Exercise ark #5.
Sampling with unequal probabilities

Oğuz–Alper, Melike & Pekarskaya, Tatsiana, Statistics Norway

October 28, 2020

Exercise 1

We have a population of 4 companies. A variable of interest is yearly turnover y. Assume that turn over for a given year for the companies 1,2,3,4 is 100, 200, 300 and 1000 millions Norwegian kr. Number employees (x) in each company is known in advance from a register. Assume that x is 20, 30, 50 and 200 for the companies. We are going to samples of size 2 using different methods to estimate the total turnover(as we know the true value is 1600). In (1)-(3) we will find estimators which does not use additional information x. In (4)-(6) we use ratio estimation.
There are three comments to the exercise:

- With sample plan is meant collection of all probabilities $p(s)$ for all possible samples s, i.e. sampling plan indicates all probabilities $p(s)$.
- With standard error (SE) of estimator is meant square root of variance and not as usual, the estimated standard error.
- With mean squared error (MSE) of an estimator $\hat{t}$ which is not biased for total t, is meant $E(\hat{t}-t)^2$. MSE can be calculated as sum of variance and square of the bias: $MSE = Var(\hat{t}) + [E(\hat{t}-t)]^2$

1. **Sampling plan 1.** Company 4 should be included and one more company is sampled from 1,2,3 with probabilities proportional to number of employees:
   - company 1: 0,2
   - company 2: 0,3
   - company 3: 0,5

   Write down the sampling plan. Calculate Horvitz-Thompson(HT) estimator and show that it is unbiased. Calculate standard error(SE) of the estimator.
2. **Sampling plan 2.** Company 4 should be included and one more company is sampled from 1,2,3 with probabilities:
   - company 1: 0,5
   - company 2: 0,3
   - company 3: 0,2

Write down the sampling plan. Calculate HT estimator and show that it is unbiased. Calculate SE of the estimator. If SE will be much larger than in (1), find an estimator without using x which will be more accurate. Calculate bias, SE and $\sqrt{MSE}$ for it.

3. **Sampling plan 3 (SRS).** We sample an SRS of 2 companies. Write down the sampling plan. Calculate HT estimator and show that it is unbiased. Calculate SE of the estimator.

4. Find and expectation of ratio estimator for sampling plan 1 (1). Calculate SE of the ratio estimator. Calculate MSE and $\sqrt{MSE}$.

5. Find and expectation of ratio estimator for sampling plan 2 (2). Calculate SE of the ratio estimator. Calculate MSE and $\sqrt{MSE}$.

6. Find and expectation of ratio estimator for sampling plan 3 (3). Calculate SE of the ratio estimator. Calculate MSE and $\sqrt{MSE}$.

7. **Compare sampling plans and estimators.** If we do not have available information on number of employees, which sampling plan (1)-(3) would you choose? Which sampling plan and estimator would you choose if there is available information on number of employees? (Bjørnstad, 2016, p.64)

Exercise 2

(R code available) The file azcounties.dat gives data from the 2000 U.S. Census on population and housing unit counts for the counties in Arizona (excluding Maricopa County and Pima County, which are much larger than the other counties and would be placed in a separate stratum). For this exercise, suppose that year 2000 population (M_i) is known and you want to take a sample of counties to estimate the total number of housing units ($t = \sum_{i \in U} t_i$). The file has the value of y_i for every county so you can calculate the population total and variance.

1. Calculate the selection probabilities ψ_i for a sample of size 1 with probability proportional to 2000 population. Find $\hat{t}_\psi$ for each possible sample, and calculate the theoretical variance $V(\hat{t}_\psi)$.

2. Repeat (1) for an equal probability sample of size 1. How do the variances compare? Why do you think one design is more efficient than the other? (Lohr, 2019, p.268)

Exercise 3

(R code available) Let's return to the situation in Exercise 3 of Session 2, in which we took an SRS to estimate the average and total numbers of refereed publications of faculty and research associates. Now, consider a probability proportional to size (pps) sample of faculty. The 27 academic units range in size from 2 to 92. We used Lahiri's method to choose 10 (primary sampling units)psus with probabilities proportional to size and with replacement, and took an SRS of four (or fewer, if $M_i < 4$) members from each psu. Note that academic unit 14 appears three times in the sample; each time it appears, a different subsample was collected.

Academic unit	M_i	ψ_i	y_{ij}
14	65	0.0805452	3, 0, 0, 4
23	25	0.0309789	2, 1, 2, 0
9	48	0.0594796	0, 0, 1, 0
14	65	0.0805452	2, 0, 1, 0
16	2	0.0024783	2, 0
6	62	0.0768278	0, 2, 2, 5
14	65	0.0805452	1, 0, 0, 3
19	62	0.0768278	4, 1, 0, 0
21	61	0.0755886	2, 2, 3, 1
11	41	0.0508055	2, 5, 12, 3

Find the estimated total number of publications, along with its standard error. (Lohr, 2019, p.269)

Exercise 4

(R code available) A two-stage unequal probability sample without replacement of size $n = 5$ from the population of statistics classes of size $N = 15$ (see Lohr, 2019, pp.247-248) is taken. The psu inclusion probabilities are proportional to the class sizes M_i. We have $M_0 = \sum_{i \in U} M_i = 647$. The data are in file classpps.dat. A sample of $m_i = 4$ ssus is selected with a simple random sampling without replacement from each sample class. The total number of hours spent studying statistics is of interest. Here, the sampling fraction n/N is 1/3, so the with-replacement variance is likely to overestimate the without-replacement variance. The joint inclusion probabilities for the psus are given in file classppsjp.dat.

1. Calculate $\hat{V}_{HT}(\hat{t}_{HT})$ and $\hat{V}_{SYG}(\hat{t}_{HT})$ for this data set.

2. SAS software approximates the without-replacement variance in unequal-probability sampling using
$$\left(1 - \frac{n}{N}\right)\hat{V}_{WR}(\hat{t}_{HT}).$$
Calculate this approximation for the class data.

3. How do these estimates compare, and how do they compare with the with replacement variance for $\hat{t}_{HT}$? (Lohr, 2019, p.270)