Brit. J. Phil. Sci. **57** (2006), 323–357

Severe Testing as a Basic Concept in a Neyman–Pearson Philosophy of Induction

Deborah G. Mayo and Aris Spanos

ABSTRACT

Despite the widespread use of key concepts of the Neyman–Pearson (N–P) statistical paradigm—type I and II errors, significance levels, power, confidence levels—they have been the subject of philosophical controversy and debate for over 60 years. Both current and long-standing problems of N–P tests stem from unclarity and confusion, even among N–P adherents, as to how a test's (pre-data) error probabilities are to be used for (post-data) *inductive inference* as opposed to *inductive behavior*. We argue that the relevance of error probabilities is to ensure that only statistical hypotheses that have passed *severe* or probative tests are inferred from the data. The *severity criterion* supplies a *meta-statistical* principle for evaluating proposed statistical inferences, avoiding classic fallacies from tests that are overly sensitive, as well as those not sensitive enough to particular errors and discrepancies.

1 *Introduction and overview*
- **1.1** *Behavioristic and inferential rationales for Neyman–Pearson (N–P) tests*
- **1.2** *Severity rationale: induction as severe testing*
- **1.3** *Severity as a meta-statistical concept: three required restrictions on the N–P paradigm*

2 *Error statistical tests from the severity perspective*
- **2.1** *N–P test $T(\alpha)$: type I, II error probabilities and power*
- **2.2** *Specifying test $T(\alpha)$ using p-values*

3 *Neyman's post-data use of power*
- **3.1** *Neyman: does failure to reject H warrant confirming H?*

4 *Severe testing as a basic concept for an adequate post-data inference*
- **4.1** *The severity interpretation of acceptance (SIA) for test $T(\alpha)$*
- **4.2** *The fallacy of acceptance (i.e., an insignificant difference): Ms Rosy*
- **4.3** *Severity and power*

5 *Fallacy of rejection: statistical vs. substantive significance*
- **5.1** *Taking a rejection of H_0 as evidence for a substantive claim or theory*
- **5.2** *A statistically significant difference from H_0 may fail to indicate a substantively important magnitude*
- **5.3** *Principle for the severity interpretation of a rejection (SIR)*

© The Author (2006). Published by Oxford University Press on behalf of *British Society for the Philosophy of Science*. All rights reserved.
doi:10.1093/bjps/axl003
For Permissions, please email: journals.permissions@oxfordjournals.org
Advance Access published on April 11, 2006

5.4 *Comparing significant results with different sample sizes in $T(\alpha)$: large n problem*
5.5 *General testing rules for $T(\alpha)$, using the severe testing concept*
6 *The severe testing concept and confidence intervals*
6.1 *Dualities between one and two-sided intervals and tests*
6.2 *Avoiding shortcomings of confidence intervals*
7 *Beyond the N–P paradigm: pure significance, and misspecification tests*
8 *Concluding comments: have we shown severity to be a basic concept in a N–P philosophy of induction?*

1 Introduction and overview

Questions about the nature and justification of probabilistic and statistical methods have long been of central interest to philosophers of science. Debates over some of the most widely used statistical tools—significance tests, Neyman–Pearson (N–P) tests and estimation—over the past 60 years, are entwined with a core philosophical question:

> 'Where should probability enter in inductive inference in science?'

In tackling this question there are two main distinct philosophical traditions from which to draw (Pearson [1950], p. 394). In one, probability is used to provide a post-data assignment of degree of probability, confirmation, or belief in a hypothesis; while in a second, probability is used to assess the reliability of a test procedure to assess and control the frequency of errors in some (actual or hypothetical) series of applications (error probabilities). We may call the former *degree of confirmation* approaches, the latter, *error probability* or *error statistical* approaches. Since the former has seemed most in sync with philosophical conceptions of inductive inference, while the latter is embodied in statistical significance tests and N–P methods, it is easy to see why conflict has abounded in the philosophical literature. The 'error probability' versus 'degree of confirmation' debates take such forms as: decision vs. inference, pre-data vs. post-data properties, long-run vs. single case, and have been discussed by numerous philosophers e.g., Earman, Fetzer, Giere, Gillies, Glymour, Hacking, Horwich, Howson, Kyburg, Levi, Peirce, Rosenkrantz, Salmon, Seidenfeld, Spielman, Urbach.[1]

As advances in computer power have made available sophisticated statistical methods from a variety of schools (N–P, Fisherian, Bayesian,

[1] A partial list among statisticians who contributed to these debates: Armitage, Barnard, Berger, Birnbaum, Cox, de Finetti, Edwards, Efron, Fisher, Good, Jeffreys, Kempthorne, LeCam, Lehmann, Lindley, Neyman, Pearson, Pratt, Savage. Excellent collections of contributions by philosophers and statisticians are Godambe and Sprott ([1971]), and Harper and Hooker ([1976]).

algorithmic), a happy eclecticism may seem to have diminished the need to resolve the philosophical underpinnings of the use and interpretation of statistical methods. However, the significance test controversy is still hotly debated among practitioners, particularly in psychology, epidemiology, ecology, and economics; one almost feels as if each generation fights the 'statistics wars' anew, with journalistic reforms, and task forces aimed at stemming the kind of automatic, recipe-like uses of significance tests that have long been deplored.[2] Moreover, the newer statistical methods involving model selection algorithms and multiple hypothesis testing do not get away from, but rather pile up applications of, significance test results. Having never resolved satisfactorily questions of the role of error probabilities, practitioners face a shortage of general principles for how—or even whether—to calculate error probabilities in such contexts.

Not that practitioners are waiting for philosophers to sort things out. We read, for instance, in a recent article in *Statistical Science*: 'professional agreement on statistical philosophy is not on the immediate horizon, but this should not stop us from agreeing on methodology' (Berger [2003], p. 2). However, the latter question, we think, turns on the former.[3] Seeking an agreement on numbers first, with the assumption that philosophy will follow, leads to 'reconciliations' that may not do justice to core principles underlying the disparate philosophies involved. In particular, using error probabilities as posterior probabilities (however ingenious the latest attempts), leads to 'hybrids' from mutually inconsistent statistical paradigms (Gigerentzer [1993]). Many Bayesian practitioners, wishing to avoid the infirmities of eliciting and depending on subjective prior probabilities, turn to developing prior 'weights' as reference points from which to calculate 'objective' posteriors. However, the various proposed 'reference' priors are themselves open to persistent problems and paradoxes (Kass and Wasserman [1996]). David Cox recently argued that their main conceptual justification is that, in a given class of cases, they lead, at least approximately, to procedures with acceptable frequentist properties (Cox [2006]), thereby raising anew the question of the nature and role of frequentist error probabilities. While not wishing to refight old battles, we propose to reopen the debate from a contemporary perspective—one that will allow developing an interpretation of tests (and associated methods) that avoids cookbooks, is inferential, and yet keeps to the philosophy of frequentist error probability statistics. Ironically, we will extract some needed threads from little discussed papers by Neyman—one

[2] The social science literature criticizing significance testing is too vast to encompass; some key sources are: Cohen ([1988]), Harlow et al. ([1997]), Morrison and Henkel ([1970]), MSERA ([1998]), Thompson ([1996]).

[3] In this article, Berger considers how to reconcile Fisher and Neyman, as well as Jeffreys; see the comments in Mayo ([2003b]).

of the key, early, controversial figures (Neyman [1955], [1956], [1957a], [1957b]).

1.1 Behavioristic and inferential rationales for Neyman–Pearson (N–P) tests

In 'Inductive Behavior as a Basic Concept of Philosophy of Science,' Jerzy Neyman ([1957a]) suggests that 'in relation to science, the philosophy of science plays the same role as anatomy and neurology play in relation to the process of walking' (pp. 7–8): to understand and improve on the proper functioning of the processes in question. In a proper 'anatomization' of the process of statistical induction, Neyman charges, 'the term ''inductive reasoning'' is a misnomer, . . . and that a better term would be something like *inductive behavior*' (p. 8)—the process of adjusting our actions to observations.

A Neyman and Pearson (N–P) test, as Neyman interprets it, is a rule of *inductive behavior*:

> To decide whether a hypothesis, H, of a given type be rejected or not, calculate a specified character, $t(\boldsymbol{x}_0)$ of the observed facts [the test statistic]; if $t(\boldsymbol{x}) > t(\boldsymbol{x}_0)$ Reject H; if $t(\boldsymbol{x}) \leq t(\boldsymbol{x}_0)$ Accept H (Neyman and Pearson [1933], p. 291).

'Accept/Reject' are identified with deciding to take specific actions, for example, rejecting H might be associated with publishing a result, or announcing a new effect. The set of outcomes that lead to 'Reject H' make up the test's *rejection (*or *critical) region*; it is specified so that:

> it may often be proved that if we behave according to such a rule . . . we shall reject H when it is true not more, say, than once in a hundred times, and in addition we may have evidence that we shall reject H sufficiently often when it is false. (Neyman and Pearson [1933], p. 291)

Why should one accept/reject statistical hypotheses in accordance with a test rule with good error probabilities? The inductive behaviorist has a ready answer:

> ***Behavioristic rationale*:** We are justified in 'accepting/rejecting' hypotheses in accordance with tests having low error probabilities because we will rarely err in repeated applications.

By and large, however, error statistical practitioners seem to regard the error probabilistic behavior of tests in (actual or hypothetical) repetitions as simply a useful way to describe the properties of tests: and these properties enable tests to suitably function for inductive inference in science. Indeed, wishing to disentangle themselves from the decision-behavior construal, most users of N–P tests favor such generic labels as hypothesis tests, statistical significance tests, or error probability methods—we will use *error statistics* for short. Their

thinking, if only implicit, is that error probabilities admit of an inferential rationale:

> ***Inferential Rationale (general)*:** Error probabilities provide a way to determine the evidence a set of data $\boldsymbol{x}_0$ supplies for making warranted inferences about the process giving rise to data $\boldsymbol{x}_0$.

The as yet unanswered question is *how do error statistical tests satisfy the inferential rationale?* How, in short, should we bridge the gap from error properties of procedures to specific inferences based on them:

Error probabilities → inference

An adequate answer requires the philosophical 'anatomist' to go beyond the traditional N–P paradigm which leaves this issue unattended, where we understand by the 'N–P paradigm' the uninterpreted statistical tools based on error probabilities. Whether this 'going beyond' is to be viewed as a reinterpretation, extension, or theoretical foundation of N–P theory, in order for it to succeed, it must address three main problems that have long been taken as obstacles for using N–P tests for inductive inference as opposed to inductive behavior; namely, that N–P tests are too:

(i) ***Coarse*:** N–P tests tell us whether to reject or accept hypotheses according to whether $t(\mathbf{x})$ falls in the test's rejection region or not, but evaluating evidence and inference post-data seem to require more data-specific interpretations.

(ii) ***Open to Fallacies*:** N–P tests give rise to fallacies of rejection (statistical significance vs. substantive significance) and of acceptance (no evidence against is not evidence for).

(iii) ***Focused on Pre-Data, Behavioristic Goals*** *(in specifying and justifying tests):* The good long-run performance characteristics of N–P tests (low type I and type II error probabilities) may conflict with criteria that seem appropriate for inference once the data are available, i.e., post-data.

By assuming the former, *degree-of-confirmation* philosophy, it is often supposed that in order for N–P methods to avoid problems (i)–(iii), pre-data error probabilities must be made to supply hypotheses with some post-data degrees of confirmation or support:

> ***Degree of confirmation rationale*:** Error probabilities may be used post-data to assign degrees of confirmation or support to hypotheses.

But error probabilities do not, nor were they intended to, supply such degrees of probability or confirmation; interpreting them as if they did yields inconsistent 'hybrids'. Post-data (posterior) degrees of probability require prior probability assignments to (an exhaustive set of) hypotheses, and N–P tests were developed to avoid reliance on such prior probabilities, however they

are interpreted (e.g., logical, subjective). One may simply posit the inferential rationale, by fiat, but this is to skirt and not answer the philosophical question (Hacking [1965]; Birnbaum [1969]). Birnbaum ([1977]) attempted an 'evidential' interpretation of N–P tests by means of his 'Confidence Concept,' but this remains a pre-data error probability notion. An attempt by Kiefer ([1977]) to deal with the coarseness problem via his notion of 'conditional' error probabilities also differs from the approach we will take.

1.2 Severity rationale: induction as severe testing

We propose to argue that N–P tests can (and often do) supply tools for inductive inference by providing methods for evaluating the *severity* or *probativeness* of tests. An inductive inference, in this conception, takes the form of inferring hypotheses or claims that survive severe tests. In the 'severe testing' philosophy of induction, the quantitative assessment offered by error probabilities tells us not 'how probable', but rather, 'how well probed' hypotheses are. This suggests how to articulate the general inferential rationale we seek:

> ***Severity rationale*:** Error probabilities may be used to make inferences about the process giving rise to data, by enabling the assessment of how well probed or how severely tested claims are, with data $\boldsymbol{x}_0$.

Although the degree of severity with which a hypothesis H has passed a test is used to determine if it is warranted to infer H, the degree of severity is not assigned to H itself: it is an attribute of the test procedure as a whole (including the inference under consideration). The intuition behind requiring **severity** is that:

> *Data $\boldsymbol{x}_0$ in test T provide good evidence for inferring H (just) to the extent that H passes severely with $\boldsymbol{x}_0$, i.e., to the extent that H would (very probably) not have survived the test so well were H false.*

Karl Popper is well known to have insisted on *severe tests*: 'Observations or experiments can be accepted as supporting a theory (or a hypothesis, or a scientific assertion) [H] *only if these observations or experiments are severe tests of the theory*' (Popper [1994], p. 89)—that is, H survived 'serious criticism'. However, Popper, and the modern day 'critical rationalists' deny they are commending a reliable process—or at least, 'they must deny this if they [are] to avoid the widespread accusation that they smuggle into their theory either inductive reasoning or some metaphysical inductive principle.' (Musgrave [1999], pp. 246–7). All we know, says Popper, is that the surviving hypotheses 'may be true'; but high corroboration is at most a report of H's past performance—we are not warranted in *relying* on it. By contrast, N–P tests will be regarded as good only insofar as they can be shown to have appropriately low error probabilities, which itself involves inductive justification. (For further discussion of critical rationalism see Mayo [2006], pp. 63–96).

1.3 Severity as a meta-statistical concept: three required restrictions on the N–P paradigm

N–P tests do not directly supply severity assessments. Having specified a null hypothesis H_0, and an alternative hypothesis H_1 (the complement of H_0) a N–P test, mathematically speaking, is simply a rule that maps each possible outcome $\boldsymbol{x} = (x_1, \ldots, x_n)$ into H_0 or H_1, so as to control at small values the probability of erroneous rejections (type I error) and erroneous acceptances (type II error). The severity principle is a *meta-statistical* principle to direct the uses of tests for the severity goal. Although N–P tests map data into two outputs, accept and reject, both may be regarded as *passing* a given statistical claim H with which data $\boldsymbol{x}$ agrees; we have then to ascertain if such agreement would occur (and how frequently) under specific denials of H. That is,

A statistical hypothesis H passes a **severe test** T with data $\boldsymbol{x}_0$ if,
(S-1) $\boldsymbol{x}_0$ agrees with H, and
(S-2) with very high probability, test T would have produced a result that accords *less* well with H than $\boldsymbol{x}_0$ does, if H were false.[4]

Our specific focus will be on cases where 'H is false' refers to *discrepancies* from parameters in a statistical model, but we will also suggest how the idea may be generalized to inferring the presence of 'an error' or flaw, very generally conceived. A main task for statistical testing is to learn, not just whether H is false, but approximately, how far from true H is, with respect to parameters in question.

The severity function has three arguments: a test, an outcome or result, and an inference or a claim. 'The severity with which inference H passes test T with outcome $\boldsymbol{x}$' may be abbreviated by:

$$\text{SEV}(\text{Test } T, \text{ outcome } \boldsymbol{x}, \text{ claim } H).$$

Granted, other terms could serve as well to bring out the essential features of our conception; the main thing is to have a notion that exemplifies the 'probative' concept, that is not already attached to other views, and to which the formal apparatus of N–P testing lends itself. The severity goal not only requires that:

(a) *'Accept/Reject' be interpreted inferentially*, as evidence of the presence or absence of departures, (appropriate to the testing context or question),

[4] Condition (**S-2**) can equivalently be written: with very low probability, test T would have produced a result that accords with H as well as (or better than) $\boldsymbol{x}_0$ does, if H were false (and a given discrepancy were present).

it requires as well that:

(b) *the test statistic $t(\boldsymbol{X})$ defines an appropriate measure of accordance or distance* (as required by severity condition **S-1**).[5]

To emphasize (b), we will use $d(\mathbf{X})$ for the test statistic; see Pearson ([1947], p. 143). Thirdly,

(c) the severity evaluation must be sensitive to the *particular outcome* $\boldsymbol{x}_0$; it must be *a post-data assessment*.

Moreover, the guide for evaluating, and possibly adjusting, error probabilities (e.g., in multiple hypothesis testing, in data mining) is whether the probativeness is altered with respect to the particular error of interest; see Mayo and Cox ([2006]).

An informal example may serve to capture the distinction between the behavioristic and severity rationales that we will be developing: Suppose a student has scored very high on a challenging test—that is, she earns a score that accords well with a student who has mastered the material. Suppose further that it would be extraordinary for a student who had not mastered most of the material to have scored as high, or higher than, she did. What warrants inferring that this score is good evidence that she has mastered most of the material? The behavioristic rationale would be that to always infer a student's mastery of the material just when they scored this high, or higher, would rarely be wrong in the long run. The severity rationale, by contrast, would be that this inference is warranted because of what the high score indicates about *this* student—mastery of the material.

From the severe-testing perspective, error probabilities have a crucial role to play in obtaining good test procedures (*pre-data*), and once the data $\boldsymbol{x}_0$ are in (*post-data*), they enable us to evaluate the probativeness or *severity* with which given hypotheses pass tests with $\boldsymbol{x}_0$. The severity criterion, we will argue, gives guidance as to what we should look for in scrutinizing N–P tests and inferences based on them; in so doing, the cluster of challenges underlying (i)–(iii) may be answered. Having the necessary impact on the controversy as it is played out in practice, however, demands not merely laying out a general principle of inference, but showing how it may be implemented. That is the goal of our discussion. We limit ourselves here to familiar classes of hypotheses tests, though our points may be extended to many classes of tests. See for example Spanos ([2006]).

[5] In an appropriate distance measure between H and $\boldsymbol{x}$, the larger $d(\boldsymbol{x})$ is the more indicative of discrepancy from H.

2 Error statistical tests from the severity perspective

Although the severity perspective directs the above restrictions/reintepretations, we retain several distinguishing features offered by the N–P (error-statistical) paradigm. To begin with, in error-statistical testing, one is asking a question about the *data generating mechanism*, framed in terms of a statistical hypothesis H. H cannot merely be an event; rather, H must assign a probability to each possible outcome $\boldsymbol{x}$, i.e., it gives the 'probability of $\boldsymbol{x}$ *under* H', abbreviated as $P(\boldsymbol{x};H)$. This notation helps also to avoid confusion with conditional probabilities in Bayes's theorem, $P(\boldsymbol{x}|H)$, where H is treated as a random variable with its own prior probabilities.[6]

The hypothesis testing question is put in terms of a null (or test) hypothesis H_0, and alternative H_1, the union of which exhausts the parameter space of the a *statistical model* which can be represented as a pair $(\mathcal{X}, \Theta)$; where $\mathcal{X}$ denotes the set of all possible values of the *sample* $\boldsymbol{X} = (X_1, \ldots, X_n)$—a set of *random variables*—one such value being the data $\boldsymbol{x}_0 = (x_1, \ldots, x_n)$, and Θ denotes the set of all possible values of the unknown *parameter(s)* θ. In hypothesis testing Θ is used as a shorthand for the family of densities indexed by θ, i.e. $\Theta := \{f(\mathbf{x}; \theta), \theta \in \Theta\}$, and the *generic form* of *null* and *alternative* hypotheses is:

$$H_0\colon \theta \in \Theta_0 \text{ vs. } H_1 : \theta \in \Theta_1, \quad \text{where } (\Theta_0, \Theta_1) \text{ constitutes a partition of } \Theta.^{7}$$

There is a test statistic $d(\boldsymbol{X})$ reflecting the distance from H_0 in the direction of H_1, such that the distribution of $d(\boldsymbol{X})$, its *sampling distribution*, evaluated under H_0, involves no unknown parameters. Because error probabilities concern the distribution of $d(\boldsymbol{X})$, evaluated under both the null and alternative hypotheses, interpreting a given result involves considering not just the observed value, $d(\boldsymbol{x}_0)$, but other possible values in $\mathcal{X}$ that could have occurred.[8]

For simplicity, we limit our focus to examples with a single unknown parameter μ, but our results apply to any hypothesis testing situation that can be viewed as a special case of the above generic form; see Spanos ([1999], ch. 14).

2.1 N–P Test $T(\alpha)$: type I, II error probabilities and power

Example. Consider a sample $\boldsymbol{X} = (X_1, \ldots, X_n)$ of size n, where each X_i is assumed to be Normal $(N(\mu, \sigma^2))$, Independent and Identically Distributed

[6] Freedman ([1995]) employs P (. ||H) to avoid the same confusion, but ';' is more familiar.

[7] Note that this precludes the artificial point against point hypothesis test that is so often the basis of criticisms (Hacking, [1965]; Royall, [1997]).

[8] By contrast, *posterior probabilities* are evaluated conditional on the particular observed value of $\mathbf{X}$, say $\boldsymbol{x}_0$. Other values that could have resulted but did not are irrelevant once $\boldsymbol{x}_0$ is in hand.

(NIID), with the standard deviation σ known, say $\sigma = 2$:

$$M\colon X_i \sim \text{NIID}\left(\mu, \sigma^2\right), \quad \text{where } -\infty < \mu < \infty,\ i = 1, 2, \ldots, n.$$

To keep the focus on the main logic, we assume that the *null and alternative hypotheses* of interest will concern the mean μ:

$$H_0\colon \mu \leqslant \mu_0 \text{ vs. } H_1\colon \mu > \mu_0.$$

The relevant test statistic is: $d(\mathbf{X}) = \left(\overline{X} - \mu_0\right)/\sigma_x$, where $\overline{X}$ is the sample mean with standard deviation $\sigma_x = (\sigma/\sqrt{n})$. Under the null, $d(\mathbf{X})$ is distributed as standard Normal, denoted by $d(\mathbf{X}) \sim N(0, 1)$.

The test is specified so that the probability of a Type I error, α, is fixed at some small number, such as 0.05 or 0.01, the *significance level* of the test:

$$\text{Type I error probability} = P(d(\mathbf{X}) > c_\alpha;\ H_0) \leq \alpha,$$

where $C_1(\alpha) = \{\boldsymbol{x}\colon d(\boldsymbol{x}) > c_\alpha\}$ denotes the *rejection region*.[9] Let $T(\alpha)$ denote the test defined by $d(\mathbf{X})$ and $C_1(\alpha)$. Having fixed the *type I error*, as the 'more important' of the two, N–P test principles then seek out the test that at the same time has a small probability of committing a *type II error* β. Since the alternative hypothesis H_1, as is typical, contains more than a single value of the parameter, it is *composite*, we abbreviate by $\beta(\mu_1)$ the type II error probability corresponding to $\mu = \mu_1$, for μ_1 values greater than μ_0, i.e., in the alternative region:

$$\begin{aligned}\text{Type II error probability (at } \mu_1) &= P(d(\mathbf{X}) \leq c_\alpha;\ \mu_1)\\ &= \beta(\mu_1), \text{ for any } \mu_1 > \mu_0.\end{aligned}$$

The 'optimality' of a N–P test of significance level α, is specified primarily in terms of minimizing $\beta(\mu_1)$ for all $\mu_1 > \mu_0$, or equivalently, maximizing the **power** (see Lehmann [1986]):

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

The above components define a N–P test $T(\alpha)$ with *significance level* α which rejects H_0 with data $\boldsymbol{x}_0$ if and only if $d(\boldsymbol{x}_0)$ is greater than c_α.

$$\text{Test } T(\alpha)\colon \text{ if } d(\boldsymbol{x}_0) > c_\alpha, \text{Reject } H_0;\ \text{if } d(\boldsymbol{x}_0) \leq c_\alpha, \text{ Accept } H_0.$$

Numerical Example. Let $\mu_0 = 12, \alpha = .025, n = 100, \sigma = 2\,(\sigma_x = 0.2)$. The test rule associated with test $T(\alpha)$ is: Reject H_0 iff $d(\boldsymbol{x}_0) > 1.96$, i.e., whenever $\bar{x} > 12.4$.

[9] A *sufficient condition* for an appropriate rejection region is that for any two significance levels α_1 and α_2 such that $0 \leqslant \alpha_1 < \alpha_2 \leqslant 1$, $C_1(\alpha_1)$ is a subset of $C_1(\alpha_2)$. This goes hand in hand with specifying a test statistic that provides an appropriate 'distance measure' as severity requires; see note 5.

Note that while test $T(\alpha)$ describes a familiar 'one-sided' test, our discussion easily extends to the case where one is interested in 'two-sided' departures: One simply combines two tests, 'one to examine the possibility that $\mu_1 > \mu_0$, the other for $\mu_1 < \mu_0$' (Cox and Hinkley [1974], p. 106, replaced θ with μ). In this case, the α level two-sided test combines both one-sided tests, each with significance level 0.5α.

2.2 Specifying Test $T(\alpha)$ using p-values

An alternative, but equivalent, way to specify the N–P test is in terms of the observed significance level or the p-value (Fisher [1935]), defined as 'the probability of a difference larger than $d(\boldsymbol{x}_0)$, under the assumption that H_0 is true,' i.e.

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

$$\text{Test } T(\alpha)\text{: if } p(\mathbf{x}_0) \leq \alpha, \text{ reject } H_0; \text{ if } p(\mathbf{x}_0) > \alpha, \text{ accept } H_0.$$

Fisherian significance tests differ from N–P tests primarily in so far as the alternative hypothesis H_1, in the N–P sense, is absent. As a result, the error probabilities are confined to those evaluated under the null, and thus, in contrast to the N–P paradigm, there is no notion of 'optimality' (based on power), associated with the choice of a test.

Fisher ([1935], [1956]), eschewed the N–P behavioristic model (which he regarded as a distortion of *his* significance tests on which it was built), preferring to report the observed p-value: if small (e.g., 0.05 or 0.01) the null hypothesis would be rejected at that level. Some prefer to simply report the p-value as a degree of inconsistency between $\boldsymbol{x}_0$ and H_0 (the 'pure significance test'): the smaller the p-value, the more inconsistent (Cox [1958]). Even N–P practitioners often prefer to report the observed p-value rather than merely whether the predesignated cut-off for rejection, c_α, has been reached, because it 'enables others to reach a verdict based on the significance level of their choice' (Lehmann, [1986], p. 70). Problems arise when p-values are interpreted illegitimately as degrees of probability of H_0 (an inconsistent hybrid). For example, a difference that is significant at level .01 does *not* mean we assign the null hypothesis a probability .01. Nevertheless, p-value reports have many inadequacies. To report p-values alone is widely disparaged as failing to assess the discrepancy or 'effect size' indicated (see Rosenthal [1994]; Thompson [1996]); nowadays it is often required (e.g., in psychology journals) that effect-size measures accompany p-values.

The severity conception retains aspects of, and also differs from, both Fisherian and Neyman–Pearsonian accounts, as traditionally understood. We wish to retain the post-data aspect of p-values—indeed extend it to other post-data error probabilities—but without forfeiting the advantages

offered by explicitly considering alternatives from the null hypothesis. A severity analysis allows both the data dependency of (post-data) error probabilities as well as an inferential report of the 'discrepancy' from the null that is warranted by data $\boldsymbol{x}_0$. It is a 'hybrid' of sorts, but it grows from a consistent inferential philosophy.[10]

3 Neyman's post-data use of power

Given the pre-data emphasis of the better known formulations of N–P theory, it is of interest to discover that, in discussing 'practice', Neyman, at times, calls attention to 'the use of the concept of the power of a test in *three* important phases of scientific research: (i) choice of a statistical test, (ii) design of an experiment, and (iii) interpretation of results' (Neyman [1957b], p. 10). Phases (i) and (ii) are pre-data. In particular, by designing $T(\alpha)$ so that POW(T(α); μ_1) = high, a tester ensures, ahead of time, that there is a high probability that the test would detect a discrepancy γ if it existed, for $\mu_1 = \mu_0 + \gamma$. That a test 'detects a discrepancy' means it rejects H_0 or reports a statistically significant (α-level) departure from H_0—perhaps a better term is that it 'signals' a discrepancy (we do not know it would do so correctly).

Phase (iii), however, is post-data. In phase (iii), 'the numerical values of probabilities of errors of the second kind are most useful for deciding whether or not the failure of a test to reject a given hypothesis could be interpreted as any sort of 'confirmation' of this hypothesis' (Neyman [1956], p. 290). To glean how Neyman intends power to be used in phase (iii), it is interesting to turn to remarks he directs at Carnap, and at Fisher, respectively.[11]

3.1 Neyman: does failure to reject *H* warrant confirming *H*?

Addressing Carnap, 'In some sections of scientific literature the prevailing attitude is to consider that once a test, deemed to be reliable, fails to reject the hypothesis tested, then this means that the hypothesis is 'confirmed' (Neyman [1955]). Calling this 'a little rash' and 'dangerous', he claims 'a more cautious attitude would be to form one's intuitive opinion only after studying the power function of the test applied' (p. 41).

[10] It is a mistake to regard the introduction of the alternative hypothesis, and with it, the notion of power, as entailing the behavioristic model of tests. While, in responding to Fisher ([1955]) distanced himself from the behavioristic construal, he described the introduction of alternative hypotheses as a 'Pearson heresy', whose aim was to put the choice of test under sounder footing. See Mayo ([1992], [1996]).

[11] To our knowledge, Neyman discusses post-data power in just the three articles cited here. Other non-behavioral signs may be found also in Neyman ([1976]) wherein he equates 'deciding' with 'concluding' and declares that his 'preferred substitute for 'do not reject H' is 'no evidence against *H* is found', both of which, being 'cumbersome' are abbreviated with 'accept *H*'. This last point is not unusual for Neyman.

> [If] the chance of detecting the presence [of discrepancy from the null], . . . is extremely slim, even if [the discrepancy is present] . . . , the failure of the test to reject H_0 cannot be reasonably considered as anything like a confirmation of H_0. The situation would have been radically different if the power function [corresponding to a discrepancy of interest] were, for example, greater than 0.95.[12] (ibid., p. 41)

Although *in theory*, once the N–P test is set up, the test is on 'automatic pilot'—H_0 is accepted or rejected according to whether $d(\boldsymbol{x}_0) > c_\alpha$—*in practice*, even behaviorist Neyman betrays a more nuanced post-data appraisal.

In an ironic retort, Neyman ([1957a]) criticizes Fisher's move from a large p-value to confirming the null hypothesis as 'much too automatic [because] . . . large values of p may be obtained when the hypothesis tested is false to an important degree. Thus, . . . it is advisable to investigate . . . what is the probability (of error of the second kind) of obtaining a large value of p in cases when the [null is false to a specified degree]' (p. 13, replaced P with p)—that is, the power of the test. *Note*: a large value of p leads to 'accept H_0', or to reporting a *non-statistically significant* difference. Furthermore, Neyman regards the post-data reasoning based on power as precisely analogous to the construal of rejection:

> [If] the probability of detecting an appreciable error in the hypothesis tested was large, say .95 or greater, *then and only then* is the decision in favour of the hypothesis tested justifiable in the same sense as the decision against this hypothesis is justifiable when an appropriate test rejects it at a chosen level of significance (Neyman [1957b], pp. 16–7).

Since he is alluding to Fisher, he combines notions from Fisherian and N–P tests in a general principle underlying the post-data use of power for interpreting 'Accept H_0', i.e., a non-significant difference $d(\boldsymbol{x}_0)$:

> (3.1) If data $d(\boldsymbol{x}_0)$ are not statistically significantly different from H_0—i.e., p is not small—and the power to detect discrepancy γ is high (low), then $d(\boldsymbol{x}_0)$ is (not) good evidence that the actual discrepancy is less than γ.

Admittedly, no such testing principle is to be found in the standard theoretical expositions of N–P testing theory.[13] The emphasis on the

[12] There are obvious similarities to the Popperian demand that hypotheses be highly corroborated. Neyman's recommendation would seem to offer a way to obtain a positive upshot from the falsificationist goals that Gillies ([1973]) looks to significance tests to provide. Mere failures to reject H_0 should not count as Popperian corroboration for H_0, but an assertion such as our 'the departure from H_0 is no greater than γ'.

[13] Early proponents of essentially the principle in (3.1) may be found in Bailey ([1971]), Cohen ([1988]), Gibbons and Pratt ([1975]), Mayo ([1983]). More recently, it arises in those wishing to reform significance tests (e.g., in psychology—see references in note 2) by supplementing them with 'effect size' measures, unaware that the seeds are already in Neyman ([1955], [1957a], [1957b]).

predesignation of tests may even seem to discourage such a post-data use of error probabilities.

Note how the stipulations in (3.1) adhere to severity requirements (**S-1**) and (**S-2**). The inference being considered for scrutiny is:

$$H\text{: 'the discrepancy (from } \mu_0\text{) is less than } \gamma\text{'}$$

which, notice, differs from H_0, unless $\gamma = 0$. The statistically insignificant result 'agrees with' H, so we have (**S-1**), and from the high power, we satisfy (**S-2**): that is, with very high probability, test T would have produced a result that accords *less* well with H than $\boldsymbol{x}_0$ does, were H false (were the discrepancy from μ_0 to exceed γ). Note that to 'accord *less* well with H' means, in this context, obtain a smaller p-value than observed. Nevertheless, severity calls for replacing the coarse assessment based on power with a data-dependent analysis.

4 Severe testing as a basic concept for an adequate post-data inference

The post-data use of power in (3.1) retains an unacceptable *coarseness:* Power is always calculated relative to the cut-off point c_α for rejecting H_0. Consider test $T(\alpha)$ with particular numerical values: $\alpha = 0.025$, $n = 100$, $\sigma = 2$ ($\sigma_x = 0.2$).

$$H_0\text{: } \mu \leq 12 \text{ vs. } H_1\text{: } \mu > 12.$$

$$\text{Reject } H_0 \text{ iff } d(\boldsymbol{x}_0) > 1.96\text{, i.e., iff } \overline{x} \geq 12.4.$$

(Equivalently, Reject H_0 iff the p-value is less than 0.025.) Suppose, for illustration, $\gamma^* = 0.2$ is deemed *substantively important* ($\mu := \mu_0 + \gamma^* = 12.2$). To determine if 'it is a little rash' to take a non-significant result, say $d(\boldsymbol{x}_0) = -1.0$, as reasonable evidence that $\gamma < \gamma^*$ (i.e., an important discrepancy is absent), we are to calculate POW($T(\alpha)$, γ^*), which is only 0.169! But why treat all values of $d(\boldsymbol{x}_0)$ in the acceptance region the same?

What if we get 'lucky' and our outcome is very much smaller than the cut-off 1.96? Intuition suggests that $d(\boldsymbol{x}_0) = -1.0$ provides better evidence for $\gamma < \gamma^*$ than $d(\boldsymbol{x}_0) = 1.95$ does. The evaluation of POW($T(\alpha)$,.2), however, will be identical for both sample realizations. In fact, were μ as large as 12.2, there is a high probability of observing a larger difference than -1. In particular, $\mathrm{P}(d(\boldsymbol{X}) > -1.0;\ 12.2) = 0.977$. This suggests that, post-data, the relevant threshold is no longer the pre-designated c_α, but $d(\boldsymbol{x}_0)$. That is, rather than calculating:

$$\textbf{Power} \text{ at } \mu = 12.2\text{: } \mathrm{P}(d(\boldsymbol{X}) > c_\alpha;\ \mu = 12.2),$$

one should calculate what may be called,

$$\textbf{Attained} \text{ (or } \textit{actual}\text{) } \textbf{Power} : \mathrm{P}(d(\boldsymbol{X}) > d(\boldsymbol{x_0}); \mu = 12.2).$$

The *attained power* against alternative $\mu = 12.2$ gives the **severity** with which $\mu < 12.2$ passes test $T(\alpha)$ when H_0 is accepted.[14] Several numerical illustrations will be shown.

4.1 The severity interpretation of acceptance (SIA) for test $T(\alpha)$

Applying our general abbreviation we write 'the **severity** with which the claim $\mu \leq \mu_1$ passes test $T(\alpha)$, with data $\boldsymbol{x}_0$':

$$\mathrm{SEV}(T(\alpha),\ d(\boldsymbol{x}_0),\ \mu \leq \mu_1),$$

where $\mu_1 = (\mu_0 + \gamma)$, for some $\gamma \geq 0$. For notational simplicity, we suppress the arguments $(T(\alpha),\ d(\boldsymbol{x}_0))$ where there is no confusion, and use the abbreviation: $\mathrm{SEV}(\mu \leq \mu_1)$—but it must be kept in mind that we are talking here of test $T(\alpha)$. We obtain a principle analogous to 3.1:

> ***SIA*:** (a): If there is a very *high* probability that $d(\boldsymbol{x}_0)$ would have been larger than it is, were $\mu > \mu_1$, then $\mu \leq \mu_1$ passes the test with *high* severity, i.e. $\mathrm{SEV}(\mu \leq \mu_1)$ is high.
>
> (b): If there is a very *low* probability that $d(\boldsymbol{x}_0)$ would have been larger than it is, even if $\mu > \mu_1$, then $\mu \leq \mu_1$ passes with *low* severity, i.e. $\mathrm{SEV}(\mu \leq \mu_1)$ is low.

We are deliberately keeping things at a relatively informal level, to aid in clarity. The explicit formula for evaluating $\mathrm{SEV}(\mu \leq \mu_1)$ in the case of a statistically insignificant result ('Accept H_0'), in the context of test $T(\alpha)$ is:

$$SEV(\mu \leq \mu_1) = P(d(\boldsymbol{X}) > d(\boldsymbol{x_0}); \mu \leq \mu_1 \text{false}) = P(d(\boldsymbol{X}) > d(\boldsymbol{x_0}); \mu > \mu_1).^{15}$$

As in the case of power, severity is evaluated at a point $\mu_1 = (\mu_0 + \gamma)$, for some $\gamma \geq 0$; yet the above holds because for values $\mu > \mu_1$ the severity

[14] We are coining 'attained power' simply to connect it with the familiar idea, for the case of 'accept H_0'. To avoid confusion, we will drop the term once the general notion of severity is in place. In the case of 'reject H_0' severity is [1 − 'attained' power].

[15] The calculations are easily obtained by means of the Standard Normal Distribution table, using the area to the right of $[d(x_0) - (\mu_1 - \mu_0)/\sigma_x] = (\overline{x} - \mu_1)/\sigma_x$ since:

$$\mathrm{SEV}(\mu \leqslant \mu_1) = P(Z > [(\overline{x} - \mu_1)/\sigma_x]) \quad (\text{where } Z \sim N(0,1)).$$

To apply this to the above example, $\overline{x} = 11.8$, so that $z = (11.8 - 12.2)/.2 = -2.0$
Hence, $\mathrm{P}(Z > -2) = .977$, i.e. the standard Normal area to the right of -2.0 is .977.

increases, i.e.

$$\mathrm{SEV}(\mu \leq \mu_1) > P(d(\mathbf{X}) > d(\boldsymbol{x}_0); \mu = \mu_1).^{16}$$

That is, the power of the test against $\mu = \mu_1$ provides a *lower bound* for severity for the inference or claim $\mu \leq \mu_1$.

It is important to emphasize that we are *not* advocating changing the original null and alternative hypotheses of the given test $T(\alpha)$; rather we are using the severe testing concept to evaluate which inferences are warrented, in this case of the form $\mu \leqslant \mu_1$—the kind of scrutiny one might especially need, as Neyman puts it, 'when we are faced with ... interpreting the results of an experiment planned and performed by someone else' (Neyman [1957b], p. 15). It is a *meta-statistical* check on various inferences one might draw using $T(\alpha)$ with data $\boldsymbol{x}_0$.

4.2 The fallacy of acceptance (i.e., an insignificant difference): Ms Rosy

A 'fallacy of acceptance' is often of concern when H_0 expresses a desirable situation such as, 'there is a *zero* increased risk' of some sort, or 'a model assumption, e.g., independence, is satisfied', and an insignificant result is interpreted too readily as positive evidence of no increased risk, or no violation of the given assumption. The test, we might say, gives *too rosy* an interpretation of the result: it would very probably overlook risk increases, and violations of interest, respectively—even were these present.

Say test $T(\alpha)$ yields the statistically insignificant result $d(\boldsymbol{x}_0) = 1.5$, i.e. $\bar{x} = 12.3$, so the test outputs 'Accept H_0' since the cut-off for rejection was 12.4. Suppose *Ms. Rosy* makes the following assertion:

> 'We may infer that any discrepancy from 12 is absent or no greater than .1.'

That is, she infers $\mu \leq 12.1$. Imagine someone critically evaluating this result wished to ask: *How severely does* $\mu \leq 12.1$ pass with $\bar{x} = 12.3$ ($d(\boldsymbol{x}_0) = 1.5$)?

The answer is: $\mathrm{SEV}(\mu \leq 12.1) = P(d(\boldsymbol{X}) > 1.5; \mu > 12.1) = .16.^{17}$

Since so insignificant a result would occur 84% of the time even if a discrepancy of .1 from H_0 exists, we would *deny* that Ms. Rosy's interpretation

[16] This inequality brings out the relationship between severity and power since for $d(\boldsymbol{x}_0) < c_\alpha$: $\mathrm{POW}(T(\alpha), \mu = \mu_1) = P(d(\boldsymbol{X}) > c_\alpha; \mu = \mu_1) = P(Z > [c_\alpha - (\mu_1 - \mu_0)/\sigma_x])$, (where $Z \sim N(0,1)$).

[17] The actual evaluation of severity takes the form:

$$P(\overline{X} > 12.3; \mu = 12.1) = P(Z > 1) = 0.16 \quad (\text{where } Z \sim N(0,1)).$$

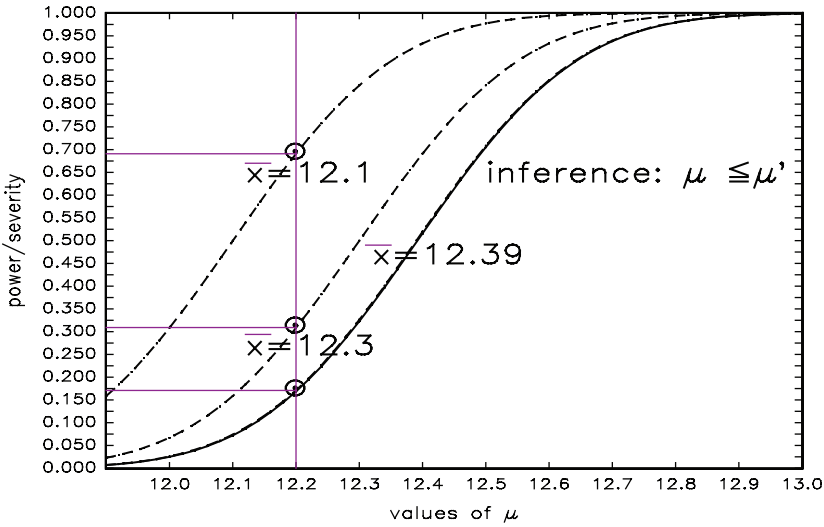

Figure 1. Case 'Accept H_0'—Power vs. Severity: the severity for $\mu \leq 12.2$ with different outcomes $\mathbf{x}_0$. Here $d(\mathbf{X}) = (\overline{X} - \mu_0)/\sigma_x$. Power curve is the solid line.
For $d(\mathbf{x}_0) = 1.95$ ($\bar{x} = 12.39$), SEV($\mu \leq 12.2$) = .171,
For $d(\mathbf{x}_0) = 1.50$ ($\bar{x} = 12.30$), SEV($\mu \leq 12.2$) = .309,
For $d(\mathbf{x}_0) = 0.50$ ($\bar{x} = 12.10$), SEV($\mu \leq 12.2$) = .691.

was warranted with severity. The general reasoning here is a straightforward application of SIA:

> If a test has a very low probability to detect the existence of a given discrepancy from μ_0, then such a negative result is poor evidence that so small a discrepancy is absent.

However, by dint of the same reasoning, we can find *some* discrepancy from H_0 that this statistically insignificant result warrants ruling out—one which very probably *would* have produced a more significant result than was observed. So even without identifying a discrepancy of importance ahead of time, the severity associated with various inferences can be evaluated. For example *the assertion that* $\mu \leq 13$ severely passes with $\overline{x} = 12.3$ ($d(\boldsymbol{x}_0) = 1.5$) since:

$$\text{SEV}(\mu \leq 13) = P(d(\boldsymbol{X}) > 1.5; \mu > 13) = 0.9997.$$

Risk-based policy controversies may often be resolved by such an assessment of negative results (Mayo [1991b]).

4.3 Severity and Power

To illustrate the evaluation of severity and its relationship to power, still keeping to the test output 'Accept H_0', consider Figure 1, showing the power

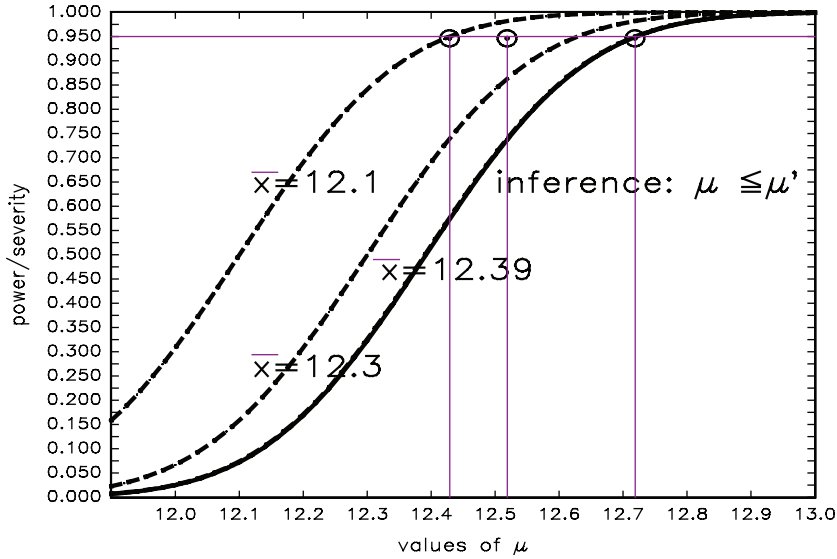

Figure 2. Case 'Accept H_0'—Power vs. Severity: the discrepancy excluded with severity .95 for $\mu \leq \mu_1$ corresponding to different outcomes $\mathbf{x}_0$. Power curve is the solid line.
For $d(\mathbf{x}_0) = 1.95$ ($\bar{x} = 12.39$), SEV($\mu \leq 12.72$) = .95,
For $d(\mathbf{x}_0) = 1.50$ ($\bar{x} = 12.30$), SEV($\mu \leq 12.63$) = .95,
For $d(\mathbf{x}_0) = 0.50$ ($\bar{x} = 12.10$), SEV($\mu \leq 12.43$) = .95.

curve (solid line), as well as the severity curves (dotted lines) corresponding to three different sample realizations $\bar{x} = 12.39$, $\bar{x} = 12.3$, $\bar{x} = 12.1$. In each case we can work with the sample mean $\bar{x}$ or the corresponding standardized distance statistic $d(\boldsymbol{x}_0)$.

In the case of $\bar{x} = 12.39$, where the observed result $d(\boldsymbol{x}_0) = 1.95$ is just inside the critical threshold $c_\alpha = 1.96$, the power curve provides a good approximation to the severity of inferring $\mu \leq \mu'$, where $\mu' = (\mu_0 + \gamma)$, for different values of the discrepancy γ. That is, the power evaluates the worst (i.e., lowest) severity values for *any* outcome that leads to 'Accept H_0' with test $T(\alpha)$. To illustrate reading the graph, the evaluations underneath Figure 1 compare the severity for inferring $\mu \leq 12.2$ for the three different samples.[18]

Figure 2 highlights a distinct use for the severity curves in Figure 1: one first chooses a high severity level, say 0.95, and then evaluates the corresponding discrepancy γ that is warranted at this pre-specified level. A handful of low and high benchmarks suffices for avoiding fallacies of acceptance.

[18] An Excel program, written by Geoff Cumming, can be used to evaluate such severity curves. This program is available at www.econ.vt.edu/spanos.

To summarize, when the null hypothesis is accepted, the goal is to be able to rule out as small a discrepancy γ from the null as possible. Restricting the analysis to power calculations allows evaluating severity for the case where $d(\boldsymbol{x}_0)$ just misses the critical threshold c_α—which, while useful, gives coarse severity assessments by treating all the results $d(\boldsymbol{x})$ below c_α the same. To avoid the 'too coarse' charge, we take account of the observed statistically insignificant result $d(\boldsymbol{x}_0)$, thereby enabling the post-data analysis to rule out values of μ even closer to μ_0.

5 Fallacy of rejection: statistical vs. substantive significance

Perhaps the most often heard, and best known, fallacy concerns taking a rejection of H_0 as evidence for a substantive claim: statistical significance is conflated with substantive significance. We need to distinguish two types of concerns.

5.1 Taking a rejection of H_0 as evidence for a substantive claim or theory

A familiar fallacy stems from reasoning that if a result is statistically significant, say at the 0.001 level, that 'one's substantive theory T, which entails the [statistical] alternative H_1, has received some sort of direct quantitative support of magnitude around .999' (Meehl [1970], p. 257). Not only does this fallaciously construe an error probability as a degree of confirmation in H_0, it erroneously conflates the statistical alternative with a substantive theory T. For example, finding a positive discrepancy from 12—which we may imagine is the mean concentration of lead in blood—would not warrant inferring a specific causal explanation. To rely on significance testing to corroborate a substantive scientific theory T, Meehl warns, is to subject T to only 'a feeble risk', and thereby violate Popperian requirements for science. In a similar vein, Imre Lakatos declares:

> After reading Meehl ([1967]) [and other psychologists] one wonders whether the function of statistical techniques in the social sciences is not primarily to provide a machinery for producing phony corroborations and thereby a semblance of 'scientific progress' where, in fact, there is nothing but an increase in pseudo-intellectual garbage. (Lakatos [1978], pp. 88–9)

The criticism here alludes to (Fisherian) significance tests. In contrast to the Fisherian tests, the N–P framework requires the null and alternative hypotheses to exhaust the parameter space of a (given) statistical model, thereby permitting only the statistical alternative H_1 to be inferred upon rejecting H_0, not a substantive theory T which might entail

H_1.[19] Even with its exhaustive space of hypotheses, fallacies of rejection can still enter the N–P paradigm, because finding a statistically significant effect, $d(\boldsymbol{x}_0) > c_\alpha$, need not be indicative of large or meaningful effect sizes.

5.2 A statistically significant difference from H_0 may fail to indicate a substantively important magnitude

In the case where $T(\alpha)$ has led to the rejection of the null hypothesis H_0 with data $\boldsymbol{x}_0$ the inference that 'passes' the test is of the form $\mu > \mu_1$, where $\mu_1 = (\mu_0 + \gamma)$, for some $\gamma \geq 0$. In other words, a statistically significant result indicates a departure from H_0 in the direction of the alternative, so severity condition (**S-1**) is satisfied: the alternative has 'survived' the test. Before we can infer, with severity, evidence of a particular positive departure, condition (**S-2**) demands we consider: How probable would so significant a result be if such a departure were absent?

Applying our general abbreviation, we write 'the severity with which test $T(\alpha)$ passes $\mu_1 > \mu_0$ with data $\boldsymbol{x}_0$' as: SEV($\mu > \mu_1$). It is evaluated by:

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

Because the assertions $\mu > \mu_1$ and $\mu \leq \mu_1$, constitute a partition of the parameter space of μ, there is a direct relationship, in test $T(\alpha)$, between the definitions of severity in the case of Accept and Reject H_0. That is,

$$\mathrm{SEV}(\mu > \mu_1) = 1 - \mathrm{SEV}(\mu \leq \mu_1).^{20}$$

As before, severity is evaluated at a point μ_1, because for any values of μ less than μ_1 the severity in test $T(\alpha)$ increases, i.e.

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

5.3 Principle for the severity interpretation of a rejection (SIR)

As with acceptances of H_0, an adequate post-data construal of 'Reject H_0' calls for a rule showing (a) the discrepancies that are well warranted, and (b) those which are not. The *severity interpretation for a rejection* of H_0, for test $T(\alpha)$ (i.e., $d(\mathbf{x}) > c_\alpha$) is this:

> ***SIR:*** (a) If there is a very *low* probability of so large a $d(\boldsymbol{x}_0)$, if $\mu \leq \mu_1$, then hypothesis $\mu > \mu_1$ passes with *high* severity, i.e.
>
> $$\mathrm{SEV}(\mu > \mu_1) \text{ is high.}$$

[19] True, the price for this is that using statistically inferred effects to learn about substantive theories demands linking, piece-meal, statistical inferences to subsequent ones, but this is a distinct issue to be dealt with separately (e.g., Mayo, [2002]).

[20] Note that to assert 'it is not the case that SEV(H) is high' does not entail that SEV(H) is low nor that SEV(not-H) is high. There may fail to be high severity for both H and for its denial. Articulating the full logic for SEV is a future project.

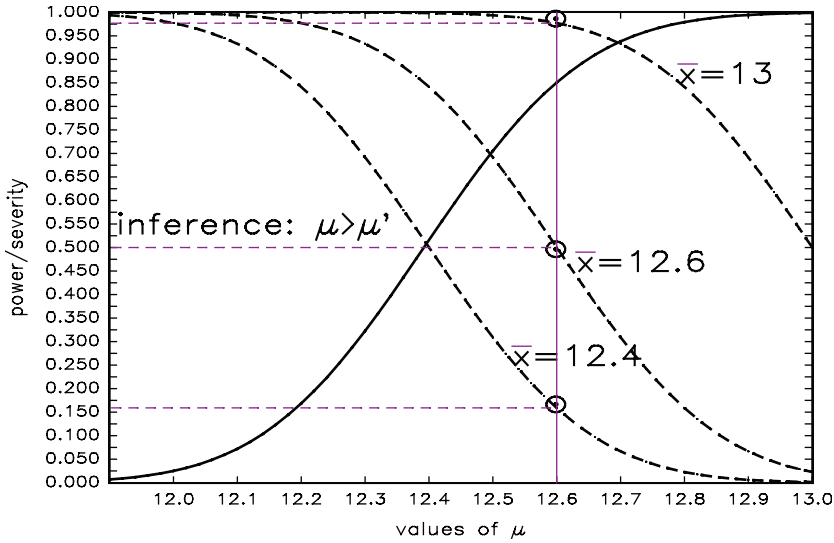

Figure 3. Case 'Reject H_0'—Power vs. Severity: the severity for $\mu > 12.2$ with different outcomes $\mathbf{x}_0$. Power curve is the solid line.

(b) If there is a very high probability of obtaining so large a $d(\boldsymbol{x}_0)$ (even) if $\mu \leq \mu_1$, then hypothesis $\mu > \mu_1$ passes with low severity, i.e.

$$\text{SEV}(\mu > \mu_1) \text{ is low.}$$

Choosing a small significance level α ensures that the inference: $\mu > \mu_0$, passes with high severity whenever we 'Reject H_0' with $d(\boldsymbol{x}_0)$.

It is instructive to observe the dramatic contrast between data-specific assessments of a rejection and the usual assessment of the power of test $T(\alpha)$ at the alternative $\mu_1 = \mu_0 + \gamma$ as in Figure 3. To illustrate how to read this graph, consider asking questions about the severity for different inferences.

A. First suppose that the outcome is $\bar{x} = 12.6$, (i.e., $d(\mathbf{x}_0) = 3.0$).

How severely does test $T(\alpha)$ pass $\mu_1 > 12.2$ with this result? The answer is .977, because: $\text{SEV}(\mu > 12.2) = P(d(\mathbf{X}) \leq 3.0; \mu_1 = 12.2) = .977$.

B. Now consider a different outcome, say, $\bar{x} = 13$, (i.e., $d(\mathbf{x}_0) = 5.0$).

How severely does test $T(\alpha)$ pass $\mu_1 > 12.2$ with this result? The answer is .9997, because: $\text{SEV}(\mu > 12.2) = P(d(\mathbf{X}) \leq 5.0; \mu_1 = 12.2) = .9997$.

Figure 3 also illustrates vividly the contrast between the relevant severity calculations (dotted curves) and power (solid line) in the case of 'reject H_0'. If $d(\boldsymbol{x})$ *has led to reject* H_0, $d(\boldsymbol{x}_0) > c_\alpha$, the severity for inferring $\mu > \mu_1$:

$$\text{SEV}(\mu > \mu_1) > 1 - \text{POW}(T(\alpha); \mu_1).$$

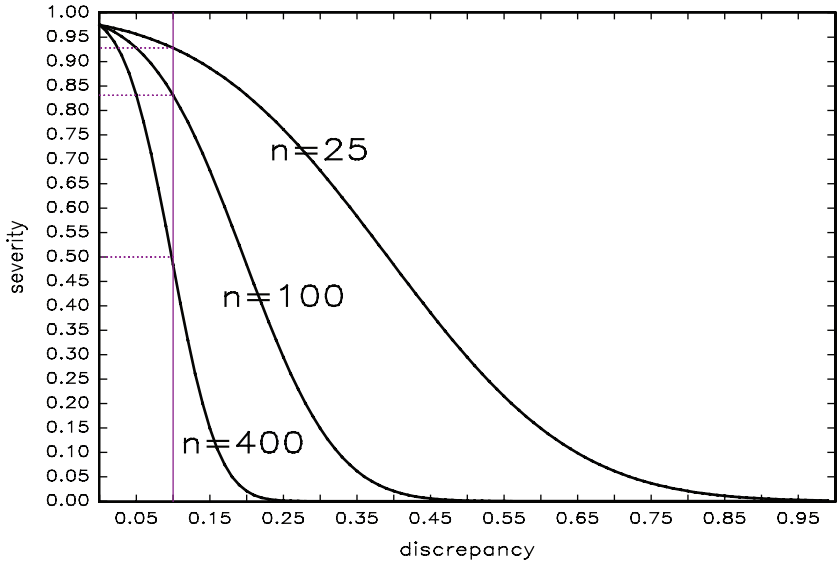

Figure 4. The severity associated with inferring $\mu > 12.1$ with the same $d(\mathbf{x}_0) = 1.96$, but different sample sizes n (the discrepancies are from 12):
for $n = 25$ and $d(\mathbf{x}_0) = 1.96$, SEV($\mu > 12.1$) = .933,
for $n = 100$ and $d(\mathbf{x}_0) = 1.96$, SEV($\mu > 12.1$) = .833,
for $n = 400$ and $d(\mathbf{x}_0) = 1.96$, SEV($\mu > 12.1$) = .500.

That is, one minus the power of the test at $\mu_1 = \mu_0 + \gamma$ provides a lower bound for the severity for inferring $\mu > \mu_1$. It follows that:

> The higher the power of the test to detect discrepancy γ, the *lower* the severity for inferring $\mu > \mu_1$ on the basis of a rejection of H_0.

This immediately avoids common fallacies wherein an α level rejection is taken as more evidence against the null, the higher the power of the test (see Section 5.4). The upshot is this: a statistically significant result with a small α level indicates, minimally, *some* discrepancy from μ_0 with high severity, 1–α; however, the larger the discrepancy one purports to have found, the *less* severely one's inference is warranted.

Notice that in the case of $\gamma = 0$, we are back to the prespecified alternative $\mu > \mu_0$; and thus, in this limiting case: SEV($\mu > \mu_0$) $> 1 - \alpha$ (Figure 4).

5.4 Comparing significant results with different sample sizes in $T(\alpha)$: large n problem

Whereas high power is desirable when evaluating a failure to reject H_0 with test $T(\alpha)$, in interpreting reject H_0, too high a power is the *problem*. An asset of the

severity requirement is that it gives a single criterion for properly interpreting both cases.[21]

Consider the common complaint that an α-significant result is indicative of different discrepancies when sample sizes differ, and that with large enough sample size, an α-significant rejection of H_0 can be very probable, even if the underlying discrepancy from μ_0 is substantively trivial. In fact, for any discrepancy γ, however small, a large enough sample size yields a high probability (as high as one likes) that the test will yield an α-significant rejection (for any α one wishes)—i.e.,

$$\mathrm{POW}(T(\alpha);\ \mu_1 = \mu_0 + \gamma) \text{ is high.}$$

N–P theory does not come with a warning about how the desideratum of high power can yield tests so sensitive that rejecting H_0 only warrants inferring the presence of a small discrepancy.

On the contrary, statistical significance at a given level is often (fallaciously) taken as more evidence against the null the larger the sample size (n).[22] In fact, it is indicative of *less* of a discrepancy from the null than if it resulted from a smaller sample size. Utilizing the severity assessment we see at once that an α-significant difference with n_1 passes $\mu > \mu_1$ *less* severely than with n_2 where $n_1 > n_2$.

5.5 General testing rules for $T(\alpha)$, using the severe testing concept

With reference to the one-sided test $T(\alpha)$, one might find it useful to define two severity rules for a metastatistical scrutiny of the N–P test outcomes: 'Accept H_0' and 'Reject H_0', corresponding to (SIA) and (SIR):

For Accept H_0:
If, with data $\mathbf{x}_0$, we accept H_0 (i.e. $d(\mathbf{x}_0) \leq c_\alpha$), then test $T(\alpha)$ passes:

(1) $\mu \leq \bar{x} + k_\varepsilon \sigma_x$ with severity $(1 - \varepsilon)$, for any $0 < \varepsilon < 1$, where $\mathrm{P}(d(\mathbf{X}) > k_\varepsilon) = \varepsilon$.[23]

[21] The 'large n problem' is also the basis for the 'Jeffreys-Good-Lindley' paradox brought out by Bayesians: even a highly statistically significant result can, as n is made sufficiently large, correspond to a high posterior probability accorded to a null hypothesis. (Good, [1983]; Edwards, Lindman, and Savage, [1963]; Lindley, [1957]). Some suggest adjusting the significance level as a function of n; the severity analysis, instead, assesses the discrepancy or 'effect size' that is, and is not, indicated by dint of the significant result.

[22] Rosenthal and Gaito ([1963]) document this fallacy among psychologists.

[23] Equivalently, rule (1) is: test $T(\alpha)$ passes $\mu \leq \mu_0 + \gamma$ with severity $(1 - \varepsilon)$, for $\gamma = (d(\mathbf{x}_0) + k_\varepsilon)\sigma_x$.

For Reject H_0:
If, with data $\mathbf{x}_0$, we reject H_0 (i.e. $d(\mathbf{x}_0) > c_\alpha$), then $T(\alpha)$ passes:
(2) $\mu > \bar{x} - k_\varepsilon \sigma_x$ with severity $(1 - \varepsilon)$, for any $0 < \varepsilon < 1$.[24]

Without setting a fixed level, one may apply the severity assessment at a number of benchmarks, to infer the extent of discrepancies that are and are not warranted by the particular dataset. In our conception of evidence, if an inference could only be said to pass a test with low severity, then there fails to be evidence for that inference (though the converse does not hold, see Note 20). A N–P tester may retain the usual test reports only supplemented by a statement of errors poorly probed. That is, knowing what is not warranted with severity becomes at least as important as knowing what is: it points to the direction of what may be tried next and of how to improve inquiries.

We emphasize that the data-specificity of the severity evaluation quantifies the extent of the discrepancy (γ) from the null that is (or is not) indicated by data $\boldsymbol{x}_0$, using the sampling distribution of the test statistic $d(\mathbf{X})$ on the basis of which all N–P error probabilities are derived. This reflects the fundamental difference between the current post-data inference and existing Bayesian accounts.

6 The severe testing concept and confidence intervals

A question that is likely to arise, especially in view of (1) and (2) in Section 5.5 is:

What is the correspondence between inferences severely passed and a Confidence Interval (CI) estimate?

Given the popularity of CI's in attempts to replace the dichotomous 'accept/reject' with a report indicating 'effect size', a brief foray into CI's seems needful.

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: $\mu = \overline{X} \pm e$.

Different sample realizations $\boldsymbol{x}$ lead to different estimates, but one can ensure, pre-data, that $(1 - \alpha)100\%$ of the time the true (fixed, but unknown) parameter value μ, whatever it may be, will be included in the interval formed. Although critics of N–P tests are at one in favoring CI's, it is important to realize that CI's are still squarely within the error-statistical paradigm. Moreover, they too are open to classic problems: they require predesignated assignments of a confidence level, and they are plagued with questions of interpretation.[25]

[24] Equivalently, rule (2) is: test $T(\alpha)$ passes $\mu > \mu_0 + \gamma$ with severity $(1 - \varepsilon)$, for $\gamma = (d(x_0) - k_\varepsilon)\sigma_x$.
[25] That is because one cannot assign the degree of confidence as a probability to the observed interval.

6.1 Dualities between one and two-sided intervals and tests

In fact there is a precise duality relationship between $(1 - \alpha)$ CI's and N–P tests: the CI contains the parameter values that would not be rejected *by the given test at the specified level of significance* (Neyman [1937]). It follows that the $(1 - \alpha)$ one-sided interval corresponding to test $T(\alpha)$ is:

$$\mu > (\overline{X} - c_\alpha \sigma_x).$$

In particular, the 97.5% CI estimator corresponding to test $T(\alpha)$ is:

$$\mu > (\overline{X} - 1.96\sigma_x).$$

Similarly, the 95% CI for μ corresponding to the two-sided test, $T(0.05)$ is:

$$(\overline{X} - 1.96\sigma_x) < \mu < (\overline{X} + 1.96\sigma_x).$$

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 $\bar{x}$. Once the estimate is formed, either the true 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.[26] Bayesian intervals introduce prior degrees of belief to get 'credibility intervals', introducing the problem of how to justify the prior from a frequentist, rather than from either a degree of belief or a priori standpoint.

6.2 Avoiding shortcomings of confidence intervals

Although CI's can be used in this way as surrogates for tests, the result is still too dichotomous to get around fallacies: it is still just a matter of whether a parameter value is inside the interval (in which case we accept it) or outside it (in which case we reject it). Consider how this is avoided by the severe testing concept.

The assertion:

$$\mu > (\bar{x} - c_\alpha \sigma_x)$$

is the observed one-sided $(1 - \alpha)$ interval corresponding to the test $T(\alpha)$, and indeed, for the particular value $\mu_1 = (\bar{x} - c_\alpha \sigma_x)$, the severity with which the inference $\mu \geq \mu_1$ passes is $(1 - \alpha)$. However, this form of inference is of interest only in the case of evaluating severity when $\boldsymbol{x}_0$ results in

[26] Although it is correct that $P\left(\left[\overline{X} - c_\alpha \sigma_x\right] < \mu\right) = (1 - \alpha)$, this probabilistic assertion no longer holds once we replace the random variable $\overline{X}$ with its observed value $\bar{x}$.

'Reject H_0'. In the case where $\boldsymbol{x}_0$ results in 'Accept H_0', the inference whose severity we wish to evaluate will rather be of the form:

$$\mu \leq (\bar{x} + c_\alpha \sigma_x).$$

Moreover, even in the case of 'Reject H_0', the CI will be importantly different from a severity assessment, although we can only discuss this here in part.

A $(1 - \alpha)$ CI, we said, corresponds to the set of null hypotheses that would not be rejected with an α-level test. But as we saw in discussing severity in the case of 'Accept H_0', the mere fact that $\boldsymbol{x}_0$ fails to reject a parameter value does not imply that $\boldsymbol{x}_0$ is evidence *for* that value. True, $\bar{x}$ is not sufficiently greater than any of the values in the CI to reject them at the α-level, but this does not imply $\bar{x}$ is good evidence for *each* of the values in the interval: many values in the interval pass test $T(\alpha)$ with very low severity with $\boldsymbol{x}_0$.

Recall the kind of question we employed severity to answer in interpreting a statistically significant result, say $d(\boldsymbol{x}_0) = 2.0$ (equivalently, $\bar{x} = 12.4$):

Does $\overline{x} = 12.4$ provide good evidence for $\mu > 12.5$?

The answer, one sees in Figure 3, is *No*, since the severity is only 0.309. However, the CI that would be formed using $d(\boldsymbol{x}_0)$ would be: $\mu > 12$. Since this interval includes 12.5, how can it be said to convey the 'No' answer, i.e., that the result *is poor* evidence for inferring $\mu > 12.5$? All values of the parameter in the CI are treated on a par, as it were. Nor does using the two-sided 95% CI cure this problem.[27] By contrast, for each value of μ_1 in the CI, there would be a different answer to the question: how severely does $\mu \geq \mu_1$ pass with $\boldsymbol{x}_0$? The CI estimation procedure sets out a fixed $(1 - \alpha)$; whereas, the severity analysis naturally leads to a sequence of inferences that are and are not warranted at different severity evaluation levels.

7 Beyond the N–P paradigm: pure significance, and misspecification tests

The concept of severe testing has been put forward elsewhere as a general account of evidence (Mayo [1996], [2004a], [2004b]); it is intended to hold in cases where severity is assessed entirely qualitatively, as in a familiar qualitative assessment of the difficulty of an exam, or quantitatively as in N–P tests—or in cases in between. Even in statistical testing, scrutinizing a N–P test from the severity perspective involves a use of background considerations (e.g., the particular error of interest as well as errors already ruled out in other studies) that is not purely formal; hence, our calling it 'meta-statistical'. Tests

[27] For $\overline{x} = 12.4$ the two-sided observed interval is $(12 < \mu < 12.8)$. Even allowing that one might entertain the inference, $12.5 < \mu < 12.8$, the CI procedure scarcely warns that evidence for this inference is poor. Note that using $\leq$ for these intervals makes no difference.

that have given rise to philosophical controversy will turn out, upon such a scrutiny, to serve poorly for the severity goal. This enables a severity scrutiny to provide a clear rationale for regarding as counterintuitive certain tests even if strictly licensed by N–P principles (e.g., certain mixed tests). Calling attention to the particular error in inference that needs to be probed before a claim is warranted with severity bears direct fruits for the knotty problems of determining which long-run is appropriate for the relevant context—a version of the philosopher's 'reference class problem' (Mayo and Kruse [2001]).

Conversely tests that do not include all the features of N–P tests may acquire a home in the severity paradigm. For example, even though a 'pure' (Fisher-type) significance test lacks an explicit alternative, it requires 'some idea of the type of departure from the null hypothesis which it is required to test' (Cox and Hinkley [1974], p. 65) which suffices to develop corresponding assessments of its ability to probe such departures (Mayo and Cox [2006]).

Consider the important category of tests to check the validity of statistical assumptions on which formal error probability assessments depend: checks for *model validation* or *misspecification tests*. Whereas N–P statistical tests take place *within* a specified (or assumed) model $\boldsymbol{M}$, when we put $\boldsymbol{M}$'s assumptions to the test, we probe *outside* $\boldsymbol{M}$, as it were; see Spanos ([1999]).

For example, in validating the model for test $T(\alpha)$, a misspecification test might have as its null hypothesis that the data constitute a realization of a random (IID) sample, and the alternative could cover all the ways these assumptions could fail. One can leave the alternative implicit in this manner, so long as unwarranted inferences are avoided. Rejecting the IID assumption may allow inferring, with severity, that the model is misspecified in *some* way or other, but it would not allow inferring, with severity, a particular alternative to IID (e.g. the presence of a particular type of dependency, say Markov). In this way, applying the severity criterion pinpoints a common fallacy in M-S testing (an instance of a fallacy of rejection of the type discussed in 5.2—see Mayo and Spanos [2004]). On the other hand, if a model's assumptions stand up to stringent probing for violations, the model may be accepted (with severity!) as adequate for the purposes of severely probing the original statistical hypotheses.

8. Concluding comments: have we shown severity to be a basic concept in a N–P philosophy of induction?

While practitioners do not see themselves as using N–P rules of behavior, the key concepts of that paradigm—type I and II errors, significance levels, power, confidence levels—are ubiquitous throughout statistical analysis.

Our goal, therefore, has been to trace out an *inferential interpretation* of tests which is consistent and in keeping with the *philosophy of error statistics*. Such an interpretation, we argued, would need to (i) avoid the coarseness of the strict model of N–P tests, wherein the same test output results regardless of where in the acceptance or rejection region $\boldsymbol{x}_0$ lies; (ii) prevent classic fallacies of acceptance and rejection, and (iii) answer the charge that it is too behavioristic and insufficiently inferential.

We have offered answers to these challenges that adhere to the key features that set tests from the error probability paradigm apart from alternative accounts: their ability to control and make use of error probabilities of tests. The key was to extend the *pre-data* error probabilities, significance level and power, to a 'customized', *post-data* assessment of the *severity* with which specific inferences pass the resulting test. A hypotheses H has severely passed a test to the extent that H would not have passed the test, or passed so well, were H false. The *data-specificity* of the severity evaluation quantifies the extent of the *discrepancy* (γ) from the null that is (or is not) indicated rather than quantifying a degree of confirmation accorded a given hypothesis. Since the interpretations are sensitive to the *actual outcome*, the limitations of just accepting or rejecting hypotheses are avoided. Fallacies of acceptance and rejection have also been explicitly dealt with.

Charge (iii) demands countering allegations that setting error probabilities are relevant only in contexts where we care about how often we can 'afford' to be wrong. From the severity perspective, the choice of the probabilities are no longer foreign to an inferential context. Pre-data, the choices for the type I and II errors reflect the goal of ensuring the test is capable of licensing given inferences severely. We set the 'worst case' values accordingly: small α ensures, minimally, that 'Reject H_0' licenses inferring *some* discrepancy from H_0; and high power against discrepancy γ ensures that failing to reject H_0 warrants $\mu < \mu_0 + \gamma$. So, while we favor reporting actual severity evaluations, even the predesignated N–P error probabilities attain a new, inferential justification.[28]

We identified some of Neyman's articles wherein he hints at such a post-data use of power; although they appear to be isolated cases. John Pratt lamented that: 'power plays virtually no role at the inference stage, and philosophies of inference in which it figures importantly are futuristic, to say the least.' (Pratt [1976], p. 781) We hope that the future is now.

[28] The challenge in (iii) may include others taken up elsewhere. Most notably, it has been charged the severity criterion conflicts with other post-data inferential criteria, e.g., likelihoods, posterior probabilities (Howson, [1995], [1997]; Achinstein [2003]). Our answer, in a nutshell, is this: the criterion leading to conflict differs from the severity criterion, and thus performs less well for the error statistician's goals (Mayo [1996], [2003b], [2005]; Mayo and Kruse [2001].)

Egon Pearson often emphasized that both he and Neyman regarded 'the ideal statistical procedure as one in which preliminary planning and subsequent interpretation were closely linked together—formed part of a single whole' (see Pearson [1962], p. 396). Critics fail to appreciate how crucial a role this entanglement plays in determining the capacity of the test procedure actually carried out. The post-data severity assessments are still based on error probabilities, but they are evaluated relative to the observed value of the test statistic. Admittedly, the N–P theory fails to articulate the principles by which to arrive at a N–P *philosophy of induction*. That is what the severe testing concept achieves. Viewing N–P tests from the severe testing perspective, we see that in scientific contexts the real value of being able to control error probabilities at small values is not the desire to have a good track record in the long run—although such a long-run justification is still available (and in several contexts may be perfectly apt). It is, rather, because of how this lets us *severely probe*, and thereby understand correctly, the process underlying the data now under consideration.

Deborah G. Mayo
Virginia Tech
Department of Philosophy
Blacksburg, VA 24061
USA
mayod@vt.edu
and
Aris Spanos
Virginia Tech
Department of Economics,
Blacksburg VA 2406
USA
aris@vt.edu

References

Achinstein, P. [2003]: *The Book of Evidence*, Oxford: Oxford University Press.

Armitage, P. [1961]: 'Contribution to the discussion in Smith, C.A.B., "Consistency in Statistical Inference and Decision"', *Journal of the Royal Statistical Society*, *B*, **23**, pp. 1–37.

Bailey, D. E. [1971]: *Probability and Statistics*, New York: Wiley.

Barnard, G. A. [1949]: 'Statistical Inference (with Discussion)', *Journal of the Royal Statistical Society*, *B*, **11**, pp. 115–139.

Barnard, G. A. [1971]: 'Scientific Inferences and Day to Day Decisions,' pp. 289-300 in V. Godambe and D. Sprott (eds.), Foundations of Statistical Inference: A Symposium, Holt, Rinehart and Winston of Canada, Toronto.

Berger, J. O. [2003]: 'Could Fisher, Jeffreys and Neyman Have Agreed?', *Statistical Science*, **18**, pp. 1–12.

Birnbaum, A. [1969]: 'Concepts of Statistical Evidence', in S. Morgenbesser, P. Suppes, and M. White (*eds.*), *Philosophy, Science, and Method: Essays in Honor of Ernest Nagel*, New York: St. Martin's Press, pp. 112–43.

Birnbaum, A. [1977]: 'The Neyman–Pearson Theory as Decision Theory, and as Inference Theory; with a Criticism of the Lindley-Savage Argument for Bayesian Theory', *Synthese*, **36**, pp. 19–49.

Carnap, R. [1962]: *Logical Foundations of Probability*, 2nd edn. The University of Chicago Press, Routledge & Kegan Paul, London.

Cohen, J. [1988]: *Statistical Power Analysis for the Behavioral Sciences*, 2nd edn., Hillsdale, NJ: Erlbaum.

Cox, D. R. [1958]: 'Some Problems Connected with Statistical Inference,' *Annals of Mathematical Statistics*, **29**, pp. 357–72.

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

Cox, D. R. [1997]: 'The Nature of Statistical Inference: Johann Bernoulli Lecture 1997', *Vierde Serie Deel*, **15**, pp. 233–42.

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

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

De Finetti, B. [1972]: *Probability, Induction and Statistics: The Art of Guessing*, New York: John Wiley & Sons.

Earman, J. [1992]: *Bayes or Bust? A Critical Examination of Bayesian Confirmation Theory*, Cambridge, MA: MIT Press.

Edwards, A. W. F. [1972]: *Likelihood: An Account of the Statistical Concept of Likelihood and its Application to Scientific Inference*, Cambridge: Cambridge University Press.

Edwards, W., Lindman, H. and Savage, L. [1963]: 'Bayesian Statistical Inference for Psychological Research', *Psychological Review*, **70**, pp. 193–242.

Efron, B. [1986]: 'Why Isn't Everyone a Bayesian?' *The American Statistician*, **40**, pp. 1–4.

Fetzer, J. H. [1981]: *Scientific Knowledge: Causation, Explanation, and Corroboration*, Dordrecht: D. Reidel.

Fisher, R. A. [1930]: 'Inverse Probability,' *Proceedings of the Cambridge Philosophical Society*, **26**, pp. 528–35.

Fisher, R. A. [1935]: 'The Logic of Inductive Inference', *Journal of the Royal Statistical Society*, **98**, pp. 39–54.

Fisher, R. A. [1955]: 'Statistical Methods and Scientific Induction,' *Journal of the Royal Statistical Society*, *B*, **17** pp. 69–78.

Fisher, R. A. [1956]: *Statistical methods and scientific inference*, Edinburgh: Oliver and Boyd.

Freedman, D. A. [1995]: 'Some Issues in the Foundations of Statistics,' *Foundations of Science*, **1**, pp. 19–83 (with discussion).

Gibbons, J. D. and Pratt, J. W. [1975]: 'P-values: Interpretation and Methodology,' *The American Statistician*, **29**, pp. 20–5.

Giere, R. N. [1976]: 'Empirical Probability, Objective Statistical Methods, and Scientific Inquiry', in W. L. Harper and C. A. Hooker (*eds.*), *Foundations of Probability Theory, Statistical Inference and Statistical Theories of Science*, Vol. II, Boston: D. Reidel, pp. 63–101.

Giere, R. N. [1984]: *Understanding Scientific Reasoning*, 2nd edn., New York: Holt, Rinehart and Winston.

Gigerenzer, G. [1993]: 'The superego, the ego, and the id in statistical reasoning,' in G. Keren and C. Lewis (*eds.*), *A Handbook of Data Analysis in the Behavioral Sciences: Methodological Issues*, Hillsdale, New Jersey: Lawrence Erlbaum Associates, pp. 311–39.

Gillies, D. A. [1973]: *An Objective Theory of Probability*, London: Methuem.

Glymour, C. [1980]: *Theory and Evidence*, Princeton: Princeton University Press.

Glymour, C., Scheines, R., Spirtes, P. and Kelly, K. [1987]: *Discovering Causal Structure: Artificial Intelligence, Philosophy of Science, and Statistical Modeling*, Orlando: Academic Press.

Godambe, V. and Sprott, D. (*eds.*) [1971]: *Foundations of Statistical Inference*, Toronto: Holt, Rinehart and Winston of Canada.

Good, I. J. [1983]: *Good Thinking*, Minneapolis: University of Minnesota Press.

Hacking, I. [1965]: *Logic of Statistical Inference*, Cambridge: Cambridge University Press.

Hacking, I. [1980]: 'The Theory of Probable Inference: Neyman, Peirce and Braithwaite', in D. H. Mellor (*ed.*), *Science, Belief and Behavior: Essays in Honour of R.B. Braithwaite*, Cambridge: Cambridge University Press, pp. 141–60.

Harlow, L. L., Mulaik, S. A. and Steiger, J. H. [1997]: *What if there were no Significance Tests?* Mahwah, NJ: Erlbaum.

Harper, W. L. and Hooker C. A. (*eds.*) [1976]: *Foundations of Probabiity Theory, Statisical Inference and Statistical Theories of Science*. Vol. 2, Dordrecht, The Netherlands: D. Reidel.

Horwich, P. [1982]: *Probability and Evidence*, Cambridge: Cambridge University Press.

Howson, C. [1995]: 'Theories of Probability,' *British Journal for the Philosophy of Science*, **46**, pp. 1–32.

Howson, C. [1997]: 'A Logic of Induction', *Philosophy of Science*, **64**, pp. 268–90.

Howson, C. and Urbach, P. [1989]: *Scientific Reasoning: The Bayesian Approach* 2nd edn, 1993, La Salle, IL: Open Court.

Jeffreys, H. [1939]: *Theory of Probability*, 3rd edn., 1961, Oxford: Oxford University Press.

Kass, R. E. and Wasserman, L. [1996]: 'Formal rules of selecting prior distributions: a review and annotated bibliography,' *Journal of the American Statistical Association,* **91**, pp. 1343–70.

Kempthorne, O. [1976]: 'Statistics and the Philosophers,' in W. L. Harper and C. A. Hooker (*eds.*), *Foundations of Probability Theory, Statistical Inference and Statistical Theories of Science*, Vol. II, Boston: D. Reidel, pp. 273–314.

Kempthorne, O. and Folks, L. [1971]: *Probability, Statistics, and Data Analysis*, Ames, IA: The Iowa State University Press.

Kiefer, J. [1977]: 'Conditional Confidence Statements and Confidence Estimators,' *Journal of the American Statistical Association*, **72**, pp. 789–827.

Kyburg, H. E. Jr. [1974]: *The Logical Foundations of Statistical Inference*, Dordrecht: Reidel.

Lakatos, I. [1978]: *The Methodology of Research Programmes*, edited by J. Worrall and G. Currie (*eds.*), vol. 1, *Philosophical Papers*, Cambridge: Cambridge University Press.

LeCam, L. [1977]: 'A Note on Metastatistics or "An Essay Toward Stating a Problem in the Doctrine of Chances" ', *Synthese*, **36**, pp. 133–60.

Lehmann, E. L. [1986]: *Testing Statistical Hypotheses*, 2nd edition, New York: Wiley:

Lehmann, E. L. [1993]: 'The Fisher and Neyman–Pearson Theories of Testing Hypotheses: One Theory or Two?' *Journal of the American Statistical Association*, **88**, pp. 1242–9.

Lehmann, E. L. [1995]: 'Neyman's Statistical Philosophy', *Probability and Mathematical Statistics*, **15**, pp. 29–36.

Levi, I. [1980]: *The Enterprise of Knowledge: An Essay on Knowledge, Credal Probability and Chance*, Cambridge, MA: MIT Press.

Lindley, D. V. [1957]: 'A Statistical Paradox', *Biometrika*, **44**, pp. 187–92.

Mayo, D. G. [1983]: 'An Objective Theory of Statistical Testing,' *Synthese*, **57**, pp. 297–340.

Mayo, D. G. [1985]: 'Behavioristic, Evidentialist, and Learning Models of Statistical Testing', *Philosophy of Science*, **52**, pp. 493–516.

Mayo, D. G. [1991a]: 'Novel Evidence and Severe Tests', *Philosophy of Science*, **58**, pp. 523–52.

Mayo, D. G. [1991b]: 'Sociological vs. Metascientific Theories of Risk Assessment', in D. Mayo and R. Hollander (*eds.*), *Acceptable Evidence: Science and Values in Risk Management*, New York: Oxford University Press, pp. 249–79.

Mayo, D. G. [1992]: 'Did Pearson Reject the Neyman–Pearson Philosophy of Statistics?', *Synthese*, **90**, pp. 233–62.

Mayo, D. G. [1996]: *Error and the Growth of Experimental Knowledge*, Chicago, IL: The University of Chicago Press.

Mayo, D. G. [2000]: 'Models of Error and the Limits of Experimental Testing', in M. Carrier, G. J. Massey and L. Reutsche (*eds.*), *Science at Century's End: Philosophical Questions on the Progress and Limits of Science*, Pittsburgh: University of Pittsburgh/University of Konstanz Press, pp. 317–44.

Mayo, D. G. [2002]: 'Theory Testing, Statistical Methodology, and the Growth of Experimental Knowledge', *Proceedings of the International Congress for Logic, Methodology, and Philosophy of Science*, Dordrecht: Kluwer.

Mayo, D. G. [2003a]: 'Severe Testing as a Guide for Inductive Learning,' in H. Kyburg (*ed.*), *Probability Is the Very Guide in Life*, Open Court, Chicago, pp. 89–117.

Mayo, D. G. [2003b]: 'Could Fisher, Jeffreys and Neyman Have Agreed? Commentary on J. Berger's Fisher Address', *Statistical Science*, **18**, pp. 19–24.

Mayo, D. G. [2004a]: 'An Error-Statistical Philosophy of Evidence,' in M. Taper and S. Lele (*eds.*), *The Nature of Scientific Evidence: Statistical, Philosophical, and Empirical Consideration*,' Chicago, IL: University of Chicago Press, pp. 79–97; 101–18.

Mayo, D. G. [2004b]: 'Evidence as Passing Severe Tests: Highly Probed vs. Highly Proved,' in P. Achinstein (*ed.*), *Scientific Evidence*, Baltimore, MD: Johns Hopkins University Press, pp. 95–127.

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

Mayo, D. G. [2006]: 'Critical Rationalism and Its Failure to Withstand Critical Scrutiny', in C. Cheyne and J. Worrall (*eds.*), *Rationality and Reality: Conversations with Alan Musgrave*, Kluwer series Studies in the History and Philosophy of Science, Dordrecht: Springer, pp. 63–96.

Mayo, D. G. and Cox, D. R. [2006]: 'Frequentist Statistics as a Theory of Inductive Inference,' forthcoming in The Second Erich L. Lehmann Symposium, Institute of Mathematical Statistics.

Mayo, D. G. and Kruse, M. [2001]: 'Principles of Inference and their Consequences', in D. Cornfield and J. Williamson (*eds.*), *Foundations of Bayesianism*, Dordrecht: Kluwer Academic Publishers, pp. 381–403.

Mayo, D. G. and Spanos, A. [2004]: 'Methodology in Practice: Statistical Misspecification Testing,' *Philosophy of Science*, **71**, pp. 1007–25.

Meehl, P. E. [1967/1970]: 'Theory-Testing in Psychology and Physics: A Methodological Paradox,' *Philosophy of Science*, **34**, pp. 103–15. Reprinted in Morrison, D. E. and Henkel, R. [1970].

Meehl, P. E. [1997]: 'The Problem is Epistemology, not Statistics: Replacing Tests by Confidence Intervals and Quantify Accuracy of Risk Numerical Predictions,' in Harlow, L. L., Mulaik, S. A. and Steiger, J. H. [1997].

Morrison, D. and Henkel, R. (*eds.*) [1970]: *The Significance Test Controversy*, Chicago, IL: Aldine.

MSERA [1998]: 'Special Issue: Statistical Significance Testing,' *Research in the Schools*, 5.

Musgrave, A. [1999]: *Essays in Realism and Rationalism*, Atlanta, GA: Rodopi.

Neyman, J. [1937]: 'Outline of a theory of statistical estimation based on the classical theory of probability,' *Philosophical Transactions of the Royal Society*, A, **236**, pp. 333–80.

Neyman, J. [1952]: *Lectures and conferences on mathematical statistics and probability*, 2nd edn., Washington, DC: U.S. Department of Agriculture.

Neyman, J. [1955]: 'The Problem of Inductive Inference,' *Communications on Pure and Applied Mathematics*, **VIII**, pp. 13–46.

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

Neyman, J. [1957a]: 'Inductive Behavior as a Basic Concept of Philosophy of Science,' *Revue d'Institute International de Statistique*, **25**, pp. 7–22.

Neyman, J. [1957b]: 'The Use of the Concept of Power in Agricultural Experimentation,' *Journal of the Indian Society of Agricultural Statistics*, **IX**, pp. 9–17.

Neyman, J. [1971]: 'Foundations of Behavioristic Statistics,' in V. Godambe and D. Sprott (*eds.*), *Foundations of Statistical Inference: A Symposium,* Toronto: Holt, Rinehart and Winston of Canada, pp. 1–13 (comments and reply, pp. 14–9).

Neyman, J. [1976]: 'Tests of Statistical Hypotheses and their use in Studies of Natural Phenomena,' *Communications in Statistics—Theory and Methods*, **5**, pp. 737–51.

Neyman, J. [1977]: 'Frequentist Probability and Frequentist Statistics,' *Synthese*, **36**, pp. 97–131.

Neyman, J. and Pearson, E. S. [1933]: 'On the problem of the most efficient tests of statistical hypotheses', *Philosophical Transactions of the Royal Society*, *A*, **231**, pp. 289–337. Reprinted in Neyman, J. and Pearson, E. S. [1966].

Neyman, J. and E. S. Pearson [1966]: *Joint Statistical Papers*, Berkeley, CA: University of California Press.

Pearson, E. S. [1947]: 'The Choice of Statistical Tests Illustrated on the Interpretation of Data Classed in a 2 × 2 Table,' *Biometrika*, **34**, pp. 139–67.

Pearson, E. S. [1950]: 'On Questions Raised by the Combination of Tests Based on Discontinuous Distributions,' *Biometrika*, **39** pp. 383–98. Reprinted in Pearson, E. S. [1966].

Pearson, E. S. [1955]: 'Statistical Concepts in Their Relation to Reality,' *Journal of the Royal Statistical Society, B*, **17**, pp. 204–7. Reprinted in Pearson, E. S. [1966].

Pearson, E. S. [1962]: 'Some Thoughts on Statistical Inference,' *Annals of Mathematical Statistics*, **33**: pp. 394–403. Reprinted in Pearson, E. S. [1966].

Pearson, E. S. [1966]: *The Selected Papers of E.S. Pearson*, Cambridge: Cambridge University Press.

Pierce, C. S. [1931–5]: *Collected Papers*, Vols. 1–6, Hartshorne and Weiss, P. (*eds.*), Cambridge: Harvard University Press.

Popper, K. [1959]: *The Logic of Scientific Discovery*, New York: Basic Books.

Popper, K. [1983]: *Realism and the aim of science*, Totowa, NJ: Rowman and Littlefield.

Popper, K. [1994]: 'The Myth of the Framework', in N. A. Notturno (*ed.*), *Defence of Science and Rationality*, London: Routledge.

Pratt, J. W. [1976]: 'A Discussion of the Question: for What Use are Tests of Hypotheses and Tests of Significance', *Communications in Statistics—Theory and Methods*, **5**, pp. 779–87.

Rosenkrantz, R. [1977]: *Inference, Method and Decision: Towards a Bayesian Philosophy of Science*, Dordrecht: Reidel.

Rosenthal, R. [1994]: 'Parametric Measures of Effect Sizes,' in H. M. Cooper and L. V. Hedges (*eds.*), *The Handbook of Research Synthesis*, Newbury, CA: Sage, pp. 231–44.

Rosenthal, R. and Gaito, J. [1963]: 'The Interpretation of Levels of Significance by Psychological Researchers,' *Journal of Psychology*, **55**, pp. 33–8.

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

Salmon, W. [1966]: *The Foundations of Scientific Inference*, Pittsburgh: University of Pittsburgh Press.

Savage, L. J. [1964]: 'The Foundations of Statistics Reconsidered', in H. E. Kyburg and H. E. Smokler (*eds.*), *Studies in Subjective Probability*, New York: John Wiley and Sons, pp. 173–88.

Seidenfeld, T. [1979]: *Philosophical Problems of Statistical Inference: Learning from R. A. Fisher*, Dordrecht: D. Reidel.

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

Spanos, A. [2006]: 'Revisiting the Omitted Variables Argument: Substantive vs. Statistical Adequacy,' forthcoming in the *Journal of Economic Methodology*.

Spielman, S. [1973]: 'A Refutation of the Neyman–Pearson Theory of Testing', *British Journal for the Philosophy of Science*, **24**, pp. 201–22.

Thompson, B. [1996]: 'AERA Editorial Policies regarding Statistical Significance Testing: Three suggested Reforms,' *Educational Researcher*, **25**, pp. 26–30.