Thoughts on “Statistical Inference as Severe Testing” by Deborah Mayo

Art B. Owen
Stanford University

January 2019

Abstract

Andrew Gelman asked me to contribute some thoughts on the book by Deborah Mayo. I'm not sure what the venue will be, but here are my thoughts. They are not written as formally as an article with references.

That book is pretty long. Here are some thoughts and ruminations written down as I read parts of it. I jumped from place to place not expecting to have time to read it all. Some of my comments are musings not directly based on parts I've read but brought up by the topics in the book and the reproducibility crisis.

For background on me, I think p-values are ok but one must understand their limitations. They get treated as if a small value clinches an argument and ends the discussion. Instead, a small value means that a completely null explanation is untenable, but it remains open whether some specific alternative that the user has in mind is the right reason for rejecting the null. The real reason might be a much desired causal outcome, an unmeasured variable, the multiplicity of testing, or some other model error such as correlations that were not properly accounted for or even non-Gaussianity. I liked Mayo's reminders that Fisher did not consider a single p-value to be decisive.

More background: my favorite statistical tools are scatterplots and confidence intervals.

I thank Jessica Hwang for taking an early look at this. I own any flaws in it.

Power and severity

There is an emphasis throughout on the importance of severe testing. It has long been known that a test that fails to reject H_0 is not very conclusive if it had low power to reject H_0. So I wondered whether there was anything more to the severity idea than that. After some searching I found on page 343 a description of how the severity idea differs from the power notion.

Suppose that one tests $H_0 : \mu = \mu_0$ versus $H_1 : \mu > \mu_0$ with a one tailed test designed to have specific power in case that $\mu = \mu_1$ for some $\mu_1 > \mu_0$. I believe that one tailed tests are more often than not a bad practice. The justifications are usually that only one alternative direction is possible or that only one alternative direction is consequential, and I think the uncertainty around those statements will not be negligible compared to the nominal level of the test. Nevertheless, for purposes of discussion, let a one tailed test be based on random data X and a statistic $d(X)$. It rejects H_0 at level α if and only if $d(X) \geqslant c_\alpha$.

A post-hoc power analysis looks at $\Pr(d(X) \geqslant c_\alpha; \mu_1)$. A severity analysis looks at $\Pr(d(X) \geqslant d(x); \mu_1)$ where x is the observed value of X. If H_0 was not rejected then $d(x) < c_\alpha$ and so

$$\Pr(d(X) \geqslant d(x); \mu_1) \geqslant \Pr(d(X) \geqslant c_\alpha; \mu_1). \tag{1}$$

The severity number comes out higher than the power number.

Mayo considers using power and severity as a post-hoc way to infer about whether $\mu < \mu_1$. Here are those two choices followed by a third that I prefer:

Power analysis: If $\Pr(d(X) \geqslant c_\alpha; \mu_1)$ is high and H_0 was not rejected, it indicates evidence that $\mu \leqslant \mu_1$.

Severity analysis: If $\Pr(d(X) \geqslant d(x); \mu_1)$ is high and H_0 was not rejected, it indicates evidence that $\mu \leqslant \mu_1$.

Confidence analysis: If the confidence interval for μ is contained within $(-\infty, \mu_1]$, it indicates evidence that $\mu \leqslant \mu_1$.

Equation (1) makes it clear that the severity analysis will find evidence that $\mu \leqslant \mu_1$ whenever the power analysis does.

Let's double check the directionality of the tests and confidence interval. If a one tailed test rejects $H_0 : \mu = \mu_0$ in favor of $H_0 : \mu > \mu_0$ then it will ordinarily also reject $H_0 : \mu = \mu_{-1}$ too for any $\mu_{-1} < \mu_0$. Then, inverting our hypothetical one tailed test will give a one sided confidence interval from $-\infty$ up to some highest value, so it could well be a subset of $(-\infty, \mu_1]$.

I don't see any advantage to severity (or to posterior power) over the confidence interval, if one is looking for evidence that $\mu \leqslant \mu_1$. One could replace the confidence interval by a posterior credible interval where that suits the problem and the user.

To make the case that severity is better than confidence, it would be necessary to explain why a value of μ_1 that is inside a confidence interval but fails a severity test should be considered implausible, and similarly, why a value of μ_1 that lies outside of the confidence interval, should nonetheless be taken as plausible if it gets a low severity value. If it can be proved that one of these outcomes is impossible then it would be enough to explain why severity is better for the other one.

The idea of estimating power post-hoc has been criticized as unnecessary. I think that it might be useful in explaining a failure to reject H_0 as the sample size being too small. A recent blog post by Andrew Gelman described how it is extremely hard to measure power post hoc because there is too much uncertainty about the effect size. Then, even if you want it, you probably cannot reliably get it. I think severity is likely to be in the same boat.

One null and two alternative hypotheses, and multi-Bayes

I liked the discussion of a remark attibuted to Senn, that when $H_0 : \mu = \mu_0$ is rejected in favor of $H_A : \mu > \mu_0$ in a test with good power when $\mu \geqslant \mu_1$, this is of course not evidence that $\mu \geqslant \mu_1$. Nobody should think it was, and I have never encountered somebody who does, but given how slippery it is to connect inferences to real world problems, some people might. The usual description of μ_1 is that of an effect so large that we would regret not detecting it. Rejecting H_0 lets us infer that $\mu > \mu_0$. If any improvement over μ_0 is enough to justify some action, then rejecting H_0 gives us confidence of having done no harm while the power calculation gives us some assurance that we won't miss a great benefit. Things are more complicated if some values $\mu \in (\mu_0, \mu_1)$ are not really better than μ_0. For instance, the benefits expected by acting as if the alternative were true could be proportional to $\mu - \mu_0$ minus some kind of switching cost. Then μ has to be enough larger than μ_0 to justify a change. Meehl wrote about this point and maybe he was not the first.

One potential remedy is to construct a test so that with high probability, the lower limit of the confidence interval is at least μ_1. This requires a new moving part: the value $\mu_2 > \mu_1$ of μ at which to make this computation. It cannot be μ_1, because if $\mu = \mu_1$ it will not be feasible to get the lower limit to be above μ_1 with high probability.

One could make μ_1 the new null. Then rejecting H_0 provides an inference in favor of the proposed change being beneficial enough. That still leaves open how to choose the second alternative value $\mu_2 > \mu_1$ under which the power is computed.

The problem seems to call for two different prior distributions. The experimenter planning the test might have a prior distribution on μ that is fairly optimistic about μ being meaningfully larger than μ_1. They might be aware of a more skeptical prior distribution on μ that somebody else, who will judge their work, holds. Then the idea is to choose an experiment informative enough that the first person believes that the second person will be convinced that $\mu \in (\mu_1, \infty)$ has high probability. That is we want

$$\Pr_{\text{experimenter}}\Bigl(\Pr_{\text{judge}}\bigl(\mu > \mu_1 \mid \text{Data}\bigr) \geqslant 1-\epsilon_1\Bigr) \geqslant 1-\epsilon_2, \tag{2}$$

for some small $\epsilon_j > 0$. The hard part of doing (2) would be in eliciting the two priors, learning first from the judge something about what it would take to be convincing, and then pinning down the experimenter's beliefs. There may also be the customary hard choice about what the variance of the future data will be. Picking the ϵ_j would also be tricky, not because of the technical challenge, but just in eliciting loss functions.

Going through this put me in mind of Jim Zidek's early 1980s work on multi-Bayesian theory. The most cited paper there is his JRSS-A paper with Weerahandri from 1981. From the abstract it looks more like it addresses formation of a consensus posterior or decision choice and is not about study design. That work is behind a Wiley pay wall so high that even Stanford's library credentials do not let me see it. I keep this in mind whenever Wiley asks me to contribute an encyclopedia article; preparing a write-only paper for them is a very low priority.

It would be a nuisance if we had to consider whether the probabilistic beliefs of the judge and experimenter were subject to some sort of statistical dependence. This seems not to be the case. Let the judge be convinced if the Data belong to some set S. Then we want the probability under the experimenter's prior distribution that the data will belong to S.

Howlers and chestnuts

I was intrigued by the collection of howlers and chestnuts. Partisans of one statistical philosophy or another might consider them one hit knockouts against another school of thought. Maybe you can reduce an opposing school of thought to the butt of an xkcd joke and then not take it seriously any longer. The problem is that the howlers have to be constructed against a straw man. For instance, a Bayesian howler against frequentist methods could have a hypothetical situation with clear and obviously important prior information ignored by the frequentist method. That won't generalize to cases with weaker and more ambiguous prior information. Likewise a frequentist howler against Bayesian methods can be misleading.

More than one school of thought has had contributions from excellent thinkers. Choosing what to do cannot be as simple as avoiding any approach with one or more howlers or chestnuts defined against it. Statistical practice is full of tradeoffs and catch-22s; no method would remain if the howlers had their say.

Two numbers smushed into one

Some attempts to fix p-value problems involve making the threshold more stringent as n increases. I think this is a bad idea. It is an attempt to smush statistical and practical significance into one decision criterion. Then nobody can undo the criterion to get back to statistical and practical significance. Presenting confidence intervals at one or more confidence levels is better. Then one can see the whole 2 × 2 table:

	Statistically significant	Statistically insignificant
Practically significant	Interesting finding	We need more data
Practically insignificant	We could have used less data	Maybe we can work with the null

Scientific and advocacy loss functions

I believe that the statistical problem from incentives is more severe than choice between Bayesian and frequentist methods or problems with people not learning how to use either kind of method properly. Incentive issues are more resistant to education and naturally promote findings that don't reproduce. (This point has been made by many others.)

We usually teach and do research assuming a scientific loss function that rewards being right. We have in mind a loss function like the scientist's loss function

Scientist's loss	Decide A	Decide not A
A true	0	1
A false	1	0

with generalizations to more than a binary choice and not necessarily equal losses. In practice many people using statistics are advocates. They behave as if, or almost as if, the loss function is

Advocate's loss	Decide A	Decide not A
A true	0	1
A false	0	1

as it would be for one side in a civil lawsuit between two companies. The loss function strongly informs their analysis, be it Bayesian or frequentist. The scientist and advocate both want to minimize their expected loss. They are lead to different methods.

The issue can afflict scientists where A is a pet theory (their own or their advisor's), people in business where A might be about their product being safe enough to use or more effective than a competitor's, and people working for social good, where A might be about results of a past intervention or the costs and benefits of a proposed change.

If you're taking the scientific point of view but somehow hoping for one outcome over the other, then your loss function starts to look like a convex combination of the above two loss functions. You could then find yourself giving results you don't like extra scrutiny compared to results that went your way. This is similar to confirmation bias, though that term is a better description of biases from one's prior belief than from the loss function.

Conversely, even somebody working as an advocate may have a loss function with a portion of scientific loss. For instance, advocating for the scientifically wrong outcome too often will in some contexts make one a less credible advocate.