STA304H1F/1003HF Fall 2018 Assignment # 3

Posted: Friday, November 2, 2018
Due: Online into Crowdmark by 10pm on Tuesday, November 27, 2018

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

Instructions:

- *Answer all three (3) questions of this assignment.*
- *Each assignment should be written up independently. Questions 1 and 2 should contain unique answers. If you work with other students on Question 3, 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:

Per Question Part

- 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

Presentation and Appendix

- 5 points: well presented, easy to read, proper English used, R code and extra output in Appendix
- 3 points: good presentation, some R code in main write-up.
- 1 point: poor presentation, handwritten, handdrawn diagrams, R code in main section.
- 0 point: illegible, missing R-codes/output

1. [30 *marks*] Consider the baseball dataset describing the population of baseball players in the data file `baseball.csv`. **Set the seed of your randomization to be the last 4 digits of your student number.**
   The R package- 'sampling', which includes the functions- strata and getdata, is useful for this question. The following R codes show how to install and load the package.

```
install.package("sampling")
#load sampling package, to use the functions- strata and getdata
library(sampling)
```

   (a) Take a stratified random sample of 150 players, using proportional allocation with the different teams as strata (teams are in column 1 of the data file). Describe how you selected the sample.

(b) Find the mean of the variable $logsal = \ln(salary)$, using your stratified sample, and give a 95% CI.

(c) Estimate the proportion of players in the data set who are pitchers, using your stratified sample, and give a 95% CI.

(d) Take a simple random sample of 150 players and repeat part (c). How does your estimate compare with that of part (c).

(e) Examine the sample variances of *logsal* in each stratum. Do you think optimal allocation would be worthwhile for this problem?

(f) Using the sample variances from (e) to estimate the population stratum variances, determine the optimal allocation for a sample in which the cost is the same in each stratum and the total sample size is 150. How much does the optimal allocation differ from proportional allocation for this scenario?

2. [15 *marks*] Use the population data set `hh18.csv` with $N = 251$ pairs of measurements of height, x and handspan, y from our class to mainly compare regression and ratio estimation for estimating the mean handspan μ_y, using information from a sample of size n =10. **Set the seed of your randomization to be the last 4 digits of your student number.**

(a) Compute a SRS estimator, a ratio estimator and a regression-based estimator of the population mean handspan μ_y.

(b) Find the error of estimation, $|\hat{\mu} - \mu_y|$ for each of the three estimators in part (a) and compare them.

(c) Compute and compare the estimated variances of the three estimates.

3. [10 *marks*] A market research firm constructed a sampling plan to estimate the weekly sales of brand **A** cereal in a certain geographic area. The firm decided to sample cities within the area and then to sample supermarkets within cities. The number of boxes of brand **A** cereal sold in a specified week is the measurement of interest. Five cities are sampled from the 20 in the area. Using the data given in the accompanying table, answer the following:

City	Number of supermarkets	Supermarkets sampled	$\bar{y}_i$	s_i^2
1	45	9	102	20
2	36	7	90	16
3	20	4	76	22
4	18	4	94	26
5	28	6	120	12

(a) Estimate the average sales for the week for all supermarkets in the area. Place a bound on the error of the estimation. Is the estimator you used unbiased?

(b) Do you have enough information to estimate the total number of boxes of cereal sold by all supermarkets in the area during the week? If so, explain how you would estimate this total, and place a bound on the error of estimation.