Probability Proportional to Size (PPS) Sampling

Sanchita Khan (STAT-012)
Suryadeep Ghosh (STAT-008)
Sumedha Guha (STAT-019)

Department of Statistics
Presidency University, Kolkata

Post Graduate (PG), First Year, Semester-II

Contents

Problem with Simple Random Sampling (SRS)

- The simple random sampling scheme provides a random sample where every unit in the population has an equal probability of selection.

Problem with Simple Random Sampling (SRS)

- The simple random sampling scheme provides a random sample where every unit in the population has an equal probability of selection.
- If the sampling units vary considerably in size, then SRS does not take into account the possible importance of the larger units in the population.

Problem with Simple Random Sampling (SRS)

- The simple random sampling scheme provides a random sample where every unit in the population has an equal probability of selection.
- If the sampling units vary considerably in size, then SRS does not take into account the possible importance of the larger units in the population.
- A large unit, i.e., a unit with a large value of Y contributes more to the population total than the units with smaller values.

Concept of Probability Proportional to Size (PPS) Sampling

- So it is natural to expect that a selection scheme which assigns more probability of inclusion in a sample to the larger units than to the smaller units would provide more efficient estimators than the estimators which provide equal probability to all the units.

Concept of Probability Proportional to Size (PPS) Sampling

- So it is natural to expect that a selection scheme which assigns more probability of inclusion in a sample to the larger units than to the smaller units would provide more efficient estimators than the estimators which provide equal probability to all the units.
- This type of sampling is known as varying probability sampling scheme or probability proportional to size (PPS) sampling.

Definition of PPS

- Probability proportional to size (PPS) sampling is a method of sampling from a finite population in which a size measure is available for each population unit before sampling and where the probability of selecting a unit is proportional to its size.

Example of situations where PPS sampling is required

- For example, in an agriculture survey, the yield depends on the area under cultivation.

Example of situations where PPS sampling is required

- For example, in an agriculture survey, the yield depends on the area under cultivation.
- So bigger areas are likely to have a larger population, and they will contribute more towards the population total, so the value of the area can be considered as the size of the auxiliary variable.

Example of situations where PPS sampling is required

- For example, in an agriculture survey, the yield depends on the area under cultivation.
- So bigger areas are likely to have a larger population, and they will contribute more towards the population total, so the value of the area can be considered as the size of the auxiliary variable.
- Also, the cultivated area for a previous period can also be taken as the size while estimating the yield of the crop.

Example of situations where PPS sampling is required

- For example, in an agriculture survey, the yield depends on the area under cultivation.
- So bigger areas are likely to have a larger population, and they will contribute more towards the population total, so the value of the area can be considered as the size of the auxiliary variable.
- Also, the cultivated area for a previous period can also be taken as the size while estimating the yield of the crop.
- Similarly, in an industrial survey, the number of workers in a factory can be considered as the measure of size when studying the industrial output from the respective factory.

PPS sampling

- $P_i = P(U_i \text{is selected at a draw}) \propto X_i$

PPS sampling

- $P_i = P(U_i \text{is selected at a draw}) \propto X_i$
- $\implies P(U_i \text{is selected at a draw}) = kX_i$

PPS sampling

- $P_i = P(U_i \text{is selected at a draw}) \propto X_i$
- $\implies P(U_i \text{is selected at a draw}) = kX_i$
- Now , $\sum_{i=1}^{n} P(U_i \text{is selected at a draw}) = 1$

PPS sampling

- $P_i = P(U_i \text{is selected at a draw}) \propto X_i$
- $\implies P(U_i \text{is selected at a draw}) = kX_i$
- Now , $\sum_{i=1}^{n} P(U_i \text{is selected at a draw}) = 1$
- $\implies k \sum_{i=1}^{n} X_i = 1$

PPS sampling

- $P_i = P(U_i \text{is selected at a draw}) \propto X_i$
- $\implies P(U_i \text{is selected at a draw}) = kX_i$
- Now , $\sum_{i=1}^{n} P(U_i \text{is selected at a draw}) = 1$
- $\implies k\sum_{i=1}^{n} X_i = 1$
- $\implies k = \frac{1}{X},$

PPS sampling

- $P_i = P(U_i \text{is selected at a draw}) \propto X_i$
- $\implies P(U_i \text{is selected at a draw}) = kX_i$
- Now , $\sum_{i=1}^{n} P(U_i \text{is selected at a draw}) = 1$
- $\implies k\sum_{i=1}^{n} X_i = 1$
- $\implies k = \frac{1}{X}$,
- Hence, $P(U_i \text{is selected at a draw}) = \frac{\boldsymbol{X_i}}{\boldsymbol{X}}, i = 1, 2, \ldots .N.$

Schemes of PPS Sampling

- Here we are going to discuss two popular schemes of PPS Sampling.

Schemes of PPS Sampling

- Here we are going to discuss two popular schemes of PPS Sampling.
  1. Cumulative Method.

Schemes of PPS Sampling

- Here we are going to discuss two popular schemes of PPS Sampling.
  1. Cumulative Method.
  2. Lahiri's Method

Cumulative Method

- First we prepare a table showing cumulative values of X_i and allocate serial numbers to the X_i's according to the cumulative values.

Cumulative Method

- First we prepare a table showing cumulative values of X_i and allocate serial numbers to the X_i's according to the cumulative values.
- **Example:**

Cumulative Method

- First we prepare a table showing cumulative values of X_i and allocate serial numbers to the X_i's according to the cumulative values.
- **Example:**
-

Unit no.	X_i	Cumulative	Range
1	X_1	X_1	$1 - X_1$
2	X_2	$X_1 + X_2$	$X_1 + 1 - X_1 + X_2$
3	X_3	$X_1 + X_2 + X_3$	$X_1 + X_2 + 1 - X_1 + X_2 + X_3$
.	.	.	.
.	.	.	.
N	X_N	$X = \sum_{i=1}^{N} X_i$	$\sum_{i=1}^{N-1} X_i + 1 - X$
Total	X		

Cumulative Method

- In general serial no. $X_1 + X_2 + ... + X_{i-1} + 1$ to $X_1 + X_2 + ... + X_i$ are alloted to the *ith* unit U_i.

Cumulative Method

- In general serial no. $X_1 + X_2 + ... + X_{i-1} + 1$ to $X_1 + X_2 + ... + X_i$ are alloted to the *ith* unit U_i.
- Then we a number r at random (i.e SRS) from 1 to X.

Cumulative Method

- In general serial no. $X_1 + X_2 + ... + X_{i-1} + 1$ to $X_1 + X_2 + ... + X_i$ are alloted to the *ith* unit U_i.
- Then we a number r at random (i.e SRS) from 1 to X.
- After that we identify U_i such that, $X_1 + X_2 + ... + X_{i-1} + 1 \leq r \leq X_1 + X_2 + ... + X_i$ and select U_i.

Cumulative Method

- $P_i = P(U_i \text{ is selected at a draw})$

Cumulative Method

- $P_i = P(U_i \text{ is selected at a draw})$
- $= \frac{X_i}{X} \quad \forall i = 1(1)N$

Lahiri's Method

- This method was proposed by Prof. D.B.Lahiri in 1951.

Lahiri's Method

- This method was proposed by Prof. D.B.Lahiri in 1951.
- In case N is large, cumulation of X_i is very tedious.

Lahiri's Method

- This method was proposed by Prof. D.B.Lahiri in 1951.
- In case N is large, cumulation of X_i is very tedious.
- In this method, first we choose an i from 1 to N.

Lahiri's Method

- This method was proposed by Prof. D.B.Lahiri in 1951.
- In case N is large, cumulation of X_i is very tedious.
- In this method, first we choose an i from 1 to N.
- Then select an U_i provisionally according to the chosen i.

Lahiri's Method

- We define $X_0 \geq max\{X_1, X_2, ..., X_N\}$

Lahiri's Method

- We define $X_0 \geq max\{X_1, X_2, ..., X_N\}$
- Choose a number r between 1 and X_0.

Lahiri's Method

- We define $X_0 \geq max\{X_1, X_2, ..., X_N\}$
- Choose a number r between 1 and X_0.
  - If $r \leq X_i$, select U_i finally.

Lahiri's Method

- We define $X_0 \geq max\{X_1, X_2, ..., X_N\}$
- Choose a number r between 1 and X_0.
  - If $r \leq X_i$, select U_i finally.
  - if $r > X_i$,do not select U_i.

Lahiri's Method

- We define $X_0 \geq max\{X_1, X_2, ..., X_N\}$
- Choose a number r between 1 and X_0.
  - If $r \leq X_i$, select U_i finally.
  - if $r > X_i$,do not select U_i.
- We repeat this entire procedure until we select the required number of samples for the analysis.

Lahiri's Method

- For a trial U_i is selected or U_i is not selected .

Lahiri's Method

- For a trial $U_i$is selected or $U_i$is not selected .
- $p_i = P(U_i \text{is selected at a trial})$

Lahiri's Method

- For a trial $U_i$is selected or $U_i$is not selected .
- $p_i = P(U_i\text{is selected at a trial})$
- $= P(i$ is selected from 1 to N and [r, selected from 1 and X_0] $r \leq X_i)$

Lahiri's Method

- For a trial $U_i$is selected or $U_i$is not selected .
- $p_i = P(U_i \text{is selected at a trial})$
- $= P(i$ is selected from 1 to N and [r, selected from 1 and X_0] $r \leq X_i)$
- $= P(\text{selected no. is } i).P(r \leq X_i|\text{selected no. is } i)$

Lahiri's Method

- For a trial $U_i$is selected or $U_i$is not selected .
- $p_i = P(U_i\text{is selected at a trial})$
- $= P(i$ is selected from 1 to N and [r, selected from 1 and X_0] $r \leq X_i)$
- $= P(\text{selected no. is } i).P(r \leq X_i|\text{selected no. is } i)$
- $=\frac{1}{N}\frac{X_i}{X_0}$

Lahiri's Method

- For a trial $U_i$is selected or $U_i$is not selected .
- $p_i = P(U_i\text{is selected at a trial})$
- $= P(i$ is selected from 1 to N and [r, selected from 1 and X_0] $r \leq X_i)$
- $= P(\text{selected no. is } i).P(r \leq X_i|\text{selected no. is } i)$
- $=\frac{1}{N}\frac{X_i}{X_0}$
- also, q=P(no selection is made at a trial)

Lahiri's Method

- For a trial $U_i$is selected or $U_i$is not selected .
- $p_i = P(U_i \text{is selected at a trial})$
- $= P(i$ is selected from 1 to N and [r, selected from 1 and X_0] $r \leq X_i)$
- $= P(\text{selected no. is } i).P(r \leq X_i|\text{selected no. is } i)$
- $=\frac{1}{N}\frac{X_i}{X_0}$
- also, q=P(no selection is made at a trial)
- $= \sum_{i=1}^{N} P(selected\, no. is\, i\, and\, r > X_i)$

Lahiri's Method

- For a trial $U_i$is selected or $U_i$is not selected .
- $p_i = P(U_i \text{is selected at a trial})$
- $= P(i$ is selected from 1 to N and [r, selected from 1 and X_0] $r \leq X_i)$
- $= P(\text{selected no. is } i).P(r \leq X_i | \text{selected no. is } i)$
- $=\frac{1}{N}\frac{X_i}{X_0}$
- also, q=P(no selection is made at a trial)
- $= \sum_{i=1}^{N} P(selected\ no. is\ i\ and\ r > X_i)$
- $= \sum \frac{1}{N}(1 - \frac{X_i}{X_0})$

Lahiri's Method

- For a trial $U_i$is selected or $U_i$is not selected .
- $p_i = P(U_i\text{is selected at a trial})$
- $= P(i$ is selected from 1 to N and [r, selected from 1 and X_0] $r \leq X_i)$
- $= P(\text{selected no. is } i).P(r \leq X_i|\text{selected no. is } i)$
- $=\frac{1}{N}\frac{X_i}{X_0}$
- also, $q{=}P(\text{no selection is made at a trial})$
- $= \sum_{i=1}^{N} P(selected\ no.\ is\ i\ and\ r > X_i)$
- $= \sum \frac{1}{N}(1 - \frac{X_i}{X_0})$
- $= 1 - \frac{\sum_{i=1}^{N} X_i}{NX_0}$

Lahiri's Method

- $= 1 - \frac{\bar{X}}{X_0}$

Lahiri's Method

- $= 1 - \frac{\bar{X}}{X_0}$
- Therefore, $P_i = P(U_i \text{is selected finally})$

Lahiri's Method

- $= 1 - \frac{\bar{X}}{X_0}$
- Therefore, $P_i = P(U_i \text{is selected finally})$
- $=p_i + qp_i + q^2 p_i + ...$

Lahiri's Method

- $= 1 - \frac{\bar{X}}{X_0}$
- Therefore, $P_i = P(U_i \text{is selected finally})$
- $= p_i + qp_i + q^2 p_i + ...$
- $= \frac{p_i}{1-q} = \frac{1}{N} \frac{X_i}{X_0} / (1 - [1 - \frac{\bar{X}}{X_0})$

Lahiri's Method

- $= 1 - \frac{\bar{X}}{X_0}$
- Therefore, $P_i = P(U_i \text{is selected finally})$
- $= p_i + qp_i + q^2 p_i + \ldots.$
- $= \frac{p_i}{1-q} = \frac{1}{N}\frac{X_i}{X_0} / (1 - [1 - \frac{\bar{X}}{X_0})$
- $= \frac{1}{N}\frac{X_i}{X_0}\frac{X_0}{\bar{X}} = \frac{X_i}{N\bar{X}} = \frac{\boldsymbol{X_i}}{\boldsymbol{X}}$, $X = \sum_{i=1}^{N} X_i$

Horvitz-Thompson (HT) Estimator

- An unbiased estimate of the population total is given by

Horvitz-Thompson (HT) Estimator

- An unbiased estimate of the population total is given by
- $L(s) = \sum_{i=1}^{n} \frac{y_j}{\pi_j} = \hat{Y}_{HT}$

Horvitz-Thompson (HT) Estimator

- An unbiased estimate of the population total is given by
- $L(s) = \sum_{i=1}^{n} \frac{y_j}{\pi_j} = \hat{Y}_{HT}$
- where π_j is the 1^{st} order inclusion probability that the j^{th} unit is included in the sample

Horvitz-Thompson (HT) Estimator

- An unbiased estimate of the population total is given by
- $L(s) = \sum_{i=1}^{n} \frac{y_j}{\pi_j} = \hat{Y}_{HT}$
- where π_j is the 1^{st} order inclusion probability that the j^{th} unit is included in the sample
- It is called Horvitz-Thompson Estimator of Population Total.

Variance of the HT estimator

- $V(\hat{Y}_{HT}) = \sum_{i=1}^{N} \frac{\pi_i(1-\pi_i)}{\pi_i^2} Y_i^2 + \sum_{i=1}^{N} \sum_{j=1}^{N} \frac{Y_i Y_j}{\pi_i \pi_j} \pi_{ij} - \pi_i \pi_j$

Variance of the HT estimator

- $V(\hat{Y}_{HT}) = \sum_{i=1}^{N} \frac{\pi_i(1-\pi_i)}{\pi_i^2} Y_i^2 + \sum_{i=1}^{N} \sum_{j=1}^{N} \frac{Y_i Y_j}{\pi_i \pi_j} \pi_{ij} - \pi_i \pi_j$
- $= \sum_{i=1}^{N} \frac{1-\pi_i}{\pi_i} Y_i^2 + \sum_{i=1}^{N} \sum_{j=1}^{N} \frac{\pi_{ij}-\pi_i\pi_j}{\pi_i\pi_j} Y_i Y_j$

Variance of the HT estimator

- $V(\hat{Y}_{HT}) = \sum_{i=1}^{N} \frac{\pi_i(1-\pi_i)}{\pi_i^2} Y_i^2 + \sum_{i=1}^{N} \sum_{j=1}^{N} \frac{Y_i Y_j}{\pi_i \pi_j} \pi_{ij} - \pi_i \pi_j$
- $= \sum_{i=1}^{N} \frac{1-\pi_i}{\pi_i} Y_i^2 + \sum_{i=1}^{N} \sum_{j=1}^{N} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j} Y_i Y_j$
- where π_{ij} is the 2^{nd} order inclusion probability that both i^{th} and j^{th} unit will be included in the sample

Variance of the HT estimator

- An alternative expression for $V(\hat{Y}_{HT})$:

Variance of the HT estimator

- An alternative expression for $V(\hat{Y}_{HT})$:
- $= \sum_{\substack{i=1 \\ i<j}}^{N} \sum_{j=1}^{N} (\pi_i \pi_j - \pi_{ij}) (\frac{Y_i}{\pi_i} - \frac{Y_i}{\pi_i})^2$

Unbiased estimator of the variance

- and an unbiased estimator of $V(\hat{Y}_{HT})$is given by,

Unbiased estimator of the variance

- and an unbiased estimator of $V(\hat{Y}_{HT})$is given by,
- $v(\hat{Y}_{HT})= \sum_{\substack{i=1 \\ i<j}}^{N} \sum_{j=1}^{N} \frac{(\pi_j \pi'_j - \pi_{jj'})}{\pi'_{jj}} (\frac{y_j}{\pi_j} - \frac{y'_j}{\pi'_j})^2$

PPSWOR

- Thus, in a PPSWOR sampling scheme

PPSWOR

- Thus, in a PPSWOR sampling scheme
- $\hat{Y}_{PPSWOR} = \sum_{j=1}^{n} \frac{y_J}{\pi_j}$

PPSWOR

- Thus, in a PPSWOR sampling scheme
- $\hat{Y}_{PPSWOR} = \sum\limits_{j=1}^{n} \frac{y_J}{\pi_j}$
- $V(\hat{Y}_{PPSWOR}) = \sum_{i=1}^{N} \sum_{j=1}^{N}_{i<j} (\pi_i \pi_j - \pi_{ij})(\frac{Y_i}{\pi_i} - \frac{Y_i}{\pi_i})^2$

PPSWOR

- Thus, in a PPSWOR sampling scheme
- $\hat{Y}_{PPSWOR} = \sum_{j=1}^{n} \frac{y_J}{\pi_j}$
- $V(\hat{Y}_{PPSWOR}) = \sum_{\substack{i=1 \\ i<j}}^{N} \sum_{j=1}^{N} (\pi_i \pi_j - \pi_{ij})(\frac{Y_i}{\pi_i} - \frac{Y_i}{\pi_i})^2$
- and $v(\hat{Y}_{PPSWOR}) = \sum_{\substack{i=1 \\ i<j}}^{N} \sum_{j=1}^{N} \frac{(\pi_j \pi'_j - \pi_{jj'})}{\pi'_{jj}} (\frac{y_j}{\pi_j} - \frac{y'_j}{\pi'_j})^2$

PPSWOR

- Thus, in a PPSWOR sampling scheme
- $\hat{Y}_{PPSWOR} = \sum_{j=1}^{n} \frac{y_J}{\pi_j}$
- $V(\hat{Y}_{PPSWOR}) = \sum_{\substack{i=1 \\ i<j}}^{N} \sum_{j=1}^{N} (\pi_i \pi_j - \pi_{ij})(\frac{Y_i}{\pi_i} - \frac{Y_i}{\pi_i})^2$
- and $v(\hat{Y}_{PPSWOR}) = \sum_{\substack{i=1 \\ i<j}}^{N} \sum_{j=1}^{N} \frac{(\pi_j \pi'_j - \pi_{jj'})}{\pi'_{jj}} (\frac{y_j}{\pi_j} - \frac{y'_j}{\pi'_j})^2$
- which is an unbiased estimator of the sampling variance

PPSWOR

- Thus, in a PPSWOR sampling scheme
- $\hat{Y}_{PPSWOR} = \sum_{j=1}^{n} \frac{y_J}{\pi_j}$
- $V(\hat{Y}_{PPSWOR}) = \sum_{\substack{i=1 \\ i<j}}^{N} \sum_{j=1}^{N} (\pi_i \pi_j - \pi_{ij})(\frac{Y_i}{\pi_i} - \frac{Y_i}{\pi_i})^2$
- and $v(\hat{Y}_{PPSWOR}) = \sum_{\substack{i=1 \\ i<j}}^{N} \sum_{j=1}^{N} \frac{(\pi_j \pi'_j - \pi_{jj'})}{\pi'_{jj}} (\frac{y_j}{\pi_j} - \frac{y'_j}{\pi'_j})^2$
- which is an unbiased estimator of the sampling variance
- Thus, an estimate of the standard error $s.e.(\hat{Y}_{PPSWOR})$ is $\sqrt{v(\hat{Y}_{PPSWOR})}$

Some Necessary Theories

- In PPSWOR sampling it may be the case that some of $\pi_i\pi_j - \pi_{ij}$may be negative ,0 and positive . But it may arise that the pairs for which $\pi_i\pi_j - \pi_{ij} < 0$ make a numerically higher contribution than the corresponding positive part and the $V_{PPSWOR}(\hat{Y}_{HT})$ may in negative. The estimator $\hat{V}(\hat{Y}_{HT})$ may also be negetive similarly.

Some Necessary Theories

- In PPSWOR sampling it may be the case that some of $\pi_i\pi_j - \pi_{ij}$may be negative ,0 and positive . But it may arise that the pairs for which $\pi_i\pi_j - \pi_{ij} < 0$ make a numerically higher contribution than the corresponding positive part and the $V_{PPSWOR}(\hat{Y}_{HT})$ may in negative. The estimator $\hat{V}(\hat{Y}_{HT})$ may also be negetive similarly.
- In case of SRSWOR,

Some Necessary Theories

- In PPSWOR sampling it may be the case that some of $\pi_i\pi_j - \pi_{ij}$may be negative ,0 and positive . But it may arise that the pairs for which $\pi_i\pi_j - \pi_{ij} < 0$ make a numerically higher contribution than the corresponding positive part and the $V_{PPSWOR}(\hat{Y}_{HT})$ may in negative. The estimator $\hat{V}(\hat{Y}_{HT})$ may also be negetive similarly.
- In case of SRSWOR,
- $\pi_i = \frac{n}{N}$ and $\pi_{ij} = \frac{n(n-1)}{N(N-1)}$, $i \neq j$.

Some Necessary Theories

- In PPSWOR sampling it may be the case that some of $\pi_i\pi_j - \pi_{ij}$may be negative ,0 and positive . But it may arise that the pairs for which $\pi_i\pi_j - \pi_{ij} < 0$ make a numerically higher contribution than the corresponding positive part and the $V_{PPSWOR}(\hat{Y}_{HT})$ may in negative. The estimator $\hat{V}(\hat{Y}_{HT})$ may also be negetive similarly.
- In case of SRSWOR,
- $\pi_i = \frac{n}{N}$ and $\pi_{ij} = \frac{n(n-1)}{N(N-1)}$, $i \neq j$.
- Then HT estimator $V(\hat{Y}_{HT}) = \frac{\sigma^2(N-n)}{(N-1)}$

Some Necessary Theories

- In PPSWOR sampling it may be the case that some of $\pi_i\pi_j - \pi_{ij}$may be negative ,0 and positive . But it may arise that the pairs for which $\pi_i\pi_j - \pi_{ij} < 0$ make a numerically higher contribution than the corresponding positive part and the $V_{PPSWOR}(\hat{Y}_{HT})$ may in negative. The estimator $\hat{V}(\hat{Y}_{HT})$ may also be negetive similarly.
- In case of SRSWOR,
- $\pi_i = \frac{n}{N}$ and $\pi_{ij} = \frac{n(n-1)}{N(N-1)}$, $i \neq j$.
- Then HT estimator $V(\hat{Y}_{HT}) = \frac{\sigma^2(N-n)}{(N-1)}$
- $\implies V_{SRSWOR}(\hat{Y}_{SRSWOR})$

Some Necessary Theories

- In PPSWOR sampling it may be the case that some of $\pi_i\pi_j - \pi_{ij}$may be negative ,0 and positive . But it may arise that the pairs for which $\pi_i\pi_j - \pi_{ij} < 0$ make a numerically higher contribution than the corresponding positive part and the $V_{PPSWOR}(\hat{Y}_{HT})$ may in negative. The estimator $\hat{V}(\hat{Y}_{HT})$ may also be negetive similarly.
- In case of SRSWOR,
- $\pi_i = \frac{n}{N}$ and $\pi_{ij} = \frac{n(n-1)}{N(N-1)}$, $i \neq j$.
- Then HT estimator $V(\hat{Y}_{HT}) = \frac{\sigma^2(N-n)}{(N-1)}$
- $\implies V_{SRSWOR}(\hat{Y}_{SRSWOR})$
- $V(\hat{Y}_{SRSWOR})$ can never be negative

Some Necessary Theories

- In PPSWOR sampling it may be the case that some of $\pi_i\pi_j - \pi_{ij}$may be negative ,0 and positive . But it may arise that the pairs for which $\pi_i\pi_j - \pi_{ij} < 0$ make a numerically higher contribution than the corresponding positive part and the $V_{PPSWOR}(\hat{Y}_{HT})$ may in negative. The estimator $\hat{V}(\hat{Y}_{HT})$ may also be negetive similarly.
- In case of SRSWOR,
- $\pi_i = \frac{n}{N}$ and $\pi_{ij} = \frac{n(n-1)}{N(N-1)}$, $i \neq j$.
- Then HT estimator $V(\hat{Y}_{HT}) = \frac{\sigma^2(N-n)}{(N-1)}$
- $\implies V_{SRSWOR}(\hat{Y}_{SRSWOR})$
- $V(\hat{Y}_{SRSWOR})$ can never be negative
- $\pi_i\pi_j - \pi_{ij} > 0$ for SRSWOR $i \neq j$

PPSWR

- In PPSWR

PPSWR

- In PPSWR
- $\pi_i = P(U_i \text{ is included in the sample})$

PPSWR

- In PPSWR
- $\pi_i = P(U_i$ is included in the sample)
- =P(U_i is selected in one of the n draws)

PPSWR

- In PPSWR
- $\pi_i = P(U_i$ is included in the sample)
- =P(U_i is selected in one of the n draws)
- $=\sum_{j=1}^{n} P(U_i$ is selected at the $j^{th} draw)$

PPSWR

- In PPSWR
- $\pi_i = P(U_i$ is included in the sample)
- =P(U_i is selected in one of the n draws)
- =$\sum_{j=1}^{n} P(U_i$ is selected at the $j^{th} draw)$
- =$\sum_{j=1}^{n} P_i$=nP_i $\qquad \sum_{i=1}^{N} P_i = 1$

PPSWR

- we write

PPSWR

- we write
- $\pi_i = nP_i$, $\sum_{i=1}^{N} P_i = 1$

PPSWR

- we write
- $\pi_i = nP_i$, $\sum_{i=1}^{N} P_i = 1$
- Then the corresponding PPSWR sampling scheme will arise

PPSWR

- we write
- $\pi_i = nP_i$, $\sum_{i=1}^{N} P_i = 1$
- Then the corresponding PPSWR sampling scheme will arise
- An unbiased estimator of the population total is

PPSWR

- we write
- $\pi_i = nP_i \ , \sum_{i=1}^{N} P_i = 1$
- Then the corresponding PPSWR sampling scheme will arise
- An unbiased estimator of the population total is
- $\hat{Y}_{PPSWR} = \frac{1}{n} \sum_{j=1}^{n} \frac{y_j}{p_j}$ and,

PPSWR

- we write
- $\pi_i = nP_i \ , \sum_{i=1}^{N} P_i = 1$
- Then the corresponding PPSWR sampling scheme will arise
- An unbiased estimator of the population total is
- $\hat{Y}_{PPSWR} = \frac{1}{n} \sum_{j=1}^{n} \frac{y_j}{p_j}$ and,
- $V_{PPSWR}(\hat{Y}_{PPSWR}) = \sum_{\substack{i=1 \\ i<j}}^{N} \sum_{j=1}^{N} P_i P_j (\frac{Y_i}{\pi_i} - \frac{Y_i}{\pi_i})^2$

PPSWR

- we write
- $\pi_i = nP_i$, $\sum_{i=1}^{N} P_i = 1$
- Then the corresponding PPSWR sampling scheme will arise
- An unbiased estimator of the population total is
- $\hat{Y}_{PPSWR} = \frac{1}{n}\sum_{j=1}^{n} \frac{y_j}{p_j}$ and,
- $V_{PPSWR}(\hat{Y}_{PPSWR}) = \sum_{i=1}^{N} \sum_{\substack{j=1 \\ i<j}}^{N} P_i P_j (\frac{Y_i}{\pi_i} - \frac{Y_i}{\pi_i})^2$
- $= \frac{1}{n} \sum\limits_{i=1}^{N} \frac{Y_i^2}{P_i^2} - Y^2$

Variance estimator

- An unbiased estimator of the variance $V(\hat{Y}_{PPSWR})$ is,

Variance estimator

- An unbiased estimator of the variance $V(\hat{Y}_{PPSWR})$ is,
- $v(\hat{Y}_{PPSWR}) = \frac{1}{n(n-1)} \left[\sum_{i=1}^{n} (\frac{y_j}{p_j})^2 - n\hat{Y}^2_{PPSWR} \right]$

Variance estimator

- An unbiased estimator of the variance $V(\hat{Y}_{PPSWR})$ is,
- $v(\hat{Y}_{PPSWR}) = \frac{1}{n(n-1)}\left[\sum_{i=1}^{n}(\frac{y_j}{p_j})^2 - n\hat{Y}^2_{PPSWR}\right]$
- An estimate of the standard error $s.e.(\hat{Y}_{PPSWR})$ is $\sqrt{v(\hat{Y}_{PPSWR})}$

Comparision between PPSWR and SRSWR

- $\hat{Y}_{SRSWR} = N\bar{y}$ where $\bar{y} = \frac{1}{n}\sum_{j=1}^{n} y_j$

Comparision between PPSWR and SRSWR

- $\hat{Y}_{SRSWR} = N\bar{y}$ where $\bar{y} = \frac{1}{n}\sum_{j=1}^{n} y_j$
- $E(\hat{Y}_{SRSWR}) = Y$ and $V(\hat{Y}_{SRSWR}) = \frac{1}{n}\left(N\sum_{i=1}^{N} Y_i^2 - Y^2\right)$

Comparision between PPSWR and SRSWR

- $\hat{Y}_{SRSWR} = N\bar{y}$ where $\bar{y} = \frac{1}{n}\sum_{j=1}^{n} y_j$
- $E(\hat{Y}_{SRSWR}) = Y$ and $V(\hat{Y}_{SRSWR}) = \frac{1}{n}\left(N\sum_{i=1}^{N} Y_i^2 - Y^2\right)$
- Now

Comparision between PPSWR and SRSWR

- $\hat{Y}_{SRSWR} = N\bar{y}$ where $\bar{y} = \frac{1}{n}\sum_{j=1}^{n} y_j$
- $E(\hat{Y}_{SRSWR}) = Y$ and $V(\hat{Y}_{SRSWR}) = \frac{1}{n}\left(N\sum_{i=1}^{N} Y_i^2 - Y^2\right)$
- Now
- $$\begin{aligned} & V(\hat{Y}_{PPSWR}) < V(\hat{Y}_{SRSWR}) \\ \leftrightarrow \; & \frac{1}{n}\left(\sum_{i=1}^{N}\frac{Y_i^2}{P_i^2} - Y^2\right) < \frac{1}{n}\left(N\sum_{i=1}^{N} Y_i^2 - Y^2\right) \\ \leftrightarrow & \sum_{i=1}^{N}\frac{Y_i^2}{P_i^2} < N\sum_{i=1}^{N} Y_i^2 \\ \leftrightarrow \; & N\sum_{i=1}^{N} Y_i^2 - \sum_{i=1}^{N}\frac{Y_i^2}{P_i^2} > 0 \end{aligned}$$

Estimating Gain: PPSWR Vs. SRSWR

- $\leftrightarrow N \sum_{i=1}^{N} Y_i^2 - \sum_{i=1}^{N} \frac{Y_i^2}{X_i} \bar{X}.N > 0$

Estimating Gain: PPSWR Vs. SRSWR

- $\leftrightarrow N \sum_{i=1}^{N} Y_i^2 - \sum_{i=1}^{N} \frac{Y_i^2}{X_i} \bar{X}.N > 0$
- $\leftrightarrow \sum_{i=1}^{N} \frac{Y_i^2}{X_i} (X_i - \bar{X}) > 0$

Estimating Gain: PPSWR Vs. SRSWR

- $\leftrightarrow N \sum_{i=1}^{N} Y_i^2 - \sum_{i=1}^{N} \frac{Y_i^2}{X_i} \bar{X}.N > 0$
- $\leftrightarrow \sum_{i=1}^{N} \frac{Y_i^2}{X_i} (X_i - \bar{X}) > 0$
- i.e if $\frac{Y^2}{X}$ and X are positively correlated in the population.

Estimating Gain: PPSWR Vs. SRSWR

- $\leftrightarrow N \sum_{i=1}^{N} Y_i^2 - \sum_{i=1}^{N} \frac{Y_i^2}{X_i} \bar{X}.N > 0$
- $\leftrightarrow \sum_{i=1}^{N} \frac{Y_i^2}{X_i} (X_i - \bar{X}) > 0$
- i.e if $\frac{Y^2}{X}$ and X are positively correlated in the population.
- Let the gain be denoted by G.

Estimating Gain: PPSWR Vs. SRSWR

- $\leftrightarrow N\sum_{i=1}^{N} Y_i^2 - \sum_{i=1}^{N} \frac{Y_i^2}{X_i}\bar{X}.N > 0$
- $\leftrightarrow \sum_{i=1}^{N} \frac{Y_i^2}{X_i}(X_i - \bar{X}) > 0$
- i.e if $\frac{Y^2}{X}$ and X are positively correlated in the population.
- Let the gain be denoted by G.
- Then,

Estimating Gain: PPSWR Vs. SRSWR

- $G = V(\hat{Y}_{SRSWR}) - V(\hat{Y}_{PPSWR})$

Estimating Gain: PPSWR Vs. SRSWR

- $G = V(\hat{Y}_{SRSWR}) - V(\hat{Y}_{PPSWR})$
- $= \frac{1}{n}\left(N\sum_{i=1}^{N} Y_i^2 - Y^2\right) - \frac{1}{n}\left(\sum_{i=1}^{N} \frac{Y_i^2}{P_i^2} - Y^2\right)$

Estimating Gain: PPSWR Vs. SRSWR

- $G = V(\hat{Y}_{SRSWR}) - V(\hat{Y}_{PPSWR})$
- $= \frac{1}{n}\left(N\sum_{i=1}^{N} Y_i^2 - Y^2\right) - \frac{1}{n}\left(\sum_{i=1}^{N} \frac{Y_i^2}{P_i^2} - Y^2\right)$
- $= \frac{1}{n}\left(N\sum_{i=1}^{N} Y_i^2 - \sum_{i=1}^{N} \frac{Y_i^2}{P_i}\right)$

Estimating Gain: PPSWR Vs. SRSWR

- $G = V(\hat{Y}_{SRSWR}) - V(\hat{Y}_{PPSWR})$
- $= \frac{1}{n}\left(N\sum_{i=1}^{N} Y_i^2 - Y^2\right) - \frac{1}{n}\left(\sum_{i=1}^{N} \frac{Y_i^2}{P_i^2} - Y^2\right)$
- $= \frac{1}{n}\left(N\sum_{i=1}^{N} Y_i^2 - \sum_{i=1}^{N} \frac{Y_i^2}{P_i}\right)$
- An unbiased estimator of the gain is,

Estimating Gain: PPSWR Vs. SRSWR

- $G = V(\hat{Y}_{SRSWR}) - V(\hat{Y}_{PPSWR})$
- $= \frac{1}{n}\left(N\sum_{i=1}^{N} Y_i^2 - Y^2\right) - \frac{1}{n}\left(\sum_{i=1}^{N}\frac{Y_i^2}{P_i^2} - Y^2\right)$
- $= \frac{1}{n}\left(N\sum_{i=1}^{N} Y_i^2 - \sum_{i=1}^{N}\frac{Y_i^2}{P_i}\right)$
- An unbiased estimator of the gain is,
- $\hat{G} = \frac{1}{n}\left(N\sum_{j=1}^{n}\frac{y_j^2}{p_j} - \sum_{j=1}^{n}\frac{y_j^2}{p_j^2}\right) = \frac{1}{n^2}\sum_{j=1}^{n}\frac{y_j^2}{p_j}(N - \frac{1}{p_j})$

Estimating Gain in efficiency: PPSWOR Vs. SRSWOR

- Probability proportional to size measures x_j (PPS) without replacement (PPSWOR) sample selection method is implemented by selecting a number, say, $n(\geq 2)$ units from U ordered as the $1^{st}, 2^{nd}, ..., n^{th}$, namely $u_1, u_2, ..., u_n$ with respective probabilities.

Estimating Gain in efficiency: PPSWOR Vs. SRSWOR

- Probability proportional to size measures x_j (PPS) without replacement (PPSWOR) sample selection method is implemented by selecting a number, say, $n(\geq 2)$ units from U ordered as the $1^{st}, 2^{nd}, ..., n^{th}$, namely $u_1, u_2, ..., u_n$ with respective probabilities.
- $$p_1, \frac{p_2}{1-p_1}, ..., \frac{p_j}{1-p_1-...-p_{j-1}} \text{ for, } j = 1, 2, ..., n$$

Estimating Gain in efficiency: PPSWOR Vs. SRSWOR

- Then, Des Raj's unbiased estimator for Y is,

Estimating Gain in efficiency: PPSWOR Vs. SRSWOR

- Then, Des Raj's unbiased estimator for Y is,
- $$t_D = \frac{1}{n}(t_1 + t_2 + ... + t_n)$$

Estimating Gain in efficiency: PPSWOR Vs. SRSWOR

- Then, Des Raj's unbiased estimator for Y is,
- $$t_D = \frac{1}{n}(t_1 + t_2 + ... + t_n)$$
- with,

Estimating Gain in efficiency: PPSWOR Vs. SRSWOR

- Then, Des Raj's unbiased estimator for Y is,
- $$t_D = \frac{1}{n}(t_1 + t_2 + ... + t_n)$$
- with,
  - $t_1 = \frac{y_1}{p_1}$

Estimating Gain in efficiency: PPSWOR Vs. SRSWOR

- Then, Des Raj's unbiased estimator for Y is,
- $$t_D = \frac{1}{n}(t_1 + t_2 + ... + t_n)$$
- with,
  - $t_1 = \frac{y_1}{p_1}$
  - $t_2 = y_1 + \frac{y_2}{p_2}(1 - p_1), ...,$

Estimating Gain in efficiency: PPSWOR Vs. SRSWOR

- Then, Des Raj's unbiased estimator for Y is,
- $$t_D = \frac{1}{n}(t_1 + t_2 + ... + t_n)$$
- with,
  - $t_1 = \frac{y_1}{p_1}$
  - $t_2 = y_1 + \frac{y_2}{p_2}(1 - p_1), ...,$
  - $t_j = y_1 + y_2 + ... + \frac{y_j}{p_j}(1 - p_1 - p_2 - ... - p_{j-1})$, $j = 1, 2, ..., n$

Estimating Gain in efficiency: PPSWOR Vs. SRSWOR

- Then, Des Raj's unbiased estimator for Y is,
- $$t_D = \frac{1}{n}(t_1 + t_2 + ... + t_n)$$
- with,
  - $t_1 = \frac{y_1}{p_1}$
  - $t_2 = y_1 + \frac{y_2}{p_2}(1 - p_1), ...,$
  - $t_j = y_1 + y_2 + ... + \frac{y_j}{p_j}(1 - p_1 - p_2 - ... - p_{j-1})$, $j = 1, 2, ..., n$
- An unbiased estimator for $V(t_D)$ is given by Des Raj (1956) as,

Estimating Gain in efficiency: PPSWOR Vs. SRSWOR

- Then, Des Raj's unbiased estimator for Y is,
- $$t_D = \frac{1}{n}(t_1 + t_2 + \ldots + t_n)$$
- with,
  - $t_1 = \frac{y_1}{p_1}$
  - $t_2 = y_1 + \frac{y_2}{p_2}(1 - p_1), \ldots,$
  - $t_j = y_1 + y_2 + \ldots + \frac{y_j}{p_j}(1 - p_1 - p_2 - \ldots - p_{j-1})$, $j = 1, 2, \ldots, n$
- An unbiased estimator for $V(t_D)$ is given by Des Raj (1956) as,
- $$v(t_D) = \frac{1}{2n^2(n-1)} \sum_{j=1,k=1}^{n} \sum_{k \neq j}^{n} (t_j - t_k)^2$$

Estimating Gain in efficiency: PPSWOR Vs. SRSWOR

- Suppose a PPSWOR sample chosen as above is at hand as $s = (u_1, ..., u_n)$ along with the values $y_1, y_2, ..., y_n$.

Estimating Gain in efficiency: PPSWOR Vs. SRSWOR

- Suppose a PPSWOR sample chosen as above is at hand as $s = (u_1, ..., u_n)$ along with the values $y_1, y_2, ..., y_n$.
- Suppose we consider a comparable strategy composed of an SRSWOR sample s_{WOR} of size n and the estimator based on it as,

Estimating Gain in efficiency: PPSWOR Vs. SRSWOR

- Suppose a PPSWOR sample chosen as above is at hand as $s = (u_1, ..., u_n)$ along with the values $y_1, y_2, ..., y_n$.
- Suppose we consider a comparable strategy composed of an SRSWOR sample s_{WOR} of size n and the estimator based on it as,
- $$N\bar{y} = \frac{N}{n} \sum_{i \in s_{WOR}} y_i$$

Estimating Gain in efficiency: PPSWOR Vs. SRSWOR

- Suppose a PPSWOR sample chosen as above is at hand as $s = (u_1, ..., u_n)$ along with the values $y_1, y_2, ..., y_n$.
- Suppose we consider a comparable strategy composed of an SRSWOR sample s_{WOR} of size n and the estimator based on it as,
- $$N\bar{y} = \frac{N}{n} \sum_{i \in s_{WOR}} y_i$$
- with variance

Estimating Gain in efficiency: PPSWOR Vs. SRSWOR

- Suppose a PPSWOR sample chosen as above is at hand as $s = (u_1, ..., u_n)$ along with the values $y_1, y_2, ..., y_n$.
- Suppose we consider a comparable strategy composed of an SRSWOR sample s_{WOR} of size n and the estimator based on it as,
- $$N\bar{y} = \frac{N}{n} \sum_{i \in s_{WOR}} y_i$$
- with variance
- $$V_{s_{WOR}}(N\bar{y}) = \frac{(N-n)N^2}{Nn(N-1)} \sum_{i=1}^{N} (y_i - \bar{Y})^2$$

Estimating Gain in efficiency: PPSWOR Vs. SRSWOR

- Suppose a PPSWOR sample chosen as above is at hand as $s = (u_1, ..., u_n)$ along with the values $y_1, y_2, ..., y_n$.
- Suppose we consider a comparable strategy composed of an SRSWOR sample s_{WOR} of size n and the estimator based on it as,
- $$N\bar{y} = \frac{N}{n} \sum_{i \in s_{WOR}} y_i$$
- with variance
- $$V_{s_{WOR}}(N\bar{y}) = \frac{(N-n)N^2}{Nn(N-1)} \sum_{i=1}^{N} (y_i - \bar{Y})^2$$
- where $\bar{y}$ denotes the sample mean.

Estimating Gain in efficiency: PPSWOR Vs. SRSWOR

- Then, an unbiased estimator for this is derived as follows: We have

Estimating Gain in efficiency: PPSWOR Vs. SRSWOR

- Then, an unbiased estimator for this is derived as follows: We have
- $$V(t_D) = E(t_D^2) - Y^2$$

Estimating Gain in efficiency: PPSWOR Vs. SRSWOR

- Then, an unbiased estimator for this is derived as follows: We have
- $$V(t_D) = E(t_D^2) - Y^2$$
- So, an unbiased estimator for Y^2 is,

Estimating Gain in efficiency: PPSWOR Vs. SRSWOR

- Then, an unbiased estimator for this is derived as follows: We have
- $$V(t_D) = E(t_D^2) - Y^2$$
- So, an unbiased estimator for Y^2 is,
- $$\hat{Y}^2 = t_D^2 - v(t_D) \text{ ... } (1)$$

Estimating Gain in efficiency: PPSWOR Vs. SRSWOR

- Then, an unbiased estimator for this is derived as follows: We have
- $$V(t_D) = E(t_D^2) - Y^2$$
- So, an unbiased estimator for Y^2 is,
- $$\hat{Y}^2 = t_D^2 - v(t_D) \text{ ... } (1)$$
- Also an unbiased estimator for $\sum_1^N y_i^2$ is $t_D(y^2)$, which is t_D as above with every y in t_D replaced by corresponding y^2. So, an unbaised estimator for $V(N\bar{y})$ is,

Estimating Gain in efficiency: PPSWOR Vs. SRSWOR

- Then, an unbiased estimator for this is derived as follows: We have
- $$V(t_D) = E(t_D^2) - Y^2$$
- So, an unbiased estimator for Y^2 is,
- $$\hat{Y}^2 = t_D^2 - v(t_D) \; ... \; (1)$$
- Also an unbiased estimator for $\sum_1^N y_i^2$ is $t_D(y^2)$, which is t_D as above with every y in t_D replaced by corresponding y^2. So, an unbaised estimator for $V(N\bar{y})$ is,
- $$v_1 = (\frac{N-n}{Nn})\frac{N^2}{N-1}[t_D(y^2) - \frac{\hat{Y}^2}{N}]$$

Estimating Gain in efficiency: PPSWOR Vs. SRSWOR

- Then, an unbiased estimator for this is derived as follows: We have
- $$V(t_D) = E(t_D^2) - Y^2$$
- So, an unbiased estimator for Y^2 is,
- $$\hat{Y}^2 = t_D^2 - v(t_D) \text{ ... } (1)$$
- Also an unbiased estimator for $\sum_1^N y_i^2$ is $t_D(y^2)$, which is t_D as above with every y in t_D replaced by corresponding y^2. So, an unbaised estimator for $V(N\bar{y})$ is,
- $$v_1 = (\frac{N-n}{Nn})\frac{N^2}{N-1}[t_D(y^2) - \frac{\hat{Y}^2}{N}]$$
- with $\hat{Y}^2$ as given in (1)

Estimating Gain in efficiency: PPSWOR Vs. SRSWOR

- Then, an unbiased estimator for this is derived as follows: We have
- $$V(t_D) = E(t_D^2) - Y^2$$
- So, an unbiased estimator for Y^2 is,
- $$\hat{Y}^2 = t_D^2 - v(t_D) \text{ ... } (1)$$
- Also an unbiased estimator for $\sum_1^N y_i^2$ is $t_D(y^2)$, which is t_D as above with every y in t_D replaced by corresponding y^2. So, an unbaised estimator for $V(N\bar{y})$ is,
- $$v_1 = (\frac{N-n}{Nn})\frac{N^2}{N-1}[t_D(y^2) - \frac{\hat{Y}^2}{N}]$$
- with $\hat{Y}^2$ as given in (1)
- Then $G_1 = v_1 - v(t_D)$ unbiasedly estimates gain in efficiency of PPSWOR over SRSWOR.

Practical Example: Data set

- The underlying table consists of the cultivated areas (study variable) and the number of persons living (size variable), for 50 villages in a tehsil. (Data Source: 'Sampling Theory and Methods' By M.N. Murthy Page: 151)

Sl. No.	Cultivated area (in acres)	No. of persons	Sl. No.	Cultivated area (in acres)	No. of persons
1	2544	3295	26	482	1058
2	428	378	27	1527	2111
3	1177	2574	28	1367	1337
4	4567	4466	29	767	827
5	2618	3915	30	1648	2535
6	4113	3249	31	2440	5820
7	4869	3462	32	2434	3378
8	2713	4918	33	1638	1877
9	2237	2461	34	61	3402
10	600	511	35	4505	5769
11	3420	6851	36	1751	3148
12	4012	4782	37	2622	2654
13	1949	3753	38	2848	4201
14	695	1299	39	3013	3523
15	1569	1816	40	1599	1714
16	4562	4942	41	2949	3479
17	2221	2383	42	2641	7420
18	2423	2836	43	1959	2681
19	608	832	44	1371	2870
20	1124	865	45	3290	4435
21	527	588	46	2526	3265
22	2767	6365	47	2935	4096
23	2770	3464	48	1109	984
24	719	941	49	2821	8200
25	607	1287	50	3678	8368

Practical Example: Cumulative Method

- First we create a table for cumulative sizes (Here, the no. of persons living)

Sl. No.	Cultivated area (in acres)	No. of persons	Range		Sl. No.	Cultivated area (in acres)	No. of persons	Range	
1	2544	3295	1	3295	26	482	1058	72234	73291
2	428	378	3296	3673	27	1527	2111	73292	75402
3	1177	2574	3674	6247	28	1367	1337	75403	76739
4	4567	4466	6248	10713	29	767	827	76740	77566
5	2618	3915	10714	14628	30	1648	2535	77567	80101
6	4113	3249	14629	17877	31	2440	5820	80102	85921
7	4869	3462	17878	21339	32	2434	3378	85922	89299
8	2713	4918	21340	26257	33	1638	1877	89300	91176
9	2237	2461	26258	28718	34	61	3402	91177	94578
10	600	511	28719	29229	35	4505	5769	94579	100347
11	3420	6851	29230	36080	36	1751	3148	100348	103495
12	4012	4782	36081	40862	37	2622	2654	103496	106149
13	1949	3753	40863	44615	38	2848	4201	106150	110350
14	695	1299	44616	45914	39	3013	3523	110351	113873
15	1569	1816	45915	47730	40	1599	1714	113874	115587
16	4562	4942	47731	52672	41	2949	3479	115588	119066
17	2221	2383	52673	55055	42	2641	7420	119067	126486
18	2423	2836	55056	57891	43	1959	2681	126487	129167
19	608	832	57892	58723	44	1371	2870	129168	132037
20	1124	865	58724	59588	45	3290	4435	132038	136472
21	527	588	59589	60176	46	2526	3265	136473	139737
22	2767	6365	60177	66541	47	2935	4096	139738	143833
23	2770	3464	66542	70005	48	1109	984	143834	144817
24	719	941	70006	70946	49	2821	8200	144818	153017
25	607	1287	70947	72233	50	3678	8368	153018	161385

Practical Example: Cumulative Method

- Here $N = 50$ and $X = \sum X_i = 161385$

Practical Example: Cumulative Method

- Here $N = 50$ and $X = \sum X_i = 161385$
- We wish to collect a sample of size 20

Practical Example: Cumulative Method

- Here $N = 50$ and $X = \sum X_i = 161385$
- We wish to collect a sample of size 20
- In order to reduce the rejection percentage we take 7 digit random numbers. $\mathrm{MOD}(10000000, 161385) = 155515$ which means 1.55515% of the random numbers get rejected. So our acceptance region is 1 to 9844485.

Practical Example: Cumulative Method

- Here $N = 50$ and $X = \sum X_i = 161385$
- We wish to collect a sample of size 20
- In order to reduce the rejection percentage we take 7 digit random numbers. $\text{MOD}(10000000, 161385) = 155515$ which means 1.55515% of the random numbers get rejected. So our acceptance region is 1 to 9844485.
- In order to get the random number r from 1 to 161385, we choose a 7 digit random number "m" within the acceptance region and and our desired r is MOD(m,161385)

Practical Example: Cumulative Method

- Here $N = 50$ and $X = \sum X_i = 161385$
- We wish to collect a sample of size 20
- In order to reduce the rejection percentage we take 7 digit random numbers. $\text{MOD}(10000000, 161385) = 155515$ which means 1.55515% of the random numbers get rejected. So our acceptance region is 1 to 9844485.
- In order to get the random number r from 1 to 161385, we choose a 7 digit random number "m" within the acceptance region and and our desired r is MOD(m,161385)
- Starting from the 1^{st} row and 1^{st} column of the Fisher Yates Table XXXIII Random Numbers (III)

Practical Example: Cumulative Method

7-digit random no.	Accept/Reject	r	Sample unit no.	Number of persons	Cultivated area (in acres)
2217686	Accept	119681	1	7420	2641
5846895	Accept	37035	2	4782	4012
2392358	Accept	132968	3	3265	2526
7022257	Accept	82702	4	5820	2440
5161094	Accept	158159	5	8368	3678
3950658	Accept	77418	6	827	767
2482034	Accept	61259	7	6365	2767
7193627	Accept	92687	8	3402	61
5946137	Accept	136277	9	4435	3290
9933755	Reject	-	-	-	-
3977327	Accept	104087	10	2654	2622
7098552	Accept	158997	11	8368	3678
530624	Accept	46469	12	1816	1569
7835162	Accept	88682	13	3378	2434
7416772	Accept	154447	14	8368	3678
3027709	Accept	122779	15	7420	2641
6187252	Accept	54622	16	2383	2221
1280624	Accept	150929	17	8200	2821
2593167	Accept	11007	18	3915	2618
1135978	Accept	6283	19	4466	4567
2305474	Accept	46084	20	1816	1569

Practical Example: Lahiri's Method

- The maximum among all the sizes is $X_0 = 8368$.

Practical Example: Lahiri's Method

- The maximum among all the sizes is $X_0 = 8368$.
- Now according to Lahiri's method we have to first choose a random number form 1 to 50. If we choose a 2 digit random number, MOD(100,50)=0 which means no rejection region.

Practical Example: Lahiri's Method

- The maximum among all the sizes is $X_0 = 8368$.
- Now according to Lahiri's method we have to first choose a random number form 1 to 50. If we choose a 2 digit random number, MOD(100,50)=0 which means no rejection region.
- We started reading random numbers from 'Random Numbers Tables by Fisher & Yates' from 11^{st} row and 11^{st} coloumn.

Practical Example: Lahiri's Method

- The maximum among all the sizes is $X_0 = 8368$.
- Now according to Lahiri's method we have to first choose a random number form 1 to 50. If we choose a 2 digit random number, MOD(100,50)=0 which means no rejection region.
- We started reading random numbers from 'Random Numbers Tables by Fisher & Yates' from 11^{st} row and 11^{st} coloumn.
- We take any 2 digit random number "m_1" and our desired population unit no. i is MOD("m_1", 50) + 1.

Practical Example: Lahiri's Method

- The maximum among all the sizes is $X_0 = 8368$.
- Now according to Lahiri's method we have to first choose a random number form 1 to 50. If we choose a 2 digit random number, MOD(100,50)=0 which means no rejection region.
- We started reading random numbers from 'Random Numbers Tables by Fisher & Yates' from 11^{st} row and 11^{st} coloumn.
- We take any 2 digit random number "m_1" and our desired population unit no. i is $\text{MOD}("m_1", 50) + 1$.
- Now again we have to select a random number between 1 and 8368.Taking a 5 digit random number, we see that only 7.952% of random numbers get rejected.

Practical Example: Lahiri's Method

- The maximum among all the sizes is $X_0 = 8368$.
- Now according to Lahiri's method we have to first choose a random number form 1 to 50. If we choose a 2 digit random number, MOD(100,50)=0 which means no rejection region.
- We started reading random numbers from 'Random Numbers Tables by Fisher & Yates' from 11^{st} row and 11^{st} coloumn.
- We take any 2 digit random number "m_1" and our desired population unit no. i is $\text{MOD}("m_1", 50) + 1$.
- Now again we have to select a random number between 1 and 8368.Taking a 5 digit random number, we see that only 7.952% of random numbers get rejected.
- The acceptance region is 1 to 92048.

Practical Example: Lahiri's Method

- The maximum among all the sizes is $X_0 = 8368$.
- Now according to Lahiri's method we have to first choose a random number form 1 to 50. If we choose a 2 digit random number, MOD(100,50)=0 which means no rejection region.
- We started reading random numbers from 'Random Numbers Tables by Fisher & Yates' from 11^{st} row and 11^{st} coloumn.
- We take any 2 digit random number "m_1" and our desired population unit no. i is MOD("m_1", 50) + 1.
- Now again we have to select a random number between 1 and 8368.Taking a 5 digit random number, we see that only 7.952% of random numbers get rejected.
- The acceptance region is 1 to 92048.
- We divide the random number by 8368 and take the remainder as our desired random number r

Practical Example: Lahiri's Method

- The maximum among all the sizes is $X_0 = 8368$.
- Now according to Lahiri's method we have to first choose a random number form 1 to 50. If we choose a 2 digit random number, MOD(100,50)=0 which means no rejection region.
- We started reading random numbers from 'Random Numbers Tables by Fisher & Yates' from 11^{st} row and 11^{st} coloumn.
- We take any 2 digit random number "m_1" and our desired population unit no. i is MOD("m_1", 50) + 1.
- Now again we have to select a random number between 1 and 8368.Taking a 5 digit random number, we see that only 7.952% of random numbers get rejected.
- The acceptance region is 1 to 92048.
- We divide the random number by 8368 and take the remainder as our desired random number r
- Starting from the 11^{th} row and 11^{th} column of the Fisher Yates Table XXXIII Random Numbers (III)

Practical Example: Lahiri's Method

2 digit r.n	Pop.unit no.	5 digit r.n.	Accept/Reject	r	Accept/Reject	Sample unit no.	No. of persons	Cultivated area (in acres)
24	25	36394	Accept	2922	Reject	-	-	-
59	10	11076	Accept	2708	Reject	-	-	-
44	45	14404	Accept	6036	Reject	-	-	-
79	30	96639	Reject	-	-	-	-	-
95	46	45981	Accept	4141	Reject	-	-	-
46	47	8654	Accept	286	Accept	1	4096	2935
48	49	20634	Accept	3898	Accept	2	8200	2821
54	5	28221	Accept	3117	Accept	3	3915	2618
82	33	52033	Accept	1825	Accept	4	1877	1638
72	23	47111	Accept	5271	Reject	-	-	-
51	2	12912	Accept	4544	Reject	-	-	-
56	7	19648	Accept	2912	Accept	5	3462	4869
29	30	30716	Accept	5612	Reject	-	-	-
91	42	36698	Accept	3226	Accept	6	7420	2641
59	10	29065	Accept	3961	Reject	-	-	-
36	37	87196	Accept	3516	Reject	-	-	-
95	46	30241	Accept	5137	Reject	-	-	-
39	40	84623	Accept	943	Accept	7	1714	1599
43	44	34278	Accept	806	Accept	8	2870	1371
74	25	51399	Accept	1191	Accept	9	1287	607
82	33	24449	Accept	7713	Reject	-	-	-
16	17	18097	Accept	1361	Accept	10	2383	2221

Practical Example: Estimation of total cultivated area

- We will use the sample collected by Cumulative method in order to estimate the total cultivated area of the Tehsil.

Practical Example: Estimation of total cultivated area

- We will use the sample collected by Cumulative method in order to estimate the total cultivated area of the Tehsil.
- The sample collected is a PPSWR sample.

Practical Example: Estimation of total cultivated area

- We will use the sample collected by Cumulative method in order to estimate the total cultivated area of the Tehsil.
- The sample collected is a PPSWR sample.
- For PPSWR sample, unbiased estimator of Y (the population total) is $\frac{1}{n}\sum_{j=1}^{n}\frac{y_j}{p_j}$ where $p_j = x_j/X$.

Practical Example: Estimation

Sl. No	Number of persons (x_j)	Cultivated area (in acres) (y_j)	p_j	y_j/p_j	$(y_j/p_j)^2$
1	7420	2641	0.045977011	57441.75	3299554643
2	4782	4012	0.029631007	135398.7077	18332810034
3	3265	2526	0.020231124	124857.124	15589301424
4	5820	2440	0.036062831	67659.69072	4577833749
5	8368	3678	0.051851163	70933.799	5031603840
6	827	767	0.005124392	149676.2938	22402992936
7	6365	2767	0.039439849	70157.46976	4922070563
8	3402	61	0.021080026	2893.734568	8373699.749
9	4435	3290	0.027480869	119719.6505	14332794718
10	2654	2622	0.016445147	159439.1372	25420838456
11	8368	3678	0.051851163	70933.799	5031603840
12	1816	1569	0.011252595	139434.5072	19441981787
13	3378	2434	0.020931313	116285.1066	13522226010
14	8368	3678	0.051851163	70933.799	5031603840
15	7420	2641	0.045977011	57441.75	3299554643
16	2383	2221	0.014765932	150413.7998	22624311180
17	8200	2821	0.050810174	55520.37622	3082512176
18	3915	2618	0.02425876	107919.7778	11646678436
19	4466	4567	0.027672956	165034.7727	27236476209
20	1816	1569	0.011252595	139434.5072	19441981787
Total				**2031529.553**	**2.44277E+11**

Practical Example: Estimation

- An unbiased estimate of the total cultivated area of the tehsil is 101576.4776 acres.

Practical Example: Estimation

- An unbiased estimate of the total cultivated area of the tehsil is 101576.4776 acres.
- On calculating $v(\hat{Y}_{PPSWR}) = \frac{1}{n(n-1)}\left[\sum_{i=1}^{n}(\frac{y_j}{p_j})^2 - n\hat{Y}^2_{PPSWR}\right]$ we get 99793388.95 as an unbiased estimate of the sampling variance

Practical Example: Estimation

- An unbiased estimate of the total cultivated area of the tehsil is 101576.4776 acres.
- On calculating $v(\hat{Y}_{PPSWR}) = \frac{1}{n(n-1)}\left[\sum_{i=1}^{n}(\frac{y_j}{p_j})^2 - n\hat{Y}^2_{PPSWR}\right]$ we get 99793388.95 as an unbiased estimate of the sampling variance
- An estimate of the standard error of the estimate is 9989.664106

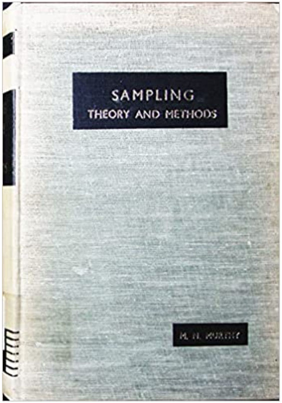
References

- 'Sampling Theory and Methods' By M.N. Murthy

References

- 'Sampling Theory and Methods' By M.N. Murthy
- https://1lib.in/book/2468591/98067c

References

- 'Sampling Theory and Methods' By M.N. Murthy
- https://1lib.in/book/2468591/98067c
- 

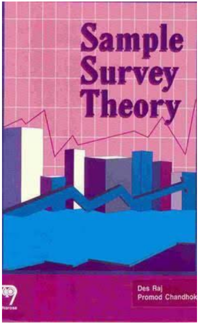
References

- 'Sample Survey Theory' By Des Raj and Promod Chandhok

References

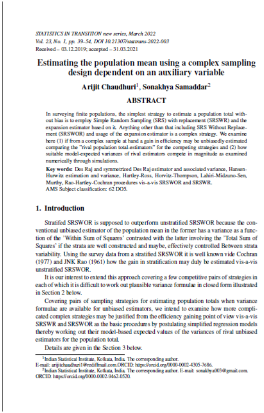
- 'Sample Survey Theory' By Des Raj and Promod Chandhok
- https://1lib.in/book/2468591/98067c

References

- 'Sample Survey Theory' By Des Raj and Promod Chandhok
- https://1lib.in/book/2468591/98067c
- 

References

- 'Estimating the population mean using a complex sampling design dependent on an auxiliary variable' by Arijit Chaudhuri and Sonakhya Samaddar.
- http://cejsh.icm.edu.pl/cejsh/element/bwmeta1.element.ojs-doi-10_21307_stattrans-2022-003



STATISTICS IN TRANSITION new series, March 2022
Vol. 23, No. 1, pp. 39–54, DOI 10.21307/stattrans-2022-003
Received – 03.12.2019; accepted – 31.03.2021

Estimating the population mean using a complex sampling design dependent on an auxiliary variable

Arijit Chaudhuri[1], Sonakhya Samaddar[2]

ABSTRACT

In surveying finite populations, the simplest strategy to estimate a population total without bias is to employ Simple Random Sampling (SRS) with replacement (SRSWR) and the expansion estimator based on it. Anything other than that including SRS Without Replacement (SRSWOR) and usage of the expansion estimator is a complex strategy. We examine here (1) if from a complex sample at hand a gain in efficiency may be unbiasedly estimated comparing the "rival population total-estimators" for the competing strategies and (2) how suitable model-expected variances of rival estimators compete in magnitude as examined numerically through simulations.

Key words: Des Raj and symmetrized Des Raj estimator and associated variance, Hansen-Hurwitz estimation and variance, Hartley-Ross, Horvitz-Thompson, Lahiri-Midzuno-Sen, Murthy, Rao-Hartley-Cochran procedures vis-a-vis SRSWOR and SRSWR.
AMS Subject classification: 62 DO5.

1. Introduction

Stratifed SRSWOR is supposed to outperform unstratified SRSWOR because the conventional unbiased estimator of the population mean in the former has a variance as a function of the 'Within Sum of Squares' contrasted with the latter involving the 'Total Sum of Squares' if the strata are well constructed and maybe, effectively controlled Between strata variability. Using the survey data from a stratified SRSWOR it is well known vide Cochran (1977) and JNK Rao (1961) how the gain in stratification may duly be estimated vis-a-vis unstratified SRSWOR.

It is our interest to extend this approach covering a few competitive pairs of strategies in each of which it is difficult to work out plausible variance formulae in closed form illustrated in Section 2 below.

Covering pairs of sampling strategies for estimating population totals when variance formulae are available for unbiased estimators, we intend to examine how more complicated complex strategies may be justified from the efficiency gaining point of view vis-a-vis SRSWR and SRSWOR as the basic procedures by postulating simplified regression models thereby working out their model-based expected values of the variances of rival unbiased estimators for the population total.

Details are given in the Section 3 below.

[1]Indian Statistical Institute, Kotkata, India. The corresponding author.
E-mail: arijitchaudhuri1@rediffmail.com.ORCID: https://orcid.org/0000-0002-4305-7686.
[2]Indian Statistical Institute, Kotkata, India. The corresponding author. E-mail: sonakhya003@gmail.com.
ORCID: https://orcid.org/0000-0002-9462-0520.

-

Acknowledgement

- We would like to express our special thanks of gratitude to our respected **Professor Biswajit Roy** who gave us the golden opportunity to do this wonderful presentation on the topic **Probability Proportional to Size (PPS) Sampling**, which also helped us in doing a lot of Research and we came to know about so many new things. We are really thankful to him.