Statistics 144 - Spring 2022

Take Home Final Exam correction P5b

Due: June 7, 2022 at 11pm

Name:

Note, the exam is open book. You are not allowed to search for solutions online or communicate with another person about the content of this exam. Any attempt to do so will be considered academic dishonesty and will be reported. The exam has 6 problems, each worth 50 points. When you are asked to extract a random subset please, submit your selected data set with your final exam.

PROBLEM 1

The data set *anthrop.csv* contains the length of the middle finger and height of N=3000 criminals. We will treat this data set as a population of size N=3000. You will select a random sample of n=200. *Note: you are each requested to select your own random sample. If there are two identical answers it will be considered cheating. The probability of this happening is so small as to be practically zero.* Calculate the following:

(a) Estimate the average height based on a simple random sample of n=200 subjects and its standard error. Find a 95% confidence interval.

(b) Determine the sample size necessary to have an absolute error of at most 2 inches. Use the whole data set to calculate the variance and compare the estimate of$n$ you get if you were to treat your current sample of n=200 as a pilot sample for a future survey.

PROBLEM 2

Use the same sample as obtained for problem 1

(a) Calculate the ratio estimate of average height and its standad error using finger length as the auxiliary variable. Calculate a 95% confidence interval.

(b) Repeat the sample size determination but this time for the ratio estimate.

(c) Repeat estimation of average height but use a regression estimate and find the standard error.

PROBLEM 3

The questions below refer to your results obtained in problems 1 and 2.

(a) Compare average height $\bar{y}_{SRS}$, $\bar{y}_r$, $\bar{y}_{reg}$ on the basis of standard error. Which estimate do you prefer and why?

(b) Compare ratio and regression estimate. Which one do you believe is more appropriate here and why. Use plots as necessary to support your argument.

PROBLEM 4

The data file *baseball.csv* contains statistics on N=797 baseball players. You are asked to select a stratified random sample of n=200 players, using proportional allocation by *LeagueID*, AL vs NL.

(a) Calculate the average salary (use log salary for your calculations and then transform to salary) for each league based on your sample and the overall average based on your stratified sample. Calculate standard error and 95% confidence interval for each estimate.

(b) Calculate the total number of games played, again for each league and the combined estimate with standard errors and 95% confidence intervals.

PROBLEM 5

Use the baseball data from problem 4.

(a) Take a one stage cluster sample (SRS of clusters) of 10 teams (clusters) and estimate average salary and the standard error plus calculate a 95% confidence interval.

(a) Repeat the calculations in (a) but this time take a one stage cluster sample of $n = 10$ clusters with replacement with probability proportional to size using Lahiri's method. Calculate average salary, standard error and 95% confidence interval. Do your results differ from part (a)? Explain!

PROBLEM 6

Suppose we have a one stage cluster sample with clusters of equal size. Show

$$V(\hat{t}_{unb}) = N^2(1 - \frac{n}{N})\frac{S_t^2}{n} \qquad \text{were} \qquad S_t^2 = \frac{1}{N-1}\sum_{i=1}^{N}(t_i - \frac{t}{N})^2$$