Solution to Exam 1: STAT 638, Fall 2015

1. A researcher randomly selects 100 white mice from a large population of such mice, and tests each one for a fairly rare genetic mutation that makes the mice unsuitable for a certain experiment. Let θ be the proportion of mice in the population with the genetic mutation. It turned out that 4 mice in the sample had the genetic mutation. As a prior for θ the researcher decides to use a beta$(1/2, 10)$ density.

(a) (8) What feature of the researcher's prior makes it seem appropriate in this situation?

The prior is concentrated near 0, and thus favors small proportions.

(b) (8) The beta$(1/2, 10)$ prior has an amount of information equivalent to that in a sample of how many mice?

$a + b = 1/2 + 10 = 10.5$

(c) (8) What is the mean of the researcher's prior?

mean $= a/(a + b) = 0.5/10.5 = 0.048$

(d) (8) Identify the researcher's posterior distribution.

$$p(\theta|\text{data}) \propto \theta^{-1/2}(1-\theta)^9 \cdot \theta^4(1-\theta)^{96} = \theta^{4.5-1}(1-\theta)^{106-1}$$

Therefore, the posterior is beta$(4.5, 106)$**.**

(e) (8) What are the mean and mode of the posterior distribution?

mean $= 4.5/110.5 = 0.041$

mode $= (4.5 - 1)/[(4.5 - 1) + (106 - 1)] = 0.032$

(f) (8) Describe how you would find a 95% credible interval for the proportion of all mice with the genetic mutation. **Note:** The interval need not be an HPD region.

Using an equal tail area approach, I would use R or some other software to find the 2.5th and 97.5th percentiles of the beta$(4.5, 106)$ distribution. Call these percentiles q_1 and q_2, respectively. Then the interval (q_1, q_2) is a 95% credible interval.

2. Suppose that $Y_1, \ldots, Y_n$ are independent and identically distributed observations from a Poisson distribution with mean θ.

(a) (8) Show that the Jeffreys prior in this case is proportional to $\theta^{-1/2} I_{(0,\infty)}(\theta)$.

$$p(\theta|\boldsymbol{y}) = e^{-n\theta}\theta^{n\bar{y}}/(y_1! \cdots y_n!) \quad \Longrightarrow \quad \log p(\theta|\boldsymbol{y}) = -n\theta + n\bar{y}\log\theta - \sum_{i=1}^{n} \log y_i! \quad \Longrightarrow$$

$$\frac{\partial \log p(\theta|\boldsymbol{y})}{\partial\theta} = -n + \frac{n\bar{y}}{\theta} \quad \Longrightarrow \quad \frac{\partial^2 \log p(\theta|\boldsymbol{y})}{\partial\theta^2} = -\frac{n\bar{y}}{\theta^2} \quad \Longrightarrow$$

$$-E\left[\frac{\partial^2 \log p(\theta|\boldsymbol{Y})}{\partial\theta^2}\right] = E\left(\frac{n\bar{Y}}{\theta^2}\right) = \frac{n\theta}{\theta^2} = n\theta$$

Therefore, the Jeffreys prior is proportional to $1/\sqrt{\theta}$.

(b) (7) Is the prior proper? Why?

No, because the integral of $1/\sqrt{\theta}$ is $2\sqrt{\theta}$, which tends to infinity as θ tends to ∞.

(c) (7) Identify the posterior corresponding to use of the Jeffreys prior. Is the posterior proper, and why?

$$p(\theta|\boldsymbol{y}) \propto \theta^{-1/2}e^{-n\theta}\theta^{n\bar{y}} = \theta^{n\bar{y}+1/2-1}e^{-n\theta}$$

So, the posterior is gamma$(n\bar{y} + 1/2, n)$, which is proper since $\bar{y}$ is always nonnegative.

3. (6) Lindley's paradox refers to

(a) a situation where a frequentist confidence interval for a parameter is drastically different than a Bayesian HPD interval.

(b) the fact that Bayes factors often overstate the signifcance of evidence against point null hypotheses.

(c) the fact that frequentist P-values often overstate the significance of evidence against point null hypotheses.

(d) situations where using noninformative priors is inappropriate.

(e) Dennis Lindley's son and daughter, both of whom have PhDs.

4. (6) It is of interest to test the hypotheses

$$H_0 : \theta = 0 \quad \text{vs.} \quad H_1 : \theta = 1$$

using the data $\boldsymbol{y}$. The two relevant values of the likelihood are $p(\boldsymbol{y}|0) = 1/4$ and $p(\boldsymbol{y}|1) = 1/10$, and the prior probabilities of H_0 and H_1 are 2/3 and 1/3, respectively. The Bayes factor and the posterior probability of H_0 are, respectively,

(a) 5 and 5/6.

(b) 5/2 and 5/6.

(c) 5 and 5/7.

(d) 5/2 and 5/7.

(e) π and e^{-1}.

5. (6) An experimenter observes data whose distribution depends on an unknown parameter θ. The posterior distribution for θ is normal with mean 10 and standard deviation 1. The experimenter wants to predict a value for Y, which is independent of the data given θ. The distribution of Y given θ is normal with mean θ and standard deviation 5. Which of the following is a valid way to generate a single value from the posterior predictive distribution of Y?

(a) Generate Y from $N(10, 1)$.

(b) Generate Y from a gamma distribution with mean 10 and standard deviation 5.

(c) Generate Y from $N(10, 25)$.

(d) Let θ be a value drawn from the prior distribution. Then Y is drawn from $N(\theta, 25)$.

(e) Let θ be a value drawn from $N(10, 1)$. Then Y is drawn from $N(\theta, 25)$.

6. (6) A good way to check whether a random sample of data fits a model $p(y|\theta)$ is to compare the

(a) empirical distribution of the data with the posterior predictive distribution.

(b) empirical distribution of the data with the posterior distribution.

(c) likelihood function with the posterior.

(d) likelihood function with the prior.

(e) prices of beef and pork.

7. (6) The data $Y_1, \ldots, Y_n$ have been observed but Y_{n+1} has not been observed. In the absence of any other information, a valid expression for the posterior predictive distribution, $m(y_{n+1}|y_1, \ldots, y_n)$, is

(a) $\int_\Theta p(y_{n+1}|\theta)p(\theta|y_1, \ldots, y_n)\, d\theta$.

(b) $\int_\Theta p(y_{n+1}|\theta)p(\theta)\, d\theta$.

(c) $\int_\Theta p(y_1, \ldots, y_n, y_{n+1}|\theta)p(\theta)\, d\theta / m(y_1, \ldots, y_n)$.

(d) $\int_\Theta p(y_1, \ldots, y_n, y_{n+1}|\theta)p(\theta|y_1, \ldots, y_n)\, d\theta$.

(e) salud!