Stat 591 – Homework 02 **Due: Monday 09/30**

Your group should submit a write-up that includes solutions the problems stated below, along with any <u>*relevant*</u> *pictures/graphs or computer code/output.*

1. (From Gelman et al., 2004, Chap. 2.) Let $X \mid \theta$ have an exponential distribution with rate θ, i.e., $f_\theta(x) = \theta e^{-\theta x}$, $x > 0$. Consider a $\mathsf{Gamma}(a, b)$ prior for θ, with PDF of the form $\pi(\theta) \propto \theta^{a-1} e^{-b\theta}$.
   (a) Suppose that we observe $X \geq 100$, but the exact value of X remains hidden; this is a right-censored observation. Find the posterior distribution of θ, given, $X \geq 100$, and write down the posterior mean and variance.
   (b) Now suppose we learn the exact value, $X = 100$. Find the posterior distribution of θ, as well as the corresponding posterior mean and variance.
   (c) Are you surprised that the posterior variance in (b), based on exact data, is larger than that in part (a), based on censored data? Why or why not?

2. (a) Problem 2.3a in [GDS], page 59.
   (b) Problem 2.21 in [GDS], page 63.

3. *Students' choice:* pick one of the two problems below.
   — Problem 2.13bc in [GDS], page 61. [Hint: For Part b(ii), find the conditional PMF for $(X_1, \ldots, X_k)$ given n and $X_1 + \cdots + X_k$; this *will not* depend on p. Then treat this conditional PMF as a likelihood function for n, given data. It is interesting that conditioning sometimes has a marginalization effect.]
   — Problem 2.14 in [GDS], page 61.

4. (Based, in part, on Problem 2.19 in [GDS], page 62.) Data $X = (X_1, X_2, X_3, X_4)$ is assumed to have, for a given probability vector $\theta = (\theta_1, \theta_2, \theta_3, \theta_4)$, a four-dimensional multinomial distribution, $\mathsf{Mult}_4(n, \theta)$, with PMF

$$f_\theta(x) = \frac{n!}{x_1! x_2! x_3! x_4!} \theta_1^{x_1} \theta_2^{x_2} \theta_3^{x_3} \theta_4^{x_4}.$$

   Note the constraints: $\sum_i X_i \equiv n$ and $\sum_i \theta_i \equiv 1$. (There's nothing special about four dimensions; this is just to be consistent with the example in part (c) below.)
   (a) As a prior for θ, consider a Dirichlet distribution,[1] denoted by $\mathsf{Dir}_4(a)$, with PDF[2]

$$\pi(\theta) = \frac{\Gamma(a_1 + \cdots + a_4)}{\Gamma(a_1) \cdots \Gamma(a_4)} \theta_1^{a_1 - 1} \theta_2^{a_2 - 1} \theta_3^{a_3 - 1} \theta_4^{a_4 - 1},$$

   where $a = (a_1, a_2, a_3, a_4)$ is a vector of positive numbers. Show that the Dirichlet prior is conjugate for the multinomial likelihood. That is, show that the posterior distribution for θ, given X, is also of the Dirichlet form, and identify the new parameter, say, a'.

[1] There is lots of information about the Dirichlet distribution on the web, e.g., `http://en.wikipedia.org/wiki/Dirichlet_distribution`

[2] This is a PDF on the *probability simplex*, the set of all probability vectors, which, in this case, is a three-dimensional subset of the full four-dimensional real space.

(b) Explain how you can simulate from a Dirichlet distribution using a gamma random number generator, such as `rgamma` in R. [Hint: Look at the `wikipedia` page for the Dirichlet distribution.] Based on this, explain how you can simulate from the posterior distribution of $g(\theta)$, where g is some real-valued function defined on the θ-space.

(c) For the agreement application in Homework 01, using the same data, draw a sample of size 3000 from the posterior distribution of κ. Then:

i. Draw a histogram of this sample to visualize the posterior distribution of κ. How does this picture look compared to that of the sampling distribution of $\hat{\kappa}$ based on the asymptotic normality and bootstrap approximations from Homework 01?

ii. Find an equi-tailed 90% credible interval for κ.

For this, you'll need to select a value for the hyper-parameter $a = (a_1, a_2, a_3, a_4)$, and you can pick anything you like, e.g., all 1's. You may also want to duplicate your calculations here for several different values of a to see if there's any difference in the results.