26

How do we choose our default methods?

Andrew Gelman

Department of Statistics, Columbia University, New York

The field of statistics continues to be divided into competing schools of thought. In theory one might imagine choosing the uniquely best method for each problem as it arises, but in practice we choose for ourselves (and recommend to others) default principles, models, and methods to be used in a wide variety of settings. This article briefly considers the informal criteria we use to decide what methods to use and what principles to apply in statistics problems.

26.1 Statistics: The science of defaults

Applied statistics is sometimes concerned with one-of-a-kind problems, but statistical methods are typically intended to be used in routine practice. This is recognized in classical theory (where statistical properties are evaluated based on their long-run frequency distributions) and in Bayesian statistics (averaging over the prior distribution). In computer science, machine learning algorithms are compared using cross-validation on benchmark corpuses, which is another sort of reference distribution. With good data, a classical procedure should be robust and have good statistical properties under a wide range of frequency distributions, Bayesian inferences should be reasonable even if averaging over alternative choices of prior distribution, and the relative performance of machine learning algorithms should not depend strongly on the choice of corpus.

How do we, as statisticians, decide what default methods to use? Here I am using the term "method" broadly, to include general approaches to statistics (e.g., Bayesian, likelihood-based, or nonparametric) as well as more specific choices of models (e.g., linear regression, splines, or Gaussian processes) and options within a model or method (e.g., model averaging, L_1 regularization, or hierarchical partial pooling). There are so many choices that it is hard to imagine any statistician carefully weighing the costs and benefits of each before

deciding how to solve any given problem. In addition, given the existence of multiple competing approaches to statistical inference and decision making, we can deduce that no single method dominates the others.

Sometimes the choice of statistical philosophy is decided by convention or convenience. For example, I recently worked as a consultant on a legal case involving audits of several random samples of financial records. I used the classical estimate $\hat{p} = y/n$ with standard error $\sqrt{\hat{p}(1-\hat{p})/n}$, switching to $\hat{p} = (y+2)/(n+4)$ for cases where $y = 0$ or $y = n$. This procedure is simple, gives reasonable estimates with good confidence coverage, and can be backed up by a solid reference, namely Agresti and Coull (1998), which has been cited over 1000 times according to `Google Scholar`. If we had been a situation with strong prior knowledge on the probabilities p, or interest in distinguishing between $p = 0.99$, 0.999, and 0.9999, it would have made sense to consider something closer to a full Bayesian approach, but in this setting it was enough to know that the probabilities were high, and so the simple $(y+2)/(n+4)$ estimate (and associated standard error) was fine for our data, which included values such as $y = n = 75$.

In many settings, however, we have freedom in deciding how to attack a problem statistically. How then do we decide how to proceed?

Schools of statistical thoughts are sometimes jokingly likened to religions. This analogy is not perfect—unlike religions, statistical methods have no supernatural content and make essentially no demands on our personal lives. Looking at the comparison from the other direction, it is possible to be agnostic, atheistic, or simply live one's life without religion, but it is not really possible to do statistics without some philosophy. Even if you take a Tukeyesque stance and admit only data and data manipulations without reference to probability models, you still need some criteria to evaluate the methods that you choose.

One way in which schools of statistics *are* like religions is in how we end up affiliating with them. Based on informal observation, I would say that statisticians typically absorb the ambient philosophy of the institution where they are trained—or else, more rarely, they rebel against their training or pick up a philosophy later in their career or from some other source such as a persuasive book. Similarly, people in modern societies are free to choose their religious affiliation but it typically is the same as the religion of parents and extended family. Philosophy, like religion but not (in general) ethnicity, is something we are free to choose on our own, even if we do not usually take the opportunity to take that choice. Rather, it is common to exercise our free will in this setting by forming our own personal accommodation with the religion or philosophy bequeathed to us by our background.

For example, I affiliated as a Bayesian after studying with Don Rubin and, over the decades, have evolved my own philosophy using his as a starting point. I did not go completely willingly into the Bayesian fold—the first statistics course I took (before I came to Harvard) had a classical perspective, and in the first course I took with Don, I continued to try to frame all the inferential

problems into a Neyman–Pearson framework. But it didn't take me or my fellow students long to slip into comfortable conformity.

My views of Bayesian statistics have changed over the years—in particular, I have become much more fond of informative priors than I was during the writing of the first two editions of *Bayesian Data Analysis* (published 1995 and 2004)—and I went through a period of disillusionment in 1991, when I learned to my dismay that most of the Bayesians at the fabled Valencia meeting had no interest in checking the fit of their models. In fact, it was a common view among Bayesians at the time that it was either impossible, inadvisable, or inappropriate to check the fit of a model to data. The idea was that the prior distribution and the data model were subjective and thus uncheckable. To me, this attitude seemed silly—if a model is generated subjectively, that would seem to be *more* of a reason to check it—and since then my colleagues and I have expressed this argument in a series of papers; see, e.g., Gelman et al (1996), and Gelman and Shalizi (2012). I am happy to say that the prevailing attitude among Bayesians has changed, with some embracing posterior predictive checks and others criticizing such tests for their low power (see, e.g., Bayarri and Castellanos, 2007). I do not agree with that latter view: I think it confuses different aspects of model checking; see Gelman (2007). On the plus side, however, it represents an acceptance of the idea that Bayesian models can be checked.

But this is all a digression. The point I wanted to make here is that the division of statistics into parallel schools of thought, while unfortunate, has its self-perpetuating aspects. In particular, I can communicate with fellow Bayesians in a way that I sometimes have difficulty with others. For example, some Bayesians dislike posterior predictive checks, but non-Bayesians mostly seem to ignore the idea—even though Xiao-Li Meng, Hal Stern, and I wrote our paper in general terms and originally thought our methods might appeal more strongly to non-Bayesians. After all, those statisticians were already using p-values to check model fit, so it seemed like a small step to average over a distribution. But this was a step that, by and large, only Bayesians wanted to take. The reception of this article was what convinced me to focus on reforming Bayesianism from the inside rather than trying to develop methods one at a time that would make non-Bayesians happy.

26.2 Ways of knowing

How do we decide to believe in the effectiveness of a statistical method? Here are a few potential sources of evidence (I leave the list unnumbered so as not to imply any order of priority):

- Mathematical theory (e.g., coherence of inference or convergence)

- Computer simulations (e.g., demonstrating approximate coverage of interval estimates under some range of deviations from an assumed model)
- Solutions to toy problems (e.g., comparing the partial pooling estimate for the eight schools to the no pooling or complete pooling estimates)
- Improved performance on benchmark problems (e.g., getting better predictions for the Boston Housing Data)
- Cross-validation and external validation of predictions
- Success as recognized in a field of application (e.g., our estimates of the incumbency advantage in congressional elections)
- Success in the marketplace (under the theory that if people are willing to pay for something, it is likely to have something to offer)

None of these is enough on its own. Theory and simulations are only as good as their assumptions; results from toy problems and benchmarks don't necessarily generalize to applications of interest; cross-validation and external validation can work for some sorts of predictions but not others; and subject-matter experts and paying customers can be fooled.

The very imperfections of each of these sorts of evidence gives a clue as to why it makes sense to care about all of them. We can't know for sure so it makes sense to have many ways of knowing.

I do not delude myself that the methods I personally prefer have some absolute status. The leading statisticians of the twentieth century were Neyman, Pearson, and Fisher. None of them used partial pooling or hierarchical models (well, maybe occasionally, but not much), and they did just fine. Meanwhile, other statisticians such as myself use hierarchical models to partially pool as a compromise between complete pooling and no pooling. It is a big world, big enough for Fisher to have success with his methods, Rubin to have success with his, Efron to have success with his, and so forth. A few years ago (Gelman, 2010) I wrote of the *methodological attribution problem*:

> "The many useful contributions of a good statistical consultant, or collaborator, will often be attributed to the statistician's methods or philosophy rather than to the artful efforts of the statistician himself or herself. Don Rubin has told me that scientists are fundamentally Bayesian (even if they do not realize it), in that they interpret uncertainty intervals Bayesianly. Brad Efron has talked vividly about how his scientific collaborators find permutation tests and p-values to be the most convincing form of evidence. Judea Pearl assures me that graphical models describe how people really think about causality. And so on. I am sure that all these accomplished researchers, and many more, are describing their experiences accurately. Rubin wielding a posterior distribution is a powerful thing, as is Efron with a permutation test or

Pearl with a graphical model, and I believe that (a) all three can be helping people solve real scientific problems, and (b) it is natural for their collaborators to attribute some of these researchers' creativity to their methods.

The result is that each of us tends to come away from a collaboration or consulting experience with the warm feeling that our methods really work, and that they represent how scientists really think. In stating this, I am not trying to espouse some sort of empty pluralism—the claim that, for example, we would be doing just as well if we were all using fuzzy sets, or correspondence analysis, or some other obscure statistical method. There is certainly a reason that methodological advances are made, and this reason is typically that existing methods have their failings. Nonetheless, I think we all have to be careful about attributing too much from our collaborators' and clients' satisfaction with our methods."

26.3 The pluralist's dilemma

Consider the arguments made fifty years ago or so in favor of Bayesian inference. At that time, there were some applied successes (e.g., I.J. Good repeatedly referred to his successes using Bayesian methods to break codes in the Second World War) but most of the arguments in favor of Bayes were theoretical. To start with, it was (and remains) trivially (but not unimportantly) true that, conditional on the model, Bayesian inference gives the right answer. The whole discussion then shifts to whether the model is true, or, better, how the methods perform under the (essentially certain) condition that the model's assumptions are violated, which leads into the tangle of various theorems about robustness or lack thereof.

Forty or fifty years ago one of Bayesianism's major assets was its mathematical coherence, with various theorems demonstrating that, under the right assumptions, Bayesian inference is optimal. Bayesians also spent a lot of time writing about toy problems, e.g., Basu's example of the weights of elephants (Basu 1971). From the other direction, classical statisticians felt that Bayesians were idealistic and detached from reality.

How things have changed! To me, the key turning points occurred around 1970–1980, when statisticians such as Lindley, Novick, Smith, Dempster, and Rubin applied hierarchical Bayesian modeling to solve problems in education research that could not be easily attacked otherwise. Meanwhile Box did similar work in industrial experimentation and Efron and Morris connected these approaches to non-Bayesian theoretical ideas. The key in any case was to use partial pooling to learn about groups for which there was only a small amount of local data.

Lindley, Novick, and the others came at this problem in several ways. First, there was Bayesian theory. They realized that, rather than seeing certain aspects of Bayes (for example, the need to choose priors) as limitations, they could see them as opportunities (priors can be estimated from data!) with the next step folding this approach back into the Bayesian formalism via hierarchical modeling. We (the Bayesian community) are still doing research on these ideas; see, for example, the recent paper by Polson and Scott (2012) on prior distributions for hierarchical scale parameters.

The second way that the Bayesians of the 1970s succeeded was by applying their methods on realistic problems. This is a pattern that has happened with just about every successful statistical method I can think of: an interplay between theory and practice. Theory suggests an approach which is modified in application, or practical decisions suggest a new method which is then studied mathematically, and this process goes back and forth.

To continue with the timeline: the modern success of Bayesian methods is often attributed to our ability using methods such as the Gibbs sampler and Metropolis algorithm to fit an essentially unlimited variety of models: practitioners can use programs such as Stan to fit their own models, and researchers can implement new models at the expense of some programming but without the need of continually developing new approximations and new theory for each model. I think that's right—Markov chain simulation methods indeed allow us to get out of the pick-your-model-from-the-cookbook trap—but I think the hierarchical models of the 1970s (which were fit using various approximations, not MCMC) showed the way.

Back 50 years ago, theoretical justifications were almost all that Bayesian statisticians had to offer. But now that we have decades of applied successes, that is naturally what we point to. From the perspective of Bayesians such as myself, theory is valuable (our *Bayesian Data Analysis* book is full of mathematical derivations, each of which can be viewed if you'd like as a theoretical guarantee that various procedures give correct inferences conditional on assumed models) but applications are particularly convincing. And applications can ultimately become good toy problems, once they have been smoothed down from years of teaching.

Over the years I have become pluralistic in my attitudes toward statistical methods. Partly this comes from my understanding of the history described above. Bayesian inference seemed like a theoretical toy and was considered by many leading statisticians as somewhere between a joke and a menace; see Gelman and Robert (2013); but the hardcore Bayesians such as Lindley, Good, and Box persisted and got some useful methods out of it. To take a more recent example, the bootstrap idea of Efron (1979) is an idea that in some way is obviously wrong (as it assigns zero probability to data that did not occur, which would seem to violate the most basic ideas of statistical sampling) yet has become useful to many and has since been supported in many cases by theory.

In this discussion, I have the familiar problem that might be called the pluralist's dilemma: how to recognize that my philosophy is just one among many, that my own embrace of this philosophy is contingent on many things beyond my control, while still expressing the reasons why I believe this philosophy to be preferable to the alternatives (at least for the problems I work on).

One way of the dilemma is to recognize that different methods are appropriate for different problems. It has been said that R.A. Fisher's methods and the associated 0.05 threshold for p-values worked particularly well for experimental studies of large effects with relatively small samples—the sorts of problems that appear over and over again in books of Fisher, Snedecor, Cochran, and their contemporaries. That approach might not work so well in settings with observational data and sample sizes that vary over several orders of magnitude. I will again quote myself (Gelman, 2010):

> "For another example of how different areas of application merit different sorts of statistical thinking, consider Rob Kass's remark: "I tell my students in neurobiology that in claiming statistical significance I get nervous unless the p-value is much smaller than 0.01." In political science, we are typically not aiming for that level of uncertainty. (Just to get a sense of the scale of things, there have been barely 100 national elections in all of U.S. history, and political scientists studying the modern era typically start in 1946.)"

Another answer is path dependence. Once you develop facility with a statistical method, you become better at it. At least in the short term, I will be a better statistician using methods with which I am already familiar. Occasionally I will learn a new trick but only if forced to by circumstances. The same pattern can hold true with research: we are more equipped to make progress in a field along directions in which we are experienced and knowledgeable. Thus, Bayesian methods can be the most effective for me and my students, for the simple reason that we have already learned them.

26.4 Conclusions

Statistics is a young science in which progress is being made in many areas. Some methods in common use are many decades or even centuries old, but recent and current developments in nonparametric modeling, regularization, and multivariate analysis are central to state-of-the-art practice in many areas of applied statistics, ranging from psychometrics to genetics to predictive modeling in business and social science. Practitioners have a wide variety of statistical approaches to choose from, and researchers have many potential directions to study. A casual and introspective review suggests that there are

many different criteria we use to decide that a statistical method is worthy of routine use. Those of us who lean on particular ways of knowing (which might include: performance on benchmark problems, success in new applications, insight into toy problems, optimality as shown by simulation studies or mathematical proofs, or success in the marketplace) should remain aware of the relevance of all these dimensions in the spread of default procedures.

References

Agresti, A. and Coull, B.A. (1998). Approximate is better than exact for interval estimation of binomial proportions. *The American Statistician*, 52:119–126.

Basu, D. (1971). An essay on the logical foundations of survey sampling, part 1 (with discussion). In *Foundations of Statistical Inference*, V.P. Godambe and D.A. Sprott, Eds. Holt, Reinhart and Winston, Toronto, pp. 203–242.

Box, G.E.P. and Tiao, G.C. (1973). *Bayesian Inference in Statistical Analysis*. Wiley, New York.

Dempster, A.P., Rubin, D.B., and Tsutakawa, R.K. (1981). Estimation in covariance components models. *Journal of the American Statistical Association*, 76:341–353.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7:1–26.

Efron, B. and Morris, C. (1975). Data analysis using Stein's estimator and its generalizations. *Journal of the American Statistical Association*, 70:311–319.

Gelman, A. (2010). Bayesian statistics then and now. Discussion of "The future of indirect evidence," by Bradley Efron. *Statistical Science*, 25:162–165.

Gelman, A., Meng, X.L., and Stern, H.S. (1996). Posterior predictive assessment of model fitness via realized discrepancies (with discussion). *Statistica Sinica*, 6:733–807.

Gelman, A. and Robert, C. (2013). "Not only defended but also applied": The perceived absurdity of Bayesian inference (with discussion). *The American Statistician*, 67(1):1–5.

Gelman, A. and Shalizi, C. (2012). Philosophy and the practice of Bayesian statistics (with discussion). *British Journal of Mathematical and Statistical Psychology*, 66:8–38.

Lindley, D.V. and Novick, M.R. (1981). The role of exchangeability in inference. *The Annals of Statistics*, 9:45–58.

Lindley, D.V. and Smith, A.F.M. (1972). Bayes estimates for the linear model. *Journal of the Royal Statistical Society, Series B*, 34:1–41.

Polson, N.G. and Scott, J.G. (2012). On the half-Cauchy prior for a global scale parameter. *Bayesian Analysis*, 7(2):1–16.

Tukey, J.W. (1977). *Exploratory Data Analysis*. Addison-Wesley, Reading, MA.