1. There are approximately 540 coronavirus testing locations in South Carolina. At the beginning of the day, officials at each location record

$$Y = \text{number of specimens tested to find the first positive case}$$

and assume Y follows a geometric distribution with probability θ. On a given day, suppose $Y_1, Y_2, ..., Y_{540}$ are observed and are modeled as iid geometric(θ) random variables. However, instead of treating θ as fixed as we have done previously, suppose θ is best regarded as a random variable itself which has a beta(α, β) prior distribution.
(a) Using the five-step algorithm in the notes, derive the posterior distribution of θ. Show each step. Is the beta(α, β) distribution a conjugate prior for θ? Explain.
(b) When you were carrying out the steps in part (a), when did you know what the posterior distribution was going to be? Explain.
(c) SC-DHEC officials believe the probability θ is small; based on population sizes and case counts from previous weeks, they believe θ is likely between 0.01 and 0.05. Give an example of a beta(α, β) prior distribution that would do a good job incorporating this *a priori* belief. Then given an example of a beta(α, β) prior that would do a bad job.
(d) Suppose that tomorrow the observed data $y_1, y_2, ..., y_{540}$ produce the sum

$$\sum_{i=1}^{540} y_i = 17295.$$

Determine the posterior distribution of θ using your good and bad prior choices in part (c). Graph both posterior distributions. If they are similar (they probably are), explain why this is true.

2. Suppose $Y_1, Y_2, ..., Y_n$ is an iid sample from

$$f_Y(y|\theta) = \begin{cases} \theta y^{\theta-1}, & 0 < y < 1 \\ 0, & \text{otherwise,} \end{cases}$$

a beta distribution with parameters $\alpha = \theta > 0$ and $\beta = 1$. In turn, suppose θ is best regarded as random with prior distribution $\theta \sim$ gamma(a, b), where a and b are known. Show the posterior distribution $g(\theta|\mathbf{y})$ is also gamma with parameters

$$\begin{aligned} a^* &= n + a \\ b^* &= \left(\frac{1}{b} - \sum_{i=1}^{n} \ln y_i \right)^{-1}. \end{aligned}$$

Does the posterior distribution depend on a sufficient statistic? Explain.

3. If S^2 is the sample variance based on an iid sample of size n from a $\mathcal{N}(\mu, \sigma^2)$ distribution, we know

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1).$$

(a) Find the distribution of S^2. Denote the pdf of S^2 by $f_{S^2|\sigma^2}(s^2|\sigma^2)$.
(b) A conjugate prior for σ^2 is the inverted gamma distribution, $\text{IG}(\alpha, \beta)$, whose pdf is given by

$$g(\sigma^2) = \begin{cases} \dfrac{1}{\Gamma(\alpha)\beta^\alpha}\dfrac{1}{(\sigma^2)^{\alpha+1}}e^{-1/\beta\sigma^2}, & \sigma^2 > 0 \\ 0, & \text{otherwise} \end{cases}$$

where $\alpha > 0$ and $\beta > 0$ are known. Show that the posterior distribution $g(\sigma^2|s^2)$ is $\text{IG}(\alpha^*, \beta^*)$, where

$$\begin{aligned} \alpha^* &= \alpha + \frac{n-1}{2} \\ \beta^* &= \left[\frac{(n-1)s^2}{2} + \frac{1}{\beta}\right]^{-1}. \end{aligned}$$

Hint: The joint distribution of S^2 and σ^2 satisfies

$$f_{S^2,\sigma^2}(s^2, \sigma^2) = f_{S^2|\sigma^2}(s^2|\sigma^2)g(\sigma^2)$$

and $g(\sigma^2|s^2)$ is proportional to $f_{S^2,\sigma^2}(s^2, \sigma^2)$.

4. SEIR models are used by epidemiologists to describe covid-19 disease severity in a population. The model consists of four different categories:

$$\begin{aligned} \text{S} &= \text{susceptible category} \\ \text{E} &= \text{exposed category} \\ \text{I} &= \text{infected category} \\ \text{R} &= \text{recovered category.} \end{aligned}$$

The four categories are mutually exclusive and exhaustive among living individuals (SEIRD models do include a fifth category for those who have died from disease). A random sample of n individuals is selected from a population (e.g., residents of Richland County) and the category status of each individual is identified. This produces the multinomial random vector

$$\mathbf{Y} \sim \text{mult}\left(n, \mathbf{p}; \sum_{j=1}^{4} p_j = 1\right),$$

where

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \end{pmatrix} \quad \text{and} \quad \mathbf{p} = \begin{pmatrix} p_1 \\ p_2 \\ p_3 \\ p_4 \end{pmatrix}.$$

The random variables Y_1, Y_2, Y_3, Y_4 record the number of individuals identified in the susceptible, exposed, infected, and recovered categories, respectively. Review the multinomial distribution from Chapter 5 (STAT 511) in case you have forgotten it.

Recall from Example 11.1 (notes) the beta distribution is a conjugate prior for the binomial. Just as the multinomial distribution can be regarded as a generalization of the binomial (to more than two categories), we need a prior distribution for $\mathbf{p}$ that is a generalization of the beta. This generalization is the Dirichlet$(\alpha_1, \alpha_2, \alpha_3, \alpha_4)$ distribution. Specifically, suppose $\mathbf{p}$ is best regarded as random with prior pdf

$$g(\mathbf{p}) = \begin{cases} \dfrac{\Gamma(\alpha_1+\alpha_2+\alpha_3+\alpha_4)}{\Gamma(\alpha_1)\Gamma(\alpha_2)\Gamma(\alpha_3)\Gamma(\alpha_4)}\, p_1^{\alpha_1-1}p_2^{\alpha_2-1}p_3^{\alpha_3-1}p_4^{\alpha_4-1}, & 0 < p_j < 1,\ \sum_{j=1}^4 p_j = 1 \\ 0, & \text{otherwise,} \end{cases}$$

where $\alpha_1 > 0$, $\alpha_2 > 0$, $\alpha_3 > 0$, and $\alpha_4 > 0$ are known.

(a) If $\mathbf{Y}|\mathbf{p} \sim \text{mult}(n, \mathbf{p}; \sum_{j=1}^4 p_j = 1)$ and $\mathbf{p} \sim g(\mathbf{p})$, show the posterior distribution $g(\mathbf{p}|\mathbf{y})$ is Dirichlet with parameters $\alpha_j^* = y_j + \alpha_j$, for $j = 1, 2, 3, 4$. *Hint:* The joint distribution of $\mathbf{Y}$ and $\mathbf{p}$ satisfies

$$f_{\mathbf{Y},\mathbf{p}}(\mathbf{y}, \mathbf{p}) = f_{\mathbf{Y}|\mathbf{p}}(\mathbf{y}|\mathbf{p})g(\mathbf{p})$$

and $g(\mathbf{p}|\mathbf{y})$ is proportional to $f_{\mathbf{Y},\mathbf{p}}(\mathbf{y}, \mathbf{p})$.

(b) A special case of the Dirichlet distribution above arises when $\alpha_1 = \alpha_2 = \alpha_3 = \alpha_4$, the so-called "symmetric Dirichlet distribution." This distribution would arise when one has no prior information to favor the count in one SEIR category over the other three. Do you think $g(\mathbf{p})$ with $\alpha_1 = \alpha_2 = \alpha_3 = \alpha_4$ would be a reasonable prior model for covid-19 in Richland County? Explain.