

Sampling Theory

Note : I'm combinig note after exam if over.

→ I performed very bad in exam. I had no understanding of statistics and i strugulled a lot.
But some how i able to pass this course, Profs. Blessing

$X \sim N(\mu, \sigma^2)$ → standard normal distribution

The symbol \sim means “is distributed as.”

$N(\mu, \sigma^2)$ denotes a Normal distribution with

- mean μ
- variance σ^2

$X \sim Exp(1) \rightarrow$ exponential distribution with rate 1, mean 1, variance 1

[Standard deviation - Wikipedia](#)

[Expected value - Wikipedia](#)

population mean

$$\mu = \frac{1}{N} \sum_{i=1}^N x_i$$

population variance

$$\sigma^2 = \frac{1}{N} \sum_{i=1}^N (x_i - \mu)^2$$

[Variance - Wikipedia](#)

$$Var(X) = E[(X - \mu)^2]$$

where $\mu = E[X]$

$$Var(X) = E[X^2 - 2\mu X + \mu^2] = E[X^2] - 2\mu E[X] + \mu^2 = E[X^2] - 2\mu^2 + \mu^2$$

μ^2 is just a constant, so $E[\mu^2] = \mu^2$

$$Var(X) = E[X^2] - \mu^2$$

$$Var(X) = E[X^2] - (E[X])^2$$

SRS = Simple Random Sampling

A sampling method where each unit of the population has an equal chance of being selected.

With Replacement (SRSWR)

Definition: After selecting an item, you put it back into the population before the next draw.

Each draw is independent, because the population size doesn't change.

A unit can be chosen more than once.

Example: Population = {A, B, C}, sample size $n = 2$

Possible samples:

(A, A), (B, B), (C, C), (A, B) (A, C), (B, A), (B, C), (C, A), (C, B)

Total = $N^n = 3^2 = 9$ ordered samples.

Without Replacement (SRSWOR)

Definition: After selecting an item, you do not put it back.

Each draw is dependent, because the population shrinks.

A unit can be chosen at most once.

Example: Population = {A, B, C}, sample size $n = 2$

Possible samples: (A, B) (A, C), (B, A), (B, C), (C, A), (C, B)

Total = $\frac{N!}{(N-n)!} = \frac{3!}{(3-2)!} = 6$ ordered samples

	SRSWR	SRSWOR
Replacement	Allowed	Not allowed
Sample size effect	Population size stays same	Population shrinks
Independence	Each draw independent	Draws are dependent
Probability of selecting a unit in one draw	$\frac{1}{N}$	Changes with each draw
Total possible ordered samples (size nnn)	N^n	$\frac{N!}{(N-n)!}$

[Sampling 03: Stratified Random Sampling](#)

[What Are The Types Of Sampling Techniques In Statistics - Random, Stratified, Cluster, Systematic](#)

Introduction to Sampling Theory

Sampling Methods- Exercises and Solutions : Pascal Ardilly and Yves Tille
([Download here through IITK Library link](#))

home.iitk.ac.in/~shalab/course432.htm
home.iitk.ac.in/~neeraj/mth432/mth432.htm

[Sampling 03: Stratified Random Sampling](#)

[What Are The Types Of Sampling Techniques In Statistics - Random, Stratified, Cluster, Systematic](#)

Unbiased estimator :

An estimator is unbiased if its expected value (mean of its sampling distribution) equals the true population parameter. $E[\hat{\theta}] = \theta$

The sample mean \bar{X} is unbiased for the population mean μ , because $E[\bar{X}] = \mu$

Biased estimator

An estimator is biased if its expected value does not equal the parameter: $E[\hat{\theta}] \neq \theta$

If you estimate variance using $S_n^2 = \frac{1}{n} \sum_{j=1}^n (X_j - \bar{X})^2$, then $E[S_n^2] = \frac{n-1}{n} \sigma^2$

So it underestimates the true variance σ^2 .

→ This is a biased estimator of variance.

To fix this, we divide by $n-1$ instead of n , $S_n^2 = \frac{1}{n-1} \sum_{j=1}^n (X_j - \bar{X})^2$ which is an unbiased estimator of variance.

Sampling Methods- Exercises and Solutions : Pascal Ardilly and Yves Tille

1. **SRSWR, SRSWOR** → Everyone equal chance, “lottery sampling”.
2. **Stratified Sampling** → Divide into groups, sample each, improves precision.
3. **Unequal Probability Sampling** → Different units have different selection chances, but use HT estimator to adjust.
4. **Horvitz–Thompson Estimator** → General tool to make estimates unbiased under unequal probability designs.

Intuition

SRS = fairness (everyone equal chance).

Stratified = fairness + efficiency (guarantee all groups represented).

Unequal probability = practicality (focus on important units), corrected by Horvitz–Thompson weights.

$X \sim N(\mu, \sigma^2) \rightarrow$ standard normal distribution

The symbol \sim means “is distributed as.”

$N(\mu, \sigma^2)$ denotes a Normal distribution with

- mean μ
- variance σ^2

$X \sim Exp(1) \rightarrow$ exponential distribution with rate 1, mean 1, variance 1

[Standard deviation - Wikipedia](#)

[Expected value - Wikipedia](#)

population mean

$$\mu = \frac{1}{N} \sum_{i=1}^N x_i$$

population variance

$$\sigma^2 = \frac{1}{N} \sum_{i=1}^N (x_i - \mu)^2$$

[Variance - Wikipedia](#)

$$Var(X) = E[(X - \mu)^2]$$

where $\mu = E[X]$

$$Var(X) = E[X^2 - 2\mu X + \mu^2] = E[X^2] - 2\mu E[X] + \mu^2 = E[X^2] - 2\mu^2 + \mu^2$$

μ^2 is just a constant, so $E[\mu^2] = \mu^2$

$$Var(X) = E[X^2] - \mu^2$$

$$Var(X) = E[X^2] - (E[X])^2$$

Cov means Covariance. It measures how two random variables move together.

For two random variable X and Y :

$$Cov(X, Y) = E[(X - E[X])(Y - E[Y])]$$

Alternative form:

$$Cov(X, Y) = E[XY] - E[X]E[Y]$$

Intuition

- If X and Y increase/decrease together, covariance is positive.
- If one increases while the other decreases, covariance is negative.
- If they are independent (not always but often), covariance is zero.

Eg :

Suppose X = hours studied, Y = exam marks.

- When study hours increase, marks also increase \rightarrow positive covariance.
- If X = hours studied, Y = hours spent gaming (and more gaming reduces marks) \rightarrow negative covariance.

Variance is just covariance with itself:

$$Var(X) = Cov(X, X)$$

1. Consider a sampling design for sampling from a population $U = (U_1, U_2, U_3)$ of three units, with study variables $Y = (Y_1, Y_2, Y_3)$ and

$$P(\underline{s}) = \begin{cases} \frac{1}{7} & \text{if } \underline{s} = (U_1, U_3) \\ \frac{2}{7} & \text{if } \underline{s} = (U_2, U_3) \\ \frac{4}{7} & \text{if } \underline{s} = (U_1, U_2, U_3) \end{cases}$$

Find

- the first order inclusion probabilities π_1, π_2 and π_3 .
- the second order inclusion probabilities π_{12}, π_{13} and π_{23}

(iii) the expected value and the variance of the estimator $\hat{\eta}(\underline{S}) = \frac{3}{n(\underline{S})} \sum_{i \in \underline{S}} Y_i$ of the population total.

- the first order inclusion probabilities π_1, π_2 and π_3 .

Only 3 samples are possible:

- $s_1 = (U_1, U_3)$
- $s_2 = (U_2, U_3)$
- $s_3 = (U_1, U_2, U_3)$

Compute First-Order Inclusion Probability

$$\pi_i = \sum_{\text{samples that include } U_i} P(s)$$

π_1 (Inclusion probability for unit U_1)

Appears in:

- $s_1 = (U_1, U_3) : \frac{1}{7}$
- $s_3 = (U_1, U_2, U_3) : \frac{4}{7}$

$$\pi_1 = \frac{1}{7} + \frac{4}{7} = \frac{5}{7}$$

π_2 (Inclusion probability for unit U_2)

Appears in:

- $s_2 = (U_2, U_3) : \frac{2}{7}$
- $s_3 = (U_1, U_2, U_3) : \frac{4}{7}$

$$\pi_2 = \frac{2}{7} + \frac{4}{7} = \frac{6}{7}$$

π_3 (Inclusion probability for unit U_3)

Appears in:

- $s_1 = (U_1, U_3) : \frac{1}{7}$
- $s_2 = (U_2, U_3) : \frac{2}{7}$
- $s_3 = (U_1, U_2, U_3) : \frac{4}{7}$

$$\pi_3 = \frac{1}{7} + \frac{2}{7} + \frac{4}{7} = \frac{7}{7} = 1$$

✓ Final Answers:

$$\pi_1 = \frac{5}{7}, \pi_2 = \frac{6}{7}, \pi_3 = 1$$

(ii) the second order inclusion probabilities π_{12} , π_{13} and π_{23}
 Compute Second-Order Inclusion Probability

$$\pi_{ij} = \sum_{\text{samples that include } U_i \text{ and } U_j} P(s)$$

π_{12} probability that both U_1 and U_2 are in the sample

Only sample 3 (U_1, U_2, U_3) includes both U_1 and U_2

$$\pi_{12} = P((U_1, U_2, U_3)) = \frac{4}{7}$$

π_{13} probability that both U_1 and U_3 are in the sample

- $s_1 = (U_1, U_3) : \frac{1}{7}$
- $s_3 = (U_1, U_2, U_3) : \frac{4}{7}$

$$\pi_{13} = \frac{1}{7} + \frac{4}{7} = \frac{5}{7}$$

π_{23} probability that both U_2 and U_3 are in the sample

- $s_2 = (U_2, U_3) : \frac{2}{7}$
- $s_3 = (U_1, U_2, U_3) : \frac{4}{7}$

$$\pi_{23} = \frac{2}{7} + \frac{4}{7} = \frac{6}{7}$$

✓ Final Answers:

$$\pi_{12} = \frac{4}{7}, \pi_{13} = \frac{5}{7}, \pi_{23} = \frac{6}{7}$$

(iii) the expected value and the variance of the estimator $\hat{\eta}(\underline{S}) = \frac{3}{n(\underline{S})} \sum_{i \in \underline{S}} Y_i$ of
 the population total.

expected value of the estimator $E(\hat{\eta}(\underline{S}))$

Compute $\hat{\eta}(\underline{s}) = \frac{3}{n(\underline{S})} \sum_{i \in \underline{S}} Y_i$ for Each Sample

$n(\underline{S})$ is the number of elements in the set S

$$E[\hat{\eta}(\underline{S})] = E\left[\frac{3}{n(\underline{S})} \sum_{i \in \underline{S}} Y_i\right]$$

$$\begin{aligned} E[\hat{\eta}(\underline{S})] &= \left(\frac{3}{2}(Y_1 + Y_3)\right)\frac{1}{7} + \left(\frac{3}{2}(Y_2 + Y_3)\right)\frac{2}{7} + \left(\frac{3}{3}(Y_1 + Y_2 + Y_3)\right)\frac{4}{7} \\ &= \frac{3}{14}(Y_1 + Y_3) + \frac{3}{7}(Y_2 + Y_3) + \frac{4}{7}(Y_1 + Y_2 + Y_3) \\ &= \left(\frac{3}{14}Y_1 + \frac{4}{7}Y_1\right) + \left(\frac{3}{7}Y_2 + \frac{4}{7}Y_2\right) + \left(\frac{3}{14}Y_3 + \frac{3}{7}Y_3 + \frac{4}{7}Y_3\right) \\ &= \left(\frac{3}{14} + \frac{4}{7}\right)Y_1 + \left(\frac{3}{7} + \frac{4}{7}\right)Y_2 + \left(\frac{3}{14} + \frac{3}{7} + \frac{4}{7}\right)Y_3 \\ &= \frac{11}{14}Y_1 + Y_2 + \frac{17}{14}Y_3 \end{aligned}$$

$$\mathbb{E}(\hat{\eta}(\underline{S})) = \frac{11}{14}Y_1 + Y_2 + \frac{17}{14}Y_3$$

variance of the estimator $Var(\hat{\eta}(\underline{S}))$:

$$Var(\hat{\eta}) = \mathbb{E}[\hat{\eta}(S)^2] - (\mathbb{E}[\hat{\eta}(S)])^2$$

$$\mathbb{E}[\hat{\eta}^2] = \frac{1}{7}\left(\frac{3}{2}(Y_1 + Y_3)\right)^2 + \frac{2}{7}\left(\frac{3}{2}(Y_2 + Y_3)\right)^2 + \frac{4}{7}\left(\frac{3}{3}(Y_1 + Y_2 + Y_3)\right)^2$$

For $s_1 = (Y_1 + Y_3)$

$$s_1 = \frac{1}{7}\left(\frac{3}{2}(Y_1 + Y_3)\right)^2 = \frac{1}{7} \cdot \frac{9}{4}(Y_1 + Y_3)^2 = \frac{9}{28}(Y_1^2 + 2Y_1Y_3 + Y_3^2)$$

For $s_2 = (Y_2 + Y_3)$

$$s_2 = \frac{2}{7}\left(\frac{3}{2}(Y_2 + Y_3)\right)^2 = \frac{2}{7} \cdot \frac{9}{4}(Y_2 + Y_3)^2 = \frac{9}{14}(Y_2^2 + 2Y_2Y_3 + Y_3^2)$$

For $s_3 = (Y_1 + Y_2 + Y_3)$

$$s_3 = \frac{4}{7}\left(\frac{3}{3}(Y_1 + Y_2 + Y_3)\right)^2 = \frac{4}{7}(Y_1^2 + Y_2^2 + Y_3^2 + 2Y_1Y_2 + 2Y_2Y_3 + 2Y_3Y_1)$$

$$\begin{aligned}
(\mathbb{E}[\hat{\eta}(\underline{S})])^2 &= \left(\frac{11}{14}Y_1 + Y_2 + \frac{17}{14}Y_3 \right)^2 \\
&= \left(\frac{11}{14}Y_1 \right)^2 + (Y_2)^2 + \left(\frac{17}{14}Y_3 \right)^2 + 2\left(\frac{11}{14}Y_1 \cdot Y_2 \right) + 2\left(Y_2 \cdot \frac{17}{14}Y_3 \right) + 2\left(\frac{17}{14}Y_3 \cdot \frac{11}{14}Y_1 \right)
\end{aligned}$$

#Pending

Rough Work

1. Sample (U_1, U_3) with probability $\frac{1}{7}$, size $n_1 = 2$

$$\hat{\eta}_1 = \frac{1}{7} \left(\frac{3}{2}(Y_1 + Y_3) \right) = \frac{3}{14}(Y_1 + Y_3)$$

1. Sample (U_2, U_3) with probability $\frac{2}{7}$, size $n_2 = 2$

$$\hat{\eta}_2 = \frac{2}{7} \left(\frac{3}{2}(Y_2 + Y_3) \right) = \frac{3}{7}(Y_2 + Y_3)$$

1. Sample (U_1, U_2, U_3) with probability $\frac{4}{7}$, size $n_3 = 3$

$$\hat{\eta}_3 = \frac{4}{7} \left(\frac{3}{3}(Y_1 + Y_2 + Y_3) \right) = \frac{4}{7}(Y_1 + Y_2 + Y_3)$$

2. Consider a sampling design for sampling from a population $U = (U_1, U_2, U_3)$ of three units, with study variables $Y = (Y_1, Y_2, Y_3)$ and

$$Q(\underline{s}) \begin{cases} k, & \text{if } n(\underline{s}) = r(\underline{s}) = 2 \\ 0, & \text{otherwise} \end{cases}$$

where k is a fixed positive constant.

Find

- (i) the first order inclusion probabilities π_1, π_2 and π_3 .
- (ii) the second order inclusion probabilities π_{12}, π_{13} and π_{23}

(iii) the expected value and the variance of the estimator $\hat{\eta}(\underline{S}) = \frac{3}{n(\underline{S})} \sum_{i \in \underline{S}} Y_i$ of the population total.

\Rightarrow

$n(\underline{s})$: sample size (count of elements in \underline{s})

$r(\underline{s})$: number of distinct units ((since here we take sets $n = r$ means no repetition)).

Since we choose exactly 2 distinct units from $\{U_1, U_2, U_3\}$ the possible samples are:

$$\{U_1, U_2\}, \{U_2, U_3\}, \{U_3, U_1\}$$

That's 3 possible samples.

The probabilities must sum to 1 : $\sum_{\underline{s}} Q(\underline{s}) = 1$

Each of the 3 samples has probability k , $3k = 1$, $k = \frac{1}{3}$

$$\hat{\eta}(\underline{S}) = \frac{3}{n(\underline{S})} \sum_{i \in \underline{S}} Y_i = \frac{3}{2} \sum_{i \in \underline{S}} Y_i \because n(\underline{s}) = 2$$

(i) the first order inclusion probabilities π_1 , π_2 and π_3 .

For π_1

U_1 appears in two samples:

$$\{U_1, U_2\} \rightarrow \text{probability } k = 1/3$$

$$\{U_3, U_1\} \rightarrow \text{probability } k = 1/3$$

$$\pi_1 = \frac{1}{3} + \frac{1}{3} = \frac{2}{3}$$

same for π_2 and π_3

(ii) the second order inclusion probabilities π_{12} , π_{13} and π_{23}

Definition : $\pi_{ij} = P$ (both unit i and j included in selected sample)

for π_{12}

U_{12} appears in one

$$\{U_1, U_2\} \rightarrow \text{probability } k = 1/3$$

$$\pi_1 = \frac{1}{3}$$

same for π_2 and π_3

(iii) the expected value and the variance of the estimator $\hat{\eta}(\underline{S}) = \frac{3}{n(\underline{S})} \sum_{i \in \underline{S}} Y_i$ of

the population total.

expected value

We have estimator $\hat{\eta}(\underline{S}) = \frac{3}{n(\underline{S})} \sum_{i \in \underline{S}} Y_i$

Under the design only samples of size 2 occur, so $n(\underline{S}) = 2$ for every possible sample. Thus

$$\hat{\eta}(\underline{S}) = \frac{3}{2} \sum_{i \in \underline{S}} Y_i$$

$$E[\hat{\eta}(\underline{S})] = E\left[\frac{3}{n(\underline{S})} \sum_{i \in \underline{S}} Y_i\right]$$

$$E[\hat{\eta}(\underline{S})] = \left(\frac{3}{2}(Y_1 + Y_3)\right)\frac{1}{7} + \left(\frac{3}{2}(Y_2 + Y_3)\right)\frac{2}{7} + \left(\frac{3}{3}(Y_1 + Y_2 + Y_3)\right)\frac{4}{7}$$

$$\mathbb{E}(\hat{\eta}(\underline{S})) = \frac{1}{3}\left[\frac{3}{2}(Y_1 + Y_2) + \frac{3}{2}(Y_2 + Y_3) + \frac{3}{2}(Y_3 + Y_1)\right] = Y_1 + Y_2 + Y_3$$

$$\begin{aligned} E[\hat{\eta}(\underline{S})] &= \left(\frac{3}{2}(Y_1 + Y_2)\right)\frac{1}{3} + \left(\frac{3}{2}(Y_2 + Y_3)\right)\frac{1}{3} + \left(\frac{3}{2}(Y_3 + Y_1)\right)\frac{1}{3} \\ &= \frac{1}{3}\left[\frac{3}{2}(Y_1 + Y_2) + \frac{3}{2}(Y_2 + Y_3) + \frac{3}{2}(Y_3 + Y_1)\right] \\ &= \frac{1}{3}\left(\frac{3}{2}\right)[(Y_1 + Y_2) + (Y_2 + Y_3) + (Y_3 + Y_1)] \\ &= \frac{1}{2}[2Y_1 + 2Y_2 + 2Y_3] \\ &= \frac{1}{2}(2)[Y_1 + Y_2 + Y_3] \\ &= Y_1 + Y_2 + Y_3 \end{aligned}$$

variance of the estimator $Var(\hat{\eta}(\underline{S}))$:

$$Var(\hat{\eta}) = \mathbb{E}[\hat{\eta}(\underline{S})^2] - (\mathbb{E}[\hat{\eta}(\underline{S})])^2$$

$$(\mathbb{E}[\hat{\eta}(\underline{S})])^2 = (Y_1 + Y_2 + Y_3)^2$$

$$\mathbb{E}[\hat{\eta}(\underline{S})^2] = \left(\frac{3}{2}(Y_1 + Y_2)\right)^2 \cdot \frac{1}{3} + \left(\frac{3}{2}(Y_2 + Y_3)\right)^2 \cdot \frac{1}{3} + \left(\frac{3}{2}(Y_3 + Y_1)\right)^2 \cdot \frac{1}{3}$$

$$\begin{aligned}
&= \frac{1}{3} \left[\left(\frac{3}{2}(Y_1 + Y_2) \right)^2 + \left(\frac{3}{2}(Y_2 + Y_3) \right)^2 + \left(\frac{3}{2}(Y_3 + Y_1) \right)^2 \right] \\
&= \frac{1}{3} \left(\frac{3}{2} \right)^2 \left[(Y_1 + Y_2)^2 + (Y_2 + Y_3)^2 + (Y_3 + Y_1)^2 \right] \\
&= \frac{3}{4} \left[(Y_1 + Y_2)^2 + (Y_2 + Y_3)^2 + (Y_3 + Y_1)^2 \right]
\end{aligned}$$

$$\begin{aligned}
Var(\hat{\eta}) &= \mathbb{E}[\hat{\eta}(S)^2] - (\mathbb{E}[\hat{\eta}(S)])^2 \\
&\left[\frac{3}{4} \left[(Y_1 + Y_2)^2 + (Y_2 + Y_3)^2 + (Y_3 + Y_1)^2 \right] \right] - [Y_1 + Y_2 + Y_3]
\end{aligned}$$

3. Let A_i , $i = 1, \dots, N$, denote the number of times the i^{th} unit U_i appears in the sample drawn from the population $U = (U_1, \dots, U_N)$ under the $SRSWOR(n)$ design. Show that

$Cov(A_i, A_j) = -\frac{n(N-n)}{N^2(N-1)}$, $i \neq j$. Also find $Cov(A_i, A_j)$, $i \neq j$, under the $SRSWR(s)$ design.

Let A_i , $i = 1, \dots, N$, denote the number of times the i^{th} unit U_i appears in the sample drawn from the population $U = (U_1, \dots, U_N)$
find $Cov(A_i, A_j)$, $i \neq j$, under the $SRSWR(s)$ design.

SRSWOR (without replacement)

When sampling without replacement each unit can appear at most once, so $A \in \{0, 1\}$

Therefore, $A_i = 1$ if unit i is selected, and 0 otherwise

$$A_i = \begin{cases} 1, & \text{if unit } i \text{ is selected} \\ 0, & \text{otherwise} \end{cases}$$

$$Cov(A_i, A_j) = \mathbb{E}[A_i A_j] - \mathbb{E}[A_i] \mathbb{E}[A_j]$$

$$P(A_i = 1) = \frac{\binom{N-1}{n-1}}{\binom{N}{n}} = \frac{n}{N},$$

$$P(A_i = 1, A_j = 1) = \frac{\binom{N-2}{n-2}}{\binom{N}{n}} \frac{\frac{(N-2)!}{(n-2)!(N-n)!}}{\frac{N!}{n!(N-n)!}} = \frac{n(n-1)}{N(N-1)}$$

(for $i \neq j$). Compute the joint probability in closed form:

Thus

$$E[A_i]E[A_j] = P(A_i = 1, A_j = 1) = \frac{n^2}{N^2}$$

$$E[A_iA_j] = P(A_i = 1, A_j = 1) = \frac{n(n-1)}{N(N-1)},$$

So

$$Cov(A_i, A_j) = \frac{n(n-1)}{N(N-1)} - \frac{n^2}{N^2} = \frac{nN(n-1) - n^2(N-1)}{N^2(N-1)} = \frac{n(N(n-1) - n(N-1))}{N^2(N-1)}$$

$$N(n-1) - n(N-1) = Nn - N - nN + n = n - N$$

$$Cov(A_i, A_j) = \frac{n(n-N)}{N^2(N-1)} = -\frac{n(N-n)}{N^2(N-1)} \quad (i \neq j)$$

SRSWR (with replacement)

In sampling with replacement:

- Each draw is independent.
- Each unit has probability $\frac{1}{N}$ of being selected on any draw.
- The number of times unit i appears in n draws, $A_i \sim \text{Binomial}\left(n, \frac{1}{N}\right)$

So for $i \neq j$ the joint distribution of (A_i, A_j) comes from multinomial:

$$(A_1, A_2, \dots, A_N) \sim \text{Multinomial}\left(n; \frac{1}{N}, \dots, \frac{1}{N}\right)$$

Properties of the multinomial distribution:

- $E[A_i] = n \cdot \frac{1}{N}$
- $Var(A_i) = n \cdot \frac{1}{N} \cdot \left(1 - \frac{1}{N}\right)$
- $Cov(A_i, A_j) = -n \frac{1}{N} \frac{1}{N} = -\frac{n}{N^2} \quad (i \neq j)$

$$Cov(A_i, A_j) = -\frac{n}{N^2} \quad (i \neq j)$$

$Cov(A_i, A_j)$ means the covariance between the random variables A_i and A_j .

[wikipedia.org/wiki/Covariance](https://en.wikipedia.org/wiki/Covariance)

$$cov(X, Y) = E[XY] - E[X]E[Y]$$

Q4. Let $y = ((X_1, Y_1), \dots, (X_N, Y_N))$, $\bar{X} = \frac{1}{N} \sum_{i=1}^N X_i$, $\bar{Y} = \frac{1}{N} \sum_{i=1}^N Y_i$ and

$C_{\underline{X}, \underline{Y}} = \frac{1}{N} \sum_{i=1}^N (X_i - \bar{X})(Y_i - \bar{Y})$. Under the SRSWR(n) design, let $\widehat{\bar{X}} = \frac{1}{n} \sum_{i \in \underline{S}} X_i$,

$\widehat{\bar{Y}} = \frac{1}{n} \sum_{i \in \underline{S}} Y_i$ and $\widehat{C}_{\underline{X}, \underline{Y}} = \frac{1}{n-1} \sum_{i \in \underline{S}} (X_i - \widehat{\bar{X}})(Y_i - \widehat{\bar{Y}})$. Find

i. the correlation coefficient between $\widehat{\bar{X}}$ and $\widehat{\bar{Y}}$.

ii. $E(\widehat{C}_{\underline{X}, \underline{Y}})$

Let $y = ((X_1, Y_1), \dots, (X_N, Y_N))$, $\bar{X} = \frac{1}{N} \sum_{i=1}^N X_i$, $\bar{Y} = \frac{1}{N} \sum_{i=1}^N Y_i$ and $C_{\underline{X}, \underline{Y}} = \frac{1}{N} \sum_{i=1}^N (X_i - \bar{X})(Y_i - \bar{Y})$

Under the SRSWR(n) design, let $\widehat{\bar{X}} = \frac{1}{n} \sum_{i \in \underline{S}} X_i$, $\widehat{\bar{Y}} = \frac{1}{n} \sum_{i \in \underline{S}} Y_i$ and $\widehat{C}_{\underline{X}, \underline{Y}} = \frac{1}{n-1} \sum_{i \in \underline{S}} (X_i - \widehat{\bar{X}})(Y_i - \widehat{\bar{Y}})$

Find $E(\widehat{C}_{\underline{X}, \underline{Y}})$

We are given

- A finite population of size $N : y = ((X_1, Y_1), \dots, (X_N, Y_N))$
- Population means:

$$\bar{X} = \frac{1}{N} \sum_{i=1}^N X_i, \quad \bar{Y} = \frac{1}{N} \sum_{i=1}^N Y_i$$

- Population covariance:

$$C_{\underline{X}, \underline{Y}} = \frac{1}{N} \sum_{i=1}^N (X_i - \bar{X})(Y_i - \bar{Y})$$

Under Simple Random Sampling With Replacement (SRSWR) of size n , the sample statistics are:

- Sample means:

$$\widehat{\bar{X}} = \frac{1}{n} \sum_{i \in \underline{S}} X_i, \quad \widehat{\bar{Y}} = \frac{1}{n} \sum_{i \in \underline{S}} Y_i$$

- Sample covariance:

$$\widehat{C}_{\underline{X}, \underline{Y}} = \frac{1}{n-1} \sum_{i \in \underline{S}} (X_i - \widehat{\bar{X}})(Y_i - \widehat{\bar{Y}})$$

to find: the correlation coefficient between $\widehat{\bar{X}}$ and $\widehat{\bar{Y}}$.

The correlation coefficient between two estimators $\widehat{\bar{X}}$ and $\widehat{\bar{Y}}$ is defined as:

$$\rho(\widehat{\bar{X}}, \widehat{\bar{Y}}) = \text{Corr}(\widehat{\bar{X}}, \widehat{\bar{Y}}) = \frac{\text{Cov}(\widehat{\bar{X}}, \widehat{\bar{Y}})}{\sqrt{\text{Var}(\widehat{\bar{X}})\text{Var}(\widehat{\bar{Y}})}}$$

We now compute each term separately under SRSWR.

Under SRSWR, every draw is independent and with equal probability. So, for sample mean:

The variance of the sample mean $\widehat{\bar{X}}$ is :

$$\text{Var}(\widehat{\bar{X}}) = \frac{1}{n} \cdot \text{Var}(X) = \frac{1}{n} \cdot \left(\frac{1}{N} \sum_{i=1}^N (X_i - \bar{X})^2 \right) = \frac{1}{n} S_X^2$$

$$\text{where } S_X^2 = \frac{1}{N} \sum_{i=1}^N (X_i - \bar{X})^2$$

Similarly :

$$\text{Var}(\widehat{\bar{Y}}) = \frac{1}{n} \cdot \text{Var}(Y) = \frac{1}{n} \cdot \left(\frac{1}{N} \sum_{i=1}^N (Y_i - \bar{Y})^2 \right) = \frac{1}{n} S_Y^2$$

The covariance between $\widehat{\bar{X}}$ and $\widehat{\bar{Y}}$ is

$$\text{Cov}(\widehat{\bar{X}}, \widehat{\bar{Y}}) = \frac{1}{n} \cdot \text{Cov}(X, Y) = \frac{1}{n} C_{\underline{X}, \underline{Y}}$$

$$\text{Corr}(\widehat{\bar{X}}, \widehat{\bar{Y}}) = \frac{\text{Cov}(\widehat{\bar{X}}, \widehat{\bar{Y}})}{\sqrt{\text{Var}(\widehat{\bar{X}})\text{Var}(\widehat{\bar{Y}})}} = \frac{\frac{1}{n} C_{\underline{X}, \underline{Y}}}{\sqrt{\frac{1}{n} S_X^2 \cdot \frac{1}{n} S_Y^2}} = \frac{C_{\underline{X}, \underline{Y}}}{\sqrt{S_X^2 \cdot S_Y^2}} = \frac{C_{\underline{X}, \underline{Y}}}{S_X \cdot S_Y} = \rho_{XY}$$

Where ρ_{XY} is the population correlation coefficient between X and Y

Q6. In a finite population $U = (U_1, \dots, U_N)$ of N units, let $Y = (Y_1, \dots, Y_N)$ be the study

variable. Suppose that Y_1 is known. A SRSWOR(n) is selected from the remaining units (U_2, \dots, U_N) and Let $\widehat{\bar{Y}}_{-1}$ be the sample mean of this sample. Let $\widehat{\bar{Y}}$ denote the sample mean based on a SRSWOR(n) from the entire population U . Consider the following two estimators of the population total $T = \sum_{i=1}^N Y_i$.

$$\text{i. } \widehat{Y}_1 = Y_1 + (N - 1)\widehat{\bar{Y}}_{-1}$$

$$\text{ii. } \widehat{Y}_2 = N\widehat{\bar{Y}}$$

Are the above two estimators unbiased for estimating the population total T ? Compare the variances of the above two estimators.

In a finite population $U = (U_1, \dots, U_N)$ of N units, let $Y = (Y_1, \dots, Y_N)$ be the study variable. Suppose that Y_1 is known. A SRSWOR(n) is selected from the remaining units (U_2, \dots, U_N) and Let $\widehat{\bar{Y}}_{-1}$ be the sample mean of this sample. Let $\widehat{\bar{Y}}$ denote the sample mean based on a SRSWOR(n) from the entire population U . Consider the following two estimators of

the population total $T = \sum_{i=1}^N Y_i$.

$$\text{i. } \widehat{Y}_1 = Y_1 + (N - 1)\widehat{\bar{Y}}_{-1}$$

$$\text{ii. } \widehat{Y}_2 = N\widehat{\bar{Y}}$$

Are the above two estimators unbiased for estimating the population total T ? Compare the variances of the above two estimators.

\Rightarrow

Estimator 1

$$\widehat{Y}_1 = Y_1 + (N - 1)\widehat{\bar{Y}}_{-1}$$

where $\widehat{\bar{Y}}_{-1}$ is the sample mean of an SRSWOR sample from $\{Y_2, \dots, Y_N\}$

Estimator 2

$$\widehat{Y}_2 = N\widehat{\bar{Y}}$$

where $\widehat{\bar{Y}}$ is the sample mean of an SRSWOR sample of size n from the entire population $\{Y_1, \dots, Y_N\}$

$$\text{i. } \hat{Y}_1 = Y_1 + (N - 1) \hat{\bar{Y}}_{-1}$$

Since Y_1 is known (i.e Constant) and expectation over the sample from $\{Y_2, \dots, Y_N\}$, we have :

$$E[\hat{Y}_1] = Y_1 + (N - 1)E\left[\hat{\bar{Y}}_{-1}\right]$$

Because the sample is drawn using SRSWOR from $\{Y_2, \dots, Y_N\}$, the sample mean is an unbiased estimator of the mean of the remaining $N - 1$ value :

$$E\left[\hat{\bar{Y}}_{-1}\right] = \frac{1}{N - 1} \sum_{i=2}^N Y_i$$

So,

$$E[\hat{Y}_1] = Y_1 + (N - 1) \left(\frac{1}{N - 1} \sum_{i=2}^N Y_i \right) = Y_1 + \sum_{i=2}^N Y_i = \sum_{i=1}^N Y_i = T$$

So, \hat{Y}_1 is unbiased.

$$\text{ii. } \hat{Y}_2 = N \hat{\bar{Y}}$$

This is the usual estimator of population total using the sample mean from a SRSWOR sample of size n from the full population of size N . We know:

$$E\left[\hat{\bar{Y}}\right] = \bar{Y} = \frac{1}{N} \sum_{i=1}^N Y_i \Rightarrow E[N \hat{\bar{Y}}] = T$$

So, \hat{Y}_2 is unbiased.

We now compute and compare: $Var(\hat{Y}_1)$ and $Var(\hat{Y}_2)$

Let's define:

Population variance:

$$S^2 = \frac{1}{N - 1} \sum_{i=1}^N (Y_i - \bar{Y})^2$$

Q7. In a finite population $U = (U_1, \dots, U_N)$ of N units, let $Y = (Y_1, \dots, Y_N)$ be the study variable. Let Y_1 and Y_N be the extreme values such that the value of Y_1 is extremely low and the value of Y_N is extremely high among Y_1, \dots, Y_N . Under the $SRSWOR(n)$ scheme, as an alternative to the sample mean $\hat{\bar{Y}}$, consider the following estimator for the population

mean:

$$\tilde{\bar{Y}} = \begin{cases} \hat{\bar{Y}} + k & \text{if the sample contains } U_1 \text{ but not } U_N \\ \hat{\bar{Y}} - k & \text{if the sample contains } U_N \text{ but not } U_1 \\ \hat{\bar{Y}} & \text{otherwise} \end{cases}$$

where k is a fixed positive constant. Find $E(\tilde{\bar{Y}})$ and $V(\tilde{\bar{Y}})$

In a finite population $U = (U_1, \dots, U_N)$ of N units, let $Y = (Y_1, \dots, Y_N)$ be the study variable. Let Y_1 and Y_N be the extreme values such that the value of Y_1 is extremely low and the value of Y_N is extremely high among Y_1, \dots, Y_N . Under the SRSWOR(n) scheme,

as an alternative to the sample mean $\hat{\bar{Y}}$, consider the following estimator for the population mean:

$$\tilde{\bar{Y}} = \begin{cases} \hat{\bar{Y}} + k & \text{if the sample contains } U_1 \text{ but not } U_N \\ \hat{\bar{Y}} - k & \text{if the sample contains } U_N \text{ but not } U_1 \\ \hat{\bar{Y}} & \text{otherwise} \end{cases}$$

where k is a fixed positive constant. Find $V(\tilde{\bar{Y}})$

Let's define:

A : Sample contains U_1 but not U_N

B : Sample contains U_N but not U_1

C : Sample contains both or neither U_1 nor U_N

Since the sampling is without replacement and the sample size is n :

Number of total samples : $\binom{N}{n}$

Number of samples containing U_1 but not U_N , choose remaining $n - 1$ units from $N - 2$

$$P(A) = \frac{\binom{N-2}{n-1}}{\binom{N}{n}}$$

Similarly,

$$P(B) = \frac{\binom{N-2}{n-1}}{\binom{N}{n}} = P(A)$$

and

$$P(C) = 1 - P(A) - P(B) = 1 - 2 \frac{\binom{N-2}{n-1}}{\binom{N}{n}}$$

$$E\left(\widehat{\bar{Y}}\right) = E\left(\widehat{\bar{Y}} + k\right) + E\left(\widehat{\bar{Y}} - k\right) + E\left(\widehat{\bar{Y}}\right)$$

$$E\left(\widehat{\bar{Y}}\right) = E\left(\widehat{\bar{Y}}\right) + k + E\left(\widehat{\bar{Y}}\right) - k + E\left(\widehat{\bar{Y}}\right)$$

$$E\left(\widehat{\bar{Y}}\right) = E\left(\widehat{\bar{Y}}\right) \frac{\binom{N-2}{n-1}}{\binom{N}{n}} + E\left(\widehat{\bar{Y}}\right) \frac{\binom{N-2}{n-1}}{\binom{N}{n}} + E\left(\widehat{\bar{Y}}\right) \left(1 - 2 \frac{\binom{N-2}{n-1}}{\binom{N}{n}}\right) = E\left(\widehat{\bar{Y}}\right) = \bar{Y}$$

Note : I did not understand

Q8. Consider a SRSWOR(2) from a population $U = (U_1, \dots, U_N)$ with study variable $Y = (Y_1, \dots, Y_N)$. Consider the following estimator of the population mean:

$$\widetilde{\bar{Y}} = \begin{cases} \frac{Y_1 + Y_2}{2} & \text{if the sample contains } U_1 \text{ and } U_2 \\ \frac{1}{2}Y_1 + \frac{2}{3}Y_3 & \text{if the sample contains } U_1 \text{ and } U_3 \\ \frac{Y_2 + Y_3}{2} & \text{if the sample contains } U_2 \text{ and } U_3 \end{cases}$$

Find $E\left(\widetilde{\bar{Y}}\right)$ and $V\left(\widetilde{\bar{Y}}\right)$.

Consider a SRSWOR(2) from a population $U = (U_1, \dots, U_N)$ with study variable $Y = (Y_1, \dots, Y_N)$. Consider the following estimator of the population mean :

$$\widetilde{\bar{Y}} = \begin{cases} \frac{Y_1 + Y_2}{2} & \text{if the sample contains } U_1 \text{ and } U_2 \\ \frac{1}{2}Y_1 + \frac{2}{3}Y_3 & \text{if the sample contains } U_1 \text{ and } U_3 \\ \frac{Y_2 + Y_3}{2} & \text{if the sample contains } U_2 \text{ and } U_3 \end{cases}$$

All Possible Samples :

With $SRSWOR(2)$ from 3 units, the possible samples are :

1. $\{U_1, U_2\}$
2. $\{U_1, U_3\}$
3. $\{U_2, U_3\}$

Since there are $\binom{3}{2} = 3$ Each has an equal probability of $\frac{1}{3}$

Corresponding Estimators

From the problem, the estimator \tilde{Y} for each sample is :

1. $\{U_1, U_2\} : \tilde{\bar{Y}} = \frac{Y_1 + Y_2}{2}$
2. $\{U_1, U_3\} : \tilde{\bar{Y}} = \frac{1}{2}Y_1 + \frac{2}{3}Y_3$
3. $\{U_2, U_3\} : \tilde{\bar{Y}} = \frac{Y_2 + Y_3}{2}$

Compute Expected Value $E(\tilde{Y})$

$$E(\tilde{\bar{Y}}) = \frac{1}{3}\left(\frac{Y_1 + Y_2}{2}\right) + \frac{1}{3}\left(\frac{1}{2}Y_1 + \frac{2}{3}Y_3\right) + \frac{1}{3}\left(\frac{Y_2 + Y_3}{2}\right)$$

$$E(\tilde{\bar{Y}}) = \frac{1}{3}(Y_1 + Y_2 + Y_3)$$

This shows that \tilde{Y} is biased, since the true population mean is :

$$\bar{Y} = \frac{1}{3}(Y_1 + Y_2 + Y_3)$$

Compute Variance $V(\tilde{Y})$

$$V(\tilde{\bar{Y}}) = E\left[\left(\tilde{\bar{Y}}\right)^2\right] - \left(E\left[\tilde{\bar{Y}}\right]\right)^2$$

$$\text{First compute each } E\left[\left(\tilde{\bar{Y}}\right)^2\right] = \frac{1}{3}\left(\frac{Y_1 + Y_2}{2}\right)^2 + \frac{1}{3}\left(\frac{1}{2}Y_1 + \frac{2}{3}Y_3\right)^2 + \frac{1}{3}\left(\frac{Y_2 + Y_3}{2}\right)^2$$

$$\left(\frac{Y_1 + Y_2}{2}\right)^2 = \frac{1}{4}(Y_1^2 + 2Y_1Y_2 + Y_2^2)$$

$$\left(\frac{1}{2}Y_1 + \frac{2}{3}Y_3\right)^2 = \frac{1}{4}Y_1^2 + \frac{2}{3}Y_1Y_3 + Y_3^2$$

$$\left(\frac{Y_2 + Y_3}{2}\right)^2 = \frac{1}{4}(Y_2^2 + 2Y_2Y_3 + Y_3^2)$$

Sampling Designs

	 SRSWR	 SRSWOR
Sampling method	Put item back after each draw	Do not put item back
Possible repeats in sample	Yes	No
Observations are i.i.d.?	Yes	No
Dependence between draws	Independent	Dependent
Distribution of draws	Same for each draw (Uniform over population)	Changes after each draw

SRS = Simple Random Sampling

A sampling method where each unit of the population has an equal chance of being selected.

With Replacement (SRSWR)

Definition: After selecting an item, you put it back into the population before the next draw.

Each draw is independent, because the population size doesn't change.

A unit can be chosen more than once.

Example: Population = {A, B, C}, sample size $n = 2$

Possible samples:

(A, A), (B, B), (C, C), (A, B) (A, C), (B, A), (B, C), (C, A), (C, B)

Total = $N^n = 3^2 = 9$ ordered samples.

Each of the n draws is independent.

On a single draw, $P(i \text{ is chosen}) = \frac{1}{N}$ So, $P(i \text{ is not chosen in one draw}) = 1 - \frac{1}{N}$

$P(i \text{ is never chosen}) = \left(1 - \frac{1}{N}\right)^n$ So,

$$\pi_i = 1 - \left(1 - \frac{1}{N}\right)^n \text{ (inclusion probability)}$$

$$\pi_{ij} = 1 - 2\left(1 - \frac{1}{N}\right)^n + \left(1 - \frac{2}{N}\right)^n$$

In SRSWR, every unit has the same inclusion probability, but it depends on n

Without Replacement (SRSWOR)

Definition: After selecting an item, you do not put it back.

Each draw is dependent, because the population shrinks.

A unit can be chosen at most once.

Example: Population = $\{A, B, C\}$, sample size $n = 2$

Possible samples: (A, B) , (A, C) , (B, A) , (B, C) , (C, A) , (C, B)

$$\text{Total} = \frac{N!}{(N-n)!} = \frac{3!}{(3-2)!} = 6 \text{ ordered samples}$$

We select exactly n distinct units out of N each subset equally likely.

Probability that unit i is chosen = proportion of subsets that include i

$$\text{Number of subsets of size } n : \binom{N}{n}$$

$$\text{Number of subsets of size } n \text{ that include unit } i : \binom{N-1}{n-1}$$

$$\pi_i = \frac{\binom{N-1}{n-1}}{\binom{N}{n}} = \frac{n}{N} \text{ (inclusion probability)}$$

$$\pi_{ij} = \frac{\binom{N-2}{n-2}}{\binom{N}{n}} = \frac{\frac{(N-2)!}{(n-2)!(N-n)!}}{\frac{N!}{n!(N-n)!}} = \frac{n(n-1)}{N(N-1)}, i \neq j$$

In SRSWOR, every unit also has the same inclusion probability, but it is exactly proportional to sample size.

	SRSWR	SRSWOR
Replacement	Allowed	Not allowed
Sample size effect	Population size stays same	Population shrinks

Independence	Each draw independent	Draws are dependent
Probability of selecting a unit in one draw	$\frac{1}{N}$	Changes with each draw
Total possible ordered samples (size n)	N^n	$\frac{N!}{(N-n)!}$

Q1. Consider a sampling design for sampling from a population $U = (U_1, U_2, U_3)$ of three units, with study variables $Y = (Y_1, Y_2, Y_3)$ and

$$P(\underline{s}) \begin{cases} \frac{1}{7} & \text{if } \underline{s} = (U_1, U_3) \\ \frac{2}{7} & \text{if } \underline{s} = (U_2, U_3) \\ \frac{4}{7} & \text{if } \underline{s} = (U_1, U_2, U_3) \end{cases}$$

Find

- (i) the first order inclusion probabilities π_1, π_2 and π_3 .
- (ii) the second order inclusion probabilities π_{12}, π_{13} and π_{23}
- (iii) the expected value and the variance of the estimator $\hat{\eta}_{\underline{S}} = \frac{3}{n(\underline{S})} \sum_{i \in \underline{S}} Y_i$ of the population total.

Q2. Repeat Problem 1 with the sampling design described by

$$Q(\underline{s}) \begin{cases} k, & \text{if } n(\underline{s}) = r(\underline{s}) = 2 \\ 0, & \text{otherwise} \end{cases}$$

where k is a fixed positive constant.

Q3. Let $A_i, i = 1, \dots, N$, denote the number of times the i^{th} unit U_i appears in the sample drawn from the population $U = (U_1, \dots, U_N)$ under the $SRSWOR(n)$ design. Show that

$Cov(A_i, A_j) = -\frac{n(N-n)}{N^2(N-1)}$, $i \neq j$. Also find $Cov(A_i, A_j)$, $i \neq j$, under the $SRSWR(s)$ design

Q4. Let $Y = ((X_1, Y_1), \dots, (X_N, Y_N))$, $\bar{X} = \frac{1}{N} \sum_{i=1}^N X_i$, $\bar{Y} = \frac{1}{N} \sum_{i=1}^N Y_i$ and

$C_{\underline{X}, \underline{Y}} = \frac{1}{N} \sum_{i=1}^N (X_i - \bar{X})(Y_i - \bar{Y})$. Under the $SRSWR(n)$ design, let $\hat{\bar{X}} = \frac{1}{n} \sum_{i \in S} X_i$,

$\hat{\bar{Y}} = \frac{1}{n} \sum_{i \in S} Y_i$ and $\hat{C}_{\underline{X}, \underline{Y}} = \frac{1}{n-1} \sum_{i \in S} (X_i - \hat{\bar{X}})(Y_i - \hat{\bar{Y}})$. Find

i. the correlation coefficient between $\hat{\bar{X}}$ and $\hat{\bar{Y}}$

ii. $E(\hat{C}_{\underline{X}, \underline{Y}})$

Q5. For sampling from a population $U = (U_1, \dots, U_N)$ with the study variable

$Y = (Y_1, \dots, Y_N)$,

consider a sampling scheme under which $SRSWR$ is continued until the sample contains d (a fixed positive integer) distinct units. Let M denote the number of selections made (i.e. M is

the sample size, that is random) and, for $r = 1, \dots, N$, $K_r \left(\sum_{r=1}^N K_r = M \right)$ denote the

frequency of the appearance of the r^{th} distinct unit in the sample. Define $\hat{\bar{Y}}_1 = \frac{1}{M} \sum_{r \in S} K_r Y_r$

and $\hat{\bar{Y}}_2 = \frac{1}{d} \sum_{r \in S} Y_r$. Show that $V(\hat{\bar{Y}}_1) \geq \sigma^2 E\left(\frac{1}{M}\right)$ and find $V(\hat{\bar{Y}}_2)$.

Q6. In a finite population $U = (U_1, \dots, U_N)$ of N units, let $Y = (Y_1, \dots, Y_N)$ be the study variable. Suppose that Y_1 is known. A $SRSWOR(n)$ is selected from the remaining units

U_2, \dots, U_N and let $\hat{\bar{Y}}_{-1}$ be the sample mean of this sample. Let $\hat{\bar{Y}}$ denote the sample mean based on a $SRSWOR(n)$ from the entire population U . Consider the following two

estimators of the population total $T = \sum_{i=1}^N Y_i$.

$$(i) \hat{Y}_1 = Y_1 + (N-1)\hat{\bar{Y}}_{-1}$$

$$(ii) \hat{Y}_2 = N\hat{\bar{Y}}$$

Q7. In a finite population $U = (U_1, \dots, U_N)$ of N units, let $Y = (Y_1, \dots, Y_N)$ be the study

variable. Let Y_1 and Y_N be the extreme values such that the value of Y_1 is extremely low and the value of Y_N is extremely high among Y_1, \dots, Y_N . Under the $SRSWOR(n)$ scheme, as

an alternative to the sample mean $\hat{\bar{Y}}$, consider the following estimator for the population mean:

$$\tilde{\bar{Y}} = \begin{cases} \hat{\bar{Y}} + k, & \text{if the sample contains } U_1 \text{ but not } U_N \\ \hat{\bar{Y}} - k, & \text{if the sample contains } U_N \text{ but not } U_1 \\ \hat{\bar{Y}}, & \text{otherwise} \end{cases}$$

where k is a fixed positive constant. Find $E(\tilde{\bar{Y}})$ and $V(\tilde{\bar{Y}})$

Q8.

1. Consider a sampling design for sampling from a population $U = (U_1, U_2, U_3)$ of three units, with study variables $Y = (Y_1, Y_2, Y_3)$ and

$$P(\underline{s}) = \begin{cases} \frac{1}{7} & \text{if } \underline{s} = (U_1, U_3) \\ \frac{2}{7} & \text{if } \underline{s} = (U_2, U_3) \\ \frac{4}{7} & \text{if } \underline{s} = (U_1, U_2, U_3) \end{cases}$$

Find

- (i) the first order inclusion probabilities π_1, π_2 and π_3 .
- (ii) the second order inclusion probabilities π_{12}, π_{13} and π_{23}

(iii) the expected value and the variance of the estimator $\hat{\eta}_{\underline{S}} = \frac{3}{n(\underline{S})} \sum_{i \in \underline{S}} Y_i$ of the population total.

\Rightarrow

first order inclusion probabilities : - $\pi_1 = \frac{5}{7}, \pi_2 = \frac{6}{7}, \pi_3 = 1$

second order inclusion probabilities : - $\pi_{12} = \frac{4}{7}, \pi_{13} = \frac{5}{7}, \pi_{23} = \frac{6}{7}$

Expected value

$$\mathbb{E}[\hat{\eta}] = \frac{3}{n(\underline{S})} \sum_{i \in \underline{S}} Y_i = \left(\frac{3}{2}(Y_1 + Y_3) \frac{1}{7} \right) + \left(\frac{3}{2}(Y_2 + Y_3) \frac{2}{7} \right) + \left(\frac{3}{3}(Y_1 + Y_2 + Y_3) \frac{4}{7} \right)$$

Here $n(\underline{S})$ are number of units. $Y_i = y_i p_i$ where y_i is possible outcomes p_i are corresponding probabilities.

$$\mathbb{E}[\hat{\eta}] = \frac{11}{14}Y_1 + Y_2 + \frac{17}{14}Y_3$$

Variance

$$Var(\hat{\eta}) = \mathbb{E}[\hat{\eta}^2] - (\mathbb{E}[\hat{\eta}])^2$$

$$Var(\hat{\eta}) = \left(\frac{1}{7} \left(\frac{3}{2}(Y_1 + Y_3) \right)^2 + \frac{2}{7} \left(\frac{3}{2}(Y_2 + Y_3) \right)^2 + \frac{4}{7} \left(\frac{3}{3}(Y_1 + Y_2 + Y_3) \right)^2 \right) - \left(\frac{11}{14}Y_1 + Y_2 + \frac{17}{14}Y_3 \right)^2$$

2. Repeat Problem 1 with the sampling design described by

$$Q(\underline{s}) \begin{cases} k, & \text{if } n(\underline{s}) = r(\underline{s}) = 2 \\ 0, & \text{otherwise} \end{cases}$$

where k is a fixed positive constant.

first order inclusion probabilities : - $\pi_1 = \pi_2 = \pi_3 = \frac{2}{3}$

second order inclusion probabilities : - $\pi_{12} = \pi_{13} = \pi_{23} = \frac{2}{3}$

Expected value

$$\mathbb{E}[\hat{\eta}] = \frac{3}{n(\underline{S})} \sum_{i \in \underline{S}} Y_i = \left(\frac{3}{2}(Y_1 + Y_2) \frac{1}{3} \right) + \left(\frac{3}{2}(Y_2 + Y_3) \frac{1}{3} \right) + \left(\frac{3}{2}(Y_1 + Y_3) \frac{1}{3} \right) = Y_1 + Y_2 + Y_3$$

Variance

$$Var(\hat{\eta}) = \mathbb{E}[\hat{\eta}^2] - (\mathbb{E}[\hat{\eta}])^2$$

$$= \left(\left(\frac{3}{2}(Y_1 + Y_2) \right)^2 \cdot \frac{1}{3} + \left(\frac{3}{2}(Y_2 + Y_3) \right)^2 \cdot \frac{1}{3} + \left(\frac{3}{2}(Y_1 + Y_3) \right)^2 \cdot \frac{1}{3} \right) - (Y_1 + Y_2 + Y_3)^2$$

3. Let A_i , $i = 1, \dots, N$, denote the number of times the i^{th} unit U_i appears in the sample drawn from the population $U = (U_1, \dots, U_N)$ under the $SRSWOR(n)$ design. Show that

$$Cov(A_i, A_j) = -\frac{n(N-n)}{N^2(N-1)}, i \neq j. \text{ Also find } Cov(A_i, A_j), i \neq j, \text{ under the}$$

SRSWR(s) design

Let A_i , $i = 1, \dots, N$, denote the number of times the i^{th} unit U_i appears in the sample drawn from the population $U = (U_1, \dots, U_N)$ under the SRSWOR(n) design.

Show that $\text{Cov}(A_i, A_j) = -\frac{n(N-n)}{N^2(N-1)}$, $i \neq j$. Also find $\text{Cov}(A_i, A_j)$, $i \neq j$, under the SRSWOR(n) design.

SRSWOR (without replacement)

When sampling without replacement each unit can appear at most once, so $A \in \{0, 1\}$

Therefore, $A_i = 1$ if unit i is selected, and 0 otherwise

$$A_i = \begin{cases} 1, & \text{if unit } i \text{ is selected} \\ 0, & \text{otherwise} \end{cases}$$

$$\text{Cov}(A_i, A_j) = \mathbb{E}[A_i A_j] - \mathbb{E}[A_i] \mathbb{E}[A_j]$$

$$\mathbb{E}[A_i] = P(A_i = 1) = \frac{\binom{N-1}{n-1}}{\binom{N}{n}} = \frac{n}{N},$$

$$\mathbb{E}[A_i, A_j] = P(A_i = 1, A_j = 1) = \frac{\binom{N-2}{n-2}}{\binom{N}{n}} = \frac{\frac{(N-2)!}{(n-2)!(N-n)!}}{\frac{N!}{n!(N-n)!}} = \frac{n(n-1)}{N(N-1)}$$

(for $i \neq j$). Compute the joint probability in closed form:

Thus

$$\mathbb{E}[A_i] \mathbb{E}[A_j] = P(A_i = 1, A_j = 1) = \frac{n^2}{N^2}$$

$$\mathbb{E}[A_i A_j] = P(A_i = 1, A_j = 1) = \frac{n(n-1)}{N(N-1)}$$

So

$$\text{Cov}(A_i, A_j) = \frac{n(n-1)}{N(N-1)} - \frac{n^2}{N^2} = \frac{nN(n-1) - n^2(N-1)}{N^2(N-1)} = \frac{n(N(n-1) - n(N-1))}{N^2(N-1)}$$

Solve further you'll get this

$$\text{Cov}(A_i, A_j) = -\frac{n(N-n)}{N^2(N-1)}$$

SRSWR (with replacement)

Later

4. Let $Y = ((X_1, Y_1), \dots, (X_N, Y_N))$, $\bar{X} = \frac{1}{N} \sum_{i=1}^N X_i$, $\bar{Y} = \frac{1}{N} \sum_{i=1}^N Y_i$ and

$C_{\underline{X}, \underline{Y}} = \frac{1}{N} \sum_{i=1}^N (X_i - \bar{X})(Y_i - \bar{Y})$. Under the SRSWR(n) design, let $\widehat{\bar{X}} = \frac{1}{n} \sum_{i \in S} X_i$,

$\widehat{\bar{Y}} = \frac{1}{n} \sum_{i \in S} Y_i$ and $\widehat{C}_{\underline{X}, \underline{Y}} = \frac{1}{n-1} \sum_{i \in S} (X_i - \widehat{\bar{X}})(Y_i - \widehat{\bar{Y}})$. Find

i. the correlation coefficient between $\widehat{\bar{X}}$ and $\widehat{\bar{Y}}$

ii. $E(\widehat{C}_{\underline{X}, \underline{Y}})$

\Rightarrow

Let $Y = ((X_1, Y_1), \dots, (X_N, Y_N))$, $\bar{X} = \frac{1}{N} \sum_{i=1}^N X_i$, $\bar{Y} = \frac{1}{N} \sum_{i=1}^N Y_i$ and $C_{\underline{X}, \underline{Y}} = \frac{1}{N} \sum_{i=1}^N (X_i - \bar{X})(Y_i - \bar{Y})$

i. the correlation coefficient between $\widehat{\bar{X}}$ and $\widehat{\bar{Y}}$

ii. $E(\widehat{C}_{\underline{X}, \underline{Y}})$

i. the correlation coefficient between $\widehat{\bar{X}}$ and $\widehat{\bar{Y}}$

[Pearson correlation coefficient - Wikipedia](#)

$$\rho_{X,Y} = \frac{Cov(X, Y)}{\sigma_X \cdot \sigma_Y}$$

$$\rho(\widehat{\bar{X}}, \widehat{\bar{Y}}) = Corr(\widehat{\bar{X}}, \widehat{\bar{Y}}) = \frac{Cov(\widehat{\bar{X}}, \widehat{\bar{Y}})}{\sqrt{Var(\widehat{\bar{X}})Var(\widehat{\bar{Y}})}} = \rho_{\underline{X}, \underline{Y}}$$

Find values and plug into formula and calculate

ii. $E(\widehat{C}_{\underline{X}, \underline{Y}})$

$$E(\widehat{C}_{\underline{X}, \underline{Y}}) = C_{\underline{X}, \underline{Y}}$$

Q5. For sampling from a population $U = (U_1, \dots, U_N)$ with the study variable

$Y = (Y_1, \dots, Y_N)$,

consider a sampling scheme under which SRSWR is continued until the sample contains d

(a fixed positive integer) distinct units. Let M denote the number of selections made (i.e. M is the sample size, that is random) and, for $r = 1, \dots, N$, $K_r \left(\sum_{r=1}^N K_r = M \right)$ denote the frequency of the appearance of the r^{th} distinct unit in the sample. Define $\widehat{Y}_1 = \frac{1}{M} \sum_{r \in S} K_r Y_r$ and $\widehat{Y}_2 = \frac{1}{d} \sum_{r \in S} Y_r$. Show that $V(\widehat{Y}_1) \geq \sigma^2 E\left(\frac{1}{M}\right)$ and find $V(\widehat{Y}_2)$.

In Process

Q6. In a finite population $U = (U_1, \dots, U_N)$ of N units, let $Y = (Y_1, \dots, Y_N)$ be the study variable. Suppose that Y_1 is known. A $SRSWOR(n)$ is selected from the remaining units U_2, \dots, U_N and let \widehat{Y}_{-1} be the sample mean of this sample. Let \widehat{Y} denote the sample mean based on a $SRSWOR(n)$ from the entire population U . Consider the following two

estimators of the population total $T = \sum_{i=1}^N Y_i$.

$$(i) \widehat{Y}_1 = Y_1 + (N - 1)\widehat{Y}_{-1}$$

$$(ii) \widehat{Y}_2 = N\widehat{Y}$$

Are the above two estimators unbiased for estimating the population total T ? compare the variances of the above two estimators.

Pending

In a finite population $U = (U_1, \dots, U_N)$ of N units, let $Y = (Y_1, \dots, Y_N)$ be the study variable

$$(i) \widehat{Y}_1 = Y_1 + (N - 1)\widehat{Y}_{-1}$$

$$(ii) \widehat{Y}_2 = N\widehat{Y}$$

Are the above two estimators unbiased for estimating the population total T ? compare the var

Q7. In a finite population $U = (U_1, \dots, U_N)$ of N units, let $Y = (Y_1, \dots, Y_N)$ be the study variable. Let Y_1 and Y_N be the extreme values such that the value of Y_1 is extremely low and the value of Y_N is extremely high among Y_1, \dots, Y_N . Under the $SRSWOR(n)$ scheme, as an alternative to the sample mean \widehat{Y} , consider the following estimator for the population mean:

$$\tilde{\bar{Y}} = \begin{cases} \hat{\bar{Y}} + k, & \text{if the sample contains } U_1 \text{ but not } U_N \\ \hat{\bar{Y}} - k, & \text{if the sample contains } U_N \text{ but not } U_1 \\ \hat{\bar{Y}}, & \text{otherwise} \end{cases}$$

where k is a fixed positive constant. Find $E(\tilde{\bar{Y}})$ and $V(\tilde{\bar{Y}})$

In a finite population $U = (U_1, \dots, U_N)$ of N units, let $Y = (Y_1, \dots, Y_N)$ be the study variable. Let Y_1 and Y_N be the extreme values such that the value of Y_1 is extremely low and the value of Y_N is extremely high among Y_1, \dots, Y_N . Under the SRSWOR(n) scheme, as an alternative to the sample mean $\hat{\bar{Y}}$, consider the following estimator for the population mean :

$$\tilde{\bar{Y}} = \begin{cases} \hat{\bar{Y}} + k, & \text{if the sample contains } U_1 \text{ but not } U_N \\ \hat{\bar{Y}} - k, & \text{if the sample contains } U_N \text{ but not } U_1 \\ \hat{\bar{Y}}, & \text{otherwise} \end{cases}$$

where k is a fixed positive constant

find $E(\tilde{\bar{Y}})$ and $V(\tilde{\bar{Y}})$

Intuition

Since Y_1 is extremely low, if the sample includes U_1 but misses the U_N , sample mean $\hat{\bar{Y}}$ will likely underestimate the population mean.

→ So we add k to correct upward.

Since Y_N is extremely high, if the sample includes U_N but misses U_1 , the sample mean $\hat{\bar{Y}}$ will likely overestimate the population mean.

→ So we add k to correct downward.

If both extremes are present (or both absent), the sample mean is more “balanced,” so no correction is needed.

$$E(\tilde{\bar{Y}}) = (k \cdot P(U_1 \in S, U_N \notin S)) + (-k \cdot P(U_1 \notin S, U_N \in S)) + E(\hat{\bar{Y}})$$

$$P(U_1 \in S, U_N \notin S) = \frac{\binom{N-2}{n-2}}{\binom{N}{n}} = \frac{\frac{(N-2)!}{(n-2)!(N-n)!}}{\frac{N!}{n!(N-n)!}} = \frac{n(n-1)}{N(N-1)}$$

Sane for $P(U_1 \notin S, U_N \in S)$ and since second term it negative so we'll left with only

$$E\left(\widehat{\bar{Y}}\right) = E\left(\widehat{\bar{Y}}\right) = \bar{Y}$$

Pending

Horvitz-Thompson Estimator, SRSWR(n)and SRSWOR(n)

Horvitz-Thompson Estimator for Total

The Horvitz-Thompson (HT) estimator for the population total is:

$$\hat{Y}_{HT} = \sum_{i \in s} \frac{Y_i}{\pi_i}$$

Where π_i is the inclusion probability of unit i , i.e., the probability that unit i is included in the sample.

Variance of HT Estimator

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

Q1. Let $Y = ((X_1, Y_1), \dots, (X_N, Y_N))$, $\bar{X} = \frac{1}{N} \sum_{i=1}^N X_i$, $\bar{Y} = \frac{1}{N} \sum_{i=1}^N Y_i$ and

$C_{\underline{X}, \underline{Y}} = \frac{1}{N} \sum_{i=1}^N (X_i - \bar{X})(Y_i - \bar{Y})$. Under the $SRSWR(n)$ design, find

(i). the correlation coefficient between \hat{X}_{HT} and \hat{Y}_{HT} , the Horvitz-Thompson estimators of $X = N\bar{X}$ and $Y = \bar{NY}$.

(ii). an unbiased estimator of $C_{\underline{X}, \underline{Y}}$.

Medium

Q2. Let $U = (B_1, B_2, \dots, B_6)$ and $Y = (Y_1, Y_2, \dots, Y_6) = (11, 21, 16, 6, 7, 11)$, where, for $i = 1, 2, \dots, 6$, B_i is the i^{th} bank in a country and Y_i is the number non-performing loans given by i^{th} bank. Consider the sampling design defind by

$S' = \{\underline{s} = (s_1, \dots, s_{n(\underline{s})}) : s_i \in \{B_1, \dots, B_6\}, i = 1, \dots, n(\underline{s}), s_1 \neq s_2 \neq \dots \neq s_{n(\underline{s})}\}$ and

$$P(\underline{s}) = \begin{cases} \frac{1}{60}, & \text{if } n(\underline{s}) = 2 \\ \frac{1}{240}, & \text{if } n(\underline{s}) = 3 \\ 0, & \text{otherwise} \end{cases}$$

Under the above design, suppose that the selected sample is (B_2, B_5, B_6) . Using Horvitz-Thompson estimator and the observed sample, estimate the total number of non-performing accounts in the population. Compute an unbiased estimate of the population variance.

Easy

Q3. Do the problem 2, Under the design

$$S' = \{\underline{s} = (s_1, \dots, s_{n(\underline{s})}) : s_i \in \{B_1, \dots, B_6\}, i = 1, \dots, n(\underline{s})\}$$

$$P(\underline{s}) = \begin{cases} \frac{1}{72}, & \text{if } n(\underline{s}) = 2 \\ \frac{1}{432}, & \text{if } n(\underline{s}) = 3 \\ 0, & \text{otherwise} \end{cases}$$

and if the observed sample is (B_2, B_3, B_2) .

Easy

Q4. Let $U = (U_1, \dots, U_N)$ and $Y = (Y_1, \dots, Y_N)$. A SRSWOR sample of size n is drawn from U and subsequently a SRSWOR subsample of size n^* is drawn from this sample. Let

$\widehat{Y}^{(2)}$ denote the sample mean based on the combined sample of size $n + n^*$ and \widehat{Y} is the sample mean based on all n units in original sample. Show that

(i) Show that $\widehat{Y}^{(2)}$ is an unbiased estimator of the population mean.

(ii) Let V_1 and V_2 respectively, be the variances of \widehat{Y} and $\widehat{Y}^{(2)}$. Show that

$$\frac{V_2}{V_1} \approx \frac{1 + 3\frac{n^*}{n}}{\left(1 + \frac{n^*}{n}\right)^2}$$

Medium

Q5. Let $Y = ((X_1, Y_1), \dots, (X_N, Y_N))$, $\bar{X} = \frac{1}{N} \sum_{i=1}^N X_i$, $\bar{Y} = \frac{1}{N} \sum_{i=1}^N Y_i$ and

$C_{\underline{X}, \underline{Y}} = \frac{1}{N} \sum_{i=1}^N (X_i - \bar{X})(Y_i - \bar{Y})$. Under the SRSWR(n) design, let \widehat{X}^* and \widehat{Y}^* denote

the sample means of X and Y variables, respectively, based on whole sample. Let

$$\widehat{C}_{\underline{X}, \underline{Y}} = \frac{1}{n-1} \sum_{i=1}^n \left(X_i - \widehat{\bar{X}}^* \right) \left(Y_i - \widehat{\bar{Y}}^* \right)$$

is the i^{th} sample unit. Find

(i) the correlation coefficient between $\widehat{\bar{X}}^*$ and $\widehat{\bar{Y}}^*$

(ii) $E(\widehat{C}_{\underline{X}, \underline{Y}})$

Medium

Q6. Let $U = (U_1, \dots, U_N)$ and $Y = (Y_1, \dots, Y_N)$. A SRSWR suppose that the coefficient of variance $C = \frac{S}{Y}$ is known. Find an estimator of the type $\widehat{\bar{Y}}^k$ and $k\widehat{\bar{Y}}^*$ for the population

mean that has the smaller mean squared error than $\widehat{\bar{Y}}^*$; here $\widehat{\bar{Y}}^*$ denotes the sample mean based on full sample. Find the relative efficiency (ratio of mean squared errors) of this

estimator relative to $\widehat{\bar{Y}}^*$

Medium

Q7. Let $U = (U_1, \dots, U_N)$ and $Y = (Y_1, \dots, Y_N)$. Suppose that we want to estimate the population proportion P of units having a given attribute.

(i) Find unbiased estimators of P under the SRSWR(n) and SRSWOR(n).

(ii) Compute variances of estimators derived in (i)

(iii) Construct a 95% confidence interval for P under SRSWOR(n) and SRSWOR(n).

Easy

Q8. Let $U = (U_1, \dots, U_N)$ and $Y = (Y_1, \dots, Y_N)$. Suppose that it is desired to estimate the population proportion P of units having rare attribute. Consider a SRSWOR scheme that is continued until m units possessing the rare attribute have been found. Let M be the total sample size of the sample. If fpc is ignored, show that

$$Pr(M = n) = \binom{n-1}{m-1} P^m (1-P)^{n-m}, \quad n = m, m+1$$

Find $E[M]$ and show that $\frac{m-1}{n-1}$ is an unbiased estimator of P .

Hard

Q9. Under SRSWR(n) note that sample mean based on the whole sample is given by

$$\widehat{\bar{Y}}^* = \frac{1}{n} \sum_{i=1}^n A_i Y_i$$

where A_i = number of times i^{th} unit appears in the sample. Using the above representation. Find the mean and the variance of \bar{Y} .

Easy

Q10. Consider $U = (U_1, U_2, U_3, U_4)$ and $Y = (Y_1, Y_2, Y_3, Y_4) = (1, 6, 6, 11)$.

- Find the probability distribution of the sample mean under $SRSWR(2)$. Find its mean and variance.
- Repeat (i) under $SRSWOR(2)$.
- Find the probability distribution of Horvitz-Thompson Estimator under $SRSWR(2)$. Find its mean and variance.
- Compare performances of different estimators described above.

Easy (i, ii)

Medium (iii, iv)

Q11. Under $SRSWR(n)$, let D denote the number of distinct unit in the sample. Let

$\hat{\bar{Y}}^*$ denote the sample mean based on the whole sample and $\hat{\bar{Y}}^{**}$ be the sample mean based on distinct units.

(i). Find the probability mass function of D .

(ii). Find $E(D)$ and $Var(D)$

(iii). Show that $E\left(\frac{1}{D}\right) \geq \frac{1}{N\left(1 - \left(1 - \frac{1}{n}\right)^n\right)}$.

(iv). Show that $\hat{\bar{Y}}^*$ and $\hat{\bar{Y}}^{**}$ are unbiased estimator of \bar{Y} . Compare their precisions.

Medium

Solutions

Q1. Let $Y = ((X_1, Y_1), \dots, (X_N, Y_N))$, $\bar{X} = \frac{1}{N} \sum_{i=1}^N X_i$, $\bar{Y} = \frac{1}{N} \sum_{i=1}^N Y_i$ and

$C_{\underline{X}, \underline{Y}} = \frac{1}{N} \sum_{i=1}^N (X_i - \bar{X})(Y_i - \bar{Y})$. Under the $SRSWR(n)$ design, find

- the correlation coefficient between \hat{X}_{HT} and \hat{Y}_{HT} , the Horvitz-Thompson estimators of $X = N\bar{X}$ and $Y = \bar{N}Y$.
- an unbiased estimator of $C_{\underline{X}, \underline{Y}}$.

⇒ **Q4. (Assignment-1)**

$$\rho_{X,Y} = \frac{Cov(X, Y)}{\sigma_X \cdot \sigma_Y}$$

$$\rho\left(\widehat{\bar{X}}_{HT}, \widehat{\bar{Y}}_{HT}\right) = Corr\left(\widehat{\bar{X}}_{HT}, \widehat{\bar{Y}}_{HT}\right) = \frac{Cov\left(\widehat{\bar{X}}_{HT}, \widehat{\bar{Y}}_{HT}\right)}{\sqrt{Var\left(\widehat{\bar{X}}_{HT}\right)Var\left(\widehat{\bar{Y}}_{HT}\right)}} = \rho_{x,y}$$

Q2. Let $U = (B_1, B_2, \dots, B_6)$ and $Y = (Y_1, Y_2, \dots, Y_6) = (11, 21, 16, 6, 7, 11)$, where, for $i = 1, 2, \dots, 6$, B_i is the i^{th} bank in a country and Y_i is the number non-performing loans given by i^{th} bank. Consider the sampling design defined by $S' = \{\underline{s} = (s_1, \dots, s_{n(\underline{s})}) : s_i \in \{B_1, \dots, B_6\}, i = 1, \dots, n(\underline{s}), s_1 \neq s_2 \neq \dots \neq s_{n(\underline{s})}\}$ and

$$P(\underline{s}) = \begin{cases} \frac{1}{60}, & \text{if } n(\underline{s}) = 2 \\ \frac{1}{240}, & \text{if } n(\underline{s}) = 3 \\ 0, & \text{otherwise} \end{cases}$$

Under the above design, suppose that the selected sample is (B_2, B_5, B_6) . Using Horvitz-Thompson estimator and the observed sample, estimate the total number of non-performing accounts in the population. Compute an unbiased estimate of the population variance.

[⇒ Follow Prof's Solution \(GPT too\)](#)

Horvitz-Thompson Estimator for Total

$$\widehat{Y}_{HT} = \sum_{i \in s} \frac{Y_i}{\pi_i}$$

Variance of HT Estimator

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

Inclusion probs : $\pi_i = \frac{5}{12}, \pi_{ij} = \frac{2}{15}$ for $i \neq j$

$$\widehat{Y}_{HT} = \sum_{i \in s} \frac{Y_i}{\pi_i} = 93.6$$

Put those values in formula and calculate

I'm writing some solution because I found myself difficulty use that n/N formula for π_i but replace $E[n] = n$. Solve further if stuck, ask GPT.

Q3. Do the problem 2, Under the design

$$S' = \{\underline{s} = (s_1, \dots, s_{n(\underline{s})}) : s_i \in \{B_1, \dots, B_6\}, i = 1, \dots, n(\underline{s})\}$$

$$P(\underline{s}) = \begin{cases} \frac{1}{72}, & \text{if } n(\underline{s}) = 2 \\ \frac{1}{432}, & \text{if } n(\underline{s}) = 3 \\ 0, & \text{otherwise} \end{cases}$$

and if the observed sample is (B_2, B_3, B_2) .

⇒ Similar to Q2 but its SRSWR

Q4. Let $U = (U_1, \dots, U_N)$ and $Y = (Y_1, \dots, Y_N)$. A SRSWOR sample of size n is drawn from U and subsequently a SRSWOR subsample of size n^* is drawn from this sample. Let $\widehat{Y}^{(2)}$ denote the sample mean based on the combined sample of size $n + n^*$ and \widehat{Y} is the sample mean based on all n units in original sample. Show that

(i) Show that $\widehat{Y}^{(2)}$ is an unbiased estimator of the population mean.

(ii) Let V_1 and V_2 respectively, be the variances of \widehat{Y} and $\widehat{Y}^{(2)}$. Show that

$$\frac{V_2}{V_1} \approx \frac{1 + 3\frac{n^*}{n}}{\left(1 + \frac{n^*}{n}\right)^2}$$

(i) use the [Law of total expectation - Wikipedia](#)

(ii) use the [Law of total variance - Wikipedia](#)

Q5. Let $Y = ((X_1, Y_1), \dots, (X_N, Y_N))$, $\bar{X} = \frac{1}{N} \sum_{i=1}^N X_i$, $\bar{Y} = \frac{1}{N} \sum_{i=1}^N Y_i$ and

$C_{\underline{X}, \underline{Y}} = \frac{1}{N} \sum_{i=1}^N (X_i - \bar{X})(Y_i - \bar{Y})$. Under the SRSWR(n) design, let \widehat{X}^* and \widehat{Y}^* denote

the sample means of X and Y variables, respectively, based on whole sample. Let

$\widehat{C}_{\underline{X}, \underline{Y}} = \frac{1}{n-1} \sum_{i=1}^n \left(X_i - \widehat{X}^* \right) \left(Y_i - \widehat{Y}^* \right)$ is the i^{th} sample unit. Find

(i) the correlation coefficient between \widehat{X}^* and \widehat{Y}^*

(ii) $E(\hat{C}_{\underline{X}, \underline{Y}})$

\Rightarrow **Q4. (Assignment-1)**

Q6. Let $U = (U_1, \dots, U_N)$ and $Y = (Y_1, \dots, Y_N)$. A SRSWR suppose that the coefficient of variance $C = \frac{S}{Y}$ is known. Find an estimator of the type $\hat{\bar{Y}}^k$ and $k\hat{\bar{Y}}^*$ for the population

mean that has the smaller mean squared error than $\hat{\bar{Y}}^*$; here $\hat{\bar{Y}}^*$ denotes the sample mean based on full sample. Find the relative efficiency (ratio of mean squared errors) of this

estimator relative to $\hat{\bar{Y}}^*$

\Rightarrow

$$\text{MSE}\left(\hat{\bar{Y}}_k\right) = E\left[\left(k\hat{\bar{Y}}^* - \bar{Y}\right)^2\right] = k^2 E\left[\left(k\hat{\bar{Y}}^*\right)^2\right] - 2kE\left[\hat{\bar{Y}}^*\right] + \bar{Y}^2$$

Q7. Let $U = (U_1, \dots, U_N)$ and $Y = (Y_1, \dots, Y_N)$. Suppose that we want to estimate the population proportion P of units having a given attribute.

(i) Find unbiased estimators of P under the $SRSWR(n)$ and $SRSWOR(n)$.

(ii) Compute variances of estimators derived in (i)

(iii) Construct a 95% confidence interval for P under $SRSWOR(n)$ and $SRSWOR(n)$

i. $E[\hat{p}] = P$

$$V_{WR}(\hat{p}) = \frac{1}{n}P(1-P), \quad V_{WOR}(\hat{p}) = \frac{1}{n} \frac{N-n}{N-1}P(1-P)$$

Q8. Let $U = (U_1, \dots, U_N)$ and $Y = (Y_1, \dots, Y_N)$. Suppose that it is desired to estimate the population proportion P of units having rare attribute. Consider a $SRSWOR$ scheme that is continued until m units possessing the rare attribute have been found. Let M be the total sample size of the sample. If fpc is ignored, show that

$$Pr(M = n) = \binom{n-1}{m-1} P^m (1-P)^{n-m}, \quad n = m, m+1$$

Find $E[M]$ and show that $\frac{m-1}{n-1}$ is an unbiased estimator of P .

Q9. Under $SRSWR(n)$ note that sample mean based on the whole sample is given by

$$\hat{\bar{Y}}^* = \frac{1}{n} \sum_{i=1}^n A_i Y_i$$

where A_i = number of times i^{th} unit appears in the sample. Using the above representation. Find the mean and the variance of \bar{Y} .

Q10. Consider $U = (U_1, U_2, U_3, U_4)$ and $Y = (Y_1, Y_2, Y_3, Y_4) = (1, 6, 6, 11)$.

i. Find the probability distribution of the sample mean under $SRSWR(2)$. Find its mean and variance.

ii. Repeat (i) under $SRSWOR(2)$.

iii. Find the probability distribution of Horvitz-Thompson Estimator under $SRSWR(2)$. Find its mean and variance.

iv. Compare performances of different estimators described above.

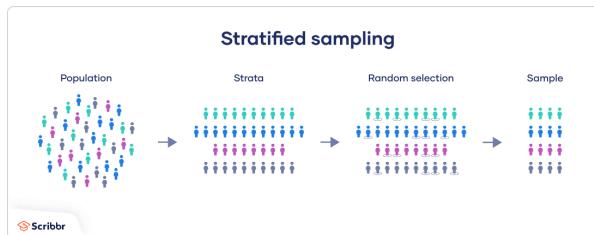
Consider $U = (U_1, U_2, U_3, U_4)$ and $Y = (Y_1, Y_2, Y_3, Y_4) = (1, 6, 6, 11)$.

Find the probability distribution of the sample mean under $SRSWR(2)$. Find its mean and va

Stratified Random Sampling

<https://home.iitk.ac.in/~shalab/sampling/chapter4-sampling-stratified-sampling.pdf>

[Stratified sampling - Wikipedia](#)



Sample Mean

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

Expectation

For both SRSWR and SRSWOR:

$$E[\bar{Y}] = \mu = \frac{1}{N} \sum_{i=1}^N y_i$$

So the sample mean is an unbiased estimator of the population mean.

Variance of Sample Mean

SRSWR

$$Var(\bar{Y}) = \frac{\sigma^2}{n},$$

where $\sigma^2 = \frac{1}{N} \sum_{i=1}^N (Y_i - \mu)^2$
(population variance).

SRSWOR

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

The extra factor $\left(1 - \frac{n}{N}\right)$ is called the finite population correction (fpc).

Equal allocation :

$$n_i = \frac{n}{k}$$

for all $i = 1, \dots, k$

n_i : sample size for stratum i
 n : total sample size,
 k : number of strata.

Proportional Allocation

$$n_i = n \frac{N_i}{N}$$

n_i : sample size allocated to stratum i
 n : total sample size,
 N_i : size of stratum i
 N : total population size,

Optimal Allocation (Neyman Allocation)

$$n_i = n \frac{n N_i S_i}{\sum_{i=1}^k N_i S_i}$$

n_i : sample size allocated to stratum i
 n : total sample size,
 N_i : size of stratum i
 S_i : standard deviation of the variable in stratum i
 N : total population size,
 k : Total number of strata

Problems

Q1. Show that $Var_{Prop}(\widehat{\bar{Y}}_{ST}) \leq Var_{Ran}(\widehat{\bar{Y}})$ provided strata means are so different that

$$\sum_{h=1}^L H_h (\bar{Y}_h - \bar{Y})^2 > \sum_{h=1}^L \left(1 - \frac{n_h}{N_h}\right) S_h^2$$

Q2. Can the sample sizes under fixed sample size proportional allocation and fixed cost proportional allocation exceed corresponding strata size? Justify your answer either through counter examples or by providing proofs. What can you say about Neyman allocation and fixed cost optimum allocations?

Q3. If $2N$ population units are allocated to two strata of the same size ($N_1 = N_2 = N$) The total sample size $2n$ is allocated to strata in proportion to their sizes and *SRSWOR* is taken from each strata. Under what conditions $Var_{Prop}\left(\widehat{\bar{Y}}_{ST}\right) \leq Var_{Ran}\left(\widehat{\bar{Y}}\right)$ hold ?

Q4. A population is divided into two strata of sizes $2N_1$ and N_2 . For the total sample size of $n = N_1 + N_2$, find conditions under which the optimum allocation is $n_1 = N_1$ and $n_2 = N_2$.

Q5. Suppose that study variable $Y \sim U(0, h)$, for some $h > 0$. The range $(0, h]$ is divided into L strata of equal sizes. A simple random sample of size $\frac{n}{L}$ is selected from each stratum. Let V_1 and V_2 denote the variances of the estimators of population mean for a simple random sample of size n and the stratified random sample, respectively. Show that

$$\frac{V_2}{V_1} = \frac{1}{L^2}$$

Solutions

Q1. Show that $Var_{Prop}\left(\widehat{\bar{Y}}_{ST}\right) \leq Var_{Ran}\left(\widehat{\bar{Y}}\right)$ provided strata means are so different that

$$\sum_{h=1}^L H_h (\bar{Y}_h - \bar{Y})^2 > \sum_{h=1}^L \left(1 - \frac{n_h}{N_h}\right) S_h^2$$

Ans 1 : [Shalab notes : PDF page 13](#)

Q2.

Q3. If $2N$ population units are allocated to two strata of the same size ($N_1 = N_2 = N$). The total sample size $2n$ is allocated to strata in proportion to their sizes and *SRSWOR* is taken from each strata. Under what conditions $\text{Var}_{Prop}\left(\widehat{\bar{Y}}_{ST}\right) \leq \text{Var}_{Ran}\left(\widehat{\bar{Y}}\right)$ hold ?

Q4. A population is divided into two strata of sizes $2N_1$ and N_2 . For the total sample size of $n = N_1 + N_2$, find conditions under which the optimum allocation is $n_1 = N_1$ and $n_2 = N_2$.

Q5. Suppose that study variable $Y \sim U(0, h)$, for some $h > 0$. The range $(0, h]$ is divided into L strata of equal sizes. A simple random sample of size $\frac{n}{L}$ is selected from each stratum.

Let V_1 and V_2 denote the variances of the estimators of population mean for a simple random sample of size n and the stratified random sample, respectively.

Show that

$$\frac{V_2}{V_1} = \frac{1}{L^2}$$

Suppose that study variable $Y \sim U(0, h)$, for some $h > 0$. The range $(0, h]$ is divided into L strata. A simple random sample of size $\frac{n}{L}$ is selected from each stratum.

Let V_1 and V_2 denote the variances of the estimators of population mean for a simple random sample of size n and the stratified random sample, respectively.

Show that

$$\frac{V_2}{V_1} = \frac{1}{L^2}$$

What is mean by study variable $Y \sim U(0, h)$
in Sampling Theory

Unequal Probability Sampling

Q1. Let P_1, \dots, P_N be non-negative constants such that $\sum_{i=1}^n P_i = 1$ Under *PPSWR*(n, P_1, \dots, P_N), show that

$$Var(\widehat{Y}_{PPSWR}) = \frac{1}{n} \sum_{1 \leq i < j \leq N} \left(\frac{Y_i}{P_i} - \frac{Y_j}{P_j} \right)^2 P_i P_j$$

Q2. Consider a population of N units. Let V_1 be variance of \widehat{Y}_{PPSWR} under $PPSWR(n, P_1, \dots, P_N)$ and V_2 be the variance of \widehat{Y}_{PPSWR} ($= \widehat{Y}_{HT}$) under $PPSWR(n, P_1, \dots, P_N)$ with $\pi_i = nP_i, i = 1, \dots, N$. Show that $V_2 \leq V_1$ iff $\pi_{ij} > \frac{n-1}{n} \pi_i \pi_j, \forall i \neq j$. Hence compare the variances of \widehat{Y}_{SRSWR} and \widehat{Y}_{SRSSWR} based on random samples of size n each.

Q3. In the context of Problem 2 above, show that there exists a $PPSWR(n, P_1, \dots, P_N)$ with

$$\pi_i = nP_i, i = 1, \dots, N.$$

Q4. For the sample size $n = 2$, show that $Var(\widehat{Y}_{DS}) \leq Var(\widehat{Y}_{PPSWR})$ where the same weight $P = (P_1, \dots, P_N)$ are used under the $PPSWR$ and $PPSWOR$. Can the above result be generalized to a general sample size n ?

Q7. Consider a population $U = (U_1, U_2, U_3, U_4, U_5)$ of five units with study variables $Y = (Y_1, Y_2, Y_3, Y_4, Y_5) = (1, 1, 2, 2, 3)$ A sample of size $n = 2$ is drawn according to the following sampling design:

$$P(s_1, s_2) = Pr((S_1, S_2) = (s_1, s_2)) = \begin{cases} \frac{1}{2} & \text{if } (s_1, s_2) = (U_1, U_2) \\ \frac{1}{6} & \text{if } (s_1, s_2) \in \{(U_3, U_4), (U_3, U_5), (U_4, U_5)\} \\ 0 & \text{otherwise} \end{cases}$$

- (a) Calculate the first and second order inclusion probabilities and find the probability distribution of the Horvitz-Thompson estimator \widehat{Y}_{HT} .
- (b) Find the probability distribution of \widehat{Y}_{HT}^2 and verify if it is an unbiased estimator of Y^2 . Find the variance of \widehat{Y}_{HT}^2 .
- (c) Is it a $PPSWOR(2, P_1, \dots, P_N)$ design for some P_i s?
- (d) Let $P_i, i = 1, \dots, 5$ denote the probability that unit U_i is selected in the first draw. Consider the Des Raj estimator \widehat{Y}_{DS} based on $PPSWOR(2, P_1, \dots, P_5)$ Find the probability

distribution, mean and variance of \hat{Y}_{DS} . Is \hat{Y}_{DS} an unbised estimator of Y .

(e) Compare the performances of \hat{Y}_{HT} and \hat{Y}_{DS} discussed above.

Solutions

Q1. Let P_1, \dots, P_N be non-negative constants such that $\sum_{i=1}^n P_i = 1$ Under $PPSWR(n,)P_1, \dots, P_N$, show that

$$Var(\hat{Y}_{PPSWR}) = \frac{1}{n} \sum_{1 \leq i < j \leq N} \left(\frac{Y_i}{P_i} - \frac{Y_j}{P_j} \right)^2 P_i P_j$$

⇒ If you do this, it will help Q2, Q3, Q4

Let P_1, \dots, P_N be non-negative constants such that $\sum_{i=1}^n P_i = 1$ Under $PPSWR(n,)P_1, \dots, P_N$

$$Var(\hat{Y}_{PPSWR}) = \frac{1}{n} \sum_{1 \leq i < j \leq N} \left(\frac{Y_i}{P_i} - \frac{Y_j}{P_j} \right)^2 P_i P_j$$

Q2. Consider a population of N units. Let V_1 be variance of \hat{Y}_{PPSWR} under

$PPSWR(n, P_1, \dots, P_N)$ and V_2 be the variance of \hat{Y}_{PPSWR} ($= \hat{Y}_{HT}$) under $PPSWR(n, P_1, \dots, P_N)$ with $\pi_i = nP_i, i = 1, \dots, N$. Show that $V_2 \leq V_1$ iff

$\pi_{ij} > \frac{n-1}{n} \pi_i \pi_j, \forall i \neq j$. Hence compare the variances of \hat{Y}_{SRSWR} and \hat{Y}_{SRSWR} based on random samples of size n each.

⇒ I think, typo in Question one of \hat{Y}_{SRSWR} should be \hat{Y}_{SRSWR}

⇒ You'll understand it better if you learn Q1

Q3. In the context of Problem 2 above, show that there exists a $PPSWR(n, P_1, \dots, P_N)$ with

$\pi_i = nP_i, i = 1, \dots, N$.

⇒ You'll understand it better if you learn Q1

Q4. For the sample size $n = 2$, show that $Var(\hat{Y}_{DS}) \leq Var(\hat{Y}_{PPSWR})$ where the same weight $\underline{P} = (P_1, \dots, P_N)$ are used under the $PPSWR$ and $PPSWOR$. Can the above result be generalized to a general sample size n ?

⇒ You'll understand it better if you learn Q1

Q7. Consider a population $U = (U_1, U_2, U_3, U_4, U_5)$ of five units with study variables $Y = (Y_1, Y_2, Y_3, Y_4, Y_5) = (1, 1, 2, 2, 3)$. A sample of size $n = 2$ is drawn according to the following sampling design:

$$P(s_1, s_2) = Pr((S_1, S_2) = (s_1, s_2)) = \begin{cases} \frac{1}{2} & \text{if } (s_1, s_2) = (U_1, U_2) \\ \frac{1}{6} & \text{if } (s_1, s_2) \in \{(U_3, U_4), (U_3, U_5), (U_4, U_5)\} \\ 0 & \text{otherwise} \end{cases}$$

- (a) Calculate the first and second order inclusion probabilities and find the probability distribution of the Horvitz-Thompson estimator \hat{Y}_{HT} .
- (b) Find the probability distribution of \hat{Y}_{HT}^2 and verify if it is an unbiased estimator of Y^2 . Find the variance of \hat{Y}_{HT}^2 .
- (c) Is it a $PPSWOR(2, P_1, \dots, P_N)$ design for some P_i s?
- (d) Let P_i , $i = 1, \dots, 5$ denote the probability that unit U_i is selected in the first draw. Consider the Des Raj estimator \hat{Y}_{DS} based on $PPSWOR(2, P_1, \dots, P_5)$. Find the probability distribution, mean and variance of \hat{Y}_{DS} . Is \hat{Y}_{DS} an unbised estimator of Y .
- (e) Compare the performances of \hat{Y}_{HT} and \hat{Y}_{DS} discussed above.

