Stat 111:
Introduction to Statistical Inference, 2017-18
Susan A. Murphy & Neil Shephard
Assignment 2

Credit will be given to the best grades from 3 individually marked questions (each question gets equal credit). Answer as many questions are you would like.

1. Consider the population of all major league baseball players as reported in USA Today on April 5, 1994. Their salaries can be found at
`https://www.dartmouth.edu/~chance/teaching_aids/data/baseball_salaries.html`.
You may find it easier to input into `R` using the csv file we put on Canvas under the files directory. The file is called `baseball.csv`. Note the population size is $K = 747$. Label their salaries by $y_1, y_2, \ldots, y_K$.

   (a) Calculate the mean salary in this population,

$$\mu = \frac{1}{K}\sum_{i=1}^{K} y_i$$

   and the standard deviation in this population,

$$\sigma = \sqrt{\frac{1}{K}\sum_{i=1}^{K}(y_i - \mu)^2}.$$

   (b) It is generally unusual to have access to the entire population: having this population allows us to calculate μ, σ. Suppose we did not have access to this population but instead we were able to collect a simple random sample of size $n = 5$ of the baseball players.
   What is random and what is fixed?

   (c) Suppose $R = 100$ baseball enthusiasts each collect a simple random sample of $n = 5$ players (let's not think of how much this would have aggravated the baseball players and let us pretend that the baseball players would have been willing to disclose their salary!). [Hint: in `R` you can use the "sample" command with this population to mimic the collection of a simple random sample by each baseball enthusiast, e.g.
   `sample(mydata$Salary,size=5,replace=T)`].
   Calculate the sample average, $\bar{Y}^*$ for each of the 100 enthusiasts. Each of the 100 sample means can be considered a proxy for μ. Calculate the standard deviation across the 100 sample averages. If this standard deviation is high then we would be reluctant to use a particular baseball enthusiast's sample average as a proxy for μ. Now suppose instead that each enthusiast collects a simple random sample of $n = 20$ baseball players; calculate the standard deviation across the 100 sample averages. Show your results and explain why these sample averages with $n = 20$ are better or not better than the sample averages based on random samples of size $n = 5$.

   (d) Repeat one more time for simple random samples of size $n = 80$. Now calculate a proxy for the standard deviation of a sample average using the 100 sample averages.

i. How should this be related to the population standard deviation σ? Do this for all three scenarios (sample sizes of $n = 5, 20, 80$).

ii. The population standard deviation, σ is the standard deviation between ***what*** in the population?

iii. The standard deviation of $\bar{Y}^*$ measures the variation between what?

iv. Explain how these two standard deviations are related, even though they are standard deviations between different things!

2. (This question uses the same data as Question 1). The bootstrap is rather amazing. Consider the population of all major league baseball players as reported in "USA Today" on April 5, 1994. Their salaries can be found at
`https://www.dartmouth.edu/~chance/teaching_aids/data/baseball_salaries.html`.
You may find it easier to input into R using the csv file we put on Canvas under the files directory. The file is called `baseball.csv`. Note the population size is $K = 747$. Label their salaries by $y_1, y_2, \ldots, y_K$. Calculate (or use your answer from Exercise 1 if you have already done this) the mean salary and the standard deviation of salaries for this population,

$$\mu = \frac{1}{K}\sum_{i=1}^{K} y_i, \quad \sigma = \sqrt{\frac{1}{K}\sum_{i=1}^{K}(y_i - \mu)^2},$$

respectively. Here we study proxying μ using a single sample from this population — in real life we usually only have access to one simple random sample!
Generate a single simple random sample of size $n = 20$ from this population. We will write this sample as $Y_1^*, ..., Y_n^*$ and report its sample average $\overline{Y}^*$. There are two ways to estimate the standard deviation of $\overline{Y}^*$.

(a) One way is to use the square root of the formula for the variance of the sample mean given before the Uber Example in Chapter 2 of the lecture notes,

$$\text{Var}^*\left(\overline{Y}^*\right) = \frac{1}{n}\sigma^2.$$

Do this for a random sample of size $n = 20$ (you will need a proxy for σ). Explain your work.

(b) If you did not know this formula the bootstrap can be used. Use $R = 5,000$ bootstrap samples (based on the $Y_1^*, ..., Y_n^*$) to do this.

(c) Compare the standard deviation calculated via bootstrap to the standard deviation calculated via the formula. Does this comparison improve if you consider random samples of size $n = 80$ instead? Show your work.

3. Let $Y \sim$Binomial$(7, 0.1)$.

(a) Derive F_Y, the cumulative distribution function for Y.
Now suppose I draw a random (i.e. i.i.d.) sample of size $n = 40$ from F_Y, denoted by $Y_1, Y_2, \ldots, Y_n$. Then the empirical distribution function is given by

$$\hat{F}_n(y) = \frac{1}{n}\sum_{i=1}^{n} 1_{Y_i \leq y}.$$

(b) Derive $\text{Var}\{\hat{F}_n(y)\}$.

(c) Next using R (or your favorite programming language), create a data set composed of random sample of $n = 40$ draws from F_Y.

i. Compute the empirical distribution function, $\hat{F}_n$ and graph $\hat{F}_n(y)$ on the y-axis versus y on the x-axis.

ii. How similar is $\hat{F}_n$ to F_Y?

(d) Draw 100 random samples of size 40 from the distribution F_Y. Compute the empirical distribution function for each of these random samples.

i. Calculate the standard deviation of $\hat{F}_n(2)$, namely $S_{\hat{F}_n(2)}$, from the 100 $\hat{F}_n(2)$'s.

ii. Compare this standard deviation to $\sqrt{\text{Var}\{\hat{F}_n(2)\}}$. The difference should be very small ($S_{\hat{F}_n(2)}$ is a good proxy for $\sqrt{\text{Var}\{\hat{F}_n(2)\}}$).
Remember to provide your code and all results.

(e) With this in mind, explain in words what you understand by $\text{Var}\{\hat{F}_n(2)\}$ (variance across what?) and compare it what you understand by $S^2_{\hat{F}_n(2)}$.

4. Consider some data $y_1, ..., y_n$. Write the sorted version of the data as $y_{(1)} \leq y_{(2)} \leq ... \leq y_{(n)}$. For simplicity assume there are no ties in the data. In this exercise (as we do throughout our lecture notes) use

$$Q_n(p) = \inf\left\{y \in R : p \leq \widehat{F}_n(y)\right\}, \quad p \in [0,1],$$

as the defintion of the p-th quantile. Recalling inf denotes infimum, the smallest value of y so that $\widehat{F}_n(y)$ is bigger than or equal to p. Here $\widehat{F}_n(y)$ is the empirical distribution function. Explain why

$$Q_n(p) = Y_{(\lceil np \rceil)},$$

where $\lceil x \rceil$, the "ceiling function", is the smallest integer above x. For 90% of the grade illustrate why using an example with $n = 4$ made up data points (hint: draw the empirical distribution function). For 100% of the grade prove it in general for any n.

5. Statisticians often use results derived using the Delta method, which is a relatively simple result from probability theory which did not appear in Joe's book. An appendix to Chapter 3 discusses the Delta method. Read that subsection. Then answer the following question.
Suppose $Y_i \overset{iid}{\sim} Exp(1/\mu)$, exponential random variables, are the lifetime of life bulbs. So $\text{E}(Y_1) = \mu$ and $\text{Var}(Y_1) = \mu^2$. Then the (Lindeberg-Lévy) Central Limit Theorem holds and the average length of the life bulbs is centered at the mean

$$\sqrt{n}\left(\overline{Y} - \mu\right) \xrightarrow{d} N(0, \mu^2), \quad \overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i,$$

as $n \to \infty$. What asymptotic distribution do these transformed statistics follow
(i) $\log(\overline{Y})$?
(ii) $\overline{Y}^{1/2}$?
(iii) $1/\overline{Y}$?

25th January 2018
To be handed in by 1.05pm, Tuesday, 13th February (week 4).