Problem Set 4: Written Part

Your Name Goes Here

Details

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 15% 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.

Problem I: Cheating

This problem is adapted from an exercise in "Introduction to Statistical Thought" by Michael Lavine (2013). Lavine writes:

> Some researchers are interested in θ, the proportion of students who ever cheat on exams. They randomly sample 100 students and ask "Have you ever cheated on a college exam?" Naturally, some students lie. Let ϕ_1 be the proportion of non-cheaters who lie and ϕ_2 be the proportion of cheaters who lie. Let X be the number of students who answer "Yes".

Define the following events:

- A is the event that a randomly sampled student actually has cheated on a college exam.
- B is the event that a randomly sampled student says they have cheated on a college exam. (They answer "Yes".)

(1) Based on the problem statement above, write the following probabilities in terms of the unknown parameters θ, ϕ_1, and ϕ_2.

- $P(A)$: the probability that a randomly sampled student has cheated on a college exam
- $P(B|A^c)$: the probability that a randomly sampled student says they have cheated on a college exam, given that they actually have not cheated

- $P(B^c|A^c)$: the probability that a randomly sampled student says they have not cheated on a college exam, given that they actually have not cheated
- $P(B|A)$: the probability that a randomly sampled student says they have cheated on a college exam, given that they actually have cheated
- $P(B^c|A)$: the probability that a randomly sampled student says they have not cheated on a college exam, given that they actually have cheated

(2) Find P(B), the probability that a randomly sampled student says they have cheated on a college exam.

(3) Specify a reasonable probability model for X, the number of the 100 students in the survey who answer "Yes". You may assume that the survey responses are independent; for example, the respondents are not friends and are not planning their survey responses together. Specify all parameters for your model in terms of θ, ϕ_1, and ϕ_2.

(4) Write down a formula for the pdf of $X|\Theta, \Phi_1, \Phi_2$ based on your model in part (3). Note that you will conduct the survey of 100 students once and record $X = x$. (You observe a single x, not $x_1, \ldots, x_{100}$.)

(5) Suppose you have specified independent Beta prior distributions for the unknown parameters θ, ϕ_1, and ϕ_2: $\Theta \sim$ Beta(α_0, β_0), $\Phi_1 \sim$ Beta(α_1, β_1), and $\Phi_2 \sim$ Beta(α_2, β_2). In this notation, α_0, β_0, α_1, β_1, α_2, and β_2 are numbers the analyst picks to specify the priors for each of the three parameters θ, ϕ_1, and ϕ_2. Write down formulas for the pdfs of these three prior distributions. You will be writing down three separate pdfs involving θ, ϕ_1, ϕ_2, α_0, β_0, α_1, β_1, α_2, and β_2.

(6) Suppose you now observe $X = x$. Write down a formula for the pdf of the joint posterior of Θ, Φ_1, and Φ_2 given $X = x$, up to a multiplicative constant of proportionality.

(7) Your real goal is to use the survey data to learn about θ, the proportion of students who have cheated. ϕ_1 and ϕ_2 are not of direct interest, they were just necessary to get a reasonable model for the data. (Parameters like this are sometimed referred to as "nuisance parameters".)

a. Suppose you want to calculate the posterior mean for the proportion of students who have cheated. Write this as a suitable integral of the joint posterior pdf $f_{\Theta,\Phi_1,\Phi_2|X}(\theta, \phi_1, \phi_2|x)$.

Hint: How can you find $f_{\Theta|X}(\theta|x)$ from $f_{\Theta,\Phi_1,\Phi_2|X}(\theta, \phi_1, \phi_2|x)$?

b. Suppose you have a sample $(\theta_1, \phi_{1_1}, \phi_{2_1}), \ldots, (\theta_m, \phi_{1_m}, \phi_{2_m})$ from the joint posterior distribution of Θ, Φ_1, and Φ_2 given X. Write how you could approximate the integral from part a using Monte Carlo integration.

c. Suppose you want to calculate the posterior probability that less than half of students have cheated. Write this as a suitable integral of the joint posterior pdf $f_{\Theta,\Phi_1,\Phi_2|X}(\theta, \phi_1, \phi_2|x)$. Set it up so you have an indicator function as part of the integrand (on the inside of the integral).

d. Write how you could approximate the integral from part c using Monte Carlo integration.