BST 5420 **Name** Key

Spring 2017

Midterm Exam

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) "We developed a study to learn about hospital overcharges in a hospital system that consists of seven different hospitals. Within each of the seven hospitals, we selected a sample of 200 patient records and examined each to see whether there was at least one overcharge."

Stratified Random Sample

Strata: Hospitals

Number of Strata: 7

Sample Size: 200 records within each sample

(b) "In order to learn about the health in the region around a lead smelter plant, we took our list of all 2158 residential addresses in the whole region. Beginning with the fifth residence (which was determined by chance) we selected every tenth residence. We then interviewed one adult from each of these residences."

Systematic sample
Sampled every $k = 10^{th}$ residence

(c) "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, and if there were duplicates, we replaced them with other values. We then sent an email questionnaire to each of the 200 students who were chosen."

Simple Random Sample
Population Size 14,252
Sample Size 200

(d) "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 210 regions (like counties). We selected 40 of these regions at random. Within each of these regions, we selected a random sample of 50 households. We then interviewed each household to determine whether bed nets were used regularly."

Two-stage cluster sample
Clusters are regions
Number of clusters: 210
Number of clusters sampled: 40
Number of households sampled in each cluster: 50

2. (4 Points: 2,2)

(a) Consider the wording of the following two questions:

(Version #1) Do you approve of an amendment that would require congress to pass a balanced budget each year, even if this would prevent the government from stimulating the economy during a recession?

(Version #2) In order to prevent the government from overspending, do you approve of an amendment that would require congress to pass a balanced budget each year?

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

#2, because the question gives a reason why you might want to not have a balanced budget.

(b) Creators would like to ask the following questions.

(Question A) Will you support an increase in state taxes for education?

(Question B) Will you support an increase in state taxes?

Which question order do you think will lead to more "Yes" answers to Question B? Asking them in the order A-B? Or asking them in the order B-A?

Question A because the question gives a reason for supporting a tax increase.

3. (10 Points: 5,5) A survey is being proposed of students in a particular middle school (grades 6, 7, and 8) regarding the number of sugared drinks that they consume per week. 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.

(a) How large of a total sample is required if we are to estimate the mean number of sugared drinks that students drink per week with a margin of error of 0.5? $c_1 = c_2 = c_3 = 1$

$e = 0.5$

$D = \frac{e^2}{4} = \frac{0.25}{4}$

$$n = \frac{\left(\sum_{h=1}^{3} \frac{N_h \sigma_h}{\sqrt{1}}\right)\left(\sum_{h=1}^{3} N_h \sigma_h \sqrt{1}\right)}{N^2 D + \sum_{h=1}^{3} N_h \sigma_h^2}$$

$$= \frac{(40\times 1 + 60\times 2 + 60\times 2)^2}{160^2 + 40\times 1 + 60\times 4 + 60\times 4}$$

$$= 36.98$$

Round up to 37

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

$$A_1 = \frac{N_1\sigma_1}{N_1\sigma_1 + N_2\sigma_2 + N_3\sigma_3} = 0.143$$

$$A_2 = \frac{N_2\sigma_2}{N_1\sigma_1 + N_2\sigma_2 + N_3\sigma_3} = 0.429 = A_3$$

$n_1 = A_1 n = 5.3$

$n_2 = A_2 n = 15.86$

$n_3 = A_3 n = 15.86$

Round to

$n_1 = 5$

$n_2 = 16$

$n_3 = 16$

4. (8 Points: 2,4,2) 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. 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, estimate 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	12.3
2	25	6	19.3
3	28	6	14.0
4	25	6	15.7

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

$$= \frac{10}{4} \left[22 \times 12.3 + 25 \times 19.3 + 28 \times 14.0 + 25 \times 15.7 \right]$$

$$= 2.5 \times 1344.6$$

$$= 3361.5$$

M is unknown

$$\hat{M} = N \bar{M}$$

$$= 10 \, \frac{22+25+28+25}{4}$$

$$= 250$$

$$\hat{\mu} = \frac{\hat{\tau}}{\hat{M}} = \frac{3361.5}{250} = 13.446$$

(c) Suppose (as is likely the case) that you need parental consent for the students to participate. If only 60% of parents consent for their child to participate, discuss the potential bias in the results. Are you likely to over- or underestimate the population mean?

Parents who refuse are likely to be less involved in their child's education. The volunteers will tend to be the ones with parental involvement and thus better students.

We would probably overestimate the students' reading skill.

5. (6 Points: 3, 3) The white-footed mouse is suspected of being a reservoir host for Lyme disease. A town 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 342 mice, tag them, and then release them. Then, one week later, they capture another 400 mice and find that 20 were tagged (i.e., they had been caught also in the first capture).

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

N = pop. size

$n_1 = 342$

$n_2 = 400$

$m = 20$

$$\hat{N} = \frac{n_1 n_2}{m} = \frac{342 \times 400}{20} = 6840$$

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

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

$$= \left(\frac{342 \times 400}{20}\right)^2 \frac{400 - 20}{20 \times 399}$$

$$= 2{,}227{,}886$$

$$\text{MOE} = 2\sqrt{\hat{V}(\hat{N})} = 2985$$

6. (6 Points: 2,2,2) The use of complex surveys is now common in public health research. Sources such as these are called *secondary data sources* since the data were not collected for any particular research project, but can be used to address many other research questions. Often the federal government or some international organization runs these data collection schemes.

(a) Give two examples of such data sources.

NHANES

National Crime Victimization Survey

(b) What information is needed to obtain point estimates of means, totals, or proportions?

① the outcomes y

② the sampling weights

(c) Describe the information that is needed in order to run models, including standard errors and margins of error, using this complex survey data in SAS.

In addition to the information from (b) we also need the stratum and cluster information for each observation.

7. (6 Points: 3,3) One question in a survey based on an SRS was "Do you favor an income tax increase to fund Medicaid expansion?" In a sample of 600 people, 240 said "Yes." After a little investigation, it was discovered that the response rates were different for men and women. The results by gender were as shown in the table below. Men and women each make up 50% of the population.

Gender	Number Responding	Number of "Yes"	Proportion of "Yes"
Male	200	60	0.30
Female	400	180	0.45
Total	**600**	**240**	**0.40**

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

$$\hat{P}_{post} = \frac{N_1}{N}\hat{P}_1 + \frac{N_2}{N}\hat{P}_2$$
$$= 0.5 \times 0.30 + 0.5 \times 0.45$$
$$= 0.375$$

(b) Find the margin of error for the poststratification estimate from part (a).

$$\hat{V}(\hat{P}_{post}) = \frac{1}{600}\left(1 - \frac{600}{\infty}\right)\left[0.5 \times 0.3 \times 0.7 + 0.5 \times 0.45 \times 0.55\right]$$
$$+ \frac{1}{600^2}\left[(1-0.5)0.3 \times 0.7 + (1-0.5) \times 0.45 \times 0.55\right]$$
$$= 0.0003819$$
$$MOE = 2\sqrt{\hat{V}(\hat{P}_{post})}$$
$$= 0.039$$