Homework Assignment 2

Elikem Asudo Tsatsu Gale-Zoyiku

2023-02-10

library(readr)  
GhanaRegions <- read.delim("C:/Users/egale/OneDrive - Ashesi University/Desktop/Statistics with Probability/GhanaRegions.txt")  
StudentsData <- read.delim("C:/Users/egale/OneDrive - Ashesi University/Desktop/Statistics with Probability/StudentsData.txt")

##Question 1

#1a. The explanatory variable is gender. The response variable is the student’s preference for cafetaria on campus.

#1b.

t <- table(StudentsData$Cafeteria,StudentsData$Gender)  
colnames(t) <- c("Female","Male")  
rownames(t) <- c("Akornor","BigBen",'Munchies')  
t

##   
## Female Male  
## Akornor 12 27  
## BigBen 45 33  
## Munchies 49 52

#1c. Column percentages are used in this example because they show the proportion of students with a particular preference for a particular cafeteria within each gender. This can help us to see if there is a difference in preferences between male and female students, and if so, what that difference is. By using column percentages, we can easily compare the proportion of students who prefer each cafeteria within each gender, making it easier to see any patterns or trends.

#1d.

t\_prop <- round(100\*prop.table(t,2),1);t\_prop

##   
## Female Male  
## Akornor 11.3 24.1  
## BigBen 42.5 29.5  
## Munchies 46.2 46.4

#1e. Yes, they do.

#1f. 29.5 percent of male students prefer Big Ben, while 42.5 percent of female students prefer Big Ben as the go-to cafeteria on campus.

#1g. 24.1 percent of male students prefer to eat from Akornor, and 11.3 percent of female students prefer Akornor as their go-to cafeteria.

#1h. Yes, I think there is.

Almost an equal percentage of male and female students prefer to eat from Munchies (46.2% and 46.4% respectively). This implies that a person’s gender does not influence their prefernce of Munchies.

There is however an interesting observation in the other 2 cafeterias, where the number of females who prefer Big Ben is almost double the number of males who do, and the number of males who prefer Akornor is more than double the number of females who prefer there. This means that gender influences students’ preferences of those two cafeterias.

##Question 2

#2a. This is an observational study. An observational study is a type of study in which the researchers observe and measure the characteristics of individuals and their outcomes without manipulating or interfering with their behavior or environment. In this case, the researchers are observing the impact of the one district one factory project in Ghana’s populous regions without manipulating or changing the project.

#2b. The sampling approach being used is multistage sampling.

#2c. The sampling frame for the study is the list of all regions in Ghana with populations greater than 1,000,000 people.

#2d.

popRegions<-GhanaRegions[GhanaRegions$Population > 1000000, ]  
popRegions

## Region Population Capital  
## 2 Ashanti 5440463 Kumasi  
## 3 Bono\xa0 1208649 Sunyani  
## 4 Bono East 1203400 Techiman  
## 5 Central 2859821 Cape Coast  
## 6 Eastern 2925653 Koforidua  
## 7 Greater Accra 5455692 Accra  
## 9 Northern 2310939 Tamale  
## 12 Upper East 1301226 Bolgatanga  
## 14 Volta 1659040 Ho  
## 15 Western 2060585 Takoradi

x<-sample(1:length(popRegions$Population),5)  
sample<-popRegions[x, ];sample

## Region Population Capital  
## 6 Eastern 2925653 Koforidua  
## 2 Ashanti 5440463 Kumasi  
## 4 Bono East 1203400 Techiman  
## 12 Upper East 1301226 Bolgatanga  
## 7 Greater Accra 5455692 Accra

#2e.

1. ‘popRegions <- GhanaRegions[GhanaRegions$Population > 1000000, ]’:

In this line, we are creating a new data frame called popRegions which includes only the regions in Ghana that have a population greater than 1,000,000 people. This is done by using row and column indexing on the original data frame GhanaRegions and selecting only the rows where the population is greater than 1 million.

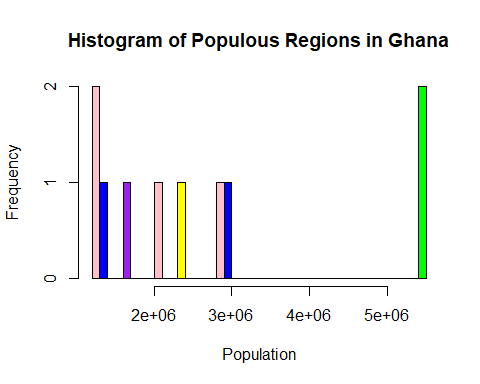
1. ‘popRegions’:

* This line is just printing the new data frame popRegions which includes only the populous regions in Ghana.

1. ‘’x <- sample(1:length(popRegions\\Population), 5)’’: In this line, we are randomly generating 5 index numbers which will be used to select 5 regions from the popRegions data frame. The sample() function is used to generate these index numbers, and the input 1:length(popRegions$Population) specifies that the index numbers should be selected from the range of the length of the population column in the popRegions data frame.
2. ‘sample <- popRegions[x, ]’: In this line, we are using the index numbers generated in step 3 to select 5 regions from the popRegions data frame and creating a new data frame called sample. The [x, ] part specifies that we are selecting the rows with the index numbers in x and all columns.
3. ‘sample’: This line is just printing the new data frame sample which includes 5 randomly selected regions from the popRegions data frame.

#2f.

hist(popRegions$Population, main="Histogram of Populous Regions in Ghana", xlab="Population", ylab="Frequency", breaks=50, col = c('pink','blue','green','yellow','purple','orange','navyblue','aquamarine'))



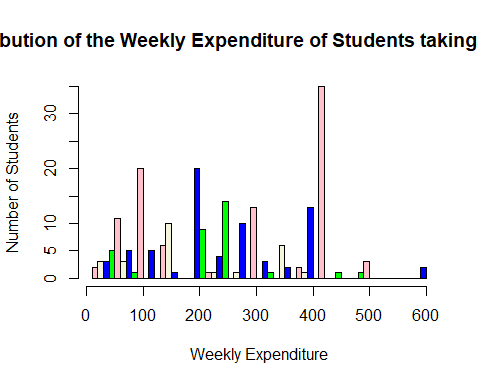
mean\_pop = mean(popRegions$Population)  
median\_pop = median(popRegions$Population)  
mode\_pop = sort(table(popRegions$Population),decreasing=TRUE)[1]  
range\_pop = max(popRegions$Population) - min(popRegions$Population)  
variance\_pop = var(popRegions$Population)  
Q1 <- quantile(popRegions$Population, probs = 0.25)  
Q3 <- quantile(popRegions$Population, probs = 0.75)  
IQR <- Q3 - Q1  
lower\_bound <- Q1 - 1.5 \* IQR  
upper\_bound <- Q3 + 1.5 \* IQR  
outliers <- popRegions$Population[popRegions$Population < lower\_bound | popRegions$Population > upper\_bound]

The histogram of the populous regions in Ghana shows that the data is not perfectly symmetrical, with a positive skew. The mean population of the regions is approximately 2.6 million, while the median is approximately 2.2 million. The mode is approximately 1.2 million. The populations range is approximately 4.25 million, and the variance is approximately 2.9 billion. The interquartile range of the populations’s distribution is approximately 1.5 million. The distribution has two outlier values, which correspond to Ashanti Region and Greater Accra Region.

This summary shows that the distribution of the populous regions in Ghana is not symmetrical, with a positive skew. The data is spread over a range of approximately 4.25 million, with a moderate amount of variation.

##Question 3

empirical\_rule\_check <- function(data) {  
   
 mean <- mean(data)  
 stddev <- sd(data)  
   
 # Calculate the interval for 68% of the data around the mean  
 interval\_68 <- c(mean - stddev, mean + stddev)  
   
 # Calculate the interval for 95% of the data around the mean  
 interval\_95 <- c(mean - 2 \* stddev, mean + 2 \* stddev)  
   
 # Calculate the interval for 99.7% of the data around the mean  
 interval\_997 <- c(mean - 3 \* stddev, mean + 3 \* stddev)  
   
 # Calculate the percentage of data that falls within each interval  
 perc\_68 <- sum(data >= interval\_68[1] & data <= interval\_68[2]) / length(data)  
 perc\_95 <- sum(data >= interval\_95[1] & data <= interval\_95[2]) / length(data)  
 perc\_997 <- sum(data >= interval\_997[1] & data <= interval\_997[2]) / length(data)  
   
 # Check if the distribution follows the empirical rule  
 if (perc\_68 >= 0.68 & perc\_95 >= 0.95 & perc\_997 >= 0.997) {  
 return("The distribution follows the empirical rule.")  
 } else {  
 return("The distribution does not follow the empirical rule.")  
 }  
   
}  
hist(StudentsData$Expenditure, col=c('pink','beige','blue','green'), main='Distribution of the Weekly Expenditure of Students taking Statistics', xlab='Weekly Expenditure', ylab='Number of Students',breaks=50)



empirical\_rule\_check(StudentsData$Expenditure)

## [1] "The distribution does not follow the empirical rule."

The distribution of the weekly expenditure of students taking Statistics is an asymmetric distribution. The distribution is left skewed. Using the empirical rule proves that this distribution is not symmetrical. This is likely because there are outliers present in the distribution.

#Question 3e.

observation <- 100  
percentage <- (sum(StudentsData$Expenditure < observation) / length(StudentsData$Expenditure))\*100  
percentage

## [1] 15.59633

It would be unusual for a Statistics student to spend less than 100 cedis on food weekly. Only 15.60% of students in the Statistics class spend below 100 cedis. The probability that a Statistics student would spend below 100 cedis is then 0.156, making it unlikely.

#Question 4

#Question 4a. The sampling method used in this study is systematic sampling. The team decided to choose the first customer that arrived at the bank after 11:00 am and then choose every fifth customer to arrive until 1:00 pm. This is a systematic method of sampling as the customers are selected at regular intervals.

#Question 4b.

# Waiting times for SilverStar Tower Branch  
silverstar\_branch <- c(9.4, 8.3, 9.4, 6.6, 3.0, 7.8, 10.5, 9.8, 9.8, 4.7, 5.9)  
mean(silverstar\_branch)

## [1] 7.745455

sd(silverstar\_branch)

## [1] 2.415104

# Waiting times for Accra Mall Branch  
accra\_mall\_branch <- c(8.4, 6.6, 9.0, 5.6, 2.2, 5.3, 6.6, 7.3, 7.5, 7.2, 6.0, 7.5, 8.9, 4.5)  
mean(accra\_mall\_branch)

## [1] 6.614286

sd(accra\_mall\_branch)

## [1] 1.830871

# Waiting times for Osu Branch  
osu\_branch <- c(7.9, 7.0, 7.1, 7.4, 6.7, 6.7, 7.3, 7.6, 6.0, 7.3, 7.2, 6.8)  
mean(osu\_branch)

## [1] 7.083333

sd(osu\_branch)

## [1] 0.4951278

The mean and standard deviation of the waiting times for the SilverStar Tower Branch are 7.745455 and 2.415104 minutes respectively. The mean and standard deviation of the waiting times for the Accra Mall Branch are 6.614286 and 1.830871 minutes respectively. The mean and standard deviation of the waiting times for the Osu Branch are 7.083333 and 0.4951278 minutes respectively.

Based on the means, the best performing branch is the Osu Branch with an average waiting time of 7.083333 minutes. The worst performing branch is the SilverStar Tower Branch with the longest average waiting time of 7.745455 minutes. The standard deviation measures the spread of the data. The smaller the standard deviation, the more the data points tend to cluster around the mean.

In this case, the Osu Branch has the smallest standard deviation of 0.4951278 minutes indicating that the waiting times of its customers tend to cluster around the mean, implying that the branch’s service is more consistent, than the Accra mall branch which has a lower average waiting time, but greater standard deviation.

SilverStar towers is the worst performing branch because their average waiting time and the standard deviation of these times are the highest, implying their service is inconsistent.

##Question 4c. #Question 4.c.i.

# Salaries of workers in SilverStar Tower Branch  
silverstar\_salaries <- c(200, 350, 220, 180, 500, 320, 370, 350, 300, 310)  
silverstar\_mean <-mean(silverstar\_salaries)  
silverstar\_mean <- sd(silverstar\_salaries)  
  
# Salaries of workers in Accra Mall Branch  
accra\_mall\_salaries <- c(1200, 2000, 3500, 3000, 2700, 2450, 1900, 2100, 2200, 2600)  
accra\_mall\_mean <-mean(accra\_mall\_salaries)  
accra\_mall\_sd <-sd(accra\_mall\_salaries)

#Question 4.c.ii.

silverstar\_sd <- sd(silverstar\_salaries)  
accra\_mall\_sd <- sd(accra\_mall\_salaries)  
silverstar\_cv <- (silverstar\_sd / silverstar\_mean)\*100  
silverstar\_cv

## [1] 100

accra\_mall\_cv <- (accra\_mall\_sd / accra\_mall\_mean)\*100  
accra\_mall\_cv

## [1] 27.00197

The salaries of the two branches could not be compared because they are of different scales and currencies, so the coefficient of variation was calculated.

Accra Mall Branch has a co-efficient of variation of approximately 27% while Silver Tower Branch has a coefficient of variation of approximately 30.5%. Therefore Accra Mall Branch’s have more equitable salaries than those of SilverStar Towers as it has a smaller coefficient of variation of salaries implying it has less variation in the salaries of workers than those of SilverStar Tower Branch.

##Question 5

#Question 5a. The study was an observational study as it involved observing and measuring the outcomes of four different groups of Ghanaians who regularly consume different amounts of kenkey. Furthermore, the study was a prospective observational study as the participants had not developed the outcome of interest as at when they were selected and the study was conducted over a three-month period, during which the participants were instructed to continue with their normal kenkey consumption habits. The study was not an experiment as there was no intervention or manipulation of the independent variable (kenkey consumption).

#Question 5b. This is an experiment because the researchers are actively manipulating the independent variable (amounts of nitrogen and phosphorus applied to the plots) to determine its effect on the dependent variable (the yield of corn on the plot). The researchers assigned random numbers to each plot because the researchers are aware of the intervention being applied to each plot so the experiment is not blind.