# Problem Set 5: Written Part

*Your Name Goes Here*

## Details

### Due Date

This assignment is due at 4:00 PM on Friday, March 8.

### How to Write Up

The written part of this assignment can be either typeset using latex or hand written.

### Grading

5% of your grade on this assignment is for turning in something legible. This means it should be organized, and any Rmd files should knit to pdf without issue.

An additional 20% of your grade is for completion. A quick pass will be made to ensure that you've made a reasonable attempt at all problems.

Across both the written part and the R part, in the range of 1 to 3 problems will be graded more carefully for correctness. In grading these problems, an emphasis will be placed on full explanations of your thought process. You don't need to write more than a few sentences for any given problem, but you should write complete sentences! Understanding and explaining the reasons behind what you are doing is at least as important as solving the problems correctly.

Solutions to all problems will be provided.

### Collaboration

You are allowed to work with others on this assignment, but you must complete and submit your own write up. You should not copy large blocks of code or written text from another student.

### Sources

You may refer to our text, Wikipedia, and other online sources. All sources you refer to must be cited in the space I have provided at the end of this problem set.

## Problem I: Spatial Organization of Chromosome (Adapted from Rice Problem 8.45)

This is a continuation of a problem from PS3. There, we found the maximum likelihood estimator and estimate; here, we will take a Bayesian perspective.

A human chromosome is a very large molecule, about 2 or 3 centimeters long, containing 100 million base pairs (Mbp). The cell nucleus, where the chromosome is contained, is in contrast only about a thousandth of a centimeter in diameter. The chromosome is packed in a series of coils, called chromatin, in association with special proteins (histones), forming a string of microscopic beads. It is a mixture of DNA and protein. In the G0/G1 phase of the cell cycle, between mitosis and the onset of DNA replication, the mitotic chromosomes diffuse into the interphase nucleus. At this stage, a number of important processes related to chromosome function take place. For example, DNA is made accessible for transcription and is duplicated, and repairs are made of DNA strand breaks. By the time of the next mitosis, the chromosomes have been duplicated. The complexity of these and other processes raises many questions about the large-scale spatial organization of chromosomes and how this organization relates to cell function. Fundamentally, it is puzzling how these processes can unfold in such a spatially restricted environment.

In the 1990's, a series of experiments (Sachs et al., 1995; Yokota et al., 1995) were conducted to learn more about the spatial organization of chromosomes. Pairs of small DNA sequences (size about 40 kbp) at specified locations on human chromosome 4 were flourescently labeled in a large number of cells. The distances between the members of these pairs were then determined by flourescence microscopy. (The distances measured were actually two-dimensional distances between the projections of the paired locations onto a plane.) The empirical distribution of these distances provides information about the nature of large-scale organization.

There has long been a tradition in chemistry of modeling the configurations of polymers by the theory of random walks. As a consequence of such a model, the two-dimensional distance should follow a Rayleigh distribution. If $X \sim \text{Rayleigh}(\theta)$ (with parameter $\theta > 0$), then the probability density function is given by

$$f(x|\theta) = \frac{x}{\theta} \exp\left(\frac{-x^2}{2\theta}\right)$$

for positive values of $x$ (and the probability density function is 0 for non-positive values of $x$). Note that I have made a slight modification to the standard form of the probability density function here. This modification will simplify the algebra you have to do for this problem, and does not change anything fundamental about the distribution.

In this problem, you will fit the Rayleigh distribution to some of the experimental results. You will find the general form of the posterior distribution in a Bayesian analysis in the written part, and fit it to the actual data set in the R part. The entire data set comprises 36 experiments in which the separation between the pairs of flourescently tagged locations ranged from 10 Mbp to 192 Mbp. In each such experimental condition, about 100–200 measurements of two-dimensional distances were determined. This exercise will be concerned just with the data from one of these experiments.

Suppose that we adopt the model $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \text{Rayleigh}(\theta)$.

**(1) Show that the Inverse-Gamma$(\alpha, \beta)$ distribution is a conjugate prior for the Rayleigh distribution, and find expressions for the parameters of the posterior distribution in terms of $\alpha$, $\beta$, $n$, and $x_1, \ldots, x_n$. You do not need to give numeric values for the posterior distribution parameters.**

Suppose we conduct a Bayesian analysis where the prior distribution for the unknown parameter of the Rayleigh$(\theta)$ distribution is $\theta \sim \text{Inverse-Gamma}(\alpha, \beta)$. With this prior distribution, the prior probability density function for $\theta$ is given by

$f_\Theta(\theta|\alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \theta^{-\alpha-1} \exp\left(-\frac{\beta}{\theta}\right)$

It is required that $\theta > 0$, $\alpha > 0$, and $\beta > 0$.

**(2) Suppose you don't know much about chromosomes, and so you don't have very strong prior beliefs about the value of $\theta$. Which of the prior distributions for $\theta$ displayed in the plot below best expresses your lack of prior knowledge? Explain why in a sentence or two.**

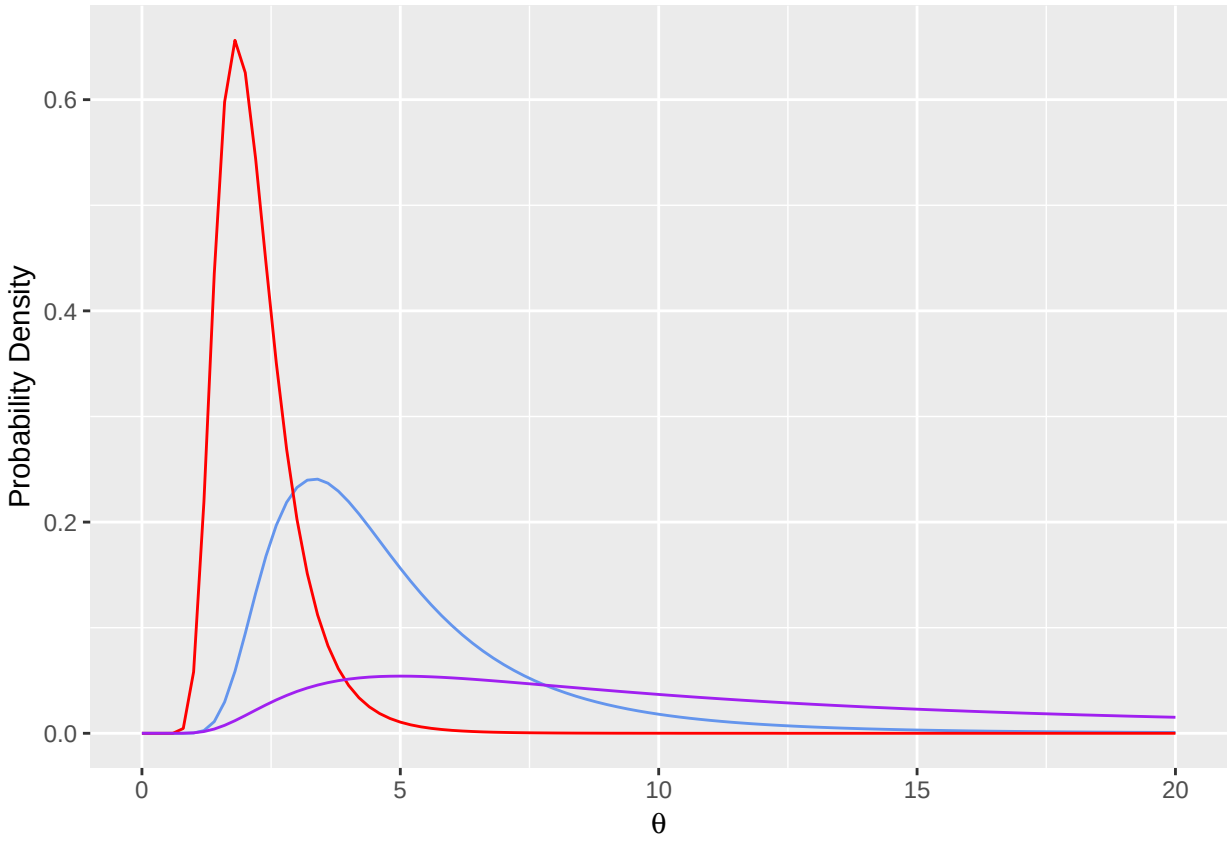
```r
library(ggplot2)

dinvgamma <- function(theta, alpha, beta) {
  log_value <- rep(-Inf, length(theta))
  inds <- which(theta > 0)

  log_value[inds] <- alpha * log(beta) - lgamma(alpha) - (alpha + 1) * log(theta[inds]) - beta/(theta[inds])

  return(exp(log_value))
}

ggplot(mapping = aes(x = c(0, 20))) +
  stat_function(fun = dinvgamma, args = list(alpha = 5, beta = 20), color = "cornflowerblue") +
  stat_function(fun = dinvgamma, args = list(alpha = 10, beta = 20), color = "red") +
  stat_function(fun = dinvgamma, args = list(alpha = 1, beta = 10), color = "purple") +
  xlab(expression(theta)) +
  ylab("Probability Density")
```




## Problem II: Multinomial Distribution

Below is some information about the multinomial distribution from the probability distributions handout. This is a distribution for a random vector $\mathbf{X} = (X_1, X_2, \ldots, X_k)$.

Suppose $\mathbf{X} \sim \text{Multinomial}(n, \theta)$, where $\theta = (\theta_1, \theta_2, \ldots, \theta_k)$.

$\mathbf{X} = (X_1, X_2, \ldots, X_k)$ is a vector of counts for how many observations fell into each of $k$ categories in a sample of $n$ independent trials where the item sampled in each trial falls into category $j$ with probability $\theta_j$. More concretely, imagine rolling a weighted die with $k$ sides $n$ times, and on each roll, side $j$ comes up with probability $\theta_j$. The vector $X$ records how many times each face of the die came up. (Note that since exactly one side of the die must come up on each roll, we must have $\sum_{j=1}^{k} \theta_j = 1$.)

The probability mass function of $X$ is

$f_{\mathbf{X}}(x|\theta) = \frac{n!}{x_1!x_2!\cdots x_k!}\theta_1^{x_1}\theta_2^{x_2}\cdots\theta_k^{x_k}$

Suppose that we observe $\mathbf{x} = (x_1, x_2, \ldots, x_k)$, and we want to estimate the vector of probabilities $\theta$. In the Bayesian setting, the most commonly used prior distribution for the vector $\mathbf{\Theta}$ is a Dirichlet distribution.

The Dirichlet distribution is a distribution on a vector of probabilities that sum to 1, such as the probabilities for each side of the die coming up in the multinomial distribution. Its parameters are a vector $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_k)$.

If $\mathbf{\Theta} = (\Theta_1, \ldots, \Theta_k)$ are jointly distributed as Dirichlet($\alpha$), then the joint pdf of $\mathbf{\Theta}$ is given by:

$f_{\mathbf{\Theta}}(\theta) = \frac{1}{B(\alpha)}\prod_{j=1}^{k}\theta_j^{\alpha_j}$

Here, $B(\alpha)$ is a complicated function of $\alpha$ that's just there to ensure that the joint pdf of $\mathbf{\Theta}$ integrates to 1.

### (1) Show that the Dirichlet is a conjugate prior for the multinomial model.

To simplify this a little, assume that you have a single observation $\mathbf{x} = (x_1, x_2, \ldots, x_k)$ from the multinomial model.