Statistics 144 - Spring 2023

Take Home Final Exam

Due: June 12, 2023 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.

You are asked to analyze the NHANES data set and answer the questions below. The data set is large and we will focus on a subset of the variables. The subset is given below:

1 Weight (bmxwt)
2 Body Mass Index (bmxbmi)
3 Waist circumference (bmxwaist)
4 Thigh circumference (bmxthicr)
5 Age
6 sdmvstra - a cluster indicator with 15 clusters which represent pseudo strata.

Problem 1

a Analyze the whole data set treating it as the population. Calculate mean, standard deviation of the population mean for Weight, Body Mass Index (BMI), Waist and Thigh Circumference.

b Discretize *Age* into 3 strata and repeat part (a) for each stratum. For the strata choose 0-14, 15-35, 36+ years.

Omit all cases with missing data. You should do this prior to sampling. That way, you avoid having incomplete observations in your sample.

Problem 2

a Take a SRS of 10% from the data set and calculate estimates for the quantities in problem 1. Calculate standard errors for your estimates and compare the true standard deviations you can calculate from the quantities in problem 1.

b Take a 10% stratified sample, stratifying on age and repeat the analysis. For the strata, use as an age cutoff 0-14, 15-35, 36+; use optimal allocation assuming constant cost and minimize variation for weight.

Problem 3

Use the clusters given and take a random sample of 2 clusters without replacement and with probability proportional to size. Calculate the Horvitz-Thompson estimator of the average of the variables and standard errors. Calculate averages making use of the fact that you know M_0. Repeat two more times for additional samples. Follow the example in the textbook with the 4 supermarkets. R code will be posted as needed.

Problem 4

a Use the original nhanes data set given and sort by BMI. Create a new data set as follows: take the first 100 observations (with the 100 smallest BMI). This is cluster 1. The next cluster will be BMI 1001-1100, cluster 3 will be BMI 2001-2100, etc until you reach cluster 10 where BMI ranges from rank 9001 to 9100. Discard the rest of the data. Calculate ICC and R_a^2 for BMI. Calculate mean and variances within clusters and the variance of the mean over the clusters.

b Repeat the analysis in part (a) but reorder the data randomly. For the randomly divided data create clusters as follows: the first 100 values are cluster 1, the next 100 values in the ist are cluster 2 etc.

c Compare your variance estimates from (a) and (b). Discuss your findings. For which of (a) or (b) would a cluster sample likely be similar in variation to a SRS? Justify your answer with an example in the textbook.