Dantzig, G.B.: On the non-existence of tests of Student's hypothesis having power functions independent of σ. Ann. Math. Statist. **11**, 186–192 (1940)
Dimitrakakis, C., Ortner, R.: Sequential Sampling. In: Decision Making Under Uncertainty and Reinforcement Learning. Intelligent Systems Reference Library, vol 223. Springer, Cham (2022). https://doi.org/10.1007/978--3-031-07614-5_5
Ghosh, B.K., Sen, P.K.: Handbook of Sequential Analysis, edited volume. Marcel Dekker, New York (1991)
Ghosh, M., Mukhopadhyay, N.: On two fundamental problems of sequential estimation. Sankhya, Series B. **38**, 203–218 (1976)
Ghosh, M. Mukhopadhyay, N.: Consistency and asymptotic efficiency of two-stage and sequential procedures. Sankhya, Series A. **43**, 220–227 (1981)
Ghosh, M., Mukhopadhyay, N., Sen, P.K.: Sequential Estimation. Wiley, New York (1997)
Lai, T.L., Siegmund, D.: A nonlinear renewal theory with applications to sequential analysis I. Ann. Statist. **5**, 946–954 (1977)
Lai, T.L., Siegmund, D.: A nonlinear renewal theory with applications to sequential analysis II. Ann. Statist. **7**, 60–76 (1979)
Lehmann, E.L.: Notes on the Theory of Estimation. University of California Press: Berkeley (1951)
Mukhopadhyay, N., Datta, S., Chattopadhyay, S.: Applied Sequential Methodologies. edited volume. Marcel Dekker, New York (2004)
Mukhopadhyay, N., de Silva, B.M.: Sequential Methods and Their Applications. CRC Press, New York (2009)
Mukhopadhyay, N., Solanky, T.K.S.: Multistage Selection and Ranking Procedures: Second-Order Asymptotics. Marcel Dekker, New York (1994)
Neyman, J. Pearson, E.S.: On the problem of the most efficient tests of statistical hypotheses. Phil. Trans. Roy. Soc., Series A. **231**, 289–337 (1933)
Ray, W. D.: Sequential confidence intervals for the mean of a normal population with unknown variance. J. Roy. Statist. Soc., Series B. **19**, 133–143 (1957)
Robbins, H.: Sequential estimation of the mean of a normal Population. In: Probability and Statistics. (Harald Cramér Volume), Ed. U. Grenander, Almquist and Wiksell, Uppsala, pp. 235–245 (1959)
Stein, C.: A two sample test for a linear hypothesis whose power is independent of the variance. Ann. Math. Statist. **16**, 243–258 (1945)
Stein, C.: Some problems in sequential estimation. Econometrica,**17**, 77–78 (1949)
Wald, A.: Sequential Analysis. Wiley, New York (1947)
Wald, A., Wolfowitz, J.: Optimum character of the sequential probability ratio test. Ann. Math. Statist. **19**, 326–339 (1948)
Woodroofe, M.: Second order approximations for sequential point and interval estimation. Ann. Statist. **5**, 984–995 (1977)

Severe Testing

Deborah G. Mayo
Department of Philosophy, Virginia Tech, Blacksburg, VA, USA

Background

The severe testing perspective in statistics grows out of a general philosophy of error-prone reasoning. It is based on a minimal principle for regarding data $\boldsymbol{x}$ as evidence for a claim C: Data fail to provide evidence for a claim C, if little if anything has been done that would have uncovered flaws in C, even if they are present. In order for $\boldsymbol{x}$ to provide evidence for a claim C, C must have passed an analysis that would probably have found C flawed, just if it is. This probability is the stringency or severity with which it has passed the test. It represents the extent to which the claim has withstood scrutiny with given data. This perspective is developed in Mayo (1991, 1996, 2018), Mayo and Cox (2006), Mayo and Spanos (2006, 2011), and Mayo and Hand (2022).

Severe Testing Philosophy

The probability that a method erroneously interprets data is an error probability, and statistical methods based on error probabilities may be called *error statistics*. Severe testing is regarded as part of a "philosophy" because it is a general conception of the interpretation and justification for formal error statistical methods. As with all statistical philosophies, it goes beyond what is typically found in statistical textbooks, but it underlies the principles and uses of a cluster of statistical methods in science. Severity relates to error statistical methods because it is possible to use formal error probabilities (e.g., type 1 and 2 errors, P-values, and confidence levels) to evaluate how severely, or inseverely, probed various statistical claims are, given the data. Its function is twofold: (1) to provide a postdata evidential

interpretation of error statistical methods and (2) to understand and avoid fallacies surrounding these methods.

Much like other concepts introduced to set out philosophies of statistical inference—credibility, support, and confirmation—severe testing relies on general assumptions concerning the underlying statistical models, and the connections between statistical inquiries and substantive questions. In the case of error statistics, checking assumptions is part of the methodology under the concept of "auditing." Until the assumptions pass an audit, there is at most an *indication* of severity (Mayo 2018). Even acknowledging that the application of formal statistical tools, in isolation from background information, fails to yield inferences that align with attractive-sounding metrics, it is important to set out the overarching statistical philosophy that guides the interpretation of these methods. A central difference between error statistical and non-error statistical methods is the former's use of the sampling distribution postdata. That is the basis for error probabilities.

Severity Requirement and Severity Principle: From Philosophy to Statistics

Popper introduced a notion of severity to contrast his account of falsification with popular accounts of confirmation. For the latter, evidence $\boldsymbol{x}$ confirms hypothesis H to the extent that the probability of H given $\boldsymbol{x}$ exceeds the prior probability of H. While confirmation theories follow Bayes rule, the notion of probability most often used was "logical" probability, defined in terms of simple first-order languages (Jeffreys 1961; Carnap 1962). That a hypothesis is highly confirmed, according to Popper, tells us only that it accords with the data: H passes the test with $\boldsymbol{x}$ to some degree. It does not tell how stringently H has passed. For Popper, "the probability of a statement ... simply does not express an appraisal of the severity of the tests a theory has passed, or of the manner in which it has passed these tests" (1959, 394–395).

Severity Requirement: For data to warrant a claim C with severity requires not just that

(S-1) H agrees with the data (H passes the test), but also

(S-2) with high probability, H would not have passed the test so well, were H false.

Even claims that are known to be true or probable, in whatever sense one chooses, may have been poorly probed by the data $\boldsymbol{x}$ at hand.

Popper never applied his intuitive notion to statistical inference. Mayo's account of severe testing employs modern statistical methods to render Popper's falsificationist philosophy relevant to contemporary practice. Where Popper's work contrasted with confirmation theories in philosophy, severity in statistics contrasts error statistics with varieties of Bayesian approaches.

The severity function in statistics has three arguments. SEV(T, $\boldsymbol{x}_0$, C) is used to abbreviate: "The severity with which claim C passes test T with outcome $\boldsymbol{x}_0$." When the testing context is clear, it can be abbreviated as SEV(C). Claim C is not limited to the result of a formal test; it may be an estimate, prediction, or other inference.

Severity Principle: An error-prone claim C is warranted by data just to the extent it has been subjected to, and passes, a test that probably would have found flaws in C, if they are present.

This probability quantifies the stringency with which C has been probed by the test method. It does not attach to the claim inferred, but to the overall method; it is a *methodological probability*.[1] When a test's formal error probabilities succeed in quantifying the capacity of tests to probe errors in inferring C, they can be used to assess how well or poorly C is warranted. The weakest variant of the severity principle merely denies there is evidence for C if the method had little if any capacity to find C flawed. This critical role is the most important function of severity.

[1] A severity assessment may also be qualitative.

Severity in Statistical Significance Tests

The severity principle captures the underlying reasoning of statistical significance tests and addresses long-standing problems and fallacies associated with it. The severity interpretation combines aspects from Neyman-Pearson (N-P) and Fisherian tests, without being guilty of the charge that they form an inconsistent hybrid (Gigerenzer 1993, 2006; Cox and Hinkley 1974).

Elements of Statistical Significance Tests

A statistical hypothesis H is a claim about some aspect of the process that might have generated the data $\boldsymbol{x}$, viewed as observed values of a random variable $\boldsymbol{X}$. Data $\boldsymbol{x}$ are used to learn about the probability distribution of $\boldsymbol{X}$ by testing various statistical hypotheses, generally about a parameter in a model M—an idealized representation of the data generation. Neyman and Pearson (N-P) called the reference hypothesis the *test hypothesis* (1933), while Fisher called it the *null hypothesis* (1935), denoted by H_0. (See Cox 1958, 1977, 2006; Cox and Hinkley 1974.)

Statistical significance tests consist of: (a) a null (or test) hypothesis H_0, couched in terms of unknown parameter θ; and (b) a test statistic $d(\boldsymbol{X})$, which reflects how well or poorly the data $\boldsymbol{x}_0$ accord with the null hypothesis H_0. The larger the value of $d(\boldsymbol{x}_0)$, the more improbable the outcome is from what is expected under H_0, with respect to the particular question being asked. An important aspect of an error statistical test is its ability to ensure the sampling distribution of the test statistic can be computed under H_0 and, generally, also under hypotheses discrepant from H_0. This enables computing (c) the statistical significance level or P-value associated with $d(\boldsymbol{x}_0)$: the probability of a worse fit with H_0 than the observed $d(\boldsymbol{x}_0)$, under the assumption that H_0 is true:

$$P\text{-value} = \Pr(d(\boldsymbol{X}) > d(\boldsymbol{x}_0); H_0).$$

The larger the value of the test statistic, the smaller the P-value. If the P-value is very small (e.g., 0.05, 0.01, 0.005), the data accord with the denial of H_0, but there are two rationales that may be given: behavioristic and evidential.

Behavioristic Versus Evidential Justification

The behavioristic rationale, as emphasized in N-P tests, reflects the goal of ensuring a low probability of erroneous inferences in a series of applications. As Cox and Hinkley put it (1974, p. 66):

> Suppose that we were to accept the available data as evidence against H_0. Then we would be bound to accept all data with a larger value of [d] as even stronger evidence. Hence, p_{obs} [the observed P-value] is the probability that we would mistakenly declare there is to be evidence against H_0, were we to regard the data under analysis as just decisive against H_0.

An evidential or inferential rationale may be obtained by appealing to "calibration" (Cox 1958, 1977): "Just as with the use of measuring instruments . . . we employ the performance features to make inferences about aspects of the particular thing that is measured" (Mayo and Cox 2006, 84).

Example. Consider the case of a random sample $\boldsymbol{X}$ of size n from a Normal distribution with unknown mean μ and, for simplicity, known variance σ^2. In a one-sided test of the hypotheses:

$$H_0: \mu \leq \mu_0 \text{ versus } H_1: \mu > \mu_0,$$

the test statistic is $d(X) = \frac{(\bar{X}-\mu_0)}{\sigma_x}$, where $\bar{X} = \frac{1}{n}\sum_{k=1}^{n} X_k$ is the sample mean, and the standard error, $\text{SE} = \left(\sigma/\sqrt{n}\right)$.[2] In the severe testing interpretation, a very small P-value (e.g., $p(\boldsymbol{x}_0) = 0.01$) indicates evidence for H_1 because H_1 has passed a severe test, provided the P-value is properly computed. The severity definition is instantiated because:

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

[2]The same rejection region follows if H_0 is 0.

However, warranting the existence of *some* positive discrepancy from μ_0 is rarely adequate for interpreting the result. A key weakness of the *P*-value is that it does not by itself give an indication of magnitude.

To remedy this, severity reasoning introduces a discrepancy parameter γ and evaluates inferences of form: H_1: $\mu > \mu_1 = (\mu_0 + \gamma)$, $\gamma > 0$. For each of a series of discrepancies of interest, there is a report of those that are well warranted, and those poorly warranted on severity grounds. The basis for doing so is summarized in (a) and (b):

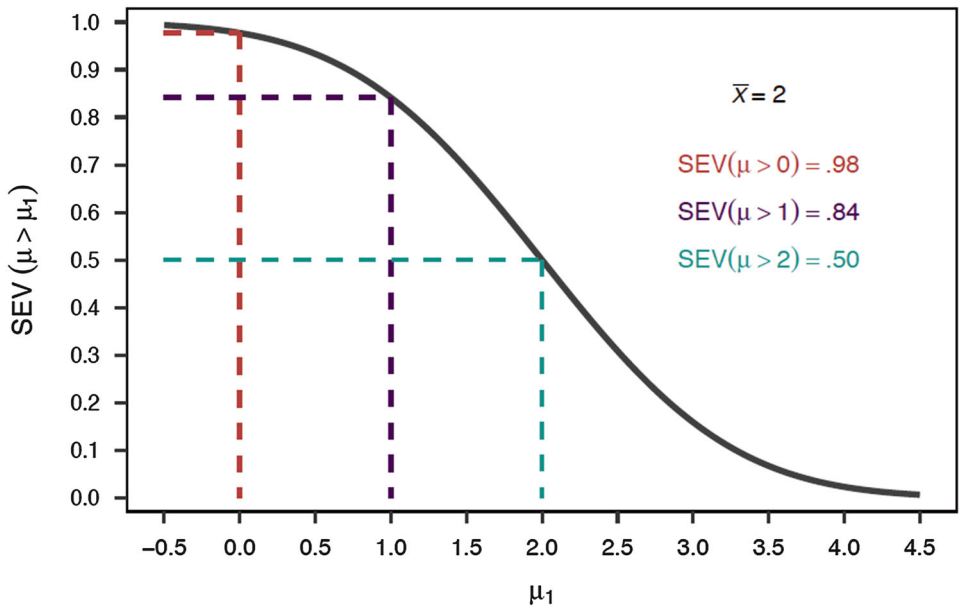
(a) If there is a very high probability (e.g., a probability > 0.95) that d would have been smaller than observed if $\mu \leq \mu_1$, then d is a good indication that $\mu > \mu_1$: SEV($\mu > \mu_1$) is high.

(b) If there is a fairly high probability that d would have been larger than observed (a probability > 0.5), even if μ is no greater than μ_1, then d is a poor indication that $\mu > \mu_1$, where $\mu_1 = \mu_0 + \gamma$: SEV($\mu > \mu_1$) is low.

The severe testing assessment is not changing the original null and alternative hypothesis, but interpreting the results in relation to the reference value in H_0. Alternatively, these stipulations might be construed as reporting the results of a series of tests of form: $\mu \leq \mu_1$ versus $\mu > \mu_1$, allowing μ_1 to vary over several values.

To have some numbers, consider H_0: $\mu \leq 0$ versus H_1: $\mu > 0$, and let the SE equal 1 [e.g., $\sigma = 10$ and $n = 100$, SE $= (\sigma/\sqrt{n})$]. The 2-SE cutoff giving a *P*-value of approximately 0.025 is 2. If d (which in this case is $\bar{x}$) is 2 or greater, the result is statistically significant at level 0.025. Suppose the observed d is 2. A useful benchmark for a poorly warranted discrepancy is a μ_1 greater than $\bar{x}$, e.g., $\bar{x} + 1$SE (3). The hypothesis $\mu > 3$ will be poorly warranted because the probability that d is even larger than 2, under the assumption that $\mu = 3$ is fairly high, 0.84. By reporting various benchmarks, tests can avoid magnitude errors in interpreting *P*-values (Fig. 1). (Numerical examples may be found in Spanos 2019.)

There are two points to note: First, inferences are to inequalities such as ($\mu > \mu_1$) and not to



Severe Testing, Fig. 1 Severity curve

points; and second, the fact that severity has the goal of testing has consequences for what counts as "high" or "low." A posterior probability to C of, say, 0.8 might be construed as fairly good evidence for C, but a severity of 0.8, corresponding to an error probability of 0.2, is poor evidence for C.

Statistically Insignificant Results

If the P-value is not small, for example, if it is greater than 0.1, there is poor evidence of a discrepancy from H_0, but it is important to avoid the classic fallacy of interpreting it as evidence *for* H_0. A P-value that is not small is generally said to be statistically insignificant.[3]

If d is statistically insignificant, the null hypothesis "passes" the test, i.e., condition (S-1) is satisfied, but the test might not have had much chance of detecting departures even if they existed. This is addressed by the second requirement for severity (S-2). We evaluate if it is good evidence for $\mu \leq \mu_1$, where $\mu_1 = \mu_0 + \gamma$, by evaluating the probability that test T+ would have produced a more statistically significant result than it did (i.e., $d(X) > d(x_0)$), if $\mu > \mu_1$:

$$\begin{aligned}&\text{SEV}(T+, x_0, \mu \leq \mu_1)\\ &= P(d(X) > d(x_0); \mu > \mu_1).\end{aligned}$$

Severity is computed at the point $\mu = \mu_1$ because $\text{SEV}(\mu \leq \mu_1)$ is even greater for values of μ less than μ_1.

Given the symmetry in this example, severity assessments for claims of form $(\mu \leq \mu_1)$ are a mirror image of those for $(\mu > \mu_1)$, provided assumptions hold. With failed assumptions, both a claim and its denial can fail to be warranted with severity.

[3]The same severity interpretation applies whether or not a tester is using a predesignated threshold for "significance." Thus, it holds regardless of which side is taken on the recent debate about using significance thresholds (Wasserstein et al. 2019).

Summary of Severity with Significant and Insignificant Results

(I) Interpreting a statistically significant result in testing H_0: $\mu \leq \mu_0$ against H_1: $\mu > \mu_0$:

$$\begin{aligned}&\text{SEV}(T+, x_0, \mu > \mu_1)\\ &= P(d(X) \leq d(x_0)\,; \mu \leq \mu_1)\\ &\text{where } \mu_1 = \mu_0 + \gamma, \gamma \geq 0.\end{aligned}$$

Compute this at the point $P(d(X) \leq d(x_0); \mu = \mu_1)$ because SEV is even greater for $\mu < \mu_1$.

(II) Interpreting a statistically insignificant result in testing H_0: $\mu \leq \mu_0$ against H_1: $\mu > \mu_0$:

$$\begin{aligned}&\text{SEV}(T+, x_0, \mu \leq \mu_1)\\ &= P(d(X) > d(x_0); \mu > \mu_1).\end{aligned}$$

Again, this is computed at the point $P(d(X) > d(x_0); \mu = \mu_1)$ because SEV is even greater for $\mu > \mu_1$.

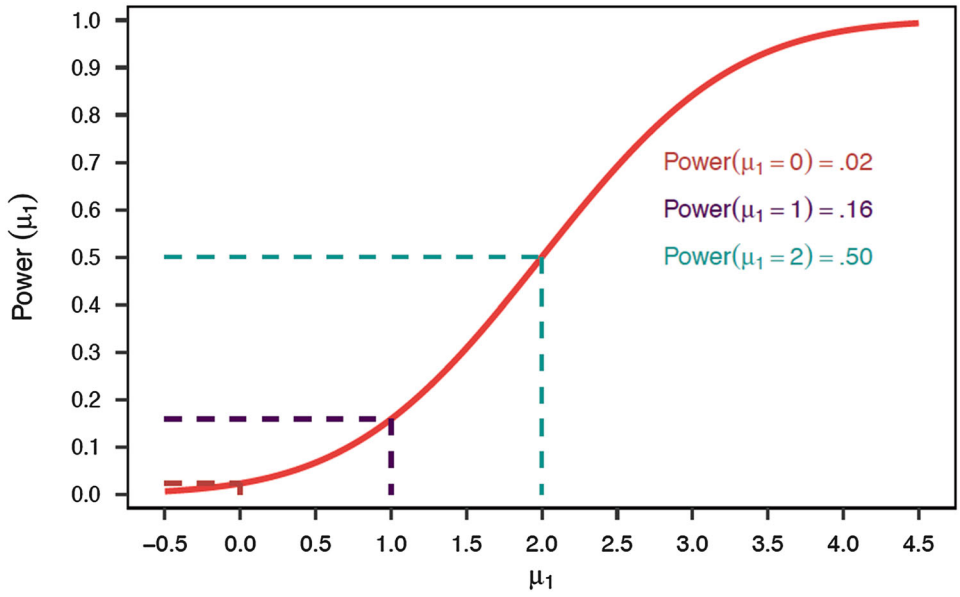
In the case of a small P-value, the concern is with interpreting it as indicating a larger effect than is actually warranted (making "mountains out of molehills"). With P-values that are not small, the concern is fallaciously inferring evidence of no discrepancy, or inferring one that is smaller than warranted. A severity assessment is designed to avoid both:

A *statistically significant* result (small P-value) licenses inferences of the form $\mu > [\mu_0 + \gamma]$, for some $\gamma \geq 0$, but with a warning that $\mu > [\mu_0 + \kappa]$ is unwarranted, for some $\kappa \geq 0$.

A *nonstatistically significant* result (P-value not small) licenses inferences of the form $\mu \leq [\mu_0 + \gamma]$, for some $\gamma \geq 0$, but with a warning that $\mu \leq [\mu_0 + \kappa]$ is unwarranted for sufficiently small values of κ.

Severity Versus Power

An N-P test sets a cutoff beyond which the test "rejects" H_0, often written as c_α. The type 1 error probability is $\Pr(d(X) > c_\alpha; H_0) = \alpha$. Fixing α at a small value, the N-P test seeks to minimize the probability of a type 2 error: erroneously failing to reject H_0, or maximize the test's power. The



Severe Testing, Fig. 2 Power curve

power of the test is relative to an alternative, yielding a *power curve*. The power curve for testing H_0: $\mu \leq 0$ where $c_\alpha = 2$ against H_1: $\mu > 0$ where $c_\alpha = 2$ is shown in Fig. 2. The power of the test T+ to detect alternative $\mu = \mu_1$ may be abbreviated as POW (μ_1).

As the POW (μ_1) increases, a result that is just statistically significant at level α (i.e., $d(x_0) = d_\alpha$) corresponds to a *decreasing* severity for $\mu > \mu_1$ (compare Figs. 1 and 2).

Error Probabilities Versus Posterior Probs

The most well-known fallacy in interpreting significance tests is to equate the *P*-value with a posterior probability on the null hypothesis. However:

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

(ii) $P(H_0 \mid d(X) \geq d(x_0))$.

The *P*-value assessment in (i) refers only to the sampling distribution of the test statistic $d(X)$; there is no use of prior probabilities, as would be necessitated in (ii). There are cases where *P*-values match posterior probabilities, notably in our one-sided test T+ with suitably diffuse priors (frequentist matching) (Pratt 1965; Casella and Berger 1987; Berger 2006; Fraser et al. 2010).

D. Mayo and D.R Cox: SEV and FEV

In their proposal to view "Frequentist Statistics as a Theory of Inductive Inference," D. Mayo and Sir D. R. Cox develop an analogous approach to a severity assessment by means of a frequentist principle of evidence (FEV) introduced for the one-sided test in Mayo and Cox (2006):

> **FEV (i):** ***y*** is (strong) evidence against H_0, i.e., (strong) evidence of discrepancy from H_0, if and only if, where H_0 a correct description of the mechanism generating ***y***, then, with high probability, this would have resulted in a less discordant result than is exemplified by ***y***.
>
> **FEV (ii)**: A moderate *p* value [e.g., greater than .1] is evidence of the absence of a discrepancy δ from H_0, only if there is a high probability the test would have given a worse fit with H_0 (i.e., smaller *p* value) were a discrepancy δ to exist (Mayo and Cox 2006, 82–84).
>
> . . . To infer the absence of a discrepancy from H_0 as large as δ we may examine the probability $\beta(\delta)$ of observing a worse fit with H_0 if $\mu = \mu_0 + \delta$. If that probability is near one then, following **FEV(ii)**, the data are good evidence that $\mu < \mu_0 + \delta$. Thus $\beta(\delta)$ may be regarded as the stringency

> or severity with which the test has probed the discrepancy δ; equivalently, one might say that $\mu < \mu_0 + \delta$ has passed a severe test.... Such an assessment is more relevant to specific data than is the notion of power, which is calculated relative to a predesignated critical value beyond which the test "rejects" the null (Mayo and Cox 2006, 88–89; Cox 2006, 25).

A discussion and extension of the Mayo/Cox joint work is Spanos 2010.

Severity Reinterprets Confidence Intervals

Neyman (1937) developed confidence intervals as inversions of tests. The confidence interval contains the values of μ that would not be rejected by $\bar{x}_0$, were they the ones under test. $\bar{x}_0$ is the observed sample mean. The one-sided lower confidence-bound CI_L corresponding to test T+: $\mu = \mu_1$ versus $\mu > \mu_1$ is $\bar{x}_0 - k_\varepsilon\sigma/\sqrt{n}$, although typically σ would be estimated. Given the duality between CIs and N-P tests, it is unsurprising that CIs inherit problems of N-P tests. The textbook frequentist justification for inferring $\mu > \text{CI}_\text{L}$ is behavioristic: It is an estimate that arose from a procedure that includes the true value of the parameter $(1 - \varepsilon)\%$ of the time. By contrast, the severity rationale for inferring $\mu > \text{CI}_\text{L}$ is counterfactual: If μ were less than or equal to the lower confidence-bound CI_L, then with high probability $(1 - \varepsilon)$, the procedure would have resulted in a sample mean that is smaller than $\bar{x}_0$. By the severity principle, it follows that $\mu > \text{CI}_\text{L}$ passes with severity $(1 - \varepsilon)$.

Another shortcoming of standard CIs is treating all members of the confidence interval on a par, whereas it is important to distinguish them. In the severity construal, each point μ' in the confidence interval corresponds to a distinct claim of form either $\mu > \mu$' or $\mu < \mu$' and is assessed with a different severity. Finally, the severity construal deals painlessly with cases where a $1 - \varepsilon$ confidence interval ($\varepsilon > 0$) contains all possible parameter values or is empty. The first case informs us that no possible parameter value is ruled out with severity $1 - \varepsilon$. The empty interval indicates model violation (see Cox and Mayo 2010, 291).

Biasing Selection Effects

When null hypotheses, test statistics, or data generation are influenced by preliminary inspection of the data, the error probabilities associated with a test may be altered in such a way as to violate severity requirements, or prevent severity from being assessed (even approximately). These are *biasing selection effects*, and they are the most common reason that results fail to replicate when independent groups set out with predesignated protocols (Benjamini 2020). Several well-known gambits—cherry picking, data dredging, multiple testing, optional stopping, and *P*-hacking—invalidate error probabilities and thereby vitiate claims to have done a good job avoiding erroneous interpretations of data. A classic example that threatens the interpretation of *P*-values in clinical trials is to data dredge until finding a subgroup of the treated group showing a large difference in the direction sought.

Another type of multiplicity is optional stopping. A famous example is the two-sided test: H_0: $\mu = 0$ against H_1: $\mu \neq 0$, where instead of fixing the sample size, the test continues sampling until $|d(X)| \geq 2\text{SE}$.[4] With probability 1, it will stop with a "nominally" significant result even though $\theta = 0$ (Edwards et al. 1963). In the same way, it can be ensured that the true parameter value is always excluded from the corresponding 95% confidence or Bayesian credible interval (Berger and Wolpert 1988, p. 81).

In cases of multiplicity, auditing the *P*-value or confidence level, if it is to be used in a severity assessment, requires a *P*-value adjustment. By contrast,

> The likelihood principle emphasized in Bayesian statistics implies, among other things, that the rules governing when data collection stops are irrelevant to data interpretation. (Edwards et al. 1963, p. 193)

Inference by Bayes rule is conditional on the data. Error probabilities, which require considering outcomes other than the one observed, do not enter.

[4]This is an example of what is called a *proper stopping rule*: The probability it will stop in a finite number of trials is 1, regardless of the true value (Savage 1962).

Those who endorse the likelihood principle often defend ignoring the stopping rule, arguing that error probabilities matter only if the concern is good performance in the long run. Severe testing gives an inferential rationale: The test has done a poor job at avoiding the error of concern: being fooled by randomness (Mayo 2018; Mayo and Hand 2022; Mayo and Kruse 2001). One is not always in the context of severe probing of error. In contexts of discovery or exploration, it might be argued, the goal is to arrive at fruitful hypotheses to subject to subsequent tests.

Conclusion

By supplementing statistical significance tests with a discrepancy parameter, and supplying a postdata quantitative assessment of well-testedness, severity avoids many classic fallacies and perceived shortcomings of statistical significance tests. At a deeper level, severe testing affords a novel philosophical standpoint about the role of probability in inference. What a severe tester seeks is not a comparative measure of belief, plausibility, or support, but the ability to falsify, statistically, claims about population effect sizes or discrepancies. The quantitative aspects arise in the form of degrees of severity and sizes of discrepancies detected or not.

Acknowledgments

I am grateful to Jean Miller for useful comments on earlier versions; I acknowledge Marcos Jiménez for providing the figures.

About the Author

Note by the Editor: The concept of Severe testing is intricately linked to the philosophical contributions of Deborah G. Mayo, a notable figure in the field of the philosophy of science. Renowned for her work in areas such as statistical inference and the philosophy of statistics, Mayo formulated the concept of severe testing as an integral component of her broader outlook on the philosophy of science. Her emphasis lies in underscoring the significance of rigorous testing and falsifiability within the realm of scientific inquiry.

Mayo's extensive body of work, including her impactful book titled *Error and the Growth of Experimental Knowledge*, delves into the principles of severe testing and explores its application in the context of scientific hypotheses and statistical methods. This concept has played a pivotal role in shaping conversations and controversies within the domains of the philosophy of science and the philosophy of statistics.

References

Benjamini, Y.: Selective inference: the silent killer of replicability. Harv. Data Sci. Rev. **2**(4) (2020). https://doi.org/10.1162/99608f92.fc62b261

Berger, J.O.: The case for objective Bayesian analysis and rejoinder. Bayesian Anal. **1**(3), 385–402 (2006).; 457–64

Berger, J., Wolpert, R.: The Likelihood Principle Lecture Notes-Monograph Series, vol. 6, 2nd edn. Institute of Mathematical Statistics (1988)

Carnap, R.: Logical Foundations of Probability, 2nd edn. University of Chicago Press, Chicago (1962)

Casella, G., Berger, R.: Reconciling Bayesian and frequentist evidence in the one-sided testing problem. J. Am. Stat. Assoc. **82**(397), 106–111 (1987)

Cox, D.R.: Some problems connected with statistical inference. Ann. Math. Stat. **29**(2), 357–372 (1958)

Cox, D.R.: The role of significance tests (with discussion). Scand. J. Stat. **4**, 49–70 (1977)

Cox, D.R., Hinkley, D.V.: Theoretical Statistics. Chapman and Hall (1974)

Cox, D.R., Mayo, D.G.: Objectivity and conditionality in frequentist inference. In: Mayo, D., Spanos, A. (eds.) Error and inference: Recent exchanges on experimental reasoning, reliability, and the objectivity and rationality of science, pp. 276–304. Cambridge University Press, New York (2010)

Edwards, W., Lindman, H., Savage, L.: Bayesian statistical inference for psychological research. Psychol. Rev. **70**(3), 193–242 (1963)

Fisher, R.A.: The Design of Experiments. Oliver and Boyd, Edinburgh (1935)

Fraser, D.A.S., Reid, N., Marras, E., Yi, G.Y.: Default priors for Bayesian and frequentist inference. J. R. Stat. Soc. Ser. B Methodol. **72**(5), 631–654 (2010)

Gigerenzer, G.: The superego, the ego, and the Id in statistical reasoning. In: Keren, G., Lewis, C. (eds.) A Handbook for Data Analysis in the Behavioral Sciences: Methodological Issues, pp. 311–339. Erlbaum, Hillsdale (1993)
Jeffreys, H.: Theory of Probability, 3rd edn. Oxford University Press, Oxford (1961)
Mayo, D.G.: Novel evidence and severe tests. Philos. Sci. **58**(4), 523–552 (1991) Reprinted (1991) in The Philosopher's Annual XIV: 203–232
Mayo, D.G.: Error and the Growth of Experimental Knowledge. University of Chicago Press, Chicago (1996)
Mayo, D.G.: Statistical Inference as Severe Testing: How to Get Beyond the Statistics Wars. Cambridge University Press, Cambridge (2018)
Mayo, D.G., Cox, D.R.: Frequentist statistics as a theory of inductive inference. In: Rojo, J. (ed.) Optimality: The Second Erich L. Lehmann Symposium Lecture notes-monograph series, pp. 77–97. Institute of Mathematical Statistics (IMS), 49, Hayward (2006)
Mayo, D.G., Hand, D.: Statistical significance and its critics: practicing damaging science, or damaging scientific practice? Synthese. **200**, 220 (2022)
Mayo, D.G., Kruse, M.: Principles of inference and their consequences. In: Cornfield, D., Williamson, J. (eds.) Foundations of Bayesianism, pp. 381–403. Kluwer Academic Publishers, Dordrecht (2001)
Mayo, D.G., Spanos, A.: Severe testing as a basic concept in a Neyman-Pearson philosophy of induction. Br. J. Philos. Sci. **57**, 323–357 (2006)
Mayo, D.G., Spanos, A.: Error statistics. In: Bandyopadhyay, P., Forster, M. (eds.) Philosophy of Statistics Handbook of the Philosophy of Science, vol. 7, pp. 153–198. Elsevier, Amsterdam (2011)
Neyman, J.: Outline of a theory of statistical estimation based on the classical theory of probability. Philos. Trans. R. Soc. Lond. A. **236**(767), 333–380 (1937) Reprinted 1967 in Early Statistical Papers of J. Neyman, University of California Press, Berkeley, 250–90
Neyman, J., Pearson, E.: On the problem of the most efficient tests of statistical hypotheses. Philos. Trans. R. Soc. Lond. A. **231**, 289–337 (1933) Reprinted in Joint Statistical Papers, of J. Neyman and E. S. Pearson. University of California Press, Berkeley (1967) 140–85
Neyman, J., Pearson, E.: Joint Statistical Papers of J. Neyman and E. S. Pearson. University of California Press (1967)
Popper, K.: The Logic of Scientific Discovery, vol. 12, p. 53. Basic Books, New York (1959)
Pratt, J. Bayesian Interpretation of Standard Inference Statements. J. R. Stat. Soc. Ser. B Methodol. **27**(2), 169–203 (1965)
Savage, L.J. (ed.): The Foundations of Statistical Inference: A Discussion, vol. 57, p. 307. Methuen, London (1962)
Spanos, A. On a new philosophy of frequentist inference: Exchanges with David Cox and Deborah Mayo. In: Mayo, D., and Spanos, A. (eds.) Error and inference: Recent exchanges on experimental reasoning, reliability, and the objectivity and rationality of science, pp.315–330. Cambridge University Press, New York (2010)
Spanos, A.: Probability Theory and Statistical Inference: Empirical Modeling with Observational Data. Cambridge University Press, Cambridge (2019)
Wasserstein, R., Schirm, A., Lazar, N.: Moving to a world beyond "$p < 0.05$". Am. Stat. **73**(1), 1–19 (2019)

Sign Test, The

Peter Sprent
Division of Mathematics, University of Dundee, Dundee, Scotland

The sign test is a nonparametric test for hypotheses about a population median given a sample of observations from that population, or for testing for equality of medians, or for a prespecified constant median difference, given paired sample (i.e. matched pairs) values from two populations. These tests are analogues of the one-sample and matched pairs t-test for means in a parametric test such as the t-test.

The sign test is one of the simplest and oldest nonparametric tests. The name reflects the fact that each more detailed observation is effectively replaced by one of the signs plus (+) or minus (−). This was basically the test used by Arbuthnot (1710) to refute claims that births are equally likely to be male or female. Records in London showed that for each of 81 consecutive years an excess of male over female births. Calling such a difference a plus, Arbuthnot argued that if births were equally likely to be of either gender, then the probability of such an outcome was, $(0.5)^{81}$, or effectively zero.

Given a sample of n observations from any population which may be discrete or continuous and not necessarily symmetric, the test is used to test a hypothesis $H_0 : M = M_0$ where M is the population median. If H_0 holds the number of values less than M_0 will have a binomial