MATH 538: Bayesian Statistics
Quiz #2 (Take-Home)
Due Date: Tuesday November 26, 2024 by 11:59pm

Important Instructions: Please read the instructions very carefully.

- There are a total of 2 questions on this Quiz. You must type ALL answers to receive full marks.
- You must do the exam individually, and cannot discuss the questions and/or the course material during the exam period with any other individual except the instructor. You can use your notes and/or the textbook, but you are not allowed to use the internet or any online source related to the course material. Copying solutions from the Internet is considered forgery and cheating.
- You should hand in your solutions (including the R/STAN codes) in one pdf file. You should also hand in your R codes in one separate .txt or .R file (or .stan). Make clear breaks/comments in your R codes to separate questions and/or different sections of a question. Both files should be uploaded **onto Canvas under Quiz #2**. Name the two files as follows:
  Solutions: YourFirstNameYourLastName-Quiz2-Solutions.pdf
  Example: ValeriePoynor-Quiz2-Solutions.pdf
  R code: YourFirstNameYourLastName-Quiz2-Rcode
  Example: ValeriePoynor-Quiz2-Rcode.txt OR ValeriePoynor-Quiz2-Rcode.R
- Upload your files by **11:59pm on Tuesday, November 26$^{\textbf{nd}}$**. The two files should be uploaded on Canvas.
- The instructor may schedule an interview with you to discuss your exam where you may be asked to elaborate on your solutions.
- **The deadline is firm, and late penalties will be applied if the submission is not received by the due date.**

1 Data Analysis

The data file *hearing.txt* is from an experiment to calibrate word lists used to measure the hearing ability of subjects. The four word lists had been designed so that they should be equally difficult to perceive, but were designed for normal-hearing subjects in an environment without background noise. The data in this experiment were collected in the presence of a noisy background. Each column is a word list, and each row is a subject. The entry is their score on that list (each subject was tested on all four lists).

1. **[50 marks]** We will consider a standard normal hierarchical model that assumes the scores (y) are independent and identically distributed within each list (i.e. assuming students are exchangeable within list). Note that this is equivalent to a one-way ANOVA model that incorporates a list random effect. We will assume conjugate priors. The full hierarchical model is given by:

$$\begin{aligned} y_{ij}|\theta_j, \sigma^2 &\sim N(\theta_j, \sigma^2) \\ \theta_j|\mu, \sigma^2 &\sim N(\mu, \sigma^2) \\ \mu &\sim N(30, 1) \\ \sigma^2 &\sim \Gamma^{-1}(2, 10) \end{aligned}$$

for $i = 1, ..., n_j$ with $n_j = 24$ for each j, and $j = 1, ..., J$ with $J = 4$.

(a) Perform an exploratory analysis on these data, describing your statistical summaries and plots. Discuss any interesting findings.

(b) Write out the joint posterior, $f(\boldsymbol{\theta}, \mu, \sigma^2|\boldsymbol{y})$.

(c) Derive the full posterior conditional distribution for θ_j. That is find the form of $f(\theta_j|\boldsymbol{\theta}_{-j}, \mu, \sigma^2, \boldsymbol{y})$ - note you may use any conjugate form derived in class without re-deriving it.

(d) Derive the full posterior conditional distribution for the hyperparameters: $f(\mu|\boldsymbol{\theta}, \sigma^2, \boldsymbol{y})$ and $f(\sigma^2|\boldsymbol{\theta}, \mu, \boldsymbol{y})$ - note you may use any conjugate form derived in class without re-deriving it.

(e) Fit the model with MCMC using RStudio (no Rstan). Show your trace plots and acf plots for μ, σ^2, and at least three θ_j's. Remove burn-in as appropriate. Be sure you obtain at least 2000 independent posterior samples.

(f) Obtain and report posterior point and interval estimates for the random student effects. Make a plot comparing the observed sample mean scores of each student to the estimated posterior means scores (i.e. posterior of θ_j's). Comment on what you see. Are the posterior estimates closer to the observed sample means or the overall observed mean score? Also discuss the variation across the list effects. What can you conclude about the lists? Do they appear to be the same level of difficulty for these students under the background conditions in which these data were collected?

(g) Provide posterior point and interval estimate as well as plots for μ and σ^2. Discuss your findings and interpretations in context of the scenario.

2. **[50 marks]** Consider the *contraceptive.csv* data that describe the number of married women sampled (N) from a particular developing country (each row). The women were asked whether or not they use modern contraceptiion techniques, and the total number who responded that they did is given by the variable Y. The variable W describe a proxy level education for each country.

(a) Perform an exploratory analysis on these data, describing your statistical summaries and plots. Discuss any interesting findings.

(b) Fit a Binomial Hierarchical Model with a non-informative hyper-prior to these data (excluding the level of education). You may use Rstudio or Rstan. Provide posterior plots, point and interval estimates, and discussions for all of your model parameters. Discuss specifically if the hierarchical model is justified.

(c) Now let's consider the level of education under the following model

$$\begin{aligned} y_j|\theta_j &\sim N(n_j, \theta_j) \\ logit(\theta_j) &= \alpha_j + \beta w_j \\ \alpha_j &\sim N(0, 100^2) \\ \beta &\sim N(0, 10^2) \end{aligned}$$

for $j = 1, ..., J$ where here, $J = 15$.

Fit this model in Rstudio or Rstan. Provide posterior plots, point and interval estimates, and discussions for your all of your model parameters. Discuss specifically if level of education is significantly associated with the proportion of women who use modern contraception. If so, in what way is it associated?