

22S:138
Bayesian Statistics

One-Parameter Models: Continued

Lecture 4
Sept. 5, 2008

Kate Cowles
374 SH, 335-0727
kcowles@stat.uiowa.edu

Bayes' rule for parameters and data

- To make probability statements about an unknown parameter θ given data \mathbf{y} , we must begin with a *model* specifying a *joint probability distribution* for θ and \mathbf{y} .
- joint probability mass function if θ and \mathbf{y} are discrete, joint probability density function if they are continuous
- in either case, we will denote the joint distribution as $p(\theta, \mathbf{y})$
- By definition

$$p(\theta, \mathbf{y}) = p(\theta) p(\mathbf{y}|\theta)$$

This is the product of

- the marginal distribution of θ — the prior
- the distribution of the data given θ — the likelihood

- by basic property of conditional probability

$$p(\theta|\mathbf{y}) = \frac{p(\theta, \mathbf{y})}{p(\mathbf{y})} = \frac{p(\theta) p(\mathbf{y}|\theta)}{p(\mathbf{y})}$$

where

$$p(\mathbf{y}) = \sum p(\theta) p(\mathbf{y}|\theta)$$

where the sum is over all possible values of θ if θ is discrete, or

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

if θ is continuous.

- Note that $p(\mathbf{y})$ does *not* depend on θ . If we are conditioning on a dataset of known values \mathbf{y} , then $p(\mathbf{y})$ may be considered a constant.
- $\frac{1}{p(\mathbf{y})}$ is the normalizing constant that makes $p(\theta|\mathbf{y})$ a valid density
 - Recall that, if $p(x)$ is a valid probability density function

$$\int_{\Omega} p(x) dx = 1$$

where Ω represents the range of all possible values of x .

- Thus, the *unnormalized* posterior is

$$p(\theta|\mathbf{y}) \propto p(\theta) p(\mathbf{y}|\theta)$$

Back to the binomial example: a “conjugate” prior

- A common way to construct a prior distribution is to assume that the prior is a member of a particular parametric family of densities.
 - Then choose the parameters of the prior so that the prior represents prior beliefs as closely as possible.
- When possible, it is very convenient analytically to choose the prior from a parametric family that has the same functional form as the likelihood function.
 - called a “conjugate” prior

Computing the posterior distribution with a conjugate prior

- Recall:

$$p(p|\mathbf{y}) \propto p(p) L(p)$$
- So:

$$p(p|\mathbf{y}) \propto p^{\alpha+y-1} (1-p)^{\beta+n-y-1}$$
- This is the kernel of another Beta density!

$$p(p|\mathbf{y}) = \text{Beta}(\alpha + y, \beta + n - y)$$
- This is the implication of conjugacy:
 - The Beta prior distribution is a conjugate family for the binomial likelihood.
 - Therefore, the posterior distribution follows the same parametric form as the prior.

- Recall the binomial likelihood:

$$L(p) = \binom{n}{y} p^y (1-p)^{n-y} \\ \propto p^y (1-p)^{n-y}, \\ 0 < p < 1$$

- The *Beta* family of densities has the same functional form: If

$$p(p) = \text{Beta}(\alpha, \beta)$$

then

$$p(p) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha) \Gamma(\beta)} p^{\alpha-1} (1-p)^{\beta-1} \\ \propto p^{\alpha-1} (1-p)^{\beta-1}, \\ 0 < p < 1$$

Choosing the parameters of a Beta distribution to match prior beliefs

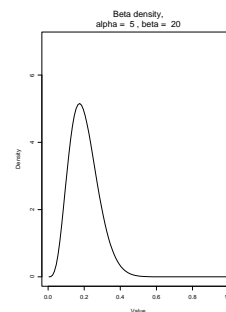
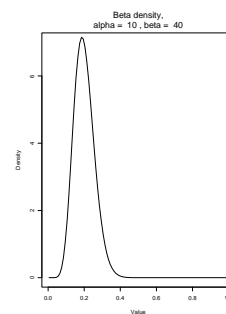
- Strategy 1: Graph some beta densities until you find one that matches your beliefs.
- Strategy 2: Note that a $\text{beta}(\alpha, \beta)$ prior is equivalent to a previously-observed dataset with $\alpha - 1$ successes and $\beta - 1$ failures.
- Strategy 3: Solve for the values of α and β that yield:
 - The desired mean (The mean of a $\text{beta}(\alpha, \beta)$ density is $\frac{\alpha}{\alpha + \beta}$).
 - The desired “equivalent prior sample size” — $\alpha + \beta - 2$
- Strategy 4: Solve for α and β after specifying:
 - What is your subjective probability that the first observation in the new data would be a success?

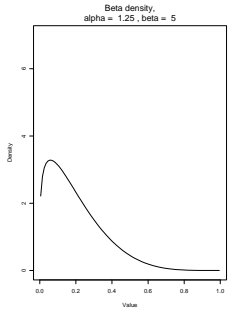
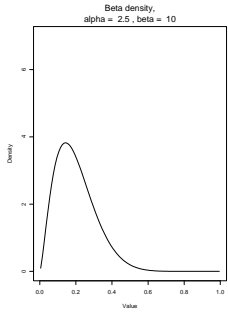
- If the first observation were to be a success, what would be your subjective probability that the *second* observation would also be a success.
- The new data *must not* be used in any way in assessing the prior!
 - We’ll see shortly why that would make inference invalid.

Back to the quitting-school-because-of-rising-tuition example

- We are attempting to construct a reasonable prior *before* we see the results of the survey of 50 UI students.
- Suppose that we read that such a survey has already been taken at ISU and
 - 25 students were interviewed
 - 5 said they would quit school; 20 said they would not
- By strategy 2, this would suggest a *beta*(6, 21) prior.
- Alternatively, we might want a prior with mean .2, the same as \hat{p} from the ISU data.
- We might want to use the ISU data but “down-weight” it, since ISU students might not be just like UI students.

- One possibility is to look at the graphs of several different beta distributions, all with the same mean .2 but with smaller and smaller “equivalent sample sizes.”





Computing and graphing the posterior distribution

- Suppose we chose the Beta(10, 40) prior.
- We then gather our own data on $N = 50$ IU students, and get $y = 7$ “successes” and $n - y = 43$ “failures.”
- Then

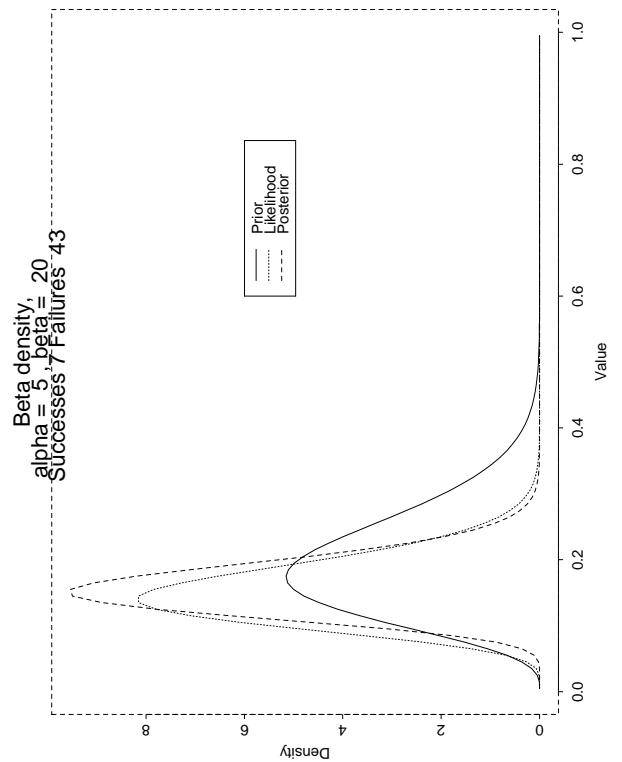
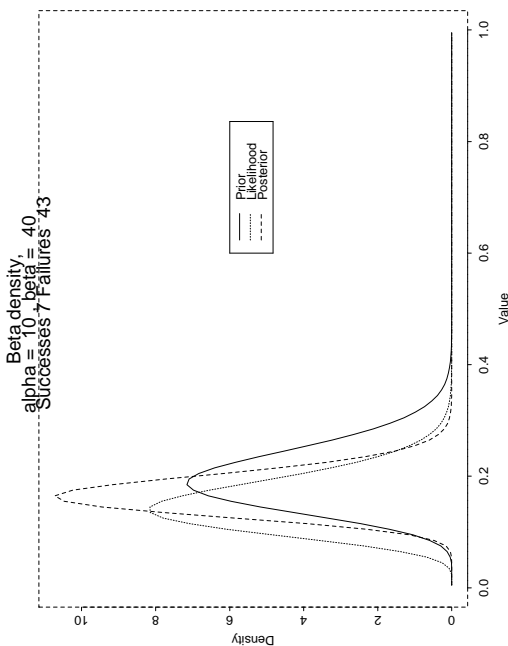
$$p(p|y) \propto p^{\alpha+y-1} (1-p)^{\beta+n-y-1}$$

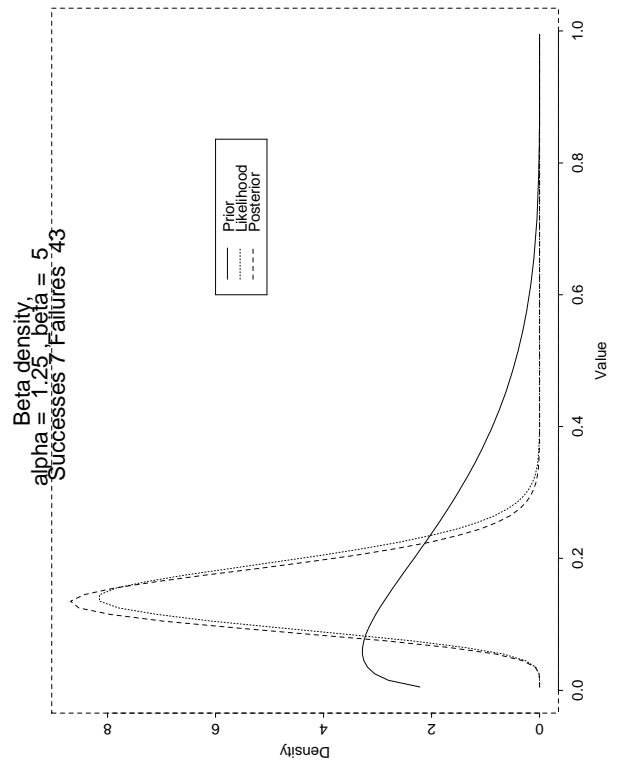
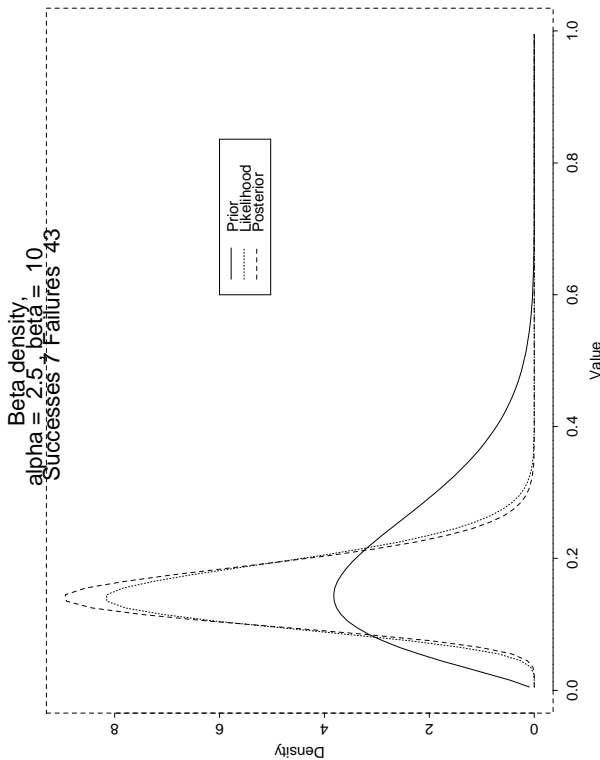
$$p^{17-1} (1-p)^{83-1}$$

This is a Beta(17, 83) density.

- With the Beta(1.25, 5) prior, the posterior would be a Beta(8.25, 48).

Plotting the prior, the likelihood, and the posterior





The posterior mean

- The mean of the posterior distribution is often used as the Bayesian *point estimate* of a parameter.
 - We will learn later that it minimizes squared error loss.
- For a beta prior and binomial likelihood, the posterior mean is:

$$E(p|y) = \frac{\alpha + y}{\alpha + y + \beta + n - y}$$

- In our example, with the Beta(10, 40) prior

$$E(p|y) = \frac{17}{100} = 0.17$$

- For a beta prior and binomial likelihood, the posterior mean is always *between* the prior mean and the value $\frac{y}{n}$ computed from the current data.

In our example, the prior mean was 0.20, and $\frac{y}{n} = 0.14$.

- In our example, if we instead had used the Beta(1.25, 5) prior

$$E(p|y) = \frac{8.25}{56.25} = 0.147$$

More on the posterior mean

- The posterior mean is a weighted average of the prior mean and the MLE \hat{p} .
- If we denote the posterior mean by μ_{post}

$$\begin{aligned}\mu_{post} &= \frac{\alpha + y}{\alpha + \beta + n} \\ &= \lambda \frac{\alpha}{\alpha + \beta} + (1 - \lambda) \frac{y}{n}\end{aligned}$$

where $\lambda = \frac{\alpha + \beta}{\alpha + \beta + n}$.

- “shrinkage” — the posterior mean shrinks the observed proportion of successes \hat{p} toward the prior mean $\frac{\alpha}{\alpha + \beta}$.
- The degree of shrinkage is controlled by
 - the size of the sum of the beta prior parameters relative to
 - the sample size $n =$ the sum of the observed number of successes plus the observed number of failures.