BST 5420
Spring 2019
Midterm Exam

Name Key

Directions: Open Book, open notes; computer allowed; internet allowed; communication with another person not allowed.

1. (12 Points: 3 each) In each of the following examples, determine whether the sample described is a

- simple random sample
- stratified random sample
- systematic sample
- one-stage cluster sample
- two-stage cluster sample

For each one, give the important sampling characteristics; for example, for a stratified sample, give the population size, each stratum size, and the sample sizes from each stratum. Some of these will be unknown, so in this case, just say "unknown".

(a) The US Census Bureau sends a short census form to every household address, and it sends a long form to every tenth address. Consider those households that get the long form.

Systematic Sample

$k = 10$

N = Population size = virtually infinite

(b) "The university was interested in student feedback regarding the operating hours and services offered by the Health Services Center. We obtained a list of all 14,252 students at the university. We then selected 200 random numbers between 1 and 14,252 using the `sample` function in R. We then sent an email questionnaire to each of the 200 students who were chosen."

Simple Random Sample

$N = 14,252$

$n = 200$

(c) "Because malaria is still a problem in a particular country, we designed a survey to determine the rate of use of bed nets to prevent mosquito bites. The country is divided in to 12 regions (like counties). We selected a simple random sample of 50 households within each of the 12 regions. We then interviewed each household in the 50 selected to determine whether bed nets were used regularly."

Stratified Random Sample

$H = 12$

$n_i = 50, \quad i = 1, 2, \ldots, 12$

$n = n_1 + n_2 + \cdots + n_H = 12 \times 50 = 600$

(d) After a surge in microcephaly cases, a study group was formed to determine the prevalence of diseases such as Zika, chikungunya, and dengue. A list of 400 towns/cities is obtained. A sample of 40 of these cities/towns is selected, and within each city/town a sample of 20 individual adult persons is selected and interviewed.

Two-stage cluster sample

N = # of clusters = 400

M_i = # households in cluster i = unknown

n = # clusters sampled = 40

m_i = # in sample from cluster i = 20

2. (3 Points)

Consider the wording of the following two questions:

(Version #1) Do you believe that every person should be required to purchase health insurance?
(Version #2) Do you believe that the government should force every person to buy health insurance?

Which do you think will get a higher proportion of "Yes"? Why?

Version #1. Saying "the government should force" is strong language. People don't want to be "forced" to do something, especially by the government.

3. (8 Points: 6,2) A stratified random sample of students in a particular middle school (grades 6, 7, and 8) regarding the number of sugared drinks that they consume per week is being proposed. Information on the grade sizes and the proposed plan are shown in the table below.

Grade	Number of Students	Sample Size
6	40	n_1
7	60	n_2
8	60	n_3

It is believed that the standard deviations in the three groups are $\sigma_1 = 1,\ \sigma_2 = 2,\ \sigma_3 = 2$. The costs for sampling each of the three groups in this stratified random sample are the same. It has been decided that a total sample of size $n = 60$ will be taken.

(a) Determine the optimal allocation of the total sample across the three grades.

$c_1 = c_2 = c_3 = 1$

$$A_1 = \frac{N_1\sigma_1}{N_1\sigma_1 + N_2\sigma_2 + N_3\sigma_3} = \frac{40\times 1}{40\times 1 + 60\times 2 + 60\times 2} = \frac{40}{280} = 0.143$$

$$A_2 = A_3 = \frac{120}{280} = 0.429$$

$$n_1 = 60\times 0.143 = 8.58$$

$$n_2 = n_3 = 60\times 0.429 = 25.74$$

Rounding up gives

$n_1 = 9$
$n_2 = 26$
$n_3 = 26$

This goes over budget by one, so we might round one of these numbers down.

(b) If the standard deviations in the three groups were really $\sigma_1 = 1,\ \sigma_2 = 2,\ \sigma_3 = 4$ what effect would this have on the allocation of the total sample? Circle one of the following. No justification is needed.

(i) n_3 would be greater than the corresponding number found in part (a).

(ii) n_3 would be less than the corresponding number found in part (a).

(iii) n_3 would be the same as the corresponding number found in part (a).

4. (11 Points: 3,5,3) You would like to estimate the reading comprehension of fourth graders at a particular school. You have a list of all 10 fifth grade classes in the district, that includes altogether 260 students. You select an SRS of 4 of these 10 classes, and you select 6 students from these five selected classes. For each of the 24 students selected, you assign several passages to read, and then observe how many questions can be answered correctly.

(a) What kind of sampling plan is this?

Two-stage cluster sample

(b) Given data in the table below, give a point estimate for the mean number of questions answered correctly for the whole school district.

Class	Number of students in the selected class	Number of students selected	$\bar{y}$ = average number of correct answers among selected students
1	22	6	13.3
2	25	6	18.3
3	28	6	12.0
4	25	6	16.7

$$\hat{\tau} = \frac{N}{n} \sum_{i \in S} M_i \bar{y}_i$$

$$= \frac{10}{4} \left[22 \times 13.3 + 25 \times 18.3 + 28 \times 12.0 + 25 \times 16.7 \right]$$

$$= \frac{10}{4} \, 1503.6$$

$$= 3759$$

$$\hat{\mu} = \frac{3759}{260}$$

$$= 14.45769$$

(c) Do you have enough information to compute the standard error of the estimate of the mean score? If so, indicate what formula you would use. If not, explain what additional information is needed. [Do not do any calculations for this problem.]

No. We would need to have the within cluster standard deviations s_i^2 as indicated in the book's equation (5.28) or the formulas in the Chapter 5 notes (slide 30)

5. (8 Points: 4,4) The white-footed mouse is suspected of being a reservoir host for Lyme disease. The town's public health department would like to estimate the number of white-footed mice in its town. (They will repeat this exercise in several months after a plan to kill most white-footed mice is carried out.) Initially, they capture 400 mice, tag them, and then release them. Then, one week later, they capture another 200 mice and find that 27 were tagged (i.e., they had been caught also in the first capture).

(a) Find a point estimate for the population size.

$n_1 = 400$

$n_2 = 200$

$m = 27$

$$\hat{N} = \frac{n_1 n_2}{m} = \frac{400 \times 200}{27} = 2963$$

(b) Find the standard error for the estimate in part (a).

$$\hat{V}(\hat{N}) = \left(\frac{n_1 n_2}{m}\right)^2 \frac{n_2 - m}{m(n_2 - 1)}$$

$$= 2963^2 \frac{173}{27 \times 199}$$

$$= 282,671$$

$$se = \sqrt{282,671}$$

$$= 531.668$$

6. (11 Points: 2,5,4) A university is interested in what proportion of faculty are satisfied with the health insurance offered as part of employment. A sample of 200 faculty was sent an on-line questionnaire and only 120 returned the survey. Suppose that across the university there are 900 faculty and the distribution among ranks is evenly divided, with one-third being Assistant Professor, one-third being Associate Professor, and one-third being Professor. The results are shown below.

Rank	Number Responding	Number "Satisfied"	Proportion "Satisfied"
Assistant Professor	40	30	0.750
Associate Professor	20	15	0.750
Professor	60	20	0.333
Total	**120**	**65**	**0.542**

(a) What is the overall response rate?

$\frac{120}{200} = 0.60 = 60\%$

(b) Do a poststratification analysis and give a point estimate for the overall proportion in the population who would answer "Yes."

$$\hat{P}_{post} = \frac{N_1}{N}\hat{P}_1 + \frac{N_2}{N}\hat{P}_2 + \frac{N_3}{N}\hat{P}_3$$

$$= \frac{1}{3}0.750 + \frac{1}{3}0.750 + \frac{1}{3}0.333$$

$$= 0.250 + 0.250 + 0.111$$

$$= 0.611$$

(c) Suppose that instead of poststratification, the researchers decided to apply two-phase (or double sampling). Explain how you would use a two-phase sample to estimate the proportion who are satisfied. Don't do any calculations for this problem, just explain how it would be done and give some indication of what formulas you would use.

Focus on the 80 nonrespondents and select some proportion of them for follow up contact. If we decided to recontact 10%, then we would have to follow up with 8 people. Then use the book's equation (8.1) or the formula on slide 14 of the Chapter 8 notes.