**MSDS 6370 Sampling Statistics Final Exam (part II)**

**Summer 2017**

**This exam is due at midnight CT on Monday, August 14, 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. (15 pts) Select the best answer to each question below: (i)-(iii) 4 points for each.

(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.

(iii) Neyman allocation is

(a) A method of selecting a sample so that units have probabilities proportional to their size

(b) A method of determining how many sample units should be chosen from each stratum so that the sample size is proportional to stratum size

(c) An estimator of the mean that is weighted by the reciprocal of the selection probability

(d) A method of allocating the sample to strata in a stratified design that provides the most advantage when strata variances differ greatly.

(iv) What are the advantages of a probability sample over non-probability sample?

(list 2) (3 points)

Probability samples allow you to use probability based statistical procedures, such as confidence intervals and hypothesis tests in drawing inferences about the population from which the sample was drawn. Nonprobability samples do not.

Probability sampling is less prone to selection bias than nonprobability sampling.

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.

1. Construct a 95% confidence interval for the total gallons of water used during the restricted watering portion of that day for the whole community.
2. 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)? \_larger\_\_\_\_\_. Carefully explain your reasoning.

Clustering by city blocks would increase the sampling error because of correlations among the city block housing units. The housing units within the same city block are not independent from one another so provides less new information.

1. 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, . 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 = 0.4. You decide to use that for planning purposes. How many blocks must they sample?

3. (10 pts) There are 25 farms which belong to 5 clusters called A, B, C, D, and E. You would like to estimate mean area of these 25 farms from a cluster design. You pick 2 clusters at random out of 5.

Specify the sampling distribution of the sample mean area of these farms based on this cluster design. Calculate the mean of the sampling distribution.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | A |  |  |  | B |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | D |  |  |  |  |  |  |  |  | E |  |  |
|  | 1 |  | 6 | 8 | 7 |  |  |  |  | 12 |  |  | C |  |  |  |  | 16 | 17 |  |  | 18 |  |  |  |  |  |  | 25 |  |  |  |
|  | 2 |  |  | 9 |  |  |  |  |  |  |  | 15 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | 24 |  |  |  |  |
|  | 3 |  |  | 10 |  |  |  |  |  |  |  |  |  |  | 14 |  |  |  |  |  |  | 19 |  |  |  |  | 22 |  |  |  |  |  |
|  | 4 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | 20 |  |  |  |  |  |  |  |  |  |  |  |
|  | 5 |  |  |  |  | 13 |  |  |  |  |  |  |  | 11 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | 21 |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | 23 |  |  |  |  |  |  |  |  |  |  |

Mean of sampling distribution = 4.68

Extra information

Cluster A has 1-5, cluster B has 6-10, cluster C has 11-15, cluster D has 16-20, and cluster E has 21-25. Assume that the area of 1×1 squares are the same. In my diagram, you can notice few differences. Please ignore it.

Farm 1 area is 1\*1 =1

Farm 13 area is 2\*8=16

Hint: Sampling distribution steps

1. Identify the sample design
2. Calculate how many possibilities
3. Each possibility calculates the corresponding statistic.
4. You can build the sampling distribution based on all the possible values.

4. (20 points) The finalexam\_parients.csv contains data from sample selected for an audit of a health care provider by using complex design (stratification and cluster). Hospitals are the strata. Patients are the clusters. There are 11426 total patients in these 9 hospitals. They select 64 patients from 9 strata and monitor all the bills from each chosen patient. Here is the table for stratum sizes and selected sample sizes.

|  |  |  |  |
| --- | --- | --- | --- |
| Stratum | Stratum size | Sample size | WT |
| 1 | 12 | 4 |  |
| 2 | 215 | 7 |  |
| 3 | 1248 | 7 |  |
| 4 | 1607 | 7 |  |
| 5 | 58 | 7 |  |
| 6 | 7116 | 8 |  |
| 7 | 335 | 10 |  |
| 8 | 827 | 10 |  |
| 9 | 8 | 4 |  |

1. Calculate weights for each selected patient and update in finalexam\_patients.csv. (You can add another column called WT to include these weights) Please submit your updated finalexam\_patients.csv via 2ds. (10 points)

weights added

1. Calculate the mean estimate for paid variable, its standard error, and the 95% confidence interval. (5 points) (You can use either SAS or R)

**analysis of stratified cluster design with fpc**

**The SURVEYMEANS Procedure**

| **Data Summary** | |
| --- | --- |
| **Number of Strata** | 9 |
| **Number of Clusters** | 64 |
| **Number of Observations** | 64 |
| **Sum of Weights** | 11426 |

| **Statistics** | | | | | |
| --- | --- | --- | --- | --- | --- |
| **Variable** | **N** | **Mean** | **Std Error of Mean** | **95% CL for Mean** | |
| **Paid** | 64 | 3133.467375 | 620.011092 | 1890.93738 | 4375.99737 |

1. Calculate the total estimate for paid variable, its standard error, and the 90% confidence interval. (5 points) (You can use either SAS or R)

**analysis of stratified cluster design with fpc**

**The SURVEYMEANS Procedure**

| **Data Summary** | |
| --- | --- |
| **Number of Strata** | 9 |
| **Number of Clusters** | 64 |
| **Number of Observations** | 64 |
| **Sum of Weights** | 11426 |

| **Statistics** | | | | |
| --- | --- | --- | --- | --- |
| **Variable** | **Sum** | **Std Error of Sum** | **90% CL for Sum** | |
| **Paid** | 35802998 | 7084247 | 23950812.8 | 47655183.6 |