Solution ark #5.
Sampling with unequal probabilities

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

October 26, 2020

Exercise 1

We have a population of 4 companies. A variable of interest is yearly turnover y. Assume that turnover 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.

Solution: Probabilities for samples in the sample plan 1:
$p(\{1,4\}) = 0.2 \quad p(\{2,4\}) = 0.3 \quad p(\{3,4\}) = 0.5$
$p(s) = 0$ for all other cases.

HT estimator where y_i - turnover for company i can be obtained as:

$$\hat{t}_{HT} = \sum_{i \in s} y_i/\pi_i$$

For $s_1 = \{1,4\}$ we get $\hat{t}^{s_1}_{HT} = y_1/\pi_1 + y_4 = 100/0.2 + 1000 = 1500$
For $s_2 = \{2,4\}$ we get $\hat{t}^{s_2}_{HT} = y_2/\pi_2 + y_4 = 200/0.3 + 1000 = 1666,67$
For $s_3 = \{3,4\}$ we get $\hat{t}^{s_3}_{HT} = y_3/\pi_3 + y_4 = 300/0.5 + 1000 = 1600$

$E(\hat{t}_{HT}) = 1500 * 0.2 + 1666.67 * 0.3 + 1600 * 0.5 = 300 + 500 + 800 = 1600 = t \implies$ the estimator is unbiased.
$V(\hat{t}_{HT}) = (1500 - 1600)^2 0.2 + (1666.67 - 1600)^2 0.3 + (1600 - 1600)^2 0.5) = 3333.33$
$SE = \sqrt{3333.33} = 57.7$

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.

Solution: Probabilities for samples in the sample plan 1:
$p(\{1,4\}) = 0.5 \quad p(\{2,4\}) = 0.3 \quad p(\{3,4\}) = 0.2$
$p(s) = 0$ for all other cases.

We use the same formula for HT estimator as in (1).
$\hat{t}^{s_1}_{HT} = y_1/\pi_1 + y_4 = 100/0.5 + 1000 = 1200$
$\hat{t}^{s_2}_{HT} = y_2/\pi_2 + y_4 = 200/0.3 + 1000 = 1666,67$
$\hat{t}^{s_3}_{HT} = y_3/\pi_3 + y_4 = 300/0.2 + 1000 = 2500$

$E(\hat{t}_{HT}) = 1200 * 0.5 + 1666.67 * 0.3 + 2500 * 0.2 = 600 + 500 + 500 = 1600 = t$
$V(\hat{t}_{HT}) = (1200 - 1600)^2 0.5 + (1666.67 - 1600)^2 0.3 + (2500 - 1600)^2 0.2) = 243333.33$
$SE = \sqrt{243333.33} = 493.3$, what is much larger than in (1), thus, we need to find a new estimator without using x.

We have two strata: a stratum with full count of units (company 4) and a stratum where sampling is taking place. We can calculate an estimate for the later stratum based on a sample and summarize it with the total from the earlier stratum. In this case the estimate for stratum $s = \{i, 4\}$ will be calculated as $\hat{t} = 3y_i + y_4$. Then the values for ratio estimators and probabilities are:

s	$\{1,4\}$	$\{2,4\}$	$\{3,4\}$
$\hat{t}_R$	1300	1600	1900
p(s)	0.5	0.3	0.2

$E(\hat{t}_R) = 1300 * 0.5 + 1600 * 0.3 + 1900 * 0.2 = 1510$
$V(\hat{t}_R) = (1300 - 1510)^2 * 0.5 + (1600 - 1510)^2 * 0.3 + (1900 - 1510)^2 * 0.2 = 54900$
$SE = 234.3$
$MSE = 549000 + 90^2 = 63000$ and $RMSE = \sqrt{MSE} = 251.0$

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.

Solution: There are possible 6 different samples of 2 companies from 4: $\{i,j\} = \{1,2\},\{1,3\},\{1,4\},\{2,3\},\{2,4\}$ and $\{3,4\}$. Each sample has a probability $p(\{i,j\}) = 1/6$.

All sampling probabilities are $n/N = 1/2$, and then

$$\hat{t}_{HT} = \sum_{i \in s} y_/\pi_i = 2 \sum_{i \in s} y_i = 4\bar{y}_s$$

$E(\hat{t}_{HT}) = 1/6 * 2(100 * 3 + 200 * 3 + 400 * 3 + 1000 * 3) = 1600$

$V(4\bar{y}_s) = E(4\bar{y}_s - 1600)^2 = \sum_{s:|s|=2}(4\bar{y}_s - 1600)^2 \frac{1}{6} =$
$= \frac{1}{6}\big[(600 - 1600)^2 + (800 - 1600)^2 + (2200 - 1600)^2 + (1000 - 1600)^2 +$
$+ (2400 - 1600)^2 + (2600 - 1600)^2\big] = 666666.667$

$\implies SE = 816.5$

NO: For this case we could also calculate SE based on formula for the variance:
$V(N\bar{y}_s) = N^2 \frac{s^2}{n}(1 - \frac{n}{N})$, where $s^2 = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - \bar{y})^2$

By using the formula we get: $s^2 = (300^2 + 200^2 + 100^2 + 600^2)/3 = 500000/3$ and
$V(4\bar{y}_s) = 4^2 \frac{500000/3}{2}(1 - \frac{1}{2}) = 4 * 500000/3 = 666666.667$.

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

Solution: The total number of employees is $20 + 30 + 50 + 200 = 300$, then ratio estimation is

$$\hat{t}_R = 300 \frac{\sum_s y_i}{\sum_s x_i}.$$

The values for ratio estimators and probabilities are:

s	$\{1,4\}$	$\{2,4\}$	$\{3,4\}$
$\hat{t}_R$	1500	1565.22	1560
p(s)	0.2	0.3	0.5

Expectation:
$E(\hat{t}_R) = 1500 * 0.2 + 1565.22 * 0.2 + 1560 * 0.5 = 300 + 469.6 + 780 = 1549.6$
$V(\hat{t}_R) = E(\hat{t}_R - 1549.6)^2 = 49.6^2 * 0.2 + 15.62^2 * 0.3 + 10.4^2 * 0.5 = 619.31$
$SE = \sqrt{619.31} = 24.9$
$MSE = 619.31 + 0.4^2 = 3159.5$ and $RMSE = \sqrt{MSE} = \sqrt{3159.5} = 56.2$
RMSE can be compared with SE for unbiased estimators.

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

Solution:
The values for ratio estimators and probabilities are:

s	$\{1,4\}$	$\{2,4\}$	$\{3,4\}$
$\hat{t}_R$	1500	1565.22	1560
p(s)	0.5	0.3	0.2

$E(\hat{t}_R) = 1531.6 \quad V(\hat{t}_R) = 999.68 \quad SE = 31.6$
$MSE = 5678.24 \quad RMSE = 75.4$

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

Solution:
The values for ratio estimators, each with probability 1/6 are:

s	$\{1,2\}$	$\{1,3\}$	$\{1,4\}$	$\{2,3\}$	$\{2,4\}$	$\{3,4\}$
$\hat{t}_R$	1800	1714.29	1500	1875	1565.22	1560

$E(\hat{t}_R) = 1669.1 \quad V(\hat{t}_R) = 18810.07 \quad SE = 137.2$
$MSE = 23584.88 \quad RMSE = 153.6$

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?

Solution: If we do not have available information on number of employees:
All HT estimator were unbiased with the standard errors:

- Utvalgsplan 1: 57.7
- Utvalgsplan 2: 493.3 (Alternativ estimator: SE = 234.2 with bias = -90 and RMSE = 251.0)
- Utvalgsplan 3: 816.5

A superior sample plan is (1): since has a higher sampling probability for bigger companies and the biggest company is necessary included.
If we do have available information on number of employees: All estimators and standard error are:

- Utvalgsplan 1 and HT-estimator: SE = 57.7, bias = 0
- Utvalgsplan 2 and HT-estimator: SE = 493.3, bias = 0 (Alternativ estimator: SE = 234.2 with bias = -90 and RMSE = 251.0)
- Utvalgsplan 3 and HT-estimator: SE = 816.5, bias = 0
- Utvalgsplan 1 and ratio estimation: SE = 24.9 and bias = -50.4; RMSE = 56.2
- Utvalgsplan 2 and ratio estimation: SE = 31.6 and bias = -68.4; RMSE = 75.4
- Utvalgsplan 3 and ratio estimation: SE = 137.2 and bias = +69.1; RMSE = 153.6

Looking at the estimator we can make the next conclusions:

- Firstly, we can notice that HT-estimator is working poorly for the cases when the sampling probability is negatively correlated with y values (See at (2)). In (2) SE is 8,5 times larger than SE in (1). Ratio estimator is much more robust. RMSE of sampling plan 2 (See at (5)) is just 1.3 times larger than (1).
- It is not reasonable to take a SRS in companies' surveys when there is a big variation of y values.
- It will be unlucky if we would choose as sampling plan with an estimator defined in (2) not looking at the fact that it is unbiased. RMSE for ratio estimation in (5) which is just 15% of SE in (2). If we do not have available additional information x if would be better to use as an estimator $\hat{t} = 3y_1 + y_4$, not looking at the fact that it is biased.
- It is not so easy to conclude here which estimator to choose. Sampling plan 1 is still a superior of the three samplings plans. Its ratio estimation (4) has a much lower SE than it's HT estimation (1), but at the same moment the bias in (4) is so big that RMSE in (4) is almost the same as SE in (1). For populations and samples with usual sizes, a bias in ratio estimation will be considerably smaller so that the ratio estimator will be preferred.

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 $V(\hat{t}_\psi)$.

Solution: The 2000 U.S. Census on population and housing unit counts for the counties in Arizona are provided. We have $n = 1$, $N = 13$ and $t = \sum_{i \in U} t_i = 572\,221$.
Select all possible samples with unequal-probabilities $\psi_i \propto M_i$, where the M_i are 2000 population. Find $\hat{t}_\psi$ for each sample, and calculate the theoretical variance $V(\hat{t}_\psi)$.

$$\hat{t}_\psi = \sum_{i \in S} \frac{t_i}{\psi_i}, \quad \psi_i = \frac{M_i}{\sum_{i=1}^{13} M_i},$$

$$V(\hat{t}_\psi) = E[(\hat{t}_\psi - t)^2] = \sum_{samples\, i} \psi_i (\hat{t}_\psi - t)^2.$$

County	t_i	ψ_i	$\hat{t}_\psi$	$\psi_i(\hat{t}_\psi - t)^2$
1	31 621	0.0572	553 292.1	20 477 223
2	51 126	0.0969	527 405.6	194 693 778
3	53 443	0.0958	558 108.6	19 071 085
4	28 189	0.0423	667 034.6	379 902 891
5	11 430	0.0276	414 597.1	684 957 758
6	3 744	0.0070	532 113.6	11 318 255
7	15 133	0.0162	932 417.7	2 105 687 838
8	80 062	0.1276	627 317.4	387 423 351
9	47 413	0.0802	590 892.8	27 974 547
10	81 154	0.1480	548 502.8	83 232 656
11	13 036	0.0316	412 582.0	805 214 838
12	81 730	0.1379	592 659.0	57 604 012
13	74 140	0.1317	562 787.3	11 723 897
Sum	572 221	1.0000		4 789 282 131

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?

Solution: Now select an SRS of size $n = 1$ and repeat the calculations in (1).
Under SRS with $n = 1$, we have $\psi_i = 1/13$ for each county. Each possible sample has also the same probability of selection.

County	t_i	ψ_i	$\hat{t}_{srs}$	$(\hat{t}_{srs} - t)^2/13$
1	31 621	0.0769	411 073	1 997 590 608
2	51 126	0.0769	664 638	656 992 453
3	53 443	0.0769	694 759	1 155 043 188
4	28 189	0.0769	366 457	3 256 832 592
5	11 430	0.0769	14 859	13 804 863 397
6	3 744	0.0769	48 672	21 084 888 877
7	15 133	0.0769	196 729	10 845 710 928
8	80 062	0.0769	1 040 806	16 890 146 325
9	47 413	0.0769	616 369	149 926 608
10	81 154	0.0769	1 055 002	17 929 037 997
11	13 036	0.0769	169 468	12 477 690 693
12	81 730	0.0769	1 062 490	18 489 514 797
13	74 140	0.0769	963 820	11 796 136 677
Sum	572 221	1.0000		1.30534×10^{11}

The design with unequal-probabilities is more efficient than the SRS. Because, population and housing unit counts are highly correlated. We have $\rho = 0.9905$. This leads to that the values of $\hat{t}_\psi = t_i/\psi_i$ do not vary so much as those under SRS from sample to sample.

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.

Solution: Academic units are selected with a pps sampling with replacement. $N = 27$ and $n = 10$. Find $\hat{t}$ and its standard error.

$$\hat{t} = \frac{1}{R}\sum_{i\in R}\frac{\hat{t}_i}{\psi_i}, \quad \hat{t}_i = \sum_{j\in R_i}\frac{M_i}{m_i}y_{ij},$$

$$\hat{V}(\hat{t}) = \frac{1}{n}\frac{1}{n-1}\sum_{i\in R}\Big(\frac{\hat{t}_i}{\psi_i} - \hat{t}\Big)^2,$$

where R denotes the set of n units in the sample, including the repeats.
$\hat{t} = 1\,371.90$,
$\widehat{SE}(\hat{t}) = 1\,179.47/\sqrt{10} = 372.98$.

psu	M_i	ψ_i	y_{ij}	$\hat{t}_i$	$\hat{t}_i/\psi_i$
14	65	0.0805452	3, 0, 0, 4	113.75	1 412.25
23	25	0.0309789	2, 1, 2, 0	31.25	1 008.75
9	48	0.0594796	0, 0, 1, 0	12.00	201.75
14	65	0.0805452	2, 0, 1, 0	48.75	605.25
16	2	0.0024783	2, 0	2.00	807.00
6	62	0.0768278	0, 2, 2, 5	139.50	1 815.75
14	65	0.0805452	1, 0, 0, 3	65.00	807.00
19	62	0.0768278	4, 1, 0, 0	77.50	1 008.75
21	61	0.0755886	2, 2, 3, 1	122.00	1 614.00
11	41	0.0508055	2, 5, 12, 3	225.50	4 438.50
Mean					1 371.90
SD					1 179.47

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 dataset.

Solution: A two-stage sample with unequal probabilities WOR. $N = 15$, $n = 5$, $m_i = 4$, and $M_0 = 647$. We have $\pi_i \propto M_i$, and $\pi_{j|i} = 4/M_i$, for $j \in S_i$. An SRSWOR is used to select the SSUs.

$$\hat{t}_{HT} = \sum_{i \in S} \frac{\hat{t}_i}{\pi_i}, \quad \hat{t}_i = \sum_{j \in S_i} \frac{y_{ij}}{\pi_{j|i}}\cdot$$

$$\hat{V}(\hat{t}_{HT}) = \hat{V}_{psu} + \hat{V}_{ssu} = \hat{V}_{psu} + \sum_{i \in S} \frac{\hat{V}(\hat{t}_i)}{\pi_i}$$

$$\hat{V}(\hat{t}_i) = M_i^2 \Big(1 - \frac{4}{M_i}\Big) \frac{s_i^2}{4}, \quad s_i^2 = \frac{1}{3} \sum_{j \in S_i} (y_{ij} - \bar{y}_i)^2\cdot$$

class	π_i	M_i	$\hat{t}_i$	s_i^2	$\hat{V}(\hat{t}_i)$
4	0.17002	22	110.00	0.1666667	16.5000
10	0.26275	34	106.25	0.7291667	185.9375
1	0.34003	44	154.00	1.6666667	733.3333
9	0.41731	54	195.75	4.2291667	2 854.6875
14	0.77280	100	200.00	0.5000000	1 200.0000

$$\sum_{i \in S} \frac{\hat{V}(\hat{t}_i)}{\pi_i} = \frac{16.5}{0.17002} + \frac{185.9}{0.26275} + \cdots + \frac{1\,200}{0.77280} = 11\,354.9\cdot$$

- The first term of $\hat{V}(\hat{t}_{HT})$ can be calculated by using either the Horvitz-Thompson (HT) estimator or the Sen-Yates-Grundy (SYG) estimator.

$$\hat{V}_{psu;HT} = \sum_{i \in S} \sum_{k \in S} \frac{(\pi_{ik} - \pi_i \pi_k)}{\pi_{ik}} \frac{\hat{t}_i}{\pi_i} \frac{\hat{t}_k}{\pi_k} = 6\,059.6\cdot$$

$$\hat{V}_{psu;SYG} = -\frac{1}{2} \sum_{i \in S} \sum_{k \neq i \in S} \frac{(\pi_{ik} - \pi_i \pi_k)}{\pi_{ik}} \Big(\frac{\hat{t}_i}{\pi_i} - \frac{\hat{t}_k}{\pi_k}\Big)^2 = 54\,784.5\cdot$$

- $\hat{V}_{HT}(\hat{t}_{HT}) = 6\,059.6 + 11\,354.9 = 17\,414.46$
- $\hat{V}_{SYG}(\hat{t}_{HT}) = 54\,784.5 + 11\,354.9 = 66\,139.41$

$\Longrightarrow$ Because of the small sample size, i.e. $n = 5$, these estimators are quite unstable. This is the reason of the big difference between the two estimates.

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

Solution: Use the fpc-adjusted with-replacement variance estimator to estimate $V(\hat{t}_{HT})$.

$$\hat{t}_{HT} = \sum_{i \in S} \frac{\hat{t}_i}{\pi_i} = \frac{110}{0.17002} + \frac{106.25}{0.26275} + \cdots + \frac{200}{0.77280} = 2\,232.15\cdot$$

$$\begin{aligned}
\hat{V}(\hat{t}_{HT}) &= (1 - \frac{n}{N})\hat{V}_{WR}(\hat{t}_{HT}) \\
&= \left(1 - \frac{n}{N}\right)\frac{n}{n-1}\sum_{i \in S}\left(\frac{\hat{t}_i}{\pi_i} - \frac{\hat{t}_{HT}}{n}\right)^2 \\
&= \left(1 - \frac{5}{15}\right)\frac{5}{4}\left[\left(\frac{110}{0.17002} - \frac{2\,232.15}{5}\right)^2 + \cdots\right] \\
&= \frac{2}{3}(97\,187.37) = 64\,791.58
\end{aligned}$$

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

Solution: Comparison of the estimates.

variance estimator	estimate of $V(\hat{t}_{HT})$	standard error
$\hat{V}_{HT}(\hat{t}_{HT})$	17 414.46	131.96
$\hat{V}_{SYG}(\hat{t}_{HT})$	66 139.41	257.18
$\hat{V}_{WR}(\hat{t}_{HT})$	97 187.37	311.75
$(1 - n/N)\hat{V}_{WR}(\hat{t}_{HT})$	64 791.58	254.54

- $\hat{V}_{HT}(\hat{t}_{HT}) < (1 - n/N)\hat{V}_{WR}(\hat{t}_{HT}) < \hat{V}_{SYG}(\hat{t}_{HT}) < \hat{V}_{WR}(\hat{t}_{HT})$
- Here, we have $\widehat{SE}_{SYG}(\hat{t}_{HT}) \approx (\sqrt{1 - n/N})\,\widehat{SE}_{WR}(\hat{t}_{HT})$