Homework #7: Due Wednesday, March 22 at 11:59PM

- Electronically submit their complete solutions in at least two separate files:
  - Upload an R Markdown file with extension .Rmd.
  - Upload the resulting **.pdf file** (NOT .html) you get after you knit the .Rmd file.
  - I should be able to knit your .Rmd file and get a similar .pdf file when I knit.
  - See file RMarkdown-HW7.Rmd for a template .Rmd file to get you going.
  - Note questions 3(a)-3(d) and 4 do not require R. You may type your work in the .Rmd file above, or if it is easier, you can scan/take a picture of written work and upload that as a separate pdf file. These questions are typed in red on the assignment.
- Assignments MUST:
  - Be your own work. Though you may collaborate with others, everyone is responsible for their own work and plagiarism of any form is not tolerated.
  - Be complete. You must provide all work and/or explanations needed to find the solution. Answers with insufficient or incomplete supporting work may lose credit.
  - Adhere to the Code of Academic Honesty.
  - Be easy to follow. Your solution to a problem must be typed using complete sentences and must include all required R code that is recreateable. You may lose credit for work that does is not fully explained or justified.
- Thanks!

1. Let X denote the morning commute time for a randomly selected person who commutes by car to work in downtown Denver. Let$Y$ denote the morning commute time for a randomly selected person who commutes by public transportation to work in downtown Denver. Let Z denote the morning commute time for a randomly selected person who commute to work by other means (such as walking, riding a bike).

   The morning commute time (in minutes) for people that commute by:

   - Car, X, is normally distributed with $\mu_X = 45$ minutes and $\sigma_X = 12$ minutes.
   - Public transportation, Y, is normally distributed with $\mu_Y = 40$ minutes and $\sigma_Y = 10$ minutes.
   - Other means, Z, is exponentially distributed with the average commute time $\mu_Z = 20$ minutes. *Recall: The rate parameter* $\lambda = \frac{1}{20}$.

   Let $X_1, X_2, \ldots, X_{12}$ denote a sample of 12 car commute times that are randomly and independently selected from X. Let $Y_1, Y_2, \ldots, Y_{10}$ denote a sample of 10 public transportation commute times that are randomly and independently selected from Y. Let $Z_1, Z_2, \ldots, Z_8$ denote a sample of 8 "by other means" commute times that are randomly and independently selected from Z.

   (a) Interpret the practical significance (in terms commute times) of the random variable W which is defined as

   $$W = \frac{12}{30}\overline{X} + \frac{10}{30}\overline{Y} + \frac{8}{30}\overline{Z}.$$

   (b) Compute the values of $E(W) = \mu_W$ and $\text{Var}(W) = \sigma_W^2$ by filling in the blanks in the R code block in the .Rmd version of this assignment. Assume that the X_i's, Y_i's, and Z_i's are all independent.

   (c) Suppose the distribution for W is normally distributed. Use the values for μ_W and σ_W^2 from the previous part to compute $P(W < 30)$.

   (d) Simulate the sampling distribution of W by completing the partially complete code in the R code block in the .Rmd file for this assignment. Plot the results of this sampling distribution in a histogram. Mark a vertical line at the value $W = 30$ in your histogram.

   (e) Use your simulation to estimate the value of $P(W < 30)$.

   (f) Why do you suspect your approximations in (c) and (e) vary so drastically? Explain in one or two complete sentences.

2. A health clinic screens every patient they see for influenza as part of their routine exam. In order to do a quality control check on the accuracy of the influenza test kits a health clinic recently received, they randomly select 60 patients and look at whether each had a positive influenza test when they visited the clinic. Let X denote the number of positive tests in the sample of 60 randomly selected tests.

   (a) Suppose a health clinic knows that 12% of the general population in the area is infected with influenza, and assume the results of individual tests are independent from each other, so each person has a 12% chance of getting a positive test. Using a binomial distribution, calculate $P(8 \leq X \leq 10)$.

   (b) Calculate the mean and variance of the number of positive tests that would occur out of a random sample of 60 influenza tests.

   (c) Use a normal distribution to approximate the binomial distribution **without a continuity correction** (and your answers from the previous part), approximate $P(8 \leq X \leq 10)$.

   (d) Now use a normal distribution to approximate the binomial distribution along **with a continuity correction** to approximate $P(8 \leq X \leq 10)$.

   (e) Compare the accuracy of the two previous answers with your answer in part (a). Which is more accurate?

3. Let sample $X_1, X_2, \ldots, X_n$ be independently chosen at random from a population $X \sim \text{Exp}\left(\frac{1}{2}\right)$ with corresponding pdf $f(x) = \frac{1}{2}e^{-\frac{x}{2}}$ for $x \geq 0$ and 0 otherwise.

   You may not use R to complete parts (a), (b), (c) and (d). You may either type your work in the .Rmd file which might not be easy since it is a lot of math equations. It might be easier to write your work on paper and upload a scanned copy in a separate file.

   (a) Find a formula for the $F(x) = P(X < x)$, the cdf of the population. Be sure to show steps of calculus involved and specify the domain.

   (b) Find a formula for $F_{X_{\text{min}}}(a) = P(X_{\text{min}} < a)$, the cdf of X_{min}. Be sure to specify the domain. *Hint: Note the following;*

   $$F_{X_{\text{min}}}(a) = P(X_{\text{min}} < a) = 1 - P(X_{\text{min}} \geq a) = 1 - P(X_1 \geq a, X_2 \geq a, \ldots, X_n \geq a).$$

   (c) Find a formula for $f_{X_{\text{min}}}$, the pdf of X_{min}. Be sure to specify the domain.

   (d) If $n = 10$, use the result from part (b) to compute $F_{X_{\text{min}}}(1) = P(X_{\text{min}} < 1)$, the probability that the smallest value in the sample is less than 1.

Complete Parts (e) and (f) on next page in the .Rmd file for this assignment.

(e) Complete the partially completed R code block in the .Rmd version of this assignment to simulate the sampling distribution for the minimum value when $n = 10$. Plot your results in a histogram, and mark a vertical line in the histogram at $X_{\text{min}} = 1$.

(f) Use your simulation to estimate the value of $P(X_{\text{min}}) < 1)$.

You may not use R to complete question 4. You may either type your work in the .Rmd file or if easier, you may write your work on paper and upload a scanned copy of your work on questions 3 and 4 in a separate file.

4. Consider the sample $\{2, 30, 110\}$.

   (a) Write out all possible bootstrap samples. You should ignore the ordering. For example a bootstrap sample of $\{2, 2, 110\}$ as the same as $\{2, 110, 2\}$, so you should not list both. *Hint: There are exactly 10 distinct bootstrap samples.*

   (b) Based on your bootstrap samples, what is the probability that bootstrap sample has a mean less than or equal to 40?

   (c) Based on your bootstrap samples, what is the probability that bootstrap sample has a maximum value less than or equal to 40?

5. Recall that the dataset *Bangladesh* contains data from 271 randomly selected groundwater samples. The dataset contains measurements on the arsenic, cobalt and chlorine concentration levels (in parts per billion, ppb).

   (a) Conduct exploratory data analysis (EDA) on the cobalt concentrations. Note there are some observations in *Bangladesh* that do not have a cobalt level recorded. We want to ignore these observations before doing EDA. The command

   ```
   cobalt <- na.omit(Bangladesh$Cobalt)
   ```

   creates a data vector called cobalt which contains only the observations that have a value entered for the variable Cobalt in *Bangladesh*. After filtering out the observation(s) with no Cobalt value(s), answer the questions below.

      i. Create a histogram of the sample data, and give the mean and standard deviation of the cobalt concentration levels of the sample.

      ii. Create a box plot of the sample data, and give the five number summary.

      iii. Create a normal quantile plot of the data and describe the shape. Is the data normal? skewed? If skewed in what direction?

   (b) Create a bootstrap distribution for the sample mean from the original sample. Use $N = 10^4$ as the number of bootstrap samples.

      i. What is the mean of the bootstrap distribution?

      ii. What is the bootstrap standard error?

      iii. What is the bootstrap estimate of the bias?

      iv. Calculate the ratio of the bootstrap bias over the bootstrap standard error. Does this exceed the rule of thumb level for having a substantial effect on the accuracy of the estimate?

   (c) Give a 95% bootstrap percentile confidence interval to estimate the mean cobalt level in all groundwater in Bangladesh. Then create a histogram of the bootstrap distribution with vertical lines indicating the cutoffs for the confidence interval.

   (d) Interpret the practical meaning of the confidence interval in a complete sentence.

6. Researchers conducted a study of primary and early secondary school children in Italy to examine the gender differences in math anxiety[1]. One of the measures used is the Abbreviated Math Anxiety Scale (AMAS), a self-reported math anxiety questionnaire. AMAS scores range from 9 to 45, with a higher score representing more math anxiety. The dataset *MathAnxiety* contains the results for a subset of the children in the original study.

   (a) How many children are in the dataset? How many of the children identify as a boy? How many of the children identify as a girl?

   (b) Create side-by-side box plots to show the math anxiety scores broken down by gender. Based on your side-by-side box plots, make three comparative statements about the data. An example of a comparative statement would be: 75% of the children that identified as a boy had a score less than 100 compared to the children that identified as a girl which has 50% with a score less than 100. (Note this is not actually a true statement. Just intended to provide an example.)

   (c) What is the difference in the observed sample mean AMAS scores between the boys and girls?

   (d) Create a bootstrap distribution for the difference in the mean anxiety score between boys and girls using the original sample. Use $N = 10^4$ as the number of bootstrap samples.

      i. What is the mean of the bootstrap distribution?

      ii. What is the bootstrap standard error?

      iii. What is the bootstrap estimate of the bias?

      iv. Calculate the ratio of the bootstrap bias over the bootstrap standard error. Does this exceed the rule of thumb level for having a substantial effect on the accuracy of the estimate?

   (e) Find a 90% bootstrap percentile confidence interval to estimate the difference in the mean AMAS scores between all boys and girls. Then create a histogram of the bootstrap distribution with vertical lines indicating the cutoffs for the confidence interval.

   (f) Based on your answer to the previous part, do you believe it is plausible that the boys and girls have different levels of math anxiety? Support your answer with an explanation.

[1] Hill et al. (2017)