Week 2 Homework: Upload to Canvas by Sunday Feb. 12th by 11:59 pm

Sampling & Designing Studies

In this assignment you will practice skills in sub-setting data and random sampling using R, answer questions about sampling, lurking variables, and study designs, as well as review skills in ***examining distributions*** and ***examining relationships***.

Preparation Tasks

1. Before beginning Homework 2 you should have completed the following OLI modules:

   Module 6: Sampling
   Module 7: Designing Studies

2. From the course drive, download the following datafiles:

   - GhanaRegions.txt
   - StudentsData.txt

RMarkdown Instructions

1. Set the working directory in RStudio to the folder where you will save RMarkdown file. Any datasets needed for the assignments should also be located in this folder. There are a number of ways to set the working directory. One way is to click Session→Set Working Directory→Choose Working Directory from the RStudio menu, and then browse the folder you will use.
2. Create a new RMarkdown file. The header should be similar to the one below:

   ```
   ---
   title: "Homework Assignment 1"
   author: "Eric Ocran"
   date: "14/01/2022"
   output: word document
   ---
   ```

   Delete everything below the metadata header (it's not needed) and then save your RMarkdown file. Be sure to save it to a folder you have set for the working directory.

3. Create a setup code chunk just below the header similar to the one below. The read.delim() code can be generated by importing the data from the environment pane and copying and pasting the code generated in the console into your code chunk (remember to delete the console prompt, + signs, and filepath).

   ```
   ```{r setup, message = FALSE}
   library(readr)

   Data<-read.delim("C:\\Users\\user\\Dropbox\\My PC (DESKTOP-
   0NQIKS4)\\Desktop\\Ashesi University\\Ashesi 2021-2022\\Semester
   2\\Statistics\\Document\\studentData.txt")

   GhanaCities<-read.delim("C:\\Users\\user\\Dropbox\\My PC (DESKTOP-
   0NQIKS4)\\Desktop\\Ashesi University\\Ashesi 2020-2021\\2020 statistics\\HW
   Assignments\\GhanaCities.txt")
   ```
   ```

Logical Operators in R

The following table of logical operators in R will be a helpful reference in this assignment.

equal to	greater than or equal to	less than or equal to	greater than	less than	not equal to	x or y	x and y	not
==	>=	<=	>	<	!=	x | y	x & y	!

Homework Instructions

1. Recall that the visual diagram for exploring the relationship between two categorical variables (C → C) is a two-way table of conditional percentages. In this question, we will explore the relationship between a statistics student's gender and their preference for the mode of teaching using a two-way table of conditional percentages. Basically, we are asking the question "is there a gender effect on student's preferences for the cafeterias on campus?"
    a. What is the explanatory variable? Response variable?
    b. In this example, the explanatory categories will be the columns and the categories of the response variable will be the rows of the two-way table. Use the sample code below to create a two-way table of counts with appropriate row and column labels.

    ```
    t <- table(data$rowVar, data$colVar); t
    colnames(t) <- c("lbl1","lbl2", …)
    rownames(t) <- c("lbl1", "lbl2", …)
    t
    ```

    c. Review page 46 on OLI. Explain why, in this example, we should use column percentages.
    d. Use the sample code below to create a two-way table of conditional percentages using column percentages.

    ```
    t_prop <- round(100*prop.table(t,2),1)
    t_prop
    ```

    e. Do the percentages in each column add up to 100%?
    f. What percentage of male students would prefer Big Ben as the go to cafeteria on campus? Female students?
    g. What percentage of male students would prefer Akornor as the go to cafeteria on campus? Female students?
    h. Based on your examination of the two-way table of conditional percentages, do you think there is a gender effect on student's preferences for the cafeterias on campus? Explain your reasoning.

2. You have been hired as a Research Assistant on a Ministry of Trade project looking at the impact of one district one factory project in Ghana's populous regions. The Lead Researcher just explained to you the sampling approach: "I will give you the list of all regions in Ghana. Randomly select 5 regions with populations greater than 1,000,000 people. Every company registered under the one district one factory project in each selected region will be surveyed."
    a. Is this an observational study or an experiment? Explain.
    b. What type of sampling is being used?
    c. What is the sampling frame for the study?
    d. Use R to randomly select 5 cities for the study.

The approach we will use is to create a new vector of populous regions with populations greater than 1,000,000 people by using [*row, column*] indexing of the original data frame, and then sample 5 regions from the new vector using randomly selected index numbers. Use the code below as a guide.

```
popRegions<-GhanaRegions[GhanaRegions$Population > 1000000, ]
popRegions
x<-sample(1:length(popRegions$Population),5)
sample<-popRegions[x, ];sample
```

e. Explain in your own words what is happening in each line of code above, or if you used a different sequence of steps (there are many ways to subset in R!) then explain the steps in your own code.

f. As part of this process, you are expected to summarize the distribution of the populous regions in Ghana. Create a histogram with proper titles and labels (test different values for the number of breaks so that the shape, center and variation are easily observed), calculate appropriate measures of center and spread (is the distribution symmetric or skew?) and then give a brief written summary of the distribution (remember, a good summary gives numerical values for center and spread with units, the data context, and addresses shape, center and variation, and identifies outliers if there are any).

3. Is the distribution of the average weekly expenditure on food that statistics students reported roughly symmetrical? And if so, does the distribution follow the empirical rule? Load the *studentData.txt* dataset from last week's assignment. Use a histogram to determine if the distribution is symmetrical and if so, calculate the mean and standard deviation of the average weekly expenditure on food reported by statistics students. To test the empirical rule, (a) determine the proportion of students below the mean, (b) the proportion within one standard deviation of the mean, (c) within two standard deviations of the mean, and (d) between three standard deviations of the mean. What can you conclude? (e) Would it be unusual for a statistics student to spend on average, less than 100 cedis on food weekly? Explain.

4. The Quality Assurance department at Ecobank was tasked with assessing the quality of customer service provided by various branches in Accra. 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. Among other things, the team measured the waiting time in minutes before each selected customer was served. The waiting times in minutes for three branches are given below.

Silverstar Tower Branch Waiting Times	9.4 8.3 9.4 6.6 3.0 7.8 10.5 9.8 9.8 4.7 5.9
Accra Mall Branch Waiting Times	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 Waiting Times	7.9 7.0 7.1 7.4 6.7 6.7 7.3 7.6 6.0 7.3 7.2 6.8

a. What type of sampling was used for the study?

b. Summarize the customer service performance for the three Ecobank branches based on the mean and standard deviation in customer waiting times for each branch. What branch had the best performance? Worst performance? Explain your reasoning. [Remember, you can use the c() function to create a list of numbers in R and assign it to a variable.]

c. The salaries of 10 randomly selected workers from 2 of the branches are displayed in the table below.

Silverstar Tower Branch (in dollars $)	200, 350, 220, 180, 500, 320, 370, 350, 300, 310
Accra Mall Branch (in cedis GH₵)	1200, 2000, 3500, 3000, 2700, 2450, 1900, 2100, 2200, 2600

i. Compute the mean and standard deviation for the salary of workers in the two branches.

ii. Based on the results in (i), what conclusion can you draw concerning the salary of Ecobank workers at the two branches?

5. Determine whether the given description corresponds to an observational study or an experiment. If observational, is the study prospective or retrospective? If an experiment, is it blind? Double blind? Not blind? Give a reason in each case.

a. A study was conducted to determine the health effects of differing amounts of kenkey in the diet for Ghanaians diagnosed with diabetes. A group of 150 Ghanaians diagnosed with diabetes who regularly eat kenkey more than five times a week, a group of 150 that regularly eat kenkey between 3 and 5 times per week, a group of 150 who regularly eat kenkey 1 or 2 times per week and a group of 150 that never eat kenkey were selected for the study. At the beginning of the study, all 600 participants were given a thorough physical exam and the severity of their diabetic symptoms measured. Participants were instructed to continue with the same amount of kenkey they regularly eat for the next three months. At the end of the study, the group that regularly ate kenkey between 3 and 5 times per week had lower diabetic symptoms on average, than the other three groups.

b. Researchers wanted to determine the combination of nitrogen and phosphorus that produces the maximum amount of corn on a plot. Amounts (in pounds) of phosphorous and nitrogen were applied to 9 plots according to the following table, where plots were given numbers and then randomly assigned to one of the combinations:

plots	7	1	3	4	9	8	6	2	5
Phosphorus	10	10	10	20	20	20	30	30	30
Nitrogen	40	50	60	40	50	60	40	50	60

Yields in plot 3 were the greatest overall.

ASSIGNMENT SUBMISSION: Knit the final version of your completed assignment to make sure it renders properly in MS Word. **Upload the RMarkdown .RMD file to Canvas.**