1. Let Y_{ij} be mutually independent random variables, $i = 1, \ldots, A$ and $j = 1, \ldots, n$, where $Y_{ij} \sim \text{beta}(\theta_i, 1)$. Recall on Day 1 you found that the asymptotic size α LRT of the hypothesis that $\theta_i = i\tau$ was

$$\phi(\boldsymbol{Y}) = \begin{cases} 1 & \text{if } TS(\boldsymbol{Y}) > \chi^2_{A-1,1-\alpha} \\ 0 & \text{otherwise} \end{cases}$$

where

$$TS(\boldsymbol{Y}) = -2n \left[\sum_{i=1}^{A} \ln(-\bar{\ell}_i) + \sum_{i=1}^{A} \ln(i\hat{\tau}) \right]$$
$$\bar{\ell}_i = \frac{\sum_{j=1}^{n} \ln(Y_{ij})}{n}$$
$$\hat{\tau} = \frac{-A}{\sum_{i=1}^{A} i\bar{\ell}_i}.$$

   (a) Using $A = 4$, $n = 3$, $\tau = 1$, and $\alpha = 0.05$ carry out a simulation study with 1000 replications to estimate the size of the test. Include your program and output.

   (b) What is the estimated size of the test? Does this test perform well? Explain how you arrived at your conclusion.

   (c) The simulation study used 1,000 replications. Is this a reasonable number of replications to estimate the size of a test? Explain why or why not.

2. A recent study shows that many of the pre-existing conditions that increase the mortality risk of COVID-19 are connected with long-term exposure to air pollution. In this problem, data on long term PM2.5 exposure (μg/m^3) and COVID-19 mortality counts are provided. The response is the cumulative COVID-19 mortality counts through May 12, 2020 for n=3109 US counties. County-level exposure to PM2.5 are calculated by averaging results from an established exposure prediction model for the years 2000–2016. In addition to PM2.5 exposure, 20 potential confounding variables are provided for the analysis.

   The data files covid.Rdata (R dataset), covid.csv (CSV text), and covid.txt (tab delimited) all contain the following variables.

`mean_pm25`:	county-level exposure to PM2.5
`Deaths`:	cumulative COVID-19 mortality counts
`Population`:	population

And some other variables that can be helpful:

`date_since:`	days since first COVID-19 case reported (a proxy for epidemic stage)
`q_popdensity:`	population density,
`older_percent:`	percent of population 65 years of age,
`prime_percent:`	percent of the population 45-64 years of age,
`mid_percent:`	percent of the population 15-44 years of age,
`poverty:`	percent of the population living in poverty,
`medhouseholdincome:`	median household income,
`pct_blk:`	percent of Black residents,
`hispanic:`	percent of Hispanic residents,
`education:`	percent of the adult population with less than high school education,
`medianhousevalue:`	median house value,
`pct_owner_occ:`	percent of owner-occupied housing,
`obese:`	percent of the population with obesity
`smoke:`	percent of current smokers,
`beds:`	number of hospital beds per unit population
`mean_summer_temp:`	average daily temperature for summer (June to September)
`mean_winter_temp:`	average daily temperature for winter (December to February)
`mean_summer_rm:`	relative humidity for summer (June to September)
`mean_winter_rm:`	relative humidity for winter (December to February)
`date_since_social:`	days since issuance of stay-at-home order for each state

The researcher is interested in analyzing the connection between COVID deaths and long term PM2.5 exposure. You are a statistical consultant giving advice to the researcher.

You need to analyze the data and write a report that includes the following (note that your report must be written to a scientist who is only familiar with basic statistics):

- Description and the objectives of the study
- Fit two appropriate models for the count data to answer the researcher's question. For each analysis,
  (a) Write out the model. Define each term and state model assumptions.
  (b) Explain why the model is appropriate.
  (c) Report the point and interval estimates of the parameter(s) that address the research question.
- The steps of the analysis and the findings. Make sure that you interpret the necessary results. Include plots and anything that is helpful to understand the analysis and results.
- Compare the analysis results from the two models you used. Explain which model is preferred and why.
- Draw a final conclusion to help the researcher understand the results of the statistical analysis.
- You also need to provide the SAR/R code in the Appendix.

Please note that the report has a strict three-page limit.