STAT 345 Spring 2018
Homework 9 - Estimation and Sampling Distributions I Name: ______________________________

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

1. Consider a population of people with high blood pressure (lets say Systolic BP = 135 mmHg). When a member of this population takes a certain medicine, their Systolic blood pressure will be reduced by by X mmHg where X is Normally distributed with mean $\mu = 12$ and sd $\sigma = 4$.

a) What is the probability that a single individual, upon taking this medicine, has their SBP decrease by less than 7 mmHg?

b) In a random sample of 5 members of this population, what is the probability that their average SBP decreases by less than 7 mmHg?

c) For a random sample of size n from this population ($n = 1, 2, \cdots 10$), make a plot (using a computer) of the probability that the average SBP decrease is less than 7 mmHg.

2. Central Limit Theorem Calculations.

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. Mean Squared Error. Suppose that X_1, X_2 and X_3 are independent random variables with mean θ and variance θ^2. Timothy, Jimothy, Kimothy and Bob each suggest a possible estimator for θ.

$$\hat{\theta}_T = X_1 \qquad \hat{\theta}_J = \frac{X_1 + 2X_2 + 3X_3}{6} \qquad \hat{\theta}_K = \frac{X_1 + X_2 + X_3}{3} \qquad \hat{\theta}_B = \frac{X_1 + X_2 + X_3}{7}$$

a) Find the Bias, Variance and MSE of each estimator. Which estimator is the "best" according to MSE?

b) Using a computer, create a plot of the MSE as a function of θ, for values of θ between 0 and 4. Use different colors for each estimator.

4. Method of Moments.

a) Professor Halfbrain conducts the following random experiment with a fair six-sided die. He rolls the die η times and records how many times he rolls a 1. He repeats this 5 times and collects the data $x_1 = 3$, $x_2 = 1$, $x_3 = 2$, $x_4 = 1$ and $x_5 = 3$. Unfortunately... he can't remember what the value of η was. Help him estimate η using Method of Moments.

b) Alf wants to know the probability of getting a match each time he swipes right on Tinder (call this parameter θ), so he conducts the following random experiment. Each day for a week, he will swipe right until he gets a match and then he stops. He collects the following data, $(18, 15, 4, 35, 17)$. Use Method of Moments to estimate θ.

c) **The Log-Normal Distribution.** Assume that $Y_1, Y_2, \cdots Y_n$ follow a Log-Normal distribution with parameters θ and ω. Recall that $E(Y) = e^{\theta+\omega^2/2}$ and $Var(Y) = e^{2\theta+\omega^2}(e^{\omega^2} - 1)$. Find $\hat{\theta}_{mom}$ and $\hat{\omega}_{mom}$, the Method Moments estimators. *Hint: Notice that* $Var(Y) = E(Y)^2(e^{\omega^2} - 1)$.

5. **Data Analysis.** The CDI dataset contains Demographic information for the 440 most populated counties in the United States. See the Chapter 6 lecture notes if you need a reminder on how to read this dataset into R. We will consider the variable "Percent of Population with a Highschool diploma". Divide this variable by 100 to give you proportions instead of percentages.

a) Create a histogram of this data. Use the option `freq=FALSE` inside the `hist()` function. You don't have to turn in this plot.

b) The Beta distribution is an excellent candidate for this distribution. The Method of Moments estimators are given by

$$\hat{\alpha} = \bar{X}\left(\frac{\bar{X}(1-\bar{X})}{S^2} - 1\right)$$
$$\hat{\beta} = (1-\bar{X})\left(\frac{\bar{X}(1-\bar{X})}{S^2} - 1\right)$$

Calculate the MoM estimates for the HS Diploma data. Create another histogram (with the `freq=FALSE` option), and use `curve()` to plot the fitted Beta distribution over the histogram. How does the fit look? Turn in this plot.

c) The CDF of a beta distribution can be computed in R as

$$F(x) = \texttt{pbeta(x, alpha, beta)}$$

Using your answer estimates from part *b*), estimate the probability that a county has between 0.6 and 0.9 of the population with a HS Diploma.

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 (truth)
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 gearbox 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 $\hat{N}_2 = 2\bar{X} - 1$.

iii) Another reasonable estimator for N is $\hat{N}_3 = \bar{X} + 1.73 \cdot S$.

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 $\hat{N}_4 = X_{(n)}\frac{k+1}{k} - 1$. *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?