An Introduction to Survey Sampling

Stanisław Jaworski, Wojciech Zieliński

Department of Econometrics and Statistics
e-mail: stanislaw_jaworski@sggw.pl
e-mail: wojciech_zielinski@sggw.pl
http://wojtek.zielinski.statystyka.info

Table of Contents

Survey sampling

Subject

In statistics, survey sampling describes the process of selecting a sample of elements from a target population to conduct a survey. The term "survey" may refer to many different types or techniques of observation. In survey sampling it most often involves a questionnaire used to measure the characteristics and/or attitudes of people (https://en.wikipedia.org/wiki/Survey_sampling).

Bibliography

Bibliography

- Ravindra Singh, Naurang Singh Mangat (1996), Elements of Survey Sampling, Originally published by Kluwer Academic Publishers in 1996, Springer Science+Business Media Dordrecht.
- Thompson M.E. (1997), Theory of Sample Surveys, Originally published by Chapman & Hall in 1997, Springer-Science+Business Media, B.Y.

Basic terms

Basic terms

- An element is a unit for which information is sought.

Basic terms

Basic terms

- An element is a unit for which information is sought.
- The population or universe is an aggregate of elements, about which the inference is to be made.

Basic terms

Basic terms

- An element is a unit for which information is sought.
- The population or universe is an aggregate of elements, about which the inference is to be made.
- Sampling units are non overlapping collections of elements of the population.

Basic terms

Basic terms

- An element is a unit for which information is sought.
- The population or universe is an aggregate of elements, about which the inference is to be made.
- Sampling units are non overlapping collections of elements of the population.
- A list of all the units in the population to be sampled is termed frame or sampling frame.

Basic terms

Basic terms

- An element is a unit for which information is sought.
- The population or universe is an aggregate of elements, about which the inference is to be made.
- Sampling units are non overlapping collections of elements of the population.
- A list of all the units in the population to be sampled is termed frame or sampling frame.
- A subset of population selected from a frame to draw inferences about a population characteristic is called a sample.

Basic terms

Basic terms

- Collection of information on every unit in the population for the characteristics of interest is known as complete enumeration or census.

Basic terms

Basic terms

- Collection of information on every unit in the population for the characteristics of interest is known as complete enumeration or census.
- The number of units (not necessarily distinct) included in the sample is known as the sample size and is usually denoted by n, whereas the number of units in the population is called population size and is denoted by N. The ratio n/N is termed as sampling fraction.

Basic terms

Basic terms

- Collection of information on every unit in the population for the characteristics of interest is known as complete enumeration or census.
- The number of units (not necessarily distinct) included in the sample is known as the sample size and is usually denoted by n, whereas the number of units in the population is called population size and is denoted by N. The ratio n/N is termed as sampling fraction.
- The method which is used to select the sample from a population is known as sampling procedure.

Basic terms

Basic terms

- If the units in the sample are selected using some probability mechanism, such a procedure is called probability sampling.

Basic terms

Basic terms

- If the units in the sample are selected using some probability mechanism, such a procedure is called probability sampling.
- The procedure of selecting a sample without using any probability mechanism is termed as nonprobability sampling.

Basic terms

Basic terms

- In with replacement (WR) sampling, the units are drawn one by one from the population, replacing the unit selected at any particular draw before executing the next draw.

Basic terms

Basic terms

- In with replacement (WR) sampling, the units are drawn one by one from the population, replacing the unit selected at any particular draw before executing the next draw.
- In without replacement (WOR) sampling, the units are selected one by one from the population, and the unit selected at any particular draw is not replaced back to the population before selecting a unit at the next draw.

Basic terms

Exercise 1

The weights (in pounds) of four children at the time of birth in a hospital:

Child	A	B	C	D
Weight	5.5	8.0	6.5	7.0

Basic terms

Exercise 1

The weights (in pounds) of four children at the time of birth in a hospital:

Child	A	B	C	D
Weight	5.5	8.0	6.5	7.0

1. Enumerate all possible WR samples of size 2. Write down values of the study variable for the sample units.

Basic terms

Exercise 1

The weights (in pounds) of four children at the time of birth in a hospital:

Child	A	B	C	D
Weight	5.5	8.0	6.5	7.0

1. Enumerate all possible WR samples of size 2. Write down values of the study variable for the sample units.
2. Enumerate all possible WOR samples of size 2, and also list the weight values for the respective sample units.

Basic terms

Exercise 2

The areas (in hectares) of six villages are given below:

Village	A	B	C	D	E	F
Area	760	343	657	550	480	935

Basic terms

Exercise 2

The areas (in hectares) of six villages are given below:

Village	A	B	C	D	E	F
Area	760	343	657	550	480	935

1. Enumerate all WR samples of size 3. Write down the values of the study variable for the sampled units.

Basic terms

Exercise 2

The areas (in hectares) of six villages are given below:

Village	A	B	C	D	E	F
Area	760	343	657	550	480	935

1. Enumerate all WR samples of size 3. Write down the values of the study variable for the sampled units.
2. List all the WOR samples of size 4 along with their area values.

Population

Basic terms

- Population $\mathcal{U} = \{u_1, u_2, \ldots, u_N\}$

Population

Basic terms

- Population $\mathcal{U} = \{u_1, u_2, \ldots, u_N\}$
- Variable $Y : \mathcal{U} \to R$

Population

Basic terms

- Population $\mathcal{U} = \{u_1, u_2, \ldots, u_N\}$
- Variable $Y : \mathcal{U} \to R$
- Population parameter $\mathcal{Y} = \{Y_1, Y_2, \ldots, Y_N\}$

Population

Basic terms

- Population $\mathcal{U} = \{u_1, u_2, \ldots, u_N\}$
- Variable $Y : \mathcal{U} \to R$
- Population parameter $\mathcal{Y} = \{Y_1, Y_2, \ldots, Y_N\}$
- Parametric function $T : \mathcal{Y} \to R$

Population

Parametric functions (examples)

Population

Parametric functions (examples)

- Total value $Y = \sum_{i=1}^{N} Y_i$

Population

Parametric functions (examples)

- Total value $Y = \sum_{i=1}^{N} Y_i$
- Mean value $\bar{Y} = \frac{1}{N} \sum_{i=1}^{N} Y_i$

Population

Parametric functions (examples)

- Total value $Y = \sum_{i=1}^{N} Y_i$
- Mean value $\bar{Y} = \frac{1}{N} \sum_{i=1}^{N} Y_i$
- Variance $\sigma^2 = \frac{1}{N-1} \sum_{i=1}^{N} (Y_i - \bar{Y})^2$

Population

Parametric functions (examples)

- Total value $Y = \sum_{i=1}^{N} Y_i$
- Mean value $\bar{Y} = \frac{1}{N} \sum_{i=1}^{N} Y_i$
- Variance $\sigma^2 = \frac{1}{N-1} \sum_{i=1}^{N} (Y_i - \bar{Y})^2$
- Maximal value $\max\{Y_i : i =, \ldots, N\}$

Population parameters

Parameters

- Any real valued function of variable values for all the population units is known as a population parameter or simply a parameter.

  Let $Y_1, Y_2, \ldots, Y_N$ be the values of the variable Y for the N units in the population.

$$\text{population mean: } \mu = \bar{Y} = \frac{1}{N}\sum_{i=1}^{N} Y_i$$

$$\text{population variance: } \sigma^2 = \frac{1}{N-1}\sum_{i=1}^{N}(Y_i - \bar{Y})^2$$

Statistics

Estimators

- A real valued function of variable values for the units in the sample is called a statistic. If it is used to estimate a parameter, it is termed as estimator.

  Let $y_1, y_2, \ldots, y_n$ be the values of the variable Y for the n units in the sample.

$$\text{sample mean: } \bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$$

$$\text{sample variance: } \hat{\sigma}^2 = s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})^2$$

Sampling design

Sampling design

We consider samples to be subsets s of $\mathcal{U} = \{1, \ldots, N\}$ and denote by $\mathcal{S}$ the collection of all subsets s of $\mathcal{U}$. A sampling design (or probability sampling design or a randomized sampling design) is formally a probability function on $\mathcal{S}$. With each sample s a probability $p(s)$ of being drawn is associated. Each $p(s)$ is a number in $[0, 1]$ and

$$\sum_{s \in \mathcal{S}} p(s) = 1.$$

Sampling design

Example 1

(A,B)	0.1667	0.1195	0.0844	0.2474	0.1355
(A,C)	0.1667	0.0066	0.1718	0.1933	0.1865
(A,D)	0.1667	0.1166	0.2186	0.0830	0.2826
(B,C)	0.1667	0.1951	0.1500	0.2824	0.0261
(B,D)	0.1667	0.2983	0.1046	0.0660	0.1231
(C,D)	0.1667	0.2639	0.2707	0.1280	0.2461

Sampling design

Inclusion probability

The inclusion probability of unit j is the probability that the unit j appears in the sample drawn:

$$\pi_j = \sum_{s:\ j \in s} p(s)$$

Sampling design

Example 1 cont

(A,B)	0.1667	0.1195	0.0844	0.2474	0.1355
(A,C)	0.1667	0.0066	0.1718	0.1933	0.1865
(A,D)	0.1667	0.1166	0.2186	0.0830	0.2826
(B,C)	0.1667	0.1951	0.1500	0.2824	0.0261
(B,D)	0.1667	0.2983	0.1046	0.0660	0.1231
(C,D)	0.1667	0.2639	0.2707	0.1280	0.2461

Sampling design

Example 1 cont

(A,B)	0.1667	0.1195	0.0844	0.2474	0.1355
(A,C)	0.1667	0.0066	0.1718	0.1933	0.1865
(A,D)	0.1667	0.1166	0.2186	0.0830	0.2826
(B,C)	0.1667	0.1951	0.1500	0.2824	0.0261
(B,D)	0.1667	0.2983	0.1046	0.0660	0.1231
(C,D)	0.1667	0.2639	0.2707	0.1280	0.2461

π_A	0.5000	0.2427	0.4748	0.5237	0.6046
π_B	0.5000	0.6129	0.3389	0.5958	0.2848
π_C	0.5000	0.4655	0.5925	0.6036	0.4587
π_D	0.5000	0.6788	0.5938	0.2769	0.6518

Sampling design

Joint inclusion probability

The joint inclusion probabilities of distinct units j and k is the probability that both units j and k appear in the sample:

$$\pi_{jk} = \sum_{s:\ j,k \in s} p(s)$$

Sampling design

Sampling design

$z_1, \ldots, z_N$: values of some variate Z

Sampling design

Sampling design

$z_1, \ldots, z_N$: values of some variate Z

E_p: expectation with respect to the sampling design p

Sampling design

Sampling design

$z_1, \ldots, z_N$: values of some variate Z
E_p: expectation with respect to the sampling design p
For the sample s the sample sum of Z equals $\sum_{j \in s} z_j$

$$E_p\left(\sum_{j \in s} z_j\right) = E_p\left(\sum_{j=1}^{N} z_j \mathbf{1}(j \in s)\right)$$

$$= \sum_{j=1}^{N} z_j E_p \mathbf{1}(j \in s) = \sum_{j=1}^{N} z_j \pi_j$$

Sampling design

Sampling design

- If $z_j \equiv 1$ than $\sum_{j \in s} z_j = n(s)$ is the sample size $n(s)$ and

$$E_p(n(s)) = \sum_{j=1}^{N} \pi_j$$

Sampling design

Sampling design

- If $z_j \equiv 1$ than $\sum_{j \in s} z_j = n(s)$ is the sample size $n(s)$ and

$$E_p(n(s)) = \sum_{j=1}^{N} \pi_j$$

- For a fixed size n design, the inclusion probabilities will sum up to n

$$E_p n = n = \sum_{j=1}^{N} \pi_j$$

Sampling design

Example 1 cont

(A,B)	0.1667	0.1195	0.0844	0.2474	0.1355
(A,C)	0.1667	0.0066	0.1718	0.1933	0.1865
(A,D)	0.1667	0.1166	0.2186	0.0830	0.2826
(B,C)	0.1667	0.1951	0.1500	0.2824	0.0261
(B,D)	0.1667	0.2983	0.1046	0.0660	0.1231
(C,D)	0.1667	0.2639	0.2707	0.1280	0.2461

Sampling design

Example 1 cont

(A,B)	0.1667	0.1195	0.0844	0.2474	0.1355
(A,C)	0.1667	0.0066	0.1718	0.1933	0.1865
(A,D)	0.1667	0.1166	0.2186	0.0830	0.2826
(B,C)	0.1667	0.1951	0.1500	0.2824	0.0261
(B,D)	0.1667	0.2983	0.1046	0.0660	0.1231
(C,D)	0.1667	0.2639	0.2707	0.1280	0.2461

π_A	0.5000	0.2427	0.4748	0.5237	0.6046
π_B	0.5000	0.6129	0.3389	0.5958	0.2848
π_C	0.5000	0.4655	0.5925	0.6036	0.4587
π_D	0.5000	0.6788	0.5938	0.2769	0.6518

Statistics

Sampling distribution

- For a given population, sampling procedure and sample size, the array of possible values of an estimator each with its probability of occurrence, is the sampling distribution of that estimator.

Sampling design

Example 1 cont

(A,B)	0.1667	0.1195	0.0844	0.2474	0.1355
(A,C)	0.1667	0.0066	0.1718	0.1933	0.1865
(A,D)	0.1667	0.1166	0.2186	0.0830	0.2826
(B,C)	0.1667	0.1951	0.1500	0.2824	0.0261
(B,D)	0.1667	0.2983	0.1046	0.0660	0.1231
(C,D)	0.1667	0.2639	0.2707	0.1280	0.2461

Sampling design

Example 1 cont

sample mean					
6.75	0.1667	0.1195	0.0844	0.2474	0.1355
6	0.1667	0.0066	0.1718	0.1933	0.1865
6.25	0.1667	0.1166	0.2186	0.0830	0.2826
7.25	0.1667	0.1951	0.1500	0.2824	0.0261
7.5	0.1667	0.2983	0.1046	0.0660	0.1231
6.75	0.1667	0.2639	0.2707	0.1280	0.2461

Sampling design

Example 1 cont

sample mean					
6.75	0.1667	0.1195	0.0844	0.2474	0.1355
6	0.1667	0.0066	0.1718	0.1933	0.1865
6.25	0.1667	0.1166	0.2186	0.0830	0.2826
7.25	0.1667	0.1951	0.1500	0.2824	0.0261
7.5	0.1667	0.2983	0.1046	0.0660	0.1231
6.75	0.1667	0.2639	0.2707	0.1280	0.2461

population mean: 6.75

Statistics

Sampling distribution

- The resultant discrepancy between the sample estimate and the population parameter value is the error of the estimate. Such an error is termed sampling error. If θ is the population parameter and $\hat{\theta}$ is its estimator then $\hat{\theta} - \theta$ is the sampling error.

Sampling design

Example 1 cont

error					
0	0.1667	0.1195	0.0844	0.2474	0.1355
-0.75	0.1667	0.0066	0.1718	0.1933	0.1865
-0.5	0.1667	0.1166	0.2186	0.0830	0.2826
0.5	0.1667	0.1951	0.1500	0.2824	0.0261
0.75	0.1667	0.2983	0.1046	0.0660	0.1231
0	0.1667	0.2639	0.2707	0.1280	0.2461

Statistics

Main properties

- The estimator $\hat{\theta}$ is unbiased for the parameter θ if

$$E_p\hat{\theta} = \theta$$

Statistics

Main properties

- The estimator $\hat{\theta}$ is unbiased for the parameter θ if

$$E_p\hat{\theta} = \theta$$

- If $E_p(\hat{\theta}) \neq \theta$ the estimator $\hat{\theta}$ is a biased estimator of θ. The bias of $\hat{\theta}$ equals

$$B_p(\hat{\theta}) = E_p(\hat{\theta}) - \theta$$

Sampling design

Example 1 cont

6.75	0.1667	0.1195	0.0844	0.2474	0.1355
6	0.1667	0.0066	0.1718	0.1933	0.1865
6.25	0.1667	0.1166	0.2186	0.0830	0.2826
7.25	0.1667	0.1951	0.1500	0.2824	0.0261
7.5	0.1667	0.2983	0.1046	0.0660	0.1231
6.75	0.1667	0.2639	0.2707	0.1280	0.2461

Sampling design

Example 1 cont

6.75	0.1667	0.1195	0.0844	0.2474	0.1355
6	0.1667	0.0066	0.1718	0.1933	0.1865
6.25	0.1667	0.1166	0.2186	0.0830	0.2826
7.25	0.1667	0.1951	0.1500	0.2824	0.0261
7.5	0.1667	0.2983	0.1046	0.0660	0.1231
6.75	0.1667	0.2639	0.2707	0.1280	0.2461
E_p	6.75	7.008	6.6653	6.7543	6.5743

Sampling design

Example 1 cont

6.75	0.1667	0.1195	0.0844	0.2474	0.1355
6	0.1667	0.0066	0.1718	0.1933	0.1865
6.25	0.1667	0.1166	0.2186	0.0830	0.2826
7.25	0.1667	0.1951	0.1500	0.2824	0.0261
7.5	0.1667	0.2983	0.1046	0.0660	0.1231
6.75	0.1667	0.2639	0.2707	0.1280	0.2461
E_p	6.75	7.008	6.6653	6.7543	6.5743
B_p	0	0.258	-0.0847	0.0043	-0.1757

Statistics

Main properties

- The sampling variance is the variance of the sampling distribution of the estimator $\hat{\theta}$:

$$Var_p(\hat{\theta}) = E_p(\hat{\theta} - E\hat{\theta})^2$$

Statistics

Main properties

- The sampling variance is the variance of the sampling distribution of the estimator $\hat{\theta}$:

$$Var_p(\hat{\theta}) = E_p(\hat{\theta} - E\hat{\theta})^2$$

- The mean square error (MSE) measures the divergence of the estimator values from the true parameter value:

$$MSE_p(\hat{\theta}) = E_p(\hat{\theta} - \theta)^2$$

Sampling design

Example 1 cont

6.75	0.1667	0.1195	0.0844	0.2474	0.1355
6	0.1667	0.0066	0.1718	0.1933	0.1865
6.25	0.1667	0.1166	0.2186	0.0830	0.2826
7.25	0.1667	0.1951	0.1500	0.2824	0.0261
7.5	0.1667	0.2983	0.1046	0.0660	0.1231
6.75	0.1667	0.2639	0.2707	0.1280	0.2461

Sampling design

Example 1 cont

6.75	0.1667	0.1195	0.0844	0.2474	0.1355
6	0.1667	0.0066	0.1718	0.1933	0.1865
6.25	0.1667	0.1166	0.2186	0.0830	0.2826
7.25	0.1667	0.1951	0.1500	0.2824	0.0261
7.5	0.1667	0.2983	0.1046	0.0660	0.1231
6.75	0.1667	0.2639	0.2707	0.1280	0.2461
Var_p	0.2708	0.2495	0.2476	0.2372	0.2514

Sampling design

Example 1 cont

6.75	0.1667	0.1195	0.0844	0.2474	0.1355
6	0.1667	0.0066	0.1718	0.1933	0.1865
6.25	0.1667	0.1166	0.2186	0.0830	0.2826
7.25	0.1667	0.1951	0.1500	0.2824	0.0261
7.5	0.1667	0.2983	0.1046	0.0660	0.1231
6.75	0.1667	0.2639	0.2707	0.1280	0.2461
Var_p	0.2708	0.2495	0.2476	0.2372	0.2514
MSE_p	0.2708	0.3161	0.2548	0.2372	0.2823

Statistics

Main properties

- Let $\hat{\theta}_1$ and $\hat{\theta}_2$ be two estimators of the parameter θ. The relative efficiency of the estimator $\hat{\theta}_2$ with respect to the estimator $\hat{\theta}_1$ is defined as

$$RE\left(\hat{\theta}_2|\hat{\theta}_1\right) = \frac{MSE(\hat{\theta}_1)}{MSE(\hat{\theta}_2)}$$

Sampling design

Sampling design

The Horvitz-Thompson estimator (HT estimator) of the population mean:

$$\hat{\mu}_{HT} = \frac{1}{N} \sum_{j \in s} \frac{y_j}{\pi_j}$$

Sampling design

Sampling design

The Horvitz-Thompson estimator (HT estimator) of the population mean:

$$\hat{\mu}_{HT} = \frac{1}{N} \sum_{j \in s} \frac{y_j}{\pi_j}$$

HT estimator is an unbiased estimator of $\mu = \frac{1}{N} \sum_{j=1}^{N} Y_j$

Sampling design

Sampling design

The Horvitz-Thompson estimator (HT estimator) of the population mean:

$$\hat{\mu}_{HT} = \frac{1}{N} \sum_{j \in s} \frac{y_j}{\pi_j}$$

HT estimator is an unbiased estimator of $\mu = \frac{1}{N} \sum_{j=1}^{N} Y_j$

We showed that $E_p(\sum_{j \in s} z_j) = \sum_{j=1}^{N} z_j \pi_j$
It is enough to put $z_j = y_j / \pi_j$

Sampling design

Sampling design

Self-weighting design: all its inclusion probabilities are equal.

Sampling design

Sampling design

Self-weighting design: all its inclusion probabilities are equal.

For the designs which are both self-weighting and of fixed size a sample mean is an unbiased estimator of the population mean.

Sampling design

Sampling design

For the sample s: the sample mean equals $\bar{y}_s = \frac{1}{n} \sum_{j \in s} y_j$

Sampling design

Sampling design

For the sample s: the sample mean equals $\bar{y}_s = \frac{1}{n} \sum_{j \in s} y_j$

Since $n = \sum_{j=1}^{N} \pi_j$ and all π_j are equal: $\pi_j = \frac{n}{N}$

Sampling design

Sampling design

For the sample s: the sample mean equals $\bar{y}_s = \frac{1}{n}\sum_{j\in s} y_j$

Since $n = \sum_{j=1}^{N} \pi_j$ and all π_j are equal: $\pi_j = \frac{n}{N}$

The HT estimator is the sample mean:

$$\hat{\mu}_{HT} = \frac{1}{N}\sum_{j\in s}\frac{y_j}{\pi_j} = \frac{1}{n}\sum_{j\in s} y_j = \bar{y}$$

Statistics

Exercise 3

Four cows in a household marked A, B, C and D respectively yield 5.00, 5.50, 6.00 and 6.50 kg of milk per day. Obtain the sampling distribution of average milk yield $\bar{y}$ based on samples of $n = 2$ cows, when the cows are selected with WOR and WR .

Statistics

Exercise 3

Four cows in a household marked A, B, C and D respectively yield 5.00, 5.50, 6.00 and 6.50 kg of milk per day. Obtain the sampling distribution of average milk yield $\bar{y}$ based on samples of $n = 2$ cows, when the cows are selected with WOR and WR .

1. Calculate $P(\bar{y} > 5.9)$.

Statistics

Exercise 3

Four cows in a household marked A, B, C and D respectively yield 5.00, 5.50, 6.00 and 6.50 kg of milk per day. Obtain the sampling distribution of average milk yield $\bar{y}$ based on samples of $n = 2$ cows, when the cows are selected with WOR and WR .

1. Calculate $P(\bar{y} > 5.9)$.
2. Find the sampling variance of $\bar{y}$.

Statistics

Exercise 3

Four cows in a household marked A, B, C and D respectively yield 5.00, 5.50, 6.00 and 6.50 kg of milk per day. Obtain the sampling distribution of average milk yield $\bar{y}$ based on samples of $n = 2$ cows, when the cows are selected with WOR and WR .

1. Calculate $P(\bar{y} > 5.9)$.
2. Find the sampling variance of $\bar{y}$.
3. Find population mean μ and $MSE(\bar{y}) = E(\bar{y} - \mu)^2$.

Statistics

Exercise 3

Four cows in a household marked A, B, C and D respectively yield 5.00, 5.50, 6.00 and 6.50 kg of milk per day. Obtain the sampling distribution of average milk yield $\bar{y}$ based on samples of $n = 2$ cows, when the cows are selected with WOR and WR .

1. Calculate $P(\bar{y} > 5.9)$.
2. Find the sampling variance of $\bar{y}$.
3. Find population mean μ and $MSE(\bar{y}) = E(\bar{y} - \mu)^2$.
4. Draw a cumulative distribution function of the average milk yield.

Statistics

Exercise 3 cont.

Four cows in a household marked A, B, C and D respectively yield 5.00, 5.50, 6.00 and 6.50 kg of milk per day. Obtain the sampling distribution of average milk yield $\bar{y}$ based on samples of $n = 2$ cows, when the cows are selected with WOR and WR .

Statistics

Exercise 3 cont.

Four cows in a household marked A, B, C and D respectively yield 5.00, 5.50, 6.00 and 6.50 kg of milk per day. Obtain the sampling distribution of average milk yield $\bar{y}$ based on samples of $n = 2$ cows, when the cows are selected with WOR and WR .

5. Check whether the estimator $\bar{y}$ of the population average milk yield is unbiased ($E\bar{y} \stackrel{?}{=} \mu$).

Statistics

Exercise 3 cont.

Four cows in a household marked A, B, C and D respectively yield 5.00, 5.50, 6.00 and 6.50 kg of milk per day. Obtain the sampling distribution of average milk yield $\bar{y}$ based on samples of $n = 2$ cows, when the cows are selected with WOR and WR .

5. Check whether the estimator $\bar{y}$ of the population average milk yield is unbiased ($E\bar{y} \stackrel{?}{=} \mu$).
6. Let $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2$ be an estimator of σ^2. Find the bias of $\hat{\sigma}^2$.

Statistics

Exercise 4

The estimated mean square errors of two estimators, $\hat{\theta}_1$ and $\hat{\theta}_2$, are 4861.79 and 5258.62 respectively. Estimate percent of the relative efficiency of estimator $\hat{\theta}_2$ with respect to $\hat{\theta}_1$.
Also point out, which of the two estimators is more efficient?

Approximate confidence interval

Definition

Let $v(\hat{\theta})$ denote an estimator of $Var(\hat{\theta})$.
If $\frac{\hat{\theta}-\theta}{\sqrt{v(\hat{\theta})}} \sim AN(0,1)$ then approximate confidence interval for θ at $1-\alpha$ confidence level is

$$\left(\hat{\theta} - u_{1-\frac{\alpha}{2}}\sqrt{v(\hat{\theta})}, \hat{\theta} + u_{1-\frac{\alpha}{2}}\sqrt{v(\hat{\theta})}\right),$$

where $u_{1-\frac{\alpha}{2}}$ is the $(1-\frac{\alpha}{2})$-quantile of the standard normal distribution.
For instance $u_{1-\frac{0.05}{2}} = u_{0.975} = 1.96$

Simple random sampling with replacement

Population mean

Unbiased estimator of population mean μ

$$\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$$

Simple random sampling with replacement

Proof of unbiasedness of sample mean (sketch)

Let $\mathcal{U} = \{1, \ldots, N\}$ and let $s = (i_1, \ldots, i_n)$ (n is a fixed sample size). In WR scheme $p(s) = 1/N^n$ and for the given s we have

$$\bar{y} = \frac{1}{n}(Y_{i_1} + \cdots + Y_{i_n}) = \frac{1}{n}\sum_{i=1}^{N} Y_i \mathbf{1}(i \in s).$$

Simple random sampling with replacement

Proof of unbiasedness of the sample mean (sketch)

$$nE(\bar{y}) = \sum_{i=1}^{N} Y_i E\mathbf{1}(i \in s) = \sum_{i=1}^{N} Y_i P(\{s : i \in s\}) =$$

Simple random sampling with replacement

Proof of unbiasedness of the sample mean (sketch)

$$nE(\bar{y}) = \sum_{i=1}^{N} Y_i E\mathbf{1}(i \in s) = \sum_{i=1}^{N} Y_i P(\{s : i \in s\}) =$$

$$= \sum_{i=1}^{N} Y_i \sum_{s:\ i \in s} p(s) = \sum_{i=1}^{N} Y_i \sum_{s:\ i \in s} \frac{1}{N^n} =$$

Simple random sampling with replacement

Proof of unbiasedness of the sample mean (sketch)

$$nE(\bar{y}) = \sum_{i=1}^{N} Y_i E\mathbf{1}(i \in s) = \sum_{i=1}^{N} Y_i P(\{s : i \in s\}) =$$

$$= \sum_{i=1}^{N} Y_i \sum_{s:\ i \in s} p(s) = \sum_{i=1}^{N} Y_i \sum_{s:\ i \in s} \frac{1}{N^n} =$$

$$= \sum_{i=1}^{N} Y_i n \cdot N^{n-1} \cdot \frac{1}{N^n} = \frac{n}{N} \sum_{i=1}^{N} Y_i = n\bar{Y}$$

Simple random sampling with replacement

Sampling variance

Sampling variance of $\bar{y}$

$$Var(\bar{y}) = \frac{1}{n}\sigma^2$$

Unbiased estimator of sampling variance

$$v(\bar{y}) = \frac{1}{n}s^2$$

Simple random sampling without replacement

Population mean

Unbiased estimator of population mean μ

$$\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$$

Simple random sampling without replacement

Proof of unbiasedness of the sample mean (sketch)

Let $\mathcal{U} = \{1, \ldots, N\}$ and let $s = (i_1, \ldots, i_n)$ (n is a fixed sample size). In WOR scheme $p(s) = 1/\binom{N}{n}$ and for the given s we have

$$\bar{y} = \frac{1}{n}(Y_{i_1} + \cdots + Y_{i_n}) = \frac{1}{n}\sum_{i=1}^{N} Y_i \mathbf{1}(i \in s).$$

Simple random sampling without replacement

Proof of unbiasedness of the sample mean (sketch)

$$nE(\bar{y}) = \sum_{i=1}^{N} Y_i E\mathbf{1}(i \in s) = \sum_{i=1}^{N} Y_i P(\{s : i \in s\}) =$$

Simple random sampling without replacement

Proof of unbiasedness of the sample mean (sketch)

$$nE(\bar{y}) = \sum_{i=1}^{N} Y_i E\mathbf{1}(i \in s) = \sum_{i=1}^{N} Y_i P(\{s : i \in s\}) =$$

$$= \sum_{i=1}^{N} Y_i \sum_{s:\ i \in s} p(s) = \sum_{i=1}^{N} Y_i \sum_{s:\ i \in s} \frac{1}{\binom{N}{n}} =$$

Simple random sampling without replacement

Proof of unbiasedness of the sample mean (sketch)

$$nE(\bar{y}) = \sum_{i=1}^{N} Y_i E\mathbf{1}(i \in s) = \sum_{i=1}^{N} Y_i P(\{s : i \in s\}) =$$

$$= \sum_{i=1}^{N} Y_i \sum_{s:\ i \in s} p(s) = \sum_{i=1}^{N} Y_i \sum_{s:\ i \in s} \frac{1}{\binom{N}{n}} =$$

$$= \sum_{i=1}^{N} Y_i \cdot \binom{N-1}{n-1} \cdot \frac{1}{\binom{N}{n}} = \frac{n}{N} \sum_{i=1}^{N} Y_i = n\bar{Y}$$

Simple random sampling without replacement

Sampling variance

Sampling variance of $\bar{y}$

$$Var(\bar{y}) = \left(\frac{1}{n} - \frac{1}{N}\right)\sigma^2$$

Unbiased estimator of sampling variance

$$v(\bar{y}) = \left(\frac{1}{n} - \frac{1}{N}\right)s^2$$

Simple random sampling without replacement

Sampling variance of $\bar{y}$ (derivation of the formula)

$$Var(\bar{y}) = E\left(\bar{y} - \bar{Y}\right)^2 = E\left[\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \bar{Y}\right)\right]^2$$

Simple random sampling without replacement

Sampling variance of $\bar{y}$ (derivation of the formula)

$$Var(\bar{y}) = E\left(\bar{y} - \bar{Y}\right)^2 = E\left[\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \bar{Y}\right)\right]^2$$

$$= \frac{1}{n^2}\left\{E\left[\sum_{i=1}^{n}\left(y_i - \bar{Y}\right)^2\right] + E\left[\sum_{i\neq j}\left(y_i - \bar{Y}\right)\left(y_j - \bar{Y}\right)\right]\right\}$$

Simple random sampling without replacement

Sampling variance of $\bar{y}$ (derivation of the formula)

$$Var(\bar{y}) = E\left(\bar{y} - \bar{Y}\right)^2 = E\left[\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \bar{Y}\right)\right]^2$$

$$= \frac{1}{n^2}\left\{E\left[\sum_{i=1}^{n}\left(y_i - \bar{Y}\right)^2\right] + E\left[\sum_{i\neq j}\left(y_i - \bar{Y}\right)\left(y_j - \bar{Y}\right)\right]\right\}$$

$$= \frac{1}{n^2}\left\{\frac{n}{N}\sum_{i=1}^{N}\left(Y_i - \bar{Y}\right)^2 + \frac{n(n-1)}{N(N-1)}\sum_{i\neq j}\left(Y_i - \bar{Y}\right)\left(Y_j - \bar{Y}\right)\right\}$$

Simple random sampling without replacement

Sampling variance of $\bar{y}$ (derivation of the formula)

$$\left[\sum_{i=1}^{N}\left(Y_i-\bar{Y}\right)\right]^2=\sum_{i=1}^{N}\left(Y_i-\bar{Y}\right)^2+\sum_{i\neq j}\left(Y_i-\bar{Y}\right)\left(Y_j-\bar{Y}\right)$$

Simple random sampling without replacement

Sampling variance of $\bar{y}$ (derivation of the formula)

$$\left[\sum_{i=1}^{N}\left(Y_i-\bar{Y}\right)\right]^2=\sum_{i=1}^{N}\left(Y_i-\bar{Y}\right)^2+\sum_{i\neq j}\left(Y_i-\bar{Y}\right)\left(Y_j-\bar{Y}\right)$$

$$0=(N-1)\sigma^2+\sum_{i\neq j}\left(Y_i-\bar{Y}\right)\left(Y_j-\bar{Y}\right)$$

Simple random sampling without replacement

Sampling variance of $\bar{y}$ (derivation of the formula)

$$\left[\sum_{i=1}^{N}\left(Y_i-\bar{Y}\right)\right]^2=\sum_{i=1}^{N}\left(Y_i-\bar{Y}\right)^2+\sum_{i\neq j}\left(Y_i-\bar{Y}\right)\left(Y_j-\bar{Y}\right)$$

$$0=(N-1)\sigma^2+\sum_{i\neq j}\left(Y_i-\bar{Y}\right)\left(Y_j-\bar{Y}\right)$$

$$\sum_{i\neq j}\left(Y_i-\bar{Y}\right)\left(Y_j-\bar{Y}\right)=-(N-1)\sigma^2$$

Simple random sampling without replacement

Sampling variance of $\bar{y}$ (derivation of the formula)

$$Var(\bar{y}) = \frac{1}{Nn}\left\{\sum_{i=1}^{N}\left(Y_i - \bar{Y}\right)^2 + \frac{n-1}{N-1}\sum_{i\neq j}\left(Y_i - \bar{Y}\right)\left(Y_j - \bar{Y}\right)\right\}$$

Simple random sampling without replacement

Sampling variance of $\bar{y}$ (derivation of the formula)

$$Var(\bar{y}) = \frac{1}{Nn}\left\{\sum_{i=1}^{N}\left(Y_i - \bar{Y}\right)^2 + \frac{n-1}{N-1}\sum_{i\neq j}\left(Y_i - \bar{Y}\right)\left(Y_j - \bar{Y}\right)\right\}$$

$$= \frac{1}{Nn}\left\{(N-1)\sigma^2 + \frac{n-1}{N-1}\sum_{i\neq j}\left(Y_i - \bar{Y}\right)\left(Y_j - \bar{Y}\right)\right\}$$

Simple random sampling without replacement

Sampling variance of $\bar{y}$ (derivation of the formula)

$$
\begin{aligned}
Var(\bar{y}) &= \frac{1}{Nn}\left\{\sum_{i=1}^{N}\left(Y_i-\bar{Y}\right)^2+\frac{n-1}{N-1}\sum_{i\neq j}\left(Y_i-\bar{Y}\right)\left(Y_j-\bar{Y}\right)\right\}\\
&= \frac{1}{Nn}\left\{(N-1)\sigma^2+\frac{n-1}{N-1}\sum_{i\neq j}\left(Y_i-\bar{Y}\right)\left(Y_j-\bar{Y}\right)\right\}\\
&= \frac{1}{Nn}\left\{(N-1)-(n-1)\right\}\sigma^2
\end{aligned}
$$

Simple random sampling without replacement

Sampling variance of $\bar{y}$ (derivation of the formula)

$$
\begin{aligned}
Var(\bar{y}) &= \frac{1}{Nn}\left\{\sum_{i=1}^{N}\left(Y_i-\bar{Y}\right)^2+\frac{n-1}{N-1}\sum_{i\neq j}\left(Y_i-\bar{Y}\right)\left(Y_j-\bar{Y}\right)\right\}\\
&= \frac{1}{Nn}\left\{(N-1)\sigma^2+\frac{n-1}{N-1}\sum_{i\neq j}\left(Y_i-\bar{Y}\right)\left(Y_j-\bar{Y}\right)\right\}\\
&= \frac{1}{Nn}\left\{(N-1)-(n-1)\right\}\sigma^2\\
&= \left(\frac{1}{n}-\frac{1}{N}\right)\sigma^2
\end{aligned}
$$

Simple random sampling

Exercise 5

In the file Doraha.xlsx there are given data related to the number of tractors in 69 serially numbered villages of Doraha development block in Punjab (India).
Select (1) WR and (2) WOR simple random sample of 10 villages.

Simple random sampling

Exercise 6

In a survey, the sample mean was computed as 796.3, and the value of the variance estimator came out to be 1016.9. Build up the confidence interval for population mean and interpret the results.

Simple random sampling

Exercise 7

An investigator has randomly selected 2000 families following WR procedure from a population of $10,000$ families. For working out a sufficiently accurate confidence interval for population mean, he/she is to guess the distribution of sample mean in absence of any information regarding the distribution of study variable in the population. Is it reasonable to assume that the sampling distribution is (a) exactly normal, (b) approximately normal or (c) not normal?

Simple random sampling

Exercise 8

From the WOR sample of 10 villages estimate the average number of tractors per village in the block of 270 tractors along with its standard error. Also, set up confidence interval for the population mean. The sample data (number of tractors) is $16, 6, 19, 18, 12, 13, 17, 8, 15, 17$. How the length of this confidence interval would change if the sample was driven with replacement?

Simple random sampling

Exercise 9

Estimate the total number of tractors in the development block of 69 villages using the samples selected in Exercise 5.

Simple random sampling

Mean estimation using distinct values

The units which get repeated while selecting the sample do not provide any additional information. The information obtained from distinct units is sufficient to estimate population mean. Let $y_1, \ldots, y_d$ be the d distinct units while selecting the sample.

Simple random sampling

Mean estimation using distinct values

Estimator of population mean μ: $\bar{y}_d = \frac{1}{d} \sum_{i=1}^{d} y_i$

Simple random sampling

Mean estimation using distinct values

Estimator of population mean μ: $\bar{y}_d = \frac{1}{d} \sum_{i=1}^{d} y_i$

Sampling variance of $\bar{y}_d$: $Var(\bar{y}_d) = \left(E(\frac{1}{d}) - \frac{1}{N}\right) \sigma^2$

Simple random sampling

Mean estimation using distinct values

Estimator of population mean μ: $\bar{y}_d = \frac{1}{d}\sum_{i=1}^{d} y_i$

Sampling variance of $\bar{y}_d$: $Var(\bar{y}_d) = \left(E(\frac{1}{d}) - \frac{1}{N}\right)\sigma^2$

Estimator of sampling variance: $v(\bar{y}_d) = (\frac{1}{d} - \frac{1}{N})s_d^2$,
where $d \geq 2$ and $s_d^2 = \frac{1}{d-1}(y_i - \bar{y}_d)^2$

Simple random sampling

Exercise 10

From the data related to the number of tractors in 69 serially numbered villages of Doraha development block in Punjab the following simple random sampling with replacement was selected:

Village	23	28	54	52	49	6	44	30	10	6	53	66	53	56	6
Tractors	7	21	11	8	38	21	29	59	10	21	12	20	12	8	21

Estimate population mean using $\bar{y}_d$ estimator. Build up the confidence interval for population mean and total. Note that the village bearing serial number 6 has been selected 3 times.

Determining sample size

Preliminary sample

Let n_1 be the size of preliminary sample selected using WOR scheme. Let $\bar{y} = \frac{1}{n_1}\sum_{i=1}^{n_1} y_i$ and $s^2 = \frac{1}{n_1-1}\sum_{i=1}^{n_1}(y_i - \bar{y})^2$ be parameters of the preliminary sample.

Determining sample size

Preliminary sample

Let n_1 be the size of preliminary sample selected using WOR scheme. Let $\bar{y} = \frac{1}{n_1}\sum_{i=1}^{n_1} y_i$ and $s^2 = \frac{1}{n_1-1}\sum_{i=1}^{n_1}(y_i - \bar{y})^2$ be parameters of the preliminary sample.

Sample size

$$n^\star = \frac{N u^2_{1-\frac{\alpha}{2}} s^2}{N d^2 + u^2_{1-\frac{\alpha}{2}} s^2} = \frac{N u^2_{1-\frac{\alpha}{2}} v^2}{N \delta^2 + u^2_{1-\frac{\alpha}{2}} v^2},$$

where $v = s/\bar{y}$, d is the given permissible error and $\delta = d/\bar{y}$.

Determining sample size

Sample size $n^\star$ (sketch of the proof)

$$P(|\bar{y} - \mu| < d) = P(\bar{y} - d < \mu < \bar{y} + d) = 1 - \alpha$$

Determining sample size

Sample size $n^\star$ (sketch of the proof)

$$P(|\bar{y} - \mu| < d) = P(\bar{y} - d < \mu < \bar{y} + d) = 1 - \alpha$$

$$P(\bar{y} - u_{1-\frac{\alpha}{2}} Var(\bar{y}) < \mu < \bar{y} + u_{1-\frac{\alpha}{2}} Var(\bar{y})) \approx 1 - \alpha$$

where $Var(\bar{y}) = (1/n - 1/N)\sigma^2 \approx (1/n - 1/N)s^2$

Determining sample size

Sample size $n^\star$ (sketch of the proof)

$$P(|\bar{y} - \mu| < d) = P(\bar{y} - d < \mu < \bar{y} + d) = 1 - \alpha$$

$$P(\bar{y} - u_{1-\frac{\alpha}{2}} Var(\bar{y}) < \mu < \bar{y} + u_{1-\frac{\alpha}{2}} Var(\bar{y})) \approx 1 - \alpha$$

where $Var(\bar{y}) = (1/n - 1/N)\sigma^2 \approx (1/n - 1/N)s^2$

$$d = u_{1-\frac{\alpha}{2}} \left(\frac{1}{n} - \frac{1}{N}\right) s^2 \;\Rightarrow\; n^\star = \frac{N u_{1-\frac{\alpha}{2}}^2 s^2}{N d^2 + u_{1-\frac{\alpha}{2}}^2 s^2}$$

Proportion

Definition

Let K units out of N possess the attribute of interest. Then population proportion

$$\theta = \frac{K}{N}$$

Proportion

Definition

Let K units out of N possess the attribute of interest. Then population proportion

$$\theta = \frac{K}{N}$$

Estimator

If k units out of n sample units possess this attribute, sample proportion is given by

$$\hat{\theta} = \frac{k}{n}$$

Simple random sampling with replacement

Population proportion

Unbiased estimator of population proportion θ

$$\hat{\theta} = \frac{k}{n}$$

Simple random sampling with replacement

Sampling variance

Sampling variance of $\hat{\theta}$:

$$Var(\hat{\theta}) = \frac{\theta(1-\theta)}{n}$$

Unbiased estimator of sampling variance

$$v(\hat{\theta}) = \frac{\hat{\theta}(1-\hat{\theta})}{n-1}$$

Simple random sampling without replacement

Population proportion

Unbiased estimator of population proportion θ

$$\hat{\theta} = \frac{k}{n}$$

Simple random sampling without replacement

Sampling variance

Sampling variance of $\hat{\theta}$:

$$Var(\hat{\theta}) = \frac{N-n}{N-1}\frac{\theta(1-\theta)}{n}$$

Unbiased estimator of sampling variance

$$v(\hat{\theta}) = \left(1-\frac{n}{N}\right)\frac{\hat{\theta}(1-\hat{\theta})}{n-1}$$

Inverse sampling

Definition

The procedure where sampling is continued until a predetermined number of units possessing the attribute are included in the sample, is known as inverse sampling.

Let n be the number of units required to be selected to obtain a predetermined number m of units possessing the rare attribute.

Inverse sampling

Population proportion

Unbiased estimator of population proportion θ

$$\hat{\theta} = \frac{m-1}{n-1}$$

Inverse sampling

Sampling variance

WOR estimator of sampling variance

$$v(\hat{\theta}) = \left(1 - \frac{n-1}{N}\right) \frac{\hat{\theta}(1-\hat{\theta})}{n-2}$$

WR estimator of sampling variance

$$v(\hat{\theta}) = \frac{\hat{\theta}(1-\hat{\theta})}{n-2}$$

Simple random sampling

Exercise 11

Punjab Agricultural University, Ludhiana, is interested in estimating the proportion of teachers who consider semester system to be more suitable as compared to the trimester system of education. A WR simple random sample of $n = 120$ teachers is taken from a total of $N = 1200$ teachers. The response is denoted by 0 if the teacher does not think the semester system is suitable, and 1 elsewhere. From the sample observations given below estimate the proportion, its standard error and the confidence interval for the proportion

Teacher	1	2	3	...	120	Total
Response	1	0	1	...	1	81

Simple random sampling

Exercise 12

A survey conducted by a student of a medical college in Ludhiana town showed that a proportion .008 of adults over 18 years of age, living in a posh colony, are suffering from tuberculosis. Another student of the same college was subsequently given an assignment to examine whether the incidence of tuberculosis infection in the adults of the same age group, living in a slum area, is on the higher side of .008?

Simple random sampling

Exercise 12 cont.

For conducting this survey, voters' lists were used as frame, and voters as the sampling units. It was decided in advance to continue with replacement simple random sampling of individuals till 10 cases of tuberculosis infection were detected. To arrive at this predetermined number of 10, the investigator had to select 380 adults from the slum area. Estimate the proportion in question.

Stratified Sampling

Stratification

Apart from increasing the sample size, another possible way to increase the precision of the estimate could be to divide the population units into certain number of groups. The groups thus formed are called strata, and the process of forming strata is known as stratification.

- The procedure of partitioning the population into groups, called strata, and then drawing a sample independently from each stratum, is known as stratified sampling.
- If the sample drawn from each stratum is random one, the procedure is then termed as stratified random sampling.

Stratified Sampling

Stratification

- N_h: total number of units in the stratum

Stratified Sampling

Stratification

- N_h: total number of units in the stratum
- $W_h = N_h/N$: the weight of the stratum

Stratified Sampling

Stratification

- N_h: total number of units in the stratum
- $W_h = N_h/N$: the weight of the stratum
- n_h: sample size the stratum

Stratified Sampling

Stratification

- N_h: total number of units in the stratum
- $W_h = N_h/N$: the weight of the stratum
- n_h: sample size the stratum
- $f_h = n_h/N_h$: sampling fraction for the stratum

Stratified Sampling

Stratification

- Y_{hi}: the value of the variable for the i-th unit in h stratum

Stratified Sampling

Stratification

- Y_{hi}: the value of the variable for the i-th unit in h stratum
- $Y_h = \sum_{i=1}^{N_h} Y_{hi}$: stratum total of the variable in the stratum

Stratified Sampling

Stratification

- Y_{hi}: the value of the variable for the i-th unit in h stratum
- $Y_h = \sum_{i=1}^{N_h} Y_{hi}$: stratum total of the variable in the stratum
- $\bar{Y}_h = \frac{1}{N_h} \sum_{i=1}^{N_h} Y_{hi}$: mean of the variable in the stratum

Stratified Sampling

Stratification

- Y_{hi}: the value of the variable for the i-th unit in h stratum
- $Y_h = \sum_{i=1}^{N_h} Y_{hi}$: stratum total of the variable in the stratum
- $\bar{Y}_h = \frac{1}{N_h} \sum_{i=1}^{N_h} Y_{hi}$: mean of the variable in the stratum
- $\bar{y}_h = \frac{1}{n_h} \sum_{i=1}^{n_h} y_{hi}$: stratum sample mean

Stratified Sampling

Stratification

- Y_{hi}: the value of the variable for the i-th unit in h stratum
- $Y_h = \sum_{i=1}^{N_h} Y_{hi}$: stratum total of the variable in the stratum
- $\bar{Y}_h = \frac{1}{N_h} \sum_{i=1}^{N_h} Y_{hi}$: mean of the variable in the stratum
- $\bar{y}_h = \frac{1}{n_h} \sum_{i=1}^{n_h} y_{hi}$: stratum sample mean
- $\sigma_h^2 = \frac{1}{N_h - 1} \sum_{i=1}^{N_h} (Y_{hi} - \bar{Y}_h)^2$: stratum mean square

Stratified Sampling

Stratification

- Y_{hi}: the value of the variable for the i-th unit in h stratum
- $Y_h = \sum_{i=1}^{N_h} Y_{hi}$: stratum total of the variable in the stratum
- $\bar{Y}_h = \frac{1}{N_h} \sum_{i=1}^{N_h} Y_{hi}$: mean of the variable in the stratum
- $\bar{y}_h = \frac{1}{n_h} \sum_{i=1}^{n_h} y_{hi}$: stratum sample mean
- $\sigma_h^2 = \frac{1}{N_h - 1} \sum_{i=1}^{N_h} (Y_{hi} - \bar{Y}_h)^2$: stratum mean square
- $s_h^2 = \frac{1}{n_h - 1} \sum_{i=1}^{n_h} (y_{hi} - \bar{y}_h)^2$: stratum sample mean square

Sampling design

Stratified random sampling

Let $\mathcal{U}$ be the union of disjoint strata $\mathcal{U}_1, \ldots, \mathcal{U}_H$. The sizes $N_1, \ldots, N_H$ of the strata are known, and $\sum_{h=1}^{H} N_h = N$. For each h separately, the design prescribes a simple random sampling of n_h draws without replacement from $\mathcal{U}_h$. In this case $\pi_j = n_h/N_h$ for $j \in \mathcal{U}_h$. Thus the HT estimator for μ is

$$\hat{\mu}_{st} = \sum_{h=1}^{H} W_h \bar{y}_h,$$

where $W_h = N_h/N$ and $\bar{y}_h$ is the mean of the sample from $\mathcal{U}_h$.

Sampling design

Stratified random sampling

$$Var\left(\sum_{j\in s} z_j\right) = \sum_{j=1}^{N} z_j^2 \pi_j(1-\pi_j) + \sum_{j\neq k}^{N} z_j z_k(\pi_{jk} - \pi_j\pi_k) \quad (1)$$

For the designs of fixed size n

$$Var\left(\sum_{j\in s} z_j\right) = \frac{1}{2}\sum_{j\neq k}^{N}(z_j - z_k)^2(\pi_j\pi_k - \pi_{jk}) \quad (2)$$

Sampling design

Sampling variance (derivation of the formula)

$$Var\left(\sum_{j\in s} z_j\right) = Var\left(\sum_{j=1}^{N} z_j \mathbf{1}(j \in s)\right) =$$

$$= \sum_{j=1}^{N} z_j^2 Var\mathbf{1}(j \in s) + \sum_{j \neq k}^{n}\sum^{n} z_j z_k Cov(\mathbf{1}(j \in s), \mathbf{1}(k \in s))$$

$$= \sum_{j=1}^{N} z_j^2 \pi_j(1-\pi_k) + \sum_{j \neq k}^{N}\sum^{N} z_j z_k (\pi_j \pi_k - \pi_{jk})$$

Sampling design

Sampling variance (derivation of the formula)

If $n(s) = n$ for every sample s then $\sum_{k=1}^{N} \mathbf{1}(k \in s) = n$. In this case

$$\sum_{k \neq j} Cov(\mathbf{1}(j \in s), \mathbf{1}(k \in s)) = Cov(\mathbf{1}(j \in s), n - \mathbf{1}(j \in s))$$

$$= -Var(\mathbf{1}(j \in s))$$

Hence

$$Var(\mathbf{1}(j \in s)) = -\sum_{k \neq j} (\pi_{jk} - \pi_j \pi_k) = \sum_{k \neq j} (\pi_j \pi_k - \pi_{jk})$$

Sampling design

Sampling variance (derivation of the formula)

$$\sum_{j=1}^{N} z_j^2 Var\mathbf{1}(j \in s) + \sum_{j \neq k}^{n}\sum^{n} z_j z_k Cov(\mathbf{1}(j \in s), \mathbf{1}(k \in s)) =$$

$$= \sum_{j=1}^{N} z_j^2 \sum_{k \neq j} (\pi_j \pi_k - \pi_{jk}) - \sum_{j \neq k}^{n}\sum^{n} z_j z_k (\pi_j \pi_k - \pi_{jk})$$

$$= \frac{1}{2} \sum_{j \neq k}^{N}\sum^{N} (z_j^2 + z_k^2)(\pi_j \pi_k - \pi_{jk}) - \sum_{j \neq k}^{n}\sum^{n} z_j z_k (\pi_j \pi_k - \pi_{jk})$$

$$= \frac{1}{2} \sum_{j \neq k}^{N}\sum^{N} (z_j - z_k)^2 (\pi_j \pi_k - \pi_{jk})$$

Simple random sampling

Exercise 13

Show that in a stratified random sampling without replacement if $n_h/N_h = n/N,$ for all h, then

$$\bar{y}_{st} = \bar{y}.$$

(in the case: $n_h = nW_h,$ for all h, we say that the allocation is proportional)

Simple random sampling

Exercise 13 cont.

It can be shown, that

$$Var\left(\sum_{j\in s} z_j\right) = \sum_{j=1}^{N} z_j^2 \pi_j(1-\pi_j) + \sum_{j\neq k}^{N} z_j z_k(\pi_{jk} - \pi_j\pi_k) \quad (3)$$

and when the design is of fixed size n

$$Var\left(\sum_{j\in s} z_j\right) = \frac{1}{2}\sum_{j\neq k}^{N}(z_j - z_k)^2(\pi_j\pi_k - \pi_{jk}) \quad (4)$$

Simple random sampling

Exercise 14

Using equation (4) calculate $Var(\bar{y})$ in simple random sampling without replacement design.

Simple random sampling

Exercise 15

Show that for Horvitz-Thompson estimator holds

$$Var(\bar{\mu}_{HT}) = \frac{1}{N^2}\left[\sum_{j=1}^{N} Y_j^2\left(\frac{1}{\pi_j} - 1\right) + \sum_{j \neq k}^{N}\sum^{N}\left(\frac{\pi_{jk}}{\pi_j\pi_k} - 1\right) Y_j Y_k\right]$$

If $n(s) = n$ for every sample s then

$$Var(\bar{\mu}_{HT}) = \frac{1}{2N^2}\sum_{j \neq k}^{N}\sum^{N}(\pi_i\pi_j - \pi_{ij})\left(\frac{Y_j}{\pi_j} - \frac{Y_k}{\pi_k}\right)^2$$

Simple random sampling

Remark

The Horvitz-Thompson estimator estimator may be seriously deficient (see Thompson (1997): example 2.5 on page 18).

Stratified Sampling

Stratified WOR simple random sampling

Let $\mathcal{U}$ be the union of disjoint strata $\mathcal{U}_1, \ldots, \mathcal{U}_H$.

Stratified Sampling

Stratified WOR simple random sampling

Let $\mathcal{U}$ be the union of disjoint strata $\mathcal{U}_1, \ldots, \mathcal{U}_H$.

The sizes $N_1, \ldots, N_H$ of the strata are known: $\sum_{h=1}^{H} N_h = N$.

Stratified Sampling

Stratified WOR simple random sampling

Let $\mathcal{U}$ be the union of disjoint strata $\mathcal{U}_1, \ldots, \mathcal{U}_H$.

The sizes $N_1, \ldots, N_H$ of the strata are known: $\sum_{h=1}^{H} N_h = N$.

For each h separately, the design prescribes a simple random sampling of n_h draws without replacement from $\mathcal{U}_h$.

Stratified Sampling

Stratified WOR simple random sampling

Unbiased estimator of the population mean μ:

$$\bar{y}_{st} = \sum_{h=1}^{H} W_h \bar{y}_h$$

Stratified Sampling

Stratified WOR simple random sampling

Variance of the estimator $\bar{y}_{st}$:

$$Var(\bar{y}_{st}) = \sum_{h=1}^{H} W_h^2 \left(\frac{1}{n_h} - \frac{1}{N_h} \right) \sigma_h^2$$

Stratified Sampling

Stratified WOR simple random sampling

Variance of the estimator $\bar{y}_{st}$:

$$Var(\bar{y}_{st}) = \sum_{h=1}^{H} W_h^2 \left(\frac{1}{n_h} - \frac{1}{N_h} \right) \sigma_h^2$$

Unbiased estimator of the variance $Var(\bar{y}_{st})$:

$$v(y_{st}) = \sum_{h=1}^{H} W_h^2 \left(\frac{1}{n_h} - \frac{1}{N_h} \right) s_h^2$$

Stratified Sampling

Stratified SRS WOR and Proportional Allocation

Let n be the overall sample size.
Proportional allocation:

$$n_h = n \cdot W_h = n \cdot \frac{N_h}{N}$$

for $h = 1, \ldots, H$

Stratified Sampling

Stratified SRS WOR and Proportional Allocation

Unbiased estimator of the population mean μ:

$$\bar{y}_{st} = \sum_{h=1}^{H} W_h \bar{y}_h$$

Stratified Sampling

Stratified SRS WOR and Proportional Allocation

Variance of the estimator $\bar{y}_{st}$:

$$Var(\bar{y}_{st}) = \left(\frac{1}{n} - \frac{1}{N}\right) \sum_{h=1}^{H} W_h \sigma_h^2$$

Stratified Sampling

Stratified SRS WOR and Proportional Allocation

Variance of the estimator $\bar{y}_{st}$:

$$Var(\bar{y}_{st}) = \left(\frac{1}{n} - \frac{1}{N}\right) \sum_{h=1}^{H} W_h \sigma_h^2$$

Unbiased estimator of the variance $Var(\bar{y}_{st})$:

$$v(\bar{y}_{st}) = \left(\frac{1}{n} - \frac{1}{N}\right) \sum_{h=1}^{H} W_h s_h^2$$

Stratified Sampling

Stratified SRS WOR fixed costs optimal allocation

Let the budget C for a survey be fixed.

Stratified Sampling

Stratified SRS WOR fixed costs optimal allocation

Let the budget C for a survey be fixed.
Let c_h be the cost of observing Y in $h-$th stratum, $h = 1, \ldots, H$

Stratified Sampling

Stratified SRS WOR fixed costs optimal allocation

Let the budget C for a survey be fixed.
Let c_h be the cost of observing Y in $h-$th stratum, $h = 1, \ldots, H$
The problem: find $n_1, \ldots, n_h$ minimizing the variance

$$Var(\bar{y}_{st}) = \sum_{h=1}^{H} W_h^2 \left(\frac{1}{n_h} - \frac{1}{N_h} \right) \sigma_h^2$$

due to constraint

$$\sum_{h=1}^{H} n_h c_h \leq C$$

Stratified Sampling

Stratified SRS WOR fixed costs optimal allocation

Solution

$$n_h = \frac{C}{\sqrt{c_h}} \cdot \frac{W_h S_h}{\sum_{k=1}^{H} W_k S_k \sqrt{c_k}}$$

Stratified Sampling

Stratified SRS WOR fixed costs optimal allocation

Solution

$$n_h = \frac{C}{\sqrt{c_h}} \cdot \frac{W_h S_h}{\sum_{k=1}^{H} W_k S_k \sqrt{c_k}}$$

Apply Lagrange multipliers

Stratified Sampling

Stratified SRS WOR fixed costs optimal allocation

If $c_h = c$ for each h, then

$$n_h = \frac{C}{c} \cdot \frac{W_h S_h}{\sum_{k=1}^{H} W_k S_k}.$$

Stratified Sampling

Stratified SRS WOR fixed costs optimal allocation

If $c_h = c$ for each h, then

$$n_h = \frac{C}{c} \cdot \frac{W_h S_h}{\sum_{k=1}^{H} W_k S_k}.$$

Neyman allocation

Stratified Sampling

Stratified SRS WOR fixed precision allocation

Let the precision V_0 of the estimation be fixed

Stratified Sampling

Stratified SRS WOR fixed precision allocation

Let the precision V_0 of the estimation be fixed

Let c_h be the cost of observing Y in $h-$th stratum, $h = 1, \ldots, H$

Stratified Sampling

Stratified SRS WOR fixed precision allocation

Let the precision V_0 of the estimation be fixed
Let c_h be the cost of observing Y in $h-$th stratum, $h = 1, \ldots, H$
The problem: find $n_1, \ldots, n_h$ minimizing the cost

$$\sum_{h=1}^{H} n_h c_h$$

due to constraint

$$Var(\bar{y}_{st}) = \sum_{h=1}^{H} W_h^2 \left(\frac{1}{n_h} - \frac{1}{N_h} \right) \sigma_h^2 \leq V_0$$

Stratified Sampling

Stratified SRS WOR fixed precision allocation

Solution

$$n_h = \frac{W_h S_h}{\sqrt{c_h}} \frac{\sum_{k=1}^{H} W_k S_k \sqrt{c_k}}{V_0 + \frac{1}{N} \sum_{k=1}^{H} W_k S_k^2}$$

Stratified Sampling

Stratified SRS WOR fixed precision allocation

Solution

$$n_h = \frac{W_h S_h}{\sqrt{c_h}} \frac{\sum_{k=1}^{H} W_k S_k \sqrt{c_k}}{V_0 + \frac{1}{N} \sum_{k=1}^{H} W_k S_k^2}$$

Apply Lagrange multipliers

Stratified Sampling

Stratified SRS WOR fixed precision allocation

If $c_h = c$ for each h, then

$$n_h = W_h S_h \frac{\sum_{k=1}^{H} W_k S_k}{V_0 + \frac{1}{N} \sum_{k=1}^{H} W_k S_k^2}$$

Stratified Sampling

Stratified SRS WOR fixed precision allocation

If $c_h = c$ for each h, then

$$n_h = W_h S_h \frac{\sum_{k=1}^{H} W_k S_k}{V_0 + \frac{1}{N} \sum_{k=1}^{H} W_k S_k^2}$$

Neyman allocation

Stratified Sampling

Exercise 16

The problem was to estimate the average time per week devoted to study in Punjab Agricultural University (PAU) library by the students of this university.

Stratified Sampling

Exercise 16

The problem was to estimate the average time per week devoted to study in Punjab Agricultural University (PAU) library by the students of this university.
The university is running undergraduate, master's degree and doctoral programs.

Stratified Sampling

Exercise 16

The problem was to estimate the average time per week devoted to study in Punjab Agricultural University (PAU) library by the students of this university.
The university is running undergraduate, master's degree and doctoral programs.
The value of the study variable is likely to differ considerably with the program.

Stratified Sampling

Exercise 16

The problem was to estimate the average time per week devoted to study in Punjab Agricultural University (PAU) library by the students of this university.
The university is running undergraduate, master's degree and doctoral programs.
The value of the study variable is likely to differ considerably with the program.
The investigator decided to divide the population of students into three strata.

Stratified Sampling

Exercise 16 cont.

$\mathcal{U}_1$: undergraduate program ($N_1 = 1300$)
$\mathcal{U}_2$: master's program ($N_2 = 450$)
$\mathcal{U}_3$: doctoral program ($N_3 = 250$)

Stratified Sampling

Exercise 16 cont.

$\mathcal{U}_1$: undergraduate program ($N_1 = 1300$)
$\mathcal{U}_2$: master's program ($N_2 = 450$)
$\mathcal{U}_3$: doctoral program ($N_3 = 250$)
WOR samples:

$$n_1 = 20,\ n_2 = 10,\ n_3 = 12$$

Stratified Sampling

Exercise 16 cont.

Stratum	Time						
undergraduate	0	1	9	4	4	4	3
	3	3	6	5	6	1	2
	2	8	2	0	10	2	
master	12	6	9	10	11	9	13
	11	8	7				
doctoral	10	14	24	15	20	14	13
	20	11	18	16	19		

Stratified Sampling

Exercise 16 cont.

1. Estimate the average time per week devoted to study by a student in PAU library.

Stratified Sampling

Exercise 16 cont.

1. Estimate the average time per week devoted to study by a student in PAU library.
2. Build up the confidence interval for this average.

Stratified Sampling

Exercise 17

A car manufacturing company has sold 2000 cars to the public through licensed dealers.

Stratified Sampling

Exercise 17

A car manufacturing company has sold 2000 cars to the public through licensed dealers.
The company is now interested in finding out the average distance traveled per week.

Stratified Sampling

Exercise 17

A car manufacturing company has sold 2000 cars to the public through licensed dealers.
The company is now interested in finding out the average distance traveled per week.
This information is likely to be helpful in fixing the warranty period for certain parts of the car.

Stratified Sampling

Exercise 17

A car manufacturing company has sold 2000 cars to the public through licensed dealers.
The company is now interested in finding out the average distance traveled per week.
This information is likely to be helpful in fixing the warranty period for certain parts of the car.
The distance traveled by a car is likely to vary with the profession of the buyer.

Stratified Sampling

Exercise 17

A car manufacturing company has sold 2000 cars to the public through licensed dealers.
The company is now interested in finding out the average distance traveled per week.
This information is likely to be helpful in fixing the warranty period for certain parts of the car.
The distance traveled by a car is likely to vary with the profession of the buyer.
The population was divided into three groups: the businessmen, employees and others.

Stratified Sampling

Exercise 17 cont.

$\mathcal{U}_1$: the businessmen ($N_1 = 825$)
$\mathcal{U}_2$: employees ($N_2 = 700$)
$\mathcal{U}_3$: others ($N_3 = 475$)

Stratified Sampling

Exercise 17 cont.

$\mathcal{U}_1$: the businessmen ($N_1 = 825$)
$\mathcal{U}_2$: employees ($N_2 = 700$)
$\mathcal{U}_3$: others ($N_3 = 475$)
The unit costs:

$$c_1 = \$4,\ c_2 = \$5.5,\ c_3 = \$6.5$$

Stratified Sampling

Exercise 17 cont.

$\mathcal{U}_1$: the businessmen ($N_1 = 825$)
$\mathcal{U}_2$: employees ($N_2 = 700$)
$\mathcal{U}_3$: others ($N_3 = 475$)
The unit costs:

$$c_1 = \$4,\ c_2 = \$5.5,\ c_3 = \$6.5$$

The total budget: $C = \$1000$

Stratified Sampling

Exercise 17 cont.

businessmen					employees			others	
656	301	575	666	746	470	281	685	712	236
400	870	525	715	560	351	410	492	679	824
526	813	310	691	475	625	240	206	665	385
774	861	650	480	399	388	636	579	319	650
780	722	470	680	635	566	422	358	840	585
812	705	460	841	560	421	517	385	421	496
805	831	483	825	704	398	451	615	666	704
525	748	310	488	774	881	380	375	848	569
401	446	489	330	533	434	326	469	410	614
806	856	576	580		405	595	612	549	253
828	387	615	811		693	401	564	602	777
					343			411	

Stratified Sampling

Exercise 17 cont.

The information on strata mean squares, from a similar survey carried out in the past for another car model, is given

$$\sigma_1^2 = 30505,\ \sigma_2^2 = 24008,\ \sigma_3^2 = 29215$$

Find minimum variance allocation.

Two-stage sampling

Introduction

$$\mathcal{U} = \bigcup_{i=1}^{H} \mathcal{U}_h, \qquad \mathcal{U}_g \cap \mathcal{U}_h = \emptyset \text{ for } g \neq h$$

$$\mathcal{U}_h = \{u_{h1}, \ldots, u_{hN_h}\}$$

$$N_1 + \cdots + N_H = N$$

Two-stage Sampling

Investigated variate

$$\bar{Y}_h = \frac{1}{N_h}\sum_{j=1}^{N_h} Y_{hj} \qquad \bar{Y} = \frac{1}{N}\sum_{h=1}^{H}\sum_{j=1}^{N_h} Y_{hj} = \sum_{h=1}^{H} W_h \bar{Y}_h$$

$$\sigma_h^2 = \frac{1}{N_h - 1}\sum_{j=1}^{N_h}(Y_{hj} - \bar{Y}_h)^2$$

$$\sigma^2 = \frac{1}{N - 1}\sum_{h=1}^{H}\sum_{j=1}^{N_h}(Y_{hj} - \bar{Y})^2$$

Two-stage Sampling

WOR + WOR scheme

- Step 1: draw $2 < M < H$ strata with respect to WOR scheme

Two-stage Sampling

WOR + WOR scheme

- Step 1: draw $2 < M < H$ strata with respect to WOR scheme
- Step 2: from drawn strata draw WOR samples

Two-stage Sampling

WOR + WOR scheme

- Step 1: draw $2 < M < H$ strata with respect to WOR scheme
- Step 2: from drawn strata draw WOR samples
- The overall sample size

$$n = \sum_{h=1}^{M} n_{(h)}$$

Two-stage Sampling

WOR + WOR scheme: two-stage estimator

$$\bar{y}_{(2)} = \frac{H}{M} \sum_{h=1}^{M} W_{(h)} \bar{y}_{(h)}$$

Two-stage Sampling

WOR + WOR scheme: two-stage estimator

$$\bar{y}_{(2)} = \frac{H}{M} \sum_{h=1}^{M} W_{(h)} \bar{y}_{(h)}$$

Properties

Two-stage Sampling

WOR + WOR scheme: two-stage estimator

$$\bar{y}_{(2)} = \frac{H}{M} \sum_{h=1}^{M} W_{(h)} \bar{y}_{(h)}$$

Properties

- $E\bar{y}_{(2)} = \bar{Y}$

Two-stage Sampling

WOR + WOR scheme: two-stage estimator

$$\bar{y}_{(2)} = \frac{H}{M} \sum_{h=1}^{M} W_{(h)} \bar{y}_{(h)}$$

Properties

- $E\bar{y}_{(2)} = \bar{Y}$
- $Var(\bar{y}_{(2)}) = \frac{H}{MN^2} \left[(H-M)\sigma^2_{(1)} + \sum_{r=1}^{H} N_r (N_r - n_r) \frac{\sigma_r^2}{n_r} \right]$

Two-stage Sampling

WOR + WOR scheme: two-stage estimator

$$\bar{y}_{(2)} = \frac{H}{M} \sum_{h=1}^{M} W_{(h)} \bar{y}_{(h)}$$

Properties

- $E\bar{y}_{(2)} = \bar{Y}$
- $Var(\bar{y}_{(2)}) = \frac{H}{MN^2} \left[(H-M)\sigma^2_{(1)} + \sum_{r=1}^{H} N_r(N_r - n_r)\frac{\sigma_r^2}{n_r} \right]$
- $\sigma^2_{(1)} = \frac{1}{2H(H-1)} \sum_{j \neq k} (T_j - T_k)^2 = \frac{1}{H-1} \sum_{j=1}^{H} (T_j - \mu_T)^2,$
  $\mu_T = \frac{1}{H} \sum_{r=1}^{H} T_r.$

Two-stage Sampling

WOR + WOR scheme: fixed costs optimal allocation

Let the budget C for a survey be fixed.

Two-stage Sampling

WOR + WOR scheme: fixed costs optimal allocation

Let the budget C for a survey be fixed.

Let c_h be the cost of observing Y in h-th stratum, $h = 1, \ldots, H$

Two-stage Sampling

WOR + WOR scheme: fixed costs optimal allocation

Let the budget C for a survey be fixed.
Let c_h be the cost of observing Y in h-th stratum, $h = 1, \ldots, H$
The problem: find $1 < M < H$ and $n_1, \ldots, n_H$ minimizing the variance

$$Var(\bar{y}_{(2)}) = \frac{H}{MN^2}\left[(H-M)\sigma^2_{(1)} + \sum_{r=1}^{H} N_r(N_r - n_r)\frac{\sigma_r^2}{n_r}\right]$$

due to constraint

$$E\left[\sum_{h=1}^{M} n_{(h)}c_{(h)}\right] \leq C$$

Two-stage Sampling

WOR + WOR scheme: fixed costs optimal allocation (simplification)

- Assumption: for all strata $\frac{n_h}{N_h} = f_2 = const$

Two-stage Sampling

WOR + WOR scheme: fixed costs optimal allocation (simplification)

- Assumption: for all strata $\frac{n_h}{N_h} = f_2 = const$
- $Var(\bar{y}_{(2)}) = \frac{1}{N^2}\frac{H^2}{M}\left[\left(1-\frac{M}{H}\right)\sigma^2_{(1)} + \frac{1-f_2}{f_2}\bar{N}\sigma^2_{(2)}\right]$

Two-stage Sampling

WOR + WOR scheme: fixed costs optimal allocation (simplification)

- Assumption: for all strata $\frac{n_h}{N_h} = f_2 = const$
- $Var(\bar{y}_{(2)}) = \frac{1}{N^2}\frac{H^2}{M}\left[\left(1 - \frac{M}{H}\right)\sigma^2_{(1)} + \frac{1-f_2}{f_2}\bar{N}\sigma^2_{(2)}\right]$
- $\sigma^2_{(2)} = \sum_{h=1}^{H} W_h\sigma^2_h; \quad \bar{N} = \frac{N}{H}$

Two-stage Sampling

WOR + WOR scheme: fixed costs optimal allocation (simplification)

- The cost of drawing stratum: c_1 (constant)

Two-stage Sampling

WOR + WOR scheme: fixed costs optimal allocation (simplification)

- The cost of drawing stratum: c_1 (constant)
- The cost of observing Y in a stratum: c_2 (constant)

Two-stage Sampling

WOR + WOR scheme: fixed costs optimal allocation (simplification)

- The cost of drawing stratum: c_1 (constant)
- The cost of observing Y in a stratum: c_2 (constant)
- The cost of survey $\hat{C} = Mc_1 + c_2 \sum_{h=1}^{M} n_{(h)}$

Two-stage Sampling

WOR + WOR scheme: fixed costs optimal allocation (simplification)

- The cost of drawing stratum: c_1 (constant)
- The cost of observing Y in a stratum: c_2 (constant)
- The cost of survey $\hat{C} = Mc_1 + c_2 \sum_{h=1}^{M} n_{(h)}$
- $E(\hat{C}) = M(c_1 + f_2 \bar{N} c_2)$

Two-stage Sampling

WOR + WOR scheme: fixed costs optimal allocation (simplification)

- The cost of drawing stratum: c_1 (constant)
- The cost of observing Y in a stratum: c_2 (constant)
- The cost of survey $\hat{C} = Mc_1 + c_2 \sum_{h=1}^{M} n_{(h)}$
- $E(\hat{C}) = M(c_1 + f_2 \bar{N} c_2)$
- Constraint: $E(\hat{C}) \leq C$

Two-stage Sampling

WOR + WOR scheme: fixed costs optimal allocation (simplification)

The problem: find M and f_2 minimizing the variance

$$Var(\bar{y}_{(2)}) = \frac{1}{N^2}\frac{H^2}{M}\left[\left(1-\frac{M}{H}\right)\sigma^2_{(1)} + \frac{1-f_2}{f_2}\bar{N}\sigma^2_{(2)}\right]$$

due to constraint

$$M(c_1 + f_2\bar{N}c_2) \leq C$$

Two-stage Sampling

WOR + WOR scheme: fixed costs optimal allocation (simplification)

$$f_2^\star = \sqrt{\frac{c_1 \sigma_{(2)}^2}{c_2(\sigma_{(1)}^2 - \bar{N}\sigma_{(2)}^2)}}$$

$$M^\star = \frac{C}{c_1 + c_2 \bar{N} f_2^\star}$$

Two-stage Sampling

WOR + WOR scheme: fixed costs optimal allocation (simplification)

$$f_2^\star = \sqrt{\frac{c_1 \sigma_{(2)}^2}{c_2(\sigma_{(1)}^2 - \bar{N}\sigma_{(2)}^2)}}$$

$$M^\star = \frac{C}{c_1 + c_2 \bar{N} f_2^\star}$$

Two-stage Sampling

WOR + WOR scheme: fixed costs optimal allocation (simplification)

$$f_2^\star = \sqrt{\frac{c_1 \sigma_{(2)}^2}{c_2(\sigma_{(1)}^2 - \bar{N}\sigma_{(2)}^2)}}$$

$$M^\star = \frac{C}{c_1 + c_2 \bar{N} f_2^\star}$$

Apply Lagrange multipliers

Two-stage Sampling

WOR + WOR scheme: fixed costs optimal allocation (simplification)

Unbaised estimator for $\sigma^2_{(2)}$:

$$\hat{\sigma}^2_{(2)} = \frac{H}{M}\sum_{h=1}^{M} W_h s_h^2$$

Unbaised estimator for $\sigma^2_{(1)}$:

$$\hat{\sigma}^2_{(1)} = \frac{1}{M-1}\sum_{r=1}^{M}(t_r - \bar{t}_M)^2 - \frac{1}{M}\sum_{r=1}^{M} N_r(N_r - n_r)\frac{s_r^2}{n_r},$$

where $t_r = N_r\bar{y}_r,\ \bar{t}_M = \frac{1}{M}\sum_{r=1}^{M} t_r$

Two-stage Sampling

WOR + WOR scheme: fixed costs optimal allocation (simplification)

Unbaised estimator for $\sigma^2_{(2)}$:

$$\hat{\sigma}^2_{(2)} = \frac{H}{M} \sum_{h=1}^{M} W_h s_h^2$$

Unbaised estimator for $\sigma^2_{(1)}$: /with $\frac{n_h}{N_h} = f_2 = const$/

$$\hat{\sigma}^2_{(1)} = \frac{1}{M-1} \sum_{r=1}^{M} (t_r - \bar{t}_M)^2 - \frac{1-f_2}{f_2} \bar{N} \hat{\sigma}^2_{(2)},$$

where $t_r = N_r \bar{y}_r$, $\bar{t}_M = \frac{1}{M} \sum_{r=1}^{M} t_r$

Multi-stage sampling

Two-stage sampling

At the first stage a sample $\mathcal{L}$ of PSU labels is taken. Then, independently for each $r \in \mathcal{L}$, a sample s_r of $n(s_r)$ elementary units is selected from $\mathcal{U}_r$ according to some scheme. Using this notation, the total sample is

$$s = \sum_{r \in \mathcal{L}} s_r$$

and $n(s) = \sum_{r \in \mathcal{L}} n(s_r)$.
The first-stage inclusion probabilities is be defined by

$$\Pi_r = P(r \in \mathcal{L})$$

The conditional inclusion probability, i.e. the probability that j is in s_r given that r is in $\mathcal{L}$, will be denoted by

Multi-stage Sampling

Exercise 18

Show that in this case HT estimator of mean μ is $\hat{\mu}_{HT} = \frac{1}{M} \sum_{r \in \mathcal{L}} \bar{y}_r$.

Multi-stage sampling

Two-stage sampling

Suppose that the design at the first stage of sampling chooses a fixed number M of PSUs and inclusion probability Π_r is proportional to the size of U_r for every $r \in \mathcal{L}$, and that if $r \in \mathcal{L}$, sampling takes place within U_r, by SRS without replacement, n_r, draws. It can be shown, that

$$Var(\hat{\mu}_{HT}) =$$

$$= \frac{1}{M^2} \sum_{r=1}^{H} \frac{\Pi_r}{n_r} \left(1 - \frac{n_r}{N_r}\right) S_r^2 + \frac{1}{2} \sum_{j \neq k} (\Pi_j \Pi_k - \Pi_{jk}) \left(\frac{T_j}{\Pi_j} - \frac{T_k}{\Pi_k}\right)^2,$$

where $T_j = \sum_{r \in U_j} Y_r$ and $S_r^2 = \sum_{r \in U_j} (Y_j - \bar{Y}_r)^2$, $\bar{Y}_r = T_r / N_r$.

Estimation over subpopulations

Introduction

It may often be impossible to obtain a frame that lists only those units in the population which are of interest. For instance, the investigator is interested in sampling households, where both husband and wife work, or he/she wants to sample households having adults over 50 years of age. However, the best frame available in both the cases is the list of all households in the target area. In this case, before any sample unit is observed, the investigator has no way of knowing whether any particular selected unit is a member of the subpopulation under consideration or not.

Estimation over subpopulations

Notions

- N: the total number of units in the population

Estimation over subpopulations

Notions

- N: the total number of units in the population
- N_1: the number of units in the subpopulation of interest

Estimation over subpopulations

Notions

- N: the total number of units in the population
- N_1: the number of units in the subpopulation of interest
- n: the number of units in the WOR simple random sample drawn from the population of size N

Estimation over subpopulations

Notions

- N: the total number of units in the population
- N_1: the number of units in the subpopulation of interest
- n: the number of units in the WOR simple random sample drawn from the population of size N
- n_1: the number of units in the sample of size n that belong to the subpopulation under consideration

Estimation over subpopulations

Notions

- N: the total number of units in the population
- N_1: the number of units in the subpopulation of interest
- n: the number of units in the WOR simple random sample drawn from the population of size N
- n_1: the number of units in the sample of size n that belong to the subpopulation under consideration
- Y_{si}: the value of study variable Y for the i-th unit of the subpopulation

Estimation over subpopulations

Notions

- N: the total number of units in the population
- N_1: the number of units in the subpopulation of interest
- n: the number of units in the WOR simple random sample drawn from the population of size N
- n_1: the number of units in the sample of size n that belong to the subpopulation under consideration
- Y_{si}: the value of study variable Y for the i-th unit of the subpopulation
- y_{si}: the value of Y for the i-th sample unit from the subpopulation

Estimation over subpopulations

Notions

The mean, total and mean square error for the target subpopulation are given by

- $\bar{Y}_s = \frac{1}{N_1} \sum_{i=1}^{N_1} Y_{si}$

Estimation over subpopulations

Notions

The mean, total and mean square error for the target subpopulation are given by

- $\bar{Y}_s = \frac{1}{N_1} \sum_{i=1}^{N_1} Y_{si}$
- $Y_s = \sum_{i=1}^{N_1} Y_{si}$

Estimation over subpopulations

Notions

The mean, total and mean square error for the target subpopulation are given by

- $\bar{Y}_s = \frac{1}{N_1} \sum_{i=1}^{N_1} Y_{si}$
- $Y_s = \sum_{i=1}^{N_1} Y_{si}$
- $S_s^2 = \frac{1}{N_1 - 1} \sum_{i=1}^{N_1} (Y_{si} - \bar{Y}_s)^2$

Mean estimation in subpopulation

Mean estimation

Let $s_s^2 = \frac{1}{n_1-1}\sum_{i=1}^{n_1}(y_{si} - \bar{y}_s)^2$

- Unbiased estimator of the subpopulation mean $\bar{Y}_s$ when N_1 is known:

$$\bar{y}_s = \frac{1}{n_1}\sum_{i=1}^{n_1} y_{si}$$

Mean estimation in subpopulation

Mean estimation

Let $s_s^2 = \frac{1}{n_1-1}\sum_{i=1}^{n_1}(y_{si} - \bar{y}_s)^2$

- Unbiased estimator of the subpopulation mean $\bar{Y}_s$ when N_1 is known:

$$\bar{y}_s = \frac{1}{n_1}\sum_{i=1}^{n_1} y_{si}$$

- Variance of estimator $\bar{y}_s$: $Var(\bar{y}_s) = [E(1/n_1) - 1/N_1]S_s^2$

Mean estimation in subpopulation

Mean estimation

Let $s_s^2 = \frac{1}{n_1-1}\sum_{i=1}^{n_1}(y_{si} - \bar{y}_s)^2$

- Unbiased estimator of the subpopulation mean $\bar{Y}_s$ when N_1 is known:

$$\bar{y}_s = \frac{1}{n_1}\sum_{i=1}^{n_1} y_{si}$$

- Variance of estimator $\bar{y}_s$: $Var(\bar{y}_s) = [E(1/n_1) - 1/N_1]S_s^2$
- Estimator of variance $Var(\bar{y}_s)$: $v(\bar{y}_s) = \left(\frac{1}{n_1} - \frac{1}{N_1}\right)s_s^2$

Mean estimation in subpopulation

Mean estimation

Let $s_s^2 = \frac{1}{n_1-1}\sum_{i=1}^{n_1}(y_{si} - \bar{y}_s)^2$

- Unbiased estimator of the subpopulation mean $\bar{Y}_s$ when N_1 is known:

$$\bar{y}_s = \frac{1}{n_1}\sum_{i=1}^{n_1} y_{si}$$

- Variance of estimator $\bar{y}_s$: $Var(\bar{y}_s) = [E(1/n_1) - 1/N_1]S_s^2$
- Estimator of variance $Var(\bar{y}_s)$: $v(\bar{y}_s) = \left(\frac{1}{n_1} - \frac{1}{N_1}\right)s_s^2$

Remark

In case of unknown N_1 it may be substituted by $n_1 N/n$.

Mean estimation in subpopulation

Exercise 19

The family planning wing of the health department of a certain state wishes to conduct a survey at a university campus for estimating the average time gap (in months) between the births of children in families having two children. The frame available, of course, lists all the 800 families of the campus. As the prior identification of the families in the population having just two children was difficult, the investigator selected a WOR random sample of 80 families. In the sample families, 32 families were found having two children. These 32 families were interviewed, and the information collected is shown in the following table

Mean estimation in subpopulation

Exercise 19 cont.

Family	Gap	Family	Gap	Family	Gap	Family	Gap
1	24	9	64	17	57	25	42
2	30	10	32	18	65	26	16
3	50	11	58	19	26	27	37
4	41	12	48	20	35	28	61
5	27	13	51	21	31	29	34
6	47	14	22	22	17	30	29
7	47	15	69	23	28	31	19
8	39	16	54	24	55	32	57

Mean estimation in subpopulation

Exercise 19 cont.

Estimate the average gap between the births of two children, and obtain confidence limits for it.

Note that we have $n_1 = 32, n = 80,$ and $N = 800$

Proportion estimation in subpopulation

Proportion estimation

Let N_1' be the number of units possessing the attribute of interest out of N_1 units

Proportion estimation in subpopulation

Proportion estimation

Let N_1' be the number of units possessing the attribute of interest out of N_1 units

Let n_1' be the number of units possessing the attribute of interest out of n_1

Proportion estimation in subpopulation

Proportion estimation

Let N_1' be the number of units possessing the attribute of interest out of N_1 units

Let n_1' be the number of units possessing the attribute of interest out of n_1

Let

$$\theta_s = \frac{N_1'}{N_1}$$

Proportion estimation in subpopulation

Proportion estimation

- Unbiased estimator of proportion θ_s in the subpopulation for known N_1: $\hat{\theta}_s = \frac{n_1'}{n_1}$

Proportion estimation in subpopulation

Proportion estimation

- Unbiased estimator of proportion θ_s in the subpopulation for known N_1: $\hat{\theta}_s = \frac{n_1'}{n_1}$
- Variance of the estimator $\hat{\theta}_s$:
  $Var(\hat{\theta}_s) = \left[N_1 E\left(\frac{1}{n_1}\right) - 1 \right] \frac{\theta_s(1-\theta_s)}{N_1 - 1}$

Proportion estimation in subpopulation

Proportion estimation

- Unbiased estimator of proportion θ_s in the subpopulation for known N_1: $\hat{\theta}_s = \frac{n_1'}{n_1}$
- Variance of the estimator $\hat{\theta}_s$:
  $$Var(\hat{\theta}_s) = \left[N_1 E\left(\frac{1}{n_1}\right) - 1 \right] \frac{\theta_s(1-\theta_s)}{N_1 - 1}$$
- Estimator of the variance $Var(\hat{\theta}_s)$:
  $$v(\theta_s) = \left(1 - \frac{n_1}{N_1}\right) \frac{\hat{\theta}_s(1-\hat{\theta}_s)}{N_1 - 1}$$

Proportion estimation in subpopulation

Exercise 20

The problem: estimate the proportion of men over 70 years of age who are still professionally active.

Proportion estimation in subpopulation

Exercise 20

The problem: estimate the proportion of men over 70 years of age who are still professionally active.
The frame of such individuals (above 70 years and alive) is not readily available.

Proportion estimation in subpopulation

Exercise 20

The problem: estimate the proportion of men over 70 years of age who are still professionally active.
The frame of such individuals (above 70 years and alive) is not readily available.
Available frame: a voters' list prepared five years back.

Proportion estimation in subpopulation

Exercise 20

The problem: estimate the proportion of men over 70 years of age who are still professionally active.
The frame of such individuals (above 70 years and alive) is not readily available.
Available frame: a voters' list prepared five years back.
The frame consisting of men expected to cross their 70th year could be prepared.

Proportion estimation in subpopulation

Exercise 20

The problem: estimate the proportion of men over 70 years of age who are still professionally active.
The frame of such individuals (above 70 years and alive) is not readily available.
Available frame: a voters' list prepared five years back.
The frame consisting of men expected to cross their 70th year could be prepared.
But, it could be possible that some of the individuals included in the frame are no more alive.

Proportion estimation in subpopulation

Exercise 20 cont.

Prepared frame: 1500 men expected to cross their 70th year.

Proportion estimation in subpopulation

Exercise 20 cont.

Prepared frame: 1500 men expected to cross their 70th year.
A sample of $n = 120$ persons was drawn from this frame following WOR .

Proportion estimation in subpopulation

Exercise 20 cont.

Prepared frame: 1500 men expected to cross their 70th year.
A sample of $n = 120$ persons was drawn from this frame following WOR .
Out of these, 14 individuals were found to have died.

Proportion estimation in subpopulation

Exercise 20 cont.

Prepared frame: 1500 men expected to cross their 70th year.
A sample of $n = 120$ persons was drawn from this frame following WOR .
Out of these, 14 individuals were found to have died.
Sample size: 106 persons.

Proportion estimation in subpopulation

Exercise 20 cont.

Prepared frame: 1500 men expected to cross their 70th year.
A sample of $n = 120$ persons was drawn from this frame following WOR .
Out of these, 14 individuals were found to have died.
Sample size: 106 persons.
Number of professionally active: 21 persons.

Proportion estimation in subpopulation

Exercise 20 cont.

Prepared frame: 1500 men expected to cross their 70th year.
A sample of $n = 120$ persons was drawn from this frame following WOR .
Out of these, 14 individuals were found to have died.
Sample size: 106 persons.
Number of professionally active: 21 persons.

Estimate the proportion in question, and derive the confidence interval for this parameter.

Probability proportional to size sampling (PPS)

Introduction

According to simple random sampling each unit in the population gets equal chance of being included in the sample. However, when the units vary considerably in size, simple random sampling does not seem to be an appropriate procedure, since it does not take into account the possible importance of the size of the unit. Under such circumstances, selection of units with unequal probabilities may provide more efficient estimators than equal probability sampling.

Probability proportional to size sampling (PPS)

Introduction

Let Y be the study variable.

Probability proportional to size sampling (PPS)

Introduction

Let Y be the study variable.
Let X be an auxiliary variable measuring the size of a unit in population.

Probability proportional to size sampling (PPS)

Introduction

Let Y be the study variable.
Let X be an auxiliary variable measuring the size of a unit in population.
Let the probability of selecting a unit to the sample is proportional to the value of X.

Probability proportional to size sampling (PPS)

Introduction

Let Y be the study variable.
Let X be an auxiliary variable measuring the size of a unit in population.
Let the probability of selecting a unit to the sample is proportional to the value of X.

This type of sampling is known as varying probability sampling or probability proportional to size (PPS) sampling.

Probability proportional to size sampling (PPS)

Cumulative Total Method

- Let X_i be the size of the i-th unit.

Probability proportional to size sampling (PPS)

Cumulative Total Method

- Let X_i be the size of the i-th unit.
- Let $X = \sum_{i=1}^{N} X_i$.

Probability proportional to size sampling (PPS)

Cumulative Total Method

- Let X_i be the size of the i-th unit.
- Let $X = \sum_{i=1}^{N} X_i$.
- Let $X_{1:N} \leq X_{2:N} \leq X_{N:N}$

Probability proportional to size sampling (PPS)

Cumulative Total Method

- Let X_i be the size of the i-th unit.
- Let $X = \sum_{i=1}^{N} X_i$.
- Let $X_{1:N} \leq X_{2:N} \leq X_{N:N}$
- Let $T_k = \sum_{j=1}^{k} X_{j:N}$

Probability proportional to size sampling (PPS)

Cumulative Total Method

- Let X_i be the size of the i-th unit.
- Let $X = \sum_{i=1}^{N} X_i$.
- Let $X_{1:N} \leq X_{2:N} \leq X_{N:N}$
- Let $T_k = \sum_{j=1}^{k} X_{j:N}$
- Choose a random number $r \in (0, X]$

Probability proportional to size sampling (PPS)

Cumulative Total Method

- Let X_i be the size of the i-th unit.
- Let $X = \sum_{i=1}^{N} X_i$.
- Let $X_{1:N} \leq X_{2:N} \leq X_{N:N}$
- Let $T_k = \sum_{j=1}^{k} X_{j:N}$
- Choose a random number $r \in (0, X]$
- Find i such $T_{i-1} < r \leq T_i$ and select i-th unit to a sample

Probability proportional to size sampling (PPS)

Cumulative Total Method

- Let X_i be the size of the i-th unit.
- Let $X = \sum_{i=1}^{N} X_i$.
- Let $X_{1:N} \leq X_{2:N} \leq X_{N:N}$
- Let $T_k = \sum_{j=1}^{k} X_{j:N}$
- Choose a random number $r \in (0, X]$
- Find i such $T_{i-1} < r \leq T_i$ and select i-th unit to a sample
- The probability of selecting the i-th population unit equals $p_i = X_i/X$.

Probability proportional to size sampling (PPS)

Lahiri's Method

1. Select a random number i from 1 to N.

Probability proportional to size sampling (PPS)

Lahiri's Method

1. Select a random number i from 1 to N.
2. Select a random number j from 1 to M, where $M \geq \max\{X_1, \ldots, X_N\}$.

Probability proportional to size sampling (PPS)

Lahiri's Method

1. Select a random number i from 1 to N.
2. Select a random number j from 1 to M, where $M \geq \max\{X_1, \ldots, X_N\}$.
3. Repeat steps (1) and (2) until $j \leq X_i$.

Probability proportional to size sampling (PPS)

Lahiri's Method

1. Select a random number i from 1 to N.
2. Select a random number j from 1 to M, where $M \geq \max\{X_1, \ldots, X_N\}$.
3. Repeat steps (1) and (2) until $j \leq X_i$.
4. Select the i-th unit to the sample.

Probability proportional to size sampling (PPS)

Estimation in PPS WR Sampling

- Unbiased estimator of population total Y:

$$\hat{Y}_{pps} = \frac{1}{n} \sum_{i=1}^{n} \frac{y_i}{p_i}$$

Probability proportional to size sampling (PPS)

Estimation in PPS WR Sampling

- Unbiased estimator of population total Y:

$$\hat{Y}_{pps} = \frac{1}{n} \sum_{i=1}^{n} \frac{y_i}{p_i}$$

- Variance of the estimator $\hat{Y}_{pps}$:

$$Var(\hat{Y}_{pps}) = \frac{1}{n} \left(\sum_{i=1}^{N} \frac{Y_i^2}{p_i} - Y^2 \right) = \frac{1}{n} \sum_{i=1}^{N} \left(\frac{Y_i}{p_i} - Y \right)^2 p_i$$

Probability proportional to size sampling (PPS)

Estimation in PPS WR Sampling

- Unbiased estimator of the variance $Var(\hat{Y}_{pps})$:

$$v(\hat{Y}_{pps}) = \frac{1}{n(n-1)} \left(\sum_{i=1}^{n} \frac{y_i^2}{p_i^2} - n\hat{Y}_{pps}^2 \right)$$
$$= \frac{1}{n(n-1)} \sum_{i=1}^{n} \left(\frac{y_i}{p_i} - \hat{Y}_{pps} \right)^2$$

Probability proportional to size sampling (PPS)

Estimation in PPS WR Sampling

- Unbiased estimator of the variance $Var(\hat{Y}_{pps})$:

$$v(\hat{Y}_{pps}) = \frac{1}{n(n-1)} \left(\sum_{i=1}^{n} \frac{y_i^2}{p_i^2} - n\hat{Y}_{pps}^2 \right)$$

$$= \frac{1}{n(n-1)} \sum_{i=1}^{n} \left(\frac{y_i}{p_i} - \hat{Y}_{pps} \right)^2$$

$$p_i = \frac{X_i}{X}$$