Mismatch between scientific theories and statistical models[1]

Andrew Gelman

24 Feb 2021

Yarkoni recommends that psychology researchers should take care to align their statistical models to the verbal theories they are studying and testing. This principle applies not just to qualitative theories in psychology but also to more quantitative sciences: there, too, mismatch between open-ended theories and specific statistical models has led to confusion.

In this comment, I would like to first put Yarkoni's paper in the context of statistical reasoning and then illustrate that the problem he discusses arises not just in psychology but in other sciences as well.

Following Popper (1934/1959) and Lakatos (1978), we can consider two basic paradigms of scientific inference:

1. *Confirmation:* You gather data and look for evidence in support of your research hypothesis. This could be done in various ways, but one standard approach is via statistical significance testing: the goal is to reject a null hypothesis, and then this rejection will supply evidence in favor of your preferred research hypothesis.
2. *Falsification:* You use your research hypothesis to make specific (probabilistic) predictions and then gather data and perform analyses with the goal of rejecting your hypothesis.

It is tempting to consider confirmationist reasoning as bad and falsificationist reasoning as good, but both have their role within good scientific practice. Mayo (1996) considers these inferential approaches as part of a larger process in which experiments are designed to test and adjudicate among competing hypotheses.

It is important to distinguish these from a third, erroneous mode of reasoning:

3. *Naive confirmationism*: You start with scientific hypothesis A, then as a way of confirming this hypothesis, the researcher comes up with null hypothesis B. Data are found which reject B, and this is taken as evidence in support of A.

In Yarkoni's terminology, hypothesis A is a verbal assertion in psychology, and null hypothesis B is a statistical model. When expressed above, naive confirmationism is an obvious logical fallacy, but it is done all the time in research published in top journals. For example, Durante et al. (2013) used survey data to claim that "the ovulatory cycle not only influences women's politics but also appears to do so differently for single women than for women in relationships" offering as evidential support the rejection of a series of statistical null hypotheses.

The difficulty here is that either the scientific hypothesis is general and non-quantitative (in which case, sure, the ovulatory cycle, like everything else, will have *some* nonzero effect, and so

[1] Discussion for *Behavioral and Brain Sciences* of "The generalizability crisis" by Tal Yarkoni.

the conformation of this vapid hypothesis tells us nothing whatsoever) or the hypothesis is quantitative—what are the purported effects, how large are they, and how persistent are they across people and across settings—in which case these effects need to be studied with a statistical model appropriate to the task, and the rejection of an empty null is irrelevant. See Gelman (2015) for further discussion of this example in the general context of varying treatment effects.

In his article, Yarkoni focuses on verbal hypotheses in psychology, but similar problems arise in other fields. For example, Chen et al. (2013) claim that a coal heating policy caused a reduction of life expectancy of 5.1 years in half of China, with the supporting evidence coming from a discontinuity regression. In this case, the scientific model is quantitative, but there is still a disconnect with the statistical model, so that the empirical claims are questionable: to put it in statistical terms, the 95% confidence intervals do not have 95% coverage, and the null hypothesis can reject much more than 5% of the time even if there is no effect (Gelman and Zelizer, 2015, Gelman and Imbens, 2019). The problem is that the statistical model makes many assumptions beyond the scientific model of the effect of pollution.

It would be usual to characterize the two above stories as statistical errors. In the ovulation example, the mistake is to make a strong conclusion from the rejection of a null hypothesis, without recognizing that in practice all statistical hypotheses are false; and in the coal heating example, the mistake is to use a flawed statistical model that overfits to patterns in the data which are not pure noise (for that, the regression would indeed have its advertised statistical properties). In addition, both cases are examples of the garden of forking paths, by which an analyst can choose among many possible statistical tests to apply to a problem, making it easier to obtain statistical significance and thus publishable results.

Following Yarkoni, though, we can see these not just as examples of poor analysis or questionable research practices, but as special cases of the general problem of mismatch between scientific and statistical models.

Moving forward, we should to recognize the limitations of any statistical model—and I say this as a person who constructs and fits these models for a living. Rather than thinking of "the model" or "the test" corresponding to a scientific or engineering question, we recommend fitting a multitude of models. Think of models as tools. When building a house, or even simply installing a shelf, we don't just use a hammer or a screwdriver or a level. We use all these tools and more. Different models and statistical tests capture different aspects of the data we observe and the underlying structure we are trying to study.

References

Chen, Y., Ebenstein, A., Greenstone. M., et al. (2013). Evidence on the impact of sustained exposure to air pollution on life expectancy from China's Huai River policy. Proceedings of the National Academy of Sciences 110, 12936-12941.

Durante, K. M., Arsena, A. R., and Griskevicius, V. (2013). The fluctuating female vote: Politics, religion, and the ovulatory cycle. Psychological Science 24, 1007-1016.

Gelman, A. (2015). The connection between varying treatment effects and the crisis of unreplicable research: A Bayesian perspective. Journal of Management 41, 632-643.

Gelman, A., and Imbens, G. (2019). Why high-order polynomials should not be used in regression discontinuity designs. Journal of Business and Economic Statistics 37, 447-456.

Gelman, A., and Zelizer, A. (2015). Evidence on the deleterious impact of sustained use of polynomial regression on causal inference. Research and Politics 2, 1-7.

Lakatos, I. (1978). Philosophical Papers. Cambridge University Press.

Mayo, D. G. (1996). Error and the Growth of Experimental Knowledge. University of Chicago Press.

Popper, K. R. (1934/1959). The Logic of Scientific Discovery. London: Hutchinson.