General Notes

- You will submit a minimum of two files, the core files must conform to the following naming conventions (including capitalization and underscores). 123456789 is a placeholder, please replace these nine digits with your nine-digit Bruin ID. The files you must submit are:

  1. *123456789_stats102c_hw5.Rmd*: Your markdown file which generates the output file of your submission.
  2. *123456789_stats102c_hw5.html/pdf*: Your output file, either a PDF or an HTML file depending on the output you choose to generate.
  3. *Included image files:* If you answer your questions with images files, you must upload them to this portal as well, or your Rmd file will not knit.
  4. Please place all of your Rmd (and image) file(s) into a single folder named 123456789_stats102c_hw5 and compress the folder into 123456789_stats102c_hw5.zip.
  5. You will submit two files; one html/pdf file (123456789_stats102c_hw5.html/pdf) and one compressed file (123456789_stats102c_hw5.zip).

  If you fail to submit any of the required core files you will receive ZERO points for the assignment. If you submit any files which do not conform to the specified naming convention, you will receive (at most) half credit for the assignment.

- **Your .Rmd file must knit**. If your .Rmd file does not knit you will receive (at most) half credit for the assignment.
  The two most common reason files fail to knit are because of workspace/directory structure issues and missing include files. To remedy the first, ensure all of the file paths in your document are relative paths pointing at the current working directory. To remedy the second, simply make sure you upload any and all files you source or include in your .Rmd file.

- Your coding should adhere to the tidyverse style guide: https://style.tidyverse.org/.

NOTE: *Everything* you need to do this assignment is here, in your class notes, or was covered in discussion or lecture.

- Please **DO NOT** look for solutions online.
- Please **DO NOT** collaborate with anyone inside (or outside) of this class.
- Please work **INDEPENDENTLY** on this assignment.
- **EVERYTHING** you submit **MUST** be 100% your, original, work. Any student suspected of plagiarizing, in whole or in part, any portion of this assignment, will be **immediately** referred to the Dean of Student's office without warning.

Problem 1: The weather on any given day on a tropical island could be rainy, sunny or cloudy, and the probability of tomorrow's weather only depends on today's weather and not any other previous days. Suppose that we obtained the transition probabilities as below.

P(Rainy tomorrow — Rainy today) = 0.6

P(Rainy tomorrow — Cloudy today) = 0.6

P(Rainy tomorrow — Sunny today) = 0.2

P(Cloudy tomorrow — Rain today) = 0.3

P(Sunny tomorrow — Cloudy today) = 0.1

P(Cloudy tomorrow — Sunny today) = 0.2

(a) Please find the transition matrix and draw a transition state diagram, with state space {1: Rainy, 2: Cloudy, 3: Sunny}.

(b) Suppose today is sunny. What is the expected weather two days from now?

(c) Is this Markov chain irreducible and aperiodic? Explain. Can you find a stationary distribution? If so, please show and explain how you find it.

Problem 2: Two urns A and B contain a total of N balls. Assume that at time t there were exactly k balls in A. At time $t+1$, an urn is selected at random in proportion to its contents (i.e., A is chosen with probability k/N and B is chosen with probability $(N-k)/N$). Then one of the N balls is randomly selected and placed in the chosen urn.

Let X_t denote the number of balls in urn A at time t, so $\{X_t, t \geq 0\}$ defines a Markov chain. Determine the transition matrix for this Markov chain.

Problem 3: Suppose the Markov chain is defined by $X^{(t+1)} = \alpha X^{(t)} + \varepsilon_t$, where $\varepsilon_t \sim N(0,1)$. Write R code to simulate this Markov chain with $X^{(0)} \sim N(0,1)$ for $t \leq 10^4$ and $\alpha = 0.7$. Check if your sample fits the distribution $N(mean = 0, sd = 1/\sqrt{(1-\alpha^2)})$.

Problem 4: We want to model the number of siblings people have in a certain population. We can model the number of siblings a person has as a Poisson random variable Y for some unknown mean parameter λ. The probability mass function of $Y \sim Pois(\lambda)$ is

$$f(y|\lambda) = \frac{e^{-\lambda}\lambda^y}{y!}, \quad \text{for } y = 0, 1, 2, \ldots.$$

Suppose, before observing any data, we model our prior beliefs about λ by a gamma distribution Gamma(α, β) with hyperparameters $\alpha, \beta > 0$, i.e,

$$\pi(\lambda) = \frac{\beta^\alpha}{\Gamma(\alpha)}\lambda^{\alpha-1}e^{-\beta\lambda}, \quad \text{for } \lambda > 0,$$

where $\Gamma(z) = \displaystyle\int_0^\infty x^{z-1}e^{-x}dx$. Note that

- The prior mean is $E(\lambda) = \dfrac{\alpha}{\beta}$.
- The prior variance is $Var(\lambda) = \dfrac{\alpha}{\beta^2}$.
- The prior mode is $\text{mode}(\lambda) = \begin{cases} \dfrac{\alpha - 1}{\beta} & \text{if } \alpha > 1, \\ 0 & \text{if } \alpha \leq 1. \end{cases}$

Suppose we observe data $y_1, y_2, \ldots, y_n \overset{iid}{\sim} Pois(\lambda)$. Let $\mathbf{y} = (y_1, y_2, \ldots, y_n)$.

(a) Show that the gamma distribution is a conjugate prior for the Poisson likelihood. More specifically, show that the posterior distribution $\pi(\lambda|y)$ is of the form

$$\pi(\lambda|y) \sim \text{Gamma}\left(\alpha + \sum_{i=1}^{n} y_i, \beta + n\right).$$

(b) Show that the posterior mean $E(\lambda|y)$ is a weighted average of the prior mean and the sample mean $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$. That is, find the weight w so that

$$E(\lambda|y) = w\frac{\alpha}{\beta} + (1 - w)\bar{y}.$$

(c) What is $\lim_{n\to\infty} E(\lambda|y)$? What does this limit represent?

(d) Use the `dgamma()` function in R to visualize the prior and posterior densities for hyperparameters $\alpha = 6, \beta = 2$ when the data is summarized by:

(i) $\sum_{i=1}^{n} y_i = 20$ and $n = 5$

(ii) $\sum_{i=1}^{n} y_i = 80$ and $n = 20$

Represent the sample mean on the plot to show how the posterior distribution is a compromise between the prior distribution and the data. Clearly indicate the different components of your plot.

Hint: For a Gamma(α, β) distribution, α is the `shape` parameter and β is the `rate` parameter.

(e) For scenarios (i) and (ii) in part (d), find and interpret 95% quantile-based credible intervals.

Problem 5: The Cauchy distribution is a continuous probability distribution. It is often used in statistics as the canonical example of a "pathological" distribution, and the density is defined as:

$$f(x|\gamma, \eta) = \frac{1}{\gamma\pi[1 + (\frac{x-\eta}{\gamma})]^2}, \ -\infty < x < \infty, \gamma > 0,$$

where η is the location parameter, specifying the location of the peak of the distribution.

(a) Please write an algorithm using the Metropolis-Hastings sampler to generate samples from the Cauchy distribution with $\gamma = 1$ and $\eta = 0$. You may use $N(\mu, \sigma)$ as the proposal distribution.

(b) Implement your algorithm in R to generate 10,000 samples and show the histogram.

(c) Compare the generated samples with `qcauchy` in R. Please conduct a exploratory analysis to determine if your samples agree with the output of `qcauchy`.

Problem 6: The Rayleigh distribution is a continuous probability density function, and is widely used for modeling the lifetime of an object depending on its age. The Rayleigh density is defined as:

$$f(x) = \frac{x}{\sigma^2} e^{x^2/(2\sigma^2)},\ x \geq 0$$

, where σ is the scale parameter of the distribution.

(a) Design a Metropolis algorithm to sample from the Rayleigh distribution with the χ^2 distribution as the proposal distribution. Implement this algorithm with σ =2, 4, and 6, respectively. Compare the performance of these three cases. Please give the acceptance rate, and plot the sample vs. the time index and the histograms.

(b) Repeat the above problem using the Gamma distribution as the proposal distribution with shape parameter as X_t and rate parameter as 1.