test, discussed above, relies on residuals because it is based on replacing their numerical values with their sign (+ or −). Likewise, the parametric M-S test for independence, placed in the context of the AR(1) model, can be shown to be equivalently based on the auxiliary autoregression in terms of the residuals:

$$\widehat{u}_k=\beta_0+\beta_1\widehat{u}_{k-1}+\varepsilon_t,\ k{=}1,2,...,n.$$

The above use of the residuals for model validation is in the spirit of the strategy in [Cox and Hinkley, 1974] to use the conditional distribution $f(\mathbf{x}\mid\mathbf{s})$ to assess the adequacy of the statistical model. What makes $f(\mathbf{x}\mid\mathbf{s})$ appropriate for assessing model adequacy is that when $\mathbf{s}$ is a sufficient statistic for θ, $f(\mathbf{x}\mid\mathbf{s})$ is free of the unknown parameter(s). The simple Poisson and Bernoulli models provide such examples [Cox, 2006, p. 33].

4.5 Further Topics, Same Logic

If one finds violations of the model assumptions then the model may need to be respecified to capture the information not accounted for, but the general discussion of respecification is beyond the scope of this entry; see [Spanos, 2006].

A distinct program of research for the error statistician is to explore the extent to which violations invalidate tests. Thanks to robustness, certain violations of the model assumptions will not ruin the validity of the test concerning the primary hypothesis.

Of particular interest in error statistics is a set of computer-intensive techniques known as resampling procedures, including permutation methods, the bootstrap and Monte Carlo simulations, which are based on empirical relative frequencies. Even without knowing the sampling distribution, one can, in effect generate it by means of these techniques. The logic underlying the generation of these simulated realizations is based on counterfactual reasoning: We ask, 'what would it be like (in terms of sampling distributions of interest) were we to sample from one or another assumed generating mechanism?' The results can then be used to empirically construct (by "brute force" some claim) the sampling distributions of any statistic of interest and their corresponding error probabilities.

This is particularly useful in cases where the sampling distribution of an estimator or a test statistic cannot be derived analytically, and these resampling methods

can be used to evaluate it empirically; see [Efron and Tibshirani, 1993]. The same pattern of *counterfactual reasoning* around which severity always turns is involved, thus unifying the methods under the error statistical umbrella. Here, however, the original data are compared with simulated replicas generated under a number of different data generating mechanisms, in order to discern the discrepancy between what was observed and "what it would be like" under various scenarios. It should be noted that model specification is distinct from model selection, which amounts to choosing a particular model within a prespecified family of models; see [Spanos, 2010].

BIBLIOGRAPHY

[Achinstein, 2010] P. Achinstein. Mill's Sins or Mayo's Errors? pp. 170-188 in *Error and Inference*, (ed.) by D. G. Mayo and A. Spanos, Cambridge University Press, Cambridge, 2010.

[Armitage, 1975] P. Armitage. *Sequential Medical Trials*, 2nd ed, Wiley, NY, 1975.

[Barnard, 1972] G. A. Barnard. The Logic of Statistical Inference (review of Ian Hacking *The Logic of Statistical Inference*), *British Journal For the Philosophy of Science*, 23: 123-132, 1972.

[Barnett, 1999] V. Barnett. *Comparative Statistical Inference*, 3rd ed.,Wiley, NY, 1999.

[Bernardo, 2005] J. M. Bernardo. Reference Analysis, pp. 17–90 in *Handbook of Statistics*, vol. 25: Bayesian Thinking, Modeling and Computation, D. K. Dey and C. R. Rao, (eds.), Elsevier, North-Holland, 2005.

[Berger, 2004] J. Berger. The Case for Objective Bayesian Analysis, *Bayesian Analysis*, 1, 1–17, 2004.

[Berger and Selke, 1987] J. Berger and T. Sellke. Testing a point-null hypothesis: the irreconcilability of significance levels and evidence, *Journal of the American Statistical Association*, 82: 112-122, 1987.

[Berger and Wolpert, 1988] J. Berger and R. Wolpert. *The Likelihood Principle*. 2d ed., Institute of Mathematical Statistics, Hayward, CA, 1988.

[Birnbaum, 1961] A. Birnbaum. Confidence Curves: An Omnibus Technique for Estimation and Testing, *Journal of the American Statistical Association*, 294: 246-249, 1961.

[Birnbaum, 1962] A. Birnbaum. On the Foundations of Statistical Inference (with discussion), *Journal of the American Statistical Association*, 57: 269-326, 1962.

[Birnbaum, 1969] A. Birnbaum. Concepts of Statistical Evidence, pp. 112-143 in S. Morgenbesser, P. Suppes, and M. White (eds.), *Philosophy, Science, and Method: Essays in Honor of Ernest Nagel*, St. Martin's Press, NY, 1969.

[Casella and Berger, 1987] G. Casella and R. Berger. Reconciling Bayesian and Frequentist Evidence in the One-Sided Testing Problem, *Journal of the American Statistical Association*, 82: 106-111, 1987.

[Cheyne and Worrall, 2006] C. Cheyne and J. Worrall, eds. *Rationality and Reality: Conversations with Alan Musgrave*, Studies in the History and Philosophy of Science, Springer, Dordrecht, 2006.

[Cox, 1958] D. R. Cox. Some Problems Connected with Statistical Inference, *Annals of Mathematical Statistics*, 29: 357-372, 1958.

[Cox, 1977] D. R. Cox. The Role of Significance Tests, (with discussion), *Scandinavian Journal of Statistics*, 4:49-70, 1977.

[Cox, 1995] D. R. Cox. Comment on "Model Uncertainty, Data Mining and Statistical Inference," by C. Chatfield, *Journal of the Royal Statistical Society*, A 158: 419-466, 1995.

[Cox, 2006] D. R. Cox. *Principles of Statistical Inference*, Cambridge University Press, Cambridge, 2006.

[Cox and Hinkley, 1974] D. R. Cox and D. V. Hinkley. *Theoretical Statistics*, Chapman and Hall, London, 1974.

[Cox and Mayo, 2010] D. R. Cox and D. G. Mayo. Objectivity and Conditionality in Frequentist Inference, pp. 276-304 in Mayo, D.G. and A. Spanos, *Error and Inference*, Cambridge University Press, Cambridge, 2010.
[Dawid, 1997] A. P. Dawid. Comments on 'Non-informative priors do not exist", *Journal of Statistical Planning and Inference*, 65: 159-162, 1997.
[Edwards *et al.*, 1963] W. Edwards, H. Lindman, and L. Savage. Bayesian Statistical Inference for Psychological Research, *Psychological Review*, 70: 193-242, 1963.
[Efron and Tibshirani, 1993] B. Efron and R. J. Tibshirani. *An Introduction to the Bootstrap*, Chapman and Hall, London, 1993.
[Fisher, 1925] R. A. Fisher. *Statistical Methods for Research Workers*, Oliver and Boyd, Edinburgh, 1925.
[Fisher, 1935] R. A. Fisher. *The Design of Experiments*, Oliver and Boyd, Edinburgh, 1935.
[Fisher, 1955] R. A. Fisher. Statistical Methods and Scientific Induction, *Journal of the Royal Statistical Society*, B, 17: 69-78, 1955.
[Fisher, 1956] R. A. Fisher. *Statistical Methods and Scientific Inference*, Oliver and Boyd, Edinburgh, 1956.
[Gigerenzer, 1993] G. Gigerenzer. The Superego, the Ego, and the Id in Statistical Reasoning, pp. 311-39 in Keren, G. and C. Lewis (eds.), *A Handbook of Data Analysis in the Behavioral Sciences: Methodological Issues*, Erlbaum, Hillsdale, NJ, 1993.
[Godambe and Sprott, 1971] V. Godambe and D. Sprott, eds. *Foundations of Statistical Inference*, Holt, Rinehart and Winston of Canada, Toronto, 1971.
[Hacking, 1965] I. Hacking. *Logic of Statistical Inference*, Cambridge University Press, Cambridge, 1965.
[Hacking, 1980] I. Hacking. The Theory of Probable Inference: Neyman, Peirce and Braithwaite, pp. 141-160 in D. H. Mellor (ed.), *Science, Belief and Behavior: Essays in Honour of R.B. Braithwaite*, Cambridge University Press, Cambridge, 1980.
[Hull, 1988] D. Hull. *Science as a Process: An Evolutionary Account of the Social and Conceptual Development of Science*, University of Chicago Press, Chicago, 1988.
[Jeffreys, 1939] H. Jeffreys. *Theory of Probability*, Oxford University Press, Oxford, 1939.
[Kadane *et al.*, 1999] J. Kadane, M. Schervish, and T. Seidenfeld. *Rethinking the Foundations of Statistics*, Cambridge University Press, Cambridge, 1999.
[Kass and Wasserman, 1996] R. E. Kass and L.Wasserman. The Selection of Prior Distributions by Formal Rules, *Journal of the American Statistical Association*, 91: 1343-1370, 1996.
[Kempthorne and Folks, 1971] O. Kempthorne and L. Folks. *Probability, Statistics, and Data Analysis*, The Iowa State University Press, Ames, IA, 1971.
[Lehmann, 1986] E. L. Lehmann. *Testing Statistical Hypotheses*, 2nd edition, Wiley, New York, 1986.
[Lehmann, 1993] E. L. Lehmann. The Fisher and Neyman-Pearson Theories of Testing Hypotheses: One Theory or Two? *Journal of the American Statistical Association*, 88: 1242-9, 1993.
[Lindley, 1997] D. Lindley. Some comments on "Non-informative priors do not exist", *Journal of Statistical Planning and Inference*, 65: 182-189, 1997.
[Mayo, 1983] D. G. Mayo. An Objective Theory of Statistical Testing, *Synthese*, 57: 297-340, 1983.
[Mayo, 1985] D. G. Mayo. Behavioristic, Evidentialist, and Learning Models of Statistical Testing, *Philosophy of Science*, 52: 493-516, 1985.
[Mayo, 1996] D. G. Mayo. *Error and the Growth of Experimental Knowledge*, The University of Chicago Press, Chicago, 1996.
[Mayo, 1997] D. G. Mayo. Severe Tests, Arguing From Error, and Methodological Underdetermination, *Philosophical Studies*, 86: 243-266, 1997.
[Mayo, 2003] D. G. Mayo. Could Fisher, Jeffreys and Neyman Have Agreed? Commentary on J. Berger's Fisher Address, *Statistical Science,* 18, 2003: 19-24, 2003.
[Mayo, 2005] D. G. Mayo. Evidence as Passing Severe Tests: Highly Probed vs. Highly Proved, pp. 95-127 in *Scientific Evidence*, P. Achinstein (ed.), Johns Hopkins University Press, 2005.
[Mayo, 2006a] D. G. Mayo. Philosophy of Statistics, pp. 802-815 in S. Sarkar and J. Pfeifer (eds.), *Philosophy of Science: An Encyclopedia*, London: Routledge, 2006.
[Mayo, 2006b] D. G. Mayo. Critical Rationalism and Its Failure to Withstand Critical Scrutiny, pp. 63-96 in C. Cheyne and J. Worrall (eds.) *Rationality and Reality: Conversations with Alan Musgrave,* Studies in the History and Philosophy of Science, Springer, Dordrecht, 2006.

[Mayo, 2008] D. G. Mayo. How to Discount Double Counting When It Counts, *British Journal for the Philosophy of Science*, 59: 857–79, 2008.
[Mayo, 2010a] D. G. Mayo. Learning from Error, Severe Testing, and the Growth of Theoretical Knowledge, pp. 28-57 in Mayo, D.G. and A. Spanos, *Error and Inference*, Cambridge University Press, Cambridge, 2010.
[Mayo, 2010b] D. G. Mayo. An Error in the Argument from Conditionality and Sufficiency to the Likelihood Principle, pp. 305-314 in Mayo, D.G. and A. Spanos, *Error and Inference*, Cambridge University Press, Cambridge, 2010.
[Mayo, 2010c] D. G. Mayo. Sins of the Epistemic Probabilist Exchanges with Peter Achinstein, pp. 189-201 in Mayo, D.G. and A. Spanos, *Error and Inference*, Cambridge University Press, Cambridge, 2010.
[Mayo, 2010d] D. G. Mayo. The Objective Epistemic Epistemologist and the Severe Tester, in *Philosophy of Science Matters: The Philosophy of Peter Achinstei*, edited by Gregory Morgan, Oxford University Press, 2010.
[Mayo and Kruse, 2001] D. G. Mayo and M. Kruse. Principles of Inference and their Consequences, pp. 381-403 in *Foundations of Bayesianism*, edited by D. Cornfield and J. Williamson, Kluwer Academic Publishers, Netherlands, 2001.
[Mayo and Spanos, 2004] D. G. Mayo and A. Spanos. Methodology in Practice: Statistical Misspecification Testing, *Philosophy of Science*, 71: 1007-1025, 2004.
[Mayo and Spanos, 2006] D. G. Mayo and A. Spanos. Severe Testing as a Basic Concept in a Neyman-Pearson Philosophy of Induction, *The British Journal for the Philosophy of Science*, 57: 323-357, 2006.
[Mayo and Cox, 2006] D. G. Mayo and D. R. Cox. Frequentist Statistics as a Theory of Inductive Inference, pp. 77-97 in in *Optimality: The Second Erich L. Lehmann Symposium*, edited by J. Rojo, Lecture Notes-Monograph Series, vol. 49, Institute of Mathematical Statistics, Beachwood, OH, 2006.
[Mayo and Spanos, 2010] D. G. Mayo and A. Spanos. *Error and Inference: Recent Exchanges on the Philosophy of Science, Inductive-Statistical Inference, and Reliable Evidence*, Cambridge University Press, Cambridge 2010.
[Neyman, 1935] J. Neyman. On the Problem of Confidence Intervals, *The Annals of Mathematical Statistics*, 6: 111-116, 1935.
[Neyman, 1952] J. Neyman. *Lectures and Conferences on Mathematical Statistics and Probability*, 2nd ed. U.S. Department of Agriculture, Washington. 1952.
[Neyman, 1956] J. Neyman. Note on an Article by Sir Ronald Fisher, *Journal of the Royal Statistical Society*, Series B (Methodological), 18: 288-294, 1956.
[Neyman, 1957] J. Neyman. Inductive Behavior as a Basic Concept of Philosophy of Science, *Revue Inst. Int. De Stat.*, 25: 7-22, 1957.
[Neyman, 1971] J. Neyman. Foundations of Behavioristic Statistics, pp. 1-19 in Godambe and Sprott, eds., 1971.
[Neyman and Pearson, 1933] J. Neyman and E. S. Pearson. On the Problem of the Most Efficient Tests of Statistical Hypotheses, *Philosophical Transactions of the Royal Society*, A, 231: 289-337, 1933.
[Neyman and Pearson, 1967] J. Neyman and E. S. Pearson. *Joint Statistical Papers of J. Neyman and E. S. Pearson*, Unversity of California Press, Berkeley, 1967.
[Pearson, 1955] E. S. Pearson. Statistical Concepts in Their Relation to Reality, *Journal of the Royal Statistical Society*, B, 17: 204-207, 1955.
[Pearson, 1962] E. S. Pearson. Some Thoughts on Statistical Inference, *Annals of Mathematical Statistics*, 33: 394-403, 1962.
[Peirce, 1931-5] C. S. Peirce. *Collected Papers*, vols. 1-6, edited by C. Hartshorne and P. Weiss, Harvard University Press, Cambridge, MA, 1931-5.
[Poole, 1987] C. Poole. Beyond the Confidence Interval, *The American Journal of Public Health*, 77: 195-199, 1987.
[Popper, 1959] K. Popper. *The Logic of Scientific Discovery*, Basic Books, NY, 1959.
[Pratt, 1977] J. W. Pratt. Decisions as Statistical Evidence and Birnbaum's Confidence Concept, *Synthese*, 36: 59-69, 1977.
[Rosenkrantz, 1977] R. Rosenkrantz. *Inference, Method and Decision: Towards a Bayesian Philosophy of Science*, Reidel, Dordrecht, 1977.
[Rosenthal and Gaito, 1963] R. Rosenthal and J. Gaito. The Interpretation of Levels of Significance by Psychological Researchers, *Journal of Psychology*, 55: 33-38, 1963.

[Royall, 1997] R. Royall. *Statistical Evidence: a Likelihood Paradigm*, Chapman and Hall, London, 1997.

[Savage, 1962] L. Savage, ed. *The Foundations of Statistical Inference: A Discussion*, Methuen, London, 1962.

[Sober, 2008] P. Sober. *Evidence and Evolution: The Logic Behind the Science*, Cambridge University Press, Cambridge, 2008.

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

[Spanos, 2000] A. Spanos. Revisiting Data Mining: 'hunting' with or without a license, *The Journal of Economic Methodology*, 7: 231-264, 2000.

[Spanos, 2007] A. Spanos. Using the Residuals to Test the Validity of a Statistical Model: Revisiting Sufficiency and Ancillarity, Working Paper, Virginia Tech, 2007.

[Spanos, 2006] A. Spanos. Where Do Statistical Models Come From? Revisiting the Problem of Specification, pp. 98-119 in *Optimality: The Second Erich L. Lehmann Symposium*, edited by J. Rojo, Lecture Notes-Monograph Series, vol. 49, Institute of Mathematical Statistics, Beachwood, OH, 2006.

[Spanos, 2010] A. Spanos. Akaike-type Criteria and the Reliability of Inference: Model Selection vs. Statistical Model Specification, forthcoming, *Journal of Econometrics*, 2010.