What can psychology’s statistics reformers learn from the error-statistical severe testing perspective?

Brian Haig
University of Canterbury

Introduction

Deborah Mayo’s ground-breaking book, *Error and the growth of statistical knowledge* (1996), which won the Lakatos Prize for outstanding work in the philosophy of science, presented the first extensive formulation of her error-statistical perspective on statistical inference. Its novelty lay in the fact that it employed ideas in statistical science to shed light on philosophical problems to do with evidence and inference.

By contrast, Mayo’s just-published book, *Statistical inference as severe testing* (SIST) (2018), focuses on problems arising from statistical practice (“the statistics wars”), but endeavors to solve them by probing their foundations from the vantage points of philosophy of science, and philosophy of statistics. The “statistics wars” to which Mayo refers concern fundamental debates about the nature and foundations of statistical inference. These wars are longstanding and recurring. Today, they fuel the ongoing concern many sciences have with replication failures, questionable research practices, and the demand for an improvement of research integrity. By dealing with the vexed problems of current statistical practice, SIST is a valuable repository of ideas, insights, and solutions designed to help a broad readership deal with the current crisis in statistics.

Psychology has been prominent among a number of disciplines suggesting statistical reforms that are designed to improve replication and other research practices. Despite being at the forefront of these reforms, psychology has, I believe, ignored the philosophy of statistics at its cost. In this post, I want to briefly consider two major proposals in psychology’s methodological literature: The recommendation that psychology should employ the so-called “new statistics” in its research practice, and the alternative proposal that it should embrace Bayesian statistics. I do this from the vantage point of the error-statistical severe testing perspective, which, for my money, is the most coherent philosophy of statistics currently available. I have enumerated what I take to be its strengths elsewhere (Haig, 2017). Of course, not everyone will want to endorse this perspective, but it surely needs to be taken seriously by methodologists and researchers, in psychology and further afield. Before concluding, I will also offer some remarks about two interesting features of the conception of science adopted in SIST, along with a brief comment about the value of the philosophy of statistics.

The new statistics

For decades, numerous calls have been made for replacing tests of statistical significance with alternative statistical methods. The new statistics, a package deal comprising effect sizes, confidence intervals, and meta-analysis, is one reform movement that has been heavily promoted in psychological circles (Cumming, 2012; 2014) as a much needed successor to null hypothesis significance testing (NHST), which is deemed to be broken-backed. Eric Eich, the recent editor of *Psychological Science*, which is the flagship journal of the Association for Psychological Science, endorsed the use of the new statistics, wherever appropriate. The new statistics might be considered the Association's current quasi-official position on statistical inference, despite the appearance of a number of public criticisms of the approach by both frequentists and Bayesians. It is noteworthy that the principal advocates of the new statistics have not publicly replied to these criticisms. Although SIST does not directly address the new statistics movement, its suggestions for overcoming the statistics wars contain insights about statistics that that can be directly employed to mount a powerful challenge to that movement.

NHST The new statisticians recommend replacing NHST with their favored statistical methods by asserting that it has several major flaws. Prominent among them are the familiar claims that NHST encourages dichotomous thinking, and that it comprises an indefensible amalgam of the Fisherian and Neyman-Pearson schools of thought. However, neither of these features applies to the error-statistical understanding of NHST. The claim that we should abandon NHST because it leads to dichotomous thinking is unconvincing because it is levelled at the misuse of a statistical test that arises from a poor understanding of its foundations. Fisher himself explicitly cautioned against such thinking. Further, SIST makes clear that the common understanding of the amalgam that is NHST is not an amalgam of Fisher's and Neyman and Pearson's thinking on the matter, especially their mature thought. Further, the error-statistical outlook can accommodate both evidential and behavioural interpretations of NHST, respectively serving *probative* and *performance* goals, to use Mayo's suggestive terms. SIST urges us to move beyond the claim that NHST is an inchoate hybrid. Based on a close reading of the historical record, Mayo argues that Fisher and Neyman and Pearson should be interpreted as compatibilists, and that focusing on the vitriolic exchanges between Fisher and Neyman prevents one from seeing how their views dovetail. Importantly, Mayo formulates the error-statistical perspective on NHST by assembling insights from these founding fathers, and additional sources, into a coherent hybrid.

Tellingly, the recommendation of the new statisticians to abandon NHST, understood as an inchoate hybrid, commits the fallacy of the false dichotomy because there exist alternative defensible accounts of NHST. The error-statistical perspective is one of these. The neo-Fisherian outlook of Hurlbert and Lombardi (2009) is another (Haig, 2017).

Confidence intervals For the new statisticians, confidence intervals replace p-valued null hypothesis significance testing. Confidence intervals are said to be more informative, and more easily understood, than p values, as well as serving the important scientific goal of estimation, which is preferred to hypothesis testing. Both of these claims are open to challenge. Whether confidence intervals are more informative than statistical hypothesis tests in a way that matters will depend on the research goals being pursued. For example, p values might properly be used to get a useful initial gauge of whether an experimental effect occurs in a particular study, before one runs further studies and reports p values, supplementary confidence intervals and effect sizes. The claim that confidence intervals are more easily understood than p values is surprising, and is not borne out by the empirical evidence (e.g., Hoekstra, Morey, Rouder, & Wagenmakers, 2014). I will speak to the claim about the greater importance of estimation under the next heading.

There is a double irony in the fact that the new statisticians criticize NHST for encouraging simplistic dichotomous thinking: As already noted, such thinking is straightforwardly avoided by employing tests of statistical significance properly, whether or not one adopts the error-statistical perspective. For another, the adoption of standard frequentist confidence intervals in place of NHST forces the new statisticians to engage in dichotomous thinking of another kind: A parameter estimate is either inside, or outside, its confidence interval.

Error-statisticians have good reason for claiming that their reinterpretation of frequentist confidence intervals is superior to the standard view. The standard account of confidence intervals adopted by the new statisticians prespecifies a single confidence interval (a strong preference for 0.95 in their case). The single interval estimate corresponding to this level provides the basis for the inference that is drawn about the parameter values, depending on whether they fall inside or outside the interval. A limitation of this way of proceeding is that each of the values of a parameter in the interval are taken to have the same probative force, even though many will have been weakly tested. By contrast, the error-statistician draws inferences about each of the obtained values according to whether they are warranted, or not, at different severity levels, thus leading to a series of confidence intervals. Crucially, the different values will not have the same probative force. Clearly, this is a more nuanced and informative assessment of parameter estimates than that offered by the standard view. Details on the error-statistical conception of confidence intervals can be found in SIST (pp. 189-201), as well as Mayo and Spanos (2011) and Spanos (2014).

Assuming that the new statisticians want to adopt a sound frequentist conception of confidence intervals, they would improve their practice by moving to the superior error-statistical understanding of such intervals.

Estimation and hypothesis tests The new statisticians claim, controversially, that parameter estimation, rather than statistical hypothesis testing, leads to better science. Their preference for estimation leads Cumming (2012) to aver that typical questions in science are *what* questions (e.g., “What is the age of the earth?”, “What is the most likely sea-level rise by 2012?”). Explanatory *why* questions and *how* questions, the latter which usually ask for information about causal mechanisms, are not explicitly considered. However, why and how questions are just as important as what questions for science. They are the sort of questions that science seeks to answer when constructing and evaluating hypotheses and theories. By contrast, SIST makes clear that, with its error-statistical perspective, statistical inference can be employed to deal with both estimation and hypothesis testing problems. It also endorses the view that providing explanations of things is an important part of science.

Bayesian statistics

Unlike the field of statistics, the Bayesian outlook has taken an age to assert itself in psychology. However, a cadre of methodologists has recently advocated the use of Bayesian statistical methods as a superior alternative to the messy frequentist practice that dominates psychology’s research landscape (e.g., Dienes, 2011; Wagenmakers, 2007). These Bayesians criticize NHST, often advocate the use of Bayes factors for hypothesis testing, and rehearse a number of other well-known Bayesian objections to frequentist statistical practice. Of course, there are challenges for Bayesians from SIST, just as there are for the new statisticians. They also need to reckon with the coherent hybrid NHST produced by the error statisticians, and argue against it if they want to justify abandoning NHST; they need to rethink whether Bayes factors can provide strong tests of hypotheses without their ability to probe errors; and they should consider, among other challenges, Mayo’s critique of the Likelihood Principle, a principle to which they often appeal when critiquing frequentist statistics.

Contrasts with the error-statistical perspective One of the major achievements of SIST is that it provides a comprehensive critical evaluation of the major variants of Bayesian statistical thinking, including the default, pragmatic, and eclectic options within the Bayesian corpus. SIST contains many challenges for Bayesians to consider. Here, I want to note three basic features of Bayesian thinking, which are rejected by the error-statistical approach of SIST:

First, the error-statistical approach rejects the Bayesian insistence on characterizing the evidential relation between hypothesis and evidence in a universal and logical manner in terms of Bayes' theorem. By contrast, it formulates the relation in terms of the substantive and specific nature of the hypothesis and the evidence with regards to their origin, modeling, and analysis. This is a consequence of a strong commitment to a contextual approach to testing using the most appropriate frequentist methods available for the task at hand.

Second, the error-statistical philosophy also rejects the classical Bayesian commitment to the subjective nature of fathoming prior probabilities in favor of the more objective process of establishing error probabilities understood in frequentist terms. It also finds the turn to objective Bayesianism unsatisfactory, as SIST makes clear.

Third, the error-statistical outlook employs probabilities to measure how effectively *methods* facilitate the detection of error, and how those methods enable us to choose between alternative hypotheses. Orthodox Bayesians are not concerned with error probabilities at all. Instead, they use probabilities to measure *belief* in hypotheses or degrees of confirmation. It is for this reason that error-statisticians will say about Bayesian methods that, without supplementation with error probabilities, they are not capable of providing stringent tests of hypotheses.

The Bayesian remove from scientific practice Two additional features of the Bayesian focus on beliefs, which have been noted by philosophers of science, draw attention to their outlook on science. First, Kevin Kelly and Clark Glymour worry that "Bayesian methods assign numbers to answers instead of producing answers outright." (2004, p. 112) Mayo agrees that we should focus on the phenomena of interest, not the epiphenomena of degrees of belief. And second, Henry Kyburg is puzzled by the Bayesian's desire to "replace the fabric of science … with a vastly more complicated representation in which each statement of science is accompanied by its probability, for each of us." (1992, p.149) Kyburg's puzzlement prompts the question, Why should we be interested in each other's probabilities? This is a question raised by David Cox about prior probabilities, and noted in SIST. I think that these legitimate expressions of concern stem from the reluctance of many Bayesians to study science itself. This Bayesian remove from science contrasts markedly with SIST's direct engagement with science. Mayo is a philosopher of science as well as statistics, and has a keen eye for scientific practice. Given that contemporary philosophers of science tend to take scientific practice seriously, it comes as no surprise that, in SIST, Mayo brings it to the fore when dealing with statistics. Indeed, her error-statistical philosophy should be seen as a significant contribution to the so-called *new experimentalism*, with its strong focus, not just on experimental practice in science, but also on the role of statistics in such practice. Her discussion of the place of frequentist statistics in the discovery of the Higgs boson in particle physics is an instructive case in point.

Taken together, these just-mentioned points of difference between the Bayesian and error-statistical philosophies constitute a major challenge to Bayesian thinking in psychology, and elsewhere, that methodologists, statisticians, and researchers need to consider.

Bayesianism with error-statistical foundations One particularly important modern variant of Bayesian thinking, which receives attention in SIST, is the *falsificationist Bayesianism* of Andrew Gelman, which received its major formulation in Gelman and Shalizi (2013). Interestingly, Gelman regards his Bayesian philosophy as essentially error-statistical in nature – an intriguing claim, given the anti-Bayesian preferences of both Mayo and Gelman's co-author, Cosma Shalizi. Gelman's philosophy of Bayesian statistics is also significantly influenced by Popper's view that scientific propositions are to be submitted to repeated criticism in the form of strong empirical tests. For Gelman, best Bayesian statistical practice involves formulating models using Bayesian statistical methods, and then checking them through hypothetico-deductive attempts to falsify and modify those models.

Both the error-statistical and neo-Popperian Bayesian philosophies of statistics extend and modify Popper's conception of the hypothetico-deductive method, while at the same time offering alternatives to received views of statistical inference. The error-statistical philosophy injects into the hypothetico-deductive method an account of statistical induction that employs a panoply of frequentist statistical methods to detect and control for errors. For its part, the Bayesian alternative involves formulating models using Bayesian statistical methods, and then checking them through attempts to falsify and modify those models. This differs from the received philosophy of Bayesian statistical modeling, which is regarded as a formal inductive process.

From the wide-ranging evaluation in SIST of the major varieties of Bayesian statistical thought on offer, Mayo concludes that Bayesian statistics needs new foundations: In short, those provided by her error-statistical perspective. Gelman acknowledges that his falsificationist Bayesian philosophy is underdeveloped, so it will be interesting to see how its further development relates to Mayo's error-statistical perspective. It will also be interesting to see if Bayesian thinkers in psychology engage with Gelman's brand of Bayesian thinking. Despite the appearance of his work in a prominent psychology journal, they have yet to do so. However, Borsboom and Haig (2013) and Haig (2018), provide sympathetic critical evaluations of Gelman's philosophy of statistics. Mayo's treatment of Gelman's philosophy brings to notice the interesting point that she is willing to allow a decoupling of statistical outlooks and their traditional philosophical foundations in favour of different foundations, which are judged more appropriate.

SIST and the nature of science

Before concluding, I want to convey some sense of how SIST extends our understanding of the nature of science. I restrict my attention to its treatment of the process of falsification and the structure of modelling, before saying something about the value of philosophy for statistics.

Falsificationism Probably the best known account of scientific method is Karl Popper's falsificationist construal of the hypothetico-deductive method, understood as a general strategy of conjecture and refutation. Although it has been roundly criticised by philosophers of science, it is frequently cited with approval by scientists, including psychologists, although they seldom consider it in depth. One of the most important features of SIST is its presentation of a falsificationist view of scientific inquiry, with error statistics serving an indispensable role. From a sympathetic, but critical, reading of Popper, Mayo endorses his strategy of developing scientific knowledge by identifying and correcting errors through strong tests of scientific claims. Making good on Popper's lack of knowledge of statistics, Mayo shows how one can properly employ a range of, often familiar, error-statistical methods to implement her all-important severity requirement. Stated minimally, and informally, this requirement says "A claim is severely tested to the extent that it has been subjected to and passes a test that probably would have found flaws, were they present". (SIST, p. xii) Further, in marked contrast with Popper, who deemed deductive inference to be the only legitimate form of inference, Mayo's conception of falsification stresses the importance of inductive, or content-increasing, inference in science. We have here, then, a viable account of falsification, which goes well beyond Popper's account with its lack of operational detail about how to construct strong tests. It is worth noting that SIST presents the error-statistical stance as a constructive interpretation of Fisher's oft-cited remark that the null hypothesis is never proved, only possibly disproved.

A hierarchy of models Building on Patrick Suppes' (1962) insight that science employs a hierarchy of models that ranges from experimental experience to theory, Mayo's (1996) error-statistical philosophy adopts a framework in which three different types of models are interconnected and serve to structure error-statistical inquiry: Primary models, experimental models, and data models. Primary models break down a research question into a set of local hypotheses that can be investigated using reliable methods. Experimental models structure the particular models at hand, and serve to link primary models to data models. And, data models generate and model raw data, as well as checking whether the data satisfy the assumptions of the experimental models. It should be mentioned that the error-statistical approach has been extended to primary models and theories of a more global nature and, in SIST, also includes a consideration of experimental design and the analysis and generation of data.

Interestingly, this hierarchy of models employed in the error-statistical perspective exhibits a structure similar to the important three-fold distinction between data, empirical phenomena, and theory (Bogen & Woodward, 1988; Haig, 2014). This related pair of three-fold distinctions accords much better with scientific practice than the coarse-grained, data-theory/model distinction that is ubiquitous in scientific talk. As with error-statistical modeling, Bogen and Woodward show that local data analysis and statistical inference connect to substantive theories via intermediate claims about experimental models and empirical phenomena. The detection of phenomena, typically empirical regularities in psychology, makes heavy-duty use of statistical methods. In this regard, SIST strongly endorses Fisher's caution against taking a single small *p* value as an indicator of a genuine effect.

Although researchers and textbook writers in psychology correctly assume that rejection of the null hypothesis implies acceptance of the alternative hypothesis, they too often err in treating the alternative hypothesis as a research, or scientific, hypothesis rather than as a statistical hypothesis. Substantive knowledge of the domain in question is required in order to formulate a scientific hypothesis that corresponds to the alternative statistical hypothesis. SIST explicitly forbids concluding that statistical significance implies substantive significance. In addition, it counsels that one should knowingly moving back and forth between the two types of significance in respect of the hierarchy of models just mentioned.

The philosophy of statistics A heartening attitude that comes through in SIST is the firm belief that a philosophy of statistics is an important part of statistical thinking. This contrasts markedly with much of statistical theory, and most of statistical practice. Through precept and practice, Mayo's work makes clear that philosophy can have a direct impact on statistical practice. Given that statisticians operate with an implicit philosophy, whether they know it or not, it is better that they avail themselves of an explicitly thought-out philosophy that serves practice in useful ways. SIST provides a strong philosophical aid to an improved understanding and use of tests of statistical significance and other frequentist statistical methods. More than this, SIST is a book on scientific methodology in the proper sense of the term. Methodology is the interdisciplinary field that draws from disciplines that include statistics, philosophy of science, history of science, as well as indigenous contributions from the various substantive disciplines. As such, it is the key to a proper understanding of statistical and scientific methods. Mayo's latest book is deeply informed about the philosophy, history, and theory of statistics, as well as statistical practice. It is for this reason that it is able to position the reader to go beyond the statistics wars.

Conclusion

SIST provides researchers and methodologists with a distinctive perspective on statistical inference. Mayo's Popper-inspired emphasis on strong tests is a welcome antidote to the widespread practice of weak hypothesis testing in psychological research that Paul Meehl labored in vain to correct. Psychologists can still learn much of value from Meehl about the nature of good statistical practice in science, but Mayo's SIST contains understandings and recommendations about sound statistical practices in science that will take them further, if they have a mind to do so. In this post, I have invited the new statisticians and Bayesians to address the challenges to their outlooks on statistics that the error-statistical perspective provides.

References

Bogen, J., & Woodward, J. (1988). Saving the phenomena. *Philosophical Review, 97*, 303-352.

Borsboom, D., & Haig, B. D. (2013). How to practice Bayesian statistics outside the Bayesian church: What philosophy for Bayesian statistical modelling? *British Journal of Mathematical and Statistical Psychology, 66,* 39-44.

Cumming, G. (2012). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis.* New York, NY: Routledge.

Cumming, G. (2014). The new statistics: Why and how. *Psychological Science, 25*, 7-29.

Dienes, Z. (2011). Bayesian versus orthodox statistics: Which side are you on? *Perspectives on Psychological Science, 6*, 274-290.

Gelman, A., & Shalizi, C. R. (2013). Philosophy and the practice of Bayesian statistics. *British Journal of Mathematical and Statistical Psychology, 66*, 8-38.

Haig, B. D. (2014). *Investigating the psychological world: Scientific method in the behavioral sciences.* Cambridge, MA: MIT Press.

Haig, B. D. (2017). Tests of statistical significance made sound. *Educational and Psychological Measurement, 77,* 489-506.

Haig. B. D. (2018). *The philosophy of quantitative methods.* New York, NY: Oxford University Press.

Hoekstra, R., Morey, R. D., Rouder, J. N., & Wagenmakers, E-J. (2014). Robust misinterpretation of confidence intervals. *Psychonomic Bulletin & Review, 21*, 1157-1164.

Hurlbert, S. H., & Lombardi, C. M. (2009). Final collapse of the Neyman-Pearson decision theoretic framework and rise of the neoFisherian. *Annales Zoologici Fennici, 46*, 311-349.

Kelly, K.T., & Glymour, C. (2004). Why probability does not capture the logic of scientific justification. In C. Hitchcock (Ed.), *Contemporary debates in philosophy of science* (pp.

94-114). Malden, MA: Blackwell.

Kyburg, H. (1992). The scope of Bayesian reasoning. In *PSA:Proceedings of the biennial meeting of the Philosophy of Science Association* (Vol. 2, pp.139-152). Chicago, IL: University of Chicago Press.

Mayo, D. G (1996). *Error and the growth of experimental knowledge*. Chicago, IL: University of Chicago Press.

Mayo, D. G. (2018*). Statistical inference as severe testing: How to get beyond the statistics wars.* New York, NY: Cambridge University Press.

Mayo, D. G., & Spanos, A. (2011). Error statistics. In P. S. Bandyopadhyay & M. R. Forster (Eds.*), Handbook of philosophy of science: Vol. 7. Philosophy of statistics* (pp.153-198). Amsterdam, the Netherlands: Elsevier.

Spanos, A. (2014). Recurring controversies about *P* values and confidence intervals revisited. *Ecology, 95*, 645-651.

Suppes, P. (1969). Models of data. In E. Nagel, P. Suppes, & A. Tarski (Eds.), *Logic, methodology, and philosophy of science: Proceedings of the 1960 International Congress* (pp. 252-261). Stanford, CA: Stanford University Press.

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of *p* values. *Psychonomic Bulletin & Review, 14*, 779-804.