Andrew Gelman
Column Editor

Ethics and the Statistical Use of Prior Information

1. The choice to not use all available information

Debates about statistical foundations can be annoying to practitioners but are important in that foundational claims are used to make general recommendations for practice.

All statistical methods allow prior information to be used in the design of a study, or in choosing what variables to include and how to transform them, or in the interpretation of results. What distinguishes Bayesian methods is the expression of prior information in the form of probability distributions on parameters in a model. But this is controversial.

Most directly, technical arguments about the efficiency of different statistical procedures translate into ethical concerns. As quantitative researchers, we are supposed to use the most accurate estimates and the most honest statements of uncertainty, using statistically inferior methods only in response to other concerns such as simplicity, cost, or substantive theory.

For example, in his "bread and peace" model, political scientist Douglas Hibbs forecasts presidential election outcomes given only two variables, one summarizing economic trends in the year or so leading up to the election and the other being a measure of military casualties during the president's term. The accuracy of the predictions is impressive given that the model is so simple—and that gives us insight into the politics of elections.

Hibbs emphasizes that his model is not intended to produce the most accurate forecasts, which would require augmenting his predictors with other information such as recent polls. It would be unethical, or at best incompetent, to present Hibbs's model as a state-of-the-art forecast given that other useful information is available. But the fitted model stands on its own terms as a statement about elections.

Hibbs's model is not Bayesian (at least, not explicitly so). It is relevant to our discussion here to illustrate that a model which is weak for one purpose can be strong for another.

Similarly, as I have written elsewhere, "There is a class of problems where, for ethical or security reasons, it is illegal or inappropriate to use all available information. For example, the Census is not allowed to release fine-grained cross-tabulations that could be used to deduce information about individual people; racial profiling cannot be used in making mortgage decisions; and students' Calculus 1 grades and SAT scores would not be used in determining their grades in Calculus 2, even though these pieces of information could be informative. So there are settings where preconceptions don't get used, but this is more of a concern of law and ethics than of statistical inference."

2. General arguments for the superiority of a Bayesian or non-Bayesian paradigm

More generally, various statisticians have argued that their preferred methods satisfy certain desirable general properties and thus should be preferred in practice. These sorts of theoretical arguments are valuable in stimulating our intuition and in suggesting ways to evaluate and improve our methods, but they do not quite apply to the practice of statistics.

For example, theoretical statistician Larry Wasserman has written, "The particle physicists have left a trail of such confidence intervals in their wake. Many of these parameters will eventually be known (that is, measured to great precision). Someday we can count how many of their intervals trapped the true parameter values and assess the coverage. The 95 percent frequentist intervals will live up to their advertised coverage claims." Whatever the strength of this claim from a theoretical perspective (not completely clear, given the difficulties of constructing confidence intervals with discrete data and in the presence of nuisance parameters), it certainly fails in practice: historically, the frequentist intervals of physical constants have not in general lived up to their advertised coverage claims, due to unavoidable systematic error (see Youden,

1962, and Henrion and Fischhoff, 1986). This historical miscalibration should not be taken as an indictment of classical methods—I have no doubt that Bayesian intervals would have similarly poor coverage—rather, it is a warning of the dangers of taking a theoretical claim out of the lab and on to the street.

From the other direction, some proponents of Bayesian methods have argued that a Bayesian approach is required for inferences to be coherent and thus it is inefficient or dishonest to use statistical procedures that cannot be interpreted in some sense as Bayesian. (Here we use "dishonest" not in a personal sense but with the meaning of calibration; from a subjective Bayesian perspective, an honest 90% interval is one that the user would be willing to place a bet at 9-1 odds.) To the extent that you find this statistical argument compelling, it makes sense to also interpret it as an ethical statement. Just as it would be (somewhat) unethical to conduct a survey and then throw away the information from one-fifth of the responses, it would be similarly morally inappropriate to avoid the use of prior information that could increase the efficiency of estimates or forecasts by 25%. That said, we do not find the argument compelling that Bayesian methods are uniformly better. In practice our models are not fixed and so the probability statements that we make do not actually form a coherent set. We are left with more general heuristics for preferring one sort of method or another (examples where a method has worked well in some application, theorems that seem reasonably applicable to real-world settings, popularity among expert practitioners, and so forth), along with the knowledge that statistical methods progress unevenly, and at any given time different methods represent the state of the art best methods in different areas of application.

3. Ethics and subjective priors

In a discussion with philosopher Deborah Mayo, statistician David Cox discusses Fisher's rule that it's ok to use prior information in design of data collection but not in data analysis. I'll argue with this point in some detail, partly because the role of prior information is central to many debates in statistics and partly because Cox's opinions deserve our respect, as the product of decades of statistical practice and introspection.

Like a lot of hundred-year-old ideas, Fisher's rule makes sense in some contexts but not in others. Consider the notorious study in which a random sample of a few thousand people was analyzed, and it was found that the most beautiful parents were 8 percentage points more likely to have girls, compared to less attractive parents. The result was statistically significant ($p<.05$) and published in a reputable journal. But in this case we have good prior information suggesting that the difference in sex ratios in the population, comparing beautiful to less-beautiful parents, is less than 1 percentage point. A (non-Bayesian) design analysis reveals that, with this level of true difference, any statistically-significant observed difference in the sample is likely to be noise. Even conditional on statistical significance, the observed difference has an over 40% chance of being in the wrong direction and will overestimate the population difference by an order of magnitude. At this point, you might well say that the original analysis should never have been done at all—but, given that it has been done, it is essential to use prior information (even if not in any formal Bayesian way) to interpret the data and generalize from sample to population.

Where did Fisher's principle go wrong here? The answer is simple—and I think Cox would agree with me here. We're in a setting where the prior information is much stronger than the data. If one's only goal is to summarize the data, then taking the difference of 8% (along with a confidence interval and even a p-value) is fine. But if you want to generalize to the population—which was indeed the goal of the researcher in this example—then it makes no sense to stop there.

Cox illustrates the difficulty in a later quote: "[Bayesians'] conceptual theories are trying to do two entirely different things. One is trying to extract information from the data, while the other, personalistic theory, is trying to indicate what you should believe, with regard to information from the data and other, prior, information treated equally seriously. These are two very different things."

Yes, but Cox is missing something important! He defines two goals: (a) Extracting information from the data. (b) A "personalistic theory" of "what you should believe." I'm talking about something in between, which is inference for the population. I think Laplace would understand what I'm talking about here. The sample is (typically) of no interest in itself, it's just a means to learning about the population.

What is the concern, then, with using prior information in data analysis? Cox says:

> There are situations where it is very clear that whatever a scientist or statistician might do privately in looking at data, when they present their information to the public or government department or whatever, they should absolutely not use prior information, because the prior opinions on some of these prickly issues of public policy can often be highly contentious with different people with strong and very conflicting views.

Maybe. But I don't think Cox even believes this statement himself if it were taken literally. For example, right now I'm working on the politically controversial problem of reconstructing historical climate from tree rings. We have a lot of prior information on the processes under which tree rings grow and how they are measured. I don't think anyone would want to just take raw numbers from core samples as a climate estimate! You need

some scientific knowledge and prior information on where these measurements came from.

So let me interpret what I think Cox was saying. I take him to be dividing any scientific inference into two parts, inside and outside. Priors are allowed in the inside work of scientific modeling, which uses lots of external information, from the basic assumptions that the data correspond to your scientific goals, through the mathematical form of the transfer function, down to details such as an assumption of normally distributed measurement errors, which might be supported based on prior experimental evidence. But Cox would prefer to avoid priors in the outside problem. In my example, I assume he'd allow prior information on the tree-ring measurement process—I don't see how you can get anywhere otherwise—but he'd rather not combine with external estimates of the temperature series. That's a tenable position. It doesn't avoid all the controversy—manipulations of the data model can map in predictable ways to changes in the final inferences—but it could make sense.

I've followed this approach in much of my own applied work, using noninformative priors and carefully avoiding the use of prior information in the final stages a statistical analysis. But that can't always be the right choice. Sometimes (as in the sex ratio example above), the data are just too weak—and a classical textbook data analysis can be misleading. Imagine a Venn diagram, where one circle is "topics that are so controversial that we want to avoid using prior information in the statistical analysis" and the other circle is "problems where the data are weak compared to prior information." If you're in the intersection of these circles, you have to make some tough choices!

Beyond this, all statistical methods require choices—assumptions, if you will. For example, Cox's proportional hazard regression is one of the most influential statistical methods of the past half-century, but additivity is a prior assumption too! It's just not possible to determine or even validate all one's choices from the data at hand. If you don't want your choices to be based on prior information, what other options do you have? You can rely on convention—using methods that appear in major textbooks and have stood the test of time—or maybe on theory. Both these meta-foundational approaches have their virtues but neither is perfect: Conventional methods are not necessarily good (as can be seen by noting that for many problems there are multiple conventional methods that give different results), and theory often doesn't help (for example classical confidence intervals and hypothesis tests are insufficient in the simple sex-ratio problem noted above).

4. A Bayesian wants everybody else to be non-Bayesian

Bayesian inference proceeds by taking the likelihoods from different data sources and then combining them with a prior distribution (or, more generally, a hierarchical model). The likelihood is key. For example, in a meta-analysis (such as the three examples in Chapter 5 of our book), you need the likelihood for each separate experiment. No funny stuff, no posterior distributions, just the likelihood. In a linear model setting, it's convenient to have unbiased estimates. I don't want everybody coming to me with their posterior distribution—I'd just have to divide away their prior distributions before getting to my own analysis.

Sort of like a trial, where the judge wants to hear what everybody saw—not their individual inferences, but their raw data. This returns us to Cox's distinction between "trying to extract information from the data" and "trying to indicate what you should believe."

In my view, we do not have an ethical obligation to avoid prior information in our analyses. In settings where prior information is particularly important, I would prefer to explicitly use that information in my analysis or, failing that, to use the prior information to aid in interpretation of results (as with our example of the 95% interval that excludes zero but nonetheless does not represent strong evidence of a positive relation in the population).

In any case, I believe we are ethically required to clearly state our assumptions and, to the best of our abilities, explain the rationales for these assumptions and our sources of information.

Further Reading

Cox, D., and Mayo, D. 2011. A statistical scientist meets a philosopher of science. *Rationality, Markets and Morals* 2, 103-114.

Gelman, A. 2008. Rejoinder to discussion of "Objections to Bayesian statistics." *Bayesian Analysis* 3, 467-478.

Henrion, M., and Fischoff, B. 1986. Assessing uncertainty in physical constants. *American Journal of Physics* 84, 791–798.

Hibbs, D. A. 2000. Bread and peace voting in U.S. presidential elections. *Public Choice* 104, 149-180.

Wasserman, L. 2008. Comment on "Objections to Bayesian statistics." *Bayesian Analysis* 3, 463-466.

Youden, W.J. 1962. *Experimentation and Measurement.* National Science Teachers' Association. Reprinted 1997 as National Institute of Science and Technology Special Publication 672. U.S. Department of Commerce.

About the Author

Andrew Gelman is a professor of statistics and political science and director of the Applied Statistics Center at Columbia University. He has received many awards, including the Outstanding Statistical Application Award from the American Statistical Association and the award for best article published in the *American Political Science Review*. He has coauthored many books; his most recent is *Red State, Blue State, Rich State, Poor State: Why Americans Vote the Way They Do.*