Homework Assignment 2

David Abeiku Saah

2023-02-12

library(readr)  
  
StudentsData <- read\_delim("StudentsData.txt",   
 delim = "\t", escape\_double = FALSE,   
 trim\_ws = TRUE)  
  
GhanaRegions <- read\_delim("GhanaRegions.txt",   
 delim = "\t", escape\_double = FALSE,   
 trim\_ws = TRUE)

### Question 1

1. Gender is the explanatory variable while the response variable is Cafeteria. This is because we are trying to check if a student’s gender determines the cafeteria they prefer to eat from.
2. A Two Way Table for Cafeteria and Gender

tbl = table(StudentsData$Cafeteria, StudentsData$Gender); tbl

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

colnames(tbl) = c("Female", "Male")  
rownames(tbl) = c("Akornor", "BigBen", "Munchies")  
tbl

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

1. Gender and Cafeteria are both categorical variables. Using raw data counts does not give a precise information about the relationship between the variables. Conditional percentages allow us to compare the distribution with a numerical summary restricted to each value of the explanatory variable. It gives the “condition” for which the explanatory variable (Gender) may “affect” the response variable (Cafeteria). Since Gender (explanatory variable) is in the column of the two-way table, we use conditional column percentages.
2. A Two Way Table for Cafeteria and Gender with conditional percentages.

tbl\_prop = round(100 \* prop.table(tbl, 2), 1)  
tbl\_prop

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

1. Column 1 = 11.3 + 42.5 + 46.2 = 100 Column 2 = 24.1 + 29.5 + 46.4 = 100 Yes, each column adds up to 100.
2. Percentage of male students who prefer BigBen: 29.5% Percentage of female students who prefer BigBen: 42.5%
3. Percentage of male students who prefer Akornor: 24.1% Percentage of female students who prefer Akornor: 11.3%
4. From the two-way table of conditional percentages, the majority of males (46.4%) and females (46.2%) prefer Munchies. Consequently, the least preferred cafeteria for both genders (24.1% for males and 11.3% for females) is Akornor. Since there is a trend of both genders preferring a particular cafeteria in similar proportions, I do not think there is a gender effect on student’s preferences for the cafeterias on campus.

### Question 2

1. This is an observational study because the researchers surveyed the companies as they went about their normal activities. Also, the researchers do not assign the factories to be used in the study. Hence, it cannot be considered as an experiment. It is an observational study.
2. The sampling method being used is cluster sampling.
3. All the regions in Ghana with populations greater than 1,000,000.
4. Randomly selecting the 5 regions for the study

pop\_regions = GhanaRegions[GhanaRegions$Population > 1000000, ]  
pop\_regions

## # A tibble: 10 × 3  
## Region Population Capital   
## <chr> <dbl> <chr>   
## 1 "Ashanti" 5440463 Kumasi   
## 2 "Bono\xa0" 1208649 Sunyani   
## 3 "Bono East" 1203400 Techiman   
## 4 "Central" 2859821 Cape Coast  
## 5 "Eastern" 2925653 Koforidua   
## 6 "Greater Accra" 5455692 Accra   
## 7 "Northern" 2310939 Tamale   
## 8 "Upper East" 1301226 Bolgatanga  
## 9 "Volta" 1659040 Ho   
## 10 "Western" 2060585 Takoradi

x = sample(1:length(pop\_regions$Population), 5)  
sample\_data = pop\_regions[x, ];sample\_data

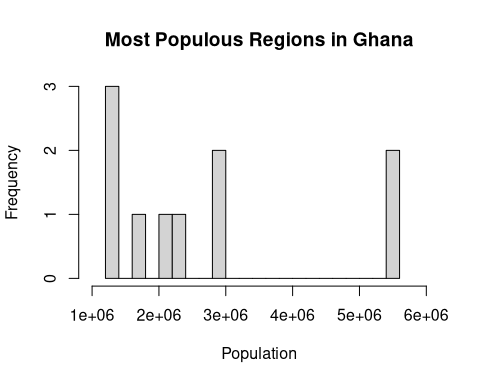
## # A tibble: 5 × 3  
## Region Population Capital   
## <chr> <dbl> <chr>   
## 1 Western 2060585 Takoradi   
## 2 Volta 1659040 Ho   
## 3 Upper East 1301226 Bolgatanga  
## 4 Northern 2310939 Tamale   
## 5 Bono East 1203400 Techiman

1. Explaining the code

* Line 1: Get all the regions in Ghana whose population is greater than 1,000,000 and store it as a vector.
* Line 2: Display pop\_regions (regions with population greater than 1,000,000).
* Line 3: Randomly select 5 numbers between 1 and number of regions greater than 1,000,000 and store it in variable x.
* Line 4: Use the index values stored in x to get the data for the randomly selected regions and store them in the sample\_data variable.

1. Summarising the distribution.

hist(pop\_regions$Population, breaks = 30, xlim = c(1000000, 6000000), main = "Most Populous Regions in Ghana", xlab = "Population");



summary(pop\_regions$Population);

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 1203400 1390680 2185762 2642547 2909195 5455692

IQR(pop\_regions$Population);

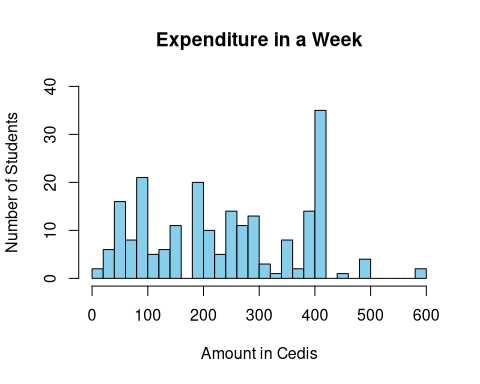
## [1] 1518516

The shape of the distribution is right-skewed. The measure of center is the median and the measure of spread is the interquartile range (IQR) because of the skewness of the distribution.

The bulk of the most populous regions are below the 3rd quartile (2,909,195 people). There is an outlier that shows that the most populated region is outside the interquartile range of the distribution.

### Question 3

hist(StudentsData$Expenditure, breaks = 40, col = "skyblue", ylim = c(0, 40), xlim = c(0, 600), main="Expenditure in a Week", xlab = "Amount in Cedis", ylab = "Number of Students");



# Find the data summary  
summary(StudentsData$Expenditure);

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 10.0 120.0 240.0 245.1 396.2 600.0

# Calculations for expenditure.  
# Expenditure is right skewed since mean > median.  
median(StudentsData$Expenditure);

## [1] 240

IQR(StudentsData$Expenditure);

## [1] 276.25

The distribution is right skewed because the mean > median. Since the distribution is not symmetric, it does not follow the empirical rule.

1. Proportion of students below the mean

# below\_mean = pnorm(mean(StudentsData$Expenditure), mean = mean(StudentsData$Expenditure), sd = sd(StudentsData$Expenditure))  
  
below\_mean = StudentsData[StudentsData$Expenditure < mean(StudentsData$Expenditure), ];  
length(below\_mean$Expenditure)/length(StudentsData$Expenditure)

## [1] 0.5045872

1. Proportion within one standard deviation

pos\_sd1 = mean(StudentsData$Expenditure) + sd(StudentsData$Expenditure)  
neg\_sd1 = mean(StudentsData$Expenditure) - sd(StudentsData$Expenditure)  
  
bet\_1\_sd = StudentsData[(StudentsData$Expenditure < pos\_sd1) & (StudentsData$Expenditure > neg\_sd1), ];  
length(bet\_1\_sd$Expenditure)/length(StudentsData$Expenditure)

## [1] 0.5

1. Proportion within two standard deviation

pos\_sd2 = mean(StudentsData$Expenditure) + 2\*sd(StudentsData$Expenditure)  
neg\_sd2 = mean(StudentsData$Expenditure) - 2\*sd(StudentsData$Expenditure)  
  
bet\_2\_sd = StudentsData[(StudentsData$Expenditure < pos\_sd2) & (StudentsData$Expenditure > neg\_sd2), ];  
  
length(bet\_2\_sd$Expenditure)/length(StudentsData$Expenditure)

## [1] 0.9908257

1. Proportion within three standard deviation

pos\_sd3 = mean(StudentsData$Expenditure) + 3\*sd(StudentsData$Expenditure)  
neg\_sd3 = mean(StudentsData$Expenditure) - 3\*sd(StudentsData$Expenditure)  
  
bet\_3\_sd = StudentsData[(StudentsData$Expenditure < pos\_sd3) & (StudentsData$Expenditure > neg\_sd3), ];  
  
length(bet\_3\_sd$Expenditure)/length(StudentsData$Expenditure)

## [1] 1

The distribution is does not follow the empirical rule because the proportions deviate from the rule. For the first standard deviation it is 0.5 opposed to 0.68. Within the two standard deviation, the proportion is 0.99 as opposed to 0.95. Within three standard deviation, the proportion is 1, as opposed to 0.997.

1. Would it be unusual for a student to spend less than 100 cedis?

below\_100 = StudentsData[StudentsData$Expenditure < 100, ];  
length(below\_100$Expenditure)/length(StudentsData$Expenditure)

## [1] 0.1559633

Since the proportion of students less than 100 cedis is 0.15, it implies that it is unusual for a statistics student to spend on average, less than 100 cedis on food weekly.

### Question 4

1. Convenience sampling.
2. Determining customer service performance

silvertower = c(9.4, 8.3, 9.4, 6.6, 3.0, 7.8, 10.5, 9.8, 9.8, 4.7, 5.9);  
accra\_mall = 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);  
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\_silvertower = mean(silvertower); mean\_silvertower

## [1] 7.745455

mean\_accra\_mall = mean(accra\_mall); mean\_accra\_mall

## [1] 6.614286

mean\_osu\_branch = mean(osu\_branch); mean\_osu\_branch

## [1] 7.083333

sd\_silvertower = sd(silvertower); sd\_silvertower

## [1] 2.415104

sd\_accra\_mall = sd(accra\_mall); sd\_accra\_mall

## [1] 1.830871

sd\_osu\_branch = sd(osu\_branch); sd\_osu\_branch

## [1] 0.4951278

The branch with the best performance would be Accra Mall. It has the lowest mean waiting time of 6.614 minutes compared to 7.745 minutes (for silvertower branch) and 7.083 minutes (for Osu branch).

The branch with the worst performance is Silverstar Tower. It has the highest mean waiting time of 7.745 minutes compared to 6.614 minutes (for Accra Mall branch) and 7.083 minutes (for Osu branch). Also, the mean waiting time for Silverstar Tower is consistent because it has a minimum spread of 0.495.

### Question 4c

silvertower\_salary = c(200, 350, 220, 180, 500, 320, 370, 350, 300, 310)  
accra\_mall\_salary = c(1200, 2000, 3500, 3000, 2700, 2450, 1900, 2100, 2200, 2600)  
  
# convert silvertower salary to cedis  
# 1 dollar = 12.2 cedis  
silvertower\_salary\_cedis = 12.2 \* silvertower\_salary

1. Calculating the mean and standard deviation

mean\_st = mean(silvertower\_salary\_cedis); mean\_st

## [1] 3782

mean\_am = mean(accra\_mall\_salary); mean\_am

## [1] 2365

sd\_st = sd(silvertower\_salary\_cedis); sd\_st

## [1] 1151.664

sd\_am = sd(accra\_mall\_salary); sd\_am

## [1] 638.5966

1. From the sample, the average salary of Silverstar Tower workers is GHS 3782 while that of Accra Mall is GHS 2365. The standard deviation of the salary of Silverstar Tower workers is GHS 1151.664. This means that it has an inconsistent salary pay and an existence of possible outliers. On the other hand, the standard deviation of Accra Mall workers’ salary is GHS 638.5966. This means that their salary is more consistent compared to that of Silverstar employees.

### Question 5

1. The study is an experiment because the researchers determined the explanatory variable (the consumption of kenkey) and assigned them to different categories of people. The experiment is not blind because both researchers and participants were aware of the treatment of the experiment.
2. The study is an experiment because the researchers controlled the amount of nitrogen and phosphorus each plot got. The experiment is not blind because the researchers were aware of the treatment.