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

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

1. Continuation of problem 3 from homework 9.. For many distributions, we can prove that the sample mean (Kimothys estimator) is the "best" *unbiased* estimator. But if we are willing to accept some bias in exchange for reduced variance, we may be able to find a better estimator (at least according to MSE). Consider a new estimator

$$\hat{\theta}_c = \frac{X_1 + X_2 + X_3}{c}$$

where c is some positive constant.

a) Find the MSE of this estimator in terms of c.

b) Use a calculus argument to find the value of c which minimizes the MSE. Is it different from Kimothy's estimator? If so how, and why does this reduce the MSE?

2. Timothy and Jimothy each have 20 chores to do. The time it takes Timothy to complete a chore is a random variable with mean 50 minutes and sd 10 minutes. The time it takes Jimothy to complete a chore is a random variable with mean 52 minutes and SD 15 minutes. Assume the time it takes to complete a chore is independent of all other chores.

a) Let T be the number of minutes it takes Timothy to complete all his chores. Using the CLT, what is the probability that he completes his chores in under 900 minutes?

b) Let J be the number of minutes it takes Jimothy to complete all his chores. Using the CLT, what is the probability that he completes his chores in under 900 minutes?

c) What is the probability that Timothy completes his chores before Jimothy? *Hint: What is the distribution of* $T - J$*?*

3. 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 can be calculated from the data), and this problem focuses on finding 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*?

4. Maximum Likelihood. The Rayleigh Distribution has probability density function,

$$f(x_i) = \frac{x_i}{\theta} e^{-x_i^2/(2\theta)},\ x_i > 0,\ \theta > 0$$

and cumulative distribution function (CDF)

$$F(x_i) = 1 - e^{-x^2/(2\theta)},\ x_i > 0,\ \theta > 0$$

a) Use the CDF to find τ, the 99^{th} percentile of the Rayleigh Distribution. *Hint: Set $F(\tau) = p$ and solve for τ.*

c) Assume that $X_1, X_2, \cdots X_n \overset{iid}{\sim} Rayleigh(\theta)$. Find the ML estimator of θ

d) Use the invariance property of the MLE to find the ML estimator of τ. What is the maximum likelihood estimate of τ when $n = 30$ and $\sum_{i=1}^{30} x_i^2 = 180$?

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 a statistic, 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. *Maximum Likelihood Data Analysis.* The goal of this problem is to repeat the analysis in **problem 5 of homework 9**, but this time using Maximum Likelihood. Assume that $X_1, X_2, \cdots X_n \overset{iid}{\sim} Beta(\alpha, \beta)$. Recall that

$$f(x_i|\alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x_i^{\alpha-1}(1-x_i)^{\beta-1}$$

a) Find the likelihood function $L(\alpha, \beta|x_1, \cdots x_n)$ by taking the product of the $f(x_i|\alpha, \beta)$. Then find the Log-likelihood function. Show all of your work for credit, but you should get:

$$\log L(\alpha, \beta|x) = n \log\left(\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\right) + (\alpha-1)\sum_{i=1}^{n} \log x_i + (\beta-1)\sum_{i=1}^{n} \log(1-x_i)$$

b) Write a function in R called `log_likelihood(theta, x)` which takes two arguments `theta` and `x` and returns the Log-likelihood. Where `theta` is a vector (α, β) and x represents the data. Once you have written this function, the Maximum Likelihood Estimates can be found by using R's optimization routine. *Note: a0 and b0 are supposed to be good starting guesses. The MoM estimates from problem* 5 *are a reasonable choice here.*

```
optim(c(a0, b0), fn=log_likelihood, x=x, control=list(fnscale=-1))
```

Report the maximum likelihood estimates $\hat{\alpha}$ and $\hat{\beta}$. Do they differ from the MoM estimates? Repeat part *b*) (i.e. produce the histogram) of problem 4 using these ML estimates.