PARAMETER ESTIMATION: BAYESIAN APPROACH.

These notes summarize the lectures on Bayesian parameter estimation.

1. Beta Distribution

We’ll start by learning about the Beta distribution, since we end up using it in some key examples. It’s just another density that’s also quite popular in statistics.

A random variable $X$ is said to have the Beta($\alpha, \beta$) distribution if its density is given by

$$f(x) = \begin{cases} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1} (1-x)^{\beta-1} & 0 \leq x \leq 1 \\ 0 & \text{otherwise.} \end{cases}$$

The $\Gamma(\cdot)$ function is the same function we have used in the Gamma distribution. It is defined as $\Gamma(r) = \int_0^\infty x^{r-1} e^{-x} dx$ for $r > 0$ and satisfies the following properties:

- $\Gamma(k) = (k-1)!$ for $k$ an integer.
- $\Gamma(1/2) = \sqrt{\pi}$
- $\Gamma(r) = (r-1)\Gamma(r-1)$
- $\frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} = \int_0^1 x^{\alpha-1} (1-x)^{\beta-1} dx$,

which tells us that the Beta density does indeed integrate to 1.

A few other important things to note:

i. $E[X] = \frac{\alpha}{\alpha+\beta}$

ii. $V(X) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$

iii. The mode of the distribution lies at $\frac{\alpha-1}{\alpha+\beta-2}$ (in class I think I had $\alpha - \beta - 1$ in the denominator, which is wrong).

iv. Beta(1,1) is just the uniform on $[0,1]$.

To understand what the Beta does it’s easiest to just look at a picture of its density - see Figure 1.

Date: December 2, 2007.
2. Bayesian Setup

In the Bayesian parametric approach we model several quantities. First of all, we model the behaviour of the data \( \text{given} \) the parameter \( \theta \). This is the “same” as the likelihood in frequentist models, and we write is as \( f(x|\theta) \). \( x \) denotes the vector of observed data \( x = \{x_1, \ldots, x_n\} \). Most often we will assume (here) that our data is a random sample (iid) and hence \( f(x|\theta) = \Pi_{i=1}^n f(x_i|\theta) \).

Next, we also model the behaviour of \( \theta \). We do this by specifying a prior pdf/pmf, \( \pi(\theta) \), on the parameter \( \theta \). This means that we no longer think of \( \theta \) as fixed, but random (and the conditioning in \( f(x|\theta) \) is proper). Alternatively, you can think of \( \pi(\theta) \) as describing our prior belief on \( \theta \).

All inference on \( \theta \) is made from the posterior on \( \theta \) given by

\[
\pi(\theta|x) = \frac{f(x|\theta)\pi(\theta)}{\int f(x|\theta)\pi(\theta)d\theta} \\
\propto f(x|\theta)\pi(\theta)
\]

The posterior distribution is a way in which we combine the data with our prior beliefs.

What numbers do you report then? You could report the mean, variance, median, mode of the distribution... whichever you think best works in your situation. However, there is some theory to back up your choice. For example, you can report a number \( \hat{\theta}_B \) which minimizes the expected loss (distance) function from \( \theta \) to \( \hat{\theta} \) under the
posterior. The loss function can be anything sensible. Most often this is chosen to be 
$$\ell(\theta, \hat{\theta}) = (\theta - \hat{\theta})^2.$$ In this case we already know that
$$\hat{\theta}_B = \arg\min_{\hat{\theta}} E\left[\left(\theta - \hat{\theta}\right)^2 \mid x\right] = E[\theta \mid x].$$
In words, this is the posterior mean. If we had instead chosen 
$$\ell(\theta, \hat{\theta}) = |\theta - \hat{\theta}|,$$ then 
$$\hat{\theta}_B = \arg\min_{\hat{\theta}} E\left[|\theta - \hat{\theta}| \mid x\right] = \text{the median of the posterior.}$$

An “estimate” of $\theta$ is never sufficient, and hence an interval which summarizes the error of our belief in $\hat{\theta}_B$ is required. Under the frequentist umbrella, this was called the confidence interval. In Bayesian data analysis, the term **credible interval** is used. This interval is simply two numbers $u, l$ so that $P(u \leq \theta \leq l \mid x) = 1 - \alpha$. Although there are no hard rules on how to choose these numbers, something sensible is called for. For example, you may choose $(u, l)$ so that this interval is the shortest one possible. This is equivalent to choosing the interval so that it sits under the highest posterior density (and then it’s called the **HPR**). Often though something simple is chosen, such as the **central** interval. If the posterior pdf is unimodal, these will probably be quite similar, if not exactly the same. If the posterior density if bimodal, you can easily see that these could be quite different.

In practice one of three things is possible:

- You “recognize” the posterior distribution; that is, it is one of the standard distributions that you are familiar with. Then working with the posterior is quite simple. In this course this is the only case we have time to look at. Often this happens when the prior and likelihood form a **conjugate** family. That is, when the posterior is in the same family of distributions as the prior.
- You don’t recognize the posterior, but are still able to calculate an exact expression for it. Calculations of the posterior mean or a credible interval (and other quantities) can either be done directly or through simulation on a computer (in this case you can always use the result from the Bonus problem on Assignment 2 to simulate the RV).
- Not only are you not able to recognize, you cannot even write the density in closed form! The problem always comes from not being able to calculate the normalizing constant $\int f(x \mid \theta) \pi(\theta) d\theta$. Miraculously, there are very sophisticated simulation methods that allow you to sample from the posterior regardless! Currently, the most widely used of these is called Markov Chain Monte Carlo (MCMC). There is even a package in R called **WinBUGS** which
One of the greatest questions in Bayesian data analysis is the choice of the prior distribution. In theory, this reflects your prior beliefs on the parameter $\theta$. In practice, this is much more difficult to achieve. A natural question to ask is how sensitive your analysis is to the choice of prior. These are of course things that Bayesian statisticians have ways of dealing with (or at least methods exist). Again, we cannot get into this in our course. However, some very nice examples where a prior is selected are given in the Lavine text, and so I strongly recommend that you read over that section of his text.

One way that we can choose a prior is to specify a noninformative one. An example of this is the Uniform$[0,1]$ distribution as a prior for the Bernoulli parameter. This can be very appealing in some situations. However, sometimes there is no way to choose a noninformative prior which is proper. There are situations where an improper noninformative prior gives a proper posterior though.

Example. Inference for Bernoulli with Beta prior. Pictures appear in Figures 2 and 3.

Example: estimating the probability of a female birth. This example is taken directly from [GCSR]. As a specific application of the binomial model, we consider the estimation of the sex ratio within a population of human births. The proportion of births that are female has long been a topic of interest both scientifically and to the lay public. Two hundred years ago it was established that the proportion of female births in European populations was less than 0.5. The currently accepted value of the proportion of female births in very large European-race populations is 0.485.

As a specific example of a factor that may influence the sex ratio, we consider the maternal condition placenta previa, an unusual condition of pregnancy in which the placenta is implanted very low in the uterus, obstructing the fetus from a normal vaginal delivery. An early study concerning the sex of placenta previa births in Germany found that of a total of 980 births, 437 were female (437/980 $\approx$ 0.446).

How much evidence does this provide for the claim that the proportion of female births in the population of placenta previa births is less than 0.485, the proportion of female births in the general population?

Assuming a uniform prior distribution for the probability of a female birth, the posterior is Beta(438, 544). Exact summaries of the posterior distribution can be obtained from the properties of the beta distribution: the posterior mean is 0.446 with the posterior standard deviation of 0.016. Exact posterior quantiles can be obtained using numerical integration of the beta density, which in practice we perform by a
computer function call; the median is 0.446 and the central 95% posterior interval is [0.415, 0.477].

**Example.** Normal example where $\sigma^2$ is known. The likelihood is Normal with mean $\mu$ and known variance. We place a prior on $\mu$ which is also normal.

**Example.** Poisson example. The likelihood is Poisson with rate $\theta$. The prior on $\theta$ is chosen from the Gamma family of distributions.

![Graphs of distributions](image)

**Figure 2.** Bernoulli example with Beta prior.
PARAMETER ESTIMATION: BAYESIAN APPROACH.

$\alpha = 0.5, \beta = 0.8, n = 10, \bar{x} = 0.5$, (blue=prior)

Figure 3. Bernoulli example with Beta prior.

REFERENCES


Prepared by Hanna Jankowski
Department of Statistics, University of Washington
Box 354322, Seattle, WA 98195-4322 U.S.A.
E-mail: hanna@stat.washington.edu