Journal for General Philosophy of Science (2020) 51:507–510
https://doi.org/10.1007/s10838-019-09486-2

BOOK REVIEW

Deborah G. Mayo: Statistical Inference as Severe Testing: How to Get Beyond the Statistics Wars

Cambridge University Press, Cambridge 2018, 486 pp, $29.99, ISBN: 9781107286184

Tom F. Sterkenburg[1]

Published online: 14 November 2019

The foundations of statistics is not a land of peace and quiet. "Tribal warfare" is perhaps putting it too strong, but it is the case that for decades now various camps and subcamps have been exchanging heated arguments about the right statistical methodology. That these skirmishes are not just an academic exercise is clear from the widespread use of statistical methods, and contemporary challenges that cry for more secure foundations: the rise of big data, the replication crisis.

One often hears that to blame are classical, frequentist methods, that lack a proper justification and are easily misused at that; so that it is all a matter of stepping up our efforts to spread the Bayesian philosophy. This not only ignores the various conflicting views *within* the Bayesian camp, but also gives too little credit to opposing philosophical perspectives. In particular, this does not do justice to the work of philosopher of statistics Deborah Mayo. Perhaps most famously in her Lakatos Award-winning *Error and the Growth of Experimental Knowledge* (1996), Mayo has been developing an account of statistical and scientific inference that builds on Popper's falsificationist philosophy and frequentist statistics. She has now written a new book, with the stated goal of helping us get beyond the statistics wars.

This work is a genuine tour de force. Mayo weaves together an extraordinary amount of philosophical themes, technical discussions, and historical anecdotes into a lively and engaging exposition of what she calls the *error-statistical* philosophy. Like few other works in the area Mayo instills in the reader an appreciation for both the interest and the significance of the topic of statistical methodology, and indeed for the importance of *philosophers* engaging with it.

That does not yet make the book an easy read. In fact, the downside of Mayo's conversational style of presentation is that it can take a serious effort on the reader's part to distill the argumentative structure and how various observations and explanations hang together. This, unfortunately, also limits its use somewhat for those intended readers that are new to the discussed topics.

✉ Tom F. Sterkenburg
tom.sterkenburg@lmu.de

[1] Munich Center for Mathematical Philosophy, LMU Munich, Munich, Germany

In the following I will summarize the book, and conclude with some general remarks. (Mayo organizes her book into "excursions" divided into "tours"—we are invited to imagine we are on a cruise—but below I will stick to chapters divided into parts.)

Chapter 1 serves as a warming-up. In the course of laying out the motivation for the book's project, Mayo introduces *severity* as a requirement for evidence. On the *weak* version of the severity criterion, one does *not* have evidence for a claim C if the method used to arrive at evidence $\boldsymbol{x}$, even if $\boldsymbol{x}$ agrees with C, had little capability of finding flaws with C even if they exist (Mayo also uses the acronym BENT: *bad evidence, no test*). On its *strong* version, if C passes a test that did have high capability of contradicting C, then the passing outcome $\boldsymbol{x}$ *is* evidence—or at least, an indication—for C. The double role for statistical inference is to identify BENT cases, where we actually have poor evidence; and, using strong severity, to mount positive arguments from coincidence.

Thus if a statistical philosophy is to tell us what we seek to quantify using probability, then Mayo's error-statistical philosophy says that this is "well-testedness" or *probativeness.* This she sets apart from *probabilism,* which sees probability as a way of quantifying plausibility of hypotheses (tenet of the Bayesian approach), but also from *performance*, where probability is a method's long-run frequency of faulty inferences (the classical, frequentist approach). Mayo is careful, too, to set her philosophy apart from recent efforts to unify or bridge Bayesian and frequentist statistics, approaches that she chastises as "marriages of convenience" that simply look away from the underlying philosophical incongruities. There is here an ambiguity in the nature of Mayo's project, that remains unresolved throughout the book: is she indeed proposing a new perspective "to tell what is true about the different methods of statistics" (p. 28), the view-from-a-balloon that might finally get us beyond the statistics wars, or should we actually see her as joining the fray with a yet different competing account? What is certainly clear is that Mayo's philosophy is much closer to the frequentist than the Bayesian school, so that an important application of the new perspective is to exhibit the flaws of the latter. In the second part of the chapter Mayo immediately gets down to business, revisiting a classic point of contention in the form of the likelihood principle.

In Chapter 2 the discussion shifts to Bayesian confirmation theory, in the context of traditional philosophy of science and the problem of induction. Mayo's diagnosis is that the aim of confirmation theory is *merely* to try to spell out inductive method, having given up on actually providing justification for it; and in general, that philosophers of science now feel it is taboo to even try to make progress on this account. The latter assessment is not entirely fair, even if it is true that recent proposals addressing the problem of induction (notably those by John Norton and by Gerhard Schurz, who both abandon the idea of a single context-independent inductive method) are still far removed from actual scientific or statistical practice. More interesting than the familiar issues with confirmation theory Mayo lists in the first part of the chapter is therefore the positive account she defends in the second.

Here she discusses falsificationism and how the error-statistical account builds and improves on Popper's ideas. We read about demarcation, Duhem's problem, and novel predictions; but also about the replicability crisis in psychology and fallacies of significance tests. In the last section Mayo returns to the question that has been in the background all this time: what is the error-statistical answer to the problem of inductive inference? By then we have already been handed a number of clues: inferences to hypotheses are arguments from strong coincidence, that (unlike "inductive" but really still deductive probabilistic logics) provide genuine "lift-off", and that (against Popperians) we are free to call warranted or justified. Mayo emphasises that the output of a

statistical inference is not a belief; and it is undeniable that for the plausibility of an hypothesis severe testing is neither necessary (the problem of after-the-fact cooked-up hypotheses, Mayo points out, is exactly that they can be so plausible) nor sufficient (as illustrated by the base-rate fallacy). Nevertheless, the envisioned epistemic yield of a (warranted) inference remains agonizingly imprecise. For instance, we read that (sensibly enough) isolated significant results do not count; but when do results start counting, and how? Much is delegated to the dynamics of the overall inquiry, as further illustrated below.

Chapter 3 goes deeper into severe testing: as employed in actual cases of scientific inference, and as instantiated in methods from classical statistics. Thus the first part starts with the 1919 Eddington experiment to test Einstein's relativity theory, and continues with a discussion of Neyman–Pearson (N–P) tests. The latter are then accommodated into the error-statistical story, with the admonition that the severity rationale goes beyond the usual behavioural warrant of N–P testing as the guarantee of being rarely wrong in repeated application. Moreover, it is stressed, the statistical methods given by N–P as well as Fisherian tests represent "canonical pieces of statistical reasoning, in their naked form as it were" (p. 150). In a real scientific inquiry these are only part of the investigator's reservoir of error-probabilistic tools "both formal and quasi-formal", providing the parts that "are integrated in building up arguments from coincidence, informing background theory, self-correcting […], in an iterative movement" (p. 162).

In the next part of Chapter 3, Mayo defends the classical methods against an array of attacks launched from different directions. Apart from some old charges (or "howlers and chestnuts of statistical tests"), these include the excusations arising from the "family feud" between adherents of Fisher and Neyman–Pearson. Mayo argues that the purported different interpretational stances of the founders (Fisher's more evidential outlook versus Neyman's more behaviourist position) are a bad reason to preclude a unified view on both methodologies. In the third part, Mayo extends this discussion to incorporate confidence intervals, and the chapter concludes with another illustration of statistical testing in actual scientific inference, the 2012 discovery of the Higgs boson.

The different parts of Chapter 4 revolve around the theme of objectivity. First up is the "dirty hands argument", the idea that since we can never be free of the influence of subjective choices, all statistical methods must be (equally) subjective. The mistake, Mayo says, is to assume that we are incapable of registering and managing these inevitable threats to objectivity. The subsequent dismissal of the Bayesian way of taking into account—or indeed embracing—subjectivity is followed, in the second part of the chapter, by a response to a series of Bayesian critiques of frequentist methods, and particularly the charge that, as compared to Bayesian posterior probabilities, *P* values overstate the evidence. The crux of Mayo's reply is that "it's erroneous to fault one statistical philosophy from the perspective of a philosophy with a different and incompatible conception of evidence or inference" (p. 265). This is certainly a fair point, but could just as well be turned against her own presentation of the error-statistical perspective as a meta-methodology. Of course, the lesson we are actually encouraged to draw is that an account of evidence in terms of severe testing is preferable to one in terms of plausibility. For this Mayo makes a strong case, in the next part, in connection to the need for tools to intercept various illegitimate research practices. The remainder of the chapter is devoted to some other important themes around frequentist methods: randomization, the trope that "all models are false", and model validation.

Chapter 5 is a relatively technical chapter about the notion of a test's *power*. Mayo addresses some purported misunderstandings around the use of power, and discusses the notion of *attained* or post-data power, combining elements of N–P and of Fisher, as part

of her severity account. Later in the chapter we revisit the replication crisis, and in the last part we are given an entertaining "deconstruction" of the debates between N–P and Fisher.

Finally, in Chapter 6, Mayo takes one last look at the probabilistic "foundations lost", to clear the way for her parting proclamation of the new probative foundations. She discusses the retreat by theoreticians from full-blown subjective Bayesianism, the shaky grounds under objective or default Bayesianism, and attempts at unification ("schizophrenia") or flat-out pragmatism. Saved till the end, fittingly, is the recent "falsificationist Bayesianism" that emerges from the writings of Andrew Gelman, who indeed adopts important elements of the error-statistical philosophy.

It seems only a plausible if not warranted inductive inference that the statistics wars will rage on for a while; but what, towards an assessment of Mayo's programme, should we be looking for in a foundational account of statistics? The philosophical attraction of the dominant Bayesian approach lies in its promise of a principled and unified account of rational inference. It appears to be too rigid, however, in suggesting a fully mechanical method of inference: after you fix your prior it is, on the standard conception, just a matter of conditionalizing. At the same time it appears to leave too much open, in allowing you to reconstruct any desired reasoning episode by suitable choice of model and prior. Mayo is very clear that her account resists the first: we are not looking for a purely formal account, a single method that can be mindlessly pursued. Still, the severity rationale is emphatically meant to be restrictive: to expose certain inferences as unwarranted. But the threat of too much flexibility is still lurking in how much is delegated to the messy context of the overall inquiry. If too much is left to context-dependent expert judgment, for instance, the account risks to forfeit its advertized capacity to help us hold the experts accountable for their inferences. This motivates the desire for a more precise philosophical conception, if possible, of what inferences count as warranted and how. What Mayo's book should certainly convince us of is the value of seeking to develop her programme further, and for that reason alone the book is recommended reading for all philosophers—not least those of the Bayesian denomination—concerned with the foundations of statistics.