SURVMETH / SURVEY 625
Applied Sampling
Winter / Spring 2025 Name: ___________________________________

Midterm Examination [PRACTICE]

IMPORTANT: Show all formulas and steps during your calculations. You are welcome to email us an Excel file or a set of code to support your work, but please refer to specific parts of the file / code in your responses. For all results and calculations, use four decimal places. Use the overleaf (the back of the previous page) if you need more room. You may also use additional sheets of 8 ½" by 11" paper, but place your name on the upper right of each additional sheet, and clearly label the problem on the sheet. Read all questions carefully and make sure to answer all parts of each question. Please contact Miss Akari Oya immediately if you have any questions or need clarification on any of the problems (815-780-7968/akarioya@umich.edu).

The exam has 11 pages in total.

1. A Probability Proportionate to estimated Size (PPeS) selection of clusters from the following list of blocks is needed. The measure of size (MOS) is the number of housing units in the block at the last Census. The "Cumulative" column is the cumulated sum of measures of size.

Block	Cluster Number	MOS	Cumulative
1		50	50
2		20	70
3		25	95
4		70	165
5		100	265
6		80	345
7		55	400
8		50	450
9		40	490
10		60	550
		550	

   a. Ensure that each cluster has a minimum Measure of Size (MOS) of at least 35. Fill in the "Cluster Number" column to show how you have formed linked clusters.

b. What is the selection interval (*k*)? What range of numbers is valid for a random start in this case? Select 3 clusters using a random start of 12. Which three clusters did you select?

c. Calculate the probability of selection for each of the three selected clusters.

d. The desired overall sampling rate (*f*) is 48/550. Calculate the required subsampling rate for each of the selected clusters that will produce an *epsem* sample of housing units.

2. Consider the following frame of blocks

Block	Cluster Number	MOS	Cumulative MOS
1		50	50
2		10	60
3		63	123
4		0	123
5		22	145
6		52	197
7		0	197
8		28	225
9		100	325
10		75	400

a. Select 2 blocks with probabilities proportionate to size, using a minimum size of 50. Use systematic selections with a random start of 75.0. If linking is necessary, identify the linked units in the appropriate column.

b. Give the probability of selection for the primary selections (i.e., the clusters formed from the blocks) made in problem (i).

c. The overall probability of selection is f = 0.05. Give the within-cluster selection rate for each cluster selected in Part 2a that will result in an epsem sample of housing units.

d. After selecting clusters, field staff are sent to selected blocks to count the number of housing units that are actually there. Suppose the (potentially linked) cluster 2 is selected and the real # of housing units is 95, calculate the expected number of housing units to be selected in cluster 2 given the actual counts.

3. Tables in Question 1 and Question 2 are the sample frame from two strata.

   - For Stratum 1: three clusters were selected, i.e., $n_1 = 3$
   - For Stratum 2: two clusters were selected, i.e., $n_2 = 2$

   a. What are stratum-specific zone size?

   b. To retain epsem with the desired overall sampling rate (*f* = 0.1), what is the targeted subsample size for each cluster in Stratum 1?

   c. To retain epsem with the desired overall sampling rate (*f* = 0.1), what is the targeted subsample size for each cluster in Stratum 2?

4. Please indicate whether each of the following statements are TRUE or FALSE by circling what you believe to be the correct answer.

   a. (TRUE / FALSE) The Taylor Series Linearization method for estimating the sampling variance of an estimated mean is necessary when selecting a simple random sample of n clusters from a total population of N clusters, and a simple random sample of m elements from the M population elements in each of the n randomly sampled clusters.

   b. (TRUE / FALSE) The use of weights in analyses (whatever introduced the need to use them) will generally tend to increase the sampling variance of estimates, and the more variable the weights are, the higher the increase.

   c. (TRUE / FALSE) Under certain assumptions, one can compute the size of a simple random sample needed to achieve a certain level of sampling variance for a given estimated mean if the only two pieces of information provided are 1) the desired level of sampling variance and 2) an estimate of the element variance.

   d. (TRUE / FALSE) With unequal-sized clusters and a fixed subsample size b within each cluster, the simple unweighted mean can still be used as an unbiased estimate of the population mean.

   e. (TRUE / FALSE) The approximation of the variance when using Taylor Series Linearization is adequate for inference if the coefficient of variation of the sample size is 0.8.

   f. (TRUE / FALSE) In determining the number of clusters selected in stratified cluster sampling, allocation proportionate to the number of elements in a stratum in the population facilitates self-weighting element selection.

g. (TRUE / FALSE) If the elements in a list are approximately continuously ordered and an odd number of units is selected using systematic random sampling, the paired selection model can be used for variance estimation.

h. (TRUE / FALSE) If one selects all possible elements within a cluster that has been selected using simple random sampling (i.e., clusters were selected with equal probability), but the clusters are of unequal sizes, selection probabilities will vary for different population elements.

i. (TRUE / FALSE) The sampling variance under "ultimate cluster" sampling mimics the sampling variance under with replacement selection of primary stage clusters and without replacement subsampling of elements within the clusters.

j. (TRUE / FALSE) Varying the subsample size m in two-stage cluster sampling by design will impact the rate of homogeneity, design effects and thus effective sample sizes.

k. (TRUE / FALSE) In theory, the estimate of sampling variance in two-stage cluster sampling includes the between-cluster and within-cluster sampling.

l. (TRUE / FALSE) A proportionate allocation design is epsem (ignoring rounding error).

5. The Michigan Department of Health has hired you to evaluate design alternatives for a new sample of housing units in the state of Michigan. The objective of the larger study is to estimate the proportion of adults between the ages of 18 and 29 with private health insurance. A two-stage cluster sample is under consideration, featuring the selection of housing units within Census block groups.

   a. Compute the size of a simple random sample of housing units (assuming one respondent per housing unit) needed to obtain a 99% confidence interval with half-width 0.025 for an expected proportion of 0.5. Assume that the degrees of freedom will be larger than 30 for this calculation, and ignore the FPC.

   b. Assuming the ultimate cluster model for a previous sample of 1,000 housing units selected from a sample of 15 block groups, the estimated between-cluster variance in proportions was 0.01. For a proportion estimated to be 0.5, what was the design effect associated with this previous design? What is the estimated value of roh for this indicator of having private health insurance?

   c. Assuming that this design effect will also apply to the new cluster sample design under consideration, how many housing units would you need to select overall in the new cluster sample to meet the same precision objectives from part 5a?

d. The Department of Health suggests that it will cost $5000 per sampled block group and $100 per housing unit for data collection. What would the optimum subsample size per block group be? Given this information and the result in part c), how many block groups would you need to sample to produce an overall sample size that meets the stated precision objectives?

e. Assume that you have the budget to purchase a commercial list of addresses within each selected block group. List two potential problems with the overall sampling frame that you can construct based on the list of block groups and the purchased commercial data. What could you do to solve these problems?

f. The vendor of the commercial data advertises the ability to purchase additional auxiliary variables for the addresses in each selected block group, and you have the budget to do this. What auxiliary information should you request? For what might you use this auxiliary information?

g. For the overall sample, there is budget available for the selection of 30 block groups. Proportionately allocate the block groups to be sampled from each of the five counties (strata) defining the target population below. You should round your answers in a way that maintains the total sample size of 30 block groups.

County	Number of Block Groups	Number of Housing Units
Wayne	250	900,000
Washtenaw	50	180,000
Oakland	150	600,000
Livingston	25	90,000
Genesee	45	200,000

h. What is your proposed sampling error estimation plan for data analysts based on this allocation in Part g?

6. An analyst computes the standard deviation of the final survey weights in a given survey data set as 40, and the mean of those final weights as 100. What is the expected increase in the sampling variance of an estimated mean if these weights are to be used in estimation? Write the correct interpretation of this expected increase in plain English, keeping in mind the assumptions underlying this calculation.