The Development of Bayesian Statistics[*]

Andrew Gelman[†]

13 Jan 2022

Abstract

The incorporation of Bayesian inference into practical statistics has seen many changes over the past century, including hierarchical and nonparametric models, general computing tools that have allowed the routine use of nonconjugate distributions, and the incorporation of model checking and validation in an iterative process of data analysis. We discuss these and other technical advances along with parallel developments in philosophy, moving beyond traditional subjectivist and objectivist frameworks to ideas based on prediction and falsification. Bayesian statistics is a flexible and powerful approach to applied statistics and an imperfect but valuable way of understanding statistics more generally.

Bayes' theorem is a mathematical identity of conditional probability, and applied Bayesian inference dates back to Laplace in the late 1700s, so what could possibly be new about it? Why does Bayesian statistics remain a research topic, 250 years later? We do not attempt any sort of historical treatment here; instead, we offer a rationalized reconstruction, laying out some of the challenges that have been recognized and addressed over the years in applying conditional probability to problems of statistical inference.

The point of this brief article is to present Bayesian statistics not as a static approach or philosophy but rather as an ever-expanding framework.

In this article, we will lay out several developments in Bayesian statistics, addressing challenges in mathematics and computation, model building, inference, and workflow: assessing, comparing, and extending models

Going beyond conjugate models with uniform priors

Let's take as a starting point the beta-binomial and normal-normal conjugate models with uniform priors, as used by Bayes, Gauss, and Laplace; see Stigler (1986). There are several ways to generalize these ideas going forward, notably we can consider informative priors; other conjugate models such

[*] For the Journal of the Indian Institute of Science. We thank the U.S. Office of Naval Research for partial support of this work.

[†] Department of Statistics and Department of Political Science, Columbia University, New York.

as gamma-Poisson; or nonlinear or nonconjugate models, in which case posterior expectations need to be expressed as integrals that cannot be expressed in closed form.

Going to informative priors with conjugate priors is an interesting example of generalization, in that it can be thought of as an entirely new idea (a new kind of prior) or it can be folded into the existing uniform-prior framework by considering the informative prior as the combination of an initially flat prior with earlier data. Even when maintaining the restriction to conjugacy, this setting is restrictive—in the binomial model, for example, it does not allow priors more variable than uniform or priors that correspond to non-integer numbers of observations. We shall encounter this sort of restriction again when considering hierarchical models.

Going beyond standard families of distributions

Around the year 1900 an important strain of statistics was the categorization of parametric models of distributions, for example the Pearson family (Pearson, 1916). It seems that there was this idea that something important could be learned from data by seeing what distribution the data came from. For example, if we think of the exponential distribution as representing waiting times for independent events, then departures from the exponential could supply insight into the underlying process, with a gamma distribution with shape 2 representing a 2-hit process. Similar reasoning can accompany departures of a count distribution from Poisson. If a variable follows a normally distribution, this corresponds to an underlying model in which it is the sum of many small independent factors, and so on.

Sufficient exposure to real data makes it clear that no fixed collection of parametric families will capture the different sorts of distributions we see with real data. A flexible and useful generalization takes us from fixed distributions to regression models. The family $p(y|\theta, x)$ is so much more flexible than $p(y|\theta)$ because the open-endedness of the predictor x allows for the marginal distribution, $p(y|theta)$, to have an arbitrary mixture form. One can also use mixture distributions, which can be thought of as conditional models where some aspects of x are unobserved. Beyond these technical advances, it is a conceptual leap to go beyond the idea of "the distribution of the data" to the use of probability distributions to represent uncertainty in predictions.

Generalizing from sample to population

Bayesian inference is usually set up assuming data have been drawn independently from the population of interest. Realistically, though, a sample can be nonrepresentative. Adjustment can be made by taking the usual formula and conditioning on adjustment variables x to yield $p(\theta|y,x) \propto p(\theta|x)p(y|\theta,x)$, which can be thought of as a conditional version of Bayesian inference (Little, 1993). Full Bayesian inference would take the form, $p(\theta|y,x) \propto p(\theta)p(x,y|\theta)$, which would require a probability model for x. Such an extension could be valuable, not just to close the theoretical loop but for two other reasons: first, the data x could be informative about the model for y; second, it opens the door to models with include predictors that are measures with error or not at all.

Consider, for example, the example of an education experiment where x include treatment assignment and pre-test score, the treatment is assigned in an imbalanced way (for example, with students who performed worse on the pre-test being more likely to get the new treatment and students who performed better on the pre-test being more likely to get the control), and y is the post-test score. Inference given x allows estimation of the treatment effect for a population of interest; going further and modeling x could supply additional information to the extent that there are similarities between the two tests. This can be seen even more clearly in a time series setting where there are multiple measurements on each person. There's no logical reason to consider a post-test as modeled and a pre-test as unmodeled.

Empirical Bayes and hierarchical models

Informative priors have gone through a series of conceptual frameworks. The cleanest idea is of nested sampling, where the data model $p(y|\theta)$ represents draws from an urn, and the prior distribution $p(\theta)$ represents a set of "urns" corresponding to different parameter values θ. When no two-stage physical sampling mechanism is available, $p(\theta)$ can be taken to represent prior uncertainty in the parameter or, equivalently, a guess of the distribution of θ across the set of problems for which the model will be used. These can be taken as the "personal" or "behavioral" interpretations. Under either case, the specification of specific probability distribution given available information requires some elicitation, and this raises the possibility that the prior could be mistaken, which could happen for many reasons: the analyst's personal knowledge could be mistaken (for example, by relying on published results that are subject to selection bias or some other unacknowledged errors), or the elicitation could have been done poorly, or the method could be applied to a different set of problems than was implied when setting up the prior reference set.

A logical next step is "empirical Bayes," where the prior distribution is estimated from the same set of data as is used to fit the model. This won't work with a single dataset but it will work with hierarchical structures with data from multiple "urns." The approach of estimating a prior from data could be seen as an extension, going outside the canonical Bayesian framework in which the prior is specified unconditionally on the data (Efron and Morris, 1972), but it can be framed as Bayesian by considering the parameters of the prior as being "hyperparameters" that are estimated jointly with the parameters of the data model (Lindley and Smith, 1972).

Beyond the possible conceptual advantages of remaining within a unified mathematical framework, the step of setting up prior estimation as hierarchical Bayes has the practical advantage of propagating uncertainty in the hyperparamers and opening the door to more elaborate models where the parameters can vary in other ways, for example over time as well as across groups, as well as connecting to ideas from data collection such as repeated measures, cluster sampling, and block designs.

Exploratory data analysis

It's good practice to look at your data, both to see clear visual patterns and to learn about the unexpected (Tukey, 1977). Any discussion of the unexpected leads to thoughts about "the expected," and this relates to statistical graphics and exploratory data analysis in two ways. First, a graph of data can be viewed as a comparison to a model, which might be explicit (as in a residual plot or quantile-quantile plot) or implicit (as when we notice some unanticipated pattern in data). Second, the existence of a model can motivate graphs that are tailored to particular concerns about data fit.

None of this is particularly Bayesian, and, indeed, the statistical literature on data visualization and exploratory data analysis has mostly been disconnected from developments in statistical modeling. What happens when we fit a model to data, look at a graph of data and fitted model, recognize a problem, and use this new understanding to reformulate the model? This hardly seems Bayesian; indeed it's counter to formal Bayesian inference in that the model is changed in light of the data. But statistical graphics can be folded into the Bayesian formalism by viewing it as posterior predictive checking, comparing replicated data y^{rep} to observed data y using the posterior predictive distribution $p(y^{\mathrm{rep}}, \theta|y)$ (Box, 1980, Rubin, 1984).

Multiple models and statistical workflow

Statistics is typically formulated in terms of estimating parameters, making predictions, or fitting models to data. When multiple models are fit, the goal will be set as choosing the best-fit model or averaging over models, which can be done using probabilistic Bayesian model averaging or using a predictive-based averaging procedure such as stacking or boosting.

But real-world statistical workflow often involves comparisons between fitted models. For example, we might obtain a simple estimate of a causal effect by comparing averages in treated and control groups, then modify this by adjusting for pre-treatment predictors, then go further by modeling selection on unobservables, then further elaborate by including treatment interactions. Along with these fits should come explanations of how and why the estimates differ (for example, "The relation between post-test and pre-test was nonlinear, and the initial regression adjustment overcorrected for differences between treatment and control group").

Again, this sort of reasoning can at first seem to be outside the Bayesian formalism, in which multiple models exist in a joint space with prior and posterior probabilities—but it can fit into an extended Bayesian framework involving inference over a network space in which each model is a node, and edges connect related models which can then be compared (Gelman et al., 2021).

Summary

Bayesian data analysis starts with a core of inference for a parameter vector with fixed prior and as such is conceptually straightforward—although not trivial, given challenges of mathematical analysis and computation. Over the past few centuries, this core has been expanded in various ways, including the use of informative priors, regression and mixture models, extrapolation to new data, estimating prior distributions from data, exploratory data analysis, and the workflow of model building, checking, improvement, and comparison. Each of these steps can at first seem to be outside the Bayesian formalism, but the core has been expanded to allow information synthesis and propagation of uncertainty at each step. In that way, Bayesian inference is not just a procedure for learning from data and models; it is also an expandable framework for the process of modeling, learning, and discovery. The references given in this article demonstrate just a few of these expansions; other important directions not discussed here include causal inference, nonparametric modeling, computation and approximation, and communication of uncertainty.

References

Box, G. E. P. (1980). Sampling and Bayes inference in scientific modelling and robustness. *Journal of the Royal Statistical Society A* **143**, 383–430.

Efron, B., and Morris, C. (1972). Limiting the risk of Bayes and empirical Bayes estimators—Part

II: The empirical Bayes case. *Journal of the American Statistical Association* **67**, 130–139.

Gelman, A., Vehtari, A., Simpson, D., Margossian, C. C., Carpenter, B., Yao, Y., Bürkner, P. C., Kennedy, L., Gabry, J., and Modrák, M. (2021). Bayesian workflow. `https://arxiv.org/abs/2011.01808`

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

Little, R. J. A. (1993). Post-stratification: A modeler's perspective. *Journal of the American Statistical Association* **88**, 1001–1012.

Pearson, K. (1916). Mathematical contributions to the theory of evolution, XIX: Second supplement to a memoir on skew variation. *Philosophical Transactions of the Royal Society A* **216**, 429–457.

Rubin, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *Annals of Statistics* **12**, 1151–1172.

Stigler, S. M. (1986). *The History of Statistics.* Cambridge, Mass.: Harvard University Press.

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