SURVEY SAMPLING – INTRODUCTION

Updated: May 6, 2019

Stanisław Jaworski

Department of Econometrics and Statistics
Warsaw University of Life Sciences
Nowoursynowska 159, PL-02-787 Warszawa
e-mail: stanislaw_jaworski@sggw.pl

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).

Contents

Types of data

- The data collected by the investigator from the original source are called **primary data**.
- If the required data had already been collected by some agencies or individuals and are now available in the published or unpublished records, these are known as **secondary data**.

Some 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 nonoverlapping 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**.
- 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**.
- 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**.
- 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.

Exercise 1.
Given below are the weights (in pounds) of 4 children at the time of birth in a hospital:

Child	A	B	C	D
Weight	5.5	8.0	6.5	7.0

a) Enumerate all possible WR samples of size 2. Also, write values of the study variable (weight) for the sample units.

b) Enumerate all possible WOR samples of size 2, and also list the weight values for the respective sample units.

Exercise 2.
Consider a population consisting of 6 villages, the areas (in hectares) of which are given below :

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

a) Enumerate all possible WR samples of size 3. Also, write the values ofthe study variable for the sampled units.

b) List all the WOR samples of size 4 along with their area values.

Estimator and its sampling distribution

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

  For instance, if $Y_1, Y_2, \ldots, Y_N$ are the values of the variable Y for the N units in the population, then

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

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

- 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**.

For instance, if $y_1, y_2, \ldots, y_n$ are the values of the variable Y for the n units in the sample, then

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

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

- 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.

- 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.

- The estimator $\hat{\theta}$ is said to be **unbiased** for the parameter θ, if $E\hat{\theta} = \theta$.

- If for an estimator $\hat{\theta}$, $E(\hat{\theta}) \neq \theta$ the estimator $\hat{\theta}$ is called a **biased estimator** of θ. The magnitude of the bias in $\hat{\theta}$ is given by $B(\hat{\theta}) = E(\hat{\theta}) - \theta$.

- The **sampling variance** is the variance of the sampling distribution of an estimator.

  Let $Var(\hat{\theta})$ denotes the variance of the sampling distribution of $\hat{\theta}$. Then $Var(\hat{\theta}) = E(\hat{\theta} - E\hat{\theta})^2$

- The **mean square error** (MSE) measures the divergence of the estimator values from the true parameter value. This can be put as

$$MSE(\hat{\theta}) = E(\hat{\theta} - \theta)^2.$$

- Let $\hat{\theta}_1$ and $\hat{\theta}_2$ be two estimators of the parameter θ. The **relative efficiency** of the estimator θ_2 with respect to the estimator θ_1 is defined as

$$RE = \frac{MSE(\hat{\theta}_1)}{MSE(\hat{\theta}_2)}.$$

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 equal probabilities and WR.

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

b) Find the sampling variance of $\bar{y}$.

c) Calculate population mean μ and find $MSE(\bar{y}) = E(\bar{y} - \mu)^2$.

d) Draw a cumulative distribution function of the average milk yield.

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

f) 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 $\hat{\sigma}^2$.

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 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

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

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

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

Simple random sampling

Estimation of population mean and total

Example 1. (Simple random sampling with replacement)
Unbiased estimator of population mean μ: $\hat{\mu} = \bar{y} = \sum_{i=1}^{n} y_i/n$.
Sampling variance of $\bar{y}$: $Var(\hat{\mu}) = \frac{1}{n}\sigma^2$.
Unbiased estimator of sampling variance $v(\hat{\mu}) = \frac{1}{n}s^2$

PROOF OF UNBIASEDNESS OF THE SAMPLE MEAN. Consider a survey population U whose units are labelled $1, \ldots, N$. Thus in the notation of sets $U = \{1, \ldots, N\}$. Denote the **sample sequence** of unit labels from U by $s = (i_1, \ldots, i_n)$, where n is a fixed sample size. Then $p(s) = 1/N^n$ and for the given s we have $\bar{y} = \frac{1}{n}(Y_{i_1} + \cdots + Y_{i_n}) = \frac{1}{n}\sum_{i=1}^{N} Y_i \mathbb{1}(i \in s)$. So

$$
\begin{aligned}
nE(\bar{y}) &= \sum_{i=1}^{N} Y_i E\mathbb{1}(i \in s) = \sum_{i=1}^{N} Y_i P(\{s : i \in s\}) = \\
&= \sum_{i=1}^{N} Y_i \sum_{s:\ i\in s} p(s) = \sum_{i=1}^{N} Y_i \sum_{s:\ i\in s} \frac{1}{N^n} = \\
&= \sum_{i=1}^{N} Y_i n \cdot N^{n-1} \cdot \frac{1}{N^n} = \frac{n}{N}\sum_{i=1}^{N} Y_i = n\bar{Y}
\end{aligned}
$$

■

Example 2. (Simple random sampling without replacement)
Unbiased estimator of population mean μ: $\hat{\mu} = \bar{y} = \sum_{i=1}^{n} y_i/n$.
Sampling variance of $\bar{y}$: $Var(\hat{\mu}) = (\frac{1}{n} - \frac{1}{N})\sigma^2$.
Unbiased estimator of sampling variance: $v(\hat{\mu}) = (\frac{1}{n} - \frac{1}{N})s^2$

PROOF OF UNBIASEDNESS OF THE SAMPLE MEAN. Denote a **sample sequence** of unit labels from U by $s = \{i_1, \ldots, i_n\}$, where n is a fixed sample size. Then $p(s) = 1/\binom{N}{n}$ and for the given s we have $\bar{y} = \frac{1}{n}(Y_{i_1} + \cdots + Y_{i_n}) =$

$\frac{1}{n}\sum_{i=1}^{N} Y_i \mathbb{1}(i \in s)$. So

$$nE(\bar{y}) = \sum_{i=1}^{N} Y_i E\mathbb{1}(i \in s) = \sum_{i=1}^{N} Y_i P(\{s : i \in s\}) =$$
$$= \sum_{i=1}^{N} Y_i \sum_{s:\ i\in s} p(s) = \sum_{i=1}^{N} Y_i \sum_{s:\ i\in s} \frac{1}{\binom{N}{n}} =$$
$$= \sum_{i=1}^{N} Y_i \cdot \binom{N-1}{n-1} \cdot \frac{1}{\binom{N}{n}} = \frac{n}{N}\sum_{i=1}^{N} Y_i = n\bar{Y}$$

■

SAMPLING VARIANCE OF $\bar{y}$ (DERIVATION OF THE FORMULA).

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

Note that

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

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

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

So

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

■

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.

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.

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 at all normal?

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?

Exercise 9.
Estimate the total number of tractors in the development block of 69 villages using the samples selected in Exercise 5.

Example 3. (Simple random sampling with replacement. 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.

Estimator of population mean: $= \sum_{i=1}^{d} y_i/d$.
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 $s_d^2 = \frac{1}{d-1}(y_i - \bar{y}_d)^2$

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.

Exercise 11.
A simple random sample of 30 housholds was drawn from a city area containing 14848 hosholds. The numbers of persons per household in the sample were as follows:

$$5, 6, 3, 3, 2, 3, 3, 3, 4, 4, 3, 2, 7, 4, 3, 5, 4, 4, 3, 3, 4, 3, 3, 1, 2, 4, 3, 4, 2, 4.$$

Estimate the total number of people in the area.

Determining sample size for estimating population mean or total. The required sample size can be determined by a two step approach. A small preliminary sample is used to estimate the population parameter values, which in turn are used to determine final sample size. The preliminary sample is then augmented by drawing additional units from the population.

Let n_1 be the size of preliminary sample selected using SRS without replacement. Let $y = \frac{1}{n_1}\sum_{i=1}^{n_1} y_i$ and $s^2 = \frac{1}{n_1-1}\sum_{i=1}^{n_1}(y_i - \bar{y})^2$ be parameters of the preliminary sample. The required sample size is

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

where $u_{1-\alpha/2}$ is a $(1-\alpha/2)-$quantile of standard normal distribution, $v = s/\bar{y}$, d is the permissible error and $\delta = d/\bar{y}$.

PROOF.

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

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

where $Var(\bar{y}) = (1/n - 1/N)S^2 \approx (1/n - 1/N)s^2$. Hence $d = u_{1-\alpha/2}(1/n_1 - 1/N)s^2$ and we can solve the resulting equation with respect to n. ■

Exercise 12.
In a study of the possible use of sampling to cut down the work in taking inventory in a stock room, a cont is made of the value of the articles on each of 36 shelves in the room.The values to the nearest dollar are as follows:

$$29, 38, 42, 44, 45, 47, 51, 53, 53, 54, 56, 56, 56, 58, 58, 59, 60, 60,$$
$$60, 60, 61, 61, 61, 62, 64, 65, 65, 67, 67, 68, 69, 71, 74, 77, 82, 85.$$

The estimate of total value made from a sample is to be correct within \$200, apart from a 1 in 20 chance. An advisor suggests that a simple random sample of 12 shelves will meet the requirements. Do you agree?

Exercise 13.
The owner of a poultry farm is interested in estimating the total weight gain, in a period of one month, for $N = 1500$ chicks kept on a new feed. For this purpose, a simple random WOR sample of $n = 25$ chicks is observed for weight gain. The sample data yielded $s^2 = 45$ gm^2. Determine the sample

size required to estimate total weight gain with two kg (=2000 gm) as margin of error.

Estimation of proportion

Let K units out of N possess the attribute of interest. Then population proportion $\theta = K/N$. If k units out of n sample units possess this attribute, sample proportion is given by $\hat{\theta} = k/n$.

Example 4. (Simple random sampling with replacement)
Unbiased estimator of population proportion θ: $\hat{\theta} = k/n$.
Sampling variance of $\hat{\theta}$: $Var(\hat{\theta}) = \frac{\theta(1-\theta)}{n}$.
Unbiased estimator of sampling variance: $v(\hat{\theta}) = \frac{\hat{\theta}(1-\hat{\theta})}{n-1}$

Example 5. (Simple random sampling without replacement)
Unbiased estimator of population proportion θ: $\hat{\theta} = k/n$.
Sampling variance of $\hat{\theta}$: $Var(\hat{\theta}) = \frac{N-n}{N-1} \cdot \frac{\theta(1-\theta)}{n}$.
Unbiased estimator of sampling variance: $v(\hat{\theta}) = (1 - \frac{n}{N})\frac{\hat{\theta}(1-\hat{\theta})}{n-1}$

Example 6. (Inverse sampling)
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.

Unbiased estimator of population proportion θ: $\hat{\theta} = \frac{m-1}{n-1}$.
Estimator of sampling variance : $\frac{\hat{\theta}(1-\hat{\theta})}{n-2}(1 - \frac{n-1}{N})$ (when sampling without replacement)
Estimator of sampling variance: $\frac{\hat{\theta}(1-\hat{\theta})}{n-2}$ (when sampling with replacement)

Exercise 14.
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 with replacement 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 suitable, and 1 if he/she does. From the sample observations given below, estimate the proportion along with the standard error of your estimate. Also, work out the confidence interval for the proportion

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

Exercise 15.
A car dealer is feeling concerned over the complaints received in the office of the manufacturer regarding the free service provided by him to the newly purchased cars. To assess the seriousness of the problem, the dealer decided to draw a **WR** random sample of 70 buyers out of the total of 1400 individuals who had purchased cars through him during the last one year. Twenty one buyers included in the sample graded service provided by him as unsatisfactory. Estimate the percentage of buyers feeling unsatisfied with the service provided, and construct a suitable level confidence interval for it.

Exercise 16.
An investigator wishes to estimate the proportion of students in a university whose fathers are graduates. To arrive at the estimate, a **WOR** simple random sample of 67 students was drawn from a total of 1400 students. On contacting the sampled students, it was found that the fathers of 46 students had not graduated. Estimate the proportion of students whose fathers were at least graduates. Also, set the confidence interval for population proportion.

Exercise 17.
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 ? 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.

Exercise 18.
The Mayor of a municipal corporation noticed an error in the calculations of general provident fund (GPF) account of an employee. Fearing that such errors might have also crept in the calculations of some other GPF accounts, he directed the audit unit of the corporation to estimate the proportion of such accounts. Expecting that the percentage of such accounts could be quite small, the investigator followed inverse sampling approach. He decided to go on sampling the accounts, using **WOR** method, from the list of all GPF accounts till 5 wrongly calculated accounts were detected. To arrive at this predetermined number of 5, he had to select 60 accounts out of a total of 1200 GPF accounts. Estimate the proportion of wrongly calculated accounts, and also find the confidence interval for it.

The mathematics of probability sampling designs

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

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

- The **inclusion probability** of unit j is defined to be the probability that j appears in the sample drawn. It is denoted by π_j and in terms of the $p(s)$ probabilities it is

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

- For distinct units j and k, let π_{jk} denote the probability that both j and k appear in the sample. Then, in terms of the $p(s)$ probabilities,

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

  The probabilities π_{jk} are called the **joint inclusion probabilities** for pairs of units in the population.

Let $z_1, \ldots, z_N$ be the values for the units in U of some real- or vector- valued variate Z. Let E_p denote expectation with respect to the sampling design p. The sample sum of Z can be written $\sum_{j\in s} z_j$ and its design expectation, by the definition, is given by

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

- With $z_j \equiv 1$ it follows that $\sum_{j\in s} z_j = n(s)$ is the sample size $n(s)$ and

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

- The expectation of sample size for a given design is the sum of its inclusion probabilities. In particular, for a fixed size n design, the inclusion probabilities will sum to n: $E_p n = n = \sum_{j=1}^{N} \pi_j$

- A sampling design is called **self-weighting** if all its *inclusion probabilities* are equal. For the designs which are both *self-weighting* and of *fixed size* a sample mean is unbiased as an estimator of the corresponding population mean. This can be seen as follows. If $\bar{y}_s = \frac{1}{n}\sum_{j\in s} y_j$ is the sample mean then from $n = \sum_{j=1}^{N} \pi_j$ and $E_p(\sum_{j\in s} z_j) = \sum_{j=1}^{N} z_j \pi_j$ it is implied that

  1. $\pi_j = n/N$
  2. $E_p(\bar{y}_s) = \mu = \frac{1}{N}\sum_{j=1}^{N} Y_j$, by substituting $z_j = y_j/n$

- Note that $E_p(\sum_{j\in s} z_j/\pi_j) = \sum_{j=1}^{N} z_j$ if $\pi_j > 0,\ \forall j$. So (substituting $z_j = y_j/N$)

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

  is an unbiased estimator of $\mu = \frac{1}{N}\sum_{j=1}^{N} Y_j$. It is called the **Horvitz-Thompson estimator** (*HT estimator*). (The estimator is taken to have value 0 if s is empty.)

- Let's consider **stratified random sampling**. Particularly let U is the union of disjoint strata $U_1, \ldots, U_H$. For example, for a human

population the strata could be age-sex groupings. 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 U_h. In this case $\pi_j = n_h/N_h$ for $j \in U_h$. 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 y in the part of the sample coming from U_h.

It can be shown, that

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

and when the design is of fixed size n

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

PROOF.

$$\begin{aligned} Var_p\left(\sum_{j\in s} z_j\right) &= Var_p\left(\sum_{j=1}^{N} z_j \mathbb{1}(j\in s)\right) = \\ &= \sum_{j=1}^{N} z_j^2 Var_p \mathbb{1}(j\in s) + \sum_{j\ \neq\ k}^{n}\sum^{n} z_j z_k Cov(\mathbb{1}(j\in s), \mathbb{1}(k\in s)) \\ &= \sum_{j=1}^{N} z_j^2 \pi_j(1-\pi_j) + \sum_{j\ \neq\ k}^{N}\sum^{N} z_j z_k(\pi_{jk} - \pi_j\pi_k) \end{aligned}$$

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

$$\begin{aligned} -\sum_{k\neq j} Cov_p(\mathbb{1}(j\in s), \mathbb{1}(k\in s)) &= -Cov_p(\mathbb{1}(j\in s), n - \mathbb{1}(j\in s)) \\ &= Var_p(\mathbb{1}(j\in s)). \end{aligned}$$

So $Var_p(\mathbb{1}(j \in s)) = -\sum_{k \neq j}(\pi_{jk} - \pi_j\pi_k) = \sum_{k \neq j}(\pi_j\pi_k - \pi_{jk})$ and

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

■

Exercise 19.
Show that in a **stratified random sampling without repacement** if $n_h/N_h = n/N,\ \forall h$, then

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

(*in the case:* $n_h = nW_h,\ \forall h$, *we say that* ***the allocation is proportional***)

Exercise 20.
Using equation (2) calculate $Var(\bar{y})$ in simple random sampling without replacement design.

Exercise 21.
Show that for **Horvitz-Thompson estimator** holds

$$Var(\hat{\mu}_{HT}) = \frac{1}{N^2}\left[\sum_{j=1}^{N} Y_j^2\left(\frac{1}{\pi_j} - 1\right) + \sum_{j \neq k}^{N}\sum^{N}\left(\frac{\pi_{jk}}{\pi_j\pi_k} - 1\right) Y_j Y_k\right] \tag{3}$$

and if $n(s) = n$ for every sample s then

$$Var(\hat{\mu}_{HT}) = \frac{1}{2N^2}\sum_{j \neq k}^{N}\sum^{N}(\pi_j\pi_k - \pi_{jk})\left(\frac{Y_j}{\pi_j} - \frac{Y_k}{\pi_k}\right)^2 \tag{4}$$

Remark 1.
The HT estimator may be seriously deficient (see Thompson (1997): example 2.5 on page 18).

Estimation over subpopulations

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.

- 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

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

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

Example 7. (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 2.
In case of unknown N_1 it may be substituted by $n_1 N/n$.

Exercise 22.
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

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

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** $N = 800$

Example 8. (Proportion estimation)
The objective in certain situations could be to estimate the proportion of units in a subpopulation possessing a specific attribute. Let $\theta_s = \frac{N_1'}{N_1}$ and $\hat{\theta}_s = \frac{n_1'}{n_1}$, where n_1' and N_1' are the number of units possessing the attribute of interest out of n_1 and N_1 units respectively.

- Unbiased estimator of proportion 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 $Var(\hat{\theta}_s)$: $v(\hat{\theta}_s) = \left(1 - \frac{n_1}{N_1}\right) \frac{\hat{\theta}_s(1-\hat{\theta}_s)}{N_1 - 1}$

Exercise 23.
Sociologist is interested in estimating the proportion of men over 70 years of age who are still actively contributing towards family income by way of doing some kind of work (business, farming, job, etc.). However, the frame of such individuals (above 70 years and alive) is not readily available. Instead, a voters' list prepared five years back is available. In this case, the frame consisting of men expected to cross their 70th year could be prepared from the available voters' list. Since the voters' list was prepared five years back, it could be possible that some of the individuals included in the frame are no more. So the frame is prepared from a five years old voters' list which consists of 1500 men expected to cross their 70th year. Obviously, the frame also includes the names of those voters who expired before, or after reaching the age of 70 years, during the preceding five years period. A sample of $n = 120$ persons was drawn from this frame following **WOR** simple random sampling. Out of these, 14 individuals were found to have died. On interviewing the remaining 106 persons, it was observed that 21 persons were still actively engaged in earning by doing some kind of work. Estimate the proportion in question, and also obtain the confidence interval for this parameter.

Sampling with varying probabilities

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. In this scheme, the units are selected with probability proportional to a given measure of size. The size measure is the value of an auxiliary

variable (say) X, which is closely associated with the study variable (say) Y. This type of sampling is known as **varying probability sampling** or **probability proportional to size (PPS) sampling**.

Methods of selecting a PPS sample

- **Cumulative Total Method.** Let the size of the i-th unit be denoted by X_i the total size for N population units being $X = \sum_{i=1}^{N} X_i$. Then, the selection procedure consists of following steps:

  1. Write down cumulative totals for the sizes X_i, $i = 1, 2, \ldots, N$.
  2. Choose a random number r, such that $1 \leq r \leq X$.
  3. Select i-th population unit if $T_{i-1} < r \leq T_i$, where $T_k = \sum_{j=1}^{k} X_j$.

  The probability of selecting the i-th population unit, using this procedure, is given by $p_i = X_i/X$.

- **Lahiri's Method.** A procedure which avoids the need for calculating cumulative totals for each unit has been given by Lahiri

  1. Select a random number (say) i from 1 to N.
  2. Select another random number (say) j, such that $1 \leq j \leq M$, where M is either equal to the maximum of the sizes X_i, $i = 1, 2, \ldots, N$, or is more than the maximum size in the population.
  3. If $j \leq X_i$ the i-th unit is selected, otherwise, the pair (i, j) of random numbers is rejected, and another pair is chosen by repeating the steps (1) and (2).

Estimation in PPSWR Sampling

- Unbiased estimator of population total $Y = N\mu$: $\hat{Y}_{pps} = \frac{1}{n} \sum_{i=1}^{n} y_i/p_i$
- Variance of estimator Y_{pps}:

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

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

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

Stratified Sampling

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**.
- N_h = total number of units in the stratum
- n_h = number of units selected in the sample from the stratum
- $W_h = \frac{n_h}{N_h}$ = proportion of the population units falling in the stratum
- $f_h = \frac{n_h}{N_h}$ = sampling fraction for the stratum
- Y_{hi} = the value of study variable for the i-th unit in the stratum $h = 1, \ldots, N_h$
- $Y_h = \sum_{i=1}^{N_h} Y_{hi}$ = stratum total for the estimation variable based on N_h units
- $\bar{Y}_h = \frac{1}{N_h} \sum_{i=1}^{N_h} Y_{hi}$ = mean for the estimation variable in the stratum
- $\bar{y}_h = \frac{1}{n_h} \sum_{i=1}^{n_h} y_{hi}$ = stratum sample mean for the estimation variable
- $S_h^2 = \frac{1}{N_h - 1} \sum_{i=1}^{N_h} (Y_{hi} - \bar{Y}_h)^2$ stratum mean square (stratum variance) based on N_h units
- $s_h^2 = \frac{1}{n_h - 1} \sum_{i=1}^{n_h} (y_{hi} - \bar{y}_h)^2$ sample mean square based on n_h sample units drawn from the stratum

Example 9. (Stratified SRS Without Replacement)
Let U is the union of disjoint strata $U_1, \ldots, U_H$. 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 an simple random sampling of n_h draws **without replacement** from U_h.

Unbiased estimator of population mean μ: $\bar{y}_{st} = \sum_{h=1}^{H} W_h \bar{y}_h$.

Variance of the estimator $\bar{y}_{st}$: $Var(\bar{y}_{st}) = \sum_{h=1}^{H} W_h^2 \left(\frac{1}{n_h} - \frac{1}{N_h}\right) S_h^2$.

Unbiased estimator of the variance $Var(\bar{y}_{st})$: $v(y_{st}) = \sum_{h=1}^{H} W_h^2 \left(\frac{1}{n_h} - \frac{1}{N_h}\right) s_h^2$.

Example 10. (Stratified SRS Without Replacement and Proportional Allocation)
The size of strata is taken into account, and the number of units are drawn in proportion to the size of strata. This means $n_h = n \cdot \frac{N_h}{N}$.

Unbiased estimator of population mean μ: $\bar{y}_{st} = \sum_{h=1}^{H} W_h \bar{y}_h$.

Variance of the estimator $\bar{y}_{st}$: $Var(\bar{y}_{st}) = \left(\frac{1}{n} - \frac{1}{N}\right) \sum_{h=1}^{H} W_h S_h^2$.

Unbiased estimator of variance $Var(\bar{y}_{st})$: $v(\bar{y}_{st}) = \left(\frac{1}{n} - \frac{1}{N}\right) \sum_{h=1}^{H} W_h s_h^2$.

Example 11. (Stratified SRS Without Replacement. Optimum/Neyman Allocation)
Case I. Very often, a survey statistician has to work within a fixed budget C. In such a situation, he is expected to minimize the variance of the estimator subject to the cost constraint.
Let c_h be the cost of observing study variable y for each unit selected in the sample from $h-$th stratum, $h = 1, \ldots, H$. The allocation $\{n_h\}$, which minimizes the variance $Var(\bar{y}_{st}) = \sum_{h=1}^{H} W_h^2 \left(\frac{1}{n_h} - \frac{1}{N_h}\right) S_h^2$ for a given cost $C = \sum_{h=1}^{H} n_h c_h$ is called **optimum allocation.**

Fixed total cost - minimum variance allocation:

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

If the cost per unit is same for all the strata, that is, $c_h = c$ for each h, optimum allocation is known as **Neyman allocation**. For this case, the

optimal allocation takes more simpler form

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

Case II. We fix the precision of the estimator at a specified level and minimize the total cost of survey. The desired level of precision can be specified in two ways. It could be done for example by fixing the value of variance $Var(\bar{y}_{st})$ at V_0.

Fixed variance - minimum cost allocation:

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

Minimum cost - Neyman allocation:

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

Exercise 24.
An assignment was given to four students attending a sample survey course. 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. Number of students registered for the three programs are 1300,450, and 250 respectively. Since the value of the study variable is likely to differ considerably with the program, the investigator divided the population of students into 3 strata: undergraduate program (stratum I), master's program (stratum II), and doctoral program (stratum III). First of the four students selected **WOR** simple random samples of sizes 20, 10, and 12 students from strata I, II, and ill respectively, so that, the total sample is of size 42. The information about weekly time devoted in library is given in the following table

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

Estimate the average time per week devoted to study by a student in PAU library. Also, build up the confidence interval for this average.

Exercise 25.

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 travelled per week by a car manufactured by the company. This information is likely to be helpful in fixing the warranty period for certain parts of the car. The addresses and telephone numbers, if installed, of all the buyers along with their occupations are available at the head office of the company. Since the distance travelled by a car is likely to vary with the profession of the buyer, the investigator divides the population into 3 groups - the businessmen (stratum I), employees (stratum II), and others (stratum III) which includes farmers, etc. Out of 2000 buyers, 825 are businessmen, 700 employees, and 475 others. The average per unit cost for collecting information is expected to be \$4 for businessmen, \$5.5 for employees, and \$6.5 for persons from other category. The total budget at hand is \$1550 which includes the overhead cost of \$1000. The observations on the study variable obtained from these three WOR simple random samples are given in the following table

Stratum I					Stratum II			Stratum III	
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	

The information on strata mean squares, from a similar survey carried out in the past for another car model, is given for strata I, II, and III respectively as $S_1^2 = 30505$, $S_2^2 = 24008$, and $S_3^2 = 29215$.

Find minimum variance allocation.

Multi-stage sampling

In many surveys the sampling is conducted in stages. The elementary units of the population are grouped to begin with into first-stage units or **primary sampling units (PSUs)**. For example, the households of a city might be grouped into city blocks of households. At the first stage of sampling, a sample of **PSUs** is taken; and subsequently elementary units are sampled from within the selected **PSUs** according to some scheme, which may itself be conducted in stages. Sampling in stages generally results in samples which are geographically clustered to some extent. This makes estimation of means and totals less efficient than for dispersed samples of the same size. However, savings in time and travel costs can be appreciable.

Let U be the union of disjoint strata $U_1, \ldots, U_H$. The strata are the primary sampling units (**PSUs**). The sizes $N_1, \ldots, N_H$ of the strata are known, and $\sum_{h=1}^{H} N_h = N$.

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 U_r according to some scheme. Using this notation, the total sample is

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

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

$$\pi_{j|r} \text{ for } r \in U_r.$$

Then the unconditional inclusion probability π_j can be computed as

$$\pi_j = \Pi_r \cdot \pi_{j|r}.$$

The **HT estimator** of μ is

$$\hat{\mu}_{HT} = \frac{1}{N} \sum_{r \in \mathcal{L}} \frac{\hat{T}_r}{\Pi_r},$$

where $\hat{T}_r = \sum_{j \in s_r} \frac{y_j}{\pi_{j|r}}$

PROOF.

$$\hat{\mu}_{HT} = \frac{1}{N} \sum_{j \in s} \frac{y_j}{\pi_j} = \frac{1}{N} \sum_{r \in \mathcal{L}} \sum_{j \in s_r} \frac{y_j}{\Pi_r \cdot \pi_{j|r}} = \frac{1}{N} \sum_{r \in \mathcal{L}} \frac{\hat{T}_r}{\Pi_r},$$

■

- Let a fixed number M of PSUs are selected at the first stage. Then

$$\sum_{r=1}^{H} \Pi_r = M$$

- If a fixed number M of PSUs are selected at the first stage and inclusion probability Π_r is proportional to the size of U_r then

$$\Pi_r = M \cdot \frac{N_r}{N}$$

PROOF. Let $\Pi_r = a \cdot N_r$, where a is the proportionality constant.So

$$M = \sum_{r=1}^{H} \Pi_r = a \sum_{r=1}^{H} N_r = a \cdot N$$

■

- If a fixed number M of PSUs are selected at the first stage and inclusion probability Π_r is proportional to the size of U_r. If additionally for $r \in \mathcal{L}$, the subsample s_r is chosen by simple random sampling without replacement, n_r draws, from U_r, then

$$\pi_{j|r} = \frac{n_r}{N_r} \text{ and } \pi_j = M \cdot \frac{N_r}{N} \cdot \frac{n_r}{N_r} = M \cdot \frac{n_r}{N}.$$

Exercise 26.
Show that in this case **HT estimator** of mean μ is $\hat{\mu}_{HT} = \frac{1}{M} \sum_{r \in \mathcal{L}} \bar{y}_r$ and $E(\hat{\mu}_{HT}) = \mu$.

Hint:

$$E(\hat{\mu}_{HT}) = E(E(\hat{\mu}_{HT}|\mathcal{L})) = E(\frac{1}{M} \sum_{r \in \mathcal{L}} \mu_r) = E(\frac{1}{M} \sum_{r=1}^{H} \mu_r \mathbb{1}(r \in \mathcal{L})).$$

Suppose that the design at the first stage of sampling chooses a fixed number M of PSUs. So $n(\mathcal{L}) = M$ for every sample $\mathcal{L}$. Suppose that if $r \in \mathcal{L}$, sampling takes place within U_r by SRS without replacement with n_r draws. It can be shown, that

$$Var(\hat{\mu}_{HT}) =$$
$$= \frac{1}{N^2} \sum_{r=1}^{H} \frac{N_r^2}{\Pi_r n_r} \left(1 - \frac{n_r}{N_r}\right) \sigma_r^2 + \frac{1}{2N^2} \sum_{j \neq k} (\Pi_j \Pi_k - \Pi_{jk}) \left(\frac{T_j}{\Pi_j} - \frac{T_k}{\Pi_k}\right)^2, \quad (5)$$

where $T_j = \sum_{r \in U_j} Y_r$ and $\sigma_r^2 = \frac{1}{1-N_r} \sum_{j \in U_r} (Y_j - \bar{Y}_r)^2$, $\bar{Y}_r = T_r/N_r$.
If additionally the inclusion probability Π_r is proportional to the size of U_r for every $r \in \mathcal{L}$ then

$$Var(\hat{\mu}_{HT}) = \\ = \frac{1}{M^2} \sum_{r=1}^{H} \frac{\Pi_r}{n_r} \left(1 - \frac{n_r}{N_r}\right) \sigma_r^2 + \frac{1}{2N^2} \sum_{j \neq k} (\Pi_j \Pi_k - \Pi_{jk}) \left(\frac{T_j}{\Pi_j} - \frac{T_k}{\Pi_k}\right)^2, \quad (6)$$

where $\Pi_r = \frac{MN_r}{N}$.

Suppose that at the first and the second stage of drawing we chose **WOR** scheme. So $\Pi_j = \frac{M}{H}$ and $\Pi_{jk} = \frac{M(M-1)}{H(H-1)}$ for every j, k. It can be shown that

$$Var(\hat{\mu}_{HT}) = \frac{H}{MN^2} \left[(H - M)\sigma_{(1)}^2 + \sum_{r=1}^{H} N_r (N_r - n_r) \frac{\sigma_r^2}{n_r} \right], \quad (7)$$

where $\sigma_{(1)}^2 = \frac{1}{2H(H-1)} \sum_{j \neq k} (T_j - T_k)^2 = \frac{1}{H-1} \sum_{j=1}^{H} (T_j - \mu_T)^2$, $\mu_T = \frac{1}{H} \sum_{r=1}^{H} T_r$
If in equation (7) we replace $\sigma_{(1)}^2$ with

$$s_{(1)}^2 = \frac{1}{M-1} \sum_{r=1}^{M} (t_r - \bar{t})^2$$

and σ_r^2 with s_r^2, we will receive an unbiased estimator of $Var(\hat{\mu}_{HT})$.

PROOF OF EQUATION (5).

$$Var(\hat{\mu}_{HT}|\mathcal{L}) = \frac{1}{N^2} \sum_{r \in \mathcal{L}} \frac{1}{\Pi_r^2} Var(\hat{T}_r|\mathcal{L}) = \frac{1}{N^2} \sum_{r \in \mathcal{L}} \frac{1}{\Pi_r^2} N_r^2 \left(\frac{1}{n_r} - \frac{1}{N_r}\right) \sigma_r^2$$

$$E(Var(\hat{\mu}_{HT}|\mathcal{L})) = \frac{1}{N^2} E\left(\sum_{r=1}^{H} \frac{1}{\Pi_r^2} N_r^2 \left(\frac{1}{n_r} - \frac{1}{N_r}\right) \sigma_r^2 \mathbb{1}(r \in \mathcal{L}) \right)$$

$$= \frac{1}{N^2} \sum_{r=1}^{H} \frac{1}{\Pi_r^2} N_r^2 \left(\frac{1}{n_r} - \frac{1}{N_r}\right) \sigma_r^2 \Pi_r$$

$$= \frac{1}{N^2} \sum_{r=1}^{H} \frac{N_r^2}{\Pi_r n_r} \left(1 - \frac{n_r}{N_r}\right) \sigma_r^2$$

$$
\begin{aligned}
Var(E(\hat{\mu}_{HT}|\mathcal{L})) &= Var\left(E\left(\frac{1}{N}\sum_{r\in\mathcal{L}}\frac{\hat{T}_r}{\Pi_r}\middle|\mathcal{L}\right)\right)\\
&= Var\left(\frac{1}{N}\sum_{r\in\mathcal{L}}\frac{1}{\Pi_r}E(\hat{T}_r|\mathcal{L})\right)\\
&= Var\left(\frac{1}{N}\sum_{r\in\mathcal{L}}\frac{1}{\Pi_r}T_r\right)\\
&= \frac{M^2}{N^2}Var\left(\frac{1}{M}\sum_{r\in\mathcal{L}}\frac{1}{\Pi_r}T_r\right)\\
&\text{see equation (4)}\\
&= \frac{M^2}{N^2}\cdot\frac{1}{2M^2}\sum_{j\neq k}(\Pi_j\Pi_k-\Pi_{jk})\left(\frac{T_j}{\Pi_j}-\frac{T_k}{\Pi_k}\right)^2
\end{aligned}
$$

It is known that

$$Var(\hat{\mu}_{HT}) = E(Var(\hat{\mu}_{HT}|\mathcal{L})) + Var(E(\hat{\mu}_{HT}|\mathcal{L})),$$

what ends the proof. ■

PROOF OF EQUATION (6). It is enough to substitute $\Pi_r = \frac{MN_r}{N}$ into the first part of equation (5) ■

PROOF OF EQUATION (7). It is enough to substitute $\Pi_j = \frac{M}{H}$ and $\Pi_{jk} = \frac{M(M-1)}{H(H-1)}$ for every j, k into equation (5) ■

Exercise 27.
The co-operative societies in an Indian state, provide loans to farmers in terms of cash and fertilizer within the sanctioned limit, which depends on the share of the individual in the co-operative society. The society declares an individual defaulter, if he/she does not repay the loan within the specified time limit. An investigator is interested in estimating the average amount of loan, per society, standing against the defaulters. The total number of co-operative societies in the state is 10126. However, the list of all the societies is not available at the state headquarter but the same is available at development block level. Therefore, it seems appropriate to use two-stage sampling for selecting sample of societies. Keeping in view the budget and

time constraints, it was decided to select 12 blocks from the total of 117 blocks and approximately 10 percent of the societies from each of the sample blocks. The information obtained from the selected societies is given in table

Block	N_r	n_r	Amount due from defaulters					
1	60	6	12.5	36.4	26	55.6	58.1	40.8
2	102	10	57.4	16.8	20.3	70.1	34.6	22.6
			44.9	28.4	17.5	33.7		
3	48	5	12.9	41.6	34.7	30.8	61.1	
4	113	11	28.7	82.4	37.3	41.9	24.7	36.6
			39.3	49.6	26	76.8	51.6	
5	92	9	44.8	42.9	51.7	28.8	36.4	40.1
			61.6	47.8	77.4			
6	57	6	31.6	24.8	69.9	44.9	59.7	38.6
7	82	8	49.6	36.9	27.3	63.6	73	44.9
			87.1	61.2				
8	96	10	53.7	34.9	41.5	43.4	56.6	28.9
			23.4	32.8	60.2	47.6		
9	53	5	41.7	54.9	33.9	27.9	46.3	
10	71	7	24.4	38.9	47.8	45	32.6	66.5
			58.3					
11	77	8	42.9	37.3	30.8	51.9	60.1	34.6
			28.4	38.3				
12	56	6	44.7	34.9	61.7	74.6	37.4	49.2

Assuming **WOR+WOR** sampling scheme estimate the average of loan.

References

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.