Homework 1

Instructions: Use the provided Quarto template to complete your assignment. Submit the PDF rendered by your Quarto document on Gradescope. You must show your work/justify your answers to receive credit.

1. Use the following information to answer each question below: $P(A) = 0.2$, $P(B) = 0.4$, $P(A \cap B) = 0.08$

   a. What is $P(A|B)$?

   b. What is $P(A \cup B)$?

   c. Are A and B mutually exclusive?

   d. Are A and B independent?

2. In 2023, 20% of a local bank's credit card holders experienced fraud that forced them to cancel their cards. The bank expects to bring in 200 new customers in 2024. They want to know how many new cards they will have to cancel, assuming the probability of fraud is consistent for both years.

   a. Which probability distribution best fits this situation?

   b. How many customers should they expect to experience fraud?

   c. What is the probability that 50 customers will experience fraud?

   d. What is the probability that no more than 50 customers will experience fraud?

3. *All about the samples*: In this exercise, you will conduct a simulation in R to evaluate the effect of sample size and sampling schemes on the sampling distribution of the mean.

We will use the dataset provided below to conduct the simulation. The dataset contains 100000 observations, so we will treat this as our population of interest. The dataset has 3 variables: 1 continuous variable (Y) and 2 categorical variables (X1 and X2).

```
population <- read.csv("https://raw.githubusercontent.com/anlane611/datasets/main/population
```

First, let's understand our population. Our main variable of interest will be Y.

a. What is the mean of Y?

b. What is the mean of Y for each level of X1? How many observations are in each level of X1?

c. What is the mean of Y for each level of X2? How many observations are in each level of X2?

Now, explore the sampling distribution of the sample mean of Y under simple random sampling.

d. Use the skeleton code below to simulate the process of collecting 1000 simple random samples, each of size 10, from the population. Generate a histogram of the 1000 sample means and describe what you observe. Does the histogram show an approximately normal distribution? What is its mean and how does it compare to the population mean?

```
SRSmeans <- data.frame(means=NA) #create an empty dataframe to store the sample means
for(i in 1:1000){
  set.seed(i) #ensure that we have a different sample each time
  SRS.sample <- __________ |> slice(sample(1:________,size=__))
  SRSmeans[i,1] <- SRS.sample |> summarise(______(Y))
}

ggplot(SRSmeans, aes(x=means)) +
  geom_histogram()+
  labs(title="_____________")
```

e. Simulate the process of collecting 1000 simple random samples, but this time each should be of size 100. Generate a histogram of the sample means. What is the difference between this histogram and the histogram you generated in part d?

Next, explore the sampling distribution of the sample mean of Y under cluster sampling.

f. Simulate the process of collecting 1000 cluster samples, where each sample is comprised of 2 randomly-selected clusters from X1. Generate a histogram of the samples means and describe what you observe. What is the mean of the sampling distribution? Is the sampling distribution approximately normal?

Finally, explore the sampling distribution of the sample mean of Y under stratified sampling.

g. Simulate the process of collecting 1000 stratified samples, where each sample is comprised of a random sample of size 100 drawn from each stratum, and X1 is the strata variable. Generate a histogram of the sample means and describe what you observe. What is the mean of the sampling distribution? Is the sampling distribution approximately normal?

h. Simulate the process of collecting 1000 stratified samples, where each sample is comprised of a random sample of size 100 drawn from each stratum, and X2 is the strata variable. Generate a histogram of the sample means and describe what you observe. What is the mean of the sampling distribution? Is the sampling distribution approximately normal?

i. Compare and contrast the results from g and h. Consider the results from parts b and c. What can you conclude about the sampling distribution of the sample mean under stratified sampling?

Bonus (5pts): Considering your results from parts f-i, what can you say about the sampling distribution under multistage sampling? You do not need to run the simulation, but your answer should be well-substantiated.