STAT 345 Spring 2018
Homework 9 - Point Estimation

Name: ______________________________

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

1. 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.

2. Continuation of Problem 1. 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?

3. Weighted Means. NOTE: Nobody is required to do this problem, it's good practice for MSE, so I'm leaving it for reference. Suppose $X_1, X_2, \cdots X_n$ are independent random variables but they are not necessarily identically distributed. Specifically, let us assume that $E(X_i) = \mu$ and $Var(X_i) = i^2$. Consider the weighted mean estimator

$$\hat{\mu}_w = \sum_{i=1}^{n} w_i X_i$$

where the w_i are constants such that $\sum_{i=1}^{n} w_i = 1$.

a) Show that $\hat{\mu}_w$ is unbiased and that the variance is $\sum_{i=1}^{n} w_i^2 i^2$.

b) If we set $w_i = 1/n$ for $i = 1, 2, \cdots n$, then we get back the original sample mean $\bar{X}$. Find the MSE of this estimator for $n = 30$ using the results from part a).

c) In this problem, we expect that X_i is closer to μ for smaller values of i, since the variance is smaller. Therefore we may want to weight these observations more heavily when estimating the mean. If we set $w_i = \frac{1/i}{\sum_{i=1}^{n} 1/i}$, then we get the following estimator for μ.

$$\hat{\mu}_w = \frac{\sum_{i=1}^{n} X_i/i}{\sum_{i=1}^{n} 1/i}$$

Using the results from part a), find the MSE of this estimator for $n = 30$. Is it smaller than the original sample mean? *Hint: The sum* $\sum_{i=1}^{n} 1/i$ *is called the* n^{th} *Harmonic number. This sum is very well approximated by* $0.577 + \ln(n)$. *You may use this approximation in your answer.*

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

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

a) Find the CDF of the Rayleigh distribution.

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

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

d) Use the invariance property of the MLE to find the MLE of $Q_{.99}$. What is the maximum likelihood estimate of $Q_{.99}$ when $n = 30$ and $\sum_{i=1}^{30} x_i^2 = 180$?

7. Challenge Problem. *Maximum Likelihood Data Analysis.* The goal of this problem is to repeat the analysis in problem 5, 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 `neg_log_likelihood(theta, x)` which takes two arguments `theta` and `x` and returns the *negative* 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: the optim() function minimizes the function fn, which is why we give it the negative log-likelihood. Also, 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=neg_log_likelihood, x=x)
```

Report the maximum likelihood estimates $\hat{\alpha}$ and $\hat{\beta}$. Do they differ from the MoM estimates? Repeat parts *b*) and *c*) of problem 5 using these estimates.