Homework Set #1 – Due 11:59pm, Tuesday, January 21.
MATH 445 Winter 2025

Note:

- This homework assignment is based on the Chapter 1 material.
- You need to submit two files (a PDF file containing your answers and an R script file containing your R code).
- Your answers must be presented in the same order as the problems given in this assignment, including all graphs and `R` outputs, if applicable (i.e., the graphs and outputs must be in the same order and must not be relegated to the back of your assignment).
- **Do not** present your answer in the R script file only. Your answers must be presented in the PDF file.
- Use your judgment to include the minimal amount of `R` code in your answers as necessary.
- All the `R` code must be well documented and submitted as a separate R script file (a file with ".R" extension).

The grading scheme is as follows:

- This homework assignment will be graded out of 30 points.
- Three out of the four problems will be selected for grading. Each of the three problems will be graded out of 10 points.
- For the problem that will not be graded, two points will be automatically deducted for each missing part.
- If the R code is missing or incomplete, at most 5 points will be deducted.

1. Let us assume that we have a sample of 400 observations, whose i-th observation is equal to i. In other words, our sample is

$$\{1, 2, 3, 4, \ldots, 398, 399, 400\}$$

(a) Suppose that 100 observations are randomly chosen without replacement. How many distinct samples (i.e., combinations) are possible? Show your answer in the "$\binom{n}{r}$" form.

(b) Suppose that 100 observations are randomly chosen with replacement. How many distinct samples (i.e., similar to combinations but with replacement) are possible? Show your answer in the "$\binom{n}{r}$" form.

(c) What is the approximate distribution of the sample means of all the distinct samples you considered in (b)? State the name of the distribution and its actual parameter values based on the given sample. Although not required, it could be helpful to use some simple `R` functions to estimate these parameter values.

2. Let

$$A = \begin{bmatrix} -0.1 & 1.5 & -0.6 & -1.4 \\ -1.2 & -0.5 & 1.0 & -1.5 \\ 0.1 & 0.5 & -0.5 & -2.0 \end{bmatrix}$$

be a 3-by-4 matrix. Using `R`, report the following.

(a) A' (transpose of A).

(b) $B = AA'$ and $C = A'A$.

(c) B^{-1} and C^{-1} (inverse matrices). For each of B^{-1} and C^{-1}, compare the answers using `solve()` and `ginv()` from the `MASS` package.

(d) If any one of them failed in (c), provide a reason why it failed.

(e) Calculate row medians and column standard deviations of A using the `apply()` function.

3. Missing observations are common in datasets. The dataset `sleep` (originally taken from the package named `VIM`, but is saved as "sleep.txt" in the żip file) is one such example. Examine the `sleep` dataset in sleep.txt whose headers are BodyWgt, BrainWgt, ... ,Danger.

(a) Store the information in the `NonD` column as a vector named `y`. Then, report `length(y)` and `sum(!is.na(y))`, and explain why they are different.

(b) Construct a new vector (say, `w`) that only stores non-missing observations (i.e., `w` is a vector without `NA`s). Report `R` code that shows how to do it (i) with `na.omit()` and (ii) without `na.omit()`.

(c) Create a data frame called `sleep17` which only stores information about the first seven columns of `sleep` (no need to report), and report their column means by ignoring `NA`s.

(d) The `boxplot()` function automatically produces a box plot for the data frame. Create a data frame called `sleep35` which only stores the third, fourth, and fifth column of `sleep` (no need to report), and report the box plot. Make sure to set the title of the box plot as "Box Plot of the Sleep Data" using the `main` argument.

(e) Using the `tapply()` function, calculate the mean sleep time (using the `Sleep` column) for each of the five danger levels (in the `Danger` column) by ignoring `NA`s.

4. The `sumdice()` function in the course notes calculates the sum of the n six-sided fair dice. Let X be a random variable for the sum of $n = 100$ six-sided fair dice. To simulate the distribution of X, one may use the `for` loop to calculate the sum many times (say, 10000 times). However, often (but not always) the `for`-loop in `R` is slow. Thus, let us think about an alternative way of simulating the distribution of X using the `rowSums()` function.
Note: Although you must provide all the `R` code (including (a), (b), and (c)) as a separate file, you only have to report the output for parts (d), (e), and (f) in your answers.

(a) Using the `sample()` function, produce a vector called `v` that stores $10000 \times 100 = 1000000$ observations by sampling observations at random from the set $\{1, 2, 3, 4, 5, 6\}$ with replacement (no need to report in your answer).

(b) Store `v` in a matrix called `vmat` that has 10000 rows and 100 columns (no need to report in your answer).

(c) Using the `rowSums()` function, calculate the sum of the dice rolls for each row (no need to report in your answer).

(d) Using (c), report the estimated mean and variance of X.

(e) Using (c), produce a histogram of X by utilizing the `hist()` function. Set `freq=FALSE` and `breaks="Scott"` in the argument. The label on the x-axis must be "sum", and make sure to set the title of the histogram as "Histogram of X".

(f) State the name and theoretical parameter values of the distribution that X approximately follows, and justify your answer. Then, compare the theoretical parameter values to the ones you reported in (d) and comment.