Exam 2

Mori Jamshidian

Spring 2024

Contents

Exam Rules and Instructions

(1) You must work on this exam individually. Any communication with others about this exam in any form is considered cheating. Should you have any questions, please send an email to mori@fullerton.edu.

(2) Copying solutions from an Internet source or any other source would be considered plagiarism and will be dealt with according to university policy.

(3) You are not allowed to distribute the questions on this exam in any form or share the questions with anyone during the exam or anytime after the due-date.

(4) Your solutions must be typewritten. You must submit a single pdf file that would include your solutions, R codes, and R outputs. Additionally submit your Rmarkdown file.

(5) Your solutions must appear in the order of the problem numbers. If you don't know the answer to a problem, write the problem number, and leave a blank space.

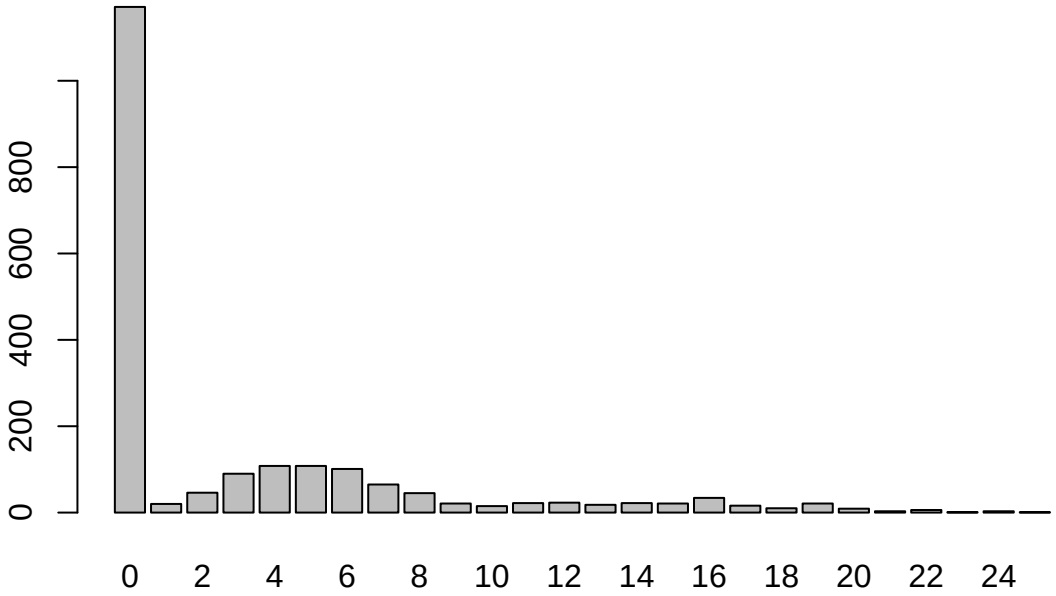
(6) I reserve the right to interview you about the exam after I grade your exam.

(7) Submit your solution to Canvas.

(8) A total of 55 points is possible.

Problem 1

We wanted to investigate marijuana use among college students. We surveyed 2000 randomly selected college students and asked them the following question: "In your best estimation, how many marijuana joints, if any, have you smoked the past thirty days?" We obtained 2000 values ranging from 0 to 27. A barplot of the data is given below.



Let x_i denote the number of joints smoked by the i-th individual surveyed. As shown in the barplot, many students said that they have not smoked at all ($x_i = 0$), and because of this large number of zeros and the fact that the proportions do not monotonically decrease, a Poisson model would not be appropriate. To model the data, we assumed there are three groups: a proportion α of students (group 1) who, for whatever reason, report that they have not smoked marijuana even if this is not true (note that some might not be comfortable revealing that they smoke marijuana). A proportion β of students (group 2) who smoke marijuana occasionally and respond truthfully. For group 2, we assume the number of joints that they smoke follows a Poisson distribution with mean μ. Finally, a proportion $1 - \alpha - \beta$ of students (group 3) who smoke marijuana more often and respond truthfully. For group 3, we assume the number of joints that they smoke follows a Poisson distribution with mean λ. Thus, we assume that each observation x_i comes from the following mixture distribution:

$$f(x_i) = \alpha \, 1_{\{x_i=0\}} + \beta \, \frac{e^{-\mu}\mu^{x_i}}{x_i!} + (1 - \alpha - \beta) \, \frac{e^{-\lambda}\lambda^{x_i}}{x_i!}$$

where $1_{\{x_i=0\}} = 1$ if $x_i = 0$, and it is zero otherwise.

Your job in this problem is to use the EM algorithm to estimate the parameters $\theta = (\alpha, \beta, \mu, \lambda)$ using the data provided in the dataset `Problem3_Data.csv`.

(a) [3 Points]

Write the (observed data) log-likelihood function.

(b) [3 Points]

Describe the complete data, and derive the the complete data log-likelihood.

(c) [3 points]

Obtain the $Q(\theta', \theta)$ function, including the formulas for the required expectations.

(d) [4 points]

Write formulas that maximize $Q(\theta', \theta)$ with respect to θ'. That is, write the update formulas for α, β, μ, and λ.

(e) [4 points]

Give the formulas for the elements of the gradient of the log-likelihood function.

(f) [9 points]

Write an R function to implement the EM algorithm, and a function to compute the gradient of the log-likelihood. Your EM function should input α, β, μ, λ, the data `x`, and a parameter `maxiter` for the maximum number of iteration. To get full credit, you must use the following instructions:

- Write a functions to compute the gradient of the log-likelihood.
- Write a function to compute the required expectations.
- Write a function that implements the EM algorithm. This function should call the gradient and the expectation functions. At each iteration print the iteration number, α, β, $1-\alpha-\beta$, μ, λ, and the norm of the gradient, in that order; use `norm(name_of_gradient, "2)`.
- Show all the functions that you write.
- Use the following parameter values to run your EM function: $\alpha = 1/3$, $\beta = 1/3$, $\mu = 1$, $\lambda = 30$, and `maxiter = 30`.
- Print all iterations.

(g) [3 Points]

Using your parameter estimates, explain your findings. Your explanation should involve interpretation of all of the parameters in the context of the problem.

(h) [2 points]

We met a student who said that she smoked 9 joints in the past 30 days. What is the probability that this student belons to each of the three groups, 1, 2, or 3? Which group would you classify this student to? Heuristic explanations don't get any points!

(i) [5 points]

Write an R function to numerically approximate the observed information matrix, and use it to obtain the standard errors of the parameter estimates.

Problem 2

Consider the function

$$g(x) = 30x^2(1-x)^2, \quad 0 < x < 1.$$

(a) [4 points]

Write an R function that uses the basic Monte Carlo method to compute the integral

$$\theta = \int_a^b g(x)dx.$$

for arbitrary values a and b that range in the interval [0,1], and provides a $100(1-\alpha)\%$ confidence interval for θ. Specifically,

- the input to your program must be a, b, number of simulated values, a seed for the random number generation, and a confidence level CL with default value of 0.95. Your program must check that the input variables $0 \leq a < b \leq 1$, and the confidence level CL is between 0 and 1. If not, the program must stop and print an error message.
- your program must generate values from uniform(a,b) for the Monte Carlo estimation.
- your program must output and print the Monte Carlo estimate of θ, its estimate of standard error, and a $100(1-\alpha)\%$ confidence interval for θ.
- After writing the program, use it to obtain the value of the integral for $a = 0.2$ and $b = 0.6$. Use 10,000 simulated values and seed 338. Your output should include the Monte Carlo estimate of θ, its estimate of standard error, and a 95% confidence interval for θ.

(b) [3 points]

Repeat part (a), but instead of generating values from uniform(a,b), suppose that you generate values from the Beta distribution with parameters $\alpha = 2$ and $\beta = 2$. Explain how you can use the beta distribution to estimate the integral. In your explanation specify what expectation you will be estimating and how you will estimate it. No R code is needed here.

(c) [2 points]

Write a formula for the variance of the estimate of θ in part (b), as a function of $Y \sim Beta(\alpha = 2, \beta = 2)$ and explain how you would estimate this variance in you R program. Do not write R code here. Only give an explanation of what R function you would use and how.

(d) [4 points]

Write an R code to implement the methods that you described in parts (b) and (c). Your program should be written for general a, b, and CL values again, and output and print the Monte Carlo estimate of θ, its estimate of standard error, and a 95% confidence interval for θ. Run your program using the same values for $a = 0.2$, $b = 0.6$, 10,000 simulated values, seed 338, and confidence level 95%. you must use the `rbeta` function to generate values from the beta distribution.

Problem 3

Consider the function

$$h(x) = \frac{x^3}{1+x^2}.$$

Let X be a random variable with density function

$$f(x) = \frac{e^{-x}}{1-e^{-1}} \quad 0 \leq x \leq 1.$$

We want to estimate $E(h(X)$, the expected value of $h(X)$.

(a) [3 points]

Write an R function that uses the basic Monte Carlo method to approximate the required integral. Specifically, generate 100,000 values from the random variable X, and use these values to estimate $E(h(X))$. Your function should output the Monte Carlo estimate of $E(h(X))$ and its estimate of standard error. Use seed 666.

(b) [3 points]

Here you repeat part (a), but use the method of Antithetic Sampling to estimate $E(h(X)] and its standard error. Write a step by step algorithm (pseudo code) for solving this problem. No R code here. [Hint: you need to generate 50,000 samples from Uniform(0,1).]

(c) [4 points]

Write an R function to implement the algorithm. Your function should output the estimate of $E(h(X))$ and its estimate of standard error. Use seed 666.

(d) [3 points]

Explain the relationship between the standard error of the estimate of $E(h(X))$ obtained using the method of Antithetic Sampling as compared to the standard error of the estimate of $E(h(X))$ obtained using the basic Monte Carlo method. Show this relationship by computing relevant quantities using your R code.