ST 411/511 Homework 1

Chimdi Chikezie

Winter 2022

Instructions

This assignment is due by 11:59 PM, January 14th on Canvas via Gradescope. **You should submit your assignment as a PDF which you can compile using the provide .Rmd (R Markdown) template.**

> Note: Create a PDF by either compiling a PDF directly (via LaTeX) or from a Word Doc. Do not submit a PDF of the HTML output as they're cumbersome and difficult to read.

Include your code in your solutions and indicate where the solutions for individual problems are located when uploading into Gradescope (Failure to do so will result in point deductions). You should also write complete, grammatically correct sentences for your solutions.

Once you've completed the assignment, and before you submit it to Gradescope, you should read the document to ensure that (1) The computed values show up in the document, (2) the document "looks nice" (i.e. doesn't have extraneous code/outputs and includes *just* the essentials), and (3) ensure the document is "easy" to read.

Goals:

1. Familiarize students with statistical vocabulary.
2. Begin learning how to decompose "scientific questions" into their statistical components.
3. Learn how to access example data sets from the `Sleuth3` R package.
4. Demonstrate that students can use functions and examples from the labs to compute solutions.
5. Demonstrate how random samples and statistical properties of estimates, like the sample mean, can be combined to learn something about an entire population.
6. Demonstrate that students can apply the Central Limit Theorem to address questions related to the sample mean.

Question 1 (6 points)

Identify the *population*, *variable*, and *parameter* of interest in the following scientific questions:

(a) (3 points) What is the average number of oranges on orange trees at Robertson's Farm?

- Population: Orange trees at Robertson's Farm.
- Variable: Number of oranges.
- Parameter: Mean (average)

(b) (3 points) What is the 20th percentile for weight of babies born in Oregon hospitals in 2018?

- Population: Babies born in Oregon hospitals in 2018.
- Variable: Weight
- Parameter: 20th percentile

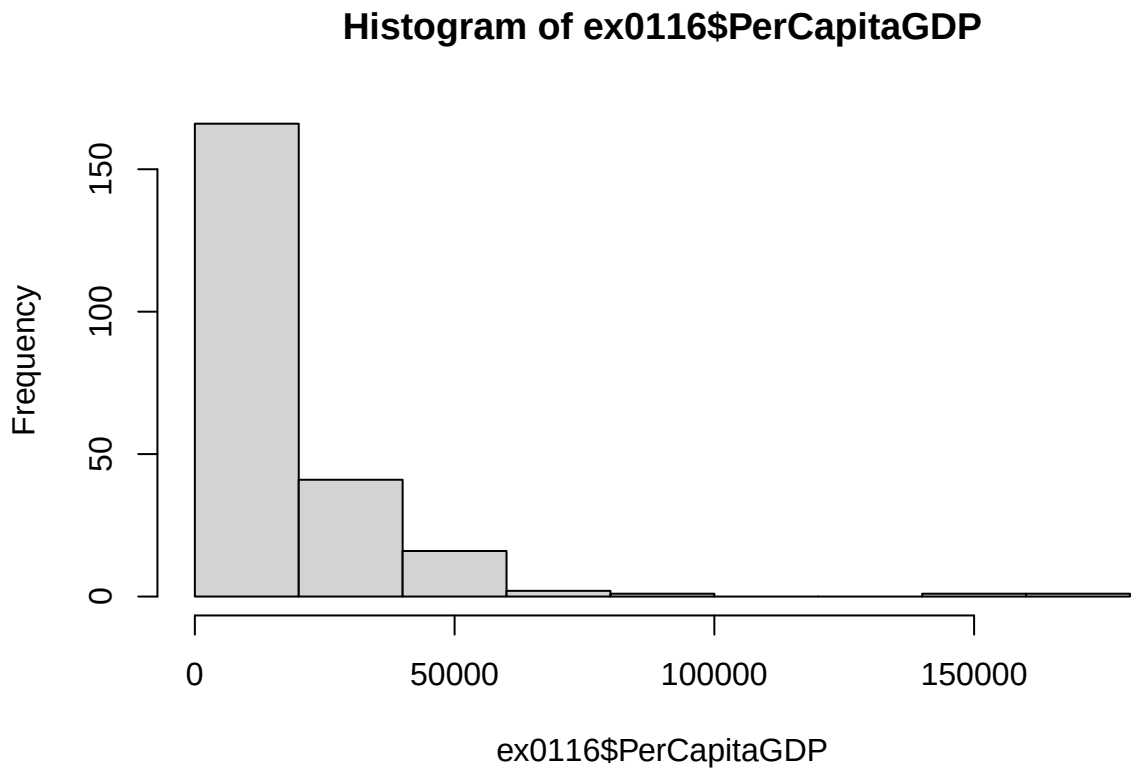
Question 2 (9 Points)

Use the following code to load the **ex0116** data set containing the gross domestic product (GDP) per capita for 228 countries in 2010.

```
library(Sleuth3)
data(ex0116)
```

(a) (2 points) Create a histogram of the population probability distribution. What do you notice about the shape of the distribution?

```
hist(ex0116$PerCapitaGDP)
```



The shape is not in any way close to a normal distribution and has a great variability. For the sample mean to be made to be approximately the population mean, it's sample size has to be increased till it starts looking like a normal distribution.

(b) (2 points) What is the *population* mean GDP per capita? What does this value describe about our population?

```
mean(ex0116$PerCapitaGDP)
```

```
## [1] 16017.54
```

Mean = 16017.54

The mean is not normal because different of great difference the mean that different samples have.

(c) (2 points) Use the following code to draw a random sample of size $n = 10$ from this population. What is the *sample* mean? What is the *sample* variance?

```
set.seed(411511)
samp1 <- sample(ex0116$PerCapitaGDP, size=10, replace=FALSE)
```

```
mean(samp1)
```

```
## [1] 10680
```

```
var(samp1)
```

```
## [1] 46179556
```

(d) (3 points) Repeat part (c) below to obtain a different random sample of size $n = 10$. What are the sample mean and sample variance from this sample? Why are these values different from those in part (c)?

> Note: Change the *number* within the `set.seed()` function to generate the same "random" sample each time. Otherwise, you'll get different random samples (Which in this case is not ideal).

```
set.seed(458511)
samp2 <- sample(ex0116$PerCapitaGDP, size=10, replace=FALSE)

mean(samp2)
```

```
## [1] 13550
```

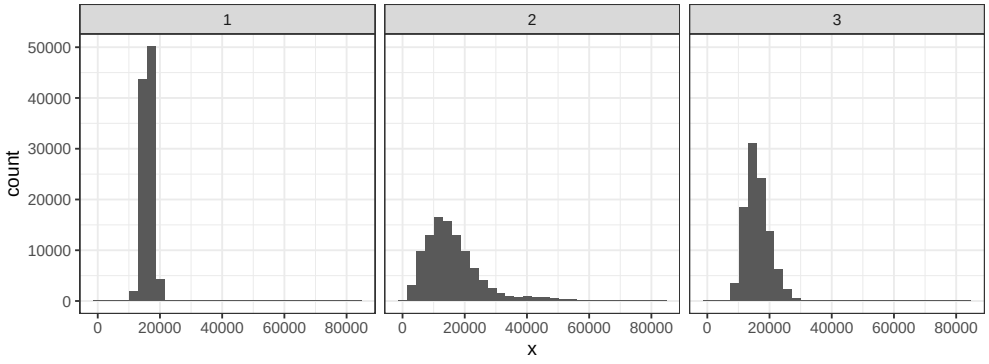
```
var(samp2)
```

```
## [1] 119736111
```

The sample mean and sample variance are different from the ones in part (C) because this sample has a set of 10 distinct numbers from those in part (C).

Question 3 (5 Points)

Consider the following three sampling distributions for the sample average from the **ex0116** GDP data used in question 2. One distribution is obtained from samples of size $n = 5$, one is obtained from samples of size $n = 25$, and one is obtained from samples of size $n = 100$.



(a) (3 points) Which histogram (1,2,3) corresponds to which sample size? Note: Base your answers on the histogram graphics and not the code itself.

- Histogram 1: Samples of size n = 100
- Histogram 2: Samples of size n = 5
- Histogram 3: Samples of size n = 25

(b) (2 points) How did you decide which histogram belongs to the different sample sizes?

I checked from their appearance. When sample size is increased, the shape of the sampling distribution will start looking more like a normal distribution because of less variability.

Question 4 (7 Points)

Recall that if a population distribution has mean μ and variance σ^2, the Central Limit Theorem says that for a sample of size n, the sample mean has an approximately Normal distribution with mean μ and variance σ^2/n.

(a) (3 points) Suppose a population has mean $\mu = 40$ and variance $\sigma^2 = 25$. What is the approximate distribution of the sample mean for samples of size $n = 20$?

$N(40, \frac{25}{20})$

(b) (4 points) Suppose a population has mean $\mu = 100$ and variance $\sigma^2 = 20$. For a sample of size $n = 10$, what is the approximate probability that the sample mean is less than 98?

Sampling distribution for the sample mean is each different sample we draw that has an average of which there are potential samples we could draw.

```
pnorm(98, mean = 100, sd = sqrt(20/10))
```

```
## [1] 0.0786496
```

The probability is 0.0786496