Of War or Peace? Essay Review of *Statistical Inference as Severe Testing*

Samuel C. Fletcher*

scfletch@umn.edu

University of Minnesota, Twin Cities

November 20, 2019

The subtitle of Deborah G. Mayo's *Statistical Inference as Severe Testing: How to Get Beyond the Statistics Wars* suggests an aim to ameliorate or transcend the acrimonious debates about the foundations of statistics in the twentieth century. These debates among scientists, statisticians, and (some) philosophers have concerned the proper answers to questions about the nature of scientific evidence and statistical methodology. Classical statistics—also known as *frequentist* statistics for its historical affinity with frequentist interpretations of probability—uses test statistics and their p-values to quantify the (in)compatibility of data with a statistical hypothesis. The methods of classical statistics thus have affinities with Popperian ideas of falsification and corroboration. Many scientists and mathematicians contributed to pushing it

*Thanks to Katie Creel, Kasey Genin, Dan Malinsky, Deborah G. Mayo, and Conor Mayo-Wilson for comments on a previous draft.

through adolescence in the 1920s and '30s, but at the center were Ronald A. Fisher, Jerzy Neyman, and Egon Pearson. Bayesian statistics, which has seen a renaissance since mid-century, gauges evidence through the change of prior probability distributions assigned to states of the world, via the famous Bayes' theorem. Philosophers may be more familiar with these latter ideas through the popularity of Bayesian epistemology, but in Bayesian statistics they often take a different form (as Mayo reveals in her chapter 6, which I discuss below). There is also likelihoodist statistics, which adopts Bayesian statistics' emphasis on the conditional probability of the observed data to measure evidence, without combining it with a prior or using Bayes' theorem. About all these, Mayo writes, "my goal is to tell what's true about statistical methods themselves" (4) and "to disentangle a jungle of conceptual issues, not to defend or criticize any given statistical school" (183). Indeed, "I will not be proselytizing for a given statistical school, so you can relax" (12).

This is only so if Mayo's excludes as a "school" her own, the titular *severe testing*. Her book is indeed best read as a substantial and important contribution to these very foundational debates, articulating and defending a particular position within classical statistics (what she calls *error statistics*) as the *correct* one for scientists—"For statistical inference in science, it is severity we seek" (437)—and for decision-makers and everyday inquirers: "The severity demand is what we naturally want as consumers of statistics" (444). Severe testing concerns the warrant of inferring statistical hypotheses from data. To describe how first requires some terminology. A *simple* (or "point") statistical hypothesis assigns a probability distribution to possible data sets. For example, in a series of independent coin tosses, there is a simple statistical hypothesis for each value, lying between zero and one, of the probability that the coin will land heads

on a single toss. These hypotheses determine the probabilities of each possible sequence of heads and tails. So far, all schools of statistical inference share this in common. Classical statistical methods add a unary "distance" function d_H on possible data X—a particular type of *test statistic*, or function of the data X—that quantifies the difference between a data set and what is expected or likely, relative to some hypothesis H.[1]

Mayo add two criteria for interpreting how such discrepancies undergird statistical evidence *for* a hypothesis. If one is testing a ("null") hypothesis H, whether simple or composite—the disjunction of some simple hypotheses—then data x provide evidence for H to the extent that:

- $d_H(x)$, the data's difference from what is expected under that hypothesis, is low, and
- $Pr_{\neg H}(d_H(X) > d_H(x))$, the probability of a difference larger than that given by the data, if that hypothesis were false, is high.

The first condition is essentially equivalent to requiring that the p-value of the test statistic forming the "distance" function be moderate or high, i.e., not close to zero. The probability in the second condition defines the *severity* of this test. Requiring high severity, a *counterfactual* condition, distinguishes Mayo's account from traditional versions of classical statistics, which instead typically appeal to the long-run error rates of testing procedures—e.g., how often one will be in error deciding to accept an

[1]This is more restricted than some approaches to classical statistical testing, which allow essentially *any* test statistic and focus on the long-run probabilities of erroneous inference with them. That d_H should actually quantify a *difference* between the data and what is likely can be found in the work of Cox (1958) and Pearson (1947).

hypothesis if the difference is sufficiently low (i.e., the type II error rate), or to accept its negation if the difference is sufficiently high (the type I error rate).[2] Severe testing is thus capable to show how data positively corroborate a hypothesis, not just conflict with it. Severity differs from the power of a test, $Pr_{\neg H}(d_H(X) > c_\alpha)$, in that it depends on the data instead of a data-independent constant c_α set by the type I error rate. (If $\neg H$ is composite, both severity and power are implicitly functions of the simple hypotheses that form it by disjunction; in such cases Mayo reports the minimum of the severity function—see footnote 2.) Thus severity is a data-dependent analog of power, like how the p-value of a test is the data-dependent analog of the test's type I error rate.

Mayo rightfully recognizes and emphasizes this difference and its epistemological consequences: "This rule will be at odds with some common interpretations of tests. . . . [But] the correct error-statistical view is this one" (201). It also differs from the perspective of classical statisticians who advocate using confidence intervals (CIs) over tests. CIs are sets of simple hypotheses selected by a certain rule, based on the data and assumptions about their possible probability distributions, that guarantees the sets contain the true simple hypothesis with a certain probability. For example, a 95% CI procedure for the coin-tossing example would produce an interval of probabilities for heads, based on the data recorded, such as $[0.4, 0.6]$ for data consisting of half heads and tails, such that the probability that the interval contains the true heads probability is 0.95. One of CIs' claimed advantages over tests is that they provide a method for selecting a *range* of hypotheses that the data best support. Mayo is sympathetic but

[2]Mayo has a separate description of the severity criteria needed for evidence of a composite hypothesis logically stronger than the negation (265–6, 351–2): essentially, the severity for each simple component of the hypothesis must be sufficiently high.

unswayed: "I don't want to step too hard on the CI champion's toes, since CIs are in the frequentist, error statistical tribe. Yet, to avoid fallacies, this standard use of CIs won't suffice" (245). The reason, Mayo avers, is that CIs formally treat all the simple hypotheses within the set they select on a par, yet they ought not if their severities differ, as they typically will. In the coin-tossing example with data consisting of half heads and tails, a heads probability of 0.5 is not equally supported as 0.6, despite both being in the same interval. In defending her position and distinguishing it from others within classical statistics, she believes it will resolve or transcend debates about the foundations of statistics: "Isolating out a particular conception of statistical inference as severe testing is a way of telling what's true about the statistics wars, and getting beyond them" (11).

Statistical Inference as Severe Testing incorporates elaborations of severe testing after Mayo's 1996 book, often done in collaboration with econometrician Aris Spanos, such as its integration with misspecification testing, the testing of auxiliary modeling assumptions in statistical inference. Another novelty is the book's audience and voice. Mayo has found a growing audience among philosophers and statisticians for public discussion of conceptual issues in statistics at her website and blog (Mayo, 2011), whose stylistic conventions allow for more literary, unconventional, and playful presentations of ideas. In the book, Mayo accordingly adopts an informal tone that shifts quite casually between different voices: first-person, both singular and plural, second-person, and third-person all find a place in the text. At times this can distract from her argument, but it occasionally augments them, as with a dramatized "theatre" production in chapter 5.7 involving Neyman, Pearson, Joseph Betrand, and Émile Borel, composed entirely of quotations from the primary sources. Nevertheless, while these stylistic choices may on balance invite a wider audience, some familiarity with the statistics wars

and Mayo's previous work would aid comprehension of certain content. Part of the reason is that Mayo sometimes varies her terminology for key ideas across the course of the book, so readers unfamiliar with the terminology may be unsure of how they relate. For example, severity is also variously labeled sensitivity (151), attained sensitivity (196), attained power (342), corroboration (343), and severe corroboration (408). She acknowledges that when it comes to the severity requirement, "I deliberately phrase it in many ways" (209) but the effect on understanding is not salutary. It does reflect the variegated phrasings of casual conversation among specialists on these topics, but those who haven't learned the local jargon will need more patience.

For philosophers of science (and any others unfamiliar with the many battles within the statistics wars) who persist, a cornucopia of rich examples and debates awaits—much more than can be outlined here. It's to Mayo's credit that she introduces to this audience interesting conceptual debates that have played out in the statistics literature, where philosophers' specialized skills often prove useful. Indeed, Mayo is often at her most insightful in her careful rebuttals of various critiques of classical statistical testing, based on the assumptions of that framework. Particular standouts in my mind were the reply to the CI advocate on behalf of statistical testers, which I already discussed above, and her analysis and rebuttal of influential arguments from Bayesian statistics that p-values exaggerate the evidence for certain hypotheses in testing (Berger and Sellke, 1987) or do not correspond with the error probabilities of tests (Berger, 2003).

There are two further little-discussed topics for which Mayo deserves recognition for bringing to readers' attention. One is historical: based in part on Erich Lehmann's posthumously published final book (2011), she argues, contrary to widespread conception, for a reconciliation of what are usually seen as rival and incompatible schools

of classical statistical testing: that of Fisher and that of Neyman and Pearson. Typically, authors present Fisher as the hard-nosed scientist who used p-values as measures of evidence against hypotheses but had no method to determine when to reject (or accept) any hypotheses; Neyman and Pearson were the quality control engineers, who only cared about making a decision to reject or accept, justifying such behavioral strategies in terms of long-term error rates. Once understood in context, however, the quotations usually pulled to indict them of propounding methods antithetical to the needs of science—by the like of philosophers such as Hacking, Levi, Howson and Urbach, Sober, and others—shows instead that Neyman and Pearson were writing initially rather in the narrow context of a debate between Betrand and Borel on (effectively) the objectivity of test statistics, and the importance of predesignation thereof. Neyman developed confidence intervals and other mathematical techniques to try to explicate Fisher's statistical deliverances; sometimes Fisher used dichotomous statistical decision procedures, and sometimes Neyman and Pearson used p-values to quantify evidence. The chasm between them was only dug over time with retrenchment of difficult personalities: "At times, Neyman exaggerates the behavioristic conception just to accentuate how much Fisher's tests need reining in. Likewise, Fisher can be spotted running away from his earlier behavioristic positions just to derogate the new N-P movement, whose popularity threatened to eclipse the statistics program that was, after all, his baby" (165). Thus accusations of inconsistency or conceptual incoherence towards modern approaches that blend these viewpoints are not based in a careful reading of history (179).

The second topic is the epistemological unclarity in modern Bayesian statistics. Although Mayo does go in for some standard criticisms of philosophers' confirmation

theory and subjective probability in scientific contexts, these are not especially new criticisms, nor ones that will likely sway convinced Bayesians. What is much more interesting is the contrast, in chapter 6.3–6.6, of the ideal image of Bayesianism in epistemology and philosophy of science with its pragmatic, even unprincipled practice in Bayesian statistics. Priors may no longer represent degrees of belief; updating solely by Bayes' rule may go by the wayside as the prior is determined in part from the data; one may change the likelihood function by running statistical tests with Bayesian p-values, or consider the prior merely as a device "to smooth the likelihood, making fitted models less sensitive to details of the data" (435). Bayesian in mathematical machinery, but frequentist in interpretation? A significant portion of self-described applied Bayesian statistics takes this attitude. In my opinion, finding an epistemological justification for this pragmatic Bayesianism is the biggest open issue in philosophy of statistics, because so much science relies on it.

Despite the richness of many of Mayo's discussions, some of them do not meet their stated goals. I'll describe two. The first concerns the so-called replication crisis in fields like social psychology and biomedicine where researchers consistently fail to reproduce many published high-profile experimental results. Mayo writes that the *statistics wars* "are the engine behind current controversies surrounding high-profile failures of replication in the social and biological sciences" (xi). Many diagnoses and treatments have been proposed, so "If we are to appraise these evidence policy reforms, a much better grasp of some central statistical problems is needed" (3). Much of the literature has focused on *institutional* causes of the replication crisis, such as lack of control over publication bias and implicit professional incentives for unjustified and questionable research practices (QRPs) that over-represent positive findings and impede independent

checks of those findings (Fidler and Wilcox, 2018). Thus, revealing the essential relevance of issues from the statistics wars, viz. those debates concerning the nature and goals of statistical methodology, would be significant and novel.

But when Mayo opens the hood to examine the engine, she shines a questioning spotlight only on familiar problems about the "adequacy of the leap from the statistical [hypothesis] to the substantive [scientific hypothesis]" and "testing the methods and measurements intended to link statistics with what [researchers] really want to know" (100). These are well-recognized cautions about mathematical modeling generally: one must be careful to ensure that statistical hypotheses rejected or corroborated actually represent well the scientific ("substantive") hypotheses of interest. That's true, but unrelated with foundational issues in scientific evidence; it establishes no new connection between replication and the statistics wars. Mayo further emphases that according to her conception of severe testing, QRPs indeed produce little evidence (104), and she reminds us that "all the fraud-busting [of questionable research] is based on error statistical reasoning (if only on the meta-level)" (21–22), but the way in which this is so does not distinguish severe testing from conventional classical statistics. It provides no new insight into the source of the replication crisis.

The second discussion that falls short of its stated goals is about objectivity in classical statistics. Although concerns about objectivity in statistics arise most commonly regarding prior probabilities in Bayesian statistics, they arise too for classical statistics when it comes to the choice of "distance" function.[3] Different functions can

[3]They also arise in the use of conditioning to eliminate nuisance parameters—when to condition, and when not to? (Cox, 1958)—but I will focus just on the case of the "distance" function.

generally yield different evidential, hence inferential, verdicts from the same data. This would be a problem for Mayo if it allowed scientists' potentially bias-inducing beliefs and interests to guide their choice of function, hence their evidential verdicts, possibilities she rejects as inimical to scientific objectivity (223). Now, there are certainly valuable criteria to impose on such functions, as Mayo discusses in chapter 3.2: unbiasedness, consistency, and monotonicity with respect to decreasing likelihood ratio—i.e., the power associated with a test using the function should increase as sample size increases and sample variability decreases. The ideal case is some sort of uniformly most powerful (UMP) test. But in general an UMP test may not exist, and the aforementioned criteria do not pick out a *unique* function, even if one is to accept them (cf. Howson and Urbach, 2006, Ch. 5). Why doesn't the resulting latitude in specification introduce a space for bias or the imposition of values?

Mayo dismisses this line of argument: "The mistake is to suppose we are incapable of critically scrutinizing how discretionary choices influence conclusions. ... [O]ur critical evaluation of what the resulting data do and do not indicate need not itself be a matter of economics, ethics, or what have you" (224). There are two problems with this response. The first is that it isn't yet clear how critical scrutiny's normative force *actually*—not just possibly—precludes the influence of non-epistemic values. It cannot be mere agreement because Mayo insists (225) that intersubjective consensus forms too thin a notion of objectivity. Here there is a large and subtle literature on scientific objectivity with which Mayo could have connected (e.g., Longino, 1990). This is a missed opportunity.[4] The second problem is that without further criteria that provably

[4]Mayo was aware of the opportunity: she discusses Longino's ideas in a few sentences in the book (236).

select a best function, there are no grounds available on which critical scrutiny could select between the different choices.[5] Until these are known, the question of severe testing's objectivity, in the above sense, is not settled.

To be completely clear, these observations are not objections to severe testing per se, only to claims that the theory is complete (xii, 437) and allows its adherents to get beyond the statistics wars. It deserves rather to be developed further, an ambition in which Mayo occasionally and rightfully asks the reader to indulge: "The more devoted amongst you will want to improve and generalize my severity curves" (350). Such a project would be worthwhile not just for philosophy of science itself, but for the philosophical foundation of the statistical methods on which so much scientific knowledge relies.

References

Berger, J. O. (2003). Could Fisher, Jeffreys and Neyman have agreed on testing? *Statistical Science 18*(1), 1–32.

Berger, J. O. and T. Sellke (1987). Testing a point null hypothesis: The irreconcilability of p values and evidence. *Journal of the American Statistical Association 82*(397), 112–122.

[5] Another option is to show that the choice within a range of acceptable functions never makes any practical difference, but this would need to explicate some interest-neutral notion of "practical difference."

Cox, D. R. (1958). Some problems connected with statistical inference. *Annals of Mathematical Statistics 29*, 357–372.

Fidler, F. and J. Wilcox (2018). Reproducibility of scientific results. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2018 ed.). Metaphysics Research Lab, Stanford University.

Howson, C. and P. Urbach (2006). *Scientific Reasoning: The Bayesian Approach* (3rd ed.). Chicago: Open Court.

Lehmann, E. L. (2011). *Fisher, Neyman, and the creation of classical statistics*. New York: Springer Science & Business Media.

Longino, H. (1990). *Science as Social Knowledge: Values and Objectivity in Scientific Inquiry*. Princeton: Princeton University Press.

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

Mayo, D. G. (2011). Error statistics philosophy. `https://errorstatistics.com/`. [Online; accessed 16 Nov. 2019].

Pearson, E. S. (1947). The choice of statistical tests illustrated on the interpretation of data classed in a 2×2 table. *Biometrika 34*(1/2), 139–167.