

22S:138
Bayesian Statistics

Inference for Proportions, continued

Lecture 6
Sept. 7, 2005

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

Prediction

- In many situations, interest focuses on predicting values of a future sample from the same population.
 - i.e. on estimating values of potentially observable but not yet observed quantities
- Example: we are considering interviewing another sample of 50 UI students in the hope of getting more evidence to present to the regents, and we would like to get an idea of how it is likely to turn out before we go to the trouble of doing so!
- So we are considering a new sample of sample size n^* , and want to estimate the probability of some particular number y^* of successes in this sample.

- If, based on the earlier survey, we actually knew the true value of the population proportion p , we'd just use the binomial probability:

$$p(y^*|p) = \binom{n^*}{y^*} p^{y^*} (1-p)^{n^*-y^*}, \quad y^* = 0, \dots, n^*$$

- But of course we still had uncertainty about p even after observing the original sample.
- All of our current knowledge about p is contained in the posterior distribution obtained using the original survey.

- The *posterior predictive probability* of some particular value of y^* in a future sample of size n^* is

$$p(y^* | y) = \int_0^1 p(y^* | p) p(p | y) dp, \quad y^* = 0, \dots, n$$

where y denotes the data from the original survey, and $p(p | y)$ is the posterior distribution based on that survey.

- For example, suppose we had used the Beta(10, 40) prior so our posterior distribution is $p(p|y) = \text{Beta}(17, 83)$. Then the posterior predictive probability of y^* successes in a future sample of size n^* is

$$p(y^* | y) = \int_0^1 p(y^* | p) \text{Beta}(p; 17, 83) dp, \quad y^* = 0, \dots$$

- This is particularly easy to compute if $n^* = 1$, in which case:

$$\begin{aligned} \text{Pr}(y^* = 1|y) &= \int_0^1 \text{Pr}(y^* = 1|p) p(p|y) dp \\ &= \int_0^1 p p(p|y) dp \\ &= E(p|y) \\ &= \frac{17}{17 + 83} \end{aligned}$$

Recognizing kernels

Normalizing constants revisited

- when trying to determine whether a function is the kernel of a standard density, consider the support
- if you do recognize a function as the kernel of a standard density, then you can easily figure out what it integrates to
- examples

$$\theta (1 - \theta)^w, \quad 0 < \theta < 1$$

Proper and Improper distributions

- a density is valid only if it integrates to one over the support of the random variable
- any “density” that integrates to a positive finite number can be “normalized” so that it integrates to one
- a density is improper if its integral is not finite
- example

$$p(\sigma) = \frac{1}{\sigma}, \quad 0 < \sigma < \infty$$

$$\theta^{w-1} \exp(-v\theta), \quad 0 < \theta < \infty$$

Noninformative or reference priors

- useful when we want inference to be unaffected by information apart from the current data
- in many scientific contexts, we would not bother to carry out an experiment unless we thought it was going to increase our knowledge significantly
 - i.e. we expect and want the likelihood to dominate the prior

The case of the binomial likelihood

- one choice of noninformative prior: $U(0, 1)$
- a disadvantage: it is not invariant under transformations
- suppose we were more interested in the logit transformation of the unknown proportion p

$$\phi = g(p) = \log\left(\frac{p}{1-p}\right)$$

than in p itself

- when we get to logistic regression later in the semester, this is exactly the quantity we will be interested in

- improper reference priors are sometimes used
 - if you do this, you must verify that the resulting posterior is proper
 - note that if the posterior is improper, it doesn't exist, so valid inference cannot be based on it
 - (in some multiple parameter models, it may be possible to make valid inference about a subset of parameters even if the posterior is improper)
- often more than one choice of reference prior for the same likelihood

- recall transformation of variables
 - if $y = g(x)$ is a one-to-one transformation of x
 - so $x = g^{-1}(y)$
 - $p_x(x)$ is the density function of x
 - we want density function of y , $p_y(y)$

$$p_y(y) = p_x(g(y)) \left| \frac{dx}{dy} \right|$$

- let's transform the uniform prior on the binomial parameter p into a prior on $\text{logit}(p)$

- uh-oh, it's not vague or uniform

Jeffreys' prior

First, recall the Fisher information

- used by frequentists in computing asymptotic variance of MLEs
- used by Bayesians in constructing one form of reference prior
- let $p(y|\theta)$ denote the probability density function of the data given the unknown parameter θ
- Fisher defined the information about a parameter provided by an experiment as

$$I(\theta|y) = -E \frac{\partial^2(\log(p(y|\theta)))}{\partial\theta^2}$$

- The expectation is taken over possible values of y for fixed θ . Since the information is an *expectation*, it depends on the *distribution* of y , not the observed value of y .

Jeffreys' prior

- If we transform the unknown parameter θ to $\phi = g(\theta)$ then

$$\frac{\partial \log(L(\phi|y))}{\partial \phi} = \frac{\partial \log(L(\theta|y))}{\partial \theta} \frac{\partial \theta}{\partial \phi}$$

- Squaring and taking expectations over values of y (note that $\frac{\partial \theta}{\partial \phi}$ does not depend on y), we get

$$I(\phi|y) = I(\theta|y) \left(\frac{\partial \theta}{\partial \phi} \right)^2$$

- So Jeffreys proposed the following reference prior

$$p(\theta) \propto \sqrt{I(\theta|y)}$$

- Since the log-likelihood $\log(L(\theta|y))$ differs from $\log(p(y|\theta))$ only by a constant, all their derivatives are equal. Thus the information can equivalently be defined as

$$I(\theta|y) = -E \frac{\partial^2 \log(L(\theta|y))}{\partial \theta^2}$$

- If there are n independent observations $\mathbf{y} = y_1, y_2, \dots, y_n$ then the probability densities multiply and the log-likelihoods add. Thus the Fisher information becomes

$$I(\theta|\mathbf{y}) = -E \frac{\partial^2(\log(L(\theta|\mathbf{y})))}{\partial \theta^2} = n I(\theta|y)$$

- Finally, it can be shown (e.g. Hogg and Craig, or Lee)

$$I(\theta|y) = E \left(\frac{\partial \log(L(\theta|y))}{\partial \theta} \right)^2$$

- advantages

- *invariance property*: no matter what scale we choose for measuring the unknown parameter, the same prior results when the parameter is transformed to any other scale
- depends on the form of the likelihood but not on the current observed data

- disadvantages

- sometimes information doesn't exist (e.g. in Cauchy distribution)
- more controversial in multiparameter setting

Jeffrey's prior for binomial likelihood

$$\log(L(p|y)) = y \log p + (n-y) \log(1-p) + \text{constant}$$

$$\frac{\partial^2 \log(L(p|y))}{\partial p^2} = -\frac{y}{p^2} - \frac{n-y}{(1-p)^2}$$

What is $E(y)$ if $y \sim \text{Binomial}(n, p)$?

So

$$I(p|y) = \frac{n}{p(1-p)}$$

Taking the square root and removing the constant n , gives

$$p(p) \propto p^{-\frac{1}{2}} (1-p)^{-\frac{1}{2}}$$

Do we recognize this density?

One more candidate noninformative prior for the binomial likelihood

- Uniform(0,1)
- Beta($\frac{1}{2}, \frac{1}{2}$)
- Beta(0, 0)
 - improper
 - will give a proper posterior *unless*
 - * either $y = 0$ or $y = n$ in current data
 - attractive feature: yields the mle $\frac{y}{n}$ as the posterior mean