Synthese (2023) 201:120
https://doi.org/10.1007/s11229-023-04128-z

INTRODUCTION

Introduction to recent issues in philosophy of statistics: evidence, testing, and applications

Molly Kao[1] · Deborah G. Mayo[2] · Elay Shech[3]

Published online: 23 March 2023

Statistics play an essential role in an extremely wide range of human reasoning. From theorizing in the physical and social sciences to determining evidential standards in legal contexts, statistical methods are ubiquitous, and thus various questions about their application inevitably arise. As tools for making inferences that go beyond a given set of data, they are inherently a means of reasoning ampliatively, and so it is unsurprising that philosophers interested in the notions of evidence and inductive inference have been concerned to utilize statistical frameworks to further our understanding of these topics. However, the field of statistics has long been the subject of heated philosophical controversy. Given that a central goal for philosophers of science is to help resolve problems about evidence and inference in scientific practice, it is important that they be involved in current debates in statistics and data science. The purpose of this topical collection is to promote such philosophical interaction. We present a cross-section of these subjects, written by scholars from a variety of fields in order to explore issues in philosophy of statistics from different perspectives.

The articles in this collection can be divided into roughly two categories. The first group contain articles by Mayo and Hand (2022), Radzvilas et al. (2021), Rubin (2021), and Spanos (2021), and are concerned mainly with foundational issues in philosophy of statistics. In particular, the authors address questions on the procedure of statistical significance testing and its accompanying concepts of p-values and significance thresholds, Bayesian versus frequentist ("classical") statistics, and Ber-

✉ Elay Shech
eshech@auburn.edu

Molly Kao
molly.kao@umontreal.ca

Deborah G. Mayo
mayod@vt.edu

[1] Université de Montréal, Montréal, QC, Canada

[2] Virginia Polytechnic Institute and State University, Blacksburg, VA, USA

[3] Auburn University, Auburn, USA

noulli's golden theorem. The second group of articles reflects on the uses of statistics in various fields, including Dethier (2022) on climate science, Di Bello (2021) on law, Gardiner and Zaharatos (2022) on epistemology, Park (2022) on medicine, and Watson (2022) on machine learning. In what follows, we offer brief summaries of the contributions to this topical collection.[1]

Two of the articles address the fraught topic of significance testing. While there has been much recent discussion questioning the use of p-values, Deborah Mayo and David Hand (2022) defend their relevance in numerous contexts, arguing that calling for an end to the use of P-value thresholds—or even the term "significance"-- only exacerbates biasing selection effects (data-dredging, optional stopping). They argue that the central criticisms of statistical significance tests arise from misunderstanding and misusing the statistical tools. They offer a reinterpretation that prevents such misinterpretations, based on the notion of severe testing. Their goal is not to point to problems of approaches with different goals, but to argue that all of the proposed alternatives put forward to replace significance testing are incapable of accomplishing the fundamental job of controlling error probabilities in distinguishing genuine from spurious effects, which is the key task of statistical significance tests. On the other hand, they do not claim that such tests are *the* correct tool for evidence, but that they are essential for this one central task. Most importantly, they emphasize why current criticisms of the use of p-values are actually damaging scientific practice; namely, that they assume philosophies of inference that are at odds with the underlying error statistics. According to Mayo and Hand, these criticisms have encouraged a groupthink that vilifies important methods without putting anything in their place that can do the intended job.

Mark Rubin's contribution (2021) also concerns statistical significance testing. However, he focuses on scientists' practice of adjusting the significance threshold (or the alpha level) during null hypothesis testing. It is generally accepted that multiple tests are more likely to yield spurious results than a test of a single hypothesis, due to the increased chance of random sampling error giving rise to at least one among a number of possible correlations. As a result, the alpha level is often adjusted to compensate for this risk. Rubin focuses on an under discussed feature of this practice, namely, the conditions in which this adjustment is appropriate or inappropriate. To this end, he draws a distinction between three types of multiple testing: disjunction testing, conjunction testing, and individual testing. The author explains the distinction between these types of testing, and he argues that alpha adjustment is only appropriate in the case of disjunction testing, in which at least one test result must be significant in order to reject the associated joint null hypothesis. Finally, Rubin elaborates on the distinction between these three types of testing, specifying when each type is warranted, looking at contexts such as tests of large families of hypotheses, as well as smaller groups of hypotheses as would be tested in a multiway ANOVA. Based on this distinction and his subsequent analysis, the author concludes that researchers

[1] This topical collection grew from a Summer Seminar in Philosophy of Statistics (July 28-August 11, 2019) at Virginia Tech run by Deborah Mayo and Aris Spanos (funded by Mayo-Chatfield private E.R.R.O.R. Fund and Virginia Tech).

should not automatically assume that alpha adjustment is necessary during multiple testing, but rather first consider what kind of testing is appropriate.

Bernoulli's "golden theorem" states that, as the number of observations n from an Independent and Identically Distributed Bernoulli sequence increases to infinity, the probability of any prespecified difference (however small) between the observed relative frequency of success and the corresponding probability goes to one as n goes to infinity (∞). However, there have been many debates regarding the significance and the interpretation of this theorem for $n < \infty$. In his article, Aris Spanos (2021) examines this theorem from the modern perspective of model-based frequentist inference. Spanos argues that several widely-accepted claims relating to the golden theorem and frequentist inference are either misleading or erroneous. These clarifications enable Spanos to address several foundational problems relating to frequentist statistical inference, including (i) Bernoulli's alleged swindle being an instance of an unwarranted claim for point and effect size estimates, (ii) frequentist error probabilities not being conditional on hypotheses, (iii) the direct vs. inverse inference problem being a contrived issue, which (iv) has been devised to link unobservable parameters to data in Bayes' rule without defining explicitly their joint distribution. He argues that a key, but neglected, contributor to the untrustworthiness of empirical evidence problem is statistical misspecification: invalid probabilistic assumptions imposed (explicitly or implicitly) on one's data (p. 13,969).

Finally, Mantas Radzvilas, William Peden and Francesco De Pretis take a different approach to the question of competing statistical methodologies, constructing a simulation to investigate different methodologies' performance. Given that the competing methodologies all have intuitively attractive properties in the long run, the authors focus on performance in the short to medium run. To do so, they focus on a straightforward decision problem based around tossing a coin with unknown bias and then placing bets. The simulation includes four "players", inspired by Bayesian statistics, frequentist statistics, Jon Williamson's version of Objective Bayesianism, and a player who simply extrapolates from observed frequencies to general frequencies (used as a baseline sample). Their results show no systematic difference in performance between the Bayesian and frequentist players, provided the Bayesian uses a flat prior and the frequentist uses a low confidence level. However, the frequentist and Williamsonian players performed poorly with high confidence levels, while the Bayesian was surprisingly harmed by biased priors. This contribution thus provides a justification of the idea that all three methodologies should be taken seriously by philosophers and practitioners of statistics, apart from the various conceptual arguments. At the same time, the simulation shows certain limitations of each of the strategies in a relatively concrete way.

While the foregoing articles are focused on issues intrinsic to statistics and statistical inference, the following turn to the analysis of these methods in context. Corey Dethier's (2022) contribution concerns the use of statistical methods in climate science. Specifically, it is common for climate scientists to treat results extracted from ensembles of climate models as if they are data generated by sampling a population by applying statistical analysis to said results. What justifies applying statistics in such cases? According to Dethier, in the same way that other statistical inferences are justified by ensuring that the assumptions constituting one's statistical model are

reliable, justified inferences based on ensembles of climate models must be based on a model that accurately represents the probabilistic relationship between ensemble-generated results/data and target. Whether applying statistics to ensembles of climate models produces reliable results will thus "vary from case to case" so that "general criticisms leveled at the very idea of applying statistics to the data generated by ensembles of models are misguided" (Dethier, 2022, 51–52).

In the context of law and litigation, when the case against the defendant rests in whole or in part on statistical evidence, it has been noted that statistical evidence may lack desirable epistemic and non-epistemic properties essential for a fair trial. In his contribution, reflecting on when statistical evidence is meant to establish a defendant's guilt, Marcello Di Bello (2021) proposes a theory of when to reject such evidence. He argues that if "there is a mismatch between the specificity of the evidence and the expected specificity of the accusation…" then evidence "should be considered insufficient to sustain a conviction unless it is adequately supplemented by other, more specific evidence" (12,252). By "specificity," for example, Di Bello intends the level of informativeness of a narrative such as the extent to which all relevant questions concerning a case may be answered. He gives the following example: "A testimony that says 'I saw the defendant run away from the crime scene' is not as specific as 'I saw the defendant's face while he was stabbing the victim in the chest'" (12,258). Moreover, Di Bello suggest that part of what drives disagreements between philosophers and legal scholars about the role of statistical evidence is the fact that said specificity is context dependent and thus varies on a case-by-case basis.

In the epistemological literature it is common to characterize knowledge as justified true belief that also meets other conditions, where "safety" and "sensitivity" are two such (typically rival) conditions. For instance, the safety condition is satisfied when a subject's belief could not have easily been false, and the sensitivity condition is satisfied when, if a subject's belief were false, said subject would not hold such a belief. In their contribution, Georgi Gardiner and Brian Zaharatos (2022) offer a unified account of the conditions of safety and sensitivity such that the two are symbiotic. They then continue to recast said account in terms of Deborah Mayo's (2018) severe testing condition. In doing so, they motivate further research into and forge fruitful connections between philosophy of statistics and contemporary epistemology. As they note, Mayo's severe testing framework "should be discussed within mainstream contemporary epistemology because it mirrors, and goes beyond, recent developments in modal epistemology" (Gardiner & Zaharatos, 2022, 369).

Clinical trials in medical research are another area where statistical notions play a crucial inferential role. John H. Park (2022) examines some of these uses, identifying places where misinterpretations of the results of clinical studies have arisen in analysis of a number of concrete examples. Park begins by describing some of the sociological features that may cause biases that contribute to faulty inferences. The author shows that despite the presence of preregistration criteria, the discussion of results may ultimately disregard some of their implications. Potential problems thus go beyond the mere practice of preregistration, but also include non adherence to the preregistered protocols at the end of the trial process and interpretation of results. To temper some of these problems, he suggests that an application of severity testing should be used, namely, by instituting tests that are "As Severe as Reasonably

Possible" (ASARP). He argues that the inclusion of ASARP testing in the protocol would make it clearer to practitioners that there are multiple features to be taken into account in statistical inferences, in essence arguing for a more holistic understanding of statistical methods in medical analyses.

Advancements in machine learning (ML), especially using deep learning techniques, have enabled algorithms to compete with human benchmarks on various tasks. ML models have become ubiquitous in both the private and public sectors of modern society. However, such models are often opaque and black-boxed in the sense that the ever-increasing complexity of these algorithms effectively shields humans from understanding how and why they work. Accordingly, a subdiscipline of computer and data science known as "interpretable machine learning" (IML) or "explainable artificial intelligence" has emerged. The idea is that IML algorithms can help identify the main factors powering ML-based statistical inference. In his contribution, David S. Watson (2022) argues that there are three conceptual challenges facing IML, namely, such algorithms "are plagued by (1) ambiguity with respect to their true target; (2) a disregard for error rates and severe testing; and (3) an emphasis on product over process" (65). He grounds the discussion in epistemology and philosophy of science, cautioning that greater care must be taken to understand the conceptual foundations of IML so that future work in this area doesn't repeat the same mistakes.

References

Dethier, C. (2022). When is an ensemble like a sample? "Model-based" inferences in climate modeling. *Synthese*, *200*, 52.

Di Bello, M. (2021). When statistical evidence is not specific enough. *Synthese*, *199*, 12251–12269.

Gardiner, G., & Zaharatos, B. (2022). The safe, the sensitive, and the severely tested: A unified account. *Synthese*, *200*, 369.

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

Mayo, D. G., & Hand, D. (2022). Statistical significance and its critics: Practicing damaging science, or damaging scientific practice? *Synthese*, *200*, 220.

Park, J. H. (2022). Current issues in medical epistemology and statistics: A view from the frontline of medicine. *Synthese*, *200*, 417.

Radzvilas, M., Peden, W., & De Pretis, F. (2021). A battle in the Statistics Wars: A simulation-based comparison of bayesian, Frequentist and Williamsonian methodologies. *Synthese*, *199*, 13689–13748.

Rubin, M. (2021). When to adjust alpha during multiple testing: A consideration of disjunction, conjunction, and individual testing. *Synthese*, *199*, 10969–11000.

Spanos, A. (2021). Bernoulli's golden theorem in retrospect: Error probabilities and trustworthy evidence. *Synthese*, *199*, 13949–13976.

Watson, D. S. (2022). Conceptual challenges for interpretable machine learning. *Synthese*, *200*, 65.