An introduction to Bayesian statistics: Inference with the Binomial Distribtution

AIMS 2009

December 14, 2009

(Updated to correct notation on Jeffreys Prior - thanks to Madeleine)
(Updated to correct calculus on posterior mode - thanks to Everyone)

1 An introduction to Bayesian Statistics

- In classical (also called Frequentist) statistics we assume that our data $x_1, x_2, \ldots, x_n$ follow a distribution $f(x|\theta)$ determined by some fixed but unknown parameter θ. The aim of inference is to make a statement about this parameter.
- In Bayesian statistics, we believe that the parameter is a random variable. We aim to make statements about its distribution, conditional on the observed data. Before performing the experiment, we summarise our knowledge about the distribution of the random parameter θ in a prior density, denoted $g(\theta)$.
- Once we observe the data, we update our beliefs about the parameter using Bayes' rule, and describe our beliefs by means of a posterior distribtuion:

$$post(\theta|x) = \frac{f(x|\theta)g(\theta)}{\int f(x|\theta)g(\theta)d\theta}$$

  where $g(\theta)$ denotes the prior and $f(x|\theta)$ denotes the Likelihood $L(\theta)$.
- As when we work with the likelihood, the value under the function is not so important, so we can often ignore many constants, and so can often carry out inference using:

$$post(\theta|x) \propto f(x|\theta)g(\theta) \tag{1}$$

- This is often summarised as:

$$\boxed{\text{Posterior} \propto \text{Likelihood} \times \text{Prior}}$$

2 Inference for a proportion

Consider the binomial density:

$$f(x|\pi) = \binom{n}{x}\pi^x(1-\pi)^{n-x}; \text{ for } x = 0, 1, \ldots, n$$

This is a function of x given π.

Conveniently, the likelihood function can be given as:

$$L(\pi|x) = \binom{n}{x}\pi^x(1-\pi)^{n-x} \text{ for } 0 \leq \pi \leq 1. \tag{2}$$

This is a function of π given x.

Now we wish to apply Bayes theorem from (1). This gives:

$$post(\pi|x) = \frac{L(\pi|x)g(\pi)}{\int_0^1 L(\pi|x)g(\pi)d\pi}$$

but we can work with the simpler:

$$post(\pi|x) \propto L(\pi|x)g(\pi).$$

If we apply a uniform prior,:

$$g(\pi) = 1 \text{ for } 0 \leq \pi \leq 1$$

we obtain an expression for the posterior as follows:

$$post(\pi|x) = \binom{n}{x}\pi^x(1-\pi)^{n-x} \tag{3}$$

Remember that this is a function of π.

This is a $Beta(a', b')$ density, where $a' = x+1$ and $b' = n-x+1$ (or our posterior is a $Beta(x+1, n-x+1)$ density.

2.1 Sketching the functions

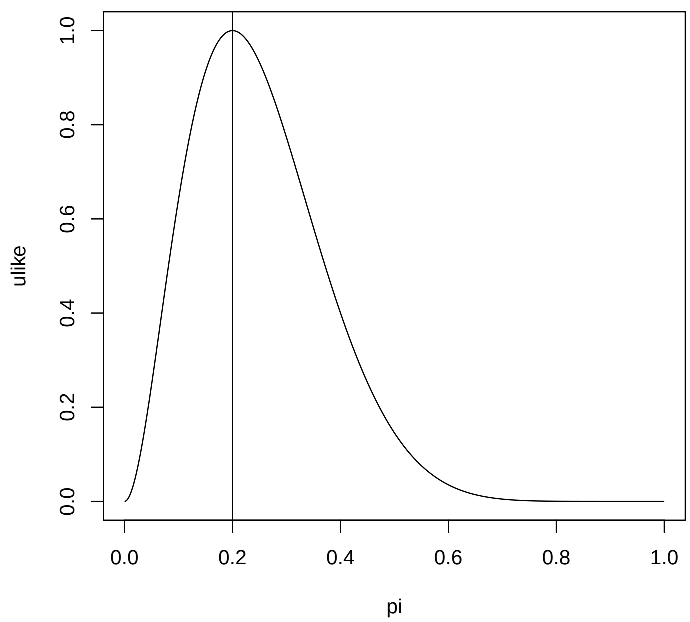
In order to explore the implications of Bayesian inference, we shall sketch these functions (prior, posterior) using a computer.

First, note that although $0 \leq \pi \leq 1$ construct a grid over that range excluding 0 and 1 but including a number values between):

Next write a short function that evaluate the posterior assuming a uniform prior. In other words, you want to evaluate the function given in (3) for all values of pi. Note that you can ignore constants and that you only need to evaluate:

$$post(\pi|x) = \pi^x(1-\pi)^{n-x}$$

You should obtain a plot as below. I have superimposed a vertical line to denote the maximum likelihood estimate.



You should see for our simple example that the posterior mode is the same as the maximum likelihood estimate. Let's now consider another prior.

3 Beta Density

The Beta density is a convenient prior as it is a "conjugate prior" and hence analytical results are available. The Beta density is given by:

$$g(\pi|a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\pi^{a-1}(1-\pi)^{b-1} \text{ for } 0 \leq \pi \leq 1 \tag{4}$$

If we multiply together our Beta prior in 4 with our binomial likelihood in 2, and *discard* any constants not involving π we get:

$$post(\pi|x) \propto \pi^{a+x-1}(1-\pi)^{b+n-x-1}$$

This is a $Beta(a', b')$ density with $a' = a + x$ and $b' = b + n - x$, i.e. a $Beta(a + x, b + n - x)$ density.

You should modify your earlier code to sketch this function.

It is also informative to plot these prior.

3.1 The Jeffreys Prior

Recall that if we transform the unknown parameter from π to ψ (more formally $\psi(\pi)$, then:

$$\frac{\partial \log L(\psi|x)}{\partial\psi} = \frac{\partial \log L(\pi|x)}{\partial\pi}\Big|\frac{d\pi}{d\psi}\Big|$$

If we square, and take expectations over values of x we find:

$$I(\pi) = I(\pi)(\Big|\frac{d\pi}{d\psi}\Big|)^2$$

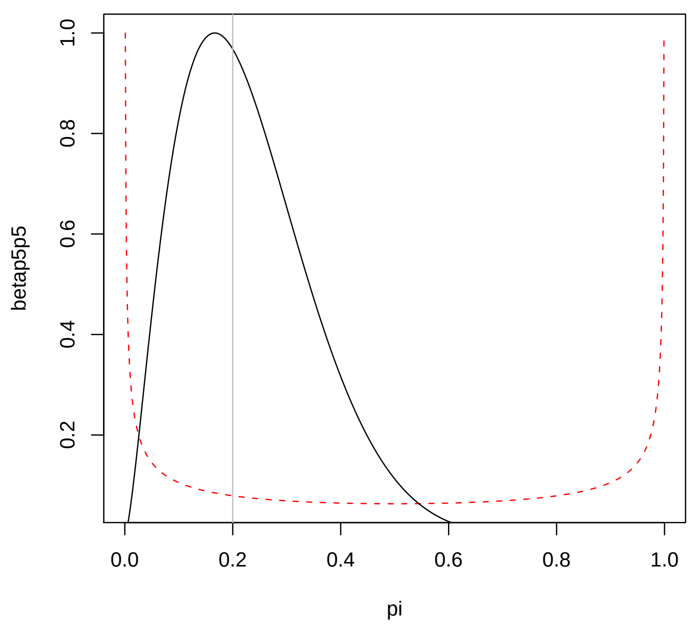
Hence the prior:

$$p(\pi) \propto \sqrt{I(\pi)}$$

if used, will after transformation yield:

$$p(\psi) \propto \sqrt{I(\psi)}$$

And so you should pay particular attention to this *Jeffrey's Prior*. Here, this is a Beta(0.5,0.5) distribution. Can you plot the posterior assuming a Jeffreys prior, and compare your results with those from the other priors.



We have plotted the "prior" (dashed) along with the posterior. It is clear here that the posterior has a very different shape.

3.2 Self-assessed exercise

We are interested in examining the effect of prior on the posterior.

- Try the following priors: Beta(0.5, 0.5), Beta(1,1), Beta(2,2), Beta(5,5)
- Try the following data: (x=2,n=10), (x=20,n=100), (x=200,n=1000)

In total, there are twelve combinations of prior and data. Carefully examine the posterior distribution.

- What is the effect of increasing the size of a and b in the prior in terms of the posterior distribution for π
- What is the effect of increasing n in the data in terms of the posterior distribution for π.

4 Summarising the Posterior

Whilst we need to note that the Posterior is a function, we need to summarise it somehow. Here are a few summary measures.

4.1 Posterior mode

This is the value that maximises the posterior distribution. For our Beta conjugate prior, we need to differentiate the corresponding posterior and set the solution to zero.

If we have:

$$post(\pi|x) \propto \pi^{(a+x)-1}(1-\pi)^{(b+n-x)-1}$$

then

$$\begin{aligned}\frac{\partial post(\pi|x)}{\partial \pi} &= \left\{[(a+x)-1]\pi^{[(a+x)-2]}\right\}\left\{(1-\pi)^{[(b+n-x)-1)]}\right\} - \\ &\quad \left\{\pi^{[(a+x)-1]}\right\}\left\{[(b+n-x)-1])(1-\pi)^{[(b+n-x)-2]}\right\}\end{aligned}$$

Which, when set equal to zero and solved gives:

$$Mode = \frac{(a+x)-1}{[(a+x)-1]+[(b+n-x)-1]}$$

Recall that the Mode for the $Beta(a,b)$ density is given by $\frac{a-1}{a+b-2}$ so we have in fact the posterior Mode for the Binomial parameter π when using a $Beta(a,b)$ prior is given by $\frac{a'-1}{a'+b'-2}$ where $a' = a+x$ and $b' = b+n-x$

4.2 Posterior median

We need to solve:

$$\int_0^{Median} g(\pi|x)d\pi = 0.5$$

Which I think in this case needs to be done numerically (or by simulation). We can define other quantiles (and the inter-quartile range) in the same way.

4.3 Posterior mean

We need to find:

$$Mean_p = \int_0^1 \pi g(\pi|x)d\pi$$

This can be solved analytically:

$$Mean_p = \frac{(a+x)}{(a+x)+(b+n-x)}$$

Again, recall that for the $Beta(a,b)$ density the mean is given by $\frac{a}{a+b}$ so we have again the result that the posterior mean for the Binomial parameter π when using a $Beta(a,b)$ prior is given by $\frac{a'}{a'+b'}$ where $a' = a+x$ and $b' = b+n-x$.

5 Posterior variance

$$Var_p(\pi|x) = \int_0^1 (\pi - m_p)^2 g(\pi|x)d\pi$$

We can obtain the analytical expression for this:

$$Var_p^(\pi|x) = \frac{(a+x)(b+n-x)}{\big((a+x)+(b+n-x)\big)^2\big((a+x)+(b+n-x)+1\big)}$$

As again, we know that the expression for the Variance of a $Beta(a,b)$ variable is $\frac{ab}{(a+b)^2(a+b+1)}$ so we see that the posterior variance for the Binomial parameter π when using a $Beta(a,b)$ prior is $\frac{a'b'}{(a'+b')^2(a'+b'+1)}$ where again $a' = a+x$ and $b' = b+n-x$

6 Credible interval

Denote our estimate of $mean_p$ by m_p and our estimate of Var_p by s_p^2. We will use a Bayesian Central Limit Theorem (but do note that this is an asymptotic statement and we have some fairly small sample sizes), and assume that the posterior is approximately Normal with mean m_p and variance s_p^2.

Accordingly, we can use the formula

$$\text{Credible Interval} = m_p \pm z_{1-\alpha/2} s_p$$

For example, for an approximation to the 95% credible interval we would use $z_{1-\alpha/2} = 1.96$. As mentioned in an earlier lecture, using ± 1.96 defines the middle 95% of the probability density of a standard normal distribution. This approximation works reasonably well for $(a+x-1) \geq 10$ and $(b+n-x-1) \geq 10$.

Next week, we shall look at working this out exactly, rather than relying on Normal approximations.

7 Notes on using the Beta Prior

Hopefully you have noticed that *our choice of prior* can alter the conclusions we might draw. Here are some notes to consider when using a Beta prior.

- When $a < b$ the density has more mass in the lower half. When $a > b$ the density has more mass in the upper half. When $a = b$ the density is symmetric. When $a = \frac{1}{2}$ more weight is given to values near zero and when $b = \frac{1}{2}$ more mass is given to values near 1.

What value of a and b gives the Uniform density?

- The mean of the beta distribution is $\frac{a}{a+b}$, and the standard deviation is $\sqrt{\frac{ab}{(a+b)^2(a+b+1)}}$. The prior also has an interpretation in terms of *equivalent sample size*. Noting that a proportion $\hat{\pi} = \frac{x}{n}$ has a variance $\frac{\pi(1-\pi)}{n}$ we can equate this variance at π_0 the prior mean to get:

  $$\frac{\pi_0(1-\pi_0)}{n_{eq}} = \frac{ab}{(a+b)^2(a+b+1)}$$

  So we can obtain the fact that

  $$n_{eq} = a + b + 1$$

So we can, if we wish, choose our prior to take account of any data that might already exist. Or we can select a Uniform prior if we don't wish to make such assertions.