ST 411/511 Homework 1

Due on January 15

Winter 2020

Instructions

This assignment is due by 11:59 PM, January 15, 2020 on Canvas via Gradescope. **You should submit your assignment as a PDF which you can compile (should you choose – recommended) using the provide .Rmd (R Markdown) template.** If you opt to not use R Markdown, please format your solutions in a similar manner as provided in this document. Include your code in your solutions and indicate where the solutions for individual problems are located when uploading into Gradescope. You should also use complete, gramatically correct sentences for your solutions.

Problems (25 points total)

Question 1

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: The population refers to all the orange trees at Robertson's farm.

Variable: The variable of interest is the number of oranges on a tree.

Parameter: Population mean is the parameter of interest.

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

Population: All the babies born in Oregon hospitals in 2018 makes up the population.

Variable: The weight of each baby is the variable.

Parameter: The population's 20th percentile of the weight is the parameter of interest.

Question 2

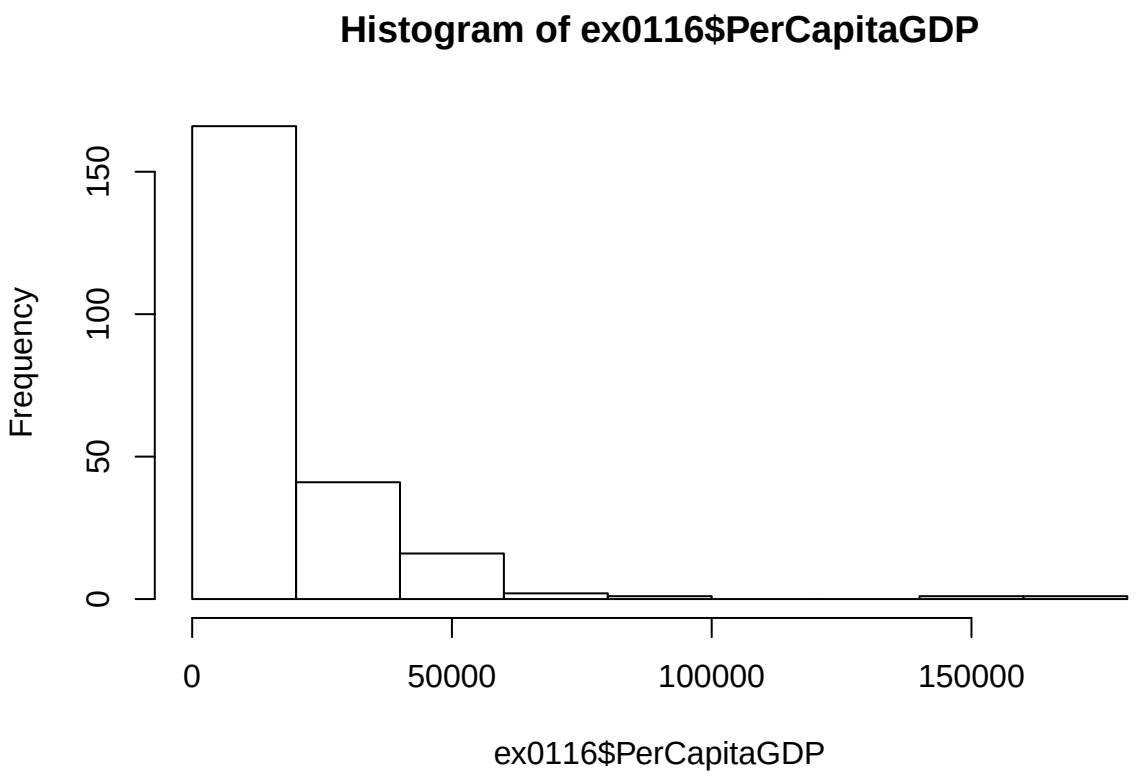
Use the following code to load the **ex0116** data set containing the gross domestic product (GDP) per capita for 228 countries in 2010.

```
#install.packages("Sleuth3",repos="http://R-Forge.R-project.org")

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?

```
#View(ex0116)
#?ex0116 --> A data frame with 228 observations on the following 3 variables.
hist(ex0116$PerCapitaGDP)
```



The distribution is not uniform and rather skewed. It looks like a half bell-curve. It shows that a majority countries have a very small Per Capita GDP.

(b) (1 point) What is the population mean GDP per capita?

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

```
## [1] 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?

```
samp1 <- sample(ex0116$PerCapitaGDP, size=10, replace=FALSE)
mean(samp1)
```

```
## [1] 25050
```

```
var(samp1)
```

```
## [1] 586218333
```

IMP : Note thar var and sd commands in R compute sample variance and sample standard distributions, not the population parameters. In general, we deal only with samples. What do you do if you have to calculate the population var and sd?

echo = False prevents the code from displaying in the Viewer pane and shows only the result such as mean or a histogram while True shows both code and the result.

(d) (3 points) Repeat part (c) 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)?

```
samp1 <- sample(ex0116$PerCapitaGDP, size=10, replace=FALSE)
mean(samp1)
```

```
## [1] 15790
```

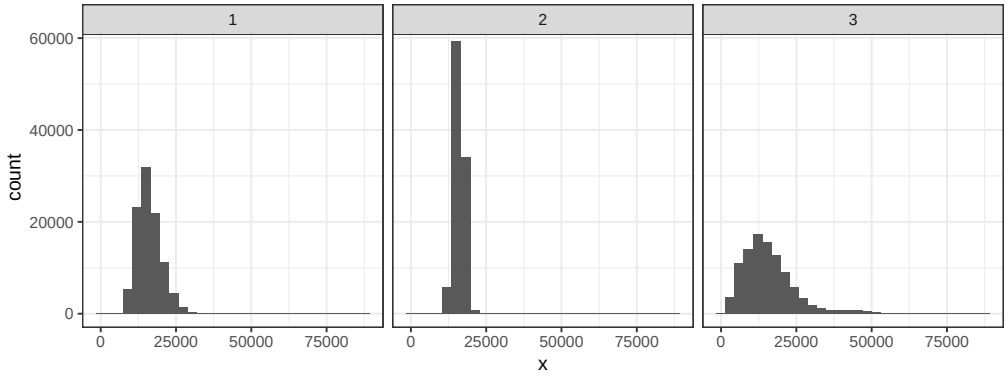
```
var(samp1)
```

```
## [1] 336267667
```

The two samples chosen are not identical (different number of samples) and thus the values affect the sample statistic differently.

Question 3

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) (2 points) Which histogram (1,2,3) corresponds to which sample size?

2nd plot is for sample size == 100 1st plot is for sample size == 25 2nd plot is for sample size == 5

(b) (2 points) How can you tell?

As n gets large, the sample's spread decreases. In the plots, I followed this trend.

Question 4

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 $n = 20$?

We can approximate it as having a normal distribution using the Central Limit Theorem. N(40, 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?

The distribution for sample mean is N(100,20/10) The approximate probability of samples in a normal distribution with mean 100 and stdev 2 below the value of 98 is given as pnorm(98, mean = 100, sd = 2)

```
pnorm(98, mean = 100, sd = 2)
```

```
## [1] 0.1586553
```

```
r = getOption("repos")
r["CRAN"] = "http://cran.us.r-project.org"
options(repos = r)
install.packages('tinytex')
```

```
## Installing package into '/home/being-aerys/R/x86_64-pc-linux-gnu-library/3.2'
## (as 'lib' is unspecified)
```

```
##
## The downloaded source packages are in
##   '/tmp/Rtmpm8gg7l/downloaded_packages'
```