Assignment 1

Data 543 Data Collection

2023 Term 1

Submit well-organized output (html/pdf) of an RMarkdown/knitr document answering the following questions. No late assignments will be accepted as I will make the solutions available immediately after the due date.

Note: The quiz will be completely closed book, so although we use R as our calculator, you will be expected to do simple calculations using a calculator in the quiz. Any calculations that require software (mean, standard deviation, etc.) will be given on the quiz.

1. You will be able to do up to and including question 2 part b subpart ii after Lab 1
2. You will be able to do the remainder of the questions after Lab 2

Due Date: Sunday, November 26, 11:59pm

Total marks: 57 marks

In this assignment I use SRS to denote Simple Random Sampling Without Replacement (SRSWOR).

1. Suppose DATA 543 has four sections taught by 4 four professors Dr. Blue, Green, Purple and Red. Which section a student enrolls in depends on a number of factors (eg. professor preference, lecture times, restrictions from other classes, etc.). Between the four professors, Dr. Green is the only one who uses clickers.

   Clickers are small handheld transmitters that students must purchase to participate in multiple-choice questions answered in class. Professor Green asked students to answer where their hometown is using their clickers. The possible answers were 'Canada' and 'Not Canada'. The results were as follows:

   - 107 students answered, with 55 selecting that their hometown is in Canada, and 52 selecting that their hometown is not in Canada.

We are interested in estimating - of all students enrolled in DATA 543 at the time the question was asked - the proportion who have a hometown in Canada. On the day the question was asked, 176 students were enrolled in Professor Green's section of DATA 543, and 608 students were enrolled in DATA 543 in total. We do not know how many students attended the lecture when Professor Green asked this question.

(a) Briefly describe, calculate, or state:
   i. (1 point) The target population;
   ii. (1 point) The population quantity of interest;
   iii. (1 point) The sample size;
   iv. (1 point) The population size;
   v. (1 point) The sample estimate of the population quantity of interest.

(b) (2 points) Describe the sampling protocol used for this study. You should specify whether it is a probability or non-probability sampling protocol, and identify the specific type of protocol used. For the specific type of protocol, refer to the list of non-probability sampling protocols and the list of probability sampling protocols covered in lecture. N.B. there may be more than one correct answer.

(c) Briefly describe:
   i. (2 points) One potential source of frame error in this study.
   ii. (2 points) One potential source of sample error in this study.
   iii. (2 points) One potential source of measurement error in this study.

2. Recently, I wondered how much time I have spent in the last year listening to podcasts. One way to estimate this would be to take a sample of the podcasts I have listened to and look at the duration of podcasts in that sample. The Podcasts.csv data file on github contains two columns named `id` and `duration`. `id` gives an identification number corresponding to individual podcasts and `duration` gives the length of that podcast (in minutes). You may consider this dataset as representing the entire population of podcasts under study. We will use μ to denote the population mean podcast duration, and σ^2 to denote the corresponding population variance.

(a) Provide each of the following:
   i. (1 point) The population size (N);
   ii. (2 points) μ **in minutes**;
   iii. (2 points) σ^2;
   iv. (2 points) The total duration of podcasts in the population $\tau = N\mu$ **in hours**.

(b) Please answer parts (b)i–v as if you had no knowledge of the full population (other than the population size). Take a sample of size $n = 30$ from the population in Podcasts.csv by simple random sampling without replacement. You may do this in R using the following commands:

```
Pod <- read.csv("Podcasts.csv")
set.seed(543)
My.Sample <- sample(Pod$duration, 30)
```

Important: Set the seed to 543

i. (2 points) Give the first five units of your sample.

ii. (3 points) Compute the sample average $\hat{\mu}$ and the sample variance for your sample.

iii. (3 points) Based on your sample construct a 95% confidence interval for μ, the (population) mean podcast duration **in minutes**.

iv. (3 points) Based on your answer to part (iii), construct a 95% confidence interval for the **total** duration of podcasts in the population **in hours**. You do not need to show your working.

v. (2 points) Provide an interpretation for the CI calculated in part (iv).

3. The 2016 United States presidential election saw an estimated turnout of 55.3%, meaning a little over half of the voting age population voted. For the purposes of this question, we will limit attention to the votes received by the two major party candidates: Donald Trump and Hillary Clinton. It has been estimated that total votes for Trump and Clinton were 62,979,860 and 65,844,952, respectively. This totals to 128,824,812 votes from an estimated voting age population of 251,107,406. If we just consider these two candidates, this converts to a turnout of 51.3%.

   We can think of an election as a (non-random) survey sample of the electorate as a whole. If we're prepared to assume voters decide whether or not to vote via simple random sampling without replacement, we can apply the various methods we have studied in this module. (You should think about how reasonable this assumption might be!) In particular, we can estimate the population mean for the proportion of the vote each candidate received.

   The United States Census Bureau divides the United States into four statistical regions: Northeast, Midwest, South, and West. We can therefore view each of these regions as individual strata, and extend our sampling assumption to suppose that voters decide to vote via simple random sampling without replacement within each of those strata.

   The following table summarizes the election data across these four regions. VAP is the 'voting age population', and can be thought of as the size of our target population in each region. (Source: http://www.electproject.org/2016g)

Region	VAP	Trump	Clinton
Northeast	44,644,808	10,108,658	13,652,408
Midwest	52,546,066	15,559,072	14,184,562
South	94,590,118	25,867,401	21,967,373
West	59,326,414	11,444,729	16,040,609

(a) (3 points) Consider the following notational definitions:

- N_h: the population size in the h^{th} region.
- $W_h = \frac{N_h}{N}$: the population stratum weight for the h^{th} region.
- n_h: the sample size in the h^{th} region. This is the total number of votes cast in that region for Donald Trump and Hillary Clinton **ONLY (we are ignoring votes cast for other candidates)**.
- $w_h = \frac{n_h}{n}$: the sample stratum weight for the h^{th} region.
- $\hat{\pi}_h = \frac{y_h}{n_h}$: the sample estimate of the proportion of voters who voted for **Donald Trump** in the h^{th} region, with y_h denoting the number of votes Donald Trump received in that region.

Create a table to summarize all the values listed above. Your table should follow this template:

Region	N_h	W_h	n_h	w_h	$\hat{\pi}_h$
Northeast					
Midwest					
South					
West					

(b) (4 points) For each of the four regions, write a brief sentence explaining whether it is over- or under-represented in the sample. In other words, compare the proportion of voters in the *sample* from each region with the proportion of voters in the *population* from each region.

(c) (3 points) Based on the voting figures, briefly discuss whether stratifying the sample into four regions seems reasonable. Provide one other stratification structure we might consider for these data (i.e., some other division of the population into strata).

(d) (2 points) Assuming we want to construct a new dataset with the same number of voters as seen is these data (i.e., $n = 128,824,812$), calculate the number of voters required in each stratum to achieve proportional allocation.

Note: Your calculations in parts (e) and (f) should be based on the sample sizes in the original dataset, rather than those calculated in part (d).

(e) (2 points) Based on these data, what is $\hat{\pi}$? More specifically, what is the overall sample estimate of π, the proportion of the voting age population who *would* vote for Donald Trump if they all voted? Note that this is *not* the stratified estimate. Convert this estimate to a percentage, written to 3 decimal places.

(f) (2 points) Based on these data, what is $\hat{\pi}_s$, the stratified sample estimate of π? Convert this estimate to a percentage, written to 3 decimal places.

(g) (2 points) Provide a reason why the stratified sample estimate in part (f) is different from the unstratified estimate in part (e).

(h) (6 points) Calculate 95% confidence intervals corresponding to your answers to parts (e) and (f). Your answers should be in the form of percentages, written to 3 decimal places.