Final Exam

Answer the following questions in a single R script called `final.R`. Answers must be given by R commands. You cannot simply look at the data set and answer the question via direct inspection. Use comments (#) to indicate which portion of your code answers which question. Be sure that you obtain the correct solutions to each question when you execute your script one line at a time from top to bottom.

Each question will be graded according to the following criteria:

- 0%: No attempts is made to answer the question.
- 25%: An attempt is made that, although unsuccessful, revealed some understanding of what the question was asking.
- 50%: Solution is incorrect, but with some modifications, could be corrected.
- 75%: Solution is incorrect, but easily resolved with minor modifications **OR** solution is correct, but obtained via convoluted reasoning or by avoiding standard approaches.
- 100%: Solution is correct and uses standard approaches.

The exam is out of 38 points. This exam will use the data set in `final_college_majors.csv` which contains employment data by various majors circa 2012. Here is a description of some of the columns in this data set:

- `Major` gives the name of a major.
- `Major_category` gives a category name. Majors are grouped into categories.
- `P25th`, `Median` and `P75th` contain the 25th, 50th, and 75th percentile for salaries in a major.
- `Unemployment_rate` gives the unemployment rate for a major.

#1) (4 points) Give a scatter plot where each point is a major, the corresponding median income is the x-coordinate of the point and the corresponding unemployment rate is the y-coordinate of the point.

#2) (4 points) Restrict to the `ACTUARIAL SCIENCE` major. Create a bar chart with a bar for each of the income percentiles (25th, 50th and 75th percentiles). The height of each bar should be the corresponding income amount. (Hint: `pivot_longer`)

#3) (4 points) Compute the average of the median incomes for each major category. Display the results in a bar chart with a bar for each of the major categories. The height of each bar should be the corresponding average median income. This plot is easier to display if you use `coord_flip()`.

#4) (4 points) Compute the average of the 25th percentile, average of the median (50th percentile), and average of the 75th percentile incomes for each major category. Display the results in a dodge bar chart with 3 bars for each of the major categories. The heights of the 3 bar should correspond to the averages of these percentiles for a major category. This plot is easier to display if you use `coord_flip()`.

#5) The science faculty hold a Christmas party each year. During this party, we play the following game. Each faculty member is assigned a Christmas carol to whistle. We record how long the faculty member is able to whistle the song without laughing. The remaining faculty member attempt to make the whistler laugh. Eli Knapp claims he can whistle "Frosty the Snowman" on average for 51 seconds without laughing with a standard deviation of 9 seconds. Let X denote the time it takes Eli to laugh when whistling "Frosty the Snowman". Suppose X has a normal distribution.

(a) (2 points) Let's believe Eli's claims. Find the probability Eli whistles for more than 60 seconds.

(b) (2 points) Suppose that in this year's party, Eli whistles for only 30 seconds. What is the probability of him performing this poorly or worse if his claims are true?

(c) (2 points) Assuming Eli's claims are true, find the 85th percentile for Eli's whistling time.

#6) Recall that $X \sim \mathcal{U}_{[0,1]}$ indicates X has a uniform distribution on the interval $[0, 1]$. Consider independent $X_1, X_2, \ldots, X_n \sim \mathcal{U}_{[0,1]}$. Let $Y = \sum_{j=1}^{n} X_j$. We take $n = 10$ in this problem.

(a) (2 points) Randomly sample Y 5000 times. Store the results in a tibble. (Hint: `runif(10)` generates 10 samples for $X \sim \mathcal{U}_{[0,1]}$.)

(b) (2 points) Display the 5000 samples of Y in a density histogram.

(c) (2 points) Recall that if $X \sim \mathcal{U}_{[0,1]}$ then $\mathrm{E}(X) = \frac{1}{2}$ and $\mathrm{Var}(X) = \frac{1}{12}$. Use this to compute $\mu_Y = \mathrm{E}(Y)$ and $\sigma_Y^2 = \mathrm{Var}(Y)$. Then overlay the density histogram in part (b) with a plot of the normal curve with the values of μ_Y and σ_Y^2 you computed.

(d) (1 point) Does the Central Limit Theorem guarantee that this density histogram should be well approximated by the normal curve you found in part (c)? Does normal curve actually approximate density histogram well?

(e) (1 point) Suppose we didn't know the exact value of n, but did have the sample data from part (a). Find an estimator for n and compute it from this sample data.

#7) (4 points) A social scientist conducts an experiment by placing a participant alone in a waiting room. The scientist then measures the amount of time it takes until the participant begins looking at their cell phone. In a random sample of $n = 40$ participants, the average time was 42 seconds with a standard deviation of 12 seconds. Find a 99% two-sided confidence interval for the population mean for the time it takes a participant to begin looking at their cell phone.

#8) (4 points) The electronics retail giant, Buy Best, takes a random sample of 20 weekend shoppers. They find that these shoppers spend an average of \$76 with a standard deviation of \$21. They also take a random sample of 20 weekday shoppers and find that the shoppers spend an average of \$64 with a standard deviation of \$20. Is there sufficient evidence to claim that weekend shoppers spend more on average than weekday shoppers? Use a significance level of $\alpha = 0.05$.

#9) (extra credit) The file `final_steel.csv` contains measurements on various steel samples.

(a) (2 points) Give a scatter plot for the `elongation` (horizontal axis) and the `‘tensile strength‘` (vertical axis).

(b) (2 points) Using a simple linear model, compute estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ from the points in part (a). Then plot the resulting regression line together with the scatter plot.