HUDM 6026 - Computational Statistics
HW 01 – Introduction

Instructions.

- You may use whatever software you like to write up your answers to the hw. For example, you might choose to use Rmarkdown for assignments that are primarily code-based; whereas, you might choose to use Word or LaTeX for more writing-heavy assignments. It's up to you.
- No matter which program you use, you should turn in a pdf or html version.
- Consult the syllabus for the rules about collaboration.

HW 01

1. A random variable X has a chi-square distribution with k degrees of freedom (df) if it can be expressed as the sum of k independent squared random normal variables. In that case, we write that $X \sim \chi^2_k$. Write a function that has two arguments, n for sample size and k for degrees of freedom. The purpose of the function is to generate a random sample of size n from a χ^2_k distribution. Importantly, you may not use the `rchisq` suite of functions. Instead, use `rnorm()` and the definition given above.

2. Use your function from above to generate a very large sample (e.g., $n = 1e7$) from χ^2_3. Display the distribution of the sample by creating a histogram in base or ggplot2. Play with the number of breaks until you settle on a number that looks good. Then, use `rchisq()` to generate a large sample of the same size and also create a histogram. If your function is working correctly, the distributions should be very nearly identical.

3. The chi-square distribution with 1 df is skewed heavily to the right. Generate 10000 sample means from iid samples of χ^2_1 of size $n = 5$ and plot the distribution of the sample means using a histogram or other univariate density plotting method. Then do the same for $n = 10$, $n = 20$, and $n = 40$. Superimpose a normal curve with mean equal to sample mean of the means and variance equal to sample variance of the means for each plot. Discuss whether this progression demonstrates the LLNs or the CLT and why.

4. Write a function to generate data (in a data frame or tibble) from the following regression model.

   $$\text{Life expectancy}_i = 71.0 - 0.28\text{Murder rate}_i + 0.05\text{HS grad rate}_i - 0.007\text{Frost days} + \epsilon_i$$

   where $\epsilon_i \overset{iid}{\sim} N(0, 0.75^2)$ and

   $$cov(\mathbf{X}) = \begin{bmatrix} 13.6 & -14.5 & -103.4 \\ -14.5 & 65.2 & 154.0 \\ -103.4 & 154.0 & 2702.0 \end{bmatrix}$$

   where predictors in the covariance matrix are in the same order as in the regression model. Once the function is created, set your seed and use it to generate a data set of size $n = 1000$. Then use `lm()` to estimate parameters and report your output in a table. Do not copy/past R output. Instead, format a nice looking table.