C. Hennig on Mayo's SIST

**Christian Hennig** (**12 April 2019 Review of SIST**) on A. Gelman's blog, *Statistical Modeling, Causal. Inference, and Social Science*.

Hennig, a statistician and my collaborator on the Beyond Subjective and Objective paper, send in *two* reviews of Mayo's book.

Here are his general comments:

What I like about Deborah Mayo's "Statistical Inference as Severe Testing"

Before I start to list what I like about "Statistical Inference as Severe Testing". I should say that I don't agree with everything in the book. In particular, as a constructivist I am skeptical about the use of terms like "objectivity", "reality" and "truth" in the book, and I think that Mayo's own approach may not be able to deliver everything that people may come to believe it could, from reading the book (although Mayo could argue that overly high expectations could be avoided by reading carefully).

So now, what do I like about it?

1) I agree with the broad concept of severity and severe testing. In order to have evidence for a claim, it has to be tested in ways that would reject the claim with high probability if it indeed were false. I also think that it makes a lot of sense to start a philosophy of statistics and a critical discussion of statistical methods and reasoning from this requirement. Furthermore, throughout the book Mayo consistently argues from this position, which makes the different "Excursions" fit well together and add up to a consistent whole.

2) I get a lot out of the discussion of the philosophical background of scientific inquiry, of induction, probabilism, falsification and corroboration, and their connection to statistical inference. I think that it makes sense to connect Popper's philosophy to significance tests in the way Mayo does (without necessarily claiming that this is the only possible way to do it), and I think that her arguments are broadly convincing at least if I take a realist perspective of science (which as a constructivist I can do temporarily while keeping the general reservation that this is about a specific construction of reality which I wouldn't grant absolute authority).

3) I think that Mayo does by and large a good job listing much of the criticism that has been raised in the literature against significance testing, and she deals with it well. Partly she criticises bad uses of significance testing herself by referring to the severity requirement, but she also defends a well understood use in a more general philosophical framework of testing scientific theories and claims in a piecemeal manner. I find this largely convincing, conceding that there is a lot of detail and that I may find myself in agreement with the occasional objection against the odd one of her arguments.

4) The same holds for her comprehensive discussion of Bayesian/probabilist foundations in Excursion 6. I think that she elaborates issues and inconsistencies in the current use of Bayesian reasoning very well, maybe with the odd exception.

5) I am in full agreement with Mayo's position that when using probability modelling, it is important to be clear about the meaning of the computed probabilities. Agreement in numbers between different "camps" isn't worth anything if the numbers mean different things. A problem with some positions that are sold as "pragmatic" these days is that often not enough care is put into interpreting what the results mean, or even deciding in advance what kind of interpretation is desired.

6) As mentioned above, I'm rather skeptical about the concept of objectivity and about an all too realist interpretation of statistical models. I think that in Excursion 4 Mayo manages to explain in a clear manner what her claims of "objectivity" actually mean, and she also appreciates more clearly than before the limits of formal models and their distance to "reality", including some valuable thoughts on what this means for model checking and arguments from models.

So overall it was a very good experience to read her book, and I think that it is a very valuable addition to the literature on foundations of statistics.

Hennig also sent some specific discussion of one part of the book:

1 Introduction

This text discusses parts of Excursion 4 of Mayo (2018) titled "Objectivity and Auditing". This starts with the section title "The myth of 'The myth of objectivity'". Mayo advertises objectivity in science as central and as achievable.

In contrast, in Gelman and Hennig (2017) we write: "We argue that the words 'objective' and 'subjective' in statistics discourse are used in a mostly unhelpful way, and we propose to replace each of them with broader collections of attributes." I will here outline agreement and disagreement that I have with Mayo's Excursion 4, and raise some issues that I think require more research and discussion.

2 Pushback and objectivity

The second paragraph of Excursion 4 states in bold letters: "The Key Is Getting Pushback", and this is the major source of agreement between Mayo's and my views (*). I call myself a constructivist, and this is about acknowledging the impact of human perception, action, and communication on our world-views, see Hennig (2010). However, it is an almost universal experience that we cannot construct our perceived reality as we wish, because we experience "pushback" from what we perceive as "the world outside". Science is about allowing us to deal with this pushback in stable ways that are open to consensus. A major ingredient of such science is the "Correspondence (of scientific claims) to observable reality", and in particular "Clear conditions for reproduction, testing and falsification", listed as "Virtue 4/4(b)" in Gelman and Hennig (2017). Consequently, there is no disagreement with much of the views and arguments in Excursion 4 (and the rest of the book). I actually believe that there is no contradiction between constructivism understood in this way and Chang's (2012) "active scientific realism" that asks for action in order to find out about "resistance from reality", or in other words, experimenting, experiencing and learning from error.

If what is called "objectivity" in Mayo's book were the generally agreed meaning of the term, I would probably not have a problem with it. However, there is a plethora of meanings of "objectivity" around, and on top of that the term is often used as a sales pitch by scientists in order to lend authority to findings or methods and often even to prevent them from being questioned. Philosophers understand that this is a problem but are mostly eager to claim the term anyway; I have attended conferences on philosophy of science and heard a good number of talks, some better, some worse, with messages of the kind "objectivity as understood by XYZ doesn't work, but here is my own interpretation that fixes it". Calling frequentist probabilities "objective" because they refer to the outside world rather than epsitemic states, and calling a Bayesian approach "objective" because priors are chosen by general principles rather than personal beliefs are in isolation also legitimate meanings of "objectivity", but these two and Mayo's and many others (see also the Appendix of Gelman and Hennig, 2017) differ. The use of "objectivity" in public and scientific discourse is a big muddle, and I don't think this will change as a consequence of Mayo's work. I prefer stating what we want to achieve more precisely using less loaded terms, which I think Mayo has achieved well not by calling her approach "objective" but rather by explaining in detail what she means by that.

## 3. Trust in models?

In the remainder, I will highlight some limitations of Mayo's "objectivity" that are mainly connected to Tour IV on objectivity, model checking and whether it makes sense to say that "all models are false". Error control is central for Mayo's objectivity, and this relies on error probabilities derived from probability models. If we want to rely on these error probabilities, we need to trust the models, and, very appropriately, Mayo devotes Tour IV to this issue. She concedes that all models are false, but states that this is rather trivial, and what is really relevant when we use statistical models for learning from data is rather whether the models are adequate for the problem we want to solve. Furthermore, model assumptions can be tested and it is crucial to do so, which, as follows from what was stated before, does not mean to test whether they are really true but rather whether they are violated in ways that would destroy the adequacy of the model for the problem. So far I can agree. However, I see some difficulties that are not addressed in the book, and mostly not elsewhere either. Here is a list.

### 3.1. Adaptation of model checking to the problem of interest

As all models are false, it is not too difficult to find model assumptions that are violated but don't matter, or at least don't matter in most situations. The standard example would be the use of continuous distributions to approximate distributions of essentially discrete measurements. What does it mean to say that a violation of a model assumption doesn't matter? This is not so easy to specify, and not much about this can be found in Mayo's book or in the general literature. Surely it has to depend on what exactly the problem of interest is. A simple example would be to say that we are interested in statements about the mean of a discrete distribution, and then to show that estimation or tests of the mean are very little affected if a certain continuous approximation is used. This is reassuring, and certain other issues could be dealt with in this way, but one can ask harder questions. If we approximate a slightly skew distribution by a (unimodal) symmetric one, are we really interested in the mean, the median, or the mode, which for a symmetric distribution would be the same but for the skew distribution to be approximated would differ? Any frequentist distribution

is an idealisation, so do we first need to show that it is fine to approximate a discrete non-distribution by a discrete distribution before worrying whether the discrete distribution can be approximated by a continuous one? (And how could we show that?) And so on.

3.2. Severity of model misspecification tests

Following the logic of Mayo (2018), misspecification tests need to be severe in ordert to fulfill their purpose; otherwise data could pass a misspecification test that would be of little help ruling out problematic model deviations. I'm not sure whether there are any results of this kind, be it in Mayo's work or elsewhere. I imagine that if the alternative is parametric (for example testing independence against a standard time series model) severity can occasionally be computed easily, but for most model misspecification tests it will be a hard problem.

3.3. Identifiability issues, and ruling out models by other means than testing

Not all statistical models can be distinguished by data. For example, even with arbitrarily large amounts of data only lower bounds of the number of modes can be estimated; an assumption of unimodality can strictly not be tested (Donoho 1988). Worse, only regular but not general patterns of dependence can be distinguished from independence by data; any non-i.i.d. pattern can be explained by either dependence or non-identity of distributions, and telling these apart requires constraints on dependence and non-identity structures that can itself not be tested on the data (in the example given in 4.11 of Mayo, 2018, all tests discover specific regular alternatives to the model assumption). Given that this is so, the question arises on which grounds we can rule out irregular patterns (about the simplest and most silly one is "observations depend in such a way that every observation determines the next one to be exactly what it was observed to be") by other means than data inspection and testing. Such models are probably useless, however if they were true, they would destroy any attempt to find "true" or even approximately true error probabilities.

3.4. Robustness against what cannot be ruled out

The above implies that certain deviations from the model assumptions cannot be ruled out, and then one can ask: How robust is the substantial conclusion that is drawn from the data against models different from the nominal one, which could not be ruled out by misspecification testing, and how robust are error probabilities? The approaches of standard robust statistics probably have something to contribute in this respect (e.g., Hampel et al., 1986), although their starting point is usually different from "what is left after misspecification testing". This will depend, as everything, on the formulation of the "problem of interest", which needs to be defined not only in terms of the nominal parametric model but also in terms of the other models that could not be rules out.

3.5. The effect of preliminary model checking on model-based inference

Mayo is correctly concerned about biasing effects of model selection on inference. Deciding what model to use based on misspecification tests is some kind of model selection, so it may bias inference that is made in case of passing misspecification tests. One way of stating the problem is to realise that in most cases the assumed model conditionally on having passed a misspecification test

does no longer hold. I have called this the "goodness-of-fit paradox" (Hennig, 2007); the issue has been mentioned elsewhere in the literature. Mayo has argued that this is not a problem, and this is in a well defined sense true (meaning that error probabilities derived from the nominal model are not affected by conditioning on passing a misspecification test) if misspecification tests are indeed "independent of (or orthogonal to) the primary question at hand" (Mayo 2018, p. 319). The problem is that for the vast majority of misspecification tests independence/orthogonality does not hold, at least not precisely. So the actual effect of misspecification testing on model-based inference is a matter that requires to be investigated on a case-by-case basis. Some work of this kind has been done or is currently done; results are not always positive (an early example is Easterling and Anderson 1978).

4 Conclusion

The issues listed in Section 3 are in my view important and worthy of investigation. Such investigation has already been done to some extent, but there are many open problems. I believe that some of these can be solved, some are very hard, and some are impossible to solve or may lead to negative results (particularly connected to lack of identifiability). However, I don't think that these issues invalidate Mayo's approach and arguments; I expect at least the issues that cannot be solved to affect in one way or another any alternative approach. My case is just that methodology that is "objective" according to Mayo comes with limitations that may be incompatible with some other peoples' ideas of what "objectivity" should mean (in which sense it is in good company though), and that the falsity of models has some more cumbersome implications than Mayo's book could make the reader believe.

(*) There is surely a strong connection between what I call "my" view here with the collaborative position in Gelman and Hennig (2017), but as I write the present text on my own, I will refer to "my" position here and let Andrew Gelman speak for himself.

References:

Chang, H. (2012) Is Water H2O? Evidence, Realism and Pluralism. Dordrecht: Springer.

Donoho, D. (1988) One-Sided Inference about Functionals of a Density. Annals of Statistics 16, 1390-1420.

Easterling, R. G. and Anderson, H.E. (1978) The effect of preliminary normality goodness of fit tests on subsequent inference. Journal of Statistical Computation and Simulation 8, 1-11.

Gelman, A. and Hennig, C. (2017) Beyond subjective and objective in statistics (with discussion). Journal of the Royal Statistical Society, Series A 180, 967–1033.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J. and Stahel, W. A. (1986) Robust statistics. New York: Wiley.

Hennig, C. (2010) Mathematical models and reality: a constructivist perspective. Foundations of Science 15, 29–48.

Hennig, C. (2007) Falsification of propensity models by statistical tests and the goodness-of-fit paradox. Philosophia Mathematica 15, 166-192.

Mayo, D. G. (2018) Statistical Inference as Severe Testing. Cambridge University Press.