8
Introduction to Bayesian Estimation

Some important applications of Bayesian statistical inference rely on computational methods. In particular, Chapters 9-XX of this book illustrate the computational role of the Gibbs sampler in Bayesian estimation. By way of preparation, this chapter introduces some of the fundamental ideas of Bayesian estimation.

Bayesian and frequentist statistical inference take fundamentally different viewpoints toward statistical decision making.

- The frequentist view of probability, and thus of statistical inference, is based on the idea of an experiment that can be repeated many times.

- The Bayesian view of probability and of inference is based on a personal assessment of probability and on observations from a single performance of an experiment.

Frequentists and Bayesians use fundamentally different procedures of estimation, and the interpretations of the resulting estimates are also fundamentally different. In practical application, both ways of thinking have advantages and disadvantages, some of which we will explore here.

Statistics is a young science. For example, interval estimation and hypothesis testing have become common in scientific research and business decision making only within the past 75 years, and then only gradually. On this time scale it seems strange to talk about "traditional" approaches. But frequentist viewpoints are currently much better established, particularly in scientific research, than Bayesian ones. Recently, the use of Bayesian methods has been increasing, partly because the Bayesian approach seems to be able to get more useful solutions than frequentist ones in some applications and partly because improvements in computation have made these methods easier–or feasible–to apply in practice. The Gibbs sampler is one computationally intensive method that is broadly applicable in Bayesian estimation.

For some of the very simple examples considered here, Bayesian and frequentist methods give similar results. But that is not the main point. We hope you will gain some appreciation that Bayesian methods are sometimes the most natural and useful ones in practice. Also, we hope you will begin to appreciate the essential role of computation in Bayesian estimation.

For most people, the starkest contrast between frequentist and Bayesian approaches is that Bayesian inference provides the opportunity—even imposes the requirement—explicitly to take into account "information" that is available before any data are collected. That is where we begin.

8.1 Prior Distributions

The Bayesian approach to statistical inference treats population parameters as random variables (not as fixed, unknown constants). The distributions of these parameters are called **prior distributions**. Often both expert knowledge and mathematical convenience play a role in selecting a particular type of prior distribution. This is easiest to explain and to understand in terms of examples. Here we introduce three examples that we carry through subsequent sections of this chapter.

Example 8.1.1 ***Election polling.*** Suppose Proposition A is on the ballot for an upcoming statewide election, and a political consultant has been hired to help manage the campaign for its passage. The proportion π of prospective voters who currently favor Proposition A is the population parameter of interest here. Based on her knowledge of the politics of the state, the consultant's judgment is that the proposition is almost sure to pass, but not by a large margin. She believes that the most likely proportion of voters in favor is 55% and that the percentage is not likely to be below 51% or above 59%.

It is reasonable to consider the beta distribution to model the expert's opinion of the proportion in favor because distributions in the beta family take values in the interval (0, 1) as do proportions. This family of distributions has density functions of the form

$$\begin{aligned} p(\pi) &= K\pi^{\alpha-1}(1-\pi)^{\beta-1} \\ &\propto \pi^{\alpha-1}(1-\pi)^{\beta-1}, \end{aligned}$$

where α, $\beta > 0$ and K is the constant such that $\int_0^1 p(\pi)\,d\pi = 1$. Here we adopt two conventions that are common in Bayesian discussions: the use of the letter p instead of f to denote a density function, and the use of the symbol $\propto$ (read "proportional to") instead of $=$ so that we can avoid specifying a constant whose exact value is unimportant to the discussion. The essential factor of the density function that remains when the constant is suppressed is called the **kernel** of the density function (or of its distribution).

A member of the beta family that corresponds reasonably well to the expert's opinion has $\alpha_0 = 330$ and $\beta_0 = 270$. (See the broken curve in Figure 1.) This is a reasonable choice of parameters for several reasons.

- This beta distribution is centered near $0.55 = 55\%$ by any of the common measures of centrality. By analytic methods one can show that the *mean* of this distribution is $\alpha_0/(\alpha_0 + \beta_0) = 330/600 = 55.00\%$ and that its *mode* is $(\alpha_0 - 1)/(\alpha_0 + \beta_0 - 2) = 329/598 = 55.02\%$. Computational methods show the *median* to be 55.01%. (The S-Plus function `qbeta(.5, 330, 270)` returns `0.5500556`.) In ???? we discuss criteria for selecting which measure of centrality to use, but here it doesn't make any practical difference.

- Also, numerical integration shows that these parameters match the expert's prior probability interval fairly well: $P\{0.51 < \pi < 0.59\} \approx 0.95$. (In S-Plus, `pbeta(.59, 330, 270) - pbeta(.51, 330, 270)` returns `0.9513758`.)

Of course, slightly different choices for α_0 and β_0 would match the expert's opinion about as well. It is not necessary to be any fussier in choosing the parameters than the expert was in specifying her hunches. Also, distributional shapes other than the beta might match the expert's opinion just as well. But we choose a member of the beta family because it makes the mathematics relatively easy in what comes later and because we have no reason to believe that the shape of our beta distribution is inappropriate here. (See Problems 1 and 2.)

If the consultant's judgments about the political situation are correct, then they may be helpful in managing the campaign. If she too often brings bad judgment to her clients, her reputation will suffer and she will be out of the political consulting business before long. Fortunately, as we will see in the next section, the details of her judgments become less important if we also have some polling data to rely upon. ◇

Example 8.1.2 ***Weighing an object.*** A construction company buys steel beams with a nominal weight of 200 lb. Experience with a particular supplier of these beams has shown that their beams very seldom weigh less than 180 or more than 220 lb. In these circumstances it may be convenient and reasonable to use $\mathsf{NORM}(200, 10)$ as the prior distribution of the weight of a randomly chosen beam from this supplier.

Usually, the exact weight of a beam is not especially important, but there are some situations in which it is crucial to know the weight of a beam more precisely. Then a particular beam is selected and weighed several times on a scale in order to determine its true weight more accurately.

Theoretically, a frequentist statistician would ignore "prior" or background experience in doing statistical inference, basing statistical decisions only on the data collected when a beam is weighed. In real life it is not so simple. For example, the design of the weighing experiment will very likely take past experience into account in one way or another. (For example, if you are going to be weighing things you need to know whether you'll be using a laboratory balance, a truck scale, or some intermediate kind of scale. And if you need more precision than the scale will give in a single measurement, you may need to weigh each object several times and take the average.) For the Bayesian statistician the explicit codification of some kinds of background information into a prior distribution is a required first step. ◇

Example 8.1.3 ***Counting mice.*** An island in the middle of a river is one of the last known habitats of an endangered kind of mouse. The mice rove about the island in ways that are not fully understood and so are taken as random.

Ecologists are interested in the average number of mice to be found in particular regions of the island. To do the counting in a region they set many traps there at night, using bait that is irresistible to mice at close range. In the morning they count and release the mice caught. It seems reasonable to suppose that almost all of the mice in the region around the trap during the previous night were caught and that the number of them on any one night has a Poisson distribution. The purpose of the trapping is to estimate the mean λ of this distribution.

Even before the trapping is done the ecologists doing this study have some information about λ. For example, even though the mice are quite shy, there have been occasional sightings of them in almost all regions of the island, so it seems likely that $\lambda > 1$. On the other hand, from what is known of the habits of the mice and the food supply in the regions, it seems unlikely that there would be as many as 25 of them in any one region at a given time.

In these circumstances, it seems reasonable to use a gamma distribution as a prior distribution for λ. This gamma distribution has the density $p(\lambda) \propto \lambda^{\alpha-1}e^{-\kappa\lambda}$, for $\lambda > 0$, where the shape parameter α and the rate parameter κ must both be positive. First, we choose a gamma distribution because it puts all of its probability on the positive half line, and λ must surely have a positive value. Second, we choose a member of the gamma family because it simplifies some important computations that we need to do later.

Using straightforward calculus, one can show that a distribution in the gamma family has mean α/κ, mode $(\alpha - 1)/\kappa$, and variance α/κ^2. These distributions are right-skewed, with the skewness decreasing as α increases.

Let's see what happens if we choose a gamma density with $\alpha_0 = 4$ and $\kappa_0 = 1/3$ as a prior distribution for λ. Reflecting the skewness, the mean 12, median 11.02, and mode 9 are noticeably different. (We obtained the median using S-Plus: `qgamma(.5, 4, 1/3)` returns `11.01618`.) Numerical methods also show that $P\{\lambda < 25\} = 0.97$. (In S-Plus, `pgamma(25, 4, 1/3)` returns `0.9662266`.) All of these values are consistent with the the expert opinions of the ecologists.

It is clear that the experience of the ecologists with the island and its endangered mice will influence the course of this investigation in many ways: dividing the island into meaningful regions, modelling the randomness of mouse movement as Poisson, deciding how many traps to use and where to place them, choosing a kind of bait that will attract mice from a region of interest but not from all over the island, and so on. The expression of some of their background knowledge as a prior distribution is perhaps a relatively small use of their expertise. But it is a necessary first step in Bayesian inference, and it is perhaps the only aspect of their expert opinion that will be explicitly tempered by the data that are collected. ◇

8.2 Data and Posterior Distributions

The second step in Bayesian inference is to collect data and to combine the information in the data with the expert opinion represented by the prior distribution. The result is a posterior distribution that can be used for inference.

Once the data are available, we can use Bayes' Theorem to compute the posterior distribution $\pi|x$. Equation (5.4) states an elementary version of Bayes' Theorem for an observed event E and a partition $\{A_1, A_2, \ldots, A_k\}$ of the sample space S. It expresses a posterior probability $P(A_j|E)$ in terms of the prior probabilities $P(A_i)$ and the conditional probabilities $P(E|A_i)$. Here we use a more general version of Bayes' Theorem involving data x and a parameter π:

$$\begin{aligned} p(\pi|x) &= \frac{p(\pi)p(x|\pi)}{\int p(\pi)p(x|\pi)\,d\pi} \\ &\propto p(\pi)p(x|\pi), \end{aligned} \tag{8.1}$$

where the integral is taken over all values of π for which the integrand is possible. (In case the distribution of π is discrete, the integral is interpreted as a sum.) The proportionality symbol $\propto$ is appropriate because the integral is a constant. (In case the distribution of π is discrete, the integral is interpreted as a sum.)

Thus the posterior distribution of $\pi|x$ is found from the prior distribution of π and the distribution of the data x given π. If π is a known constant, $p(x|\pi)$ is the density function of x; we might integrate it with respect to

x to evaluate the probability $P(x \in A) = \int_A p(x)\,dx$. However, when we use (8.1) to find a posterior, we know the data x, and we view $p(x|\pi)$ as a function of π. When viewed in this way, $p(x|\pi)$ is called the **likelihood function** of π. (Technically, the likelihood function is defined only up to a positive constant.)

A convenient summary of of our procedure for finding the posterior distribution with relationship (8.1) is to say

$$\text{POSTERIOR} \propto \text{PRIOR} \times \text{LIKELIHOOD}.$$

We now illustrate this procedure for each of the examples of the previous section.

Example 8.2.1 ***Election Polling (continued).*** Suppose that n randomly selected subjects express opinions on Proposition A. What is the likelihood function, and how do we use it to find the posterior distribution? If the value of π were known, the number x of the respondents in favor of Proposition A is a random variable with the binomial distribution: $\binom{n}{x}\pi^x(1-\pi)^{n-x}$, for $x = 0, 1, 2, \ldots, n$. So, now that we have data x, $p(x|\pi) \propto \pi^x(1-\pi)^{n-x}$ is the likelihood function of π.

Display (8.1) gives the posterior distribution

$$\begin{aligned} p(\pi|x) &\propto \pi^{\alpha_0-1}(1-\pi)^{\beta_0-1} \times \pi^x(1-\pi)^{nx} \\ &\propto \pi^{\alpha_0+x-1}(1-\pi)^{\beta_0+nx-1}, \end{aligned}$$

where we recognize the last line as the kernel of a beta distribution with parameters $\alpha_n = \alpha_0 + x$ and $\beta_n = \beta_0 + n - x$. It is easy to find the posterior in this case because the (beta) prior distribution we selected has a functional form that is similar to that of the (binomial) distribution of the data, yielding a (beta) posterior. In this case we say that the beta is a **conjugate prior** for binomial data.

Recall that the parameters of the prior beta distribution are $\alpha_0 = 330$ and $\beta_0 = 270$. If $x = 620$ of the $n = 1000$ respondents favor Proposition A, then the posterior has a beta distribution with parameters $\alpha_0 + x = 950$ and $\beta_0 + n - x = 650$. Look at Figure 1 for a visual comparison of the prior and posterior distributions. The density curves were plotted with the following S-Plus script. (Using `lines` is one way to plot more than one curve on the same axes.)

```
x <- seq(.45,.7,.001)
y <- dbeta(x,330,270);  z <- dbeta(x,950,650)
plot(x,z,type="l",ylim=c(0,35),
     xlab="Proportion in Favor",ylab="Density")
lines(x,y,lty=4,col=2)
```

The posterior mean is $950/(950 + 650) = 59.4\%$, a Bayesian point estimate of the actual proportion of the population currently in favor of Proposition A. Also, according to the posterior distribution, $P\{0.570 <$

$\pi < 0.618\} = 0.95$, so that a 95% **posterior probability interval** for the proportion in favor is (57.0%, 61.8%). (In S-Plus, `qbeta(.025, 950, 650)` returns `0.5695848`, and `qbeta(.975, 950, 650)` returns `0.6176932`.)

This probability interval resulting from Bayesian estimation is a straightforward probability statement. Based on the combined information from her prior distribution and from the polling data, the political consultant now believes it is very likely that between 57% and 62% of the population currently favors Proposition A. In contrast to a frequentist "confidence" interval, the consultant can use the probability interval without the need to view the poll as a repeatable experiment. ◇

Example 8.2.2 ***Weighing a beam (continued).*** Suppose that a particular beam is selected from among the beams available. Recall that, according to our prior distribution, the weights of beams in this population is $\mathsf{NORM}(200, 10)$ so that $\mu_0 = 100$ pounds and $\sigma_0 = 10$ pounds. The beam is weighed $n = 5$ times on a balance that gives unbiased, normally distributed readings with a standard deviation of $\sigma = 1$ pound. Denote the data by $\mathbf{x} = (x_1, \ldots, x_n)$, where the x_i are independent $\mathsf{NORM}(\mu, \sigma)$, and μ is the parameter to be estimated. Such data have the likelihood function

$$p(\mathbf{x}|\mu) \propto \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2\right],$$

where the distribution of μ is determined by the prior, and $\sigma = 1$ is known. Then after some algebra (see Problem 6), the posterior is

$$p(\mu|\mathbf{x}) \propto p(\mu)p(\mathbf{x}|\mu) \propto \exp[-(\mu - \mu_n)^2/2\sigma_n^2],$$

which is the kernel of $\mathsf{NORM}(\mu_n, \sigma_n)$, where

$$\mu_n = \frac{\frac{1}{\sigma_0^2}\mu_0 + \frac{n}{\sigma^2}\bar{x}}{\frac{1}{\sigma_0^2} + \frac{n}{\sigma_0}} \quad \text{and} \quad \sigma_n^2 = \frac{1}{\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}}.$$

It is common to use the term **precision** to refer to the reciprocal of a variance. If we define $\tau_0 = 1/\sigma_0^2$, $\tau = 1/\sigma^2$, and $\tau_n = 1/\sigma_n^2$, then the last two equations become

$$\mu_n = \frac{\tau_0}{\tau_0 + n\tau}\mu_0 + \frac{n\tau}{\tau_0 + n\tau}\bar{x} \quad \text{and} \quad \tau_n = \tau_0 + n\tau.$$

Thus, we say that the posterior precision is the sum of the precisions of the prior and the data, and that the posterior mean is a precision-weighted average of the means of the prior and the data.

In our example, $\tau_0 = 0.01$, $\tau = 1$, and $\tau_n = 5.01$. And the weights are $0.01/5.01 \approx 0.002$ for the prior mean and $5/5.01 \approx 0.998$ for the mean of the data. Thus, the posterior precision is almost entirely due to the precision of the data, and the value of the posterior mean is almost entirely due to the mean of the sample. In this case, the sample of five relatively high-precision observations is enough to concentrate the posterior and diminish the impact of the prior. (See Problem 7 and Figure 2 for the computation of the posterior mean and a posterior probability interval.) ♢

Example 8.2.3 ***Counting mice (continued).*** Suppose that a region of the island is selected where the gamma distribution with parameters $\alpha_0 = 4$ and $\kappa_0 = 1/3$ is a reasonable prior for λ. The prior density is $p(\lambda) \propto \lambda^{\alpha_0 - 1} e^{-\kappa_0 \lambda}$.

Over a period of about a year, traps are set out on $n = 50$ nights with the total number of captures $t = \sum_{i=1}^{50} x_i = 256$ for an average of 5.12 mice captured per night. Thus the Poisson likelihood function of the data is $p(\mathbf{x}|\lambda) \propto \prod_{i=1}^{n} \lambda^{x_i} e^{-\lambda} = \lambda^t e^{-n\lambda}$.

Then the posterior distribution is

$$p(\lambda|\mathbf{x}) \propto \lambda^{\alpha_0-1}e^{-\kappa_0\lambda} \times \lambda^t e^{-n\lambda} = \lambda^{\alpha_0+t-1}e^{-(\kappa_0+n)\lambda},$$

in which we recognize the kernel of the beta distribution with parameters $\alpha_n = \alpha_0 + t$ and $\kappa_n = \kappa_0 + n$. Thus the posterior mean is $\alpha_n/\kappa_n = (\alpha_0 + t)/(\kappa_0 + n)$. For our particular prior and data, the posterior mean is $(4+256)/(\frac{1}{3}+50) = 260/50.33 = 5.17$. Based on the posterior, a 95% probability interval for λ is $(4.56, 5.81)$. (In S-Plus, `qgamma(.025, 260, 50.33)` returns `4.557005`, and `qgamma(.975, 260, 50.33)` returns `5.812432`.) The prior and posterior densities are shown in Figure 3. ♢

8.3 Problems

1. In practice, the Beta family of distributions offers a rich variety of shapes for modeling priors to match expert opinion.

   (a) Beta densities are defined on the *open* unit interval. Show that parameter α controls behavior of the density function near 0. In particular, find $p(0^+)$ and $p'(0^+)$ in each of the following five cases: $\alpha < 1$, $\alpha = 1$, $1 < \alpha < 2$, $\alpha = 2$, and $\alpha > 2$. Evaluate each limit as being 0, positive and finite, ∞, or $-\infty$. (As usual, 0^+ means to take the limit as the argument approaches 0 through positive values.)

   (b) By symmetry, parameter β controls behavior of the density function near 1. Thus, combinations of the parameters yield 25 cases, each with its own "shape" of density. In which of these 25 cases does the density have a unique mode in $(0, 1)$? The number of possible inflection points of a beta density curve is 0, 1, or 2. For each of the 25 cases, give the number of inflection points.

(c) The S-Plus script below plots examples of each of the 25 cases, scaled vertically (with `top`) to show the properties in parts (a) and (b) about as well as can be done and yet show most of each curve. Compare this matrix of plots with your results above (α-cases are rows, β-cases are columns). In this display, which three of the 25 densities can be made assymetrical by choosing $\alpha \neq \beta$?

```
alpha <- c(.5, 1, 1.2, 2, 5);  beta  <- alpha
par(mfrow=c(5,5))    # Formats 5 x 5 matrix of plots
x <- seq(.001,.999,.001)
for (i in 1:5)
   {
   for (j in 1:5)
     {
     top <- .2 +1.2 * max(dbeta(c(.05,.2,.5,.8,.95),
                                       alpha[j],beta[i]))
     plot(x,dbeta(x,alpha[i],beta[j]),
          type="l", ylim=c(0,top), xlab="", ylab="")
     }
   }
```

2. In Example 8.1.1, we require a prior distribution with $\mathrm{E}(\pi) \approx 0.55$ and $P\{0.51 < \pi < 0.59\} \approx 0.95$. How might we find suitable parameters α and β for such a beta distributed prior?

(a) For a beta distribution, the mean is $\mu = \alpha/(\alpha + \beta)$, and the variance is $\sigma^2 = \alpha\beta/[(\alpha + \beta)^2(\alpha + \beta + 1)]$. Also, for unimodal and roughly symmetrical distributions on π the Empirical Rule states that $P\{\mu - 2\sigma < \pi < \mu + 2\sigma\} \approx 0.95$. Use these facts to find approximate values of α and β satisfying the requirements.

(b) The following S-Plus script finds integer values of α and β that may come close to satisfying the requirements, and then checks to see how well they succeed.

```
alpha <- 1:2000         # Trial values of alpha
beta <- .818*alpha      # Corresponding values of beta

# Vector of probabilities for interval (.51, .59)
prob <- pbeta(.59, alpha, beta)
        - pbeta(.51, alpha, beta)
prob.err <- abs(.95 - prob)  # Errors for probabilities

# Results: Target parameter values
t.alpha <- alpha[prob.err==min(prob.err)]
t.beta <- round(.818*t.alpha)
t.alpha; t.beta

# Checking: Achieved mean and probability
a.mean <- t.alpha/(t.alpha + t.beta)
```

```
a.mean
a.prob <- pbeta(.59, t.alpha, t.beta)
          - pbeta(.51, t.alpha, t.beta)
a.prob
```

What assumptions about α are inherent in the script? Why do we use $\beta = 0.818\alpha$? What values of α and β are returned? For integer values of the parameters, how close do we get to the desired values of $\mathrm{E}(\pi)$ and $P\{0.51 < \pi < 0.59\}$?

(c) If the desired mean is 0.56 and the desired probability in the interval $(0, 51, 0.59)$ is 90%, what values of the parameters are returned by a suitably modified script?

3. In Example 8.1.1, we require a prior distribution with $\mathrm{E}(\pi) \approx 0.55$ and $P\{0.51 < \pi < 0.59\} \approx 0.95$. If we were willing to use nonbeta priors, how might we find ones that meet these requirements?

   (a) If we were willing to use a normal distribution, what parameters μ and σ would satisfy the requirements?

   (b) Suppose we were willing to use a density function in the shape of an isosceles triangle. What equations for its sides would satisfy the requirements?

   (c) Plot three priors on the same axes: the beta density of Example 8.1.1 and the results of parts (a) and (b). Do you think the expert would object strongly to any of these probability models of her feelings about the distribution of π? (Use the method in Example 8.2.1 to put several plots on the same axes. Experiment: If `v <- c(0, 1, 1, 2, 2, 3)` and `w <- c(0, 0, 1, 1, 0, 0)`, then what does `lines(v, w)` add to an existing `plot`?)

4. Computational methods are often necessary if we multiply the kernels of the prior and likelihood and then can't recognize the result as the kernel of a known distribution. This can occur, for example, when we don't use a conjugate prior. We illustrate several computational methods using the polling situation of Examples 8.1.1 and 8.2.1 where we seek to estimate the parameter π.

   To begin, suppose we know the beta prior $p(\pi)$ (with $\alpha = 330$ and $\beta = 270$) and the binomial likelihood $p(x|\pi)$ (for $x = 620$ subjects in favor out of $n = 1000$ responding). But we have *not* been clever enough to notice the convenient beta form of the posterior $p(\pi|x)$. We wish to compute the posterior estimate of centrality $\mathrm{E}(\pi|x)$ and the posterior probability $P\{\pi > .6|x\}$ of a "big margin" in favor of the ballot proposition.

   From the *equation* in (8.1), we have $\mathrm{E}(\pi|x) = \int_0^1 \pi p(\pi)p(x|\pi)\,d\pi/D$ and $P(\pi > 0.6|x) = \int_{0.6}^1 p(\pi)p(x|\pi)\,d\pi/D$, where the denominator is