STAT 345 Fall 2019
Challenge Homework - More Confidence Intervals Name: ____________________

FOR THIS HOMEWORK, YOU WILL NEED TO PRINT YOUR CODE AND ATTACH IT TO THE BACK OF YOUR SUBMISSION.

1. t -procedures. Consider n observations $X_1, X_2, \cdots X_n$ which are iid samples from a Log Normal distribution with population mean μ and population skew γ. The following code can be used to simulate n observations from this population.

```
generate_data <- function(n, mean, skew){
  zeta <- optim(skew^.67, function(z) ((z+2)*sqrt(z-1)-skew)^2,
                method='Brent', lower=1, upper=1e9)$par
  omega <- sqrt(log(zeta))
  theta <- log(mean) - omega^2/2
  x <- rlnorm(n, theta, omega)
}
```

For the rest of this problem, assume that $\mu = 25$.

a) Use the provided function to simulate $n = 30$ observations from the population with $\gamma = 0.5$. Make a histogram of the data, and explain why the t-procedures assumptions are (roughly) met.

b) Use t-procedures to construct a 95% confidence interval for μ. Interpret this interval.

c) Perform a *simulation study* by repeating the process above (a) and b)) $M = 10,000$ times. You can use the following code as a "skeleton". Count the number of times that the 95% CI actually succeeds in capturing the true value of μ. How close is this "coverage" to the desired value of 95%?

```
M <- 10000   #Number of simulations
n <- 30      #Sample size
skew <- .5   #Population skewness
count <- 0.  #Number of times CI succeeds
#Start simulations
for(i in 1:M){
  #Simulate data
  x <- generate_data(n, 25, skew)
  #Calculate CI endpoints here (use t-procedures)

  #Check to see if CI captures the true value
  if(change_this_part <- TRUE){
    count <- count + 1
  }
}
print(100*count/M)
```

d) Repeat your simulation study for sample sizes of $n = 100$ and $n = 500$. What happens and why?
e) Repeat the entire simulation study (for all three values of n) after increasing the skew to $\gamma = 10$. Explain what happens and why. Repeat this again for $\gamma = 50$.

2. The Bootstrap and the Accelerated Bootstrap. The iris data set is a famous data set consisting of measurements for 150 iris flowers. The "petal length" of these 150 flowers can be obtained by typing `iris$Petal.Length` in R. Create a histogram of the data, and note that the data is "bimodal".
The Bimodality coefficient is a statistic which can be calculated from the data as

$$BC = \frac{\hat{\gamma}^2 + 1}{\hat{\kappa} + 3\left(\frac{(n-1)^2}{(n-2)(n-3)} - 1\right)}$$

where $\hat{\gamma}$ and $\hat{\kappa}$ are the sample *skew* and sample *kurtosis* of the data respectively

$$\hat{\gamma} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_x}\right)^3 \qquad \hat{\kappa} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_x}\right)^4.$$

a) Write a function in R which calculates the bimodality coefficent from the data. What is the estimated value of BC for the petal length data?

b) Roughly speaking, larger values of BC indicate bimodality, where 0.47 is often used as a cut-off (this is the value of BC for a uniform distribution). Does the estimate of BC for the petal length data seem to capture the bimodality of the data?

c) Use the Bootstrap algorithm to approximate the sampling distribution of BC for the petal length data. Provide a histogram of the bootstrap distribution and also calculate a 95% Bootstrap CI for the true value of the bimodality coefficient. Do you feel confident in asserting that the true value of BC is greater than 0.47?

d) **Challenge:** For (even more) challenge points, construct a CI using *accelerated bootstrap* techinque. (Link on course web-page).

3. Asymptotic normality of the MLE. Consider $X_1, X_2, \cdots X_n \overset{iid}{\sim} Beta(\theta, 1)$. The density function corresponding to this distribution is

$$f(x_i|\theta) = \theta x_i^{\theta-1},\ 0 < x_i < 1$$

a) Find the MLE of θ. (Hint: We did this one in class). **b)** Find the asymptotic standard deviation of $\hat{\theta}$.

$$SD(\hat{\theta}) \approx \left(-E\left[\frac{d^2 \log f(x_i|\theta)}{d\theta^2}\right]\right)^{-1/2}$$

c) Use the fact that, as $n \to \infty$,

$$Z = \frac{\hat{\theta} - \theta}{SD(\hat{\theta})/\sqrt{n}} \overset{approx}{\sim} N(0,1)$$

to construct a 95% CI for θ¿
d) Simulate data `rbeta(300, 4, 1)` and use your answer to c) to calculate a 95% CI for θ. Did it capture the true value?