Confirmationist and Falsificationist Paradigms in Statistical Practice

Andrew Gelman

There is a divide in statistics between classical frequentist and Bayesian methods. Classical hypothesis testing is generally taken to follow a falsificationist, Popperian philosophy in which research hypotheses are put to the test and rejected when data do not accord with predictions. Bayesian inference is generally taken to follow a confirmationist philosophy in which data are used to update the probabilities of different hypotheses. We disagree with this conventional Bayesian-frequentist contrast: We argue that classical null hypothesis significance testing is actually used in a confirmationist sense and in fact does not do what it purports to do; and we argue that Bayesian inference cannot in general supply reasonable probabilities of models being true. The standard research paradigm in social psychology (and elsewhere) seems to be that the researcher has a favorite hypothesis A. But, rather than trying to set up hypothesis A for falsification, the researcher picks a null hypothesis B to falsify, which is then taken as evidence in favor of A. Research projects are framed as quests for confirmation of a theory, and once confirmation is achieved, there is a tendency to declare victory and not think too hard about issues of reliability and validity of measurements.

Instead, we recommend a falsificationist Bayesian approach in which models are altered and rejected based on data. The conventional Bayesian confirmation view blinds many Bayesians to the benefits of predictive model checking. The view is that any Bayesian model necessarily represents a subjective prior distribution and as such could never be tested. It is not only Bayesians who avoid model checking. Quantitative researchers in political science, economics, and sociology regularly fit elaborate models without even the thought of checking their fit.

We can perform a Bayesian test by first assuming the model is true, then obtaining the posterior distribution, and then determining the distribution of the test statistic under hypothetical replicated data under the fitted model. A posterior distribution is not the final end, but is part of the derived prediction for testing. In practice, we implement this sort of check via simulation.

Posterior predictive checks are disliked by some Bayesians because of their low power arising from their allegedly "using the data twice". This is not a problem for us: it simply represents a dimension of the data that is virtually automatically fit by the model.

What can statistics learn from philosophy? Falsification and the notion of scientific revolutions can make us willing to check our model fit and to vigorously investigate anomalies rather than treat prediction as the only goal of statistics. What can the philosophy of science learn from statistical practice? The success of inference using elaborate models, full of assumptions that are certainly wrong, demonstrates the power of deductive inference, and posterior predictive checking demonstrates that ideas of falsification and error statistics can be applied in a fully Bayesian environment with informative likelihoods and prior distributions.