STA304H1F/1003HF Fall 2018 Assignment # 2

Given: Friday, October 5, 2018
Due: Online into Crowdmark by 10pm on Thursday, October 25, 2018

Note: E-mail submissions will NOT be accepted. Late assignments will be subjected to a penalty of 20% per day late. Submission will not be allowed beyond 48 hours of the due date.

Instructions:

- *Answer all three (3) questions of this assignment.*
- *Each assignment should written up independently. Questions 2 and 3 should contain unique answers. If you work with other students on Question 1, indicate the names of the students on your solutions.*
- *Presentation of solutions is important. Assignments should be word-processed and presently neatly.*
- *Use proper statistical terminology and write in plain English.*
- *Supporting materials, such as R codes and extraneous output should be placed in an Appendix.*
- *Compile your entire solution, including your Appendix, as a PDF document (Word, LaTeX or Rmarkdown can be your base).*

Grading: The grand total is 60 marks. Each of the 11 parts is worth 5 marks and appendix/presentation of results is worth 5 marks. A general marking scheme for each part is given below:

- 5 points: complete and correct answers
- 4 points: answers with minor problems
- 3 points: good answers that are unclear, contain some mistakes, missing components
- 2 points: poor answers with some value
- 1 point: irrelevant answers
- 0 point: unanswered questions or **if instructions are not followed**

1. Suppose that $\{y_1, y_2, \ldots, y_n\}$ denotes a simple random sample without replacement from a finite population of values, with population mean $\mu = \sum_{i=1}^{N} y_i/N$ and population variance $\sigma^2 = \sum_{i=1}^{N}(y_i - \mu)^2/N$. The sample mean and sample variance are defined as $\bar{y} = \sum_{i=1}^{n} y_i/n$ and $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$.

   (i) Show that $\mathrm{E}(s^2) = \frac{N}{N-1}\sigma^2$.

   (ii) Using s^2, find an unbiased estimator of $\mathrm{V}(\bar{y})$ where $\mathrm{V}(\bar{y}) = \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right)$.

   (iii) Show that the sample variance $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$ can be written as $s^2 = \frac{n}{n-1}\hat{p}(1-\hat{p})$ for $\{0,1\}$ data, where $\hat{p} = \sum_{i=1}^{n} y_i/n$ is the sample proportion.

2. Consider the data set on certify.csv (description on certify.pdf) about 1994 Survey of American Statistical Association Membership on Certification. Use R to answer the following questions. **Set the seed of your randomization to be the last 4 digits of your student number.**

(a) Take a **simple random sample without replacement** of size 100 and estimate the proportion of respondents who think that ASA should develop some form of certification (YES=1).

(b) Compute $\widehat{Var}(\hat{p})$ and create a 95% confidence interval for the population proportion.

(c) What should be the sample size n if we set a margin of error B=5%.

(d) Recalculate parts (a) and (b) using the sample size obtained in part (c). What is the range of the confidence interval? (Range=Upper limit - Lower limit)

3. Suppose the data, "goals.csv" available in Quercus, consist of the grades that all STA304/1003 L0101 Fall 2018 students are aiming to achieve. Use R to answer the following parts. **Set the seed of your randomization to be the last 4 digits of your student number.**

(a) Find the population mean and population variance. Draw a histogram to describe your population.

(b) Select a simple random sample of 10 grades. Draw a histogram to describe your sample. Compute the sample mean and sample variance.

(c) Repeat part (b) 49 additional times and combine to have 50 sample means. Draw a histogram to describe your 50 sample means. Now find the sample mean and sample variance of the 50 sample means. Compare with your histogram, mean and variance to those in parts (a) and (b). Include in your appendix the index positions of your 50 samples.

(d) Specify completely two potential stratification variables for grade goal; describe all levels of the variables. Discuss one advantage and one disadvantage of stratifying in comparison to simple random sampling.