Homework #6

Answer the following questions in a single R script called `hw06.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.

#1) We introduced the uniform distribution on an interval $[a, b]$ in Worksheet #5. Since it is one of the simplest continuous distributions, we didn't bother learning any of R's built-in functions for this distribution. That is about to change!

(a) Consider $X \sim \mathcal{U}_{[0,3]}$. Research the `runif` command. Use it to generate 1000 values for X. Store these values in a tibble. Display the results using a histogram.

(b) Modify the plot from part (a) so that the histogram and probability density function (`dunif`) are displayed together.

(c) Let's denote the random values generated in part (a) by x_i. Compute the following quantities from the tibble in part (a): $\max(x_i)$, $\frac{1001}{1000}\max(x_i)$ and $2\overline{x}$.

(d) Let's repeat the experiment in part (a) 1000 times. For each such experiment, compute $\max(x_i)$, $\frac{101}{100}\max(x_i)$ and $2\overline{x}$. Store these results in a tibble.

(e) Display the results of part (d) in overlapping histograms for $\max(x_i)$, $\frac{1001}{1000}\max(x_i)$ and $2\overline{x}$. Be sure to set `alpha = 0.5` for each histogram so that one histogram doesn't obscure another. Also set `binwidth=0.01`. Also be sure to choose different colors for each histogram.

(f) Compute the (sample) mean and standard deviation for the $\max(x_i)$, $\frac{1001}{1000}\max(x_i)$ and $2\overline{x}$. Display these results in a tibble (hint: `pivot_longer` and `summarize` can do this nicely).

(g) If you did part (e) correctly, you should see that $2\overline{x}$ has a bell-shaped histogram. This is a consequence of the Central Limit Theorem. In worksheet #7, we're told that if $X \sim \mathcal{U}_{[0,3]}$ then $\mathrm{E}(X) = \frac{0+3}{2} = 1.5$ and $\mathrm{Var}(X) = \frac{1}{12}(3-0)^2 = 0.75$. Plot the density histogram for $2\overline{x}$ along with the probability density function for the appropriate normal distribution. (Hint: $\mathrm{Var}(2\overline{X}) = 4\mathrm{Var}(\overline{X})$)

(h) Let $2\overline{X}$ denote the random variable corresponding to $2\overline{x}$. A consequence of the Central Limit Theorem is that $2\overline{X} \sim \mathcal{N}(\mu, \sigma^2)$. In part (g), you should have (essentially) computed μ and σ^2. With that in mind, find $P(\mu - .01 \leq 2\overline{X} \leq \mu + .01)$.

(i) Let $Y = \max(X_i)$ denote the random variable corresponding to $\max(x_i)$. Let $Z = Y/3$. Thus Z takes on values in $[0, 1]$. One can show that Z has a **beta distribution**. The beta distribution has two parameters: α and β. For Z, we have $\alpha = 1000$ and $\beta = 1$. In R, α and β are specified by the arguments `shape1` and `shape2` (respectively). Plot the probability density function for this beta distribution (given by `dbeta` in R) along with a density histogram for the values of $\max(x_i)$.

(j) Compute $P(\mu - .01 \leq \max(X_i) \leq \mu + .01) = P\left(\frac{\mu - .01}{3} \leq Z \leq \frac{\mu + .01}{3}\right)$.

(k) Compute $P\left(\mu - .01 \leq \max\left(\frac{1001}{1000} X_i\right) \leq \mu + .01\right)$.

#2) This exercise will have you use `hw06_creditcard.csv`. This file contains a list of all credit card transactions made in September 2013 by European cardholders. You can find a more thorough description of this data set at `https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud`. Ignore columns `V1` through `V28`. These columns contain numeric data that essentially anonymized cardholder information. The columns we will focus on are `Amount` and `Class`. `Amount` contains the amount of a transaction. `Class` has either a 0 or 1. A 0 indicates the transaction was legitimate. A 1 indicates the transaction was fraudulent.

(a) Store the above dataset as a tibble. Filter this dataset so that there are only amounts less than 2000 (euros). We do this to avoid outliers that might otherwise complicate our analysis. Plot a density histogram for the amounts in legitimate transactions. Then plot along with it a density histogram for the amounts in fraudulent transactions. Make sure the histograms have different colors and that they are semi-transparent. (Hint: You will need to reset the `data` in `geom_histogram` when plotting the density histogram for fraudulent transactions.)

(b) Compute the mean and standard deviation for legitimate transaction amounts.

(c) Compute the mean and standard deviation for fraudulent transaction amounts.

Interlude: There are 491 fraudulent transactions with amounts less than 2000. Suppose we didn't know these transactions were fraudulent. By looking at their mean only, can we tell that they are abnormal in comparison to means taken from legitimate transactions amount? The following parts have you figure this out.

(d) Suppose our data set is stored as `df`. The following code snippet randomly samples 491 legitimate transaction amounts and takes their sample mean:

```
df %>% filter(Class==0) %>% select(Amount) %>% pull %>% sample(size=491)
    %>% mean
```

Repeat this sampling 2000 times and store the results in a tibble. I recommend using the `rowwise` function to do this. This tibble may take a while to compute. Display these sample means in a density histogram.

(e) The density histogram in part (e) should be bell-shaped. This is a consequence of the Central Limit Theorem. In part (b), you should have found estimates for the population mean μ and standard deviation σ for legitimate transaction amounts. Assume these are the true values. Overlay the density histogram with the appropriate probability density function.

(f) Let $\overline{X}$ denote the random variable that gives the sample mean from a sample of 491 transactions from the population of legitimate transactions. Let $\overline{x}_{\text{fraud}}$ denote the mean of the 491 fraudulent transaction amounts. Find $P(\overline{X} \geq \overline{x}_{\text{fraud}})$.

(g) Find a value q so that $P(\overline{X} \geq q) = 0.00001$.