Due date: Friday May 14, by 5pm
You must submit your assignment electronically and as a single file via the LMS page for this subject. Your solutions must include your workings. **Note that this assignment is worth** 10% **of your overall mark for this subject.**

In submitting your work, you are consenting that it may be copied and transmitted by the University for the detection of plagiarism. Please start with the following statement of originality, which must be included near the top of your submitted assignment:

"This is my own work. I have not copied any of it from anyone else."

1. **Minnesota traffic volume.** For this question, you will be using this data set from the UCI Machine Learning Repository:

   `https://archive.ics.uci.edu/ml/datasets/Metro+Interstate+Traffic+Volume`

   From the URL, click on the "Data Folder" link and then download `Metro_Interstate_Traffic_Volume.csv.gz` and extract the csv file.[1]

   When importing the CSV file to R, make sure that the "Heading" option is set to "Yes".

   (a) Give a short written description of the contents of this data set. You should include information about how the data is organised, what the sample variables are, and how many samples there are.

   (b) Plot a histogram of the data in the `traffic_volume` column, using breaks = 30. Submit the R code and the histogram. Ensure your histogram has appropriate labels on each axis and a suitable title. You can set the title by changing the `main` parameter.

   (c) Use R to determine a 95% confidence interval for the average traffic volume. Submit the R code you used *and* the 95% confidence interval you find.

   Note: make use of included functions such as `sd` and `mean`; you may also find the `nrow` function useful.

2. **Daily chickenpox rates in Hungary.** For this question, you will be using this data set from the UCI Machine Learning Repository:

   `https://archive.ics.uci.edu/ml/datasets/Hungarian+Chickenpox+Cases`

   Download the data files, and locate the spreadsheet `hungary_chickenpox.csv` in the zip file.

   (a) Give a short written description of the contents of this data set.

   (b) Using R, determine an approximate 95% confidence interval for the average number of weekly chickenpox cases across all of Hungary.

   Note: you are doing this for the entire population of Hungary, so you will need to sum together the number of cases in each row. You are required to do this in R—do not modify the .csv file. The `rowSums` function will be useful. You will also need to exclude the first column from the sum; you are advised to store the result of the command `hungary_chickenpox[,2:21]` in another variable and sum the rows of that variable.

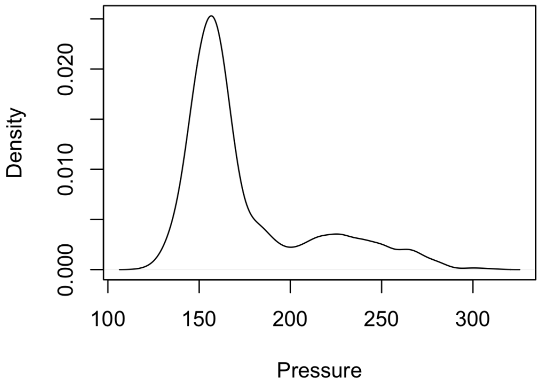
   (c) The county of Baranya comprises approximately 4% of Hungary's population, and the county of Somogy comprises approximately 3% of Hungary's population. Assuming that the number of cases of chickenpox is spread uniformly amongst the population of Hungary, use your answer to (b) to calculate the approximate average number of weekly cases in each of Baranya and Somogy.

   (d) Now use R and the data from the appropriate columns in the original data set to compute 95% confidence intervals for the weekly average number of chickenpox cases in each of Baranya and Somogy.

   (e) Compare your answers to (c) and (d). What, if anything, can you conclude about the assumption from part (c) that cases are spread uniformly amongst the population of Hungary?

[1] If you are unsure how to extract a .gz file, the program 7-Zip is recommended. Download here: `https://www.7-zip.org/`

3. **Mixture distributions**. Consider again the `pressure3.csv` file from the Week 7 Lab. Load the file into R, and view the histogram as we did in the lab. You may observe that there appear to be two "peaks" in the histogram; one around 150 and one around 225. This is made more evident by viewing the *kernel density estimate* of the probability density function by running the command `plot(density(pressure3$Pressure))`, which yields the following graph:



This suggests that the probability density function for the pressure data may be appropriately described by *two* (or more) underlying distributions. This is not an unusual idea. For example, consider human height: the average height of human males is higher than that of human females, and therefore the height of each subpopulation is distributed around a different mean. When the probability density function is defined by a combination of distributions, it is known as a *mixture distribution.*

(a) For a simple example of a mixture distribution, consider the following game: a player flips a coin; if the result is heads, roll a six-sided die, and if the result is tails, roll a 12-sided die. The player's score is the number shown on whichever die they roll. Let X denote the random variable corresponding to the player's score in this game.

Assume that heads and tails on the coin are equally likely, and each outcome on each dice is equally likely.

i. What is the sample space for X?

ii. Determine the probability mass function for X. Give your answer in table form.

iii. Calculate $E(X)$ and $\text{Var}(X)$.

(b) Now consider a mixture of two normal distributions. Let X_1 and X_2 be normally distributed random variables given by $X_1 \sim N(\mu_1, \sigma_1)$ and $X_2 \sim N(\mu_2, \sigma_2)$. A third random variable, Y, is determined as follows:

- with probability p, the value of Y follows the distribution of X_1;
- otherwise, the value of Y follows the distribution of X_2.

The probability density function for Y is then given by:

$$f_Y(x) = pf_1(x) + (1-p)f_2(x),$$

where $f_1(x)$ and $f_2(x)$ are, respectively, the probability distribution functions for X_1 and X_2.

To gain some intuition for this, consider the discussion of height from earlier. In that case, p may represent the probability of a person being male, at which point their height is determined by X_1; otherwise, they are female with probability $1-p$ and their height is determined by X_2.

i. Write an R function `dmixed` that implements the probability density function for R. You will need to use the `dnorm` function, and your function will require arguments x, p_1, μ_1, μ_2, σ_1 and σ_2.

ii. Using the fact that

$$\int_{-\infty}^{\infty} af(x) + bg(x)\,dx = a\int_{-\infty}^{\infty} f(x)\,dx + b\int_{-\infty}^{\infty} g(x)\,dx,$$

show that $E(Y) = p_1\mu_1 + p_2\mu_2$.

iii. Using the histogram of the pressure data from `pressure3.csv`, and overlaying a plot of the density function from (a) on this histogram, use trial and error to find values of μ_1, μ_2, σ_1, σ_2 and p that closely resemble the data from pressure3.csv.[2] Include the plot and the values you chose in your submission.

[2] Note: exactness is not required here; a suitable approximation that resembles the histogram by eye will be awarded full marks. But you should try to make the mean as calculated by (ii) as close as possible to the result of `mean(pressure3$Pressure)`. If you are interested in a more formal, algorithmic approach, refer to Chapter 8.4 of Matloff, N. S. (2020). *Probability and statistics for data science: Math + R + Data*, available on the La Trobe library website.