ST 411/511 Homework 1

Pavan Sai Nallagoni

Summer 2022

Instructions

This assignment is due by 11:59 PM, July 22nd 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: All the orange trees at Robertson's Farm
- Variable: Number of oranges on orange trees at Robertson's Farm
- Parameter: Mean (average)

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

- Population: All the babies born in Oregon hospitals in 2018
- Variable: The weight of babies born in Oregon hospitals in 2018
- Parameter: 20th percentile

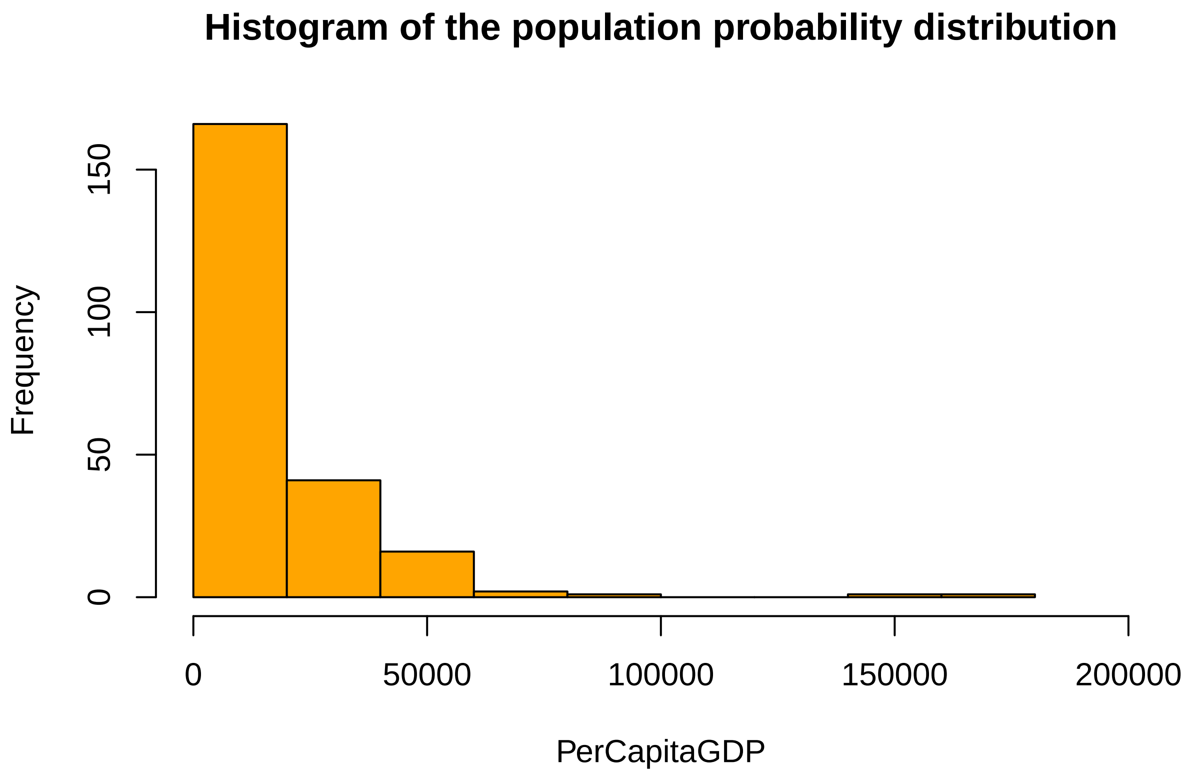
Question 2 (9 Points)

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?

```
# names(ex0116)
hist(ex0116$PerCapitaGDP, main = "Histogram of the population probability distribution",
     xlab = "PerCapitaGDP", xlim = c(0,200000),
     col = 'orange')
```



Shape of the distribution:

The shape of the distribution is a right-skewed histogram which shows us that there is a higher frequency of population(countries) with PerCapitaGDP ranging from (0, 2000](approximately) than the rest of the PerCapitaGDP (>20000).

As the per capita GDP increases, the population frequency keeps decreasing such that there is a less number of countries as we go right. Therefore, from the histogram, we can assume that the mean can be somewhere around 10000 since most numbers of countries were populated here (Mean can be in the interval (0 - 20000]).

(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

The mean tells us the average PerCapitaGDP of the given population(Countries) is 16017.54. It means that there's a probability of a higher frequency of population(Countries) whose PerCapitaGDP is equal to (or somewhere around) 16017.54.

(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(412432)
samp2 <- sample(ex0116$PerCapitaGDP, size=10, replace=FALSE)

mean(samp2)
```

```
## [1] 11860
```

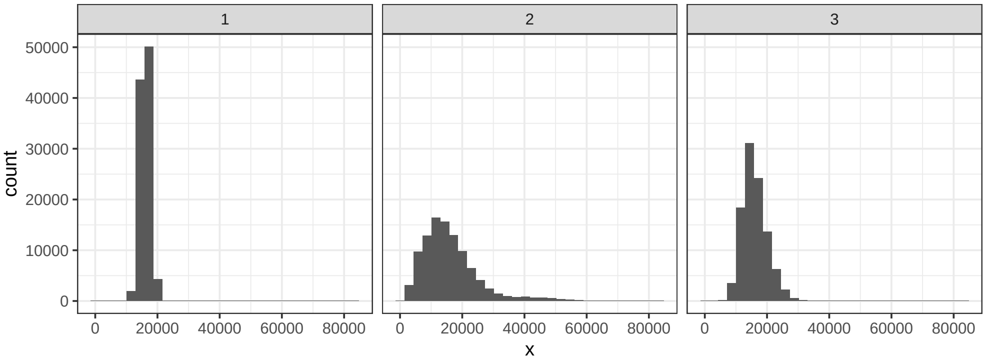
```
var(samp2)
```

```
## [1] 185349333
```

The sample mean is different from the mean in part(c). It is because in (d), we set the seed to a different value that generates new random numbers in the sample that are different from (c). It means we have a new set of numbers in the sample compared to (c). Thus, these new values in the sample will give us a new mean and variance.

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: The sample size for this histogram is 100
- Histogram 2: The sample size for this histogram is 5
- Histogram 3: The sample size for this histogram is 25

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

With small sample sizes, different samples can have very different mean. As a result, the variance increases, and the histogram looks wider.

However, as the number of samples increases, there will be much less variability and the histogram converges to the center (mean). It is because, as we increase the size of the samples, it will represent the population more closely such that the estimated mean will be closer to the true mean. Therefore, the larger the sample size, the lesser the width of the histogram (5 > 25 > 100).

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 $n = 20$?

The approximate distribution of the sample mean for samples of size $n = 20$ is approximately N(μ, σ^2/n) which is equal to N(40, 25/20) = N(40, 1.25).

(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?

The sampling distribution of the sampling mean for samples of size $n = 10$ is approximately N(μ, σ^2/n) which is equal to N(100, 20/10) = N(100,2).

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

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

Therefore, the approximate probability that the sample mean is less than 98 is pnorm(98, μ, sd) which is equal to 0.0786496.