Bayes Estimation

January 20, 2006

1 Introduction

Our general setup is that we have a random sample $\mathbf{Y} = (Y_1, \ldots, Y_n)$ from a distribution $f(y|\theta)$, with θ unknown.

Our goal is to use the information in the sample to estimate θ. For example, suppose we are trying to determine the average height of all male UK undergraduates (call this θ). We observe 100 male undergraduates and find their average height $\bar{y}$ to be 69.74 inches. One possible question to ask is whether it is likely that θ is between 69 and 70 inches. This question might be interpreted as whether there is a high probability that θ is between 69 and 70. In other words we want to determine the probability, given our observed data, that θ is between 69 and 70. Formally, we want to determine

$$\Pr(69 < \theta < 70|\mathbf{Y} = \mathbf{y})$$

Using Bayes Theorem

$$\Pr(69 < \theta < 70|\mathbf{Y} = \mathbf{y}) = \frac{\Pr(\mathbf{Y} = \mathbf{y}|69 < \theta < 70)\Pr(69 < \theta < 70)}{\Pr(\mathbf{Y} = \mathbf{y})} \tag{1}$$

Notice this expression contains the quantity $\Pr(69 < \theta < 70)$. That is the probability that θ is between 69 and 70 *without* conditioning on the data. **To make probability statements about a parameter after observing data, you have to make probability statements about a parameter before observing data**. That statement motivates a split in statistical methods. In Bayesian inference, the *prior probabilities* are specified and then

Bayes theorem is used to make probability statements about the parameter as in equation (1). In frequentist inference such *prior probabilities* are considered nonsensical. The parameter θ is considered an unknown constant, not a random variable. Since it is not random, making probability statements doesn't make sense. A counterargument to this is that even if it is a constant, since it is unknown we may view it as a random variable. It might be one value, it might be another, it might be a third. This uncertainty may be considered randomness. Such arguments can and have continued for many years and are very interesting.

HOWEVER, if you are just interested in determining θ, Bayesian and frequentist methods both offer promising paths toward a solution. Often the two methods generate extremely similar answers anyway, making any argument about which one is better nearly meaningless from the standpoint of whether the method arrives at the correct value of θ. Specifically, often the MSEs of the two methods are identical or nearly identical. This is the viewpoint I am going to follow in this course. There are certain problems where the frequentist solution (usually Maximum Likelihood Estimation) is easier to follow, other problems where the Bayesian solution is easier to follow. Thus, a knowledge of both methods is useful.

2 Bayesian Estimation

If we have decided we are willing to specify *prior probabilities* about θ, some thought must be given as to what are reasonable values. The first step in Bayesian estimation is to formulate a *prior distribution*, $\pi(\theta)$, on θ. This prior distribution allows us to compute $\Pr(\theta \in A)$ for any set A.

The prior distribution is intended to represent the uncertainty about θ. Often you have very little information about θ, suggesting this prior should be very diffuse. For example, if we are trying to guess the average height (in feet) of male students at UK, we may know enough to realize the most people are between 5 and 6.5 feet tall, and therefore the mean should be between 5 and 6.5 feet, but we may not want to be more specific than that. We wouldn't, for example, want to specify $\pi(\theta) = N(5.8, 0.000001)$. Even though 5.8 feet may be a good guess, this prior places almost all its mass between 5.799995 and 5.800005 feet, indicating we are almost sure, *before seeing any data*, then the mean height is in this range. I'm personally not that sure, so I might choose a much more diffuse prior, such as setting $\pi(\theta) = Uni(5, 6.5)$, indicating that I'm sure the mean height is between 5 and 6.5 feet but every value in there seems about as likely as any other. Another possible prior is $\pi(\theta) = N(5.8, 0.4)$. This prior places about 95% of its mass between 5 and 6.6 feet, which is reflective of my uncertainty.

For the purposes of this class, I will specify the prior distribution for you to use. Fortunately, for many problems, including problems involving a simple random sample from most common distributions, all reasonably diffuse priors perform similarly. However, in some problems the choice of prior is extremely important, and there is a considerable amount of research on this question.

After a prior has been specified, we compute the *posterior distribution* of θ, from which all inferences will be made. Using Bayes Theorem, the conditional density of θ is

$$\pi(\theta|y_1, \ldots, y_n) = \frac{f(y_1, \ldots, y_n|\theta)\pi(\theta)}{f(y_1, \ldots, y_n)} = \frac{\left[\prod_i f(y_i|\theta)\right]\pi(\theta)}{\int_\Omega \left[\prod_i f(y_i|\theta)\right]\pi(\theta)d\theta} \tag{2}$$

where Ω is the entire parameter space. Often I will refer to the *Likelihood function*, which is defined as

$$L(y_1, \ldots, y_n|\theta) = \prod_i f(y_i|\theta)$$

Thus, equation (2) may also be written

$$\pi(\theta|y_1, \ldots, y_n) = \frac{L(y_1, \ldots, y_n|\theta)\pi(\theta)}{\int_\Omega L(y_1, \ldots, y_n|\theta)\pi(\theta)d\theta} \tag{3}$$

The posterior $\pi(\theta|Y_1, \ldots, Y_n)$ is a distribution over θ and has all the usual properties of a distribution. In particular

1. The posterior distribution integrates to 1.

$$\int \pi(\theta|\mathbf{y})d\theta = 1$$

2. We may compute the posterior probability that θ is in a set A by

$$P(\theta \in A|\mathbf{y}) = \int_A \pi(\theta|\mathbf{y})d\theta$$

3. The posterior distribution has a mean and variance, just like any other distribution. If we have to make a guess as to the exact value of θ, one commonly used guess is the *posterior mean*

$$E_\pi[\theta|\mathbf{y}] = \int \theta\pi(\theta|\mathbf{y})d\theta$$

Example 1

Suppose that $Y_1, \ldots, Y_n \sim Exp(\theta)$ (with density $\theta \exp\{-\theta y\}$ on the interval $(0, \infty)$). We observe $y_1 = 3$, $y_2 = 6$, and $y_3 = 10$ and the prior on θ is $\pi(\theta) = 3\theta^2/19$ on the interval $[2, 3]$. To use equation (3) we need to compute the likelihood

$$L(y_1, \ldots, y_n|\theta) = \prod_i \theta e^{-\theta y} = \theta^n e^{-\theta \sum_i y_i}$$

Placing this and $\pi(\theta)$ in equation (3), the posterior density is

$$\pi(\theta|y_1, \ldots, y_n) = \frac{\theta^n e^{-\theta \sum_i y_i}(3\theta^2/19)}{\int_\Omega \theta^n e^{-\theta \sum_i y_i}(3\theta^2/19)d\theta}$$

over the range $[2, 3]$ (the same range as the prior. Outside this interval the prior density is 0 and therefore the posterior density is 0 as well). Performing some algebra, this simplifies to

$$\pi(\theta|y_1, \ldots, y_n) = \frac{\theta^{n+2} e^{-\theta \sum_i y_i}}{\int_2^3 \theta^{n+2} e^{-\theta \sum_i y_i} d\theta}$$

Now we must determine the value of the denominator. Often (see the next section) this calculation is easy. Here it is not. There is no analytical formula for the denominator, even in terms of Gamma functions (the integral is from 2 to 3, not 0 to ∞). However, we do know $n = 3$ and $\sum_i y_i = 19$, so the integral is

$$\int_2^3 \theta^5 e^{-19\theta} d\theta$$

This can be done using numerical integration techniques such as Simpson's rule or the trapezoidal rule. Doing this on my computer, I found the integral to be $6.06194e - 17$, indicating the entire posterior distribution is

$$\pi(\theta|\mathbf{y}) = \frac{\theta^5 e^{-19\theta}}{6.06194e-017}$$

over the range $[2,3]$. To compute the posterior mean, we use

$$\int \theta\pi(\theta|\mathbf{y})d\theta = \int_2^3 \theta\frac{\theta^5 e^{-19\theta}}{6.06194e-017}d\theta$$

I computed this quantity by numerical integration as well, finding the answer to be 2.0601.

2.1 Avoiding Integration

Often the most difficult (tedious, annoying, take your pick) part of Bayesian inference is computing the integral in the denominator of the posterior distribution. Sometimes numerical techniques are even required as in the previous example. However, often the integral can be **completely ignored**. WOOHOO! The posterior distribution is

$$\pi(\theta|y_1,\ldots,y_n) = \frac{L(y_1,\ldots,y_n|\theta)\pi(\theta)}{\int_\Omega L(y_1,\ldots,y_n|\theta)\pi(\theta)d\theta}$$

Let's just name

$$C = \frac{1}{\int_\Omega L(y_1,\ldots,y_n|\theta)\pi(\theta)d\theta}$$

for the moment (we are going to use C as a catchall name for any term not involving θ). Thus

$$\pi(\theta|y_1,\ldots,y_n) = (C)L(y_1,\ldots,y_n|\theta)\pi(\theta)$$

Often, we can compute the posterior distribution using the following steps

1. Simplify $L(\mathbf{y}|\theta)\pi(\theta)$ as far as possible. I will usually be specifying "nice" priors for you that allow the likelihood and prior to combine.

2. Pull all terms *not* involving θ out into C, the normalizing constant.

3. The remaining terms form the *kernel* of the distribution for θ. You must now try to recognize what distribution the kernel represents.

Example

Let $Y_1, \dots, Y_n \sim \mathit{Geometric}(p)$ (with the parameterization where the possible values are 0, 1, 2, and so on) and let p have a Beta(2,3) prior, so the prior density is

$$\frac{\Gamma(5)}{\Gamma(2)\Gamma(3)} p^1 (1-p)^2 = 12p(1-p)^2$$

The likelihood is

$$\prod_i p(1-p)^{y_i} = p^n (1-p)^{\sum_i y_i}$$

The product of the likelihood and prior (the numerator of the posterior distribution) is

$$p^n (1-p)^{\sum y_i} 12p(1-p)^2 = 12p^{n+1}(1-p)^{\sum y_i + 2}$$

Rewriting the normalizing constant in the denominator as C, we have the posterior distribution of p is

$$\pi(p|y_1, \dots, y_n) = (C)12p^{n+1}(1-p)^{\sum y_i + 2}$$

Notice that the 12 does not depend on p, it's just a number, so we may pull it into C (remember C is just a unknown constant), leaving the kernel

$$p^{n+1}(1-p)^{\sum y_i + 2}$$

Only one distribution has this kernel. Looking through the distributions, we find a Beta distribution has a kernel of the form

$$p^{\alpha-1}(1-p)^{\beta-1}$$

Matching up parameter values, the posterior distribution is a $\mathit{Beta}(n+2, 3+\sum y_i)$ distribution. The Bayes estimate is the posterior mean, which for a $\mathit{Beta}(n+2, 3+\sum y_i)$ is $(n+2)/(\sum y_i + n + 5)$.

Suppose we wished to use a general $Beta(\alpha, \beta)$ prior. We would like a formula for the posterior in terms of α and β. We proceed as before, finding the prior density to be

$$\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}p^{\alpha-1}(1-p)^{\beta-1}$$

The likelihood is unchanged, so the product of the prior and likelihood simplifies is

$$p^n(1-p)^{\sum y_i}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}p^{\alpha-1}(1-p)^{\beta-1} = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}p^{n+\alpha-1}(1-p)^{\sum y_i+\beta-1}$$

The prior parameters α and β are treated as fixed constants (eventually we will give them numerical values, we are just deriving a general formula for the moment). Thus the Gamma functions in front may be considered part of the normalizing constant C, leaving the kernel

$$p^{n+\alpha-1}(1-p)^{\sum y_i+\beta-1}$$

Using the same reasoning as before, this is the kernel of a $Beta(n+\alpha, \sum y_i+\beta)$ distribution, with posterior mean

$$\frac{n+\alpha}{\sum y_i+n+\alpha+\beta}\ .$$