CHAPTER 3

Simple Random Sampling

3.1 INTRODUCTION

Simple random sampling is the simplest, popular, and commonly used probability sampling. In this chapter we will consider simple random sampling with (SRSWR) and without replacement (SRSWOR) procedures. Various results on simple random sampling have been presented in detail. Methods of estimating population mean, proportion, and their standard errors (SEs) have been proposed. Estimators of the domain mean, proportions, and their variances have been derived. Methods of determining confidence intervals for mean and proportions have also been proposed. Use of the Rao–Blackwell technique for improvement of estimators based on SRSWR procedure has been presented. The method of inverse sampling for estimating population proportion has also been discussed.

3.2 SIMPLE RANDOM SAMPLING WITHOUT REPLACEMENT

3.2.1 Sampling Scheme

On the first draw a unit is selected from the population of N units at random (with probability $1/N$). On the second draw, another unit is selected at random from the remaining $N-1$ units, which were not selected in the first draw. In general, at the rth draw, a unit is selected at random with probability $p_i(r) = \dfrac{1}{N-(r-1)}$ from $N-(r-1)$ units, which were not selected in the earlier $r-1$ draws where $r = 1,\ldots, n$ and n is the required sample size. So, for an SRSWOR sampling scheme, the probability of selection of an ordered sample $s_o = (i_1 \to i_2, \ldots, \to i_n)$ is $p(s_o) = \dfrac{1}{N}\cdot\dfrac{1}{N-1}\cdots\dfrac{1}{N-n+1}$, where the unit i_k is selected at the kth draw. Hence the probability of selection of an unordered sample $s = (j_1,\ldots, j_k,\ldots, j_n)$ obtained from s_o by arranging their label in ascending order as $j_1 < \cdots < j_k \cdots < j_n$ is given by

$$p(s) = n!\cdot\frac{1}{N}\cdot\frac{1}{N-1}\cdots\frac{1}{N-n+1} = 1\Big/\binom{N}{n}$$

Survey Sampling Theory and Applications
ISBN 978-0-12-811848-1
http://dx.doi.org/10.1016/B978-0-12-811848-1.00003-0

Therefore, the total number of distinct unordered samples of size n is $\binom{N}{n}$.

Theorem 3.2.1
The unconditional probability of selection of a unit at any draw is $1/N$.

Proof
The probability of selection of an ordered sample $s_o = (i_1 \to i_2, \ldots, \to i_n)$ is $p(s_o) = 1 \Big/ \left[n! \binom{N}{n} \right]$.

Hence the unconditional probability of selection of the ith unit at kth draw $= p(s_o) \times$ number of ways we can permute $(n-1)$ integers $(i_1, \ldots, i_{k-1}, i_{k+1}, \ldots, i_n)$ taking from the $(N-1)$ integers $1, \ldots, i-1, i+1, \ldots, N = \binom{N-1}{n-1}(n-1)!p(s_o) = \frac{1}{N}$.

3.2.2 Estimation of Population Mean and Variance

Theorem 3.2.2
Let $\overline{y}(s) = \sum_{i \in s} y_i/n$ be the sample mean based on an unordered sample s. Then,
(i) $\overline{y}(s)$ is an unbiased estimator for the population mean $\overline{Y}$
(ii) Variance of $\overline{y}(s)$ is

$$V(\overline{y}(s)) = \left(\frac{1}{n} - \frac{1}{N}\right) S_y^2 = \frac{1}{n}(1 - f_n) S_y^2$$

where $S_y^2 = \sum_{i=1}^{N} \left(y_i - \overline{Y}\right)^2 \Big/ (N-1)$ and $f_n = n/N$ is known as a sampling fraction.
(iii) An unbiased estimator of $V(\overline{y}(s))$ is

$$\widehat{V}(\overline{y}(s)) = \frac{1}{n}(1 - f_n) s_y^2$$

where $s_y^2 = \sum_{i \in s} \left(y_i - \overline{y}(s)\right)^2 \Big/ (n-1)$ is the sample variance.

Proof
(i) Let $\mathfrak{S}$ be the set of all possible $\binom{N}{n}$ unordered samples of size n and

$$I_{si} = \begin{cases} 1 & \text{if } i \in s \\ 0 & \text{if } i \notin s \end{cases}.$$

Then, $\overline{y}(s) = \frac{1}{n} \sum_{i=1}^{N} I_{si} y_i$

and

$$E(\overline{y}(s)) = \sum_{s\in\mathfrak{S}} \overline{y}(s)p(s) = \frac{1}{n}\sum_{s\in\mathfrak{S}}\sum_{i=1}^{N}\gamma_i I_{si} \Bigg/ \binom{N}{n}$$

$$= \frac{1}{n}\sum_{i=1}^{N}\gamma_i \sum_{s\in\mathfrak{S}} I_{si} \Bigg/ \binom{N}{n}$$

Now noting $\sum_{s\in\mathfrak{S}} I_{si}$ = total number of unordered samples containing the unit $i = \sum_{s\supset i} = \binom{N-1}{n-1}$,

we get $E(\overline{y}(s)) = \frac{1}{n}\sum_{i=1}^{N}\gamma_i \binom{N-1}{n-1} \Big/ \binom{N}{n}$

$$= \frac{1}{N}\sum_{i=1}^{N}\gamma_i = \overline{Y}.$$

(ii) $V(\overline{y}(s)) = E(\overline{y}(s))^2 - \overline{Y}^2$

$$= \frac{1}{n^2}E\left(\sum_{i=1}^{N} I_{si}\gamma_i^2 + \sum_{i\neq j}^{N}\sum_{=1}^{N} I_{si}I_{sj}\gamma_i\gamma_j\right) - \overline{Y}^2$$

$$= \frac{1}{n^2}\sum_{s\in\mathfrak{S}}\frac{1}{\binom{N}{n}}\left(\sum_{i=1}^{N} I_{si}\gamma_i^2 + \sum_{i\neq j}^{N}\sum_{=1}^{N} I_{si}I_{sj}\gamma_i\gamma_j\right) - \overline{Y}^2$$

$$= \frac{1}{n^2\binom{N}{n}}\left(\sum_{i=1}^{N}\gamma_i^2\sum_{s} I_{si} + \sum_{i\neq}^{N}\sum_{j=1}^{N}\gamma_i\gamma_j\sum_{s} I_{si}I_{sj}\right) - \overline{Y}^2$$

Noting $\sum_{s} I_{si}I_{sj} = \sum_{s\supset i,j}$ = number of unordered samples containing the units i and j $(i \neq j) = \binom{N-2}{n-2}$, we can verify part (ii) of the theorem.

(iii) $(n-1)E\left(s_\gamma^2\right) = E\left(\sum_{i\in s}\gamma_i^2\right) - nE(\overline{\gamma}(s))^2$

$$= \sum_{s\in\mathfrak{S}}\sum_{i=1}^{N} I_{si}\gamma_i^2 \Big/ \binom{N}{n} - n\left(V(\overline{\gamma}(s)) + \overline{\Upsilon}^2\right)$$

Now noting

$$\sum_{s\in\mathfrak{S}}\sum_{i=1}^{N} I_{si}\gamma_i^2 \Big/ \binom{N}{n} = \sum_{i=1}^{N}\gamma_i^2 \binom{N-1}{n-1} \Big/ \binom{N}{n} = \frac{n}{N}\sum_{i=1}^{N}\gamma_i^2, \qquad \text{we}$$

find $E\left(s_\gamma^2\right) = S_\gamma^2$, i.e., s_γ^2 is an unbiased estimator of S_γ^2. Hence, part (iii) of the theorem is proved.

Theorem 3.2.3

The first two order inclusion probabilities π_i and $\pi_{ij}(i \neq j)$ for an SRSWOR design with sample size n are n/N and $n(n-1)/\{N(N-1)\}$, respectively.

Proof

$$\pi_i = \sum_{s\in\mathfrak{S}} I_{si}p(s)$$

$$= \sum_{s\in\mathfrak{S}} I_{si} \Big/ \binom{N}{n}$$

$$= \binom{N-1}{n-1} \Big/ \binom{N}{n} = n/N$$

and

$$\pi_{ij} = \sum_{s\in\mathfrak{S}} I_{si}I_{sj}p(s)$$

$$= \sum_{s\in\mathfrak{S}} I_{si}I_{sj} \Big/ \binom{N}{n}$$

$$= \binom{N-2}{n-2} \Big/ \binom{N}{n} = \{n(n-1)/N(N-1)\}$$

The Horvitz–Thomson estimator for the population total Y for SRSWOR sampling design is given by

$$\widehat{Y}_{ht} = \sum_{i\in s} \gamma_i/\pi_i = N\,\overline{\gamma}(s)$$

Using Theorem 3.2.3, we can prove Theorem 3.2.2 in the following alternative way:

(i) $E(\overline{y}(s)) = \frac{1}{N} E\left(\widehat{Y}_{ht}\right)$

$= \frac{Y}{N} = \overline{Y}$

(ii) $V[\overline{y}(s)] = \frac{1}{N^2} V\left(\widehat{Y}_{ht}\right)$

$= \frac{1}{N^2}\left[\frac{1}{2}\sum_{i\neq}^{N}\sum_{j}^{N}(\pi_i\pi_j - \pi_{ij})\left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j}\right)^2\right]$

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

and

(iii) An unbiased estimator of $V(\overline{y}(s))$ is

$$\widehat{V}(\overline{y}(s)) = \frac{1}{N^2}\widehat{V}_{YG}$$

$$= \frac{1}{N^2}\left[\frac{1}{2}\sum_{i\neq}\sum_{j\in s}\frac{(\pi_i\pi_j - \pi_{ij})}{\pi_{ij}}\left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi j}\right)^2\right]$$

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

Remark 3.2.1

Since $\widehat{Y}_{ht} = \sum_{i\in s} y_i/\pi_i = N\,\overline{y}(s)$ is admissible for $Y = N\overline{Y}$, the sample mean $\overline{y}(s)$ is also an admissible estimator of the population mean $\overline{Y}$.

Theorem 3.2.4

Let s be a sample of size n selected from a population of size N by SRSWOR method, s_1 a subsample of size $m\,(\leq n)$ selected from s by SRSWOR method, and $\overline{y}(s_1) = \sum_{i\in s_1} y_i/m$. Then,

(i) $E(\overline{y}(s_1)) = \overline{Y}$

(ii) $V(\overline{y}(s_1)) = \left(\frac{1}{m} - \frac{1}{N}\right) S_y^2$

(iii) An unbiased estimator of $V(\overline{y}(s_1))$ is

$$\widehat{V}(\overline{y}(s)) = \left(\frac{1}{m} - \frac{1}{N}\right) s_{1y}^2, \text{where } s_{1y}^2 = \sum_{i \in s_1} \left(y_i - \overline{y}(s_1)\right)^2 / (m-1)$$

(iv) $Cov(\overline{y}(s_1), \overline{y}(s)) = V(\overline{y}(s)) = \left(\frac{1}{n} - \frac{1}{N}\right) S_y^2$

where $Cov[\overline{y}(s_1), \overline{y}(s)]$ = covariance between $\overline{y}(s_1)$ and $\overline{y}(s)$.

Proof

(i) $E(\overline{y}(s_1)) = E[E(\overline{y}(s_1)|s)] = E(\overline{y}(s)) = \overline{Y}$

(ii) $$\begin{aligned} V(\overline{y}(s_1)) &= E[V(\overline{y}(s_1)|s)] + V[E(\overline{y}(s_1)|s)] \\ &= \left(\frac{1}{m} - \frac{1}{n}\right) E\left(s_y^2\right) + V(\overline{y}(s)) \\ &= \left(\frac{1}{m} - \frac{1}{n}\right) S_y^2 + \left(\frac{1}{n} - \frac{1}{N}\right) S_y^2 \\ &= \left(\frac{1}{m} - \frac{1}{N}\right) S_y^2 \end{aligned}$$

(iii) $$\begin{aligned} Cov(\overline{y}(s_1), \overline{y}(s)) &= E(\overline{y}(s_1)\overline{y}(s)) - \{E(\overline{y}(s_1))\}\{E(\overline{y}(s))\} \\ &= E[E(\overline{y}(s_1)|s)\overline{y}(s)] - \overline{Y}^2 \\ &= E(\overline{y}(s))^2 - \overline{Y}^2 \\ &= V(\overline{y}(s)) \\ &= \left(\frac{1}{n} - \frac{1}{N}\right) S_y^2 \end{aligned}$$

Theorem 3.2.5

Let s be a sample of size n selected from a population U of size N by the SRSWOR method, s^* be a sample of size m selected from $U - s$ (the set of units not belonging to s) by the SRSWOR method, and $\overline{y}(s^*) = \sum_{i \in s^*} y_i / m$.

Then,

(i) $E(\overline{y}(s^*)) = \overline{Y}$

(ii) $V(\overline{y}(s^*)) = \left(\frac{1}{m} - \frac{1}{N}\right) S_y^2$

(iii) An unbiased estimator of $V(\overline{y}(s^*))$ is

$$\widehat{V}(\overline{y}(s^*)) = \left(\frac{1}{m} - \frac{1}{N}\right) s_y^{*2}$$

(iv) $Cov(\overline{y}(s^*), \overline{y}(s)) = -\dfrac{S_y^2}{N}$

where $s_y^{*2} = \sum_{i \in s^*} \left(y_i - \overline{y}(s^*)\right)^2 / (m-1)$.

Proof

(i) $E(\overline{y}(s^*)) = E[E(\overline{y}(s^*)|s)]$

$$= E\left(\sum_{i \in U-s} y_i/(N-n)\right)$$

$$= E(Y - n\overline{y}(s))/N - n$$

$$= \overline{Y}$$

(ii) $$V(\overline{y}(s^*)) = E[V(\overline{y}(s^*)|s)] + V[E(\overline{y}(s^*)|s)] \tag{3.2.1}$$

Now $V(\overline{y}(s^*)|s) = \left(\dfrac{1}{m} - \dfrac{1}{N-n}\right) S_{U-s,y}^2$

where $S_{U-s,y}^2 = \dfrac{1}{N-n-1}\left[\sum_{i \in U-s} y_i^2 - \dfrac{1}{(N-n)}\left(\sum_{i \in U-s} y_i\right)^2\right]$.

$$E\left(S_{U-s,y}^2\right) = \frac{1}{N-n-1} E\left[\left(\sum_{i \in U} y_i^2 - \sum_{i \in s} y_i^2\right) - \frac{1}{(N-n)}\left(N\overline{Y} - n\overline{y}(s)\right)^2\right]$$

$$= \frac{1}{N-n-1}\left[\left(\sum_{i \in U} y_i^2 - \frac{n}{N}\sum_{i \in U} y_i^2\right)\right.$$

$$\left. - \frac{1}{(N-n)} E\left\{\left(N\overline{Y}\right)^2 + n^2(\overline{y}(s))^2 - 2\, n\ N\ \overline{Y}\ \overline{y}(s)\right\}\right]$$

$$= \frac{1}{N-n-1}\left[\left(\frac{N-n}{N}\sum_{i \in U} y_i^2\right)\right.$$

$$\left. - \frac{1}{(N-n)}\left\{(N-n)^2\ \overline{Y}^2 + n^2 V(\overline{y}(s))\right\}\right] = S_y^2$$

Hence,

$$E[V(\overline{y}(s^*)|s)] = \left(\frac{1}{m} - \frac{1}{N-n}\right) S_y^2 \tag{3.2.2}$$

Furthermore,

$$V[E(\bar{y}(s^*)|s)] = V\left[\frac{N\bar{Y} - n\bar{y}(s)}{N-n}\right] = \left(\frac{n}{N-n}\right)^2 V(\bar{y}(s)) = \left(\frac{1}{N-n} - \frac{1}{N}\right) S_y^2 \tag{3.2.3}$$

Finally substituting (3.2.2) and (3.2.3) in (3.2.1), we can verify part (ii) of the theorem.

(iii) $$E\left(s_y^{*2}\right) = E\left[E\left(s_y^{*2}|s\right)\right]$$

$$= \frac{1}{m-1} E\left[\sum_{i\in s^*} y_i^2 - m(\bar{y}(s^*))^2\right]$$

$$= \frac{1}{m-1}\left[E\left\{E\left(\sum_{i\in s^*} y_i^2|s\right)\right\} - m\left\{V(\bar{y}(s^*) + \bar{Y}^2\right\}\right]$$

$$= \frac{1}{m-1}\left[E\left(\frac{m}{N-n}\sum_{i\in U-s} y_i^2\right) - m\left\{\left(\frac{1}{m} - \frac{1}{N}\right)S_y^2 + \bar{Y}^2\right\}\right]$$

$$= \frac{1}{m-1}\left[\frac{m}{N-n}\left\{\sum_{i\in U} y_i^2 - E\left(\sum_{i\in s} y_i^2\right)\right\} - m\left\{\left(\frac{1}{m} - \frac{1}{N}\right)S_y^2 + \bar{Y}^2\right\}\right]$$

$$= \frac{1}{m-1}\left[\frac{m}{N-n}\left(\sum_{i\in U} y_i^2 - \frac{n}{N}\sum_{i\in U} y_i^2\right) - m\left\{\left(\frac{1}{m} - \frac{1}{N}\right)S_y^2 + \bar{Y}^2\right\}\right]$$

$$= S_y^2 \tag{3.2.4}$$

Using (3.2.4) we verify $E\left[\widehat{V}(\bar{y}(s^*))\right] = V[\bar{y}(s^*)]$

(iv) $$Cov[\bar{y}(s^*), \bar{y}(s)] = E[\bar{y}(s^*), \bar{y}(s)] - [E\{\bar{y}(s^*)\}][E\{\bar{y}(s)\}]$$

$$= E[E\{\bar{y}(s^*)|s\}\bar{y}(s)] - Y^2$$

$$= \frac{1}{N-n}\left[E\left\{N\bar{Y} - n\bar{y}(s)\right\}\bar{y}(s)\right] - \bar{Y}^2$$

$$= \frac{1}{N-n}\left[N\bar{Y}^2 - nV\{\bar{y}(s)\} - n\bar{Y}^2\right] - \bar{Y}^2$$

$$= -S_y^2/N$$

3.2.3 Estimation of Population Covariance

Let y_i and x_i be the value of the variables y and x for the ith unit of the population U. Then the finite population covariance between x and y is defined as

$$S_{xy} = \sum_{i=1}^{N} (y_i - \overline{Y})(x_i - \overline{X})/(N-1) \text{ where } \overline{X} = \sum_{i=1}^{N} x_i/N.$$

Let a sample s of size n be selected by SRSWOR. Let $\overline{x}(s)$ and $\overline{y}(s)$ be the sample mean of x and y, respectively. The sample covariance between x and y is defined as

$$s_{xy} = \sum_{i \in s} \{x_i - \overline{x}(s)\}\{y_i - \overline{y}(s)\}/(n-1).$$

Theorem 3.2.6

$$E(s_{xy}) = S_{xy}$$

Proof

$$\begin{aligned} E(s_{xy}) &= E\left[\sum_{i \in s} x_i y_i - n\overline{x}(s)\overline{y}(s)\right] \Big/ (n-1) \\ &= E\left[\frac{1}{n}\sum_{i \in s} x_i y_i - \frac{1}{n(n-1)} \sum_{i \neq} \sum_{j \in s} x_i y_j\right] \\ &= \left[\frac{1}{n}\sum_{i \in U} x_i y_i \sum_{s \supset i} - \frac{1}{n(n-1)} \sum_{i \neq} \sum_{j \in U} x_i y_j \sum_{s \supset i,j}\right] \Big/ \binom{N}{n} \\ &= \frac{1}{N}\sum_{i \in U} x_i y_i - \frac{1}{N(N-1)} \sum_{i \neq} \sum_{j \in U} x_i y_j \\ &= S_{xy} \end{aligned}$$

Theorem 3.2.7

(i) The covariance between $\overline{y}(s)$ and $\overline{x}(s)$ is

$$\begin{aligned} Cov\{\overline{x}(s), \overline{y}(s)\} &= \left(\frac{1}{n} - \frac{1}{N}\right) S_{xy} \\ &= \left(\frac{1}{n} - \frac{1}{N}\right) \rho S_x S_y \end{aligned}$$

(ii) An unbiased estimator of $Cov\{\overline{x}(s), \overline{y}(s)\}$ is

$$Cov\{\overline{x}(s), \overline{y}(s)\} = \left(\frac{1}{n} - \frac{1}{N}\right) s_{xy}$$
$$= \left(\frac{1}{n} - \frac{1}{N}\right) \widehat{\rho} s_x s_y$$

where

ρ = population correlation coefficient between x and $y = S_{xy}/(S_x S_y)$

and

$\widehat{\rho}$ = sample correlation coefficient between x and $y = s_{xy}/(s_x s_y)$

Proof

$$\text{(i) } Cov\{\overline{x}(s), \overline{y}(s)\} = E\{\overline{x}(s)\overline{y}(s)\} - [E\{\overline{x}(s)\}][E\{\overline{y}(s)\}]$$
$$= E\left(\sum_{i\in s} x_i\right)\left(\sum_{i\in s} y_i\right) \Big/ n^2 - \overline{X}\,\overline{Y}$$
$$= E\left(\sum_{i\in s} x_i y_i + \sum_{i\neq} \sum_{j\in s} x_i y_j\right) \Big/ n^2 - \overline{X}\,\overline{Y}$$
$$= \frac{1}{\binom{N}{n}} \sum_{s} \left(\sum_{i\in s} x_i y_i + \sum_{i\neq} \sum_{j\in s} x_i y_j\right) \Big/ n^2 - \overline{X}\,\overline{Y}$$
$$= \left(\sum_{i=1}^{N} x_i y_j \sum_{s\supset i} + \sum_{i\neq}^{N} \sum_{j=1}^{N} x_i y_j \sum_{s\supset i,j}\right) \Big/ \left\{n^2 \binom{N}{n}\right\} - \overline{X}\,\overline{Y}$$
$$= \left\{\sum_{i=1}^{N} x_i y_i \binom{N-1}{n-1} + \sum_{i\neq}^{N} \sum_{j=1}^{N} x_i y_j \times \binom{N-2}{n-2}\right\} \Big/ \left\{n^2 \binom{N}{n}\right\} - \overline{X}\,\overline{Y}$$
$$= \frac{1}{Nn}\left(\sum_{i=1}^{N} x_i y_i + \frac{n-1}{N-1} \sum_{i\neq}^{N} \sum_{j=1}^{N} x_i y_j\right) - \overline{X}\,\overline{Y}$$
$$= \left(\frac{1}{n} - \frac{1}{N}\right) S_{xy}$$

(ii) The result follows from Theorem 3.2.6.

An Alternative Proof of the Theorem 3.2.7

(i) Writing $k_i = y_i - x_i$ for every $i \in U$ and $\overline{k}(s) = \sum_{i \in s} k_i/n = \overline{y}(s) - \overline{x}(s)$, we get

$$V\left\{\overline{k}(s)\right\} = V\{\overline{y}(s)\} + V\{\overline{x}(s)\} - 2Cov\{\overline{y}(s), \overline{x}(s)\} \text{ and hence}$$

$$Cov\{\overline{y}(s), \overline{x}(s)\} = \left[V\{\overline{y}(s)\} + V\{\overline{x}(s)\} - V\left\{\overline{k}(s)\right\}\right]/2 \tag{3.2.5}$$

Now using Theorem 3.2.2, we get

$$V\left\{\overline{k}(s)\right\} = \left(\frac{1}{n} - \frac{1}{N}\right) S_k^2$$

where

$$S_k^2 = \sum_{i=1}^{N} \left(k_i - \overline{K}\right)^2/(N-1) = S_y^2 + S_x^2 - 2S_{xy}, \overline{K} = \overline{Y} - \overline{X} \text{ and}$$

$$S_x^2 = \sum_{i=1}^{N} \left(x_i - \overline{X}\right)^2/(N-1) \tag{3.2.6}$$

Finally, substituting $V\{\overline{x}(s)\} = \left(\frac{1}{n} - \frac{1}{N}\right) S_x^2$, $V\{\overline{y}(s)\} = \left(\frac{1}{n} - \frac{1}{N}\right) S_y^2$ and $S_k^2 = S_y^2 + S_x^2 - 2S_{xy}$ in (3.2.5), we can verify the theorem.

(ii) Let $s_k^2 = \frac{1}{n-1} \sum_{i \in s} \left\{k_i - \overline{k}(s)\right\}^2 = s_x^2 + s_y^2 - 2s_{xy}$, where $s_x^2 = \frac{1}{n-1} \sum_{i \in s} \{x_i - \overline{x}(s)\}^2$. Theorem 3.2.2 yields

$$E\left(s_k^2\right) = S_k^2$$

i.e., $E\left(s_x^2\right) + E\left(s_y^2\right) - 2E(s_{xy}) = S_x^2 + S_y^2 - 2S_{xy}$;

i.e., $E(s_{xy}) = S_{xy}$

3.2.4 Estimation of Population Proportion

Sometimes we need to estimate the proportion of the population that possesses a certain attribute A, such as smoking, drug addiction, or unemployment. In such a situation, we take $y_i = 1$ if the ith unit belongs to the group A and $y_i = 0$ otherwise (if the ith unit does not belong to the group A). So in this case, $Y = N_A$ = total number of units possessing the attribute A and $\overline{Y} = N_A/N = \pi_A$ = proportion of units in the population belonging to the group A; $\overline{y}(s) = n_A/n = \widehat{\pi}_A$ = proportion of units in the sample that belong to the group A and n_A is the total number of units in the sample that belong to the group A. Now noting

(i) $S_y^2 = \frac{1}{(N-1)} \sum_{i=1}^{N} (y_i - \overline{Y})^2 = \frac{1}{(N-1)} \left(\sum_{i=1}^{N} y_i - N\overline{Y}^2 \right)$

$= \frac{N\{\pi_A(1-\pi_A)\}}{N-1}$ (since $y_i = 0$ or 1),

(ii) $s_y^2 = \sum_{i \in s} \{y_i - \overline{y}(s)\}^2 / (n-1) = \frac{n}{n-1} \widehat{\pi}_A(1-\widehat{\pi}_A)$,

and using Theorem 3.2.2, we have the following.

Theorem 3.2.8

For an SRSWOR sampling of size n,

(i) $\widehat{\pi}_A$ is an unbiased estimator for the population proportion π_A.

(ii) Variance of $\widehat{\pi}_A$ is

$$V(\widehat{\pi}_A) = \frac{N-n}{n(N-1)} \pi_A(1-\pi_A)$$

and

(iii) An unbiased estimator of $V(\widehat{\pi}_A)$ is

$$\widehat{V}(\widehat{\pi}_A) = \frac{N-n}{N(n-1)} \widehat{\pi}_A(1-\widehat{\pi}_A).$$

3.2.5 Estimation of Domain Mean and Total

Suppose we want to estimate $Y_D = \sum_{i \in D} y_i$, the domain total of the y-values of a section D of the population U. Let N_D be the population domain size = total number of units in the domain D, then $\overline{Y}_D = Y_D/N_D$ = population domain mean of y and $S_{yD}^2 = \sum_{i \in D} (y_i - \overline{Y}_D)^2/(N_D - 1)$ = population domain variance of y. Let $s_D (\subset s)$ be the set of units in the sample s, which belong to D, and $n_D (\leq n)$ be the total numbers of units in s_D, $\overline{y}(s_D) = \sum_{i \in s_D} y_i/n_D$ = sample domain mean and $s_{yD}^2 = \sum_{i \in s_D} \{y_i - \overline{y}(s_D)\}^2/(n_D - 1)$ be the sample domain variance.

Let us define $z_i = d_i y_i$, where $d_i = 1$ if the ith unit belongs to D and $d_i = 0$ otherwise. Then, $Z = \sum_{i=1}^{N} z_i = Y_D$ = population domain total, $\overline{Y}_D = Y_D/N_D$ = population domain mean, $\pi_D = N_D/N$ = population domain proportion, and $\overline{Z} = Z/N = \pi_D \overline{Y}_D$; the sample domain

mean $= \overline{y}(s_D) = \sum_{i \in s_D} y_i / n_D$ and the sample domain proportion $= \widehat{\pi}_D = n_D/n$. Furthermore,

$$S_z^2 = \sum_{i \in U} \left(z_i - \overline{Z}\right)^2 / (N-1)$$

$$= \left[\sum_{i \in D} y_i^2 - N\left(\pi_D \overline{Y}_D\right)^2 \right] \Big/ (N-1)$$

$$= (N_D - 1) S_{yD}^2 / (N-1) + N_D \overline{Y}_D^2 (1 - \pi_D)/(N-1)$$

$$= \left[(N\pi_D - 1) S_{yD}^2 + N\pi_D (1-\pi_D) \overline{Y}_D^2 \right] \Big/ (N-1),$$

$$\overline{z}(s) = \sum_{i \in s} z_i / n = \sum_{i \in s_D} y_i / n = \widehat{\pi}_D \overline{y}(s_D) \text{ and } s_z^2 = \sum_{i \in s} \left[z_i - \overline{z}(s)\right]^2 / (n-1)$$

$$= (n\widehat{\pi}_D - 1) s_{yD}^2 \Big/ (n-1) + n\widehat{\pi}_D (1 - \widehat{\pi}_D) \left\{\overline{y}(s_D)\right\}^2 / (n-1)$$

Now using Theorem 3.2.2, we derive the following results:

Theorem 3.2.9

(i) $\widehat{Y}_D = N\,\overline{z}(s) = N\widehat{\pi}_D \overline{y}(s_D)$ is an unbiased estimator of the domain total Y_D

(ii) The variance of $\widehat{Y}_D$ is

$$V\left[\widehat{Y}_D\right] = N^2 \left(\frac{1}{n} - \frac{1}{N}\right) S_z^2$$

$$= N^2 \left(\frac{1}{n} - \frac{1}{N}\right) \left[\frac{(N\pi_D - 1)}{N-1} S_{yD}^2 + \frac{N\pi_D (1-\pi_D)}{N-1} \overline{Y}_D^2 \right]$$

(iii) An unbiased estimator of $V\left[\widehat{Y}_D\right]$ is

$$\widehat{V}\left[\widehat{Y}_D\right] = N^2 \left(\frac{1}{n} - \frac{1}{N}\right) s_z^2$$

$$= N^2 \left(\frac{1}{n} - \frac{1}{N}\right) \left[\frac{(n\widehat{\pi}_D - 1)}{n-1} s_{yD}^2 + \frac{n\widehat{\pi}_D (1 - \widehat{\pi}_D)}{n-1} \left\{\overline{y}(s_D)\right\}^2 \right]$$

Remark 3.2.2

In case the domain size N_D is known, one can easily get an unbiased estimator for the domain mean $\overline{Y}_D$ as $\widehat{\overline{Y}}_D = \widehat{Y}_D / N_D = \dfrac{\widehat{\pi}_D}{\pi_D} \overline{y}(s_D)$. But in

most situations, N_D is unknown, and we can use the sample mean $\overline{y}(s_D)$ as an unbiased estimator of $\overline{Y}_D$. The following theorem gives the detailed properties of $\overline{y}(s_D)$.

Theorem 3.2.10

(i) $E[\overline{y}(s_D)] = \overline{Y}_D$

(ii) $V[\overline{y}(s_D)] = \left[E\left(\frac{1}{n_D}\right) - \frac{1}{N_D}\right] S_{yD}^2$

(iii) $\widehat{V}[\overline{y}(s_D)] = \left(\frac{1}{n_D} - \frac{1}{N_D}\right) s_{yD}^2$

where $V[\overline{y}(s_D)]$ and $\widehat{V}[\overline{y}(s_D)]$ are respectively the variance of $\overline{y}(s_D)$ and an unbiased estimator of $V[\overline{y}(s_D)]$.

Proof

For fixed n_D, we regard $\overline{y}(s_D)$ as a random sample of the entire domain D selected by SRSWOR method. Hence

$$E\{\overline{y}(s_D)|n_D\} = \overline{Y}_D, V\{\overline{y}(s_D)|n_D\} = \left(\frac{1}{n_D} - \frac{1}{N_D}\right) S_{yD}^2 \text{ and } E\left\{s_{yD}^2 \middle| n_D\right\} = s_{yD}^2 \tag{3.2.7}$$

Now using (3.2.7), we can find

(i) $E[\overline{y}(s_D)] = E[\overline{y}(s_D)|n_D] = \overline{Y}_D$

(ii) $V[\overline{y}(s_D)] = E[V\{\overline{y}(s_D)|n_D\}] + V[E\{\overline{y}(s_D)|n_D\}]$

$$= \left[E\left(\frac{1}{n_D}\right) - \frac{1}{N_D}\right] S_{yD}^2 + V\left(\overline{Y}_D\right)$$

$$= \left[E\left(\frac{1}{n_D}\right) - \frac{1}{N_D}\right] S_{yD}^2$$

(iii) $E\left[\widehat{V}\{\overline{y}(s_D)\}\right] = E\left[E\left(\widehat{V}\{\overline{y}(s_D)|n_D\}\right)\right]$

$$= \left[E\left(\frac{1}{n_D}\right) - \frac{1}{N_D}\right] S_{yD}^2$$

$$= V[\overline{y}(s_D)]$$

Example 3.2.1

Table 3.2.1 gives the height, weight, and gender of 30 students who participated in a sport.

Table 3.2.1

Serial no. of students	Height (cm) x	Weight (kg) y	Gender Male = 1 Female = 0	Serial no. of students	Height (cm) x	Weight (kg) y	Gender Male = 1 Female = 0
1	155	60	0	16	165	70	1
2	160	70	1	17	140	50	0
3	170	72	1	18	165	70	1
4	150	56	0	19	168	72	1
5	170	70	1	20	160	65	1
6	175	75	0	21	158	58	0
7	160	70	1	22	160	60	0
8	160	70	0	23	155	50	1
9	155	55	0	24	170	75	1
10	150	60	1	25	145	78	1
11	158	65	0	26	156	65	1
12	170	80	1	27	156	62	0
13	165	75	1	28	175	42	1
14	165	70	1	29	170	75	1
15	160	72	1	30	150	60	0

Select a sample of size 8 by the SRSWOR method. From the selected sample, (i) estimate the mean height and weight of students (male and female combined) and obtain the variances of the estimator used. Estimate the SEs of the estimators. (ii) Estimate the proportion of the male and female students with their SEs. (iii) Estimate the mean weight of the male and female students and estimate their SEs. (iv) Give an unbiased estimator of the covariance of height and weight of the students (male and female combined).

To select 30 students from a list of 60 students, we consider three sets of two-digit random numbers. The first set consists of random numbers 01−30; the second set 31−60, and the third set 61−90, respectively. We associate random numbers 01, 02,…, 30 of the first set to the student serial numbers 1, 2,…, 30, respectively. Random numbers 31, 32,…, 60 of the second set are associated with the units 1, 2,…, 30, respectively. Similarly, random numbers 61, 62,…, 90 of the third set are associated with the units 1, 2,…, 30, respectively. From the page of a random number table, we select the random numbers row wise as follows:

Random number selected	45	45	27	54	46	66	60	24	67	74
Unit selected	15	—	27	24	16	6	30	—	7	14

Here the symbol "—" indicates selection of no unit because the unit 15 and 24 are already selected in earlier draws.

So, the selected ordered sample is $s_o = (15, 27, 24, 16, 6, 30, 7, 14)$. After arranging the units in ascending order of label, the unordered sample obtained is $s = (6, 7, 14, 15, 16, 24, 27, 30)$. The data obtained from the selected sample s is as follows:

Serial no. of students	Height (cm) x	Weight (kg) y	Gender Male = 1 Female = 0
6	175	75	0
7	160	70	1
14	165	70	1
15	160	72	1
16	165	70	1
24	170	75	1
27	156	62	0
30	150	60	0

(i) Estimated mean height and weight (male and female combined) are given by $\overline{x}_s = 162.625$ cm and $\overline{y}_s = 69.25$ kg, respectively. The variance of $\overline{x}_s$ is $V(\overline{x}_s) = (1/n - 1/N)S_x^2 = (1/8 - 1/30) \times 73.981 = 6.781\ \text{cm}^2$. The variance of $\overline{y}_s$ is $V(\overline{y}_s) = (1/n - 1/N)S_y^2 = (1/8 - 1/30) \times 83.374 = 7.642\ \text{kg}^2$

The estimated SEs for $\overline{x}_s$ and $\overline{y}_s$ are $se(\overline{x}_s) = \sqrt{\widehat{V}(\overline{x}_s)} = \sqrt{(1/n - 1/N)s_x^2} = \sqrt{(1/8 - 1/30) \times 62.267} = 2.389$ cm and $se(\overline{y}_s) = \sqrt{\widehat{V}(\overline{y}_s)} = \sqrt{(1/n - 1/N)s_y^2} = \sqrt{(1/8 - 1/30) \times 30.5} = 1.672$ kg, respectively.

(ii) The estimated proportions of male and female students are given by $\widehat{\pi}_m = 5/8 = 0.625$ and $\widehat{\pi}_f = 3/8 = 0.375$. Estimated SEs of $\widehat{\pi}_m$ and $\widehat{\pi}_f$ are given respectively $Se(\widehat{\pi}_m) = \sqrt{\frac{N - n}{N(n - 1)}\widehat{\pi}_m(1 - \widehat{\pi}_m)} = \sqrt{\frac{30 - 8}{30 \times 7} \times 0.625 \times (1 - 0.625)} = 0.156$ and $Se(\widehat{\pi}_f) = \sqrt{\frac{N - n}{N(n - 1)}\widehat{\pi}_f(1 - \widehat{\pi}_f)} = \sqrt{\frac{30 - 8}{30 \times 7} \times 0.375 \times (1 - 0.375)} = 0.156$

(iii) Selected male and female students are given by $s_1 = (7, 14, 15, 16, 24)$ and $s_2 = (6, 27, 30)$, respectively. Estimated mean weights of the male and female students are given by $\overline{y}_{s_1} = 71.4$ kg and $\overline{y}_{s_2} = 65.7$ kg. The sample variances of the weights of male and female students are $s_{1y}^2 = 4.8$ and $s_{2y}^2 = 66.33$. Estimated SEs for the male and females are $Se(\overline{y}_{s_1}) = \sqrt{\widehat{V}(\overline{y}_{s_1})} = \sqrt{(1/5 - 1/8)s_{1y}^2} = \sqrt{(1/5 - 1/8) \times 4.8} = 0.6$ kg and $Se(\overline{y}_{s_2}) = \sqrt{\widehat{V}(\overline{y}_{s_2})} = \sqrt{(1/3 - 1/8)s_{2y}^2} = \sqrt{(1/3 - 1/8) \times 66.33} = 3.717$ kg.

(iv) Unbiased estimate of the covariance of the height and weight of the students (male and female combined) = sample covariance $= s_{xy} = \sum_{i \in s}\{x_i - \overline{x}(s)\}\{y_i - \overline{y}(s)\}/(n - 1) = \left(\sum_{i \in s} x_i y_i - 8 \times 162.625 \times 69.25\right)\Big/7 = 38.964$.

3.3 SIMPLE RANDOM SAMPLING WITH REPLACEMENT

3.3.1 Sampling Scheme

On the first draw, a unit is selected from a population of N units at random, i.e., with probability $1/N$. For drawing the second unit, the unit selected on the first draw is returned to the population and a unit is selected at random again from the entire population with probability $1/N$. So the unit selected on the first draw may appear again in the second draw. The procedure is continued till a sample of n units is selected. So, in an SRSWR, the probability of selection of the ith unit at the kth draw is $p_i(k) = 1/N$; $i = 1,\ldots, N$, $k = 1,\ldots, n$ and the probability of selection of an ordered sample $s_o = (i_1 \to i_2,\cdots, \to i_n)$ is $p(s_o) = 1/N^n$.

3.3.2 Estimation of the Population Mean and Variance

Let y_i be the value of the variable under study y for the ith ($i = 1, 2,\ldots, N$) unit of the population and $y_{(r)}$ the value of the variable y for the unit that is selected at the rth ($r = 1, 2,\ldots, n$) draw. Clearly, $y_{(r)} = y_j$ if the rth draw produces the jth unit, $j = 1,\ldots, N$. Let $\overline{y}(s_o) = \sum_{r=1}^{n} y_{(r)}/n$ and $\widehat{\sigma}_y^2 = \sum_{r=1}^{n} \{y_{(r)} - \overline{y}(s_o)\}^2/(n-1)$ denote the sample mean and sample variance, respectively, based on the ordered sample s_o of size n units with repetition. Then we have the following theorem:

Theorem 3.3.1

(i) $\overline{y}(s_o)$ is an unbiased estimator for the population mean $\overline{Y}$.

(ii) The variance of $\overline{y}(s_o)$ is

$$V[\overline{y}(s_o)] = \sigma_y^2/n$$

(iii) An unbiased estimator of $V[\overline{y}(s_o)]$ is

$$\widehat{V}[\overline{y}(s_o)] = \widehat{\sigma}_y^2/n$$

where $\sigma_y^2 = \sum_{i=1}^{N} (y_i - \overline{Y})^2/N$

Proof

(i) $E[\overline{y}(s_o)] = \sum_{r=1}^{n} E(y_{(r)})/n = \overline{Y}$ since, $E(y_{(r)}) = \sum_{r=1}^{N} y_i/N = \overline{Y}$.

(ii) $V[\overline{y}(s_o)] = V\left(\sum_{r=1}^{n} y_{(r)}/n\right)$

$$= \left[\sum_{r=1}^{n} V(y_{(r)}) + \sum_{r\neq}^{n}\sum_{r'}^{n} Cov(y_{(r)}, y_{(r')})\right] \Big/ n^2$$

$$= \sigma_y^2/n$$

since, $V(y_{(r)}) = E(y_{(r)} - \overline{Y})^2 = \sum_{i=1}^{N} (y_i - \overline{Y})^2/N = \sigma_y^2$ and $Cov(y_{(r)}, y_{(r')}) = 0$ for $r \neq r'$ as the draws are independent.

(iii) $$E\left(\widehat{\sigma}_y^2/n\right) = \frac{1}{n(n-1)}\left[\sum_{r=1}^{n} E\left(y_{(r)}^2\right) - nE\{\overline{y}(s_o)\}^2\right]$$

$$= \frac{1}{n(n-1)}\left[n\left(\sigma_y^2 + \overline{Y}^2\right) - n\left(\frac{\sigma_y^2}{n} + \overline{Y}^2\right)\right]$$

$$= V[\overline{y}(s_o)]$$

3.3.3 Estimation of Population Proportion

Let $y_i = 1$ if the ith unit belongs to the group A, and let $y_i = 0$ if the ith unit does not belong to the group A. In this case $Y = N_A$ = total number of units that possess the attribute A and $\overline{Y} = N_A/N = \pi_A$ = proportion of units in the population belonging to the group A; $\overline{y}(s_o) = n_A/n = \widehat{\pi}_A$ = proportion of units in the sample s_o belonging to A, where n_A is the total number of units in the sample that fall in group A. Now noting that

(i) $\sigma_y^2 = \sum_{i=1}^{N} (y_i - \overline{Y})^2/N = \sum_{i=1}^{N} y_i/N - \overline{Y}^2 = \pi_A(1 - \pi_A)$ since $y_i = 0$ or 1; and

(ii) $\widehat{\sigma}_y^2 = \sum_{r=1}^{n} \{y_{(r)} - \overline{y}(s_o)\}^2/(n-1) = \frac{n}{n-1}\widehat{\pi}_A(1 - \widehat{\pi}_A)$

we have the following theorem:

Theorem 3.3.2

(i) $\widehat{\pi}_A$ is an unbiased estimator for the population proportion π_A.

(ii) Variance of $\widehat{\pi}_A$ is $V(\widehat{\pi}_A) = \dfrac{\pi_A(1 - \pi_A)}{n}$

(iii) An unbiased estimator of $V(\widehat{\pi}_A)$ is $\widehat{V}[\widehat{\pi}_A] = \dfrac{\widehat{\pi}_A(1 - \widehat{\pi}_A)}{n-1}$

Remark 3.3.1

It is important to note that Theorems 3.3.1 and 3.3.2 can be obtained from Theorems 3.2.2 and 3.2.8 when N is sufficiently large compared to n so that the finite population correction term $f_n = n/N$ is ignored.

Example 3.3.1

From the list of 30 students given in the Example 3.2.1, select a sample of size 8 by the SRSWR method. From the selected sample, (i) estimate the mean height and weight of students (male and female combined) and obtain

the variances of the estimator used. Estimate the SEs of the estimators. (ii) Estimate the proportion of the male and female students with their SEs.

Here we associate serial number of students to the two-digit random numbers 01−90 as in Example 3.2.1, and we select the random numbers as follows:

Random number selected	45	45	27	54	46	66	60	24
Unit selected	15	15	27	24	16	6	30	24

Here the selected SRSWR sample is $s_o = (15, 15, 27, 24, 16, 6, 30, 24)$ and the data corresponding to the sample s_o are given below:

Serial no. of students	Height (cm)	Weight (kg)	Gender Male = 1 Female = 0
15	160	72	1
15	160	72	1
27	156	62	0
24	170	75	1
16	165	70	1
6	175	75	0
30	150	60	0
24	170	75	1

(i) Estimated mean height and weight of the 30 students are given by $\overline{x}_s = 163.25$ cm and $\overline{y}_s = 70.125$ kg. The variance of $\overline{x}_s$ is $\sigma_x^2/n = (N-1)S_x^2/(nN) = 29 \times 73.981/(30 \times 8) = 8.939$ cm^2. The variance of $\overline{y}_s$ is $V(\overline{y}_s) = \sigma_y^2/n = (N-1)S_y^2/(nN) = 29 \times 83.374/(30 \times 8) = 10.074$ kg^2. The estimated SEs for $\overline{x}_s$ and $\overline{y}_s$ are, respectively, $se(\overline{x}_s) = \sqrt{\widehat{V}(\overline{x}_s)} = \sqrt{s_x^2/n} = \sqrt{68.785/8} = 2.932$ cm and $se(\overline{y}_s) = \sqrt{\widehat{V}(\overline{y}_s)} = \sqrt{s_y^2/n} = \sqrt{35.267/8} = 2.099$ kg.

(ii) The estimated proportions of male and female students are $\widehat{\pi}_m = 5/8 = 0.625$ and $\widehat{\pi}_f = 3/8 = 0.375$. Estimated SEs of $\widehat{\pi}_m$ and $\widehat{\pi}_f$ are, respectively, $Se(\widehat{\pi}_m) = \sqrt{\dfrac{\widehat{\pi}_m(1-\widehat{\pi}_m)}{n-1}} = \sqrt{\dfrac{0.625(1-0.625)}{7}} = 0.183$ and $Se(\widehat{\pi}_f) = \sqrt{\dfrac{\widehat{\pi}_f(1-\widehat{\pi}_f)}{n-1}} = \sqrt{\dfrac{0.375(1-0.375)}{7}} = 0.183$

3.3.4 Rao–Blackwellization

It follows from Section 2.7.3 that the estimator $\overline{y}(s_o)$ is inadmissible because it is based on ordered data, which may consist of repletion of units, hence $\overline{y}(s_o)$ is not a function of a sufficient statistic. Let $s = (j_1,\ldots, j_\nu)$ denote the unordered sample obtained by taking distinct ν units $j_1,\ldots, j_\nu$ with $j_1<\cdots<j_\nu$. The unordered sample s is a sufficient statistic. Hence, we can improve the inefficient estimator $\overline{y}(s_o)$ by applying the Rao–Blackwellization technique. The improved estimator is given by $E[\overline{y}(s_o)|s] = \overline{y}_s = \sum_{i\in s} y_i/\nu =$ the sample mean based on the distinct units. The details have been given in the following theorems.

Theorem 3.3.3

Let $\overline{y}_s = \sum_{i\in s} y_i/\nu = \sum_{k=1}^{\nu} y_{j_k}/\nu$ be the sample mean based on the distinct units of s_o. Then,

(i) $E\left[\overline{y}_s\right] = E[\overline{y}(s_o)] = \overline{Y}$

(ii) $V\left[\overline{y}_s\right] \le V[\overline{y}(s_o)]$

Proof

Let $n_i(s_o)$ denote the number of times the ith unit appears in s_o. Then writing,

$$\begin{aligned}\overline{y}(s_o) &= \sum_{r=1}^{n} y_{(r)}/n \\ &= \sum_{i=1}^{N} n_i(s_o)y_i/n \\ &= \sum_{k=1}^{\nu} n_{j_k}(s_o)y_{j_k}/n\end{aligned}$$

where $n_{j_1}(s_o), \ldots, n_{j_\gamma}(s_o)$ denote the number of times the distinct units $j_1,\ldots, j_\nu$ appear in s_o.

Now noting $E\left(n_{j_k}(s_o)|s\right) = n/\nu$ we get

(i) $E[\overline{y}(s_o|s)] = \sum_{k=1}^{\nu} E\left(n_{j_k}(s_o)\big|s\right)y_{j_k}/n = \overline{y}_s$

and

(ii) $V[\overline{y}(s_o)] = V[E\{\overline{y}(s_o)|s\}] + E[V\{\overline{y}(s_o)|s\}] \ge V[E\{\overline{y}(s_o)|s\}] = V\left[\overline{y}_s\right]$

Theorem 3.3.4

Let s be an unordered sample derived by taking distinct units from an ordered sample s_o and $\overline{y}_s = \sum_{i \in s} y_i/\nu$, where ν is the number of distinct units of s_o. Then,

(i) Variance of $\overline{y}_s$ is

$$V(\overline{y}_s) = \left[E\left(\frac{1}{\nu}\right) - \frac{1}{N}\right] S_y^2$$

(ii) An unbiased estimator for $V(\overline{y}_s)$ is

$$\widehat{V}(\overline{y}_s) = \left(\frac{1}{\nu} - \frac{1}{N}\right) s_\nu^2$$

where $s_\nu^2 = \dfrac{1}{\nu - 1} \sum_{i \in s} (y_i - \overline{y}_s)^2$ is the sample variance of the set of distinct units of s_o.

Proof

Let $E(\overline{y}_s|\nu)$ and $V(\overline{y}_s|\nu)$ be the conditional expectation and conditional variance of $\overline{y}(s)$ over the variation of all possible unordered samples of size ν. Then, given ν, s is an SRSWOR sample of size ν from the entire population U. Hence

$$E(\overline{y}_s|\nu) = \overline{Y}, V(\overline{y}_s|\nu) = \left(\frac{1}{\nu} - \frac{1}{N}\right) S_y^2 \text{ and } E(s_\nu^2|\nu) = S_y^2 \qquad (3.3.1)$$

Using (3.3.1), we get

$$\begin{aligned} E(\overline{y}_s) &= E\{E(\overline{y}_s|\nu)\} = \overline{Y}, \\ V(\overline{y}_s) &= E\{V(\overline{y}_s|\nu)\} + V\{E(\overline{y}_s|\nu)\} \\ &= E\left(\frac{1}{\nu} - \frac{1}{N}\right) S_y^2 + V(\overline{Y}) \\ &= \left[E\left(\frac{1}{\nu}\right) - \frac{1}{N}\right] S_y^2 \end{aligned}$$

and

$$\begin{aligned} E[\widehat{V}(\overline{y}_s)] &= E\left[\left(\frac{1}{\nu} - \frac{1}{N}\right) E(s_\nu^2|\nu)\right] \\ &= \left[E\left(\frac{1}{\nu}\right) - \frac{1}{N}\right] S_y^2. \end{aligned}$$

Remark 3.3.2

(i) The exact probability distribution of ν was derived by Feller (1957) as

$$Prob\{\nu = j\} = \frac{1}{N^n}\binom{N}{j}\sum_{r=0}^{j}(-1)^r\binom{j}{r}(j-r)^n; j = 1, \ldots, \min(n, N) \tag{3.3.2}$$

(ii) The expression of $E\left(\frac{1}{\nu}\right)$ was derived by Pathak (1961) as

$$E\left(\frac{1}{\nu}\right) = \frac{1^{n-1} + 2^{n-1} + \cdots + N^{n-1}}{N^n} \tag{3.3.3}$$

(iii) Substituting (3.3.3) in Theorem 3.3.4, we get

$$V(\bar{y}_s) = \frac{1^{n-1} + 2^{n-1} + \cdots + (N-1)^{n-1}}{N^n} S_y^2$$

Neglecting terms of degree greater than $(1/N)^2$, Murthy (1967) obtained an approximate expression of $V(\bar{y}_s)$ as

$$V(\bar{y}_s) = \left(\frac{1}{n} - \frac{1}{2N} + \frac{n-1}{12N^2}\right) S_y^2$$

Theorem 3.3.5

Let ν_{s_o} be the number of distinct units in an ordered sample s_o of size n selected by the SRSWR method. Then,

(i) $E(\nu_{s_o}) = N[1 - \alpha]$ and

(ii) The variance of ν_{s_o} is

$$V(\nu_{s_o}) = N(N-1)[1 - 2\alpha + \beta] - \{N(1-\alpha)\}\{N(1-\alpha) - 1\}$$

where $\alpha = \left(1 - \frac{1}{N}\right)^n$ and $\beta = \left(1 - \frac{2}{N}\right)^n$.

Proof

The inclusion probability of the ith unit $= \pi_i =$ probability of selection of the ith unit in any of the n draws $= 1 -$ probability that the ith unit will not be selected from any of the n draws $= 1 - \left(1 - \frac{1}{N}\right)^n = 1 - \alpha$.

Inclusion probability of the ith and jth $(i \neq j)$ unit $= \pi_{ij} =$ probability of selection of both the ith and jth units in n draws $= 1$—at least one of the units i and j will not be selected in n draws

$$= 1 - \left\{ \left(1 - \frac{1}{N}\right)^n + \left(1 - \frac{1}{N}\right)^n - \left(1 - \frac{2}{N}\right)^n \right\} = 1 - 2\alpha + \beta.$$

Now using Theorem 1.3.1, we get $E(\nu_{s_o}) = \sum_{i=1}^{N} \pi_i = N(1-\alpha)$ and

$$V(\nu_{s_o}) = \sum_{i \neq}^{N} \sum_{j=1}^{N} \pi_{ij} - E(\nu_{s_o})\{E(\nu_{s_o} - 1)\}$$
$$= N(N-1)(1 - 2\alpha + \beta) - N(1-\alpha)\{N(1-\alpha) - 1\}.$$

3.4 INTERVAL ESTIMATION

In this section, we will present the formulae for determining confidence intervals for the population mean $\overline{Y}$ and population proportion π_A based on SRSWOR and SRSWR sampling schemes. In general, we cannot determine exact confidence intervals as no distribution of the parameter vector $\mathbf{y}$ is assumed. However, approximate confidence intervals are determined assuming that the central limit theorem holds when the sample size is large. The validity of the normality assumption and requirement of the sample size are given in detail by Cochran (1977) and Sukhatme et al. (1984). For smaller sample sizes, we use t distribution.

3.4.1 Confidence Intervals for Mean and Proportion

3.4.1.1 Large Sample Size

In case the sample size n is large ($n \geq 35$) and the population variance S_y^2 or σ_y^2 is known, $(1-\alpha)100\%$ confidence interval for $\overline{Y}$ is given by

$$\overline{y}(s) \pm z_{\alpha/2}\sqrt{V[\overline{y}(s)]} = \overline{y}(s) \pm z_{\alpha/2} S_y \sqrt{\frac{N-n}{Nn}} \quad \text{(for SRSWOR)}$$

and

$$\overline{y}(s_o) \pm z_{\alpha/2}\sqrt{V[\overline{y}(s_o)]} = \overline{y}(s_o) \pm z_{\alpha/2}\frac{\sigma_y}{\sqrt{n}} \quad \text{(for SRSWR)}$$

where $z_{\alpha/2}$ is the upper $\alpha/2$ point of standard normal distribution, S_y and σ_y denote the positive square root of S_y^2 and σ_y^2, respectively.

In most situations S_y^2 or σ_y^2 is unknown. In this case S_y^2 and σ_y^2 are replaced by their unbiased estimators, and we set $(1-\alpha)100\%$ interval for the population mean $\overline{Y}$ as

$$\overline{y}(s) \pm z_{\alpha/2}\sqrt{\widehat{V}[\overline{y}(s)]} = \overline{y}(s) \pm z_{\alpha/2} s_y \sqrt{\frac{N-n}{Nn}} \quad \text{(for SRSWOR)}$$

and

$$\overline{y}(s_o) \pm z_{\alpha/2}\sqrt{\widehat{V}[\overline{y}(s_o)]} = \overline{y}(s_o) \pm z_{\alpha/2}\frac{s_y}{\sqrt{n}} \quad \text{(for SRSWR)}$$

where s_y is the positive square root of s_y^2.

For a large n we set $(1-\alpha)100\%$ interval for the population proportion π_A as

$$\widehat{\pi}_A \pm z_{\alpha/2}\sqrt{\frac{N-n}{N}\cdot\frac{\widehat{\pi}_A(1-\widehat{\pi}_A)}{n-1}} \quad \text{(for SRSWOR)}$$

and

$$\widehat{\pi}_A \pm z_{\alpha/2}\sqrt{\frac{\widehat{\pi}_A(1-\widehat{\pi}_A)}{n-1}} \quad \text{(for SRSWR)}$$

3.4.1.2 Small Sample Size

For small sample size $n < 35$, $z_{\alpha/2}$ is replaced by $t_{\alpha/2,n-1}$ in all the formulae aforementioned, where $t_{\alpha/2,n-1}$ is the upper $(\alpha/2)100\%$ point of t distribution with degrees of freedom (df) $n-1$.

Example 3.4.1

For Example 3.2.1, $N = 30$, $n = 8$, and the sample s are selected by the SRSWOR method. Here the sample size $n = 8$ is small. Hence, a 95% confidence interval for the mean height of the students is $\overline{x}(s) \pm t_{0.025,7}s_x\sqrt{\frac{N-n}{Nn}} = 162.625 \text{ cm} \pm 2.365 \times 2.389 \text{ cm} = (156.975 \text{ cm}, \ 168.275 \text{ cm})$. Similarly for Example 3.2.1, a 90% confidence interval for the proportion of male students would be $\widehat{\pi}_m \pm t_{0.05,7}\sqrt{\frac{N-n}{N}\frac{\widehat{\pi}_m(1-\widehat{\pi}_m)}{n-1}} = 0.625 \pm 1.895 \times 0.156 = (0.329, 0.920)$.

Example 3.4.2

From a population of 10,000 adult males of a certain community, a sample of 500 males were selected by SRSWOR method, and 125 males were found to be HIV positive. Find a 95% confidence interval for the proportion of HIV-positive persons (π_A) in the community.

Here both the sample and population size are very large, and so we use the following formula for the confidence interval for π_A.

$$\widehat{\pi}_A \pm z_{\alpha/2}\sqrt{\frac{\widehat{\pi}_A(1-\widehat{\pi}_A)}{n}} = 0.25 \pm 1.96\sqrt{\frac{0.25(1-0.25)}{500}}$$
$$= (0.212, 0.288)$$

3.5 DETERMINATION OF SAMPLE SIZE

In any large- or small-scale survey, one has to determine the sample size, which will fulfill the objective of the survey. To get a representative sample, it is expected that the sample size should increase with the population size. In general, the sampling error decreases with the increase of sample size. However, the cost of a survey increases with the increase of sample size. The cost also increases with the increase of the size of the questionnaire because the investigator requires more time to complete a long questionnaire. The size of the sample is determined, considering available cost, time, and the degree of precision needed for estimation of parameters under consideration. In this section we will consider how one can determine an appropriate sample size for SRSWR and SRSWOR sampling designs.

3.5.1 Consideration of the Cost of a Survey

Suppose the cost of a survey is expressed by a simple cost function $C = C_o + c\,n$ where C_o is the overhead cost of the survey, which is fixed for a survey, i.e., independent of the sample size, and c is the average cost for surveying a unit. Hence for a fixed available cost $C = C^*$ of a survey, one should select sample size $n = (C^* - C_o)/c$.

3.5.2 Consideration of the Efficiency of Estimators

3.5.2.1 Given Variance

The magnitude of the sampling error of an estimator is determined by its variance. Hence, one can determine the sample size that yields some specific value of variance, V_0, for example. To estimate the population mean $\overline{Y}$, if one uses the sample mean as an estimator and expects that the sample mean will have a specific value of its variance V_0, then n can be derived from the relation

$$V\{\overline{y}(s)\} = \left(\frac{1}{n} - \frac{1}{N}\right)S_y^2 = V_0 \text{ i.e., } n = \left(\frac{V_0}{S_y^2} + \frac{1}{N}\right)^{-1} \text{(for SRSWOR)}$$

and

$$V\{\overline{y}(s_o)\} = \frac{1}{n}\sigma_y^2 = V_0 \text{ i.e., } n = \frac{\sigma_y^2}{V_0} \quad \text{(for SRSWR)}$$

To determine the value of n, one needs to know the value of S_y^2 $\left(\text{or } \sigma_y^2 = \frac{N-1}{N}S_y^2\right)$, which is generally unknown. Hence, it is customary to use a value of S_y^2 either from the past survey (or experience) or replace S_y^2 by its estimate s_y^2 determined from a pilot survey.

3.5.2.2 Given Coefficient of Variation

The coefficient of variation (CV) of a population characteristic y is defined by $C_y = \sigma_y / \overline{Y}$. It is found that the value of the CV of a population characteristic is much more stable (not subject to as much change over time) than the population variance or mean. Generally, the CV of a population is available from a past survey. In case it is not known, a reliable estimator of CV may be obtained from a pilot survey. The CV of an estimator t is defined by $C_t = \sqrt{Var(t)}/E(t)$. In case C_y is known, we can find the sample size n, which yields a specific value of CV of the estimator t, $C_t = C_{to}$ (say) as follows:

For SRSWOR with $t = \overline{y}(s)$, $C_{\overline{y}(s)} = C_{to}$ yields

$$C_{\overline{y}(s)} = \frac{\sqrt{\frac{N-n}{Nn}S_y^2}}{\overline{Y}} = \frac{\sqrt{\frac{N-n}{n(N-1)}\sigma_y^2}}{\overline{Y}} = \sqrt{\frac{N-n}{n(N-1)}}\,C_y = C_{to}, \text{ which}$$

in turn gives

$$n = \frac{N\{C_y\}^2}{(N-1)C_{to}^2 + \{C_y\}^2}$$

For SRSWR with $t = \overline{y}(s_o)$, $C_{\overline{y}(s_o)} = C_o$ yields the required sample size

$$n = \left(\frac{C_y}{C_{to}}\right)^2$$

3.5.2.3 Given Margin of Permissible Error

Here, sample size is determined by assigning the probability to a certain level, $(1-\alpha)$ (say) for the maximum permissible error (difference between the estimated and the true value of the parameter) to a certain value d. For

instance, let d be the permissible error, maximum acceptable difference between the estimator t, and the population mean $\overline{Y}$. The sample size is determined from the relation

$$Prob\left\{\left|t-\overline{Y}\right| \le d\right\} = 1-\alpha \tag{3.5.1}$$

The above equation is equivalent to

$$Prob\left\{\frac{\left|t-\overline{Y}\right|}{\sqrt{V(t)}} \le \frac{d}{\sqrt{V(t)}}\right\} = 1-\alpha$$

Assuming the sample size to be so large, enabling $z = \dfrac{\left|t-\overline{Y}\right|}{\sqrt{V(t)}}$ distributed $N(0,1)$, one can determine the value of n using the relation

$$\frac{d}{\sqrt{V(t)}} = z_{\alpha/2} \tag{3.5.2}$$

For an SRSWOR design with $t = \overline{y}(s)$, Eq. (3.5.1) yields

$$n = \left[\frac{1}{N} + \left(\frac{d}{S_y z_{\alpha/2}}\right)^2\right]^{-1} = \left[\frac{1}{N} + \frac{N-1}{N}\left(\frac{k}{C_y z_{\alpha/2}}\right)^2\right]^{-1} \tag{3.5.3}$$

where $k = d/\overline{Y}$

For an SRSWR design with $t = \overline{y}(s_o)$, Eq. (3.5.1) yields

$$n = \left(\frac{\sigma_y\, z_{\alpha/2}}{d}\right)^2 = \left(\frac{C_y\, z_{\alpha/2}}{k}\right)^2 \tag{3.5.4}$$

3.5.3 Use of Chebyshev Inequality

Here we determine sample size by keeping permissible error to a certain level with probability exceeding a certain preassigned value $(1-\alpha)$. Let t be an unbiased estimator for $\overline{Y}$. Using the Chebyshev Inequality, we have

$$Prob\left[\left|t-\overline{Y}\right| \le d\right] \ge 1 - \frac{V(t)}{d^2} \tag{3.5.5}$$

The sample size n is determined from the relation $1 - \dfrac{V(t)}{d^2} = 1-\alpha$ which is equivalent to

$$\frac{V(t)}{d^2} = \alpha \tag{3.5.6}$$

For an SRSWOR design with $t = \overline{y}(s)$, Eq. (3.5.5) yields

$$n = \left[\frac{1}{N} + \alpha\frac{N-1}{N}\left(\frac{d}{\sigma_y}\right)^2\right]^{-1} = N[1 + \gamma^2\alpha(N-1)]^{-1} \tag{3.5.7}$$

where $d = \gamma\ \ \sigma_y$

For an SRSWR design with $t = \overline{y}\,(s_o)$, Eq. (3.5.6) yields

$$n = \frac{\sigma_y^2}{\alpha d^2} = \frac{1}{\alpha\gamma^2} \tag{3.5.8}$$

Substituting $\alpha = 0.05$ and $\gamma = 1$ in (3.5.8), we get $n = 20$, i.e., selection of $n = 20$ ensures

$$Prob\left[\left|\overline{y}(s_o) - \overline{Y}\right| \le \sigma_y\right] \ge 0.95$$

Remark 3.5.1

Determination of sample size using Chebyshev Inequality does not require the assumption of the normality of estimator t used for estimating the parameter $\overline{Y}$.

Example 3.5.1

It is decided to estimate the mean consumption of food per household in a region of 1000 households. A pilot study reveals that the average consumption of food per family is \$2500 with a SE \$1250. Determine the minimum sample size required to ensure that the permissible error is 5% of the true value with 99% confidence coefficient under (i) SRSWOR and (ii) SRSWR procedures.

Here $N = 1000$, $\overline{Y} = \$2500$, $S_y = \$1250$, and $d = 0.05\ \overline{Y}$.

(i) For SRSWOR, using (3.5.3) we get the required sample size as

$$n = \left[\frac{1}{N} + \left(\frac{d}{S_y z_{\alpha/2}}\right)^2\right]^{-1} = \left[\frac{1}{1000} + \left(\frac{0.05 \times 2500}{1250 \times z_{.005}}\right)^2\right]^{-1}$$

$$= \left[\frac{1}{1000} + \left(\frac{0.05 \times 2500}{1250 \times 2.576}\right)^2\right]^{-1} = 399$$

(ii) For SRSWR, the formula (3.5.4) yields

$$n = \left(\frac{\sigma_y z_{\alpha/2}}{d}\right)^2 = \left(\frac{1250 \times 2.576}{0.05 \times 2.500}\right)^2 = 664$$

Here we assumed $\sigma_y = S_y$ as $N(=1000)$ is large.

Example 3.5.2

It is known from the past survey that the HIV infection, STD, and TB rates in a locality were 25%, 5%, and 15%, respectively. To find if there has been any improvement of health-care system to reduce the infection rates of HIV, STD, and TB, an SRSWR survey is proposed in a community consisting of 16,000 people. Determine the minimum sample size required, which ensures a permissible error of 10% of the true value with 95% confidence for measuring each of the infections.

Let π_x, π_y, and π_z be the population infection rates for HIV, STD, and TB, respectively and let $\widehat{\pi}_x$, $\widehat{\pi}_y$ and $\widehat{\pi}_z$ be the sample proportions.

Here we find n for which

$$Prob\{|\widehat{\pi}_t - \pi_t| \le d\} = 1 - \alpha \text{ for } t = x, y \text{ and } z \tag{3.5.9}$$

where $d = \pi_t \times 0.1$ and $1 - \alpha = 0.95$.

Assuming the population size $N(=16{,}000)$ is large, Eq. (3.5.9) yields

$$Prob\left\{|z| \le \frac{d}{\sqrt{V(\widehat{\pi}_t)}}\right\} = 1 - \alpha \tag{3.5.10}$$

where $z = (\widehat{\pi}_t - \pi_t)/\sqrt{V(\widehat{\pi}_t)}$ and $V(\widehat{\pi}_t) = \pi_t(1 - \pi_t)/n$

Now equating $\dfrac{d}{\sqrt{V(\widehat{\pi}_t)}} = z_{\alpha/2}$, we get

$$n = \left[\sqrt{\pi_t(1 - \pi_t)} z_{\alpha/2}/d\right]^2 = \left[\sqrt{\pi_t((1 - \pi_t)} \times 1.96/d\right]^2$$

Taking π_t as the previous infection rate, we determine the sample sizes as follows:

Infection	π_t	d	Required sample size n
HIV	0.25	0.025	1152.48
STD	0.05	0.005	7299.04
TB	0.15	0.015	2176.90

Considering the table above, we have to take the maximum sample size, which is 7299 to meet the requirement.

3.6 INVERSE SAMPLING

In estimating π, the proportion of units of a population possessing a rare characteristic A such as color blindness or having suffered from the hepatitis viral infection, the conventional sample proportion may not serve as an efficient estimator because very often the sample proportion becomes zero and hence its variance becomes very large. Haldane (1946) proposed an inverse sampling scheme, which is more effective and excludes zero as an estimator of π. In inverse sampling, units are selected one by one by the SRSWOR or SRSWR method until a specified number m of the units is selected from the rare population A.

3.6.1 Simple Random Sampling Without Replacement

Let us suppose that a population consists of N units of which $N_A(=N\pi)$ units possess certain rare characteristics A and that the remaining $N_B = N - N_A$ do not possess this characteristic. Let X be the number of units required to be drawn to get m units that possess characteristic A. Here X is a random variable whose probability distribution depends on m and N_A. The probability distribution of X is given by

$$P(X = x) = f(x|N_A, m)$$

$$= \text{Probability of getting } (m-1) \text{ units bearing characteristic } A \text{ in the first } x-1 \text{ draws and at the } x\text{th draw one unit is selected from the group } A.$$

$$= \frac{\binom{N_A}{m-1}\binom{N_B}{(x-1)-(m-1)}}{\binom{N}{x-1}} \frac{N_A-(m-1)}{N-(x-1)};$$

$$x \geq m, m+1, \ldots$$

Theorem 3.6.1

(i) An unbiased estimator of π is

$$\widehat{\pi} = \frac{m-1}{x-1}$$

(ii) An unbiased estimator for the variance of $\widehat{\pi}$ is

$$\widehat{V}(\widehat{\pi}) = \frac{\widehat{\pi}(1-\widehat{\pi})}{x-2}\left(1 - \frac{x-1}{N}\right)$$

Proof

(i) $$E(\widehat{\pi}) = \sum_{x \geq m} \frac{m-1}{x-1} f(x|N_A, m)$$

$$= \sum_{x \geq m} \frac{m-1}{x-1} \frac{\binom{N_A}{m-1}\binom{N_B}{(x-1)-(m-1)}}{\binom{N}{x-1}} \frac{N_A-(m-1)}{N-(x-1)}$$

$$= \frac{N_A}{N} \sum_{x \geq m} \frac{\binom{N_A-1}{m-2}\binom{N_B}{(x-2)-(m-2)}}{\binom{N-1}{x-2}}$$

$$\times \frac{(N_A-1)-(m-2)}{(N-1)-(x-2)}$$

$$= \pi \sum_{z \geq m-1} \frac{\binom{N_A-1}{m-2}\binom{N_B}{(z-1)-(m-2)}}{\binom{N-1}{z-1}}$$

$$\times \frac{(N_A-1)-(m-2)}{(N-1)-(z-1)}$$

$$= \pi \sum_{z \geq m-1} f(z|N_A-1, m-1)$$

$$= \pi \tag{3.6.1}$$

(ii) Following (3.6.1) we get

$$E\left(\frac{m-1}{x-1}\frac{m-2}{x-2}\right) = \frac{N_A}{N}\frac{N_A-1}{N-1} \sum_{z \geq m-2} f(z|N_A-2, m-2)$$

$$= \pi \frac{\pi N - 1}{N-1}$$

$$\text{i.e.,}\quad \frac{(N-1)}{N}E\left\{\frac{m-1}{x-1}\left(\frac{m-2}{x-2}+\frac{1}{N-1}\right)\right\}=\pi^2 \tag{3.6.2}$$

Finally using (3.6.2), an unbiased estimator for the variance of $\widehat{\pi}$ is obtained as

$$\begin{aligned}\widehat{V}(\widehat{\pi}) &= \widehat{\pi}^2 - \text{unbiased estimator of } \pi^2\\ &= \widehat{\pi}^2 - \frac{(N-1)}{N}\frac{m-1}{x-1}\left(\frac{m-2}{x-2}+\frac{1}{N-1}\right)\\ &= \widehat{\pi}\left[\widehat{\pi}-\frac{N-1}{N}\frac{\widehat{\pi}(x-1)-1}{x-2}-\frac{1}{N}\right]\\ &= \frac{\widehat{\pi}(1-\widehat{\pi})}{x-2}\left(1-\frac{x-1}{N}\right)\end{aligned}$$

3.6.2 Simple Random Sampling With Replacement

In inverse sampling with the SRSWR method, we select units one by one with replacement until m units possessing characteristic A are selected. Let X be the required number of draws. Then the probability distribution of X follows negative binomial distribution with parameter π and is given by

$$P(X=x)=f(x|\pi,m)=\binom{x-1}{m-1}\pi^x(1-\pi)^{x-m};x=m,m+1,\ldots$$

Here we get

$$\begin{aligned}E\left(\frac{m-1}{x-1}\right) &= \sum_{x\geq m}\frac{m-1}{x-1}f(x|\pi,m)\\ &= \sum_{x\geq m}\frac{m-1}{x-1}\binom{x-1}{m-1}\pi^x(1-\pi)^{x-m}\\ &= \pi\sum_{x\geq m}\binom{x-2}{m-2}\pi^{x-1}(1-\pi)^{(x-1)-(m-1)}\\ &= \pi\sum_{z\geq m-1}\binom{z-1}{m-1-1}\pi^z(1-\pi)^{z-(m-1)}\\ &= \pi\sum_{z\geq m-1}f(z|\pi,m-1)\\ &= \pi\end{aligned} \tag{3.6.3}$$

Similarly, we get

$$E\left(\frac{m-1}{x-1}\,\frac{m-2}{x-2}\right)=\pi^2 \tag{3.6.4}$$

The expressions (3.6.3) and (3.6.4) yield

Theorem 3.6.2

(i) An unbiased estimator of π is

$$\widehat{\pi}=\frac{m-1}{x-1}$$

(ii) An unbiased estimator for the variance of $\widehat{\pi}$ is

$$\widehat{V}(\widehat{\pi})=\frac{m-1}{x-1}\left(\frac{m-1}{x-1}-\frac{m-2}{x-2}\right)$$

$$=\frac{\widehat{\pi}(1-\widehat{\pi})}{x-2}$$

3.7 EXERCISES

3.7.1 The following table gives the weight of six students in a first-year class

Serial no of students	1	2	3	4	5	6
Weight (in kg)	55	60	75	77	60	55

(i) Write down all possible samples of size 3, which can be selected by the SRSWOR method.

(ii) Compute the sample means of all the possible samples and construct the frequency distribution of the sample mean.

(iii) From the frequency distribution of the sample mean, verify that the sample mean is an unbiased estimator of the population mean and the sample variance follows the formula given in Theorem 3.2.2.

(iv) Compute the sample variances for all the samples, make a frequency distribution of sample variance, and verify that the sample variance is an unbiased estimator of the population variance.

3.7.2 The ages of five school-going children are given in the following table.

Serial no of children	1	2	3	4	5
Age in years	5	6	5	4	6

(i) Write down all possible samples of size 2 that can be selected by SRSWR method.
(ii) Compute the frequency distribution of the sample means $\overline{y}(s_o)$ and sample variances.
(iii) Show that the sample mean and sample variances are unbiased estimators of the population mean and population variance σ_y^2 respectively.
(iii) Find the frequency distribution of sample means of distinct units $\overline{y}(s)$ and compute the variance of $\overline{y}(s)$.
(iv) Find the frequency distribution of the distinct units $\nu(s)$ and compute (i) $E\{\nu(s)\}$, (ii) $V\{\nu(s)\}$, and (iii) $E\{1/\nu(s)\}$.

3.7.3 The following table gives the marks of Mathematics and Statistics of 30 students by gender in the first-year class of the University of Botswana.

Serial no of students	Marks on Mathematics	Marks on Statistics	Gender	Serial no of students	Marks on Mathematics	Marks on Statistics	Gender
1	50	53	1	16	36	60	1
2	90	81	1	17	71	55	0
3	61	85	1	18	72	57	1
4	38	68	0	19	35	82	0
5	45	73	0	20	62	62	1
6	67	62	1	21	81	20	1
7	40	34	1	22	74	55	0
8	50	49	1	23	88	32	0
9	89	59	1	24	55	59	1
10	77	84	1	25	69	61	1
11	67	23	0	26	32	66	1
12	82	73	0	27	37	74	1
13	48	61	1	28	66	92	1
14	71	37	0	29	34	68	0
15	42	35	1	30	35	71	1

0, female; *1*, male.

(i) Select a sample of 10 students by SRSWR method and give unbiased estimates of the average marks in Mathematics and Statistics and also estimate their variances.
(ii) Compute an unbiased estimate of the difference between the average marks in Mathematics and Statistics with its SE.
(iii) Give an estimate of the population CV of Mathematics marks.
(iv) Estimate the population proportion of female students and estimate its SE. Compute a 95% confidence interval for the proportion of male students in the class.
(v) Give unbiased estimates of the average marks for Mathematics and Statistics of the male students and give unbiased estimates of their variances.
(vi) Find a 90% confidence interval for the difference between the average marks for Mathematics and Statistics of the entire class.

3.7.4 From a list of 100 households, 10 households were selected by the SRSWOR method and the following information was obtained.

Household	Head of household	Household income (in $)	Household size	Education	Expenditure on transport (in $)
1	M	2000	4	Primary	200
2	M	3000	2	Primary	250
3	F	4500	5	Middle	600
4	M	8000	3	High	500
5	F	2000	2	Primary	100
6	F	5000	4	Middle	150
7	M	7500	5	High	300
8	M	4000	3	High	250
9	F	5000	4	Middle	200
10	F	6000	2	High	300

M, Male, *F*, Female.

(i) Estimate the average (a) household income and (b) expenditure on transport and calculate unbiased estimates of the variances of the estimators used.
(ii) Estimate the proportion of people having primary, middle and high education. Give unbiased estimates for the variance–covariance matrix of the proposed estimators.
(iii) Estimate 95% confidence intervals for the population proportion of male-headed and female-headed households.

(iv) Give unbiased estimates of the total household income for the income groups 0—4000, 4001—6000, and 6000, respectively. Also give unbiased estimates of the average income for these income groups when the proportions of persons belonging to these income groups are known, which are 0.4, 0.3, and 0.3, respectively.

(v) Estimate the average income of the female-headed household for the income group 0—5000. Is your estimator biased or unbiased? Explain why.

(vi) Estimate the population correlation coefficient between the household income and household size? Is the estimator unbiased?

3.7.5 From the student population given in 3.7.3, select a sample of size 5 by the SRSWOR method and estimate the following:

(i) The average marks for Mathematics and Statistics and compute their SEs.

(ii) Estimate the proportion of male students and obtain a 95% confidence interval for the proportion.

(iii) Estimate the difference between the average marks for Mathematics and Statistics of the female students and obtain a 90% confidence interval for the difference of the mean marks.

3.7.6 Suppose samples s^* and s of sizes m and n are selected from the entire population U of size N and from $U - s^*$ of size $N - m$ by SRSWOR method. Find the optimum value of w for which the variance of the weighted estimator $\overline{y}_w = w\overline{y}_{s^*} + (1 - w)\overline{y}_s$ of the population mean attains a minimum value.

3.7.7 A sample of 70 migrant workers is selected from a population of 500 migrant workers and they are classified into the following age groups:

Age group	<15	16—50	51—65	>65
No of workers	15	20	30	5
Average daily wages in $	30	50	55	40

(i) Estimate the average daily wage for all the migrant workers (μ) and give the SE of your estimator. Obtain a 95% confidence interval for μ.

(ii) Estimate the proportion of workers in each age group and estimate the variance and covariance matrix of the estimators.

(iii) Estimate the proportion of workers of age less than or equal to 50. Find a 90% confidence interval of this proportion.

3.7.8 Let s be a sample of size n selected from a population U by the SRSWR method and s_m be the set of distinct units in s. Suppose a sample s_u of size u is selected from $U - s_m$ by the SRSWOR method and $\overline{y}_u$ is the sample mean of s_u. Show that $\overline{y}_u$ is an unbiased estimator of the mean of the population U. Derive the variance of $\overline{y}_u$.

3.7.9 Let s and s_m be the same as in 3.7.8 but s'_u be a subsample of size u selected from s_m by the SRSWOR method and $\overline{y}'_u$ be the sample mean of s'_u. Show that $\overline{y}'_u$ is an unbiased estimator of the population mean. Derive the variance of $\overline{y}'_u$.

3.7.10 A survey is conducted to find the prevalence of HIV/AIDS infection in a certain community consisting of 150 residents. Given that an estimate of the prevalence obtained from the pilot study is 25%, determine the sample size required so that the permissible error is 1% of the true value with 95% confidence coefficient if the sample is selected by (i) SRSWOR and (ii) SRSWR methods.

3.7.11 It is known from the past survey that the CV of household income is 15%. Determine the required sample size so that the CV of the estimate of the household mean income is less than 10% when the sample is selected from a population of 500 households by the (i) SRSWOR and (ii) SRSWR methods.

3.7.12 Estimate the minimum sample size required to estimate the population mean household expenditure on food to ensure that the error of an estimator is less than 1% of the true value of the population mean with probability 0.99, when the sample is selected from a population of 500 by the (i) SRSWR and (ii) SRSWOR methods, given that the sample mean and the sample standard deviation of the expenditure on food were obtained as \$4000 and \$625, respectively, from the pilot survey.