Mayo, Deborah. *Statistical Inference as Severe Testing How to Get Beyond the Statistics Wars.* Cambridge: Cambridge University Press, 2018. xv + 471 pp.

For more than three decades, Deborah Mayo has defended *severe testing*, a framework for evaluating the strength of scientific evidence that simultaneously provides philosophical foundations for (variants of) classical statistical methods. In her newest book, *Statistical Inference as Severe Testing*, Mayo (1) extends and refines her views on severe-testing, (2) takes aim at researchers who claim that the reproducibility crisis is a result of the use of classical statistics, and (3) exposes some contemporary Bayesian methods as lacking philosophical foundations. The book is engaging, sometimes funny, and often insightful. After reviewing the basics of severe testing, I summarize three lessons from Mayo's book that are valuable for philosophers and scientists alike. I then discuss one way in which I think Mayo's arguments could be tightened.

Mayo argues that one has evidence for a hypothesis to the extent the hypothesis has passed a *severe test*. More precisely, she argues that "for data to warrant a hypothesis H 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" (92). Mayo's two criteria are inspired by the Neyman-Pearsonian maxim that good statistical tests minimize Type I and Type II errors, but they are novel in at least two ways.

First, Mayo's criteria are applicable outside of statistical contexts. For example, let H be the hypothesis that the Cleveland Browns played an excellent team last weekend, and suppose our data is that the Cleveland Browns lost their game. If the Cleveland Browns lose to all opponents regardless of ability, then our data fails to satisfy S2 and so does not severely test H. Since S1 and S2 are applicable in both statistical and non-quantitative contexts, Mayo's criteria unify the logic of statistical testing with qualitative logics of evidence.

Second, severity is a function of data, unlike size and power, the two core quantities of orthodox Neyman-Pearsonian methodology. For this reason, Mayo's two criteria closely resemble Nozick's *adherence* and *sensitiv-*

ity conditions when those conditions are relativized to "methods" [Nozick, 1981].[1]

Mayo draws several important lessons from this seemingly simple, two-pronged framework. Here, I discuss three that recur throughout the newest book. First, Mayo carefully distinguishes between statistical hypotheses (e.g., about the frequency of a trait in a population) and "substantive" scientific hypotheses about, for example, causation or the existence of a theoretical entity (84, 93). Conflating the two can lead one to mistakenly infer that a substantive scientific hypothesis has been severely tested, when in fact the evidence against the hypothesis is stronger than that for it.

As a vivid example, Mayo discusses a series of psychological experiments used to defend the hypothesis H that heterosexual "men's implicit self-esteem is lower when a [female] partner succeeds than when a [female] partner fails" (101). Mayo carefully distinguishes four statistical hypotheses that are plausibly consequences of $\neg H$ and then argues that the data were sufficient to reject only one of the four; call it $\neg H_1$. Mayo concludes that the experimental data failed to fit H, as researchers could not reject three statistical hypotheses that were compatible with $\neg H$. Further, Mayo argues that the *statistical* evidence *against* $\neg H_1$ (i.e., the purported evidence *for* $\neg\neg H = H$) was perfectly compatible with the falsity of the *substantive* hypothesis H. So the data also failed to satisfy S2.

Second, Mayo argues that, whereas the severe-tester has a unified account of testing both statistical hypotheses *and* modeling assumptions, likelihoodists and Bayesians lack a cogent story about when to reject a bad statistical model (298-320). Some likelihoodists, for example, argue that measures of evidential strength must be comparative and that the likelihood ratio $P(E|H_0)/P(E|H_1)$ quantifies the degree to which evidence E fits a hypothesis H_0 better than H_1 [Edward, 1984, Royall, 1997]. It is easy to construct cases in which such a likelihood ratio is enormous, because the the data fits H_0 much better than H_1, and yet the data fit neither statistical hypothesis well. If H_0 and H_1 are the only statistical hypotheses under investigation, one might erroneously conclude that one has good evidence for H_0. The right conclusion, according to the severe tester, is to recognize that the two hypotheses do not exhaust logical space and to reject that underlying modeling assumption instead.

Naive Bayesian reasoning falls victim to the same criticism. According

[1] For a discussion of the relationship between Nozick's simple versions of adherence and sensitivity and Neyman-Pearsonian methodology, see [Mayo-Wilson, 2018] and [Mayo-Wilson, 2020]. See [Fletcher and Mayo-Wilson, 2020] for a comparison of the methods-relative versions of Nozick's conditions and Mayo's theory.

to some Bayesians, experimenters ought to assign prior probabilities to a fixed set of hypotheses $\mathcal{H}$ and then update their degrees of belief by conditionalization. Again, suppose the data fit all hypotheses in $\mathcal{H}$ poorly but fits one $H \in \mathcal{H}$ much better than its rivals. Then a naive Bayesian's posterior degree of belief in H might be nearly one. Again, the right conclusion, according to the severe tester, is to consider more hypotheses. But orthodox Bayesianism is silent about when to do that.

Finally, Mayo argues that several novel Bayesian approaches to statistical inference lack foundations (400-415). These parts of the book will be of special interest to philosophers who are unfamiliar with the substantial differences between Bayesianism in theory and in practice. According to textbook philosophical presentations, prior and posterior probabilities represent an experimenter's beliefs. However, some contemporary Bayesians have given up on the myth that the infinitely-precise densities that appear in journal articles are appropriate idealizations of any person's beliefs. Those Bayesians, however, disagree about what justifies the choice of prior (e.g., transparent communication vs. frequentist coverage properties) and the use of further Bayesian mathematical machinery.

Mayo levels several different criticisms against these contemporary Bayesian approaches, but I think all share the same underlying philosophical problem. Rationality, so says standard decision theory, requires one to maximize expected utility *relative to one own's degrees of belief*. So some contemporary "Bayesian" methods require a new theory of rationality to justify the behavior of an experimenter whose decisions employ posterior probabilities that do not even approximately represent her (or perhaps anyone's) beliefs.

There are other important lessons scattered through Mayo's engaging and fairly accessible book. That leads me to my main criticism: accessibility sometimes comes at the cost of precision. I focus on one example that is directly relevant to Mayo's analysis of the causes of the replicability crisis, a central topic of the book.

Mayo repeatedly suggests that a hypothesis is not severely tested if it is the result of *data-dredging*. "Data dredging" is one of several perjorative terms (including p-hacking) that some statistical reformers use, and it refers to a loosely defined set of purportedly objectionable practices for finding associations in data. Often, "data-dredging" describes the practice of subjecting a single data set to many (sometimes hundreds or thousands of) statistical tests in an attempt to find a statistically significant result that is publishable.

Critics of data-dredging allege that, because there is always a chance of finding a spurious association between two variables, one should be wary of "dredged" hypotheses. For severe testers, that default skepticism is justifiable: if an experiment is not designed to test a hypothesis, it is unlikely to be a severe test by chance. However, intuitions about data-dredging depend upon how the practice is described. Searching through data for patterns sounds exactly like what scientists *should* do!

Mayo illustrates the dangers of data-dredging via several thought experiments. For example, she imagines a pharmaceutical company that hires scientists to find the benefits of a recently developed drug (267-268). When the hired researchers find no evidence that the drug is salutary in any of ways the manufacturer had hoped, they mine their data to find *some* salutary effect. Sure enough, they find the drug has an "impressive benefit on factor B" and begin marketing the drug with that promise. Mayo concludes that the hired scientists fail to subject the hypothesis to a severe test because they would have found an effect no matter what.

Unfortunately, Mayo never explicitly applies her two criteria for severity to this example. Which hypothesis exactly is being tested? Here are two candidates. Let H_1 be the hypothesis, "The drug has some *some* salutary effect" and let H_2 be the hypothesis "The drug has the *specific* salutary effect found on B." Even if the drug manufacturer is guaranteed to find *some* effect (i.e., H_1 is not severely tested), it does not follow that the manufacturer lacks evidence that the drug has the *specific* effect found by mining the data (i.e., it does not follow that H_2 is not severely tested). When I grade student logic exams, I am guaranteed to find that *some* student scored highest, but I do not thereby lack evidence that Roshan scored best when I tally her score and see it is highest. Similarly, cigarette manufacturers cannot avoid liability for causing lung cancer by (1) collecting enormous amounts of data, (2) investigating the relationship between cigarette-smoking and hundreds of health problems, and then (3) claiming they were destined to find some statistically significant effect or another.[2]

Some accusations of data-dredging, I claim, are instances of a fallacy that Mayo typically distinguishes for her readers with great care: the conflation of the pre-experimental reliability of a test with the post-experimental evidence provided by a particular data set.[3] The history of science is replete with examples of unlikely, happy discoveries, where the data unequivocally support a hypothesis that scientists did not intend to test. In accessible, vivid prose, Mayo illustrates why one should *typically* be wary of data-

[2] See [Mayo-Wilson, 2018] for similar criticisms of theories of knowledge that employ only pre-sample measures of reliability.

[3] For different criticisms of this attitude towards data dredging, see [Kotzen, 2013].

dredging, but readers will be disappointed if they wish to learn how Mayo's two criteria can be used to distinguish between (a) hypotheses that are illicitly "dredged" and (b) hypotheses that are supported by one's data, despite being formulated post experiment.

Quibbles aside, Mayo's newest book is a yet another deep and engaging contribution to the philosophy of statistics and philosophy of science. Most importantly, it shows how foundational philosophical debates matter to pressing issues in science policy.

Conor Mayo-Wilson
Department of Philosophy
University of Washington

References

Anthony William Fairbank Edward. *Likelihood.* CUP Archive, 1984.

Samuel Fletcher and Conor Mayo-Wilson. Evidence in Classical Statistics. *Under Review*, 2020.

Matthew Kotzen. Multiple Studies and Evidential Defeat. *Nous*, 47(1): 154–180, 2013.

Conor Mayo-Wilson. Epistemic Closure in Science. *Philosophical Review*, 127(1):73–114, 2018.

Conor Mayo-Wilson. An epistemic defense of interval estimation. *British Journal for Philosophy of Science*, Forthcoming, 2020.

Robert Nozick. *Philosophical explanations.* Harvard University Press, 1981.

Richard Royall. *Statistical evidence: a likelihood paradigm*, volume 71. Chapman & Hall/CRC, 1997.