STA304H1S/1003HS Winter 2021 Assignment # 1

Given: Monday, January 25, 2021
Due: Online into Quercus (as a Quercus Quiz) by 10pm on Thursday, February 4, 2021
Note: E-mail submissions will NOT be accepted. Late assignments will be accepted but subject to a 1% penalty of the total assignment marks per hour late. Submission will not be allowed beyond 48 hours of the due date.

Instructions:

- *Answer all three (3) questions of this assignment. Submit your answers, links and/or files into the Quercus Quiz named 'Assignment 1'.*
- *Each assignment should be written up independently. Questions 2 and 3 should contain unique answers. You are allowed to work with at most one classmate of your choice on Question 1. If you do, please include the full official name of the student on your solutions.*
- *Use proper statistical terminology and proper English language to ensure that your answers are clear.*
- *Submit an R script or RMarkdown file with your codes for us to reproduce your answers for Question 3.*

Grading: The grand total is 42 marks. A general rubric for each part is given below.

Level	Description
Exceptional 3 points	Complete, correct and clearly written answers. Answers model individual preparation and academic honesty (where applicable).
Good 2 points	Good answers that are unclear, contain few mistakes or missing components. Answers demonstrate some individual preparation and some academic honesty (where applicable).
Minimal 1 points	Poor answers or many missing components. Most answers do not demonstrate individual preparation or academic honesty (where applicable).
Non-participant 0 points	Missing or incomprehensible answers. Answers are not academically integral.

1. (9 marks) Find a recent sample survey published online over the last 13 months about a human or non-human Canadian population. *Note that you must include an active link to an appropriate survey for the grading of this part. The question will be awarded 0 marks if the link is not active or if the survey is improper.*

   (a) (3 marks) Describe the survey and at least one key question of interest. Describe the target population, sampled population, sampling frame (if possible), sampling design, sample size and observation unit.

   (b) (3 marks) State at least one concluding result drawn from the survey by the authors? Do you think that the conclusions are generalizable to the target population? Explain.

   (c) (3 marks) Describe two possible sources of error in the survey and propose possible ways to alleviate the errors.

2. (12 marks) Read section 2.6 of the textbook - *Elementary Survey Sampling*, Scheaffer et al (2012). Choose a human target population from one or all of University of Toronto's campuses and a population parameter of interest (for example, a mean, a proportion or a total). Suppose you are interested in planning a sample survey and you establish a questionnaire regarding your variable of interest. Answer the following:

(a) (3 marks) Clearly state your objective. Define your target population. Describe a suitable sampling frame. Describe your sampling design. Determine your method(s) of measurement.

(b) (3 marks) Write an appropriate open-ended question related to your objective.

(c) (3 marks) Write a closed question that is an alternative to the one in the part above.

(d) (3 marks) Write an example of a poorly worded question to your answer in part (b) or (c) above. Explain the error incurred in the wording.

3. (21 marks) Use all of the digits of your UofT student number as your population of interest. Produce and upload an R script or RMarkdown file to do the following parts. *Note that your codes must reproduce your answers, else this question will be awarded 0 marks.* Where appropriate, round your final answers to 2 decimal places. All plots produced must have properly labelled axes and be given an appropriate title which includes your student number.

   (a) (3 marks) Numerically describe the population of digits by finding the mean, median, variance, sum and proportion of digits like the first digit in your number.

   (b) (3 marks) Produce a frequency histogram of your population.

   (c) (3 marks) Produce the frequency histogram of the sampling distribution of the sample mean based on drawing samples of size 3 **without replacement** from your population. (Hint: The base R function choose() and the combn() function in the R package *combinat* can be used to answer this part.)

   (d) (3 marks) Compute the mean and variance of the sample means of part (c).

   (e) (3 marks) **Set the seed of your randomization to be your student number.** Simulate the sampling distribution of the sample mean by using 250, 000 samples of size 3 based on drawing **with replacement** from your population. Produce the frequency histogram of the empirical sampling distribution of the sample mean.

   (f) (3 marks) Compute the mean and variance of the sample means based on part (e).

   (g) (3 marks) Compare the results of part (c) and (e) to the theoretical properties of the sample mean based on samples of size 3. Explain the similarities or differences found.