STAT 345 Spring 2018
Homework 10 - Sampling Distributions Name: _____

Please adhere to the homework rules as given in the Syllabus.

1. The number of eggs laid by Green Sea Turtles is approximately normal with mean 100 and standard deviation 20.

a) What is the probability that a single Green Sea Turtle lays more than 108 eggs?

b) In a random sample of 25 Green Sea Turtles, what is the probability that the sample average of number of eggs laid is greater than 108?

2. Central Limit Theorem.

a) Suppose that $X_1, X_2, \cdots X_{100}$ are independent and identically distributed Exponential RV's with $\lambda = 1$. Find the mean and variance of $\bar{X}$. Use the CLT to approximate $P(\bar{X} > 1.05)$.

b) Suppose that $X_1, X_2, \cdots X_{144}$ are independent and identically distributed (continuous) Uniform distributed RV's with $a = 0$ and $b = 6$. Use the CLT and then work backwards to find x such that $P(\bar{X} < x) = 0.05$.

c) Suppose that $X_1, X_2 \cdots X_{36}$ are independent and identically distributed Normally distributed RV's with mean $\mu = 0$ and standard deviation σ. Also assume that $P(\bar{X} > 1) = 0.01$. What is the value of σ?

3. Let $X_1, X_2, \cdots X_n$ be discrete iid random variables with the following PMF.

$$f(x_i) = \frac{\tau(\ln(1/\tau))^{x_i}}{x_i!}, x_i = 0, 1, 2, \cdots, \quad \tau \in (0, 1)$$

a) Define the statistics,

$$T_i = \begin{cases} 1, & if\ X_i = 0 \\ 0, & otherwise \end{cases}$$

What is the sampling distribution of T_i? What is the sampling distribution of $\sum_{i=1}^{n} T_i$?

b) Consider the estimator $\hat{\tau} = \frac{1}{n}\sum_{i=1}^{n} T_i$. Find the bias and variance of $\hat{\tau}$.

c) If $n = 10$ and $\tau = 0.9$, find $P(\hat{\tau} \leq 0.8)$.

4. Exponential Distribution. Let $X_1, X_2, \cdots X_n \overset{iid}{\sim} Exp(\frac{1}{\beta})$. Note that this is an alternative parameterization where the CDF is

$$F(x) = 1 - e^{-x/\beta}$$

This parameterization is convenient, especially for estimation problems, because $E(X) = \frac{1}{1/\beta} = \beta$. Let $X_{(1)}$ be the minimum of $X_1, X_2, \cdots X_n$. This is a statistic (since it only depends on the data), and this problem focuses on fidning the *sampling distribution* of $X_{(1)}$.

a) Let $F_1(x)$ be the CDF of $X_{(1)}$. Find $F_1(x)$.

b) Find $f_1(x)$, the PDF of $X_{(1)}$. What is the distribution of $X_{(1)}$?

c) Suppose we are interested in estimating β. Consider the estimator $\tilde{\beta} = nX_{(1)}$. Show that $\tilde{\beta}$ is unbiased for β.

d) Find the variance of $\tilde{\beta}$. Is this estimator *consistent*?

5. Central Limit Theorem and the Bootstrap. R contains a built in dataset called `rivers` which contains the length of the 141 "Major" rivers North America. This data can be found by just typing `rivers` into RStudio. Take a random sample of $n = 30$ of these rivers with the following R code: (*Note: setting the seed just ensures that everybody gets the same "random" sample. Makes it easier to grade.)*

```
set.seed(117)
x <- sample(rivers, 30, replace=FALSE)
```

The CLT says that the sampling distribution of $\bar{X}$ should be approximately normal. But this depends on the skew of the original data and the sample size n.

a) Make a histogram of the sampled data. Comment on it's skew.

The Bootstrap lets us approximate the sampling distribution of the data, and gives us a way of checking if the sample size is large enough so that the sampling distribution is indeed Normal.

b) Approximate the sampling distribution of the sample mean of the river data using a Bootstrap by filling in the following code. Make a Histogram and QQ-plot of the "Bootstrap distribution". Is the sample size large enough for the CLT to apply here?

```
M <- 10000
boot_samples <- rep(NA, M)
for(i in 1:M){
   temp <- sample(x, length(x), replace=TRUE)
   boot_samples[i] <- #FILL IN THE REST OF THIS LINE
}
#MAKE A HISTOGRAM OF THE BOOTSTRAP DISTRIBUTION HERE
#MAKE A QQ-PLOT OF THE BOOTSTRAP DISTRIBUTION HERE
```

c) Repeat parts *b*) and *c*), but keep all $n = 141$ observations. Does the CLT apply now?

Comment about R: If you include the line `par(mfrow=c(1,3))` *at the beginning of your code, Rstudio will automatically place the three plots side by side for you.*

6. Challenge Problem. *The German Tank Problem.* During World War II, the Allies made substantial efforts to determine the extent of German production. They approached this in two different ways: conventional intelligence gathering and statistical estimation. In many cases, such as the estimating the number of German tanks, the statistical approach wins by a mile.

Month	Statistical estimate	Intelligence estimate	German records
June 1940	169	1000	122
June 1941	244	1550	271
August 1942	327	1550	342

At this point, the Germans were labeling their tanks with serial numbers, sequentially from 1 to N where N is the number of tanks the Germans had. The Allies were interested in estimating N. Consider the following scenario.

- Assume that the true value of N is 342.
- Assume that the Allies captured $k = 10$ tanks and assume that any tank is equally likely to be captured.
- Let $X_1, X_2, \cdots X_{10}$ be the serial numbers of the 10 tanks that were captured by the Allies. You can simulate this data in `R` by typing `x <- sample(342, 10, replace=F)`.

In this problem we will consider 4 estimators.

i) The MLE for N is the maximum observation, $\hat{N}_1 = X_{(n)}$.

ii) The MoM for N is simply $\hat{N}_2 = 2\bar{X}$.

iii) Another reasonable estimator for N is $\hat{N}_3 = \bar{X} + 1.74 \cdot S$.

iv) The "Max + average gap" estimator is obtained by taking the MLE and adjusting it so that it becomes unbiased. This estimator of N is $\hat{N}_4 = X_{(n)}(k+1)/k - 1$. *Note: This is the estimator which was actually used by the Allies. It is also the so-called UMVUE, the minimum variance unbiased estimator.*

The Problem: Construct approximate sampling distributions in the form of a histogram for each of these 4 estimators.

1. Simulate data by typing `x <- sample(342, 10, replace=F)`.
2. Calculate each of the 4 estimators based on this sample.
3. Save these values, and repeat this process 1000 times. Use the sample code on the course web-page or see me if you need help with this step.
4. Create a Histogram showing the sampling distribution for each estimator. On each plot, include a vertical line showing the true value $N = 342$. Use the `xlim` argument to set the x-axis scale to be the same for each plot.
5. Calculate the estimated bias and variance of each estimator.
6. Based on 4 and 5, compare the estimators. Which one would you prefer?