SIMPLE RANDOM SAMPLING

A procedure for selecting a sample of size n out of a finite population of size N in which each of the possible distinct samples has an equal chance of being selected is called **random sampling or simple random sampling.**

We may have two distinct types of simple random sampling as follows:

i) Simple random sampling with replacement $(srswr)$.

ii) Simple random sampling without replacement $(srswor)$.

Simple random sampling with replacement $(srswr)$

In sampling with replacement a unit is selected from the population consisting of N units, its content noted and then returned to the population before the next draw is made, and the process is repeated n times to give a sample of n units. In this method, at each draw, each of the N units of the population gets the same probability $\frac{1}{N}$ of being selected. Here the same unit of the population may occur more than once in the sample (order in which the sample units are obtained is regarded). There are N^n samples, and each has an equal probability $\frac{1}{N^n}$ of being selected.

Note: If order in which the sample units are obtained is ignored (unordered), then in such case the number of possible samples will be

$$^NC_n + N(1 + {}^{N-1}C_1 + {}^{N-1}C_2 + \cdots + {}^{N-1}C_{n-2}).$$

Simple random sampling without replacement $(srswor)$

Suppose the population consist of N units, then, in simple random sampling without replacement a unit is selected, its content noted and the unit is not returned to the population before next draw is made. The process is repeated n times to give a sample of n units. In this method at the $r-$th drawing, each of the $N-r+1$ units of the population gets the same probability $\frac{1}{N-r+1}$ of being included in the sample. Here any unit of the population cannot occur more than once in the sample (order is ignored). There are ^NC_n possible samples, and each such sample has an equal probability $\frac{1}{^NC_n}$ of being selected.

Example: For a population of size $N=5$ with values 1, 3, 6, 8 and 9 make list of all possible samples of size $n=3$ by both the methods [$srswr$ (unordered) and $srswor$].

Solution: By the sampling wr, the number of possible samples will be

$$^NC_n + N(1 + {}^{N-1}C_1 + \cdots + {}^{N-1}C_{n-2}) = {}^5C_3 + 5(1 + {}^4C_1) = 35,$$ which are as follows:

(1, 1, 1), (1, 1, 3), (1, 1, 6), (1, 1, 8), (1, 1, 9), (1, 3, 3), (1, 3, 6), (1, 3, 8), (1, 3, 9), (1, 6, 6), (1, 6, 8), (1, 6, 9), (1, 8, 8), (1, 8, 9), (1, 9, 9), (3, 3, 3), (3, 3, 6), (3, 3, 8), (3, 3, 9), (3, 6, 6), (3, 6, 8), (3, 6, 9), (3, 8, 8), (3, 8, 9), (3, 9, 9), (6, 6, 6), (6, 6, 8),(6, 6, 9), (6, 8, 8), (6, 8, 9), (6, 9, 9), (8, 8, 8), (8, 8, 9), (8, 9, 9), (9, 9, 9).

By the sampling wor, the number of possible samples will be ${}^N C_n = {}^5 C_3 = 10$, which are as follows:

(1, 3, 6), (1, 3, 8), (1, 3, 9), (1, 6, 8), (1, 6, 9), (1, 8, 9), (3, 6, 8), (3, 6, 9), (3, 8, 9), (6, 8, 9).

Theory of simple random sampling with replacement

N, population size.

n, sample size.

Y_i, value of the $i-$th unit of the population.

y_i, value of the $i-$th unit of the sample.

$Y = \sum_{i=1}^{N} Y_i$, population total.

$\bar{Y} = \frac{1}{N}\sum_{i=1}^{N} Y_i$, population mean.

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

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

$S^2 = \frac{1}{N-1}\sum_{i=1}^{N}(Y_i - \bar{Y})^2 = \frac{1}{N-1}\left(\sum_{i=1}^{N} Y_i^2 - N\bar{Y}^2\right)$, population mean square.

$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} y_i^2 - n\bar{y}^2\right)$, sample mean square.

Theorem: In $srswr$, the sample mean $\bar{y}$ is an unbiased estimate of the population mean $\bar{Y}$ i.e. $E(\bar{y}) = \bar{Y}$ and its variance $V(\bar{y}) = \frac{N-1}{nN} S^2 = \frac{\sigma^2}{n}$.

Proof: It is immediately seen that

$E(\bar{y}) = E\left(\frac{1}{n}\sum_{i=1}^{n} y_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(y_i)$. By definition,

$E(y_i) = \sum_{i=1}^{N} Y_i \Pr(y_i = Y_i) = \frac{1}{N}\sum_{i=1}^{N} Y_i = \bar{Y}$, since y_i can take any one of the values $Y_1, \cdots, Y_N$ each with probability $1/N$.

Therefore,

$$E(\bar{y}) = \frac{1}{n}\sum_{i=1}^{n} \bar{Y} = \bar{Y}.$$

To obtain the variance, we have

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

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

Justification of the above result can see by taking particular case, i.e. as

$\left[\sum_{i=1}^{n}\{y_i - E(y_i)\}\right]^2 = \left(\sum_{i=1}^{n} a_i\right)^2 = (a_1+a_2+\ldots+a_n)^2$. Put $n=3$, then,

$$(a_1+a_2+a_3)^2 = a_1^2+a_2^2+a_3^2+a_1a_2+a_1a_3+a_2a_1+a_2a_3+a_3a_1+a_3a_2 = \sum_{i=1}^{3}a_i^2 + \sum_{\substack{i,j\\ i\neq j}}^{3}a_ia_j\ .$$

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

$$= \frac{1}{n^2}\sum_{i=1}^{n}V(y_i) + \frac{1}{n^2}\sum_{\substack{i,j\\ i\neq j}}^{n}Cov(y_i,y_j) \qquad (2.1)$$

Consider

$$V(y_i) = E(y_i-\bar{Y})^2 = \sum_{i=1}^{N}(Y_i-\bar{Y})^2\ \Pr(y_i=Y_i)$$

$= \frac{1}{N}\sum_{i=1}^{N}(Y_i-\bar{Y})^2$, since y_i can take any one of the values $Y_1,\cdots,Y_N$ each with probability $1/N$.

$$= \sigma^2 = \frac{N-1}{N}S^2,\ \text{since}\ S^2 = \frac{1}{N-1}\sum_{i=1}^{N}(Y_i-\bar{Y})^2 \qquad (2.2)$$

and

$$Cov\ (y_i,y_j) = E[(y_i-\bar{Y})(y_j-\bar{Y})] = \sum_{i,j}^{N}(Y_i-\bar{Y})(Y_j-\bar{Y})\Pr(y_i=Y_i,y_j=Y_j)\ .$$

In this case y_j can take any one of the values $Y_1,\cdots,Y_N$ with probability $1/N$ irrespective of the values taken by y_i, because old composition of the population remain the same throughout the sampling process due to the sampling with replacement. In other words for $i\neq j$, y_i and y_j are independent, so that

$$\Pr(y_i = Y_i, y_j = Y_j) = \Pr(y_i = Y_i)\Pr(y_j = Y_j) = \frac{1}{N}\times\frac{1}{N} = \frac{1}{N^2}.$$

Hence,

$$Cov(y_i, y_j) = \frac{1}{N^2}\sum_{i,j}^{N}(Y_i - \bar{Y})\,(Y_j - \bar{Y}) = \frac{1}{N^2}\sum_{i=1}^{N}(Y_i - \bar{Y})\sum_{j=1}^{N}(Y_j - \bar{Y}) = 0. \qquad (2.3)$$

Substitute the values of equations (2.2) and (2.3) in equation (2.1), we get

$$V(\bar{y}) = \frac{1}{n^2}\sum_{i=1}^{n}\frac{N-1}{N}S^2 = \frac{N-1}{nN}S^2 = \frac{\sigma^2}{n}.$$

Corollary: $\hat{Y} = N\,\bar{y}$ is an unbiased estimate of the population total Y with its variance $V(\hat{Y}) = \frac{N^2\sigma^2}{n} = \frac{N(N-1)}{n}S^2$.

Proof: By definition,

$$E(\hat{Y}) = E(N\,\bar{y}) = N\,E(\bar{y}) = N\,\bar{Y} = N\frac{1}{N}\sum_{i=1}^{N}Y_i = Y$$

and $V(\hat{Y}) = V(N\,\bar{y}) = N^2\,V(\bar{y}) = \frac{N^2\sigma^2}{n} = \frac{N(N-1)}{n}S^2$.

Remarks:

i) The standard error (SE) of $\bar{y}$ is $SE(\bar{y}) = \sqrt{V(\bar{y})} = \frac{\sigma}{\sqrt{n}} = S\sqrt{\frac{N-1}{nN}}$.

ii) The standard error $\hat{Y}$ is $SE(\hat{Y}) = \sqrt{V(\hat{Y})} = \frac{N\sigma}{\sqrt{n}} = S\sqrt{\frac{N(N-1)}{n}}$.

Theorem: In *srswr*, sample mean square s^2 is an unbiased estimate of the population variance σ^2 i.e. $E(s^2) = \sigma^2$.

Proof: By definition

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

To obtain $E(y_i^2)$ and $E(\bar{y}^2)$, note that

$V(y_i) = E(y_i^2) - \bar{Y}^2$, so that

$E(y_i^2) = \sigma^2 + \bar{Y}^2$, since $V(y_i) = (N-1)S^2/N = \sigma^2$.

and

$V(\bar{y}) = E(\bar{y}^2) - \bar{Y}^2$, so that

$$E(\bar{y}^2)=\frac{\sigma^2}{n}+\bar{Y}^2, \text{ since } V(\bar{y})=\left(\frac{N-1}{nN}\right)S^2=\frac{\sigma^2}{n}, \text{ for } srswr.$$

Therefore,

$$E(s^2)=\frac{1}{n-1}\left[\sum_{i=1}^{n}(\sigma^2+\bar{Y}^2)-n\left(\frac{\sigma^2}{n}+\bar{Y}^2\right)\right]=\sigma^2=\left(\frac{N-1}{N}\right)S^2.$$

Example: In a population with $N=5$, the values of Y_i are 8, 3, 11, 4 and 7.

a) Calculate population mean $\bar{Y}$, variance σ^2 and mean sum square S^2.

b) Enumerate all possible samples of size 2 by the replacement method and verify that

i) Sample mean $\bar{y}$ is unbiased estimate of population mean $\bar{Y}$ i.e. $E(\bar{y})=\bar{Y}$.

ii) $N\bar{y}$ is unbiased estimate of population total Y i.e. $E(N\bar{y})=Y$.

iii) $V(\bar{y})=\dfrac{(N-1)S^2}{nN}=\dfrac{\sigma^2}{n}$, and

iv) $E(s^2)=\left(\dfrac{N-1}{N}\right)S^2=\sigma^2$.

Solution:

a) We know that

$$\bar{Y}=\frac{1}{N}\sum_{i=1}^{N}Y_i=6.6,\ \sigma^2=\frac{1}{N}\sum_{i=1}^{N}Y_i^2-\bar{Y}^2=8.24 \text{ and } S^2=\frac{1}{N-1}\left(\sum_{i=1}^{N}Y_i^2-N\bar{Y}^2\right)=10.3.$$

b) Form a table for calculation as below:

Samples	$\bar{y}_i$	$\bar{y}_i^2$	$N\bar{y}_i$	s_i^2	Samples	$\bar{y}_i$	$\bar{y}_i^2$	$N\bar{y}_i$	s_i^2
(8, 8)	8.0	64.00	40.0	0.0	(11, 4)	7.5	56.25	37.5	24.5
(8, 3)	5.5	30.25	27.5	12.5	(11, 7)	9.0	81.00	45.0	8.0
(8, 11)	9.5	90.25	47.5	4.5	(4, 8)	6.0	36.00	30.0	8.0
(8, 4)	6.0	36.00	30.0	8.0	(4, 3)	3.5	12.25	17.5	0.5
(8, 7)	7.5	56.25	37.5	0.5	(4, 11)	7.5	56.25	37.5	24.5
(3, 8)	5.5	30.25	27.5	12.5	(4, 4)	4.0	16.00	20.0	0.0
(3, 3)	3.0	9.00	15.0	0.0	(4, 7)	5.5	30.25	27.5	4.5
(3, 11)	7.0	49.00	35.0	32.0	(7, 8)	7.5	56.25	37.5	0.5
(3, 4)	3.5	12.25	17.5	0.5	(7, 3)	5.0	25.00	25.0	8.0
(3, 7)	5.0	25.00	25.0	8.0	(7, 11)	9.0	81.00	45.0	8.0
(11, 8)	9.5	90.25	47.5	4.5	(7, 4)	5.5	30.25	27.5	4.5
(11, 3)	7.0	49.00	35.0	32.0	(7, 7)	7.0	49.00	35.0	0.0
(11, 11)	11.0	121.00	55.0	0.0					

i) $E(\bar{y})=\frac{1}{n'}\sum_{i=1}^{n'}\bar{y}_i=\frac{1}{25}\times 165=6.6=\bar{Y}$, where n' is the number of sample.

ii) $E(N\,\bar{y})=\frac{1}{n'}\sum_{i=1}^{n'}N\,\bar{y}_i=33$ or $E(N\,\bar{y})=N\,E(\bar{y})=33$.

iii) $V(\bar{y})=\frac{1}{n'}\sum_{i=1}^{n'}\bar{y}_i^2-\bar{Y}^2=4.12$.

Now,

$$\frac{(N-1)\,S^2}{nN}=4.12\text{, and }\quad \frac{\sigma^2}{n}=4.12\text{, therefore,}$$

$$V(\bar{y})=\frac{(n-1)\,S^2}{nN}=\frac{\sigma^2}{n}=4.12.$$

iv) $E(s^2)=\frac{1}{n'}\sum_{i=1}^{n'}s_i^2=\frac{1}{25}\times 206=8.24$ (1a)

and $\frac{(N-1)\,S^2}{N}=8.24$ (2a)

In view of equation (1a) and (2a), we get

$$E(s^2)=\frac{(N-1)\,S^2}{N}=\sigma^2=8.24.$$

Theory of simple random sampling without replacement

Theorem: In *srswor*, sample mean $\bar{y}$ is an unbiased estimate of the population mean $\bar{Y}$ i.e. $E(\bar{y})=\bar{Y}$ and its variance is $V(\bar{y})=\left(\frac{N-n}{nN}\right)S^2$.

Proof: As in *srswr*,

$$E(\bar{y})=\bar{Y}\text{, and}\,V(\bar{y})=\frac{1}{n^2}\sum_{i=1}^{n}V(y_i)+\frac{1}{n^2}\sum_{\substack{i,j\\ i\neq j}}^{n}Cov(y_i,y_j), \tag{2.4}$$

where $V(y_i)=\frac{N-1}{N}S^2$, for each i. (2.5)

Consider

$$Cov(y_i,y_j)=E[(y_i-\bar{Y})(y_j-\bar{Y})]=\sum_{i,j}^{N}(Y_i-\bar{Y})(Y_j-\bar{Y})\Pr(y_i=Y_i,y_j=Y_j).$$

In this case y_j can take any one of the values except Y_i, the value which is known to have already been assumed by y_i, with equal probability $\frac{1}{N-1}$, so that for $i\neq j$,

$$\Pr(y_i = Y_i, y_j = Y_j) = \Pr(y_i = Y_i)\Pr(y_j = Y_j \mid y_i = Y_i) = \frac{1}{N} \times \frac{1}{N-1}.$$

Hence,

$$Cov(y_i, y_j) = \frac{1}{N(N-1)} \sum_{i,j}^{N} (Y_i - \bar{Y})\ (Y_j - \bar{Y})$$

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

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

$$= -\frac{1}{N(N-1)} \sum_{i=1}^{N} (Y_i - \bar{Y})^2 = -\frac{S^2}{N} \tag{2.6}$$

Substitute the values of equations (2.5) and (2.6) in equation (2.4), we get

$$V(\bar{y}) = \frac{1}{n^2}\ n \left(\frac{(N-1)S^2}{N} \right) + \frac{1}{n^2}\ n(n-1) \left(-\frac{S^2}{N} \right) = \frac{(N-1)}{nN} S^2 - \frac{n-1}{n\,N} S^2$$

$$= \left(\frac{N-n}{nN} \right) S^2 = \left(1 - \frac{n}{N} \right) \frac{S^2}{n} = (1-f) \frac{S^2}{n},$$

where $f = \frac{n}{N}$ is called the sampling fraction and the factor $(1-f)$ is called the finite population correction (fpc). If the population size N is very large or if n is small corresponding with N, then $f = \frac{n}{N} \to 0$ and consequently $fpc \to 1$.

Alternative expression

$$V(\bar{y}) = \left(\frac{N-n}{nN} \right) S^2 = \left(\frac{1}{n} - \frac{1}{N} \right) S^2.$$

Corollary: $\hat{Y} = N\,\bar{y}$ is an unbiased estimate of the population total Y with its variance $V(\hat{Y}) = N^2 (1-f) S^2 / n$.

Proof:

By definition,

$$E(\hat{Y}) = E(N\,\bar{y}) = N\,E(\bar{y}) = N\,\bar{Y} = N\,\frac{1}{N} \sum_{i=1}^{N} Y_i = Y$$

and

$$V(\hat{Y}) = V(N\,\bar{y}) = N^2 \left(\frac{N-n}{nN} \right) S^2 = N^2 (1-f)\ \frac{S^2}{n}.$$

Remarks

i) The standard error of $\bar{y}$ is $SE(\bar{y}) = S\sqrt{\frac{N-n}{nN}} = S\sqrt{\frac{1-f}{n}} = S\sqrt{\left(\frac{1}{n}-\frac{1}{N}\right)}$.

ii) The standard error $\hat{Y}$ is $SE(\hat{Y}) = N\,S\sqrt{\frac{N-n}{nN}} = N\,S\sqrt{\frac{1-f}{n}} = N\,S\sqrt{\left(\frac{1}{n}-\frac{1}{N}\right)}$.

For large population $fpc = (1-f) \to 1$, then

i) $V(\bar{y}) = \frac{S^2}{n}$, and $SE(\bar{y}) = \frac{S}{\sqrt{n}}$.

ii) $V(\hat{Y}) = \frac{N^2 S^2}{n}$, and $SE(\hat{Y}) = \frac{N\,S}{\sqrt{n}}$.

Theorem: In *srswor*, sample mean square s^2 is an unbiased estimate of the population mean square S^2 i.e. $E(s^2) = S^2$.

Proof: By definition,

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

To obtain $E(y_i^2)$ and $E(\bar{y}^2)$, note that

$$V(y_i) = E(y_i^2) - \bar{Y}^2, \text{ so that}$$

$$E(y_i^2) = \frac{N-1}{N}S^2 + \bar{Y}^2, \text{ since } \quad V(y_i) = (N-1)S^2/N.$$

and $V(\bar{y}) = E(\bar{y}^2) - \bar{Y}^2$, so that

$$E(\bar{y}^2) = \left(\frac{N-n}{nN}\right)S^2 + \bar{Y}^2, \text{ since } V(\bar{y}) = \left(\frac{N-n}{nN}\right)S^2, \text{ for } srswr.$$

Therefore,

$$E(s^2) = \frac{1}{n-1}\left[\sum_{i=1}^{n}\left(\frac{N-1}{N}S^2 + \bar{Y}^2\right) - n\left(\frac{N-n}{nN}S^2 + \bar{Y}^2\right)\right]$$

$$= \frac{1}{n-1}[n(N-1) - (N-n)]\,\frac{S^2}{N} = \frac{1}{n-1}(n-1)\,N\,\frac{S^2}{N} = S^2.$$

Example: A random sample of $n = 2$ households was drawn from a small colony of $N = 5$ households having monthly income (in rupees) as follows:

Households:	1	2	3	4	5
Income (in thousand rupees):	8	6.5	7.5	7	6

a) Calculate population mean $\bar{Y}$, variance σ^2 and mean sum square S^2.

b) Enumerate all possible samples of size $n = 2$ by the without replacement method and verify that

i) Sample mean $\bar{y}$ is unbiased estimate of population mean $\bar{Y}$ i.e. $E(\bar{y}) = \bar{Y}$.

ii) $N\,\bar{y}$ is unbiased estimate of population total Y i.e. $E(N\,\bar{y}) = Y$.

iii) $V(\bar{y}) = \dfrac{(N-n)\,S^2}{nN}$, and

iv) $E(s^2) = S^2$.

Solution:

a) We know that

$$\bar{Y} = \frac{1}{N}\sum_{i-1}^{N} Y_i = 7\,,\quad \sigma^2 = \frac{1}{N}\sum_{i=1}^{N} Y_i^2 - \bar{Y}^2 = 0.5\,, \text{ and } S^2 = \frac{1}{N-1}\left(\sum_{i=1}^{N} Y_i^2 - N\,\bar{Y}^2\right) = 0.625\,.$$

b) Form a table for calculation as below:

Samples	$\bar{y}_i$	$\bar{y}_i^2$	$N\,\bar{y}_i$	s_i^2	Samples	$\bar{y}_i$	$\bar{y}_i^2$	$N\,\bar{y}_i$	s_i^2
(8, 6.5)	7.25	52.563	36.25	1.125	(8, 7.5)	7.75	60.063	38.75	0.125
(8, 7)	7.50	56.250	37.50	0.500	(8, 6)	7.00	49.000	35.00	2.000
(6.5, 7.5)	7.00	49.000	35.00	0.500	(6.5, 7)	6.75	45.563	33.75	0.125
(6.5, 6)	6.25	39.063	31.25	0.125	(7.5, 7)	7.25	52.563	36.25	0.125
(7.5, 6)	6.75	45.563	33.75	1.125	(7, 6)	6.50	42.250	32.50	0.500

i) $E(\bar{y}) = \dfrac{1}{n'}\displaystyle\sum_{i=1}^{n'} \bar{y}_i = 7 = \bar{Y}$, where n' is the number of sample.

ii) $E(N\,\bar{y}) = \dfrac{1}{n'}\displaystyle\sum_{i=1}^{n'} N\,\bar{y}_i = 35\,,$ or $E(N\,\bar{y}) = N\,E(\bar{y}) = 35\,.$

iii) $V(\bar{y}) = \dfrac{1}{n'}\displaystyle\sum_{i=1}^{n'} (\bar{y}_i - \bar{Y})^2 = \dfrac{1}{n'}\displaystyle\sum_{i=1}^{n'} \bar{y}_i^2 - \bar{Y}^2 = 0.1875$, and $\dfrac{(N-n)\,S^2}{nN} = 0.1875\,.$

Therefore,

$$V(\bar{y}) = \frac{(N-n)\,S^2}{nN} = 0.1875\,.$$

iv) $E(s^2) = \dfrac{1}{n'}\displaystyle\sum_{i=1}^{n'} s_i^2 = 0.625 = S^2$.

Property: $V(\bar{y})$ under *srswor* is less than the $V(\bar{y})$ under *srswr*.

Proof:

Under *srswor*, $$V(\bar{y}) = \frac{N-n}{nN}\,S^2 \tag{2.7}$$

and under *srswr*, $V(\bar{y})=\frac{\sigma^2}{n}=\frac{N-1}{nN}S^2$ (2.8)

Comparing (2.7) and (2.8), we note that $(N-1)>(N-n)$, which is always the case $\frac{N-1}{nN}S^2>\frac{N-n}{nN}S^2$.

Example: In a population $N=5$, the values are 2, 4, 6, 8 and 10, then for a *srs* size $n=3$, show that $V(\bar{y})_{srswor}<V(\bar{y})_{srswr}$.

Solution: We know that

$$V(\bar{y})_{srswor}=\frac{N-n}{nN}S^2, \text{ and } V(\bar{y})_{srswr}=\frac{N-1}{nN}S^2,$$

where, $S^2=\frac{1}{N-1}\sum_{i=1}^{N}(Y_i-\bar{Y})^2=10$ and $\bar{Y}=\frac{1}{N}\sum_{i=1}^{N}Y_i=6$.

Thus,

$$V(\bar{y})_{srswor}=\frac{4}{3},\ V(\bar{y})_{srswr}=\frac{8}{3}, \text{ and therefore,}$$

$$V(\bar{y})_{srswor}<V(\bar{y})_{srswr}.$$

Theorem: Let *srswor* sample of size n is drawn from a population of size N. Let $T=\sum_{i=1}^{n}\alpha_i y_i$ is a class of linear estimator of $\bar{Y}$, where $\alpha_i's$ are coefficient attached to sample values, then,

i) The class T is linear unbiased estimate class if $\sum_{i=1}^{n}\alpha_i=1$.

ii) The sample mean $\bar{y}$ is the best linear unbiased estimate.

Proof:

i) $E(T)=E\left(\sum_{i=1}^{n}\alpha_i y_i\right)=\sum_{i=1}^{n}\alpha_i\,E(y_i)=\sum_{i=1}^{n}\alpha_i\bar{Y}=\bar{Y}$, iff $\sum_{i=1}^{n}\alpha_i=1$.

ii) $V(T)=E\left(\sum_{i=1}^{n}\alpha_i y_i-\bar{Y}\right)^2$, under $\sum_{i=1}^{n}\alpha_i=1$.

$$=E\left[\left(\sum_{i=1}^{n}\alpha_i y_i\right)^2-2\bar{Y}\left(\sum_{i=1}^{n}\alpha_i y_i\right)+\bar{Y}^2\right]=E\left(\sum_{i=1}^{n}\alpha_i y_i\right)^2-\bar{Y}^2.$$

Consider,

$$E\left(\sum_{i=1}^{n}\alpha_i y_i\right)^2=E\left(\sum_{i=1}^{n}\alpha_i^2 y_i^2+\sum_{i\neq j}^{n}\alpha_i\alpha_j\,y_i y_j\right)=\sum_{i=1}^{n}\alpha_i^2\,E(y_i^2)+\sum_{i\neq j}^{n}\alpha_i\alpha_j\,E(y_i y_j)$$

Now

$$E(y_i^2) = \frac{1}{N}\sum_{i=1}^{N} y_i^2 \text{ , note that}$$

$$(N-1)S^2 = \sum_{i=1}^{N}(y_i - \bar{Y})^2 = \sum_{i=1}^{N} y_i^2 - N\bar{Y}^2 \quad \text{or} \quad \sum_{i=1}^{N} y_i^2 = (N-1)S^2 + N\bar{Y}^2 .$$

Thus,

$$E(y_i^2) = \frac{1}{N}(N-1)S^2 + \bar{Y}^2$$

and $\displaystyle E(y_i y_j) = \sum_{i\neq j}^{N} y_i \Pr(i)\, y_j \Pr(j|i) = \frac{1}{N}\frac{1}{N-1}\sum_{i\neq j}^{N} y_i\, y_j$.

Note that

$$\left(\sum_{i=1}^{N} y_i\right)^2 = \sum_{i=1}^{N} y_i^2 + \sum_{i\neq j}^{N} y_i y_j = (N-1)S^2 + N\bar{Y}^2 + \sum_{i\neq j}^{N} y_i y_j$$

$$\Rightarrow \quad \sum_{i\neq j}^{N} y_i y_j = N^2\bar{Y}^2 - (N-1)S^2 - N\bar{Y}^2 .$$

Hence,

$$E(y_i y_j) = \frac{1}{N}\frac{1}{N-1}[N^2\bar{Y}^2 - (N-1)S^2 - N\bar{Y}^2] = \bar{Y}^2 - S^2/N .$$

and

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

$$= S^2\sum_{i=1}^{n}\alpha_i^2 - \frac{S^2}{N}\sum_{i=1}^{n}\alpha_i^2 + \bar{Y}^2\sum_{i=1}^{n}\alpha_i^2 + \left(1 - \sum_{i=1}^{n}\alpha_i^2\right)\left(\bar{Y}^2 - \frac{S^2}{N}\right)$$

$$= S^2\sum_{i=1}^{n}\alpha_i^2 + \bar{Y}^2 - \frac{S^2}{N} .$$

Thus,

$$V(T) = S^2\sum_{i=1}^{n}\alpha_i^2 - \frac{S^2}{N}, \text{ since } \sum_{i=1}^{n}\alpha_i^2 = \sum_{i=1}^{n}\left(\alpha_i - \frac{1}{n}\right)^2 + \frac{1}{n}, \text{ under condition } \sum_{i=1}^{n}\alpha_i = 1,$$

then

$$V(T) = S^2\left[\sum_{i=1}^{n}\left(\alpha_i - \frac{1}{n}\right)^2 + \left(\frac{1}{n} - \frac{1}{N}\right)\right].$$

Therefore, we note that $V(T)$ will be minimum, if $\sum_{i=1}^{n}\left(\alpha_i - \frac{1}{n}\right)^2 = 0$, where $\alpha_i = \frac{1}{n}$, for all $i = 1, 2, \cdots, n$, and $T = \frac{1}{n}\sum_{i=1}^{n} y_i = \bar{y}$.

OR

Differentiating variance function with respect to α_i and equating to zero, we get

$$\frac{\partial}{\partial \alpha_i} V(T) = 2S^2\left(\alpha_i - \frac{1}{n}\right) = 0 \quad \Rightarrow \alpha_i = \frac{1}{n}, \text{ for all } i = 1, 2, \cdots, n, \text{ and } T = \frac{1}{n}\sum_{i=1}^{n} y_i = \bar{y}.$$

Simple random sampling applied to qualitative characteristics

Suppose a random sample of size n is drawn from a population of size N, for which the proportion of individuals having a character C (attribute) is P. Thus, in the population, NP members are with a particular character C and NQ members with the character $not - C$ (e.g. in sampling from a population of persons, we may have persons who are smokers and non-smokers, honest and dishonest, below poverty line and above poverty line etc.). Let a be the number of members in the sample having the character C, then the sample proportion $p = \frac{a}{n}$.

To obtain the expectation and variance of sample proportion, first we change the attribute to variable by adopting the following procedure.

We assign to the $i-$th member of the population the value Y_i, which is equal to 1 if this member possesses the character C and is equal to 0 otherwise. In this way, we get a variable y, which has

Population total $= \sum_{i=1}^{N} Y_i = NP = A$.

Population mean $= \frac{1}{N}\sum_{i=1}^{N} Y_i = \frac{NP}{N} = P$.

Population variance $= \frac{1}{N}\sum_{i=1}^{N}(Y_i - P)^2 = \frac{1}{N}\sum_{i=1}^{N} Y_i^2 - P^2 = \frac{NP}{N} - P^2 = PQ$.

Mean square of population $= \frac{1}{N-1}\sum_{i=1}^{N}(Y_i - P)^2 = \frac{1}{N-1}\left(\sum_{i=1}^{N} Y_i^2 - NP^2\right)$

$$= \frac{NP - NP^2}{N-1} = \frac{NPQ}{N-1}.$$

Similarly, assign to the $i-$th member of the sample the value y_i, which is equal to 1 if this member possesses the character C and is equal to 0 otherwise, then

Sample total $= \sum_{i=1}^{n} y_i = np = a$, and Sample mean $= \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{a}{n} = p$.

Mean square for sample $= \dfrac{1}{n-1}\sum_{i=1}^{n}(y_i - p)^2 = \dfrac{1}{n-1}\left(\sum_{i=1}^{n} y_i^2 - np^2\right)$

$$= \frac{1}{n-1}(np - np^2) = \frac{npq}{n-1}.$$

Case I) Random sampling with replacement

On replacing $\bar{Y}$ by P, Y by NP, $\bar{y}$ by $p = \dfrac{a}{n}$, S^2 by $\dfrac{NPQ}{N-1}$ and σ^2 by PQ in the expressions obtained in expectation and variance of the estimates of population mean and population total, we find

i) $E(p) = E(\bar{y}) = \bar{Y} = P$. This shows that sample proportion p is an unbiased estimate of population proportion P and $V(p) = V(\bar{y}) = \dfrac{\sigma^2}{n} = \dfrac{PQ}{n}$.

ii) $E(\hat{A}) = E(Np) = N\,E(p) = NP = A$, means that $Np = \hat{A}$ is an unbiased estimate of $NP = A$ and

$$V(\hat{A}) = V(\hat{Y}) = N^2 V(\bar{y}) = \frac{N^2\sigma^2}{n} = \frac{N^2 PQ}{n}.$$

Theorem: $\hat{V}(p) = v(p) = \dfrac{pq}{n-1}$ is an unbiased estimate of $V(p) = \dfrac{PQ}{n}$.

Proof: $E[\hat{V}(p)] = E\left(\dfrac{pq}{n-1}\right) = E\left(\dfrac{n}{n}\dfrac{pq}{n-1}\right) = \dfrac{1}{n}E\left(\dfrac{npq}{n-1}\right)$

$= \dfrac{PQ}{n}$, since in *srswr* $E(s^2) = \sigma^2 = PQ$ and $s^2 = \dfrac{npq}{n-1}$.

Corollary: $\hat{V}(\hat{A}) = \hat{V}(Np) = N^2\,\hat{V}(p) = N^2\,\dfrac{pq}{n-1}$ is an unbiased estimate of $V(\hat{A}) = N^2\,\dfrac{PQ}{n}$.

Remarks

i) The standard error (SE) of p is $SE\,(p) = \sqrt{PQ/n}$.

ii) The standard error of $\hat{A}$ is $SE(\hat{A}) = N\sqrt{PQ/n}$.

Case II) Random sampling without replacement

Results are:

i) $E(p) = E(\bar{y}) = \bar{Y} = P$. This shows that sample proportion p is an unbiased estimate of population proportion P and $V(p) = V(\bar{y}) = \dfrac{N-n}{nN}S^2 = \left(\dfrac{N-n}{nN}\right)\dfrac{NPQ}{N-1} = \left(\dfrac{N-n}{N-1}\right)\dfrac{PQ}{n}$.

ii) $E(\hat{A}) = E(Np) = N\,E(p) = NP = A$, means that Np is an unbiased estimate of NP and

$$V(\hat{A}) = V(\hat{Y}) = N^2 V\,(\bar{y}) = N^2\left(\frac{N-n}{nN}\right)S^2 = N^2\left(\frac{N-n}{nN}\right)\frac{NPQ}{N-1} = N^2\left(\frac{N-n}{N-1}\right)\frac{PQ}{n}.$$

Theorem: $\hat{V}(p)=v(p)=\left(\frac{N-n}{n-1}\right)\frac{pq}{N}$ is an unbiased estimate of $V(p)=\left(\frac{N-n}{N-1}\right)\frac{PQ}{n}$.

Proof: $E[\hat{V}(p)]=E\left[\left(\frac{N-n}{n-1}\right)\frac{pq}{N}\right]=E\left[\left(\frac{N-n}{nN}\right)\frac{npq}{n-1}\right]=\left(\frac{N-n}{nN}\right)E\left(\frac{npq}{n-1}\right)$

$=\left(\frac{N-n}{N-1}\right)\frac{PQ}{n}$, since in *srswor* $E(s^2)=S^2=\frac{NPQ}{N-1}$ and $s^2=\frac{npq}{n-1}$.

Corollary: $\hat{V}(\hat{A})=\hat{V}(Np)=N^2\,\hat{V}(p)=N\left(\frac{N-n}{n-1}\right)pq$ is an unbiased estimate of $V(\hat{A})=N^2\left(\frac{N-n}{N-1}\right)\frac{PQ}{n}$.

Remarks

The standard error (SE) of p is $SE(p)=\sqrt{\left(\frac{N-n}{N-1}\right)\frac{PQ}{n}}$ and the standard error of $\hat{A}$ is $SE(\hat{A})=N\sqrt{\left(\frac{N-n}{N-1}\right)\frac{PQ}{n}}$.

Example: A list of 3000 voters of a ward in a city was examined for measuring the accuracy of age of individuals. A random sample of 300 names was taken, which revealed that 51 citizens were shown with wrong ages. Estimate the total number of voters having a wrong description of age in the list and estimate the standard error.

Solution: Given $N=3000$, $n=300$, $a=51$, and $p=\frac{a}{n}=0.17$, then, $\hat{A}=N\,p=510$.

i) If *srswr*, is considered, the estimate of the standard error is given by

$$Est[SE(\hat{A})]=N\sqrt{\frac{pq}{n-1}}=65.1696\cong 65.$$

ii) If *srswor*, is considered, the estimate of the standard error is given by

$$Est[SE(\hat{A})]=\sqrt{N\left(\frac{N-n}{n-1}\right)pq}=61.8246\cong 62.$$

Confidence interval (Interval estimations)

After having the estimate of an unknown parameter (which is rarely equal to parameter), it becomes necessary to measure the reliability of the estimate and to construct some confidence limits with a given degree of confidence. An estimate of a population parameter given by two numbers between which the parameter may be considered to lie is called an interval estimate, i.e. an interval estimate of a parameter θ is an interval of the form $L\le\theta\le U$, where L and U depends on the sampling distribution of $\hat{\theta}$.

To choose L and U for any specified probability $1-\alpha$, where L, such that $\Pr(L\le\theta\le U)=1-\alpha$. An interval $L\le\theta\le U$, computed for a particular sample, is called a $(1-\alpha)100\%$ confidence interval, the quantity $(1-\alpha)$ is called the confidence coefficient or the degree of confidence, and the end points L and U are called the lower and upper confidence limits. For instance, when $\alpha = 0.05$ the degree of confidence is 0.95 and we get a 95% confidence interval.

Limits in case of simple random sampling with replacement

1. **Confidence limit for population mean:** It is usually assumed that the estimator $\bar{y}$ is normally distributed about the corresponding population values, i.e. $\bar{y} \sim N(\bar{Y}, \sigma^2 / n)$.

   Since the tables are available for standard normal variable, so that we transform the values normal to standard normal as $Z = \dfrac{\bar{y} - \bar{Y}}{\sigma / \sqrt{n}} \sim N(0,1)$.

   By definition,

   $$\Pr(|Z| \le Z_{\alpha/2}) = 1 - \alpha \quad \text{or} \quad \Pr(-Z_{\alpha/2} \le Z \le Z_{\alpha/2}) = 1 - \alpha$$

   or $\Pr\left(-Z_{\alpha/2} \le \dfrac{\bar{y} - \bar{Y}}{SE(\bar{y})} \le Z_{\alpha/2} \right) = 1 - \alpha$

   or $\Pr[-Z_{\alpha/2}\, SE(\bar{y}) \le \bar{y} - \bar{Y} \le Z_{\alpha/2}\, SE(\bar{y})] = 1 - \alpha$

   or $\Pr[\bar{y} - Z_{\alpha/2}\, SE(\bar{y}) \le \bar{Y} \le \bar{y} + Z_{\alpha/2}\, SE(\bar{y})] = 1 - \alpha$.

   The probability being $(1 - \alpha)$, the interval $\Pr[\bar{y} - Z_{\alpha/2}\, SE(\bar{y}) \le \bar{Y} \le \bar{y} + Z_{\alpha/2}\, SE(\bar{y})] = 1 - \alpha$ will include $\bar{Y}$, i.e. $\bar{y} \pm Z_{\alpha/2}\, \sigma / \sqrt{n}$ will include $\bar{Y}$.

2. **Confidence limit for population total:** On the same above lines, we see that

   $\Pr[N\bar{y} - Z_{\alpha/2}\, SE(\hat{Y}) \le Y \le N\bar{y} + Z_{\alpha/2}\, SE(\hat{Y})] = 1 - \alpha$

   The probability being $(1 - \alpha)$, the interval, $N\bar{y} \pm Z_{\alpha/2}\, N\sigma / \sqrt{n}$ will include Y.

Note: If the sample size is less than 30, and population variance is unknown, **Student**$-t$ is used, instead of standard normal.

3. **Confidence limit for population proportion:** As above, we see that

   $\Pr[p - Z_{\alpha/2}\, SE(p) \le P \le p + Z_{\alpha/2}\, SE(p)] = 1 - \alpha$

   The probability being $(1 - \alpha)$, the interval, $p \pm Z_{\alpha/2} \sqrt{PQ / n}$ will include P.

Limits in case of simple random sampling without replacement

1. **Confidence limit for population mean:** Here also the distribution of the estimate based on the sample as distributed normally, i.e. $\bar{y} \sim N(\bar{Y}, (1 - f) S^2 / n)$, then, $Z = \dfrac{\bar{y} - \bar{Y}}{S\sqrt{(1 - f)/n}} \sim N(0,1)$. By definition,

   $\Pr[\bar{y} - Z_{\alpha/2}\, SE(\bar{y}) \le \bar{Y} \le \bar{y} + Z_{\alpha/2}\, SE(\bar{y})] = 1 - \alpha$.

   The probability being $(1 - \alpha)$, the interval $[\bar{y} - Z_{\alpha/2}\, SE(\bar{y}) \le \bar{Y} \le \bar{y} + Z_{\alpha/2}\, SE(\bar{y})]$ will include $\bar{Y}$,i.e. $\bar{y} \pm Z_{\alpha/2}\, S\sqrt{(1 - f)/n}$ will include $\bar{Y}$.

2. **Confidence limit for population total:** As in *srswr*, we see that

   $\Pr[N\,\bar{y}-Z_{\alpha/2}\,SE(\hat{Y})\le Y\le N\,\bar{y}+Z_{\alpha/2}\,SE(\hat{Y})]=1-\alpha$. The probability being $(1-\alpha)$, the interval, $N\,\bar{y}\pm Z_{\alpha/2}\,NS\sqrt{(1-f)/n}$ will include Y.

Note: If the sample size is less than 30, and population variance is unknown, **Student** $-t$ is used, instead of standard normal.

3. **Confidence limit for population proportion:** As in *srswr*, we see that

   $\Pr[p-Z_{\alpha/2}\,SE(p)\le P\le p+Z_{\alpha/2}\,SE(p)]=1-\alpha$. The probability being $(1-\alpha)$, the interval , $p\pm Z_{\alpha/2}\sqrt{\left(\frac{N-n}{N-1}\right)\frac{PQ}{n}}$ will include P.

Example: In a library, there are 4500 members who can borrow the books. A random sample of 16 persons was taken and number of books borrowed by them during a month was recorded as follows:

2, 3, 10, 0, 5, 7, 13, 1, 6, 23, 18, 12, 6, 0, 1 and 7. Estimate the average number of books borrowed by each member during a month and obtain 95% confidence interval.

Solution: Given $N=4500$, $n=16$

Estimate of population mean $\hat{\bar{Y}}=$ Sample mean $\bar{y}=\frac{1}{n}\sum_{i=1}^{n}y_i=7.125$.

Since sample size is small and variance of population is unknown, so the interval is defined as

$$\bar{y}\pm t_{\alpha/2,n-1}\,S\sqrt{\frac{N-n}{nN}}=\bar{y}\pm t_{\alpha/2,n-1}\frac{S}{\sqrt{n}}\text{, as population size is very large.}$$

S^2 is unknown, it can be replaced by its estimator $s^2=\frac{1}{n-1}\left(\sum_{i=1}^{n}y_i^2-n\,\bar{y}^2\right)=44.25$.

Therefore,

Upper confidence limit $=7.125+2.131\times\frac{6.652}{\sqrt{16}}=10.668853\cong 11$, and

Lower confidence limit $=7.125-2.131\times\frac{6.652}{\sqrt{16}}=3.58\cong 4$.

Example: In a mess, it was observed that leftover cost a lot. A survey was conducted to find out the optimum quantity for each item. A random sample of 10 inmates showed that they taken 4, 5, 2, 3, 1, 7, 2, 3, 4, 4 slices of bread in their breakfast. If there are 120 breakfasts are to be served every day, estimate the number of slices required every day. Also obtain a 95% confidence interval for it.

Solution: Given $N=120$, $n=10$, and $\bar{y}=\frac{1}{n}\sum_{i=1}^{n}y_i=3.5$, then

Estimate of population total $\hat{Y}=N\,\bar{y}=420$.

Since sample size is small and variance of population is unknown, so that, confidence limit

$$N\,\bar{y} \pm t_{\alpha/2,\,n-1}\, NS\sqrt{(1-f)/n}$$

Since S^2 is unknown, so it can be replaced by its unbiased estimator

$$s^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} y_i^2 - n\,\bar{y}^2\right) = 2.94444\,.$$

Hence,

Upper confidence limit $= 420 + 2.262 \times 120 \times 1.716\sqrt{\left(1 - \frac{10}{120}\right)/10} = 561.02517 \cong 561$

and

Lower confidence limit $= 420 - 141.02517 = 278.97483 \cong 279\,.$

Example: 100 villages were selected under *srswor* from a list of 1521 villages. It was found that 19 of the selected villages where illegally occupied by some landlords. Estimate all such villages occupied by the landlords out of the total 1521 villages and 95% confidence interval.

Solution: Given $N = 1521$, $n = 100$, and $a = 19$, then, $p = 0.19$

Estimate of number of village illegally occupied by landlords in the population of villages $\hat{A} = N\,p = 288.99 \cong 289\,.$

Since sample size is > 30 and variance of population proportion is unknown, then, confidence limit will be

$N\,p \pm Z_{\alpha/2}\, SE(\hat{A})$, where, $SE(\hat{A})$ is unknown, so it can be replaced by its unbiased estimator

$$\sqrt{N\left(\frac{N-n}{n-1}\right)pq} = 57.964667\,.$$

Thus,

Upper confidence limit $= 289 + 1.96 \times 57.964667 = 412.5227 \cong 413$, and

Lower confidence limit $= 289 - 1.96 \times 57.964667 = 165.4773 \cong 165\,.$

Example: A simple random sample of 30 households was drawn without replacement from a city area containing 14848 households. The number of persons per household in the sample were as follows: 5, 6, 3, 3, 2, 3, 3, 3, 4, 4, 3, 2, 7, 4, 3, 5, 4, 4, 3, 3, 4, 3, 3, 1, 2, 4, 3, 4, 3 and 4. Estimate the average and total number of people in the area and compute the probability that these estimates are with in $\pm 10\%$ of the true value.

Solution: Given $N = 14848$, and $n = 30$, then,

Estimate of the population total $Y = N\,\bar{y} = 14848 \times \frac{105}{30} = 51968\,.$ Assuming that the population values are normally distributed, so that, $N\,\bar{y} \sim N\left(Y, NS\sqrt{\frac{1-f}{n}}\right)$, thus,

$$\Pr(\text{ Estimate lies with in 10\% of the true value})$$

$$= \Pr(Y - 10\% \text{ of } Y \le N\bar{y} \le Y + 10\% \text{ of } Y)$$

$$= P(0.9\,Y \le N\bar{y} \le 1.1Y) = P(N\bar{y} \le 1.1Y) - P(N\bar{y} \le 0.9Y)$$

We shall use the result that

$$N\bar{y} \sim N\left(Y, NS\sqrt{\frac{1-f}{n}}\right), \text{ so that } Z = \frac{N\bar{y} - Y}{NS\sqrt{\frac{1-f}{n}}} \sim N(0,1)$$

$$\Pr(N\bar{y} \le 1.1Y) = \Pr\left(\frac{1}{1.1} N\bar{y} \le Y\right) = \Pr\left(\frac{10}{11} N\bar{y} \le Y\right)$$

$$= \Pr\left(\frac{10}{11} N\bar{y} + \frac{1}{11} N\bar{y} \le Y + \frac{1}{11} N\bar{y}\right)$$

$$= \Pr\left(N\bar{y} \le Y + \frac{1}{11} N\bar{y}\right) = \Pr\left(N\bar{y} - Y \le \frac{1}{11} N\bar{y}\right)$$

$$= \Pr\left(\frac{N\bar{y} - Y}{NS\sqrt{\frac{1-f}{n}}} \le \frac{N\bar{y}}{11NS\sqrt{\frac{1-f}{n}}}\right) = \Pr(Z \le 1.457) = 0.9279.$$

Similarly,

$$\Pr(N\bar{y} \le 0.9Y) = \Pr\left(\frac{10}{9} N\bar{y} \le Y\right) = \Pr\left(\frac{10}{9} N\bar{y} - \frac{1}{9} N\bar{y} \le Y - \frac{1}{9} N\bar{y}\right)$$

$$= \Pr\left(N\bar{y} - Y \le -\frac{1}{9} N\bar{y}\right) = \Pr\left(\frac{N\bar{y} - Y}{NS\sqrt{\frac{1-f}{n}}} \le -\frac{N\bar{y}}{9NS\sqrt{\frac{1-f}{n}}}\right)$$

$$= \Pr(Z \le -1.78) = 0.0375.$$

Therefore, the required probability $0.9279 - 0.0375 = 0.8904$.

Estimation of sample size

In planning a sample survey for estimating the population parameters, the preliminary thing is how to determine the size of the sample to be drawn. Following ways can do it:

a) **Specify the precision in terms of margin of error:** The margin of error, which is permissible in the estimate, is known as permissible error. It is taken as the maximum difference between the estimate and the parametric value that can be tolerated. Suppose an error d on either side of the parameter value $\bar{Y}$ can be tolerated in the estimate $\bar{y}$ based on the sample values. Thus the permissible error in the estimate $\bar{y}$ is specified by

$$\bar{y} = \bar{Y} \pm d \text{ or } \bar{y} - \bar{Y} = \pm\, d \quad \text{or} \quad |\bar{y} - \bar{Y}| = d.$$

Since $|\bar{y}-\bar{Y}|=d$ differ from sample to sample, so this margin of error can be specified in the form of probability statement as:

$$\Pr[|\bar{y}-\bar{Y}|\geq d]=\alpha \quad \text{or} \quad \Pr[|\bar{y}-\bar{Y}|\leq d]=1-\alpha\,. \tag{2.9}$$

Where α is small and it is the risk that we are willing to bear if the actual difference is greater than d. This α is called the level of significance and $(1-\alpha)$ is called level of confidence or confidence coefficient.

As the population is normally distributed, so the sample mean will also follow the normal distribution i.e. $\bar{y}\sim N[\bar{Y},V(\bar{y})]$, then $Z=\dfrac{\bar{y}-\bar{Y}}{\sqrt{V(\bar{y})}}\sim N(0,1)$.

For the given value of α we can find a value Z_α of standard normal variate from the standard normal table by the following equation:

$$\Pr\left[\frac{|\bar{y}-\bar{Y}|}{\sqrt{V(\bar{y})}}\geq Z_{\alpha/2}\right]=\alpha \quad \text{or} \quad \Pr\,[|\bar{y}-\bar{Y}|\geq\sqrt{V(\bar{y})}\,Z_{\alpha/2}]=\alpha \tag{2.10}$$

Comparing the equation (2.9) and (2.10), we get

$$d=Z_{\alpha/2}\sqrt{V(\bar{y})}\,,\ \text{so that}\ d^2=Z_{\alpha/2}^2\,V(\bar{y})=Z_{\alpha/2}^2\left(\frac{1}{n}-\frac{1}{N}\right)S^2\,.$$

$$\Rightarrow\quad 1=\frac{Z_{\alpha/2}^2\,S^2}{d^2}\left(\frac{1}{n}-\frac{1}{N}\right)=n_0\left(\frac{1}{n}-\frac{1}{N}\right),\ \text{where}\ n_0=\frac{Z_{\alpha/2}^2\,S^2}{d^2} \tag{2.11}$$

$$\text{or}\quad 1=\frac{n_0}{n}-\frac{n_0}{N}\quad\Rightarrow\quad\frac{n_0}{n}=1+\frac{n_0}{N}\quad\text{or}\quad n=\frac{n_0}{1+\dfrac{n_0}{N}} \tag{2.12}$$

If N is sufficiently large, then $n\cong n_0$ and for unknown S^2, some rough estimate of S^2 can be used in relation's (2.12) and (2.11).

b) **Specify the precision in terms of margin of** $V(\bar{y})$ i.e. we have to find sample size n such that $V(\bar{y})=V$ (given). As in case of margin of error,

$$d=Z_{\alpha/2}\sqrt{V(\bar{y})}\quad\Rightarrow\quad V(\bar{y})=\frac{d^2}{Z_{\alpha/2}}\,,\ \text{and}\ n_0=\frac{Z_{\alpha/2}^2\,S^2}{d^2}=\frac{S^2}{V(\bar{y})}$$

Therefore, $n_0=S^2/V$, and hence n can be obtained by relation (2.12).

c) **Specify the precision in terms of coefficient of variation of $\bar{y}$:**

$$\text{Let}\ CV(\bar{y})=e=\frac{\sqrt{V(\bar{y})}}{\bar{Y}}\quad\Rightarrow\quad\frac{V(\bar{y})}{\bar{Y}^2}=e^2\quad\text{or}\quad V(\bar{y})=e^2\,\bar{Y}^2 \tag{2.13}$$

Substitute equation (2.13) in relation (2.11), we get,

$$n_0=\frac{S^2}{e^2\,\bar{Y}^2}\,,\ \text{and hence}\ n\ \text{from (2.12).}$$

Remark

i) To get n such that the margin of error in the estimate $\hat{Y}=N\,\bar{y}$ of the population total Y is d', then, $|\hat{Y}-Y|=d'$ or $|N\bar{y}-N\bar{Y}|=d'$, or $N|d|=d'$ or $N^2 d^2=d'^2$, or $d^2=\dfrac{d'^2}{N^2}$.

Therefore,

$$n_0=\left(\frac{N\,Z_{\alpha/2}\,S}{d'}\right)^2$$, and n can be obtained by the relation (2.12).

ii) To find n for $\hat{A}=N\,\bar{y}$ with precision specified as $V(\hat{A})=V$ i.e. $V(\hat{A})=N^2\,V(\bar{y})=V'$

$\Rightarrow\quad V(\bar{y})=\dfrac{V'}{N^2}$, and $n_0=\dfrac{N^2\,S^2}{V'}$, then, n from (2.12).

Example: For a population of size $N=430$ roughly we know that $\bar{Y}=19$, $S^2=85.6$ with srs, what should be the size of sample to estimate $\hat{\bar{Y}}$ with a margin of error 10% of $\bar{Y}$ apart chance is 1 in 20.

Solution: Margin of error in the estimate $\bar{y}$ of $\bar{Y}$ is given, i.e.

$\bar{y}=\bar{Y}\pm 10\%$ of $\bar{Y}$ or $|\bar{y}-\bar{Y}|=10\%$ of $\bar{Y}=\dfrac{19}{10}=1.9$, so that

$\Pr[|\bar{y}-\bar{Y}|\ge 1.9]=\dfrac{1}{20}=0.05$, and $n_0=\dfrac{Z_{\alpha/2}^2\,S^2}{d^2}=\dfrac{(1.96)^2\times 85.6}{(1.9)^2}=91.091678$.

Therefore,

$$n=\frac{n_0}{1+\frac{n_0}{N}}=75.168\cong 75.$$

Example: In the population of 676 petition sheets. How large must the sample be if the total number of signatures is to be estimated with a margin of error of 1000, apart from a 1 in 20 chance? Assume that the population mean square to be 229.

Solution: Let Y be the number of signature on all the sheets. Let $\hat{Y}$ is the estimate of Y. Margin of error is specified in the estimate $\hat{Y}$ of Y as

$|\hat{Y}-Y|=1000$, so that, $\Pr[|\hat{Y}-Y|\ge 1000]=\dfrac{1}{20}=0.05$.

We know that

$n=\dfrac{n_0}{1+\frac{n_0}{N}}$, here, $n_0=\left(\dfrac{N\,Z_{\alpha/2}\,S}{d'}\right)^2=\left(\dfrac{676\times 1.96}{1000}\right)^2 229=402.01385$, and hence

$n=252.09\cong 252$.

Estimation of sample size for proportion

a) **When precision is specified in terms of margin of error:** Suppose size of the population is N and population proportion is P. Let a srs of size n is taken and p be the corresponding sample proportion and d is the margin of error in the estimate p of P. The margin of error can be specified in the form of probability statement as

$$\Pr[|p-P| \geq d] = \alpha \quad \text{or} \quad \Pr[|p-P| \leq d] = 1-\alpha \tag{2.14}$$

As the population is normally distributed, so $\bar{y} \sim N[P, V(p)]$, then $Z = \dfrac{p-P}{\sqrt{V(p)}} \sim N(0,1)$. For the given value of α we can find a value Z_α of the standard normal variate from the standard normal table by the following relation:

$$\Pr\left[\frac{|p-P|}{\sqrt{V(p)}} \geq Z_{\alpha/2}\right] = \alpha \quad \text{or} \quad \Pr[|p-P| \geq \sqrt{V(p)}\, Z_{\alpha/2}] = \alpha \tag{2.15}$$

Comparing equation (2.14) and (2.15), the relation which gives the value of n with the required precision of the estimate p of P is given by

$$d = Z_{\alpha/2}\sqrt{V(p)} \quad \text{or} \quad d^2 = Z^2_{\alpha/2} V(p) = Z^2_{\alpha/2}\left(\frac{N-n}{N-1}\right)\frac{PQ}{n}, \text{ as sampling is } srswr.$$

$$\Rightarrow\ 1 = \frac{Z^2_{\alpha/2}\,PQ}{d^2}\left(\frac{N-n}{n(N-1)}\right) = n_0\,\frac{N-n}{n(N-1)}, \text{ where } n_0 = \frac{Z^2_{\alpha/2}\,PQ}{d^2} = \frac{PQ}{V(p)} \tag{2.16}$$

$$\text{or} \quad \frac{N-1}{n_0} = \frac{N-n}{n} = \frac{N}{n} - 1 \quad \Rightarrow \quad \frac{N}{n} = 1 + \frac{N-1}{n_0}$$

$$\text{or} \quad n = \frac{N}{1+\dfrac{N-1}{n_0}} = \frac{N\,n_0}{n_0+(N-1)} = \frac{n_0}{\dfrac{n_0}{N}+\dfrac{N-1}{N}} = \frac{n_0}{1+\dfrac{n_0}{N}} \tag{2.17}$$

If N is sufficiently large, then $n \cong n_0$

b) If precision is specified in terms of $V(p)$ i.e. $V(p) = V$ (given).

Substituting $V(p) = V$ in relation (2.16) we get, $n_0 = \dfrac{PQ}{V}$, and hence n can be obtained by relation (2.17).

c) When precision is given in terms of coefficient of variation of p

Let

$$CV(p) = e = \frac{\sqrt{V(p)}}{P} \quad \Rightarrow \quad \frac{V(p)}{P^2} = e^2, \quad \text{or} \quad V(p) = e^2 P^2 \tag{2.18}$$

Substitute equation (2.18) in relation (2.16), we get,

$$n_0 = \frac{PQ}{e^2 P^2} = \frac{Q}{e^2 P} = \frac{1}{e^2}\left(\frac{1}{P} - 1\right), \text{ and hence } n \text{ is given by the relation (2.17).}$$

Remarks

i) To get n, if the margin of error in the estimate $\hat{A}=Np$ of the population total $A=NP$ is d', then,

$$|\hat{A}-A|=d' \text{ or } |N\,p-N\,P|=d', \text{ or } N|\,d\,|=d', \text{ or } N^2\,d^2=d'^2, \text{ or } d^2=\frac{d'^2}{N^2}.$$

Thus,

$$n_0=\left(\frac{N\,Z_{\alpha/2}\,PQ}{d'}\right)^2, \text{ and } n \text{ can be obtained by the relation (1.17).}$$

ii) To find n, for $\hat{A}=Np$ with precision specified as $V(\hat{A})=V$ i.e. $V(\hat{A})=N^2\,V(p)=V'$, so that, $V(p)=\frac{V'}{N^2}$, substitute this value in equation (2.16), we get, $n_0=\frac{N^2\,PQ}{V'}$, and n is given by relation (2.17).

Example: In a population of 4000 people who were called for casting their votes, 50% returned to the poll. Estimate the sample size to estimate this proportion so that the marginal error is 5% with 95% confidence coefficient.

Solution: Margin of error in the estimate p of P is given by

$|\,p-P\,|=0.05$, then $\Pr[|\,p-P\,|\geq 0.05]=0.05$.

We know that

$$n_0=\frac{Z_{\alpha/2}^2\,PQ}{d^2}=\frac{(1.96)^2\times 0.5\times 0.5}{0.0025}=384.16\cong 384, \text{ and hence,}$$

$$n=\frac{n_0}{1+(n_0\,/\,N)}=350.498\cong 351.$$

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

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

The estimate of total value made from a sample is to be correct within \$200, apart from a 1 in 20 chance. An advisor suggests that a simple random sample of 12 shelves will meet the requirements. Do you agree? $\sum Y_i=2138$, and $\sum Y_i^2=131\,682$.

Solution: It is given that $\sum_i Y_i=2138$, $\sum_i Y_i^2=131\,682$, and $N=36$, then

$$S^2=\frac{1}{N-1}\left[\sum_i Y_i^2-N\bar{Y}^2\right]=\frac{1}{36-1}\left[131\,682-36\left(\frac{2138}{36}\right)^2\right]=134.5, \text{ and}$$

$$|\,\hat{Y}-Y\,|\leq 200, \text{ then, } \Pr[|\,\hat{Y}-Y\,|\leq 200]=\frac{1}{20}=0.05.$$

We know that

$$n = \frac{n_0}{1+\frac{n_0}{N}}, \text{ here } \quad n_0 = \left(\frac{N\, Z_{\alpha/2}}{d}\right)^2 S = \left(\frac{36\times 1.96}{200}\right)^2 134.5 = 16.7409 \text{, and therefore,}$$

$$n = 11.42765 \cong 12 .$$

Determination of sample size in decision problems (Another approach)

Let $l(z)$ denote the amount of loss (in monetary terms) that will be incurred in a decision through an error of amount z in the estimate. Let $f(z)$ denote the probability density function of z. Then the expected loss for a given sample size n will be

$$L(n) = E[l(z)] = \int l(z)\, f(z)\, dz$$

If $C(n)$ is the cost of a sample of size n then the most economic sample size will be that which minimize the sum of cost and expected loss. Thus the problem of determination of the sample size can be stated as

Find n such that $\phi(n) = C(n) + L(n)$ is minimum.

Exercise: If the loss function due to an error in $\bar{y}$ is proportional to $|\, \bar{y} - \bar{Y}\,|$ and if the total cost of the survey is $C = c_0 + c_1 n$, show that with simple random sampling, ignoring the fpc, the most economical value of n is $\left(\frac{\lambda\, S}{c_1\sqrt{2\pi}}\right)^{2/3}$, where λ is a constant.

Solution: Given $l(z) \propto |\, \bar{y} - \bar{Y}\,|$, then, $l(z) = \lambda\, |\, \bar{y} - \bar{Y}\,|$, and

$$\bar{y} \sim N\left(\bar{Y}, \frac{S^2}{n}\right), \text{when } fpc \text{ is ignored } V(\bar{y}) = \frac{S^2}{n}, \Rightarrow \quad z = (\bar{y} - \bar{Y}) \sim N\left(0, \frac{S^2}{n}\right), \text{ so that}$$

$$f(z) = \frac{1}{(S/\sqrt{n})\,\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{z}{S/\sqrt{n}}\right)^2\right] = \frac{1}{(S/\sqrt{n})\,\sqrt{2\pi}} \exp\left(-\frac{n\,z^2}{2\,S}\right)$$

Now

$$|\, z\,| = |\, \bar{y} - \bar{Y}\,| = \bar{y} - \bar{Y} = z\,, \quad \text{if} \quad \bar{y} > \bar{Y}\,.$$

$$|\, z\,| = |\, \bar{y} - \bar{Y}\,| = \bar{y} - \bar{Y} = -z\,, \; \text{if} \quad \bar{y} < \bar{Y}\,.$$

Thus, the expected loss

$$L(n) = \int_{-\infty}^{\infty} \lambda |\, z\,|\; f(z)\, dz = \int_{-\infty}^{0} \lambda |\, z\,|\; f(z)\, dz + \int_{0}^{\infty} \lambda |\, z\,|\; f(z)\, dz$$

$$= -\int_{-\infty}^{0} \lambda\, z\; f(z)\, dz + \int_{0}^{\infty} \lambda\, z\; f(z)\, dz = 2\int_{0}^{\infty} \lambda\, z\; f(z)\, dz$$

$$= 2\int_{0}^{\infty} \lambda\, z \frac{1}{(S/\sqrt{n})\,\sqrt{2\pi}} \exp\,(-n\,z^2 / 2\,S)\, dz$$

Put $\frac{n z^2}{2 S^2} = t$, then $\frac{2 n z}{2 S^2} dz = dt$ or $z\, dz = \frac{S^2}{n} dt$.

Therefore,

$$L(n) = 2\int_0^\infty \frac{\lambda S^2}{n} \frac{1}{(S/\sqrt{n})\,\sqrt{2\pi}} e^{-t}\, dt = \frac{2\lambda\, S}{\sqrt{2 n\pi}} \int_0^\infty e^{-t}\, dt = \frac{2\lambda S}{\sqrt{2 n \pi}}, \text{ as } \int_0^\infty e^{-t}\, dt = 1.$$

To determine the value of n, consider the function

$$\phi(n) = L(n) + C(n) = c_0 + c_1 n + \frac{2\lambda S}{\sqrt{2\pi}} n^{-1/2}$$

Differentiate this function with respect to n, we get

$$\frac{\partial \phi}{\partial n} = 0 = c_1 - \frac{1}{2}\left(\frac{2\lambda S}{\sqrt{2\pi}}\right) n^{-3/2} \quad \text{or} \quad \frac{\lambda S}{\sqrt{2\pi}} n^{-3/2} = c_1$$

$$\text{or} \quad n^{-3/2} = \frac{c_1\sqrt{2\pi}}{\lambda S} \quad \text{or} \quad n = \left(\frac{\lambda S}{c_1\sqrt{2\pi}}\right)^{2/3}.$$

Exercise: With a loss function $l(z) = \lambda z^2$ and a cost function $C = c_0 + c_1 n$. Show that using *srs* the most economic value of the sample size n to estimate the population mean $\bar{Y}$ is $\left(\frac{\lambda S^2}{c_1}\right)^{1/2}$, where $z = \bar{y} - \bar{Y}$, $\bar{y}$ is the sample mean used to estimate $\bar{Y}$.

Solution: Given $l(z) = \lambda z^2$ quadratic loss function. By definition

$L(n) = E[\lambda z^2] = \lambda E(z^2)$. Consider,

$V(z) = E[z - E(z)]^2 = E(z^2)$, since $E(z) = E(\bar{y} - \bar{Y}) = 0$.

Also

$$V(z) = V(\bar{y} - \bar{Y}) = V(\bar{y}) = \left(\frac{1}{n} - \frac{1}{N}\right) S^2.$$

Therefore,

$E(z^2) = \frac{S^2}{n} - \frac{S^2}{N}$, and the expected loss $L(n) = \frac{\lambda S^2}{n} - \frac{\lambda S^2}{N}$.

To determine the value of n, consider the function

$$\phi(n) = L(n) + C(n) = \frac{\lambda S^2}{n} - \frac{\lambda S^2}{N} + c_0 + c_1 n$$

Differentiate this function with respect to n, we get

$$\frac{\partial \phi}{\partial n} = 0 = -\frac{\lambda S^2}{n^2} + c_1, \quad \text{or} \quad \frac{\lambda S^2}{n^2} = c_1, \text{or} \quad n = \left(\frac{\lambda S^2}{c_1}\right)^{1/2}.$$

Exercise: The selling price of a lot of standing timber is UW, where U is the price per unit volume and W is the volume of timber on the lot. The number N of logs on the lot is counted, and the average volume per log is estimated from a simple random sample of n logs. The estimate is made and paid for by the seller and is provisionally accepted by the buyer. Later, the buyer finds out the exact volume purchased, and the seller reimburses him if he has paid for more than was delivered. If he has paid for less than was delivered, the buyer does not mention the fact.

Construct the seller's loss function. Assuming that the cost of measuring n logs is cn, find the optimum value of n. The standard deviation of the volume per log may be denoted by S and the fpc ignored.

Solution: Let $\hat{W}$ be the estimated total volume of the timber. The error in the estimate is $\hat{W} - W$.

If $\hat{W} - W = z > 0$ sellers loss is zero, i.e. $l(z) = 0$

If $\hat{W} - W = z < 0$ sellers loss is $-Uz$, i.e. $l(z) = -Uz$.

When fpc is ignored $V(\hat{W}) = N^2 S^2 / n$, then

$$\hat{W} \sim N\left(W, \frac{N^2 S^2}{n}\right), \quad \text{or} \quad z = (\hat{W} - W) \sim N\left(0, \frac{NS}{\sqrt{n}}\right), \text{ so that}$$

$$f(z) = \frac{1}{(NS/\sqrt{n})\ \sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{z}{NS/\sqrt{n}}\right)^2\right] = \frac{1}{(NS/\sqrt{n})\ \sqrt{2\pi}} \exp\left(-\frac{nz^2}{2N^2S^2}\right)$$

Thus, the expected loss

$$L(n) = \int_{-\infty}^{\infty} l(z)\ f(z)\, dz = \int_{-\infty}^{0} (-Uz) \frac{1}{(NS/\sqrt{n})\ \sqrt{2\pi}} \exp\left(-\frac{nz^2}{2N^2S^2}\right) dz$$

$$= -\int_{-\infty}^{0} Uz \frac{1}{(NS/\sqrt{n})\ \sqrt{2\pi}} \exp\left(-\frac{nz^2}{2N^2S^2}\right) dz$$

$$= \int_{0}^{\infty} Uz \frac{1}{(NS/\sqrt{n})\ \sqrt{2\pi}} \exp\left(-\frac{nz^2}{2N^2S^2}\right) dz$$

Put $\dfrac{nz^2}{2N^2S^2} = t$, then $\dfrac{2nz}{2N^2S^2} dz = dt$ or $z\,dz = \dfrac{N^2S^2}{n} dt$.

Therefore,

$$L(n) = \int_{0}^{\infty} \frac{UN^2S^2}{n} \frac{1}{(NS/\sqrt{n})\ \sqrt{2\pi}} e^{-t}\, dt = \frac{UNS}{\sqrt{2n\pi}} \int_{0}^{\infty} e^{-t}\, dt = \frac{UNS}{\sqrt{2n\pi}}, \text{ as } \int_{0}^{\infty} e^{-t}\, dt = 1.$$

To determine the value of n, consider the function

$$\phi(n) = L(n) + C(n) = cn + \frac{UNS}{\sqrt{2\pi}} n^{-1/2}.$$

Differentiate this function with respect to n, we get

$$\frac{\partial\phi}{\partial n}=0=c-\frac{1}{2}\left(\frac{UNS}{\sqrt{2\pi}}\right)n^{-3/2} \quad \text{or} \quad \frac{UNS}{2\sqrt{2\pi}}n^{-3/2}=c$$

or $n^{-3/2}=\dfrac{2c\sqrt{2\pi}}{UNS}$ or $n=\left(\dfrac{UNS}{2c\sqrt{2\pi}}\right)^{2/3}$.

Exercise: With certain populations, it is known that the observations Y_i are all zero on a portion QN of N units $(0<Q<1)$. Sometimes with varying expenditure of efforts, these units can be found and listed, so that they need not be sampled. If σ^2 is the variance of Y_i in the original population and σ_0^2 is the variance when all zeros are excluded, then show that $\sigma_0^2=\dfrac{\sigma^2}{P}-\dfrac{Q}{P^2}\bar{Y}^2$, where $P=1-Q$, and $\bar{Y}$ is the mean value of Y_i for the whole population.

Solution: Given $Y_1,Y_2,\cdots,Y_{NP},Y_{NP+1},\cdots,Y_N$ (first NP units not zero, and rest NQ units which are all zero). Thus, $\bar{Y}=\dfrac{1}{N}\sum_{i=1}^{N}Y_i$, population mean, and $\bar{Y}_{NP}=\dfrac{1}{NP}\sum_{i=1}^{NP}Y_i$, $\bar{Y}_{NQ}=\dfrac{1}{NQ}\sum_{i=1}^{NQ}Y_i=0$, also, $\sum_{i=1}^{N}Y_i=\sum_{i=1}^{NP}Y_i$, and $\sum_{i=1}^{N}Y_i^2=\sum_{i=1}^{NP}Y_i^2$, so that $N\bar{Y}=NP\,\bar{Y}_{NP}$, or $\bar{Y}_{NP}=\dfrac{1}{P}\bar{Y}$. By definition,

$$\sigma^2=\frac{1}{N}\sum_{i=1}^{N}(Y_i-\bar{Y})^2=\frac{1}{N}\sum_{i=1}^{N}Y_i^2-\bar{Y}^2, \quad \text{or} \quad N\sigma^2=\sum_{i=1}^{N}Y_i^2-N\bar{Y}^2.$$

Similarly, $NP\sigma_0^2=\sum_{i=1}^{NP}Y_i^2-NP\bar{Y}_{NP}^2$.

Thus,

$$N(\sigma^2-P\sigma_0^2)=NP\bar{Y}_{NP}^2-N\bar{Y}^2=NP\frac{1}{P^2}\bar{Y}^2-N\bar{Y}^2=N\left(\frac{1}{P}-1\right)\bar{Y}^2=N\left(\frac{Q}{P}\right)\bar{Y}^2.$$

Therefore,

$$P\sigma_o^2=\sigma^2-\left(\frac{Q}{P}\right)\bar{Y}^2 \quad \text{or} \quad \sigma_o^2=\frac{\sigma^2}{P}-\frac{Q}{P^2}\bar{Y}^2.$$

Exercise: From a random sample of n units, a random sub-sample of n_1 units is drawn without replacement and added to the original sample. Show that the mean based on $(n+n_1)$ units is an unbiased estimator of the population mean, and that ratio of its variance to that of the mean of the original n units is approximately $\dfrac{1+3n_1/n}{(1+n_1/n)^2}$, assuming that the population size is large.

Solution: Let the sample mean based on n, n_1, and $n+n_1$ elements are denoted by $\bar{y}_n$, $\bar{y}_{n_1}$, and $\bar{y}_{n+n_1}$ respectively, and are defined as $\bar{y}_n = \frac{1}{n}\sum_{i=1}^{n} y_i$, $\bar{y}_{n_1} = \frac{1}{n_1}\sum_{i=1}^{n_1} y_i$, and $\bar{y}_{n+n_1} = \frac{n\,\bar{y}_n + n_1\,\bar{y}_{n_1}}{n+n_1}$. We have to show $E(\bar{y}_{n+n_1}) = \bar{Y}$, in this case the expectation is taken in two stages,

i) when n is fixed

ii) over all expectation

$$E\,(\bar{y}_{n+n_1}) = \frac{1}{n+n_1} E\,(n\,\bar{y}_n + n_1\,\bar{y}_{n_1}) = \frac{1}{n+n_1} E[n\,\bar{y}_n + n_1\,E(\bar{y}_{n_1} | \,n)]$$

$$= \frac{1}{n+n_1} E\,(n\,\bar{y}_n + n_1\,\bar{y}_n)\text{, since } n_1 \text{ is a sub-sample of the sample of size } n.$$

$$= \frac{1}{n+n_1} (n\bar{Y} + n_1\,\bar{Y}) = \bar{Y}\,.$$

To obtain the variance

$$V(\bar{y}_{n+n_1}) = E\,(\bar{y}_{n+n_1} - \bar{Y})^2 = E\left(\frac{n\,\bar{y}_n + n_1\,\bar{y}_{n_1}}{n+n_1} - \bar{Y}\right)^2$$

$$= \frac{1}{(n+n_1)^2} E[n\,\bar{y}_n + n_1\,\bar{y}_{n_1} - (n+n_1)\bar{Y}]^2$$

$$= \frac{1}{(n+n_1)^2} E[n\,\bar{y}_n - n\bar{Y} + n_1\,\bar{y}_{n_1} - n_1\,\bar{Y}]^2$$

$$= \frac{1}{(n+n_1)^2} E[n(\bar{y}_n - \bar{Y}) + n_1\bar{y}_{n_1} - n_1\bar{y}_n + n_1\bar{y}_n - n_1\bar{Y}]^2$$

$$= \frac{1}{(n+n_1)^2} E[(n+n_1)(\bar{y}_n - \bar{Y}) + n_1(\bar{y}_{n_1} - \bar{y}_n)]^2$$

$$= \frac{1}{(n+n_1)^2} [(n+n_1)^2\; E(\bar{y}_n - \bar{Y})^2 + n_1^2\; E(\bar{y}_{n_1} - \bar{y}_n)^2]\text{, as samples are drawn independently.}$$

$$= \frac{1}{(n+n_1)^2} [(n+n_1)^2\, V(\bar{y}_n) + n_1^2\; E\{E(\bar{y}_{n_1} - \bar{y}_n)^2 | \,n\}]$$

$$= \frac{1}{(n+n_1)^2} \left[(n+n_1)^2\; V(\bar{y}_n) + n_1^2\; E\left\{\left(\frac{1}{n_1} - \frac{1}{n}\right) s_n^2\right\}\right]$$

$$= \frac{1}{(n+n_1)^2} \left[(n+n_1)^2\; V(\bar{y}_n) + n_1^2 \left(\frac{n-n_1}{n_1 n}\right) S^2\right]$$

$$= \frac{1}{(n+n_1)^2}\left[(n+n_1)^2 V(\bar{y}_n) + \frac{n_1(n-n_1)}{n} S^2\right] = V(\bar{y}_n) + \frac{n_1(n-n_1)}{n(n+n_1)^2} S^2 .$$

Therefore,

$$\frac{V(\bar{y}_{n+n_1})}{V(\bar{y}_n)} = 1 + \frac{n_1(n-n_1)}{n(n+n_1)^2 V(\bar{y}_n)} S^2 \cong 1 + \frac{n_1(n-n_1)}{n(n+n_1)^2 S^2/n} S^2$$

$$= \frac{(n+n_1)^2 + n_1(n-n_1)}{(n+n_1)^2} = \frac{n^2 + n_1^2 + 2n_1 n + n_1 n - n_1^2}{(n+n_1)^2}$$

$$= \frac{n^2 + 3n_1 n}{(n+n_1)^2} = \frac{1+(3n_1/n)}{(1+n_1/n)^2} .$$

Exercise: A simple random sample of size $n = n_1 + n_2$ with mean $\bar{y}$ is drawn from a finite population, and a simple random subsample of size n_1 is drawn from it with mean $\bar{y}_1$. Show that

i) $V(\bar{y}_1 - \bar{y}_2) = S^2[(1/n_1) + (1/n_2)]$, where $\bar{y}_2$ is mean of the remaining n_2 units in the sample,

ii) $V(\bar{y}_1 - \bar{y}) = S^2[(1/n_1) - (1/n)]$,

iii) $Cov(\bar{y}, \bar{y}_1 - \bar{y}) = 0$.

Repeated sampling implies repetition of the drawing of both the sample and subsample.

Solution:

i) In repeated sampling the given procedure is equivalent to draw subsamples of sizes n_1 and n_2 independently, thus

$V(\bar{y}_1 - \bar{y}_2) = V(\bar{y}_1) + V(\bar{y}_2)$, since $Cov(\bar{y}_1, \bar{y}_2) = 0$

$= S^2[(1/n_1) + (1/n_2)]$, ignoring fpc.

ii) $$\bar{y} = \frac{n_1\bar{y}_1 + n_2\bar{y}_2}{n_1 + n_2} \Rightarrow \bar{y}_1 - \bar{y} = \bar{y}_1 - \frac{n_1\bar{y}_1 + n_2\bar{y}_2}{n_1 + n_2}$$

or $$\bar{y}_1 - \bar{y} = \frac{n_1\bar{y}_1 + n_2\bar{y}_1 - n_1\bar{y}_1 - n_2\bar{y}_2}{n_1 + n_2} = \frac{n_2(\bar{y}_1 - \bar{y}_2)}{n} .$$

Therefore,

$$V(\bar{y}_1 - \bar{y}) = V\left(\frac{n_2(\bar{y}_1 - \bar{y}_2)}{n}\right) = \frac{n_2^2}{n^2} V(\bar{y}_1 - \bar{y}_2) = \frac{n_2^2}{n^2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right) S^2$$

$$= \frac{n_2^2}{n^2}\left(\frac{n_1 + n_2}{n_1 n_2}\right) S^2 = \frac{n_2}{n_1 n} S^2 = \frac{n - n_1}{n_1 n} S^2 = \left(\frac{1}{n_1} - \frac{1}{n}\right) S^2 .$$

iii) $Cov(\bar{y}, \bar{y}_1 - \bar{y}) = E[\bar{y}(\bar{y}_1 - \bar{y})] - E(\bar{y})E(\bar{y}_1 - \bar{y})$

$$= E(\bar{y}\,\bar{y}_1 - \bar{y}^2) - \bar{Y} \times 0 = E(\bar{y}\,\bar{y}_1) - E(\bar{y}^2) \qquad (1)$$

Consider

$$E(\bar{y}\,\bar{y}_1) = E\left(\frac{n_1\bar{y}_1 + n_2\bar{y}_2}{n}\,\bar{y}_1\right) = E\left(\frac{n_1}{n}\bar{y}_1^2 + \frac{n_2}{n}\bar{y}_1\bar{y}_2\right)$$

$$= \frac{n_1}{n}E(\bar{y}_1^2) + \frac{n_2}{n}E(\bar{y}_1)E(\bar{y}_2)$$

$$= \frac{n_1}{n}[V(\bar{y}_1) + \bar{Y}^2] + \frac{n_2}{n}\bar{Y}^2 = \frac{n_1}{n}\left(\frac{S^2}{n_1} + \bar{Y}^2\right) + \frac{n_2}{n}\bar{Y}^2$$

$$= \frac{S^2}{n} + \frac{n_1}{n}\bar{Y}^2 + \frac{n_2}{n}\bar{Y}^2 = \frac{S^2}{n} + \bar{Y}^2 \qquad (2)$$

Now

$$V(\bar{y}) = E(\bar{y}^2) - \bar{Y}^2 \quad \text{or} \quad E(\bar{y}^2) = V(\bar{y}) + \bar{Y}^2 = \frac{S^2}{n} + \bar{Y}^2 \qquad (3)$$

In view of equations (1), (2), and (3), we get

$$Cov(\bar{y},\, \bar{y}_1 - \bar{y}) = \left(\frac{S^2}{n} + \bar{Y}^2\right) - \left(\frac{S^2}{n} + \bar{Y}^2\right) = 0\,.$$

Exercise: A population has three units U_1, U_2 and U_3 with variates Y_1, Y_2 and Y_3 respectively. It is required to estimate the population total Y by selecting a sample of two units. Let the sampling and estimation procedures be as follows:

Sample (s)	$P(s)$	Estimator t	Estimator t'
(U_1, U_2)	$1/2$	$Y_1 + 2Y_2$	$Y_1 + 2Y_2 + Y_1^2$
(U_1, U_3)	$1/2$	$Y_1 + 2Y_3$	$Y_1 + 2Y_3 - Y_1^2$

Prove that both t and t' are unbiased for Y and find their variances. Comment on the estimators.

Solution: By definition

$$E(t) = \sum_i t_i\; p(t_i) = \frac{1}{2}(Y_1 + 2Y_2 + Y_1 + 2Y_3) = Y\,.$$

This shows that estimator t is unbiased for Y.

$$E(t^2) = \frac{1}{2}[(Y_1 + 2Y_2)^2 + (Y_1 + 2Y_3)^2] = \frac{1}{2}(Y_1^2 + 4Y_2^2 + 4Y_1Y_2 + Y_1^2 + 4Y_3^2 + 4Y_1Y_3)$$

$$= Y_1^2 + 2Y_2^2 + 2Y_3^2 + 2Y_1Y_2 + 2Y_1Y_3\,.$$

Therefore,

$$V(t) = E(t^2) - [E(t)]^2 = Y_1^2 + 2Y_2^2 + 2Y_3^2 + 2Y_1Y_2 + 2Y_1Y_3 - (Y_1 + Y_2 + Y_3)^2$$

$$= Y_2^2 + Y_3^2 - 2Y_2Y_3 = (Y_2 - Y_3)^2\,.$$

Similarly,

$E(t') = \sum_i t_i' \, p(t_i') = \frac{1}{2}(Y_1 + 2Y_2 + Y_1^2 + Y_1 + 2Y_3 - Y_1^2) = Y$, hence, t' is unbiased for Y.

$$E(t'^2) = \frac{1}{2}[(Y_1 + 2Y_2 + Y_1^2)^2 + (Y_1 + 2Y_3 - Y_1^2)^2]$$

$$= \frac{1}{2}(Y_1^4 + 2Y_1^3 + Y_1^2 + 4Y_1^2 Y_2 + 4Y_1 Y_2 + 4Y_2^2 + Y_1^4 - 2Y_1^3$$

$$+ Y_1^2 - 4Y_1^2 Y_3 + 4Y_1 Y_3 + 4Y_3^2)$$

$$= Y_1^4 + Y_1^2 + 2Y_1^2 Y_2 + 2Y_1 Y_2 + 2Y_2^2 - 2Y_1^2 Y_3 + 2Y_1 Y_3 + 2Y_3^2 .$$

Therefore,

$$V(t') = E(t'^2) - [E(t')]^2$$

$$= Y_1^4 + Y_1^2 + 2Y_1^2 Y_2 + 2Y_1 Y_2 + 2Y_2^2 - 2Y_1^2 Y_3 + 2Y_1 Y_3 + 2Y_3^2 - (Y_1 + Y_2 + Y_3)^2$$

$$= (Y_2 - Y_3)^2 + Y_1^2 (Y_1^2 + 2Y_2 - 2Y_3)$$

$$= V(t) + Y_1^2 (Y_1^2 + 2Y_2 - 2Y_3) .$$

We conclude that both linear estimator t and quadratic estimator t' are unbiased; among which estimator has minimum variance depends on the variate values.