**Homework 9. Selection of a Sample Design**

**MSDS 6370**

**Objective:**

* For the student to learn how to select a sample design to provide a solution to the question being asked.
* For the student to learn more about calculating the probabilities of selection for sample units and their design weights.

**Introduction**

Asynchronous Lecture 8 included a discussion of calculating probabilities of selection and design weights for sample units for use in forming estimates using the sample data. This assignment builds on that discussion and extends it by considering the sample designs a company might use to decide if its employees are complying with its policy for using email.

**Exercise 1**

1. A large manufacturing company has policies against employees using its email system for certain purposes (e.g., non-work communication, ones discussing certain types of proprietary information, etc.). Each email can be classified into one of 4 categories: a non-violation, or one of 3 categories of violation. Though the company has automated monitoring in place, they would like to supplement this for accuracy by sampling the emails and having a human inspector. Going forward, they will produce estimates of the proportion of emails in each category each month based on the sample results. This will allow them to both monitor the policy violation rate over time, and to compare results with their automated monitoring system.

Their sampling plan is the following: They will select 2 working days per month at random and evaluate all emails for those days.

Match each of the following sampling concepts with its realization/type in this application. (Not all of the items in the Realization column will be used.) Enter your answers on the Results page.

|  |  |  |  |
| --- | --- | --- | --- |
| sampling concept |  | Realization |  |
| 1. population | b. | a. All working days in month |  |
| 2. cluster | a. | b. All emails in month |  |
| 3. sampling frame | c. | c. All emails in a day |  |
| 4. parameter of interest | f. | d. Stratified sample |  |
| 5. sample design | e. | e. Cluster sample |  |
|  |  | f. Proportion |  |
|  |  | g. Standard deviation |  |
|  |  | h. Mean |  |

**Exercise 2.**

Refer to Exercise 1 above. The company found that it was too costly to examine all emails in the selected days. Therefore, they decided to sample them. They considered two possible designs.

Design A. Randomly sample 2 working days per month. Select every 10th email (by time) on that day, after selecting one of the first 10 randomly.

Design B. Randomly sample 2 working days per month. Randomly select 100 emails sent on that day.

For each design, answer the following questions.

i. What is the probability of selection for an email in a month with 20 working days on a day with 500 emails sent?

A: 1/10 B: 1/5

ii. What is the weight for an email in a month with 20 working days on a day with 500 emails sent?

A: 10 B: 5

iii. Is the design always an EPSEM (equal probability of selection) design? Why or why not?

A: Yes, the number of days per month will not change for any email in the month and selecting every tenth email doesn’t depend on the number of emails in each day. (unless there were less than 10 emails in a day)

B: No, the probability of selection depends on the number of emails in the day.

**Exercise 3**

**Introduction**

Asynchronous Lecture 9 focused on replication methods of variance estimation for complex survey designs. In particular, the discussion focused on the jackknife and BRR methods of variance estimation. This assignment builds on that discussion and extends it by exploring more options in SAS PROC SURVEYMEANS for implementing variance estimation.

1. In the video this week, you are asked to compute the jackknife and BRR estimates of variance for the toy data we examined.

**Submit the spreadsheet with the completed calculations for homework this week.**

Fill in the last two rows of the table below:

|  |  |
| --- | --- |
| method of estimation for standard error | Standard error estimate |
| Taylor linearization without fpc | 0.009348 |
| Taylor linearization with fpc (as given in video) | 0.007637 |
| SAS computed jackknife estimate | 0.009425 |
| SAS computed BRR estimate | 0.009447 |
| jackknife estimate from spreadsheet calculation | 0.009424724 |
| BRR estimate from spreadsheet calculation | 0.009446672 |