MSDS 6370 Sampling Statistics Final Exam
Spring 2017

This exam is due at midnight CT on Monday, April 17, 2017. You are to work independently on this exam. You may not consult other people. However, you may use course materials. (Please take SMU Honor code seriously)

1. (8 pts) Select the best answer to each question below:

(i) What is the purpose of using a poststratification adjustment?
A. To reduce the variance of the estimator due to undercoverage.
B. To reduce the variance of the estimator due to the sample design's unequal selection probabilities.
C. To reduce the bias of the estimator due to undercoverage.
D. To reduce the bias of the estimator due to the sample design's unequal selection probabilities.

(ii) A SRS of 100 patients from a particular doctor's practice is chosen in order to estimate the total unpaid charges for all patients. All bills from the past 12 months for each sampled patient are selected, and amount of the unpaid charges on each bill is recorded. The doctor billed a total of 900 patients during the year.

This sample (all bills from sampled patients) can be thought of as which of the following? __ (enter a, b, or c)

(a) a simple random sample of bills

(b) a cluster sample of bills

(c) a stratified sample of bills.

2.(15 pts) A simple random sample of 100 of the 1000 housing units in a small community is sampled. Their water meters are monitored during the restricted watering portion of one day in order to estimate the total water usage in the community for that day, which fell in the drought season. The sample mean and sample variance are found to be $\bar{y} = 130.0$ gal., $s^2 = 125{,}000$.

(a) Construct a 95% confidence interval for the total gallons of water used during the restricted watering portion of that day for the whole community.

(b) Suppose that the city decided to save money next year in data collection by selecting the sample of 100 by using city blocks as a frame. Suppose there are 100 blocks in the community, each with an average of 10 housing units. Their plan was to randomly select 10 blocks and sample all the houses on those blocks, instead of the plan in part (a). Would the margin of error for this design be likely to be larger or smaller than what you calculated in part (a)? _____. Carefully explain your reasoning.

(c) You are the statistical consultant for the community. They ask you to determine how many blocks they would need to sample in order to achieve the SAME margin of error as they did in the analysis in (a), but using the city blocks as sampling units. You tell them you would need to know the value of the intra-cluster correlation, ρ. They don't have any data on that; however, water usage is highly correlated with size of the lot for the housing unit, for which $\rho = 0.4$. You decide to use that for planning purposes. How many blocks must they sample?

3. (12 pts) The SMU student directory contains 94 pages of student listings. Each page has 115 lines of text and 4 columns of names. The number of students listed on each page can vary, since different students have a different number of lines.

Examples of sample listings in SMU directory

ABABTAIN Eman Abdulrahman A
Masters……xxx/xxx-xxxx
Lyle SOFT-MS
9030 Southwestern Blvd Apt 3231
Dallas TX 75214-1542
ABALOS LIRA Jose Pedro
Masters……xxx/xxx-xxxx
Guildhall DGLVLD-MIT
6400 Ohio Dr Apt 321 Plano TX 75024-2659
Los Ginkos 13425 Santiago, RM 7591532
CHILE
ABBAH Ucha Chinyere
First Year……..xxx/xxx-xxxx
Dedman ENGUN-PMJ
PO Box 751347 Dallas TX 75275-1347
5632 Arlington Park Dr
Dallas TX 75235-6202
ABBAS Sabrina First Year xxx/xxx-xxxx
Dedman CRCOMP-PMJ
PO Box 751378 Dallas TX75275-1347
1700 Windemere Dr Plano TX 75093-2844

1 page, 4 columns and 115 lines

115 lines

Consider the following sampling plans for selecting a sample of SMU students listed in the directory. ***For each, note whether it is a probability design or not, and explain why***.
For each of the sample designs you identified as probability samples, specify how you would define the features in PROC SURVEYMEANS so that a valid estimator of the mean and its standard error can be computed. In particular, for each so designated design, specify (i) the weight for each sampled student; (ii) Do you include the STRATUM statement in PROC SURVEYMEANS? If you include STRATUM statement then what is the STRATUM variable? (iii) Do you include CLUSTER statement in PROC SURVEYMEANS?
If you include CLUSTER statement then what is the CLUSTER variable?

(a) You use a random number table to select a simple random sample (SRS) of 10 of the 94 pages. Then you select all the students on the selected pages into your sample. Is this a probability sample? Circle yes or no. Why?

(b) For every one of the 94 pages in the directory, you select a SRS of 2 of the 4 columns. Then you select at random 1 of the 115 lines from each of those 2 columns. Then you select the student to whom that line pertains. Circle yes or no. Why?

(c) You select a SRS of 40 of the 94 pages. Then you select the first and last students listed on each selected page. Circle yes or no. Why?

4. (10 pts)

A population of 16 balls (ordered as shown below) is made up of 7 white balls and 9 red balls as shown below:

R R W R W R W W R R R W W R R W

A systematic sample of size 3 is selected from the population and the proportion of red balls in the population is estimated from the sample using as the estimator

$$\hat{P} = \text{the proportion of red balls in the sample}.$$

Find the sampling distribution of $\hat{P}$.

5. (15 pts) For this exercise, you will examine the HRS data (download from Asynchronous Lecture 11.4.1).(Same data set you used for lab11).

Consider the question of whether or not diabetes (DIABETES) is independent of race/ethnicity (RACECAT). Using the Rao-Scott Chi-square test, determine if diabetes and race/ethnicity are independent or not. Include your SAS code and output, as well as 6 steps of hypothesis.