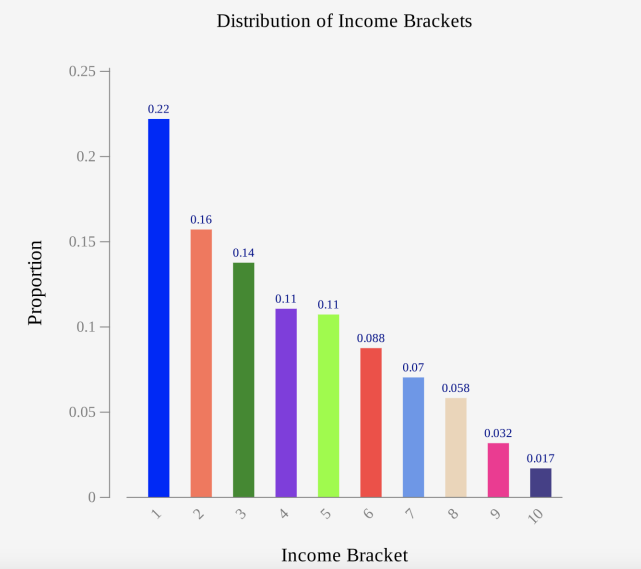
# MATH 538: Bayesian Statistics
# Midterm Exam (Take-Home)
# Due Date: Thursday November 7, 2024 by 5:00pm

**Important Instructions:** Please read the instructions very carefully.

- There are a total of 2 questions on this exam. 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 codes) in one pdf file. You should also hand in your R codes in one separate .Rmd, .txt or .R file. 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 Exam 1 Take Home assignment**. Name the two files as follows:
  **Solutions:** YourFirstNameYourLastName-Exam-Solutions.pdf
  Example: ValeriePoynor-Exam-Solutions.pdf
  **R code:** YourFirstNameYourLastName-Exam-Rcode
  Example: ValeriePoynor-Exam-Rcode.txt OR ValeriePoynor-Exam-Rcode.R
- Upload your files by **by 5:00pm, Thursday November** $7^{th}$. 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. **[30 marks]** [FINANCIAL APPLICATION] "The Pareto Distribution principle was first employed in Italy in the early 20th century to describe the distribution of wealth among the population. In 1906, Vilfredo Pareto introduced the concept of the Pareto Distribution when he observed that 20% of the pea pods were responsible for 80% of the peas planted in his garden. He related this phenomenon to the nature of wealth distribution in Italy, and he found that 80% of the country's wealth was owned by about 20% of its population." - `https://corporatefinanceinstitute.com/resources/economics/pareto-distribution/`. In this problem, we consider the top 20% highest incomes for US adults in 1996, income20train.csv. For your viewing pleasure here is a Pareto graph of the income distribution of all income brackets of the training data.1: > \$250000, 2: (\$250000, \$200000), 3: (\$200000, \$175000), 4: (\$175000,\$150000), 5: (\$150000, \$120000), 6: (\$120000 ,\$100000),7: (\$100000, \$70000), 8: (\$70000, \$40000), 9: (\$40000, \$30000), 10: (\$30000, \$0).




The model:

$$\begin{aligned} Y_i &\sim Pareto(\alpha, \beta) \quad \text{for i} = 1...\text{n} \\ p(\alpha, \beta) &\propto 1 \quad \text{(improper noninformative prior)} \end{aligned}$$

In class, you derived the following conditional posterior distributions:

$$\begin{aligned} \alpha|\beta, \boldsymbol{y} &\sim \Gamma(n+1, \sum_{1=1}^{n}(\log(y_i)) - n\log(\beta)) \\ \beta|\alpha, \boldsymbol{y} &\sim Mono(n\alpha + 1, min\{y_1, ...y_n\}) \end{aligned}$$

| Distribution | Density | Support |
|---|---|---|
| Pareto: Pareto($\alpha$, $\beta$) | $p(\theta\|\alpha,\beta) = \alpha\beta^{\alpha}\theta^{-(\alpha+1)}\boldsymbol{I}_{\{\theta>\beta\}}$ | $0 < \theta, \alpha, \beta$ |
| Monotonic: $Mono(\alpha, \beta)$ | $p(\theta\|\alpha,\beta) = \frac{\alpha}{\beta^{\alpha}}\theta^{\alpha-1}\boldsymbol{I}_{\{0<\theta<\beta\}}$ | $0 < \theta, \alpha, \beta$ |

(a) Perform an exploratory analysis of these data (plots ,summaries, etc.). Provide discuss in context of the problem.

(b) Explicitly write out the Gibbs sampling algorithm in your text (not your Rcode). Be specific about your index of iteration, as this is really the only thing missing from the conditional distributions provided.

(c) Apply your Gibbs sampler in R to obtain 2000 INDEPENDENT samples from the posterior distribution, $p(\alpha, \beta|\boldsymbol{y})$. Now, in order to sample from the Monotonic distribution, we will need to apply Inverse CDF sampling. I did not put this on the study guide, so I wrote a sampling function for you, see below. To sample from the Pareto distribution, use the *EnvStats* R package with function rpareto(n, location, shape). Provide trace plots, acf plots, and discuss burnin and thinning (if needed). Provide a bivariate scatterplot plot of your final parameter samples with contour lines added.

```
rmono <- function(n, alpha, beta){
  u = runif(n)
  x = exp( log(beta) + ((1/alpha)*log(u)))
            return(x)
}
```

(d) Using you posterior samples, provide plots for posterior marginal distributions of $p(\alpha|data)$ and $p(\beta|data)$ and summarize your findings with posterior point estimates and credible intervals. The $\beta$ parameter has a practical interpretation here, so be sure to interpret your posterior interval in context.

(e) Use you posterior samples of $\alpha$ and $\beta$ to obtain 2000 samples from your posterior predictive distribution of the top 20% incomes. Provide a density plot of your posterior predictive samples along with point and 95% interval estimates. Using the income20test.csv data, did your interval estimate capture about 95% of the incomes for these new income values? That is, does the Pareto distribution seem to b a fairly accurate model for the top 20% of incomes for US adults? Discuss your findings in context and provide any additional plot/summaries you feel important for your discussion.

‘

2. **[40 marks]**[SURVIVAL APPLICATION] Do rats eat themselves to grave earlier? The *rat.csv* dataset (Berger et. al. 1988) consists of survival times (in years) of rats from two experimental groups: the "Ad libitum group" (coded as 0) is comprised of 90 rats who were allowed to eat freely (as much and as often as desired), whereas the "Restricted group" (coded as 1) includes 106 rats that were placed on a restricted diet (restricted eating time and amount). Consider the following model for these data:

$$\begin{aligned} y_i &\sim Weibull(4, \sigma_i) \quad i = 1, ..., 196 \\ \sigma_i &= log(\beta_0 + \beta_1 x_i) \\ p(\beta_0, \beta_1) &\propto 1 \end{aligned}$$

where $y_i$'s are the survival times for each rat, $x_i$'s indicate the experimental group for each rat, and $\lambda_i$ is the rate parameter of the gamma distribution. The model can be interpreted as assuming the survival times of the two groups follow a Weibull distribution with same shape parameter, but with scale that is dependent on the group. Note $\sigma$ acts like the "lifetime" parameter in this scenario, such that the larger the value of $\sigma$, the longer the lifespan. Please refer to the stats R library (?dweibull) for the pdf form of the Weibull distribution.

(a) Perform an exploratory analysis of these data (plots, summaries, etc.). Provide discuss in context of the problem.

(b) Perform MCMC using a Metropolis-Hastings algorithm with a bivariate normal proposal distribution.

(c) Provide trace plots and autocorrelation plots of the posterior samples of $\beta_0$ and $\beta_1$. Discuss your results. Report your acceptance ratio, burnin, and thinning/ effective sample size. Tune your MCMC to maximize your acceptance while achieving good mixing (i.e. low autocorrelation). Report your tuning parameter (covariance matrix).

(d) Obtain marginal and joint posterior plots of 2000 independent posterior samples of $\beta_0$ and $\beta_1$ . Also report the MAP and 95% posterior credible estimates for these model parameters. Is there a significant effect of the experimental group on the survival times? Elaborate.

(e) In survival analysis, the hazard rate function is of special interest, as it provides the rate of risk of failure at particular time, given survival up to that

time. For continuous survival time, the hazard rate function is defined as the density function, $f(y|4,\sigma)$, divided by the survival function, $S(y|4,\sigma)$ (which is simply the probability of survival beyond a particular time point OR 1 minus the CDF). For our model, $S(y|4,\sigma) = \exp[-(y/\sigma)^4]$ and, thus, our hazard rate function is given by, $h(y|4,\sigma) = (4/\sigma)(y/\sigma)^3$. In the regression setting, researchers are often interested in the hazards ratio (HR) of two experimental groups: HR $= h(y|4,\sigma(x=1))/h(y|4,\sigma(x=0))$. What is the HR expression for our model? Use your posterior samples to obtain the MAP estimate an 95% posterior credible for the HR given these data. If the risk of failure for each of the groups are the same, then the HR would be 1. Use your 95% posterior credible for the HR to determine if one group has a significantly higher risk of death, and if so, state which one. Discuss the contextual research question: do rats tend to eat themselves to an earlier grave?

(f) [challenge] A more common way for the Weibull to be utilized in the regression setting is through the log-linear family of survival models:

$$\log(Y) \quad = \quad \beta_0 + \beta_1 x + \tau Z$$

where $\beta_0$ and $\beta_1$ are the coefficients of the regression, $\tau > 0$ is a scale, and $Z$ is the error distribution. Prove using a transformation of variables from $Z$ to $Y$ that if that $Z$ has a standard (minimum) extreme value (EV) distribution, $EV(0,1)$ (Note, the density form for $EV(W|\mu,\tau) = \frac{1}{\tau}\exp[(w-\mu)/\tau]\exp[-\exp[(w-\mu)/\tau]]$), then $Y$ follows a Weibull distribution. What are the expressions for the shape, $\alpha$, and scale, $\sigma$, of the Weibull distribution in terms of $\beta_0, \beta_1$, and $\tau$?