

Assigned 19 Oct 2009; due 27 Oct 2009. *You may study the questions/answers with others, but what you hand in for credit must be your own work.*

1. Let  $\theta_A$  and  $\theta_B$  be the average number of children of men in their 30s with and without bachelor's degrees, respectively.
  - (a) Using a Poisson sampling model, a  $G(2, 1)$  prior for each  $\theta$  and the data linked from the course web page, obtain 5,000 samples of  $\tilde{Y}_A$  and  $\tilde{Y}_B$  from the posterior predictive distribution of the two populations. Plot the Monte Carlo approximations to these two posterior predictive densities.
  - (b) Find the 95% quantile-based posterior credible intervals for  $\theta_A - \theta_B$  and  $\tilde{Y}_A - \tilde{Y}_B$ . Describe in words the differences between the two populations using these quantities and the plots in (a), along with any other results that may interest you.
  - (c) Obtain the empirical distribution (and/or mass) of the data in group  $B$ . Compare this to the Poisson model with mean  $\theta = 1.4$ . Do you think the Poisson model is a good fit? Why or why not?
  - (d) For each of the 5,000  $\theta_B$ -values you sampled, sample  $n_B = 218$  Poisson random variables and count the number of 0s and 1s in each of the 5,000 simulated datasets. You should now have two sequences of length 5,000 each, one sequence counting the number of people having zero children for each of the 5,000 datasets, the other counting the number of people with one child. Plot the two sequences against one another (one on the  $x$ -axis, one on the  $y$ -axis). Add points to the plot point marking how many people in the observed dataset had zero children and one child, respectively. Using this plot, describe the adequacy of the Poisson model.
2. Starting from independent uniform random variables ( $U \sim U(0, 1)$ ), give an algorithm to simulate independent draws from a Logistic distribution, having density

$$f(x) = \frac{e^{-x}}{(1 + e^{-x})^2} \quad \text{for } x \in \mathbb{R}.$$

Write a function in R to generate samples from the Logistic based on your algorithm, and use some graphical summaries to double-check that they have the correct distribution. Finally, use the samples to estimate  $P(X \in (2, 3))$ .

3. Consider a random variable  $X$  having the density

$$f(x) = e^{-(x+1)} + (e - 1)e^{-ex}, \quad x \in (0, \infty).$$

- (a) Design a rejection sampling algorithm for obtaining independent draws of  $X$  based only upon samples  $U \sim U(0, 1)$ . Comment on the efficiency of your algorithm. Implement the algorithm in R and use it double-check your work (as in Question 2).
  - (b) Describe a method which is more efficient than your rejection sampling algorithm, but still uses only samples  $U \sim U(0, 1)$ , and justify why it is more efficient.
  - (c) Design an importance sampling (IS) algorithm for calculating the  $P(X \in (2, 3))$ . Compare (empirically) the variance of IS estimator to the plug-in estimators for the same quantity that may be based upon the samplers you created in parts (a–b).
4. A data file linked on the course web page contains data on the amount of time students from three high schools spent on studying or homework during an exam period. Analyze data from each of these schools separately, using the normal model with a conjugate prior distribution in which  $\{\mu_0 = 5, \sigma_0^2 = 4, \kappa_0 = 1, \nu_0 = 2\}$ , and compute or approximate the following:
- (a) posterior means and 95% credible intervals for the mean  $\mu$  and standard deviation  $\sigma$  from each school.
  - (b) the posterior probability that  $\mu_i < \mu_j < \mu_k$  for all six permutations  $\{i, j, k\}$  of  $\{1, 2, 3\}$ .
  - (c) the posterior probability that  $\tilde{Y}_i < \tilde{Y}_j < \tilde{Y}_k$  for all six permutations  $\{i, j, k\}$  of  $\{1, 2, 3\}$ , where  $\tilde{Y}_i$  is a sample from the posterior predictive distribution of school  $i$ .
  - (d) Compute the posterior probability that  $\mu_1$  is bigger than both  $\mu_2$  and  $\mu_3$ , and the posterior probability that  $\tilde{Y}_1$  is bigger than both  $\tilde{Y}_2$  and  $\tilde{Y}_3$ .