PLSC 502: "Statistical Methods for Political Research"

Bayesian Inference: A Very Brief Introduction
December 1, 2016

Introduction

Our initial discussion of probability theory and inference relied – whether we knew it or not – on what statisticians refer to as the *frequentist* approach to probability. In that view, the probability of an event:

- Can be thought of as the long-run relative frequency of that event in repeated, independent "trials." This means that
- The probability of that event is a *fixed* but *unknown* quantity, which we want to learn about by examining our (sample) data.

This is a reasonably natural (if somewhat abstract) way of thinking about probability. However, an alternative approach – one with very deep roots in philosophy and mathematics – is the idea of *subjective probability.* Together with an elementary theorem of probability and modern computing power, the subjective view of probability has become a central part of statistics and other quantitative methods. We'll first review the theorem – Bayes' theorem – and then discuss more generally the idea of subjective probability, before moving on to a discussion of Bayesian estimation and inference. There won't be a lot of data or examples at this point; rather, the idea is to get you familiar with the concepts and intuition of the Bayesian approach.[1]

Bayesian Inference: A Conceptual Overview

Suppose we have some quantity θ – a mean, a median, a correlation coefficient, a regression parameter, or whatever – that we are interested in knowing. We'll assume that θ has a range of possible values defined by the set Θ; so, for example, if θ is a Pearson's r, then $\Theta = [-1, 1]$. To learn about θ, we have some data (called, generically, Y) which can tell us something about θ. The goal is to make probability statements about θ on the basis of Y, and to do so in a way that (a) uses the information contained in Y optimally, and (b) is logically and mathematically consistent.

The *sampling density* for the observed Y is defined as $\Pr(Y|\theta)$; think of this as the joint probability of observing all the values of Y we observe in the data, conditional on whatever θ is. A *probability model* for the data consists of a distributional assumption about $\Pr(Y|\theta)$ along with the range of possible values for θ (that is, Θ). So, for example, if Y is data

[1] This presentation draws on a couple of excellent introductory lectures on Bayesian statistics prepared by my Bayesian friend Andrew Martin; see (e.g.) here for details.

on SAT scores among high school students, we might (reasonably) assume that they are more-or-less normally distributed; if, in contrast, Y consists of data on survey respondents' "agree/disagree" responses to a question about gun control, it might be more reasonable to think of $\Pr(Y|\theta)$ as following a binomial distribution.

Note, however, that what we really want to know is $\Pr(\theta|Y)$, the probability distribution of θ conditional on Y. If we think of Y as fixed, and instead consider $\Pr(Y|\theta)$ as a function of θ, then we refer to $\Pr(Y|\theta)$ as the *likelihood function*, written $L(\theta|Y)$. As we already mentioned a bit before, the *maximum likelihood* estimate of θ is the value of θ that maximizes $L(\theta|Y)$.

Bayes' Theorem / Rule / Law

How do we get from $\Pr(Y|\theta)$ to $\Pr(\theta|Y)$? First, recall that the basic definition of conditional probability states that, for two events A and B,

$$\Pr(A|B) = \frac{\Pr(A \cap B)}{\Pr(B)} \tag{1}$$

and

$$\Pr(B|A) = \frac{\Pr(A \cap B)}{\Pr(A)}. \tag{2}$$

Equation (2) can be rearranged to express the joint probability of A and B as

$$\Pr(A \cap B) = \Pr(B|A)\Pr(A). \tag{3}$$

If we substitute (3) for $\Pr(A \cap B)$ in Equation (1), we get

$$\Pr(A|B) = \frac{\Pr(B|A)\Pr(A)}{\Pr(B)}. \tag{4}$$

Equation (4) is known as *Bayes' Theorem* (or, sometimes, as *Bayes Rule* or even *Bayes' Law*). In words, it means that the conditional probability of two events A and B is equal to the product of the conditional probability of B given A and the marginal probability of A, divided by the marginal probability of B.

In the context of our parameter θ and our data Y, we can begin by thinking of θ as analogous to A in the above example, and Y as akin to B. We can then write

$$\begin{aligned} \Pr(\theta|Y) &= \frac{\Pr(\theta \cap Y)}{\Pr(Y)} \\ &= \frac{\Pr(Y|\theta)\Pr(\theta)}{\Pr(Y)}. \qquad (5) \end{aligned}$$

Each component of (5) has a corresponding name:

- As we noted above, $\Pr(Y|\theta)$ is the *sampling density* of the data.
- $\Pr(\theta)$ is typically known as the *prior density* of θ (usually shortened to just "the prior"). It is the marginal (pre-data) probability density of the parameter θ.
- The left-hand side, $\Pr(\theta|Y)$ is known as the *posterior density* of θ (again, almost always shortened to just "the posterior"). It is the conditional probability density of θ, where the conditioning is on Y.
- Finally, $\Pr(Y)$ is the marginal probability of Y, the data. Since, for a single sample, Y is fixed, we can (and people often do) rewrite (5) as:

$$\Pr(\theta|Y) \propto \Pr(Y|\theta)\Pr(\theta). \tag{6}$$

We discuss what this means in more concrete terms below.

Subjective Probability

Equations (4) and (5) are logically true statements;[2] they hold irrespective of one's views of probability. To conduct Bayesian inference, however (and for reasons we'll see in a bit), one typically adopts a *subjectivist* view of probability.

Under this view,[3] probability is best thought of as a state of knowledge, the *degree of belief* that a particular proposition is true. That is, the probability of an event is the analyst's belief in its likelihood of occurring.

Subjective probability shouldn't be unfamiliar to anyone who has ever entered a lottery, placed a bet, or otherwise subjected something of value to a stochastic process. It's entirely reasonable, in such a view, for individuals to have different estimates of some event's probability of occurring (e.g., a particular horse winning a race), even though from a frequentist perspective there is only one "true" probability of that event (and, a Bayesian would say, there is no objective "probability" of a particular horse winning a particular race; the horse either wins, or it doesn't).

In a subjective probability context, we can think of $\Pr(\theta)$ as our prior / "pre-data" estimate of the value/distribution of θ (since it does not "depend on" Y), and $\Pr(\theta|Y)$ as our posterior / "post-data" estimate. Seen in this way, (5) provides a means for updating our beliefs about the value of θ on the basis of the information contained in Y. So, for example, (6) states that:

[2] Except that they aren't; see here.

[3] Subjective probability has been around for a while, but saw most of its development in the 20th century, in works by people like de Finetti (1930), Ramsey (1931), Jeffreys (1931, 1939) and Savage (1954).

> *The posterior density of θ is proportional to the product of the prior density of θ and the data Y.*

To the extent that the data are informative about θ, the posterior estimate will be (in many ways) "better" than the prior.

Bayesian Data Analysis: A Conceptual Overview

At a conceptual / intuitive level, then Bayesian data analysis proceeds in (basically) five steps:

1. **Set up a probability model for the data.** That is, begin by specifying the conditional probability distribution for the data [$f(Y|\theta)$], on the basis of what one knows about those data.

   - This includes things like
     - picking the probability distribution for Y,
     - specifying the model (i.e., which covariates to include, as well as interactions and the like, which functional form to use, and so forth), and
     - placing restrictions (if any) on the parameter space Θ.
   - At this stage, theory and knowing one's data is particularly important.

2. **Posit one's prior beliefs.** This is a critical difference between frequentist and Bayesian practices: Bayesians are necessarily explicit about their prior belief about $\Pr(\theta)$. Note a few things about this:

   - A prior is nothing more than the researcher's *a priori expectation* about the value(s) of θ.
   - More specifically, it is the researcher's pre-data belief about the *distribution* of θ (remember that, in the Bayesian framework, θ is itself a random variable). That means that the prior is usually specified in terms of a *distribution* of values for θ.
   - So, for example, if θ is a correlation coefficient (r), and we didn't have any particular prior expectation about what r might be, we might set a prior of

     $$f(r) \sim U(-1, 1)$$

     that is, as a uniform distribution between 0 and 1. In contrast, if θ is π, the probability parameter of a binomial (as in something like a coin toss), and we had a good reason to think that the value of π was 0.5 (i.e., that the coin was "fair"), then we might specify

     $$f(\pi) \sim N(0.5, 0.25)$$

     that is, a Normal distribution with a mean of 0.5 and a variance of 0.25.

- A prior that has especially large variance is often referred to as a "flat" (or "diffuse" or "vague") prior. Such priors indicate that the researcher has a great deal of *ex ante* uncertainty about the value of θ.
- As a practical matter, a common choice of a prior distribution is what is known as a *conjugate* prior. It's a bit involved to get into right now, but suffice it to note a few things:
    - A conjugate prior is a prior that follows a distribution which – when combined with the likelihood – yields a posterior that follows the same distribution.
    - So, for example, if one assumes a Bernoulli distribution for one's sampling density / likelihood (as we would do if we were, say, modeling coin flips or other binary events), then a prior that followed a Beta distribution[4] when combined with a Bernoulli likelihood, yields a posterior that is also Beta-distributed.
    - Every likelihood distribution – discrete or continuous – has an associated conjugate prior distribution; the use of a conjugate prior almost always simplifies computation of the posterior.

3. **Calculate the posterior distribution using Bayes' Theorem.** Once one has (a) the prior for θ, (b) the sampling density, and (c) the data, then calculation of $\Pr(\theta|Y)$ is (in theory) straightforward. We won't go into this in any detail, but the machinery for doing this involves simulations.

    - Gibbs sampling +
    - Metropolis-Hastings →
    - Markov-Chain Monte Carlo ("MCMC").

    We won't talk about any of these in detail; suffice it (for now) to note that each involves simulating many, many draws from the posterior distribution, using Bayes' Theorem, and then summarizing the resulting quantities.

4. **Summarizing the posterior density.** This usually involves calculating quantities of interest to the researcher, including things like:

    - The posterior *expected value* of θ :

$$\mathrm{E}(\theta|Y) = \int_{\Theta} \theta \Pr(\theta|Y) d\theta.$$

[4] Recall that the Beta is a two-parameter distribution with support on the unit interval, of the form $\mathcal{B}(X) = f(X|\alpha,\beta) = \frac{X^{\alpha-1}(1-X)^{\beta-1}}{\int_0^1 u^{\alpha-1}(1-u)^{\beta-1}\,du}$.

- The posterior *variance* of θ:

$$\mathrm{Var}(\theta|Y) = \int_{\Theta} [\theta - \mathrm{E}(\theta|Y)]^2 \, \mathrm{Pr}(\theta|Y) d\theta.$$

- $100 \times (1-\alpha)\%$ credible intervals for θ (similar to confidence intervals in the frequentist world, but with some nicer properties).[5]

5. **Conduct post-estimation model checking.** This is also beyond what we want to cover here, but generally includes checking for convergence of the MCMC chain and the like.

Advantages to the Bayesian Approach

There are a number of advantages to a Bayesian approach to inference.[6] The central advantage of the Bayesian approach can be summed up as:

Bayesian inference directly quantifies uncertainty.

That is, both through the specification of a prior distribution for each parameter, and through the way Bayes' Theorem incorporates the information contained in the data, the Bayesian approach is always explicit about uncertainty. For example, a more explicit/small-variance prior will tend to "dominate" the data, particularly if the number of observations is small; conversely, a diffuse prior will tend to be dominated by the information in the data.

[5] In particular, a credible interval can be interpreted as the interval in which the parameter θ has a (posterior) probability of falling of $(1-\alpha)$. So, for example, if a calculated 95% credible interval for a Pearson's correlation coefficient r equal to $[0.60, 0.76]$, that can be interpreted as meaning that the posterior probability of r falling within the range $[0.60, 0.76]$ is 0.95. In other words, credible intervals are what we'd like confidence intervals to be.

[6] In fact, some would argue that there's really no reason to be anything *other* than Bayesian when it comes to data analysis and statistics...

More generally, advantages of Bayesian inference include:

- **Bayesian data analysis provides direct quantities of interest to researchers.** That is, we learn $\Pr(\theta|Y)$, which is/are the thing(s) we typically care about. Moreover, they do so while providing intuitive summaries of those quantities, and with no appeal to asymptotic theory.
- **Bayesian methods of inference are logically consistent, and very intuitive to most people.**
- **Bayesian methods allow the incorporation of prior information in a mathematically and logically consistent way.**
- **Bayesian methods allow the fitting of models that are more complex that might otherwise be estimable.**
- **Bayesian approaches are practically flexible; for example, they allow for simple means of pooling data across sources, making model comparisons, and dealing with missing data.**

Some Disadvantages

The Bayesian approach is not without its potential disadvantages. Common concerns include:

- **The inherent subjectivity of choosing priors, and the sensitivity of one's results to the priors chosen.**
- **The computational difficulty/complexity of Bayesian estimation.**
- **The difficulty in knowing whether one's simulations have "converged."**
- **The lack of simple-to-use software for doing Bayesian analysis.**