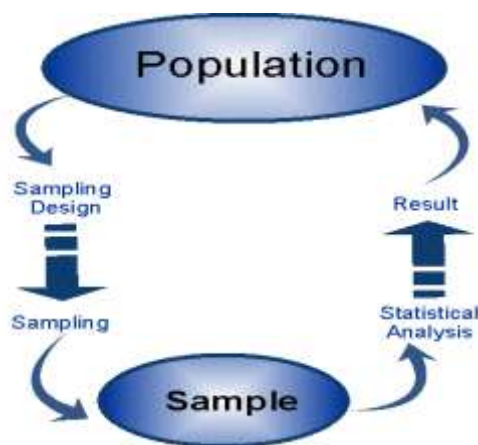


Course Handouts
Sampling Techniques
Virtual University of Pakistan

BASIC CONCEPTS**Q . Define Sampling Techniques?**

Ans: The data is very important for decision making. Appropriate method of data collection is desired. How best to obtain the data or sample. How best to use the data/sample to estimate the characteristic of the whole population. There are two parts of any sampling strategy. First is the selection procedure. Second is the estimation procedure.

**Q .Define sampling?**

Ans: The method of selecting the sample is called sampling design.

Q . Define sample?

Ans: A perfect sample would be a version of population, mirroring every characteristic of population. (Lohr, 2nd edition). A **good sample** is a small but representative part of the population.

Q . What is difference between Observation and Sampling Units?

Ans: A unit that can be selected for a sample is called sampling unit. An object on which a measurement is taken is called observation unit.

Q . Define sampling frame?

Ans: The sampling frame is the list/map from which the potential sampling units are drawn. For example, List of all the class rooms , Map of area containing farms.

Q . Distinguish between parameter and statistic?

Ans: A parameter is any summary number described the whole population. A **Statistic** is any summary number obtained from a sample.



**Q . Define sampling error?**

Ans: The errors in the estimates because sample is the part of population. The difference between the parameter and statistic is considered a sampling error.

Sampling Error= Statistic-Parameter.

Q . Define non-sampling error?

Ans: These errors can occur even the whole population is studied. Non-sampling error cannot attributed to the sampling.

Q . What is selection bias?

Ans: When a random process is not considered in selection of sample. Some part of the population is ignored during selection of sample. Some results of study may not be accurate.

Q . What is measurement error?

Ans: When the response has a tendency to differ from the true value in one direction. problem with the measurement instrument.

Q . Define Nonresponse?

Ans: The individuals chosen for the sample are not ready to participate in the survey. It is called nonresponse. This is a type of selection bias.

Q . What is probability sample?

Ans: This is a sample which is obtained when each unit of the population has some chances for being included into the sample. This is obtained by a random process.

Example: A bag consists on one white and one black ball.

White ball
1

Black ball
1

$$P(\text{White}) = \frac{\text{Number of white balls}}{\text{Number of total balls}}$$

$$P(\text{White}) = \frac{1}{2}$$

Q . Define randomness?

Ans: When a unit has some chances of selection in the sample. When we are not certain about the selection of any unit

Q . How many types of probability sampling?

Ans: Simple random sampling, stratified sampling, cluster sampling, systematic sampling, multistage /multiphase sampling.

Q. Define Non-probability sampling?

Ans: No random process is followed for the selection of units. The population is evenly distributed with respect to the characteristics of interest.

Q . What is convenience sampling?

Ans: Respondents are selected according to convenience of researcher. This is also called accidental sampling, opportunity sampling or grab sampling.

Q. Define purposive sampling?

Ans: The units which are appropriate to meet the objective of study. The selection is based on the judgment of researcher.

Q. What is quota sampling?

Ans: The population is divided into subgroups on the basis of similar characteristics. The subgroups are called “Quota”. The nonrandom selection is made in each quota.

Q . Define snow-ball sampling?

Ans: This is also called chain sampling so it works like a chain. The selected subject asked for the assistance to reach the other subjects.

Q. Define simple random sampling?

Ans: Each member of the population has an equal probability of being included in the sample. Each sample of size 'n' has an equal probability of being selected in the sample.

Q. What is the difference between sampling with replacement and without replacement?

Ans: Sampling With Replacement: If the unit is replaced before the selection of next unit then sampling is with replacement (WR).

$$\text{Total samples} = N^n, \quad P(S) = \frac{1}{N^n}$$

Example: Consider a population of size $N=3$, as A, B, C

The possible samples of size $n=2$ are

$S_1=(A,A)$, $S_2=(A,B)$, $S_3=(A,C)$, $S_4=(B,A)$, $S_5=(B,B)$, $S_6=(B,C)$, $S_7=(C,A)$, $S_8=(C,B)$, $S_9=(C,C)$

$$P(S_i) = \frac{1}{N^n} = \frac{1}{3^2} = \frac{1}{9}$$

Sampling without Replacement: If the unit is not replaced before the selection of next unit then sampling is without replacement (WOR).

$$\text{Total Samples} = {}^N C_n, \quad P(S) = \frac{1}{{}^N C_n}$$

Example: Consider a population of size $N=3$, as A, B, C

The possible samples of size $n=2$ are $S_1=(A,B)$, $S_2=(A,C)$, $S_3=(B,C)$

$$P(S_1) = P(S_2) = P(S_3) = \frac{1}{3} = \frac{1}{{}^3 C_2} = \frac{1}{{}^N C_n}$$

$${}^N C_n = \frac{N!}{n!(N-n)!}$$

Q. Define stratified sampling?

Ans: When population is heterogeneous with respect to characteristics of interest. Divide the population into homogeneous subgroups called strata. Select the simple random sample from each subgroup.

Q. Define linear systematic sampling?

Ans: First unit is selected randomly from first 'k' units. Every k^{th} unit is included in the sample. The interval 'k' is calculated by dividing the population size to the sample size.

Q . Define circular systematic sampling?

Ans: 'k' will be decided by rounding down to the integer nearest to N/n . First unit will be selected at random from 1 to N units. Every k^{th} unit is included in the sample.

Q . Define cluster sampling?

Ans: A cluster is the sampling unit consisting on the observation units. Any sampling method can be used for selection of clusters. All the units within a cluster are studied.

SIMPLE RANDOM SAMPLING

Q . Prove that sample mean is unbiased estimator for population mean under simple random sampling with/without replacement.

Ans: Unbiased Estimator: The estimator is unbiased if its expected value is equal to the population parameter

$$E(\text{estimator}) = \text{Parameter}$$

The expectation of a random variable is defined as the sum of the products of the probabilities and the variable.

$$E(y_i) = \sum_{i=1}^N p_i Y_i$$

Theorem: In simple random sampling with replacement and without replacement, the sample mean, $\bar{y}$, is an unbiased estimator of the population mean $\bar{Y}$.

Under SRSWR: $\bar{y}$ is unbiased estimate of $\bar{Y}$ when

$$E(\bar{y}) = \bar{Y}$$

n= sample size

N=population size

y_i = observe variable on the ith sample unit.

The population mean $\bar{Y}$ is

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

An estimator of population total is

$$y' = \frac{N}{n} \sum_{i=1}^n y_i$$

$\bar{y}$ is unbiased estimate of $\bar{Y}$ when

$$E(\bar{y}) = \bar{Y}$$

The expectation of $\bar{y}$, by definition is,

$$\begin{aligned}
 E(\bar{y}) &= E\left(\frac{1}{n} \sum_{i=1}^n y_i\right) \\
 &= \frac{1}{n} \sum_{i=1}^n E(y_i) \\
 &= E(y_i) \\
 E(y_i) &= \sum_{i=1}^N p_i Y_i \\
 &= \sum_{i=1}^N \frac{1}{N} Y_i \\
 &= \frac{1}{N} \sum_{i=1}^N Y_i \\
 &= \bar{Y} \\
 E(\bar{y}) &= \bar{Y}
 \end{aligned}$$

Under SRSWOR: Since there are $\binom{N}{n} = m$ possible distinct samples for without replacement, then

$$E(\bar{y}) = \sum_{k=1}^m \left[\sum_{i=1}^n \frac{y_i}{n} \right] / \binom{N}{n}$$

Now in $\binom{N}{n}$ possible samples, each unit is appearing $\binom{N-1}{n-1}$ times,

$$\begin{aligned}
 E(\bar{y}) &= \sum_{k=1}^m \left[\sum_{i=1}^n \frac{y_i}{n} \right] / \binom{N}{n} \\
 &= \frac{1}{n} \binom{N-1}{n-1} \sum_{i=1}^N Y_i / \binom{N}{n} \\
 &= \frac{1}{N} \sum_{i=1}^N Y_i \\
 &= \bar{Y}
 \end{aligned}$$

Q . Derive the Variance of Sample Mean under Simple Random Sampling (SRS) with replacement and without replacement?

Ans: Variance of Sample Mean Under SRSWOR:

The variance of the sample mean for simple random sampling without replacement, $\bar{y}_{wor}$, is

$$\begin{aligned} Var(\bar{y}_{wor}) &= \frac{N-n}{N} \frac{S^2}{n} \\ &= (1-f) \frac{S^2}{n} \end{aligned}$$

$$\text{where, } f = \frac{n}{N}$$

Proof:

We know that

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

The variance of this estimator is

$$Var(\bar{y}) = \frac{1}{n^2} Var\left(\sum_{i=1}^n y_i\right)$$

Since the variance of the sum of the random variables is equal to the sum of the variance of random variables plus the sum of the covariance.

$$Var(\bar{y}) = \frac{1}{n^2} \left[\sum_{i=1}^n Var(y_i) + \sum_{\substack{i,j=1 \\ j \neq i}}^n Cov(y_i, y_j) \right]$$

$$Var(y_i) = E(y_i - \bar{Y})^2 = E(y_i^2) - \bar{Y}^2$$

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

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

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

$$= \frac{N-1}{N} S^2$$

$$\begin{aligned}
 Cov(y_i y_j) &= E(y_i y_j) - \bar{Y}^2 \\
 &= \frac{1}{N(N-1)} \sum_{i,j=1}^N \sum_{j \neq i} Y_i Y_j - \bar{Y}^2 \\
 &= \frac{1}{N(N-1)} \left[\left(\sum_{i=1}^N Y_i \right)^2 - \sum_{i=1}^N Y_i^2 \right] - \bar{Y}^2 \\
 &= \frac{1}{N(N-1)} \left[\left(\sum_{i=1}^N Y_i \right)^2 - \sum_{i=1}^N Y_i^2 \right] - \bar{Y}^2 \\
 &= \frac{1}{N(N-1)} \left[\left(\sum_{i=1}^N Y_i \right)^2 - \sum_{i=1}^N Y_i^2 - \frac{\left(\sum_{i=1}^N Y_i \right)^2}{N^2} N(N-1) \right] \\
 Cov(y_i y_j) &= -\frac{1}{N(N-1)} \left(\sum_{i=1}^N Y_i^2 - \frac{\left(\sum_{i=1}^N Y_i \right)^2}{N} \right) = -S^2 / N
 \end{aligned}$$

After substitution we get

$$\begin{aligned}
 Var(\bar{y}) &= \frac{1}{n^2} \left[\sum_{i=1}^n \frac{N-1}{N} S^2 - \sum_{i,j=1}^n \sum_{j \neq i} \frac{S^2}{n} \right] \\
 &= \frac{1}{n^2} \left[n \frac{N-1}{N} S^2 - \frac{n(n-1)}{n} S^2 \right] = \frac{N-n}{N} \frac{S^2}{n}
 \end{aligned}$$

Variance of sample mean for SRSWR:

The variance of the sample mean for simple random sampling with replacement, $\bar{y}_{wr}$, is

$$\begin{aligned}
 Var(\bar{y}_{wr}) &= \frac{N-1}{N} \frac{S^2}{n} \\
 &= \left(1 - \frac{1}{N}\right) \frac{S^2}{n}
 \end{aligned}$$

Proof: We know that

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

The variance of this estimator is

$$Var(\bar{y}) = \frac{1}{n^2} Var\left(\sum_{i=1}^n y_i\right)$$

Since the variance of the sum of the random variables is equal to the sum of the variance of random variables plus the sum of the covariance

$$\begin{aligned} Var(\bar{y}) &= \frac{1}{n^2} \left[\sum_{i=1}^n Var(y_i) \right] \\ Var(y_i) &= E(y_i - \bar{Y})^2 = E(y_i^2) - \bar{Y}^2 \\ &= \frac{1}{N} \sum_{i=1}^N Y_i^2 - \bar{Y}^2 \\ &= \frac{N-1}{N} S^2 \\ Var(\bar{y}) &= \frac{1}{n^2} \left[\sum_{i=1}^n \frac{N-1}{N} S^2 \right] \\ Var(\bar{y}) &= \frac{N-1}{Nn} S^2 \end{aligned}$$

Q. Derive unbiased estimator of variance under SRS with replacement and without replacement.

Ans: For simple random sampling without replacement, s^2 is an unbiased estimator of S^2 and for simple random sampling with replacement s^2 is an unbiased estimator of $S^2 (N-1)/N$.

Proof: For both simple random sampling without replacement and simple random sampling with replacement, we have

$$\begin{aligned} (n-1)s^2 &= \sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n \{ (y_i - \bar{Y}) - (\bar{y} - \bar{Y}) \}^2 \\ &= \sum_{i=1}^n (y_i - \bar{Y})^2 + \sum_{i=1}^n (\bar{y} - \bar{Y})^2 - 2 \sum_{i=1}^n (y_i - \bar{Y}) (\bar{y} - \bar{Y}) \end{aligned}$$

$$\begin{aligned}
 &= \sum_{i=1}^n (y_i - \bar{Y})^2 + n(\bar{y} - \bar{Y})^2 - 2(\bar{y} - \bar{Y}) \sum_{i=1}^n (y_i - \bar{Y}) \\
 &= \sum_{i=1}^n (y_i - \bar{Y})^2 + n(\bar{y} - \bar{Y})^2 - 2n(\bar{y} - \bar{Y})^2 \\
 &= \sum_{i=1}^n (y_i - \bar{Y})^2 - n(\bar{y} - \bar{Y})^2
 \end{aligned}$$

Taking expectation

$$(n-1) E(s^2) = n E(y_i - \bar{Y})^2 - n E(\bar{y} - \bar{Y})^2$$

As we know that $E(y_i - \bar{Y})^2 = \frac{N-1}{N} S^2$, therefore

$$(n-1) E(s^2) = n \frac{N-1}{N} S^2 - n E(\bar{y} - \bar{Y})^2.$$

Now $E(\bar{y} - \bar{Y})^2$ is the variance of the sample mean, which is $\left(1 - \frac{n}{N}\right) \frac{S^2}{n}$ for simple random sampling without replacement and is $\left(1 - \frac{1}{N}\right) \frac{S^2}{n}$ for simple random sampling with replacement.

Hence for simple random sampling without replacement

$$\begin{aligned}
 E(s^2) &= \frac{n}{n-1} \frac{N-1}{N} S^2 - \left(\frac{N-n}{N}\right) \frac{n}{n-1} \frac{S^2}{n} \\
 &= S^2
 \end{aligned}$$

and for simple random sampling with replacement

$$\begin{aligned}
 E(s^2) &= \frac{n}{n-1} \frac{N-1}{N} S^2 - \frac{N-1}{N} \frac{n}{n-1} \frac{S^2}{n} \\
 &= \frac{N-1}{N} S^2
 \end{aligned}$$

Standard error

The standard error of the sample mean $\bar{y}$ is

$$SE(\bar{y}) = \sqrt{Var(\bar{y})}$$

$$= \begin{cases} S\sqrt{\frac{1-f}{n}} & \text{for srswor} \\ S\sqrt{\frac{1}{n}\left(1-\frac{1}{N}\right)} & \text{for srswr} \end{cases}$$

Q . What is R?

Ans: R was initially written by **Ross Ihaka** and **Robert Gentleman** at the Department of Statistics of the University of Auckland. The R software is derived from an original set of notes describing the S and S-Plus environments.

Downloading R <https://cran.r-project.org/>



- CRAN
- Mirrors
- What's new?
- Task Views
- Search
- About R
- R Homepage
- The R Journal
- Software
- R Sources
- R Binaries
- Packages
- Other
- Documentation
- Manuals
- FAQs
- Contributed

The Comprehensive R Archive Network

Download and Install R

Precompiled binary distributions of the base system and contributed packages, **Windows and Mac** users most likely want one of these versions of R:

- [Download R for Linux](#)
- [Download R for \(Mac\) OS X](#)
- [Download R for Windows](#)

R is part of many Linux distributions, you should check with your Linux package management system in addition to the link above.

Source Code for all Platforms

Windows and Mac users most likely want to download the precompiled binaries listed in the upper box, not the source code. The sources have to be compiled before you can use them. If you do not know what this means, you probably do not want to do it!

- The latest release (Friday 2017-06-30, Single Candle) [R-3.4.1 tar.gz](#), read [what's new](#) in the latest version.
- Sources of [R alpha and beta releases](#) (daily snapshots, created only in time periods before a planned release).
- Daily snapshots of current patched and development versions are [available here](#). Please read about [new features and bug fixes](#) before filing corresponding feature requests or bug reports.
- Source code of older versions of R is [available here](#).
- Contributed extension [packages](#)



CRAN
[Mirrors](#)
[What's new?](#)
[Task Views](#)
[Search](#)

About R
[R Homepage](#)
[The R Journal](#)

Software
[R Sources](#)
[R Binaries](#)
[Packages](#)
[Other](#)

Documentation
[Manuals](#)
[FAQs](#)

R for Windows

Subdirectories:

[base](#) Binaries for base distribution (managed by Duncan Murdoch). This is what you want to [install R for the first time](#).

[contrib](#) Binaries of contributed CRAN packages (for R >= 2.11.x; managed by Uwe Ligges). There is also information on [third party software](#) available for CRAN Windows services and corresponding environment and make variables.

[old contrib](#) Binaries of contributed CRAN packages for outdated versions of R (for R < 2.11.x; managed by Uwe Ligges).

[Rtools](#) Tools to build R and R packages (managed by Duncan Murdoch). This is what you want to build your own packages on Windows, or to build R itself.

Please do not submit binaries to CRAN. Package developers might want to contact Duncan Murdoch or Uwe Ligges directly in case of questions / suggestions related to Windows binaries.

You may also want to read the [R FAQ](#) and [R for Windows FAQ](#).

Note: CRAN does some checks on these binaries for viruses, but cannot give guarantees. Use the normal precautions with downloaded executables.



CRAN
[Mirrors](#)
[What's new?](#)
[Task Views](#)
[Search](#)

About R
[R Homepage](#)
[The R Journal](#)

Software
[R Sources](#)
[R Binaries](#)
[Packages](#)
[Other](#)

R-3.4.1 for Windows (32/64 bit)

[Download R 3.4.1 for Windows](#) (75 megabytes, 32/64 bit)

[Installation and other instructions](#)

[New features in this version](#)

If you want to double-check that the package you have downloaded matches the package distributed by CRAN, you can compare the [md5sum](#) of the .exe to the [fingerprint](#) on the master server. You will need a version of md5sum for windows: both [graphical](#) and [command line versions](#) are available.

Frequently asked questions

- [Does R run under my version of Windows?](#)
- [How do I update packages in my previous version of R?](#)
- [Should I run 32-bit or 64-bit R?](#)

Please see the [R FAQ](#) for general information about R, and the [R Windows FAQ](#) for Windows-specific information.

other builds


A screenshot of the R GUI (32-bit) window. The 'R Console' pane is active, displaying the following text:

```
R version 3.3.3 (2017-03-06) -- "Another Cabbage"
Copyright (C) 2017 The R Foundation for Statistical Computing
Platform: i386-w64-mingw32/i386 (32-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help,
Type 'q()' to quit R.

[Previously saved workspace restored]

> |
```

A screenshot of the R GUI (32-bit) window with the 'File' menu open. The menu options are: Search R code..., New script (highlighted), Open script..., Display file(s)..., Load Workspace..., Save Workspace... (Ctrl+S), Load History..., Save History..., Change dir..., Print... (Ctrl+P), Save to File..., and Exit. The background shows the same R console text as the previous screenshot.



R Commands	<code>y<-c(1,2,3,4,5)</code>
#indexing	<code>Y[2]</code>
#Mean	<code>mean(y)</code>
#Standard Deviation	<code>sd(y)</code>
#Variance	<code>var(y)</code>

Q . How sampling is done by using R?

Ans: Sampling With replacement

Sample (y, size, replace=TRUE)

Sampling With replacement

Sample (y, size, replace=FALSE)

OR

Sample(y, size)

Consider a Population consisting on

111, 150, 121, 198, 112, 136, 114, 129, 117, 115, 186, 110, 121, 115, 114

Enter population data in R

```
yp <- c(111, 150, 121, 198, 112, 136, 114, 129, 117, 115, 186, 110, 121, 115, 114)
```

Taking sample by SRSWR in R

```
ys <- sample(yp,5)
```

Output: 114 121 111 186 150

Mean of sample `mean(ys)`

Output: 136.4

Standard Deviation `sd(ys)`

Output: 31.73

```
yp <- c(11,150, 121, 198, 12, 136, 14, 129, 17, 115, 186, 110, 121, 15, 14)
#population size          N<-length(yp)
#sample of size 5        ys <- sample(yp,5)
#sample size             n<-length(ys)
#mean of the sample      mys<-mean(ys)
#variance of the sample  vys<-var(ys)
#variance of ybar        vybar<-(1-n/N) *var(yp)/n
#standard error          sdr<-sqrt(vybar)
#estimate of population total ept<-N*mys
#estimated variance       vept<-N^2*vybar
#estimated standard error sdept<-sqrt(vept)
yp <- c(11,150, 121, 198, 12, 136, 14, 129, 17, 115, 186, 110, 121, 15, 14)
#sample of size 5        ys <- sample(yp,5, replace=TRUE)
```

Q. Define the Simulation Study for Sampling Strategy?

Ans: The simulation study is useful to evaluate a sampling strategy. We can generate the populations considering specific situations. Generate the population. The sample of size ' n ' is obtained ' k ' times. From each sample the estimator is obtained. The variance of ' k ' estimators is calculated for examining the efficiency.

Consider a Population consisting on

111, 150, 121, 198, 112, 136, 114, 129, 117, 115, 186, 110, 121, 115, 114

Enter population data in R

```
yp <- c(111,150, 121, 198, 112, 136, 114, 129, 117, 115, 186, 110, 121, 115, 114)
```

```
Taking sample by SRSWOR in R      ys <- sample(yp,5)
                                   mean(ys)
```

Now we will take ' k ' samples in R.

R program for k Samples

Suppose, $k=10$, $n=5$.

The 'for loop' is used to repeat the statements.

```
m1<-c();
yp <- c(111,150, 121, 198, 112, 136, 114, 129, 117, 115, 186, 110,
121, 115, 114)
```

```
for (i in 1:10){  
  s <- sample(yp,5)  
  m1[i] <- mean(s)  
}
```

Output

Sample means with 10 repetitions

149.8 129.2 140.8 132.4 118.2 117.6 118.4 118.0 132.6 132.6

k Samples

Suppose, $k=10000$, $n=5$.

The 'for loop' is used to repeat the statements.

```
m1<-c();  
yp <- c(111,150, 121, 198, 112, 136, 114, 129, 117, 115, 186, 110, 121, 115, 114)  
for (i in 1:10000){  
  s <- sample(yp,5)  
  m1[i] <- mean(s)  
}var(m1)
```

Output

On first run, the result was 101.7741

On second run, the result was 102.1403

On third run, the result was 100.2455

Q . Obtain the 10000 random sample of size 6 under SRSWOR using the following Population and find the mean.

111, 150, 121, 198, 112, 136, 114, 129, 117, 115, 186, 110, 121, 115, 114

Ans:

Given population is

111, 150, 121, 198, 112, 136, 114, 129, 117, 115, 186, 110, 121, 115, 114

Enter population data in R

```
yp <- c(111,150, 121, 198, 112, 136, 114, 129, 117, 115, 186, 110, 121, 115, 114)
```

Here, $k=15000$, $n=6$.

The 'for loop' is used to repeat the statements.

```
m1<-c();
```

```
yp <- c(111,150, 121, 198, 112, 136, 114, 129, 117, 115, 186, 110, 121, 115, 114)
for (i in 1:10000){
  s <- sample(yp,6)
  m1[i] <- mean(s)
}
var(m1)
```

Out put

On first run, the result was	75.4034
On second run, the result was	75.9143
On third run, the result was	75.6871

Q . Obtain 1000 random number through normal distribution with mean 0 and variance 1 as population. Obtain the 10000 random sample of size 6 under SRSWOR using the population and find the mean.

Ans: 1000 values with mean=0 and standard deviation=1.

```
rnorm(n,mean,sd)
yp <- rnorm(1000,0,1)
Suppose, k=10000, n=5.
m1<-c();
yp <- rnorm(1000,0,1)
for (i in 1:10000){
  s <- sample(yp,5)
  m1[i] <- mean(s)
}
var(m1)
```

Output

On first run, the result was	0.2044265
On second run, the result was	0.1914794
On third run, the result was	0.198996

Q . Explain the Estimation of Sample Size for Mean Estimation by using an example?

Ans: Sample size Estimation

Let 'd' be the margin of error with some probability α by which sampling value differs from population value. Specifying a maximum allowable difference 'd' between the estimate and the true value and allowing for a small probability α .

The probability of the margin of error being less than d is given by

$$P\left[|\bar{y} - \bar{Y}| > d \right] = \alpha$$

We assume that sample mean is normally distributed, i.e.

$$t = \frac{\bar{y} - \bar{Y}}{S.E(\bar{y})}$$

$$t = \frac{|\bar{y} - \bar{Y}|}{S.E(\bar{y})} = \frac{d}{S.E(\bar{y})}$$

$$d = |\bar{y} - \bar{Y}| = t \cdot S.E(\bar{y})$$

Consider the case of sampling without replacement.

$$d = t \sqrt{\frac{N-n}{N} \frac{S^2}{n}}$$

$$d^2 = t^2 \left(\frac{N-n}{N} \frac{S^2}{n} \right)$$

$$\frac{d^2 N n}{t^2} = (N S^2 - n S^2)$$

$$n = \frac{(t S / d)^2}{1 + \frac{(t S / d)^2}{N}}$$

When the population size is large

$$n = (t S / d)^2$$

Example: A quality manager is interested to estimate the mean diameter of bolts produced in the last week. Determine the sample size to obtain 90% confidence level for population if the error in the estimate will not be more than 3cm. A pilot sample yields a standard deviation of 30 cm.

Solution:

$$n = (t S / d)^2$$

$$n = \frac{(1.645)^2 (30)^2}{(3)^2}$$

$$n = 270.60$$

Sample size when cost is involved

Cost plays major role in conduct of surveys

$$C = C_0 + n C_1$$

$$C = \text{Total Cost}$$

$$C_0 = \text{Fixed Cost}$$

$$C_1 = \text{Cost Per Unit}$$

$$d(\bar{y} - \bar{Y})^2$$

The loss due to sample mean is not equal to population mean where d is a constant.

$$L(n) = E(C_0 + nC_1) + E[d(\bar{y} - \bar{Y})^2]$$

$$= C_0 + C_1E(n) + d E(\bar{y} - \bar{Y})^2$$

$$L(n) = C_0 + C_1n + d \text{Var}(\bar{y})$$

$$L(n) = C_0 + C_1n + d \frac{S^2}{n}$$

Derivative with respect to n and equating to zero will produce optimum value as

$$n = \sqrt{dS^2 / C_1}$$

Q . Estimation of sample size for estimation of proportion?

Ans: Suppose we have N population units i.e. $Y_1, Y_2, \dots, Y_i, \dots, Y_N$
 $y_i = 1$ if i th unit possesses a certain attribute and 0 otherwise.

The population proportion is defined

The sample proportion is

$$\bar{Y} = \sum_{i=1}^N Y_i / N = A / N = P$$

$$\bar{y} = \sum_{i=1}^n y_i / n = a / n = p$$

Since Y_i takes the values as 1 and 0.

$$\sum_{i=1}^N Y_i^2 = A = NP$$

The same is the case for sample.

$$\sum_{i=1}^n y_i^2 = a = np$$

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

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

$$S^2 = \frac{N}{N-1} P(1-P)$$

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

Similarly

$$s^2 = npq / (n-1)$$

For SRSWOR

$$\text{var}(p_{\text{wor}}) = \frac{N-n}{N} \frac{pq}{n-1},$$

For SRSWR

$$\text{var}(p_{\text{wr}}) = \frac{N-1}{N} \frac{pq}{n-1},$$

When fpc is ignored

$$\text{var}(p) = pq / (n-1)$$

For SRSWOR

$$\text{VAR}(\bar{y}_{\text{wor}}) = \frac{N-n}{Nn} S^2,$$

$$\text{VAR}(p_{\text{wor}}) = \frac{N-n}{Nn} \frac{NPQ}{N-1},$$

$$\text{VAR}(p_{\text{wor}}) = \frac{N-n}{n} \frac{PQ}{N-1},$$

Standard Error of Proportion

$$S.E(p_{\text{wor}}) = \sqrt{\frac{N-n}{n} \frac{PQ}{N-1}}$$

For the large sample size

$$\text{var}(p) = pq / n$$

The standard error

$$S.E(p) = \sqrt{pq / n}$$

Sample size Estimation

$$t = \frac{|p - P|}{S.E(p)} = \frac{d}{S.E(p)}$$

$$d = t \times S.E(p)$$

$$n = \frac{t^2 PQ / d^2}{1 + \frac{1}{N} (t^2 PQ / d^2 - 1)}$$

When sample size is large.

$$n = t^2 \frac{pq}{d}$$

$$d = t \times S.E(p)$$

$$d^2 = t^2 \times \text{VAR}(p)$$

$$d^2 = t^2 \left[\frac{N-n}{N-1} \frac{PQ}{n} \right]$$

$$d^2 = \left[\frac{t^2 NPQ}{n(N-1)} - \frac{t^2 PQ}{N-1} \right]$$

$$\frac{t^2 NPQ}{(N-1)} = n \left[d^2 + \frac{t^2 PQ}{N-1} \right]$$

$$n = \frac{\frac{t^2 NPQ}{(N-1)}}{\left[d^2 + \frac{t^2 PQ}{N-1} \right]}$$

$$n = \frac{\frac{t^2 NPQ}{(N-1)}}{\frac{1}{N-1} \left[(N-1)d^2 + t^2 PQ \right]}$$

$$n = \frac{t^2 NPQ}{\left[(N-1)d^2 + t^2 PQ \right]}$$

$$n = \frac{\frac{t^2 PQ}{d^2}}{\frac{1}{N} \left[(N-1) + \frac{t^2 PQ}{d^2} \right]}$$

$$n = \frac{t^2 PQ / d^2}{1 + \frac{1}{N} (t^2 PQ / d^2 - 1)}$$

Q. Derive the Confidence Interval for mean and proportion. Also explain with the help of R.

Ans: Confidence Interval for Mean (Known variance)

The interval estimation of population mean

$$\bar{y} \pm z_{1-\alpha/2} S.E(\bar{y})$$

Lower Limit

$$\bar{y} - z_{1-\alpha/2} S.E(\bar{y})$$

Upper Limit

$$\bar{y} + z_{1-\alpha/2} S.E(\bar{y})$$

Confidence Interval for Mean (Unknown variance)

The interval estimation of population mean when population standard deviation is not known

$$\bar{y} \pm t_{1-\alpha/2, v} S.E(\bar{y})$$

Example: The XYZ Company produces cold drink diet cans with standard deviation of the amount poured into cans by automatic filling machine is 1.4 ml (milliliter). A random sample is taken of the amount of filling in cans were 281, 278, 276, 282, 280, 279, 278, 280. Suppose that population of filling amount follows normal distribution. Determine 95% confidence interval for the mean amount in all cans filled by the machine.

Solution:

Given

$$\sigma = 1.4$$

$$\bar{y} = \frac{\sum y}{n} = 279.25$$

$$S.E(\bar{y}) = \frac{\sigma}{\sqrt{n}} = \frac{1.4}{\sqrt{8}} = 0.495$$

The 95% confidence limits will be

$$\bar{y} \pm 1.96 S.E(\bar{y})$$

The Lower Limit

$$279.25 - 1.96(0.495) = 278.2998$$

The Upper Limit

$$279.25 + 1.96(0.495) = 280.2201$$

Confidence Interval of Mean Using R

```
yp <- c(111, 150, 121, 198, 112, 136, 114, 129, 117, 115, 186, 110, 121, 115, 114)
```

```
#population size      N<-length(yp)
```

```
#sample of size 5     ys <- sample(yp,5)
```

```
#sample size          n<-length(ys)
```

```
#mean of the sample    mys<-mean(ys)
```

```
#variance of the sample      vys<-var(ys)
#variance of ybar            vybar<-var(yp)/n
#standard error              sdr<-sqrt(vybar)
```

The interval estimation of population mean when population standard deviation is not known

```
error <- qnorm(0.975)*sdr
```

The respective lower and upper limits are

```
left <- mys-error
right <- mys+error
```

Confidence Interval for Proportion

The interval estimation of population pro-portion is

$$p \pm z_{1-\alpha/2} S.E(p)$$

A simple random sample of 80 students is taken from a population of 470 students in a department. The total number of smokers in the sample was 22.

Confidence Interval of Smokers

Smoker are 22

$$p = \frac{22}{80} = 0.275$$

$$q = 1 - p = 0.725$$

$$S.E(p) = \sqrt{\frac{pq}{n}} = 0.0499$$

The 95% confidence limits will be

$$p \pm z_{1-\alpha/2} S.E(p)$$

$$p \pm 1.96 S.E(p)$$

The Lower Limit

$$0.275 - 1.96(0.0499) = 0.1772$$

The Upper Limit

$$0.275 + 1.96(0.0499) = 0.3728$$

STRATIFIED SAMPLING

Notations

$$N = N_1 + N_2 + N_3 + \dots + N_h$$

Similarly,

$$n = n_1 + n_2 + n_3 + \dots + n_h$$

$$\bar{Y}_h = \frac{\sum_{i=1}^{N_h} Y_{hi}}{N_h} = \text{Stratum mean}$$

Sample mean of stratum

$$\bar{y}_h = \frac{\sum_{i=1}^{n_h} Y_{hi}}{n_h}$$

As in Simple Random Sampling

$$E(\bar{y}_h) = \bar{Y}_h$$

Sample Mean

$$\text{Var}(\bar{y}_h) = \frac{N_h - n_h}{N_h} \frac{S_h^2}{n_h}$$

$$\bar{y}_{st} = \sum_{h=1}^k N_h \bar{y}_h / N = \sum_{h=1}^k W_h \bar{y}_h$$

$$S_h^2 = \frac{1}{N_h - 1} \sum_{i=1}^{N_h} (Y_{hi} - \bar{Y}_h)^2$$

Estimated population total

$$s_h^2 = \frac{1}{n_h - 1} \sum_{i=1}^{n_h} (y_{hi} - \bar{y}_h)^2$$

$$y'_{st} = N \bar{y}_{st}$$

Q. Prove that sample mean is unbiased estimator of population mean under stratified sampling.

In stratified sampling, the sample mean is an unbiased estimator of population mean i.e.

$$E(\bar{y}_{st}) = \bar{Y}_{st}$$

As in Simple Random Sampling

$$E(\bar{y}_h) = \bar{Y}_h$$

Taking expectation of the sample mean of stratified sample, we have:

$$E(\bar{y}_{st}) = \frac{1}{N} \sum_{h=1}^K N_h E(\bar{y}_h) = \bar{Y}_{st}$$

$$E(\bar{y}_h) = \bar{Y}_h$$

Similarly it can be proved that

$$E(y'_{st}) = Y$$

Q . In how many allocation methods variance of sample mean can be derived in stratified sampling?

Ans: Allocation of Sample Size

The variance of $\bar{y}_{st}$ can be derived by using following allocation methods:

- i. Arbitrary Allocation,
- ii. Proportional Allocation, and
- iii. Optimum Allocation.

Q . Drive the variance of sample mean in stratified sampling using arbitrary allocation?

Ans: Arbitrary Allocation

The total sample is allocated arbitrarily among the strata

Theorem

The variance of sample mean, $\bar{y}_{st}$ for stratified random sampling for finite population sampling is

$$Var(\bar{y}_{st}) = \frac{1}{N^2} \sum_{h=1}^K \left[N_h (N_h - n_h) \frac{S_h^2}{n_h} \right]$$

$$Var(\bar{y}_{st}) = \frac{1}{N^2} \sum_{h=1}^k N_h (N_h - n_h) \frac{S_h^2}{n_h}$$

$$Var(\bar{y}_{st}) = \frac{1}{N^2} \sum_{h=1}^k N_h^2 (N_h - n_h) \frac{S_h^2}{N_h n_h}$$

$$Var(\bar{y}_{st}) = \sum_{h=1}^k \frac{N_h^2}{N^2} (N_h - n_h) \frac{S_h^2}{N_h n_h}$$

We know that the mean for stratified random sampling design is:

$$\bar{y}_{st} = \frac{1}{N} \sum_{h=1}^k N_h \bar{y}_h$$

$$Var(\bar{y}_{st}) = \frac{1}{N^2} \sum_{h=1}^k N_h^2 Var(\bar{y}_h)$$

$$\begin{aligned} Var(\bar{y}_h) &= \frac{N_h - n_h}{N_h} \frac{S_h^2}{n_h} \\ Var(\bar{y}_{st}) &= \sum_{h=1}^k \frac{N_h^2}{N^2} (N_h - n_h) \frac{S_h^2}{N_h n_h} \\ Var(\bar{y}_{st}) &= \sum_{h=1}^k W_h^2 S_h^2 \frac{(N_h - n_h)}{N_h n_h} \\ Var(\bar{y}_{st}) &= \frac{1}{N^2} \sum_{h=1}^k \frac{N_h^2}{n_h} S_h^2 - \frac{1}{N^2} \sum_{h=1}^k N_h S_h^2 \\ Var(\bar{y}_{st}) &= \frac{1}{N^2} \sum_{h=1}^k N_h (N_h - n_h) \frac{S_h^2}{n_h} \\ Var(\bar{y}_{st}) &= \sum_{h=1}^k \frac{W_h^2 S_h^2}{n_h} - \sum_{h=1}^k \frac{W_h S_h^2}{N} \end{aligned}$$

For large value of N, the above equation transferred to

$$\begin{aligned} Var(\bar{y}_{st}) &= \sum_{h=1}^k W_h^2 S_h^2 / n_h \\ Var(\bar{y}_{st}) &= \sum_{h=1}^k \frac{W_h^2 S_h^2}{n_h} - \sum_{h=1}^k \frac{W_h S_h^2}{N} \end{aligned}$$

Variance of total

$$Var(y'_{st}) = Var(N \bar{y}_{st})$$

$$Var(y'_{st}) = \sum_{h=1}^k \frac{N_h^2}{n_h} S_h^2 - \sum_{h=1}^k N_h S_h^2$$

Q . Derive the Variance of Sample Mean in Stratified Sampling using Proportional Allocation?

Ans: Its Originally proposed by Bowley (1926) . If sampling fraction in all the strata is same, then the allocation is termed as proportional allocation. The sample size of hth stratum in this case given by

$$\begin{aligned} n_h / n &= N_h / N \\ n_h &= \frac{n}{N} N_h = n W_h \end{aligned}$$

$$Var(\bar{y}_{st}) = \sum_{h=1}^k \frac{N_h^2}{N^2} (N_h - n_h) \frac{S_h^2}{N_h n_h}$$

$$\begin{aligned} Var(\bar{y}_{st}) &= \sum_{h=1}^k W_h^2 S_h^2 \frac{(N_h - n_h)}{N_h n_h} \\ Var_{prop}(\bar{y}_{st}) &= \sum_{h=1}^k W_h^2 S_h^2 \frac{\left(N_h - \frac{nN_h}{N}\right)}{N_h \frac{nN_h}{N}} \\ Var_{prop}(\bar{y}_{st}) &= \sum_{h=1}^k W_h^2 S_h^2 \frac{(N - n)}{nN_h} \\ Var_{prop}(\bar{y}_{st}) &= \sum_{h=1}^k W_h^2 S_h^2 \frac{(N - n)}{nN_h} \end{aligned}$$

Q. Derive the variance of sample mean in stratified sampling using optimum allocation

Ans: The purpose of optimum allocation is to allocate n_h in such a way that minimum variance is achieved for a minimum cost. n_h are chosen either to minimize $Var(\bar{y}_{st})$ for a fixed sample size or cost is minimized for given variance.

The two aspect of optimum allocation are

- Sample size is proportional to stratum size and standard deviation of stratum (Neyman Allocation).
- Sample size is inversely proportional to cost.

Sample Size to Minimum Variance for Fixed Cost

In stratified random sampling $Var(\bar{y}_{st})$ will be minimum subject to the cost when n_h is proportional $W_h S_h / \sqrt{c_h}$ to i.e.

$$n_h = \left(W_h S_h / \sqrt{c_h} \right) / \left[\sum_{h=1}^K W_h S_h / \sqrt{c_h} \right]$$

The variance of sample mean $\bar{y}_{st}$ for stratified random sampling is

$$Var(\bar{y}_{st}) = \sum_{h=1}^k W_h^2 S_h^2 / n_h - \sum_{h=1}^k W_h^2 S_h^2 / N$$

where C = total cost,

$$C = C_0 + \sum_{h=1}^k C_h n_h$$

We introduce Lagrange's multiplier i.e.

$$Var(\bar{y}_{st}) = \sum_{h=1}^k \frac{W_h^2 S_h^2}{n_h} - \sum_{h=1}^k \frac{W_h S_h^2}{N}$$

Partially differentiating w.r.t. n_h and equating to zero

$$-\frac{W_h^2 S_h^2}{n_h^2} + \lambda C_h = 0$$

$$n_h \sqrt{\lambda} = W_h S_h / \sqrt{C_h}$$

$$n_h = \frac{W_h S_h / \sqrt{C_h}}{\sqrt{\lambda}}$$

$$\sum_{h=1}^k n_h = n = \frac{\sum_{h=1}^k W_h S_h / \sqrt{C_h}}{\sqrt{\lambda}}$$

$$\frac{n_h}{n} = \frac{W_h S_h / \sqrt{C_h}}{\sum_{h=1}^k W_h S_h / \sqrt{C_h}}$$

$$\frac{n_h}{n} = \frac{W_h S_h / \sqrt{C_h}}{\sum_{h=1}^k W_h S_h / \sqrt{C_h}}$$

$$n_h = \frac{n W_h S_h / \sqrt{C_h}}{\sum_{h=1}^k W_h S_h / \sqrt{C_h}}$$

Variance under Optimum Allocation

Minimum Variance for Fixed Cost

$$Var(\bar{y}_{st}) = \sum_{h=1}^k \frac{W_h^2 S_h^2}{n_h} - \sum_{h=1}^k \frac{W_h S_h^2}{N}$$

$$n_h = \frac{n W_h S_h / \sqrt{C_h}}{\sum_{h=1}^k W_h S_h / \sqrt{C_h}}$$

$$Var_{\min}(\bar{y}_{st}) = \frac{1}{n} \left(\sum_{h=1}^k W_h S_h / \sqrt{c_h} \right) \left(\sum_{h=1}^k W_h S_h \sqrt{c_h} \right)$$

Neyman Allocation

$$n_h = \frac{n W_h S_h / \sqrt{C_h}}{\sum_{h=1}^k W_h S_h / \sqrt{C_h}}$$

If $C_1 = C_2 = \dots = C$ then the cost function will be $C = C_0 + nC$

$$n_h = \frac{n W_h S_h / \sqrt{C}}{\sum_{h=1}^k W_h S_h / \sqrt{C}}$$

$$n_h = \frac{n W_h S_h}{\sum_{h=1}^k W_h S_h}$$

$$Var_{\min}(\bar{y}_{st}) = \frac{1}{n} \left(\sum_{h=1}^k W_h S_h \right)^2 - \frac{1}{N} \sum_{h=1}^k W_h S_h^2$$

$$Var_{\min}(\bar{y}_{st}) = \frac{1}{n} \left(\sum_{h=1}^k W_h S_h / \sqrt{c_h} \right) \left(\sum_{h=1}^k W_h S_h \sqrt{c_h} \right) - \frac{1}{N} \sum_{h=1}^k W_h S_h^2$$

For large N

$$Var_{\min}(\bar{y}_{st}) \simeq \frac{1}{n} \left(\sum_{h=1}^k W_h S_h \right)^2$$

Example:

Consider a population of size 700 consisting on three strata such that $N_1=100$, $N_2=250$ and $N_3=350$. The required sample size is 18. The sample size from stratum-1, stratum-2 and stratum-3 is arbitrarily decided as 4, 8 and 6, respectively.

The sample from each stratum is chosen as

Stra-1	Stra-2	Stra-3
1	7	23
3	12	14
2	8	20

5	5	22
	11	24
	10	17
	9	
	12	

	Stra-1	Stra-2	Stra-3
	1	7	23
	3	12	14
	2	8	20
	5	5	22
		11	24
		10	17
		9	
		12	
mean	2.75	9.25	20
N_h	100	250	350
S_h	1.708	2.493	3.847
n_h	4	8	6

	Stra-1	Stra-2	Stra-3
mean	2.75	9.25	20
N_h	100	250	350
S_h	1.7078	2.4928	3.8471
n_h	4	8	6

$$\bar{y}_{st} = \sum_{h=1}^k W_h \bar{y}_h = \sum_{h=1}^k N_h \bar{y}_h / N$$

$$\bar{y}_{st} = \frac{1}{N} [N_1 \bar{y}_1 + N_2 \bar{y}_2 + N_3 \bar{y}_3]$$

$$\bar{y}_{st} = 13.70$$

$$W_1^2 S_1^2 \frac{(N_1 - n_1)}{N_1 n_1} + W_2^2 S_2^2 \frac{(N_2 - n_2)}{N_2 n_2} + W_3^2 S_3^2 \frac{(N_3 - n_3)}{N_3 n_3}$$

$$Var(\bar{y}_{st}) = \sum_{h=1}^3 W_h^2 S_h^2 \frac{(N_h - n_h)}{N_h n_h}$$

	Stra-1	Stra-2	Stra-3
W_h	0.1429	0.3571	0.5
N_h	100	250	350
S_h²	2.9168	6.2143	14.800
n_h	4	8	6

$$Var(\bar{y}_{st}) = 0.7162$$

Confidence Interval for Mean

The interval estimation of population mean

$$\bar{y}_{st} \pm z_{1-\alpha/2} S.E(\bar{y}_{st})$$

$$13.7 \pm 1.96(0.8463)$$

Lower Limit

$$13.7 - 1.96(0.8463) = 11.41$$

Upper Limit

$$13.7 + 1.96(0.8463) = 14.73$$

Example:

Ans: Consider a population of size 700 consisting on three strata such that $N_1=100$, $N_2=250$ and $N_3=350$. The required sample size is 18.

The sample size from stratum-1, stratum-2 and stratum-3 is

$$n_h = \frac{n}{L} = \frac{18}{3} = 6$$

The sample from each stratum is chosen as

Stra-1	Stra-2	Stra-3
1	7	23
3	12	14
2	8	20
5	5	22
4	11	24
3	10	17

	Stra-1	Stra-2	Stra-3
	1	7	23
	3	12	14
	2	8	20
	5	5	22
	4	11	24
	3	10	17
mean	3	8.83	20
N_h	100	250	350
S_h	1.4142	2.6394	3.8471
n_h	6	6	6

	Stra-1	Stra-2	Stra-3
mean	3	8.83	20
N_h	100	250	350
S_h	1.4142	2.6394	3.8471
n_h	6	6	6

$$\bar{y}_{st} = \sum_{h=1}^k W_h \bar{y}_h = \sum_{h=1}^k N_h \bar{y}_h / N$$

$$\bar{y}_{st} = \frac{1}{N} [N_1 \bar{y}_1 + N_2 \bar{y}_2 + N_3 \bar{y}_3]$$

$$\bar{y}_{st} = 13.58$$

$$Var(\bar{y}_{st}) = \sum_{h=1}^3 W_h^2 S_h^2 \frac{(N_h - n_h)}{N_h n_h}$$

$$W_1^2 S_1^2 \frac{(N_1 - n_1)}{N_1 n_1} + W_2^2 S_2^2 \frac{(N_2 - n_2)}{N_2 n_2} + W_3^2 S_3^2 \frac{(N_3 - n_3)}{N_3 n_3}$$

	Stra-1	Stra-2	Stra-3
W_h	0.1429	0.3571	0.5
N_h	100	250	350
S_h²	2	6.967	14.800
n_h	6	6	6

$$\text{Var}(\bar{y}_{st}) = 0.7163$$

Confidence Interval for Mean

The interval estimation of population mean

$$\bar{y}_{st} \pm z_{1-\alpha/2} S.E(\bar{y}_{st})$$

$$13.58 \pm 1.96(0.8701)$$

Lower Limit

$$13.58 - 1.96(0.8701) = 11.36$$

Upper Limit

$$13.58 + 1.96(0.8701) = 14.78$$

Example:

Ans: Consider a population of size 700 consisting on three strata such that $N_1=100$, $N_2=250$ and $N_3=350$. The required sample size is 18. First we will allocate the sample size to each stratum according to proportional allocation.

Sample Size Allocation

$$N_1=100, N_2=250, N_3=350, N=700, n=18$$

$$n_h = \frac{n}{N} N_h$$

$$n_1 = \frac{n}{N} N_1 = \frac{18}{700} 100 = 3 \quad n_2 = \frac{n}{N} N_2 = \frac{18}{700} 250 = 6$$

$$n_3 = \frac{n}{N} N_3 = \frac{18}{700} 350 = 9$$

The sample from each stratum is chosen as

Stra-1	Stra-2	Stra-3
1	5	23
2	11	14
5	10	20
	9	22
	12	24
	6	17
		23
		21

		19
--	--	----

	Stra-1	Stra-2	Stra-3
	1	5	23
	2	11	14
	5	10	20
		9	22
		12	24
		6	17
			23
			21
			19
mean	2.67	8.83	19.86
N_h	100	250	350
n_h	3	6	9

	Stra-1	Stra-2	Stra-3
mean	2.67	8.83	19.86
N_h	100	250	350
n_h	3	6	9

$$\bar{y}_{st} = \sum_{h=1}^k W_h \bar{y}_h = \sum_{h=1}^k N_h \bar{y}_h / N$$

$$\bar{y}_{st} = \frac{1}{N} [N_1 \bar{y}_1 + N_2 \bar{y}_2 + N_3 \bar{y}_3]$$

$$\bar{y}_{st} = 13.46$$

	Stra-1	Stra-2	Stra-3
N_h	100	250	350
n_h	3	6	9
S_h²	4.333	7.77	12.476

$$Var(\bar{y}_{st}) = \sum_{h=1}^3 W_h^2 S_h^2 \frac{(N_h - n_h)}{N_h n_h}$$

$$W_1^2 S_1^2 \frac{(N_1 - n_1)}{N_1 n_1} + W_2^2 S_2^2 \frac{(N_2 - n_2)}{N_2 n_2} + W_3^2 S_3^2 \frac{(N_3 - n_3)}{N_3 n_3}$$

$$\text{Var}(\bar{y}_{st}) = 0.527$$

Confidence Interval for Mean

The interval estimation of population mean

$$\bar{y}_{st} \pm z_{1-\alpha/2} S.E(\bar{y}_{st})$$

$$13.46 \pm 1.96(0.726)$$

Lower Limit

$$13.46 - 1.96(0.726) = 12.47$$

Upper Limit

$$13.46 + 1.96(0.726) = 15.32$$

Q . Give an example to explain how Sample Size is obtained in Optimum Allocation?

Ans: A manufacturing company is interested to conduct a survey about a certain product from three towns (say A, B, and C) of a city. The towns are different from each other with respect to the household income. The number of houses in Town A, B, and C are 170, 135, and 80, respectively.

The company finds that obtaining an observation cost from town A or B is same as Rs.500 (i.e. $c_1 = c_2 = 500$). The cost per observation in the town C is Rs. 800. (i.e. $c_3 = 800$).

$$S_1 \approx 3, S_2 \approx 7, S_3 \approx 10$$

The overall sample size with certain margin or error is 30. Find the sample size from each Town (stratum)

$$n_1, n_2, n_3 = ?$$

Sample Size in Optimum

	Town-A	Town-B	Town-C
S_h	3	7	10
N_h	170	135	80
C_h	500	500	800

$$n_h = n \left(\frac{N_h S_h / \sqrt{c_h}}{\sum_{h=1}^H N_h S_h / \sqrt{c_h}} \right)$$

	Stra-1	Stra-2	Stra-3
S_h	3	7	10
N_h	170	135	80
C_h	500	500	800
N_hS_h	510	945	800
N_hS_h/sqrt(C_h)	22.8078	42.2616	28.2843

$$\sum_{h=1}^3 N_h S_h / \sqrt{c_h} = 93.35$$

$$n_h = n \left(\frac{N_h S_h / \sqrt{c_h}}{\sum_{h=1}^3 N_h S_h / \sqrt{c_h}} \right)$$

$$n_1 = n \left(\frac{N_1 S_1 / \sqrt{c_1}}{\sum_{h=1}^3 N_h S_h / \sqrt{c_h}} \right)$$

$$n_1 = 30 \left(\frac{22.8078}{93.35} \right) = 7.33 \approx 7$$

$$n_2 = n \left(\frac{N_2 S_2 / \sqrt{c_2}}{\sum_{h=1}^3 N_h S_h / \sqrt{c_h}} \right)$$

$$n_2 = 30 \left(\frac{42.2616}{93.35} \right) = 13.58 \approx 14$$

$$n_3 = n \left(\frac{N_3 S_3 / \sqrt{c_3}}{\sum_{h=1}^3 N_h S_h / \sqrt{c_h}} \right)$$

$$n_3 = 30 \left(\frac{28.2843}{93.35} \right) = 9.09 \approx 9$$

$$n_1 = 7, n_2 = 14, n_3 = 9$$

Q . By using an example explain the Variance for Optimum Allocation

Ans:

$$n_1 = 7, n_2 = 14, n_3 = 9$$

Stra-1	Stra-2	Stra-3
1	5, 7	23
2	11, 6	14

5	10, 8	20
4	9, 12	22
3	12, 11	24
6	8, 9	17
10	11, 13	23
		21 19

	Stra-1	Stra-2	Stra-3
Mean	4.4286	9.4286	20.333
N_h	170	135	80
n_h	7	14	9

$$\bar{y}_{st} = \sum_{h=1}^k W_h \bar{y}_h = \sum_{h=1}^k N_h \bar{y}_h / N \quad \bar{y}_{st} = \frac{1}{N} [N_1 \bar{y}_1 + N_2 \bar{y}_2 + N_3 \bar{y}_3]$$

$$\bar{y}_{st} = 9.47$$

	Stra-1	Stra-2	Stra-3
Mean	4.4286	9.4286	20.3333
N_h	170	135	80
n_h	7	14	9
S_h²	8.9523	5.8022	10.5

$$Var(\bar{y}_{st}) = \sum_{h=1}^3 W_h^2 S_h^2 \frac{(N_h - n_h)}{N_h n_h}$$

$$W_1^2 S_1^2 \frac{(N_1 - n_1)}{N_1 n_1} + W_2^2 S_2^2 \frac{(N_2 - n_2)}{N_2 n_2} + W_3^2 S_3^2 \frac{(N_3 - n_3)}{N_3 n_3}$$

$$Var(\bar{y}_{st}) = 0.329$$

Confidence Interval

The interval estimation of population mean

$$\bar{y}_{st} \pm z_{1-\alpha/2} S.E(\bar{y}_{st})$$

$$9.49 \pm 1.96(0.57399)$$

Lower Limit

$$9.49 - 1.96(0.57399) = 8.36$$

Upper Limit

$$9.49 + 1.96(0.57399) = 10.61$$

Q . Give an example to explain how Sample Size is obtained in Neyman Allocation?

Ans: A manufacturing company is interested to conduct a survey about a certain product from three towns (say A, B, and C) of a city. The towns are different from each other with respect to the household income. The number of houses in Town A, B, and C are 170, 135, and 80, respectively.

The company finds that obtaining an observation cost from town A or B or C is same.

$$S_1 \approx 3, S_2 \approx 7, S_3 \approx 10$$

The overall sample size with certain margin of error is 30. Find the sample size from each Town (stratum)

$$n_1, n_2, n_3 = ?$$

	Town-A	Town-B	Town-C
S_h	3	7	10
N_h	170	135	80
C_h	500	500	500

$$n_h = n \left(\frac{N_h S_h}{\sum_{h=1}^3 N_h S_h} \right)$$

	Stra-1	Stra-2	Stra-3
S_h	3	7	10
N_h	170	135	80
C_h	500	500	800
$N_h S_h$	510	945	800

$$\sum_{h=1}^3 N_h S_h = 2255$$

$$n_h = n \left(\frac{N_h S_h}{\sum_{h=1}^3 N_h S_h} \right)$$

$$n_1 = n \left(\frac{N_1 S_1}{\sum_{h=1}^3 N_h S_h} \right)$$

$$n_1 = 30 \left(\frac{510}{2255} \right) = 6.78 \approx 7$$

$$n_3 = n \left(\frac{N_3 S_3}{\sum_{h=1}^3 N_h S_h} \right)$$

$$n_2 = 30 \left(\frac{945}{2255} \right) = 12.57 \approx 12$$

$$n_2 = n \left(\frac{N_2 S_2}{\sum_{h=1}^3 N_h S_h} \right) \quad n_3 = 30 \left(\frac{800}{2255} \right) = 10.64 \approx 11$$

$$n_1 = 7, n_2 = 12, n_3 = 11$$

Q . Derive the Variance for Neyman Allocation by using an example?

Ans:

$$n_1 = 7, n_2 = 12, n_3 = 11$$

Stra-1	Stra-2	Stra-3
1	5, 7	23
2	11, 6	14
5	10, 8	20
4	9, 12	22
3	12, 11	24
6	8, 9	17
10		23
		21, 19 20, 18

	Stra-1	Stra-2	Stra-3
Mean	4.4286	9.0000	20.0909
N_h	170	135	80
n_h	7	12	11

$$\bar{y}_{st} = \sum_{h=1}^k W_h \bar{y}_h = \sum_{h=1}^k N_h \bar{y}_h / N \quad \bar{y}_{st} = \frac{1}{N} [N_1 \bar{y}_1 + N_2 \bar{y}_2 + N_3 \bar{y}_3]$$

$$\bar{y}_{st} = 9.29$$

	Stra-1	Stra-2	Stra-3
Mean	4.4286	9.0000	20.0909
N_h	170	135	80
n_h	7	12	11
S_h²	8.9524	5.2727	8.8909

$$Var(\bar{y}_{st}) = \sum_{h=1}^3 W_h^2 S_h^2 \frac{(N_h - n_h)}{N_h n_h}$$

$$W_1^2 S_1^2 \frac{(N_1 - n_1)}{N_1 n_1} + W_2^2 S_2^2 \frac{(N_2 - n_2)}{N_2 n_2} + W_3^2 S_3^2 \frac{(N_3 - n_3)}{N_3 n_3}$$

$$Var(\bar{y}_{st}) = 0.3184$$

Confidence Interval

The interval estimation of population mean

$$\bar{y}_{st} \pm z_{1-\alpha/2} S.E(\bar{y}_{st})$$

$$9.29 \pm 1.96(0.5643)$$

Lower Limit

$$9.29 - 1.96(0.5643) = 8.18$$

Upper Limit

$$9.29 + 1.96(0.5643) = 10.39$$

Q . Give Comparison of Allocation Methods?

Ans: Variance (Rand) & Variance (Prop)

$$Var_{ran}(\bar{y}) = \frac{N-n}{Nn} S^2$$

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

$$S^2 = \frac{\sum_{h=1}^k \sum_{i=1}^{N_h} (Y_{hi} - \bar{Y})^2}{N-1}$$

$$Var_{ran}(\bar{y}) = \frac{N-n}{Nn} \frac{\sum_{h=1}^k \sum_{i=1}^{N_h} (Y_{hi} - \bar{Y})^2}{N-1}$$

$$\sum_{h=1}^k \sum_{i=1}^{N_h} (Y_{hi} - \bar{Y})^2 = \sum_{h=1}^k \sum_{i=1}^{N_h} [(Y_{hi} - \bar{Y}_h) + (\bar{Y}_h - \bar{Y})]^2$$

$$\sum_{h=1}^k \sum_{i=1}^{N_h} [(Y_{hi} - \bar{Y}_h)^2 + (\bar{Y}_h - \bar{Y})^2 + (Y_{hi} - \bar{Y}_h)(\bar{Y}_h - \bar{Y})]$$

$$\sum_{h=1}^k \sum_{i=1}^{N_h} (Y_{hi} - \bar{Y}_h)^2 + \sum_{h=1}^k \sum_{i=1}^{N_h} (\bar{Y}_h - \bar{Y})^2$$

$$\text{as } \sum_{h=1}^k (\bar{Y}_h - \bar{Y}) = 0$$

$$(N-1)S^2 = \sum_{h=1}^k (N_h - 1)S_h^2 + \sum_{h=1}^k N_h (\bar{Y}_h - \bar{Y})^2$$

for the large N

$$NS^2 = \sum_{h=1}^k N_h S_h^2 + \sum_{h=1}^k N_h (\bar{Y}_h - \bar{Y})^2$$

$$S^2 = \sum_{h=1}^k \frac{N_h}{N} S_h^2 + \sum_{h=1}^k \frac{N_h}{N} (\bar{Y}_h - \bar{Y})^2$$

$$= \sum_{h=1}^k W_h S_h^2 + \sum_{h=1}^k W_h (\bar{Y}_h - \bar{Y})^2$$

$$\theta S^2 = \theta \sum_{h=1}^k W_h S_h^2 + \theta \sum_{h=1}^k W_h (\bar{Y}_h - \bar{Y})^2$$

$$V_{ran} = V_{prop} + \theta \sum_{h=1}^k W_h (\bar{Y}_h - \bar{Y})^2$$

$$V_{ran} \geq V_{prop}$$

Variance (Prop) & Variance (Opt)

$$Var_{prop}(\bar{y}_{st}) = \frac{N-n}{Nn} \sum_{h=1}^k W_h S_h^2$$

$$Var_{opt}(\bar{y}_{st}) = \frac{1}{n} \left(\sum_{h=1}^k W_h S_h^2 \right)^2 - \frac{1}{N} \sum_{h=1}^k W_h S_h^2$$

$$Var(\bar{y}_{prop}) - Var(\bar{y}_{opt})$$

$$= \frac{1}{n} \left[\sum_{h=1}^k W_h S_h^2 - \left(\sum_{h=1}^k W_h S_h \right)^2 \right]$$

$$Var(\bar{y}_{prop}) - Var(\bar{y}_{opt}) = \frac{1}{n} \left[\sum_{h=1}^k W_h S_h^2 - 2 \left(\sum_{h=1}^k W_h S_h \right)^2 + \left(\sum_{h=1}^k W_h S_h \right)^2 \right]$$

$$= \frac{1}{n} \left[\sum_{h=1}^k W_h S_h^2 - 2 \left(\sum_{h=1}^k W_h S_h \right) \left(\sum_{h=1}^k W_h S_h \right) + \sum_{h=1}^k W_h \left(\sum_{h=1}^k W_h S_h \right)^2 \right]$$

$$Var(\bar{y}_{prop}) - Var(\bar{y}_{opt}) = \frac{1}{n} \sum_{h=1}^k W_h \left(S_h - \sum_{h=1}^k W_h S_h \right)^2$$

$$Var(\bar{y}_{prop}) = Var(\bar{y}_{opt}) + \frac{1}{n} \sum_{h=1}^k W_h \left(S_h - \sum_{h=1}^k W_h S_h \right)^2$$

Comparison

$$V_{prop} \geq V_{opt}$$

$$V_{ran} \geq V_{prop}$$

$$V_{ran} \geq V_{prop} \geq V_{opt}$$

Example:

All the 80 farms in a population are stratified by farm size. The expenditure on the insecticides used during the last year by each farmer is presented in table

Large farmers		Medium farmers				Small farmers		
75	76	55	40	51	28	35	31	26
65	79	45	38	55	47	28	38	32
86	62	35	33	41	61	36	42	18
57	92	30	43	48	35	40	33	16
45	50	42	53	54	31	25	29	
69	48	38	37	36	23	18	25	
48	77	40	52	44		28	35	
60	60	36	39	47		32	26	
55	64	48	46	39		13	30	
66	58	46	42	41		19	37	

- Select a stratified sample of 24 farmers by using equal allocation
- Compute the variance of sample mean under simple random sampling without replacement.
- Compute the variance of sample mean under stratified sampling using equal allocation.
- Compare the variances

Ans:

Population Mean

It is given that $N=80$,

$n = 24$, $N_1 = 20$, $N_2 = 36$, and $N_3 = 24$.

$W_1 = 0.25$, $W_2 = 0.45$, and $W_3 = 0.30$.

$$\bar{Y} = \sum_{i=1}^N Y_i / N = \frac{(75 + 65 + \dots + 16)}{80} = 47.79$$

Variance Under SRSWOR

The overall standard deviation is

$$Var(\bar{y}) = \frac{N-n}{Nn} S^2$$

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

$$= 268.68$$

$$Var(\bar{y}) = \left(\frac{80-24}{80 \times 24} \right) 268.68$$

$$Var(\bar{y}) = 7.84$$

Variance under Stratified Sampling

	Stra-1	Stra-2	Stra-3
W_h	0.25	0.45	0.30
N_h	20	36	24
S_h^2	169.52	70.56	61.45
n_h	8	8	8

$$Var(\bar{y}_{st}) = \sum_{h=1}^3 W_h^2 S_h^2 \frac{(N_h - n_h)}{N_h n_h}$$

$$W_1^2 S_1^2 \frac{(N_1 - n_1)}{N_1 n_1} + W_2^2 S_2^2 \frac{(N_2 - n_2)}{N_2 n_2} + W_3^2 S_3^2 \frac{(N_3 - n_3)}{N_3 n_3}$$

Under SRSWOR $Var(\bar{y}_{st}) = 2.64$

$$Var(\bar{y}) = 7.84$$

Under Stratified Sampling for equal allocation

$$Var(\bar{y}_{st}) = 2.64$$

Q . Give an example to explain the Comparison between Proportional Allocation and Simple Random Sampling?

Ans : All the 80 farms in a population are stratified by farm size. The expenditure on the insecticides used during the last year by each farmer is presented in table below (Source: Elements of Survey Sampling by Singh and Mangat).

Large farmers		Medium farmers				Small farmers		
75	76	55	40	51	28	35	31	26
65	79	45	38	55	47	28	38	32
86	62	35	33	41	61	36	42	18
57	92	30	43	48	35	40	33	16
45	50	42	53	54	31	25	29	
69	48	38	37	36	23	18	25	
48	77	40	52	44		28	35	
60	60	36	39	47		32	26	
55	64	48	46	39		13	30	
66	58	46	42	41		19	37	

Population Mean

It is given that $N=80$,

$n = 24$, $N_1 = 20$, $N_2 = 36$, and $N_3 = 24$.

$W_1 = 0.25$, $W_2 = 0.45$, and $W_3 = 0.30$.

$$\bar{Y} = \sum_{i=1}^N Y_i / N = \frac{(75+65+.....+16)}{80} = 47.79$$

Variance under SRSWOR

The overall standard deviation is

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

$$= 268.68$$

$$Var(\bar{y}) = \frac{N-n}{Nn} S^2$$

$$Var(\bar{y}) = \left(\frac{80-24}{80 \times 24} \right) 268.68$$

$$Var(\bar{y}) = 7.84$$

Proportional Allocation

$N_1=20$, $N_2=36$, $N_3=24$, $N=80$, $n=24$

$$n_1 = \frac{n}{N} N_1 = \frac{24}{80} 20 = 6$$

$$n_2 = \frac{n}{N} N_2 = \frac{24}{80} 36 = 10.8$$

$$n_3 = \frac{n}{N} N_3 = \frac{24}{80} 24 = 7.2$$

Variance under Proportional

	Stra-1	Stra-2	Stra-3
W_h	0.25	0.45	0.30
N_h	20	36	24
S_h²	169.52	70.56	61.45
n_h	6	11	7

$$Var(\bar{y}_{st}) = \sum_{h=1}^3 W_h^2 S_h^2 \frac{(N_h - n_h)}{N_h n_h}$$

$$W_1^2 S_1^2 \frac{(N_1 - n_1)}{N_1 n_1} + W_2^2 S_2^2 \frac{(N_2 - n_2)}{N_2 n_2} + W_3^2 S_3^2 \frac{(N_3 - n_3)}{N_3 n_3}$$

$$1.2361 + .9021 + .5596$$

Comparison of Variances Under SRSWOR

$$Var(\bar{y}_{st}) = 2.698$$

$$Var(\bar{y}) = 7.84$$

Under Stratified Sampling for proportional allocation

$$Var(\bar{y}_{st}) = 2.698$$

$$RE = 290.5$$

Example:

Comparison between Neyman Allocation and Simple Random Sampling?

Ans: Population Mean

It is given that N=80,

n = 24, N1 = 20, N2 = 36, and N3 = 24.

W1 = 0.25, W2 = 0.45, and W3 = 0.30.

$$\bar{Y} = \sum_{i=1}^N Y_i / N = \frac{(75 + 65 + \dots + 16)}{80} = 47.79$$

Variance under SRSWOR

The overall standard deviation is

$$Var(\bar{y}) = \frac{N-n}{Nn} S^2$$

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

$$= 268.68$$

$$Var(\bar{y}) = \left(\frac{80-24}{80 \times 24} \right) 268.68$$

$$Var(\bar{y}) = 7.84$$

Variance under Neyman Allocation

	Stra-1	Stra-2	Stra-3
W_h	0.25	0.45	0.30
N_h	20	36	24
S_h²	169.52	70.56	61.45

$$n_1 = n \left(\frac{N_1 S_1}{\sum_{h=1}^3 N_h S_h} \right) = 8.3$$

$$n_2 = n \left(\frac{N_2 S_2}{\sum_{h=1}^3 N_h S_h} \right) = 9.7$$

$$n_3 = n \left(\frac{N_3 S_3}{\sum_{h=1}^3 N_h S_h} \right) = 6$$

	Stra-1	Stra-2	Stra-3
W_h	0.25	0.45	0.30
N_h	20	36	24
S_h²	169.52	70.56	61.45
n_h	8	10	6
Var($\bar{y}_{st}$)	$= \sum_{h=1}^3 W_h^2 S_h^2 \frac{(N_h - n_h)}{N_h n_h}$		

$$W_1^2 S_1^2 \frac{(N_1 - n_1)}{N_1 n_1} + W_2^2 S_2^2 \frac{(N_2 - n_2)}{N_2 n_2} + W_3^2 S_3^2 \frac{(N_3 - n_3)}{N_3 n_3}$$

$$.7946 + 1.0320 + .6913$$

$$Var(\bar{y}_{st}) = 2.5179$$

**Comparison of Variances
Under SRSWOR**

$$Var(\bar{y}) = 7.84$$

Under Stratified Sampling for Neyman allocation

$$Var(\bar{y}_{st}) = 2.5179$$

$$RE = 311.2$$

Stratified Sampling with R

Q. Define the following data in R and perform stratified sampling. Calculate the parameters for each stratum. Also find the mean of population.

Large farmers		Medium farmers				Small farmers		
75	76	55	40	51	28	35	31	26
65	79	45	38	55	47	28	38	32
86	62	35	33	41	61	36	42	18
57	92	30	43	48	35	40	33	16
45	50	42	53	54	31	25	29	
69	48	38	37	36	23	18	25	
48	77	40	52	44		28	35	
60	60	36	39	47		32	26	
55	64	48	46	39		13	30	
66	58	46	42	41		19	37	

Ans:

Defining Strata in R

- `str1<-c(75,76,65,79,86,62,57,92,45,50,69,48,48,77,60,60,55,64,66,58)`
- `str2<-c(55,40,51,45,38,55,35,33,41,30,43,48,42,53,54,38,37,36,40,52,44,36,39,47,48,46,39,46,42,41,28,47,61,35,31,23)`
- `str3<-c(35,31,26,28,38,32,36,42,18,40,33,16,25,29,18,25,28,35,32,26,13,30,19,37)`

Parameters

Mean and Standard Deviation

```
y<-c(str1,str2,str3)
m_p=mean(y)
```

Output

sd_p=sd(y)

m_p43.7875

sd_p16.39

Defining Terms

Defining Stratum Size

$N1=length(str1)$

$N2=length(str2)$

$N3=length(str3)$

$N=N1+N2+N3$

Defining Weights

$W1=N1/N; W2=N2/N;$

$W3=N3/N$

$n=24$

Mean & Standard Deviation

Mean of Strata

$m_st1=mean(str1)$

$m_st2=mean(str2)$

$m_st3=mean(str3)$

Standard Deviation of Strata

$sd_st1=sd(str1)$

$sd_st2=sd(str2)$

$sd_st3=sd(str3)$

Output

64.6,42.19,28.83

13.01,8.4,7.84

Variance & Stratified Mean

Variances of Strata

$$\text{var_st1}=\text{var}(\text{str1})$$

$$\text{var_st2}=\text{var}(\text{str2})$$

$$\text{var_st3}=\text{var}(\text{str3})$$

Mean of population using stratified population

$$m_{yst}=(1/N)*$$

$$(N1*m_{st1} + N2*m_{st2} + N3*m_{st3})$$

Q . Define the following data in R and perform stratified sampling using the proportional allocation using the sample size 24. Also find the mean and variance.

Large farmers		Medium farmers				Small farmers		
75	76	55	40	51	28	35	31	26
65	79	45	38	55	47	28	38	32
86	62	35	33	41	61	36	42	18
57	92	30	43	48	35	40	33	16
45	50	42	53	54	31	25	29	
69	48	38	37	36	23	18	25	
48	77	40	52	44		28	35	
60	60	36	39	47		32	26	
55	64	48	46	39		13	30	
66	58	46	42	41		19	37	

Ans:

Sample Size Allocation

$$n1=\text{round}(n*(N1/N))$$

$$n2=\text{round}(n*(N2/N))$$

$$n3=\text{round}(n*(N3/N))$$

$$\text{Term1}=(W1^2)*\text{var_st1}*(N1-n1)/(N1*n1)$$

$$\text{Term2}=(W2^2)*\text{var_st2}*(N2-n2)/(N2*n2)$$

Output

$$24=6+11+7$$

0.25,0.45,0.3

Variance under Proportional Allocation

Term3=(W3^2)*var_st3*

(N3-n3)/(N3*n3)

vp_prop=Term1+Term2+Term3

Output

2.698

Sampling from Stratum

s1=sample(str1,n1)

s2=sample(str2,n2)

s3=sample(str3,n3)

ms_st1=mean(s1)

ms_st2=mean(s2)

ms_st3=mean(s3)

m_prop=?

Output

73.17,42.55,27.14

45.58

Estimated Variance

vars_st1=var(s1);vars_st2=var(s2);vars_st3=var(s3)

Term1= (W1^2)*vars_st1*(N1-n1)/(N1*n1)

Term2=(W2^2)*vars_st2*(N2-n2)/(N2*n2)

Term3=(W3^2)*vars_st3*(N3-n3)/(N3*n3)

vs_prop =Term1+Term2+Term3

output

> vs_prop

[1] 2.57464

Q . Define the following data in R and perform stratified sampling using the Neyman allocation using the sample size 24. Also find the mean and variance.

Large farmers		Medium farmers				Small farmers		
75	76	55	40	51	28	35	31	26
65	79	45	38	55	47	28	38	32
86	62	35	33	41	61	36	42	18
57	92	30	43	48	35	40	33	16
45	50	42	53	54	31	25	29	
69	48	38	37	36	23	18	25	
48	77	40	52	44		28	35	

60	60	36	39	47	32	26
55	64	48	46	39	13	30
66	58	46	42	41	19	37

Ans:

Defining Strata in R

```
str1<-c(75,76,65,79,86,62,57,92,45,50,69,48,48,77,60,60,55,64,66,58)
str2<-c(55,40,51,45,38,55,35,33,41,30,43,48,42,53,54,38,37,36,40,52,44,36,39,47,48,46,39,46,42
,41,28,47,61,35,31,23)
str3<-c(35,31,26,28,38,32,36,42,18,40,33,16,25,29,18,25,28,35,32,26,13,30,19,37)
```

Parameters

Mean and Standard Deviation

```
y<-c(str1,str2,str3)
m_p=mean(y)
sd_p=sd(y)
m_p 43.7875
sd_p 16.39
```

Defining Terms

Defining Stratum Size

```
N1=length(str1)
N2=length(str2)
N3=length(str3)
N=N1+N2+N3
```

Defining Weights

```
W1=N1/N; W2=N2/N; W3=N3/N
n=24
```

Output

```
80=20+36+24
0.25,0.45,0.3
```

Sample Size Allocation

```
nn1=round(n*(N1*sd_st1/sum))
nn2=round(n*(N2*sd_st2/sum))
nn3=round(n*(N3*sd_st3/sum))
Term1=(W1^2)*var_st1*
(N1-n1)/(N1*n1)
Term2=(W2^2)*var_st2*
(N2-n2)/(N2*n2)
```

Output

```
24=8+10+6
0.79,1.03,0.69
```

Variance Under Neyman Allocation

```
Term3=(W3^2)*var_st3*
(N3-n3)/(N3*n3)
```

vp_Ney=Term1+Term2+Term3

Output

2.52

Sampling From Strata

```
s1=sample(str1,n1)
s2=sample(str2,n2)
s3=sample(str3,n3)
ms_st1=mean(s1)
ms_st2=mean(s2)
ms_st3=mean(s3)
m_Ney=?
```

Output

68.13, 40.2,26.83
43.17

Estimated Variance

```
vars_st1=var(s1);vars_st2=var(s2);vars_st3=var(s3)
Term1= (W1^2)*vars_st1*(N1-nn1)/(N1*nn1)
Term2= (W2^2)*vars_st2*(N2-nn2)/(N2*nn2)
Term3=(W3^2)*vars_st3*(N3-nn3)/(N3*nn3)
```

Estimated Variance

vs_Ney=Term1+Term2 +Term3

Output

> vs_Ney

Q. Define the following data in R and perform stratified sampling. Calculate the parameters for each stratum. Also find the mean of population.

Stratum	
1	12, 14, 19, 22
2	362, 441, 456, 482, 444, 472,
3	124, 189, 142, 165, 135, 140

N=16;N₁=4;N₂=6;N₃=6; n₁=2;n₂=3;n₃=3;n=8;

Y<-c(12,14,19,22,362,441,456,482,444,472,124,189,142,165,135,140)

Y1<-c(12,14,19,22)

Y2<-c(362,441,456,482,444,472)

Y3<-c(124,189,142,165,135,140)

N=16;N₁=4;N₂=6;N₃=6; n₁=2;n₂=3;n₃=3;n=8;

```
w1=N1/N; w2=N2/N;w3=N3/N;
y1=c();y2=c();y3=c();yst=c(); for(i in 1:10000){
sa1=sample(Y1,n1)
sa2=sample(Y2,n2)
sa3=sample(Y3,n3)
y1[i]=mean(sa1);
y2[i]=mean(sa2);
y3[i]=mean(sa3);
yst[i]=w1*y1[i]+w2*y2[i]+w3*y3[i];
}
mean(yst); var(yst)
```

Output

226.138

56.25539

mean(y)

$$vp=(w1^2)*var(Y1)*(N1-n1)/(N1*n1) + (w2^2)*var(Y2)*(N2-n2)/(N2*n2) +$$
$$(w3^2)*var(Y3)*(N3-n3)/(N3*n3)$$

Output

226.1875

Vp=56.12526

Q . Generate the stratified population consisting on three strata such that stratum-1 is normally distributed with mean 10 and standard deviation 2 with 100 values, stratum-2 is normally distributed with mean 100 and standard deviation 2 with 500 values, stratum-3 is normally distributed with mean 500 and standard deviation 2 with 1000 values. Find the mean and variance for this population using the method of stratified sampling.

Ans: Simulation Study

$N1=100;N2=500;N3=1000;n=50$

$Y1<-rnorm(N1,mean=10,sd=2)$

$Y2<-rnorm(N2,mean=100,sd=2)$

$Y3<-rnorm(N3,mean=500,sd=2)$

$Y<-c(Y1,Y2,Y3)$

```

y1=c();y2=c();y3=c();yst=c();
N=N1+N2+N3;
w1=N1/N;w2=N2/N;w3=N3/N;
n1=round(n*w1)
n2=round(n*w2)
n3=round(n*w3)

```

Looping

```

for(i in 1:10000){
sa1=sample(Y1,n1)
sa2=sample(Y2,n2)
sa3=sample(Y3,n3)
y1[i]=mean(sa1);
y2[i]=mean(sa2);
y3[i]=mean(sa3);
yst[i]=w1*y1[i]+w2*y2[i]+w3*y3[i];
}
mean(yst); var(yst)

```

Output

```

Mean(yst)=344.36
Var(yst) = 0.08

```

Variance of Mean Using R

```

mean(y)
vp=(w1^2)*var(Y1)*(N1-n1)/(N1*n1) + (w2^2)*var(Y2)*(N2-n2)/(N2*n2) +
(w3^2)*var(Y3)*(N3-n3)/(N3*n3)

```

Q Generate a population of size 1000 consisting on three strata such that

- 200 values for stratum-1 from normal distribution with mean=2 and standard deviation=3.
- 300 values for stratum-2 from normal distribution with mean=10 and standard deviation=9.

- 500 values for stratum-3 from normal distribution with mean=30 and standard deviation=5.
- Allocate the sample size to each stratum by Neyman Allocation where n=50.
- Select the sample from each stratum and estimate the mean of population.

Ans:

Defining Population

```
N1=200;N2=300;N3=500 ; n=50
```

```
Y1<-rnorm(N1,mean=2,sd=3)
```

```
Y2<-rnorm(N2,mean=10,sd=9)
```

```
Y3<-rnorm(N3,mean=30,sd=5)
```

```
Y<-c(Y1,Y2,Y3)
```

Sample Size Under Neyman Allocation

```
N=N1+N2+N3;
```

```
w1=N1/N;w2=N2/N;w3=N3/N
```

```
sum=w1*3+w2*9+w3*5
```

```
n1=round(n*w1*3/sum)
```

```
n2=round(n*w2*9/sum)
```

```
n3=round(n*w3*5/sum)
```

Looping

```
y1=c();y2=c();y3=c();yst=c();
```

```
for(i in 1:10000){
```

```
sa1=sample(Y1,n1)
```

```
sa2=sample(Y2,n2)
```

```
sa3=sample(Y3,n3)
```

```
y1[i]=mean(sa1);
```

```
y2[i]=mean(sa2);
```

```
y3[i]=mean(sa3);
```

```
yst[i]=w1*y1[i]+w2*y2[i]+w3*y3[i];
```

```
} mean(yst); var(yst)
```

Output

```
Mean(yst)=344.36
```

$$\text{Var}(\text{yst}) = 0.08$$

Q. Find the mean and variance of Proportion in Stratified Sampling

Ans: Suppose we have N population units i.e. $Y_1, Y_2, \dots, Y_i, \dots, Y_N$

$y_i = 1$ if i th unit possesses a certain attribute and 0 otherwise.

The population proportion is defined

$$\bar{Y} = \sum_{i=1}^N Y_i / N = A / N = P$$

The sample proportion is

$$\bar{y} = \sum_{i=1}^n y_i / n = a / n = p$$

Since Y_i takes the values as 1 and 0.

$$\sum_{i=1}^N Y_i^2 = A = NP$$

The same is the case for sample.

$$\sum_{i=1}^n y_i^2 = a = np$$

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

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

$$S^2 = \frac{N}{N-1} P(1-P)$$

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

Similarly $s^2 = npq / (n-1)$

Unbiased Variance Estimator

For SRSWOR

$$\text{VAR}(\bar{y}_{\text{wor}}) = \frac{N-n}{Nn} S^2,$$

$$\text{VAR}(p_{\text{wor}}) = \frac{N-n}{Nn} \frac{NPQ}{N-1},$$

$$VAR(p_{wor}) = \frac{N-n}{n} \frac{PQ}{N-1},$$

Proportion Estimation

$$p_{st} = \frac{1}{N} \sum_{h=1}^k p_h N_h$$

For single stratum

$$A_h = N_h P_h,$$

$$S_h^2 = N_h P_h Q_h / (N_h - 1),$$

$$Var(p_h) = \frac{N_h - n_h}{(N_h - 1)n_h} P_h Q_h,$$

$$p_{st} = \frac{1}{N} \sum_{h=1}^k p_h N_h$$

$$Var(p_{st}) = \frac{1}{N^2} \sum_{h=1}^K N_h^2 Var(p_h)$$

$$Var(p_h) = \frac{N_h - n_h}{N_h - 1} \frac{P_h Q_h}{n_h}$$

$$Var(p_{st}) = \frac{1}{N^2} \sum_{h=1}^k \frac{N_h^2 (N_h - n_h)}{N_h - 1} \frac{P_h Q_h}{n_h}$$

$$Var(p_{st}) = \sum_{h=1}^k W_h^2 \frac{(N_h - n_h)}{N_h - 1} \frac{P_h Q_h}{n_h}$$

If $N_h - 1 \sim N_h$ and n_h/N_h is ignored

$$Var(p_{st}) = \sum_{h=1}^k W_h^2 P_h Q_h / n_h$$

Estimator of Variance

$$var(p_{st}) = \sum_{h=1}^k W_h^2 \frac{(N_h - n_h)}{N_h} \frac{p_h q_h}{n_h - 1}$$

Variance in Case Of Proportional Allocation

$$Var_{prop}(\bar{y}_{st}) = \frac{N-n}{Nn} \sum_{h=1}^k W_h S_h^2$$

$$S_h^2 = N_h P_h Q_h / (N_h - 1)$$

$$Var_{prop}(p_{st}) = \frac{N-n}{Nn} \sum_{h=1}^k \frac{W_h N_h P_h Q_h}{(N_h - 1)}$$

Variance in Case of Neyman Allocation

$$Var_{opt}(\bar{y}_{st}) = \frac{1}{n} \left(\sum_{h=1}^k W_h S_h \right)^2 - \frac{1}{N} \sum_{h=1}^k W_h S_h^2$$

$$S_h^2 = N_h P_h Q_h / (N_h - 1)$$

$$Var_{opt}(p_{st}) = \frac{1}{n} \left(\sum_{h=1}^k W_h \sqrt{N_h P_h Q_h / (N_h - 1)} \right)^2 - \frac{1}{N} \sum_{h=1}^k W_h S_h^2$$

For large stratum size i.e. $N_h - 1 \sim N_h$

$$Var_{opt}(p_{st}) = \frac{1}{n} \left(\sum_{h=1}^k W_h \sqrt{P_h Q_h} \right)^2 - \frac{1}{N} \sum_{h=1}^k W_h S_h^2$$

Example:

The management of a local newspaper is to decide whether it should continue with the publication of 'Children Column', which had been introduced on experimental basis. For this purpose, it is imperative to estimate the proportion of readers who would favor its continuance.

The frame consists of readers who had stayed with the paper for the last six months. Since different attitudes are expected from the urban and rural readers, the population is stratified into urban readers and rural readers. In the population, there are 73000 urban readers and 30280 rural readers.

Sample Size Allocation

$$N_1=73000, N_2= 30280, N=103280, n=1016$$

$$n_1 = \frac{1016}{103280} 73000 \approx 718$$

$$n_2 = \frac{1016}{103280} 30280 \approx 298$$

The investigator selected WOR simple random samples of 718 respondents from stratum I (urban readers) and 298 readers from stratum II (rural readers). The number of individuals who favor continuation of the column was 570 from stratum I and 143 from stratum II.

Sample Size Allocation

$$x_1 = 570, x_2 = 143, n_1 = 718, n_2 = 298$$

$$p_1 = \frac{x_1}{n_1}, p_2 = \frac{x_2}{n_2}$$

Out put

$$P_1 = 0.7939$$

$$P_2 = 0.4799$$

Proportion Estimation

$$N_1 = 73000, N_2 = 30280, N = 103280, n = 1016$$

$$p_{st} = \frac{1}{N} \sum_{h=1}^k p_h N_h$$

$$p_{st} = \frac{1}{N} [p_1 N_1 + p_2 N_2]$$

Output

$$P_1 = 0.7939$$

$$P_2 = 0.4799$$

$$P_{st} = 0.7018$$

Example:

The management of a local newspaper is to decide whether it should continue with the publication of 'Children Column', which had been introduced on experimental basis. For this purpose, it is imperative to estimate the proportion of readers who would favor its continuance.

The frame consists of readers who had stayed with the paper for the last six months. Since different attitudes are expected from the urban and rural readers, the population is stratified into urban readers and rural readers. In the population, there are 73000 urban readers and 30280 rural readers.

The investigator selected WOR simple random samples of 718 respondents from stratum I (urban readers) and 298 readers from stratum II (rural readers). The number of individuals who favor continuation of the column was 570 from stratum I and 143 from stratum II.

Sample Size Allocation

$x_1 = 570, x_2 = 143, n_1 = 718, n_2 = 298$

$$p_1 = \frac{x_1}{n_1} = \frac{570}{718} = 0.7939$$

$$p_2 = \frac{x_2}{n_2} = \frac{143}{298} = 0.4799$$

Output

$P_1 = 0.7939$

$P_2 = 0.4799$

$N_1 = 73000, N_2 = 30280$

$N = 103280, n = 1016$

$$W_1 = \frac{N_1}{N} = \frac{73000}{103280} = 0.7068$$

$$W_2 = \frac{N_2}{N} = \frac{30280}{103280} = 0.2932$$

Proportion Estimation

$N_1 = 73000, N_2 = 30280, N = 103280, n = 1016$

$$p_{st} = \frac{1}{N} \sum_{h=1}^k p_h N_h$$

$$p_{st} = \frac{1}{N} [p_1 N_1 + p_2 N_2]$$

$$p_{st} = 0.7018$$

Estimated Variance

	Stra-1	Stra-2
N_h	73000	30280
n_h	718	298
p_h	0.7939	0.4799

$$var(p_{st}) = \sum_{h=1}^k W_h^2 \frac{(N_h - n_h)}{N_h} \frac{p_h q_h}{n_h - 1}$$

$$Var(p_{st}) = W_1^2 \frac{(N_1 - n_1)}{N_1} \frac{p_1 q_1}{n_1 - 1} + W_2^2 \frac{(N_2 - n_2)}{N_2} \frac{p_2 q_2}{n_2 - 1}$$

Output

$W_s = 0.7068, 0.2932$

$= .0001129 + .0000715 = .0001844$

Confidence Interval for Proportion

The 95% confidence limits will be

$$p_{st} \pm 1.96S.E(p_{st})$$

The Lower Limit

$$0.7018 - 1.96(0.0136) = 0.1772$$

The Upper Limit

$$0.7018 + 1.96(0.0136) = 0.6752$$

Systematic Sampling

Introduction:

First unit is selected randomly from first 'k' units and rest of the units are selected automatically. Systematic sampling has many types but we discuss the commonly used methods i.e. the linear and circular systematic sampling.

Linear Systematic Sampling

Group	Sample Composition
1	1, k+1, 2k+1, ..., (i-1)k+1, ... (n-1)k+1
2	2, k+2, 2k+2, ..., (i-1)k+2, ... (n-1)k+2
‘	‘ ‘ ‘ ‘
‘	‘ ‘ ‘ ‘
‘	‘ ‘ ‘ ‘
r	r, k+r, 2k+r, (i-1)k+r, ... (n-1)k+r
‘	‘ ‘ ‘ ‘
‘	‘ ‘ ‘ ‘
‘	‘ ‘ ‘ ‘
k	k, 2k, 3k, ..., ik, Nk

Every k^{th} unit is included in the sample.

The interval 'k' is calculated by dividing the population size to the sample size.

The sampling interval will be

$$k=N/n$$

The feature of operational convenience is prominent.

No risk to miss the large part of population.

Example

A certain company claims about their daily production in numbers as

125, 135, 157, 192, 151, 175, 164, 169, 147, 150, 138, 167, 155, 159, 139, 147, 149, 158.

We are interested to select the systematic sample of size 3

Here we have $N=18$ with $n=3$, so $k=6$

125, 135, 157, 192, 151, 175, 164, 169, 147, 150, 138, 167, 155, 159, 139, 147, 149, 158.

Random Start	Serial Number	Sampled values
1	1,7,13	125,164,155
2	2,8,14	135,169,159
3	3,9,15	157,147,139
4	4,10,16	192,150,147
5	5,11,17	151,138,149
6	6,12,18	175,167,158

$N=15$, $n=3$

125, 135,

157, 192, 151,

175,164,169, 147,

150,138, 167, 155, 159,139

Q . Prove that sample mean is unbiased estimator of population mean

Group	Sample Composition	Sample Mean
1	1, k+1, 2k+1,..., (i-1)k+1, ... (n-1)k+1	$\bar{y}_1$
2	2, k+2, 2k+2,..., (i-1)k+2,... (n-1)k+2	$\bar{y}_2$

⋮	⋮ ⋮ ⋮ ⋮	⋮
⋮	⋮ ⋮ ⋮ ⋮ ⋮	⋮
⋮	⋮ ⋮ ⋮ ⋮ ⋮	⋮
r	r, k+r, 2k+r, (i-1)k+r, ... (n-1)k+r	$\bar{y}_r$
⋮	⋮ ⋮ ⋮ ⋮ ⋮	⋮
⋮	⋮ ⋮ ⋮ ⋮ ⋮	⋮
⋮	⋮ ⋮ ⋮ ⋮ ⋮	⋮
k	k, 2k, 3k, ..., ik, ..., Nk	$\bar{y}_k$

Population Mean

The population mean is given by

$$\bar{Y} = \frac{1}{nk} \sum_{r=1}^k \sum_{i=1}^n y_{ri},$$

Expectation of Sample Mean

From the table, we can see that

$$\bar{y}_r = \frac{1}{n} \sum_{i=1}^n y_{ri}$$

$$E(\bar{y}_{sy}) = \frac{1}{k} \sum_{r=1}^k \bar{y}_r,$$

$$E(\bar{y}_{sy}) = \frac{1}{k} \sum_{r=1}^k \frac{1}{n} \sum_{i=1}^n y_{ri},$$

$$E(\bar{y}_{sy}) = \frac{1}{nk} \sum_{r=1}^k \sum_{i=1}^n y_{ri},$$

$$\bar{Y} = \frac{1}{nk} \sum_{r=1}^k \sum_{i=1}^n y_{ri},$$

$$E(\bar{y}_{sy}) = \bar{Y}$$

Example

A certain company claims about their daily production in numbers as 125, 135, 157, 192, 151, 175, 164, 169, 147, 150, 138, 167, 155, 159, 139, 147, 149, 158. We are interested to select the systematic sample of size 3. Here we have N=18 with n=3, so k=6
125, 135, 157, 192, 151, 175, 164, 169, 147, 150, 138, 167, 155, 159, 139, 147, 149, 158.

Random Start	Sampled values	Serial Number
1	125,164,155	1,7,13
2	135,169,159	2,8,14
3	157,147,139	3,9,15
4	192,150,147	4,10,16
5	151,138,149	5,11,17
6	175,167,158	6,12,18

Sampled values	Sampled Means
125,164,155	148
135,169,159	154.333
157,147,139	147.667
192,150,147	163
151,138,149	146
175,167,158	166.667

Mean=154.28

Q. Derive the Variance of Sample Mean Under Systematic Sampling?

Ans: Linear Systematic Sampling

Group	Sample Composition	Sample Mean
1	1, k+1, 2k+1,..., (i-1)k+1, ... (n-1)k+1	$\bar{y}_1$
2	2, k+2, 2k+2,..., (i-1)k+2,... (n-1)k+2	$\bar{y}_2$
⋮	⋮ ⋮ ⋮ ⋮ ⋮	⋮
⋮	⋮ ⋮ ⋮ ⋮ ⋮	⋮
⋮	⋮ ⋮ ⋮ ⋮ ⋮	⋮
r	r, k+r, 2k+r, (i-1)k+r, ... (n-1)k+r	$\bar{y}_r$

‘	‘ ‘ ‘ ‘ ‘	‘
‘	‘ ‘ ‘ ‘ ‘	‘
‘	‘ ‘ ‘ ‘ ‘	‘
k	k, 2k, 3k,...,ik.... Nk	$\bar{y}_k$

Variance of Sample Mean

Expressions of Variance

$$V(\bar{y}_{sy}) = \frac{1}{k} \sum_{r=1}^k (\bar{y}_r - \bar{Y})^2$$

$$Var(\bar{y}_{sy}) = \frac{N-1}{N} S^2 - \frac{k(n-1)}{N} S_w^2$$

$$S_w^2 = \frac{1}{k(n-1)} \sum_{r=1}^k \sum_{i=1}^n (y_{ri} - \bar{y}_r)^2$$

The variance of sample mean is given by

$$V(\bar{y}_{sy}) = \frac{1}{k} \sum_{r=1}^k (\bar{y}_r - \bar{Y})^2$$

$$S^2 = \frac{\sum_{r=1}^k \sum_{i=1}^n [y_{ri} - \bar{Y}]^2}{(nk-1)}$$

$$(nk-1)S^2 = \sum_{r=1}^k \sum_{i=1}^n [y_{ri} - \bar{Y}]^2$$

$$(nk-1)S^2 = \sum_{r=1}^k \sum_{i=1}^n [y_{ri} - \bar{Y}]^2$$

$$= \sum_{r=1}^k \sum_{i=1}^n [y_{ri} - \bar{y}_r + \bar{y}_r - \bar{Y}]^2$$

$$= \sum_{r=1}^k \sum_{i=1}^n (y_{ri} - \bar{y}_r)^2 + \sum_{r=1}^k \sum_{i=1}^n (\bar{y}_r - \bar{Y})^2$$

$$(nk-1)S^2 = \sum_{r=1}^k \sum_{i=1}^n (y_{ri} - \bar{y}_r)^2 + \sum_{r=1}^k \sum_{i=1}^n (\bar{y}_r - \bar{Y})^2$$

Group	Sample Composition	Sample Mean
1	1, k+1, 2k+1,..., (i-1)k+1, ... (n-1)k+1	$\bar{y}_1$

2	2, k+2, 2k+2,..., (i-1)k+2,... (n-1)k+2	$\bar{y}_2$
⋮	⋮ ⋮ ⋮ ⋮	⋮
⋮	⋮ ⋮ ⋮ ⋮ ⋮	⋮
⋮	⋮ ⋮ ⋮ ⋮ ⋮	⋮
r	r, k+r, 2k+r, (i-1)k+r, ... (n-1)k+r	$\bar{y}_r$
⋮	⋮ ⋮ ⋮ ⋮ ⋮	⋮
⋮	⋮ ⋮ ⋮ ⋮ ⋮	⋮
⋮	⋮ ⋮ ⋮ ⋮ ⋮	⋮
k	k, 2k, 3k,... ,ik Nk	$\bar{y}_k$

$$(nk-1)S^2 = \sum_{r=1}^k \sum_{i=1}^n (y_{ri} - \bar{y}_r)^2 + \sum_{r=1}^k \sum_{i=1}^n (\bar{y}_r - \bar{Y})^2$$

$$(nk-1)S^2 = k(n-1)S_w^2 + n \sum_{r=1}^k (\bar{y}_r - \bar{Y})^2$$

$$Var(\bar{y}_{sy}) = \frac{N-1}{N} S^2 - \frac{k(n-1)}{N} S_w^2$$

$$Var(\bar{y}_{sy}) = \frac{N-1}{N} S^2 - \frac{k(n-1)}{N} S_w^2$$

Example:

The heights of the 30 trees from a certain area of a forest are given by

40, 38,36,35,32,35,32,30,31,29,37,41,30,28,24,25,26,27,29,32,34,36,21,19,17,22,35,
28,29,31

Select a systematic random sample of size 5.

Estimate the mean of the population

Estimate the variance.

We have

N=30, n=5;k=6

Random Start	Serial Number
1	1,7,13,19,25
2	2,8,14,20,26
3	3,9,15,21,27
4	4,10,16,22,28
5	5,11,17,23,29
6	6,12,18,24,30

Serial Number	Sampled Values
1,7,13,19,25	40, 32,30,29,17
2,8,14,20,26	38,30,28,32,22
3,9,15,21,27	36,31,24,34,35
4,10,16,22,28	35,29,25,36,28
5,11,17,23,29	32,37,26,21,29
6,12,18,24,30	35,41,27,19,31

Sampled Values	Sample Means
40,32,30,29,17	29.6
38,30,28,32,22	30
3	32
6,31,24,34,35	30.6
35,29,25,36,28	29
32,37,26,21,29	30.6
35,41,27,19,31	

Mean and Variance

$$E(\bar{y}_{sy}) = \frac{1}{k} \sum_{r=1}^k \bar{y}_r,$$

$$V(\bar{y}_{sy}) = \frac{1}{k} \sum_{r=1}^k (\bar{y}_r - \bar{Y})^2$$

$$\text{Sum}=181.8$$

$$\text{Mean}=30.3$$

$$\text{var}=0.89$$

$$\text{Sumdev}=5.34$$

Q . Drive the following expression of variance of sample mean under Systematic Sampling.

Proof:

$$V(\bar{y}_{sy}) = \frac{1}{k} \sum_{r=1}^k (\bar{y}_r - \bar{Y})^2$$

$$\rho_w = \frac{\frac{1}{nk(n-1)} \sum_{r=1}^k \sum_{i \neq u=1}^n \sum_{i=1}^n (y_{ri} - \bar{Y})(y_{ru} - \bar{Y})}{\left(\frac{N-1}{nk} \right) S^2} \left[\frac{\sum_{i=1}^n y_i}{\sum_{i=1}^n (y_{ri} - \bar{Y})} \right]^2$$

The intra class correlation between the pairs of units that are in the same systematic sample is

$$\begin{aligned} V(\bar{y}_{sy}) &= \frac{1}{n^2 k} \sum_{r=1}^k \left[\sum_{i=1}^n (y_{ri} - \bar{Y})^2 + \sum \sum (y_{ri} - \bar{Y})(y_{ru} - \bar{Y}) \right] \\ V(\bar{y}_{sy}) &= \frac{1}{n^2 k} \left[\sum_{r=1}^k \sum_{i=1}^n (y_{ri} - \bar{Y})^2 + \sum_{r=1}^k \sum_{i \neq u=1}^n \sum (y_{ri} - \bar{Y})(y_{ru} - \bar{Y}) \right] \\ V(\bar{y}_{sy}) &= \frac{1}{n^2 k} \left[(nk-1)S^2 + \sum_{r=1}^k \sum_{i \neq u=1}^n \sum (y_{ri} - \bar{Y})(y_{ru} - \bar{Y}) \right] \\ V(\bar{y}_{sy}) &= \frac{1}{n^2 k} \left[(nk-1)S^2 + \sum_{r=1}^k \sum_{i \neq u=1}^n \sum (y_{ri} - \bar{Y})(y_{ru} - \bar{Y}) \right] \\ V(\bar{y}_{sy}) &= \frac{1}{n^2 k} \left[(nk-1)S^2 + \sum_{r=1}^k \sum_{i \neq u=1}^n \sum (y_{ri} - \bar{y}_i)(y_{ru} - \bar{y}_u) \right] \end{aligned}$$

$$\rho_w(n-1)(nk-1)S^2 = \sum_{r=1}^k \sum_{i \neq u=1}^n \sum (y_{ri} - \bar{y}_i)(y_{ru} - \bar{y}_u) \quad \rho_w = \frac{E(y_{ri} - \bar{Y})(y_{ru} - \bar{Y})}{E(y_{ri} - \bar{Y})^2}$$

$$\begin{aligned} V(\bar{y}_{sy}) &= \frac{1}{n^2 k} \left[(nk-1)S^2 + \rho_w(n-1)(nk-1)S^2 \right] \\ V(\bar{y}_{sy}) &= \frac{(nk-1)S^2}{n^2 k} [1 + \rho_w(n-1)] \\ V(\bar{y}_{sy}) &= \frac{(nk-1)S^2}{n^2 k} [1 + \rho_w(n-1)] \end{aligned}$$

OR

$$V(\bar{y}_{sy}) = \frac{(nk-1)}{nk} \frac{S^2}{n} [1 + \rho_w(n-1)]$$

OR

$$V(\bar{y}_{sy}) = \frac{(N-1)S^2}{Nn} [1 + \rho_w(n-1)]$$

Q. Describe the comparison between SRS and Systematic Sampling on th basis of variance of sample mean?

Ans:

$$Var(\bar{y}_{sy}) = \frac{N-1}{N} S^2 - \frac{k(n-1)}{N} S_w^2$$

$$Var(\bar{y}_{srs}) = \frac{N-n}{Nn} S^2$$

$$Var(\bar{y}_{srs}) - Var(\bar{y}_{sy}) = \frac{N-n}{Nn} S^2 - \frac{N-1}{N} S^2 + \frac{k(n-1)}{N} S_w^2$$

$$\left(\frac{N-n}{Nn} - \frac{N-1}{N} \right) S^2 + \frac{(n-1)}{n} S_w^2$$

$$Var(\bar{y}_{srs}) - Var(\bar{y}_{sy}) = \frac{(n-1)}{n} (S_w^2 - S^2)$$

more efficient than $\bar{y}_{SRS}$ when $S_{wy}^2 > S^2$.

less efficient than $\bar{y}_{SRS}$ when $S_{wy}^2 < S^2$.

$$R.E = \frac{Var(\bar{y}_{srs})}{Var(\bar{y}_{sy})}$$

$$R.E = \frac{\frac{N-n}{Nn} S^2}{\frac{(N-1)S^2}{Nn} [1 + \rho_w(n-1)]}$$

$$R.E = \frac{1}{\frac{(N-1)}{N-n} [1 + \rho_w(n-1)]}$$

$$R.E = \frac{Var(\bar{y}_{srs})}{Var(\bar{y}_{sy})}$$

$$R.E = \frac{1}{\frac{(N-1)}{N-n} [1 + \rho_w(n-1)]}$$

$$\rho_w < -\frac{1}{nk-1}$$

$$\rho_w > -\frac{1}{nk-1}$$

Q. Derive the Variance of Stratified Sampling in Systematic setting?

Ans:

Group	Sample Composition	Sample Mean
1	1, k+1, 2k+1,..., (i-1)k+1, ... (n-1)k+1	$\bar{y}_1$
2	2, k+2, 2k+2,..., (i-1)k+2,... (n-1)k+2	$\bar{y}_2$
⋮	⋮ ⋮ ⋮ ⋮	⋮
⋮	⋮ ⋮ ⋮ ⋮ ⋮	⋮
⋮	⋮ ⋮ ⋮ ⋮ ⋮	⋮
r	r, k+r, 2k+r, (i-1)k+r, ... (n-1)k+r	$\bar{y}_r$
⋮	⋮ ⋮ ⋮ ⋮ ⋮	⋮
⋮	⋮ ⋮ ⋮ ⋮ ⋮	⋮
⋮	⋮ ⋮ ⋮ ⋮ ⋮	⋮
k	k, 2k, 3k,... ,ik Nk	$\bar{y}_k$

Stratified Sampling in Setting of Sys Sampling

$$\bar{y}_{str} = \frac{1}{N} \sum_{h=1}^L N_h \bar{y}_h$$

$$\bar{y}_{str} = \frac{1}{nk} \sum_{j=1}^n k \bar{y}_j$$

$$\bar{y}_{str} = \frac{1}{n} \sum_{j=1}^n \bar{y}_j$$

$$V(\bar{y}_{str}) = \frac{1}{n^2} \sum_{j=1}^n V(\bar{y}_j)$$

$$V(\bar{y}_{str}) = \frac{1}{n^2} \sum_{j=1}^n \frac{k-1}{k} S_j^2$$

$$V(\bar{y}_{str}) = \frac{k-1}{n^2 k} \sum_{j=1}^n S_j^2$$

$$V(\bar{y}_{str}) = \frac{k-1}{nk} S_{wst}^2$$

$$V(\bar{y}_{str}) = \frac{nk-n}{n.nk} S_{wst}^2$$

$$V(\bar{y}_{str}) = \frac{N-n}{nN} S_{wst}^2$$

$$S_j^2 = \frac{1}{k-1} \sum_{i=1}^k (y_{ij} - \bar{y}_{str})^2 = \frac{N-n}{nN} S_{wst}^2$$

is the mean sum of squares of units in the j^{th} stratum.

$$S_{wst}^2 = \frac{1}{n} \sum_{j=1}^n S_j^2 = \frac{1}{n(k-1)} \sum_{i=1}^k \sum_{j=1}^n (y_{ij} - \bar{y}_j)^2$$

is the mean sum of squares within strata (or rows).

Q . Describe a comparison between Stratified and Systematic Sampling on the basis of variance of sample mean.

Ans:

$$V(\bar{y}_{sy}) = \frac{1}{k} \sum_{r=1}^k (\bar{y}_r - \bar{Y})^2$$

$$V(\bar{y}_{sy}) = \frac{1}{k} \sum_{r=1}^k \left[\frac{\sum_{i=1}^n y_i}{n} - \frac{\sum_{i=1}^n \bar{y}_i}{n} \right]^2$$

$$V(\bar{y}_{sy}) = \frac{1}{n^2 k} \sum_{r=1}^k \left[\sum_{i=1}^n (y_{ri} - \bar{y}_i) \right]^2$$

$$V(\bar{y}_{sy}) = \frac{1}{n^2 k} \sum_{r=1}^k \left[\sum_{i=1}^n (y_{ri} - \bar{y}_i)^2 + \sum \sum (y_{ri} - \bar{y}_i)(y_{ru} - \bar{y}_u) \right]$$

$$V(\bar{y}_{sy}) = \frac{1}{n^2 k} \left[\sum_{r=1}^k \sum_{i=1}^n (y_{ri} - \bar{y}_i)^2 + \sum_{r=1}^k \sum_{i \neq u=1}^n \sum (y_{ri} - \bar{y}_i)(y_{ru} - \bar{y}_u) \right]$$

$$V(\bar{y}_{sy}) = \frac{1}{n^2 k} \left[(nk-1)S^2 + \sum_{r=1}^k \sum_{i \neq u=1}^n \sum (y_{ri} - \bar{y}_i)(y_{ru} - \bar{y}_u) \right]$$

Alternative form of variance

$$V(\bar{y}_{sy}) = \frac{1}{n^2 k} \left[(nk-1)S^2 + \sum_{r=1}^k \sum_{i \neq u=1}^n \sum (y_{ri} - \bar{y}_i)(y_{ru} - \bar{y}_u) \right]$$

The intra class correlation between the pairs of units that are in the same systematic sample is

$$\rho_{wst} = \frac{E(y_{ri} - \bar{y}_i)(y_{ru} - \bar{y}_u)}{E(y_{ri} - \bar{y}_i)^2}$$

$$\rho_w = \frac{\frac{1}{nk(n-1)} \sum_{r=1}^k \sum_{i \neq u=1}^n \sum (y_{ri} - \bar{y}_i)(y_{ru} - \bar{y}_u)}{\left(\frac{N-1}{nk} \right) S_{wst}^2}$$

$$\rho_{wst} (n-1)(nk-1)S_{wst}^2 = \sum_{r=1}^k \sum_{j \neq u=1}^n \sum (y_{rj} - \bar{y}_j)(y_{ru} - \bar{y}_u)$$

$$V(\bar{y}_{sy}) = \frac{1}{n^2 k} \left[(N-n)S_{wst}^2 + \sum_{r=1}^k \sum_{j \neq u=1}^n \sum (y_{rj} - \bar{y}_j)(y_{ru} - \bar{y}_u) \right]$$

$$V(\bar{y}_{sy}) = \frac{1}{n^2 k} \left[(N-n)S_{wst}^2 + \rho_{wst} (n-1)(N-n)S_{wst}^2 \right]$$

$$V(\bar{y}_{sy}) = \frac{(N-n)S_{wst}^2}{nN} \left[1 + \rho_{wst} (n-1) \right]$$

$$Var(\bar{y}_{str}) = \frac{N-n}{Nn} S_{wst}^2$$

$$Var(\bar{y}_{srs}) - Var(\bar{y}_{sy}) = \frac{(N-n)S_{wst}^2}{nN} \left[1 + \rho_{wst} (n-1) \right] - \frac{(N-n)S_{wst}^2}{nN}$$

$$R.E = \frac{Var(\bar{y}_{str})}{Var(\bar{y}_{sy})}$$

$$R.E = \frac{1}{\left[1 + \rho_{wst} (n-1) \right]}$$

Q . State the Stratified Sampling in terms of systematic sampling for population with linear trend?

Ans: The population increase according to linear trend. The variance of sample mean for SRSWOR is

$$Var(\bar{y}_{wor}) = \frac{(k-1)(nk+1)}{12}$$

For Stratified Sampling in Terms of Systematic Sampling

$$Var(\bar{y}_{str}) = \frac{N-n}{Nn} S_{wst}^2$$

$$S_{wst}^2 = \frac{\sum_{r=1}^k (y_r - \bar{Y}_k)^2}{k-1}$$

$$\begin{aligned} & \sum_{r=1}^k (y_r - \bar{Y}_k)^2 \\ &= \sum_{r=1}^k (y_r^2 + \bar{Y}_k^2 - 2y_r \bar{Y}_k) \\ &= \left(\sum_{r=1}^k y_r^2 + k\bar{Y}_k^2 - 2\bar{Y}_k \sum_{r=1}^k y_r \right) \\ &= \left(\sum_{r=1}^k y_r^2 + k\bar{Y}_k^2 - 2k\bar{Y}_k^2 \right) \\ &= \sum_{r=1}^k y_r^2 - k\bar{Y}_k^2 \\ &= \sum_{r=1}^k y_r^2 - \frac{\left(\sum_{r=1}^k y_r \right)^2}{k} \end{aligned}$$

$$\sum_{i=1}^N y_i = \sum_{i=1}^N i = 1 + 2 + \dots + N$$

$$\sum_{r=1}^k y_r = \sum_{r=1}^k r = 1 + 2 + \dots + k$$

$$\sum_{r=1}^k r = \frac{k(k+1)}{2}$$

$$\sum_{r=1}^k r^2 = \frac{k(k+1)(2k+1)}{6}$$

$$= \sum_{r=1}^k r^2 - \frac{\left(\sum_{r=1}^k r \right)^2}{k}$$

$$\left[\frac{k(k+1)(2k+1)}{6} - \frac{(k(k+1))^2}{4k} \right]$$

$$\frac{k(k+1)}{2} \left[\frac{(2k+1)}{3} - \frac{(k+1)}{2} \right]$$

$$= \frac{k(k+1)}{2} \left[\frac{(k-1)}{6} \right]$$

$$\sum_{r=1}^k (y_r - \bar{Y}_k)^2 = \frac{k(k+1)}{2} \left[\frac{(k-1)}{6} \right]$$

$$S_{wst}^2 = \frac{\sum_{r=1}^k (y_r - \bar{Y}_k)^2}{k-1}$$

$$S_{wst}^2 = \frac{k(k+1)}{12}$$

$$Var(\bar{y}_{str}) = \frac{N-n}{Nn} S_{wst}^2$$

$$Var(\bar{y}_{str}) = \frac{nk-n}{n^2k} S_{wst}^2$$

$$Var(\bar{y}_{str}) = \frac{k^2-1}{12n}$$

Q . Describe the variance of systematic sampling for population with linear trend.

Ans: Variance of sample mean under Systematic Sampling

$$Var(\bar{y}_{sy}) = \frac{\sum_{r=1}^k (\bar{y}_r - \bar{Y})^2}{k}$$

$$= \frac{1}{k} \left[\sum_{r=1}^k y_r^2 - \frac{\left(\sum_{r=1}^k y_r \right)^2}{k} \right]$$

$$\sum_{i=1}^N y_i = \sum_{i=1}^N i = 1 + 2 + \dots + N$$

$$\sum_{r=1}^k y_r = \sum_{r=1}^k r = 1 + 2 + \dots + k$$

$$\sum_{r=1}^k r = \frac{k(k+1)}{2}$$

$$\sum_{r=1}^k r^2 = \frac{k(k+1)(2k+1)}{6}$$

$$= \frac{1}{k} \left[\sum_{r=1}^k y_r^2 - \frac{\left(\sum_{r=1}^k y_r \right)^2}{k} \right]$$

$$= \frac{1}{k} \left[\sum_{r=1}^k r^2 - \frac{\left(\sum_{r=1}^k r \right)^2}{k} \right]$$

$$= \frac{1}{k} \left[\frac{k(k+1)(2k+1)}{6} - \frac{(k(k+1))^2}{4k} \right]$$

$$= \frac{k(k+1)}{2k} \left[\frac{(2k+1)}{3} - \frac{(k+1)}{2} \right]$$

$$= \frac{k(k+1)}{2k} \left[\frac{(k-1)}{6} \right]$$

$$= \frac{k(k+1)(k-1)}{12k}$$

$$= \frac{(k^2 - 1)}{12}$$

When n=1

$$Var(\bar{y}_{wor}) = \frac{(k-1)(nk+1)}{12}$$

$$Var(\bar{y}_{sy}) = \frac{k^2 - 1}{12n}$$

$$Var(\bar{y}_{sy}) = \frac{(k^2 - 1)}{12}$$

When $n > 1$

$$Var(\bar{y}_{wor}) = \frac{(k-1)(nk+1)}{12}$$

$$Var(\bar{y}_{str}) = \frac{k^2-1}{12n}$$

$$Var(\bar{y}_{sy}) = \frac{(k^2-1)}{12}$$

$$Var(\bar{y}_{str}) < Var(\bar{y}_{sy}) < Var(\bar{y}_{wor}) \quad \text{for } n > 1$$

$$Var(\bar{y}_{str}) \leq Var(\bar{y}_{sy}) \leq Var(\bar{y}_{wor}) \quad \text{for } n \geq 1$$

Example: A certain company claims about their daily production as

125, 135, 157, 192, 151, 175, 164, 169, 147, 150, 138, 167, 155, 159, 139, 147, 149, 158.

Select the systematic sample of size 3 using R language. Also find mean of sample.

Ans:

Here we have $N=18$ with $n=3$, so $k=6$

```
y<-c(125, 135,157,192,151,175,164,169,147,150,138,167,155,159,139,147,149,158)
```

```
  y<-c(125, 135,157,192,
```

```
    151,175,164,169,147,
```

```
    150,138,167,155,159,139,147,149,158)
```

```
n=3;N=length(y)
```

```
k=N/n
```

```
start <- sample(1:k, 1)
```

```
s <- seq(start, N, k)
```

```
sys.sample<-y[s]
```

```
mean(sys.sample)
```

Output

```
First run=154.33
```

```
2nd run=148
```

Example: 2

The heights of the 30 trees from a certain area of a forest are given by

40, 38,36,35,32,35,32,30,31,29,37,41,30,28,24,25,26,27,29,32,34,36,21,19,17,22,35,28,29,31

Select a systematic random sample of size 5. Also find sample mean.

Ans:

```
N=30, n=5;k=6
```

```
pop<-c(40, 38,36,35,32,35,32,30,31,29,37,41,30,28,24,25,26,27,29,32,34,  
36,21,19,17,22,35,28,29,31)
```

```
n=5;N=length(pop);
```

```
k=N/n
```

```
start <- sample(1:k, 1)
```

```
s <- seq(start, N, k)
```

```
sys.sample<-pop[s]
```

```
mean(sys.sample)
```

```
var(sys.sample)
```

```
sd(sys.sample)
```

Output

```
29.6(8.26)
```

```
29(sd=6.04)
```

Example: A certain company claims about their daily production as

125, 135, 157,192,151,175,164,169,147,150,138,167,155,159,139,147,149,158.

We are interested to select the systematic sample of size 3.

Obtain 10000 samples using systematic sampling. Find the mean of each sample. Find the mean of means and variance of means.

Ans:

Here we have $N=18$ with $n=3$, so $k=6$

```
pop<-c(125, 135,157,192,151,175,164,169,147,150,138,167,155,159,139,147,149,158)
```

```
  n=3;N=length(pop);
  k=N/n
  for(i in 1:10000)
  {
    start <- sample(1:k, 1)
    s <- seq(start, N, k)
    sys.sample<-pop[s]
    m[i]<- mean(sys.sample)
  }mean(m)
var(m)
```

Out put

Mean=154.29

Var=63.6

Example

The heights of the 30 trees from a certain area of a forest are given by

40, 38,36,35,32,35,32,30,31,29,37,41,30,28,24,25,26,27,29,32,34,36,21,19,17,22,35,
28,29,31

Select a systematic random sample of size 5.

Obtain 5000 samples using systematic sampling.

Find the mean of each sample.

Find the mean of means and variance of means.

Ans

We have

$N=30$, $n=5$; $k=6$

```
pop<-c(40,38,36,35,32,35,32,30,31,29,37,41,30,28,24,25,26,27,29,32,34,36,21,
19,17,22,35,28,29,31)
```

```
  n=5;N=length(pop);k=N/n
  for(i in 1:5000)
  {
```

```
start <- sample(1:k, 1)
s <- seq(start, N, k)
sys.sample<-pop[s]
m[i]<- mean(sys.sample)
}
mean(m)
var(m)
```

Output

Mean=30.29

Var=0.89

Example:

Generate a population of size 1000 values from normal distribution with mean=2 and standard deviation=3.

Select the 10000 samples each of size 50 using systematic sampling technique and estimate the mean of each sample.

Find the mean and variance of 10000 means.

Ans:

```
N=1000; n=50;k=N/n;m=c();
pop<-rnorm(N,mean=2,sd=3)
for(i in 1:10000)
{
start <- sample(1:k, 1)
s <- seq(start, N, k)
sys.sample<-pop[s]
m[i]=mean(sys.sample)
-} mean(m);var(m)
```

Out put

2.01667

0.1929605

Example: 2

Generate a population of size 500 values from normal distribution with mean=20 and standard deviation=10.

Select the 5000 sample each of size 50 using the systematic sampling technique and estimate the mean of each sample. Find the mean and variance of 5000 means.

Ans:

```
N=500; n=50;k=N/n;m=c();  
pop<-rnorm(N,mean=20,sd=10)  
for(i in 1:5000)  
{  
  start <- sample(1:k, 1)  
  s <- seq(start, N, k)  
  sys.sample<-pop[s]  
  m[i]=mean(sys.sample)  
}  
mean(m);var(m);
```

Output

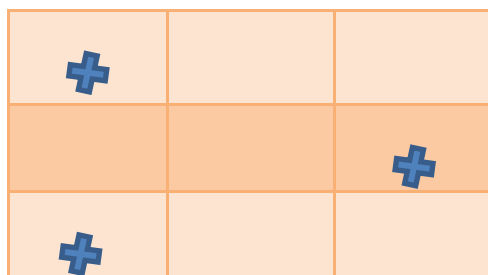
```
19.93  
1.456223
```

Cluster Sampling

Introduction:

➤ Cluster Sampling

- A cluster is the sampling unit consisting on the observation units.
- Any sampling method can be used for selection of clusters.
- All the units within a cluster are studied.
- Nine clusters each of same size.



➤ Clusters Settings

Cluster	1,	2	3	...,	j,	...	M	Sample Mean
1	y ₁₁	y ₁₂	y ₁₃	...,	y _{1j} ,	...	y _{1M}	$\frac{y_{1.}}{M} = \bar{y}_1$
2	y ₂₁	y ₂₂	y ₂₃	...,	y _{2j} ,	...	y _{2M}	
‘	‘		‘		‘		‘	‘
‘	‘		‘		‘		‘	‘
i	y _{i1}	y _{i2}	y _{i3}	...,	y _{ij} ,	...	y _{iM}	$\frac{y_{i.}}{M} = \bar{y}_i$
‘	‘		‘		‘		‘	‘
‘	‘		‘		‘		‘	‘
N	y _{N1}	y _{N2}	y _{N3}	...,	y _{Nj} ,	...	y _{NM}	$\bar{y}_N$

Total	y ₁	y ₁	y ₁	y ₁	$\bar{Y}$
-------	----------------	----------------	----------------	----------------	-----------

➤ **Notations**

- Suppose we have a population of N Clusters each of size M.
- NM= Total elements in population.
- y_{ij} = observation value of j^{th} element in i^{th} cluster.
- nM= Total elements in sample.
- $y_i = y_i =$ Total of i^{th} cluster.

$$y_i = \sum_{j=1}^M y_{ij}$$

- $\bar{y}$ = Sample mean per cluster.

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

- $\bar{\bar{y}}$ = Overall sample mean.

$$\bar{\bar{y}} = \frac{\sum_{j=1}^M \sum_{i=1}^n y_{ij}}{Mn}$$

- $\bar{\bar{Y}}$ = Overall population mean.

$$\bar{\bar{y}} = \frac{\sum_{j=1}^M \sum_{i=1}^n y_{ij}}{Mn}$$

- Let y_i be the total of i^{th} cluster.

$$\bar{\bar{y}} = \frac{\bar{y}}{M}$$

➤ **Unbiased mean estimator**

$$M E(\bar{\bar{y}}) = E(\bar{y})$$

$$M E(\bar{\bar{y}}) = \bar{Y}$$

$$E(\bar{\bar{y}}) = \bar{\bar{Y}}$$

Q. Find the variance of Sample Mean in Cluster Sampling considering equal cluster sizes.

Ans:

➤ **Unbiased mean estimator**

- The unbiased mean estimator is,

$$\bar{\bar{y}} = \frac{\sum_{i=1}^n \sum_{j=1}^M y_{ij}}{Mn}$$

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

➤ **Variance of mean estimator**

$$V(\bar{\bar{y}}) = \left(\frac{1-f}{n} \right) S_b^2$$

$$S_b^2 = \frac{\sum_{i=1}^N (\bar{y}_i - \bar{\bar{Y}})^2}{N-1}$$

$$V(\bar{\bar{y}}) = \left(\frac{1-f}{n} \right) \frac{\sum_{i=1}^N (\bar{y}_i - \bar{\bar{Y}})^2}{N-1}$$

$$\sum_{i=1}^N (\bar{y}_i - \bar{\bar{Y}})^2 = \sum_{i=1}^N \left[\frac{1}{M} \sum_{j=1}^M (y_{ij} - \bar{\bar{Y}}) \right]^2$$

$$\sum_{i=1}^N \left[\frac{1}{M} \sum_{j=1}^M (y_{ij} - \bar{\bar{Y}}) \right]^2 = \frac{1}{M^2} \sum_{i=1}^N \sum_{j=1}^M (y_{ij} - \bar{\bar{Y}})^2 + \frac{1}{M^2} \sum_{i=1}^N \sum_{j \neq k=1}^n \sum (y_{ij} - \bar{\bar{Y}})(y_{ik} - \bar{\bar{Y}})$$

$$\sum_{i=1}^N \left[\frac{1}{M} \sum_{j=1}^M (y_{ij} - \bar{\bar{Y}}) \right]^2 = \frac{1}{M^2} \left[(NM-1)S^2 + \sum_{i=1}^N \sum_{j \neq k=1}^n \sum (y_{ij} - \bar{\bar{Y}})(y_{ik} - \bar{\bar{Y}}) \right]$$

$$\sum_{i=1}^N \left[\frac{1}{M} \sum_{j=1}^M (y_{ij} - \bar{\bar{Y}}) \right]^2 = \frac{1}{M^2} \left[(NM-1)S^2 + \sum_{i=1}^N \sum_{j \neq k=1}^n \sum (y_{ij} - \bar{\bar{Y}})(y_{ik} - \bar{\bar{Y}}) \right]$$

The interclass correlation between the elements within a cluster,

$$\rho_w = \frac{E(y_{ij} - \bar{\bar{Y}})(y_{ik} - \bar{\bar{Y}})}{E(y_{ij} - \bar{\bar{Y}})^2}$$

$$\rho_w = \frac{\frac{1}{NM(M-1)} \sum_{i=1}^N \sum_{j \neq k=1}^n \sum (y_{ij} - \bar{\bar{Y}})(y_{ik} - \bar{\bar{Y}})}{\left(\frac{NM-1}{NM} \right) S^2}$$

$$\rho_w (M-1)(NM-1)S^2 = \sum_{i=1}^N \sum_{j \neq k=1}^n \sum (y_{ij} - \bar{\bar{Y}})(y_{ik} - \bar{\bar{Y}})$$

$$\sum_{i=1}^N \left[\frac{1}{M} \sum_{j=1}^M (y_{ij} - \bar{\bar{Y}}) \right]^2 = \frac{1}{M^2} \left[(NM-1)S^2 + \rho_w (M-1)(NM-1)S^2 \right]$$

$$\begin{aligned}
 V(\bar{y}) &= \left(\frac{1-f}{n} \right) \frac{\sum_{i=1}^N (\bar{y}_i - \bar{Y})^2}{N-1} = \frac{1}{M^2} \left(\frac{1-f}{n} \right) \frac{\sum_{i=1}^N \left[\sum_{j=1}^M (y_{ij} - \bar{Y}) \right]^2}{N-1} \\
 &= \frac{1}{M^2} \left(\frac{1-f}{n} \right) \frac{(NM-1)S^2 + \rho_w(M-1)(NM-1)S^2}{N-1} \\
 V(\bar{y}) &= \frac{1}{M^2} \left(\frac{1-f}{n} \right) \\
 &\quad \frac{(NM-1)S^2}{N-1} [1 + \rho_w(M-1)] \\
 V(\bar{y}) &= \left(\frac{S^2}{nM} \right) [1 + \rho_w(M-1)] \quad NM-1 = NM, \\
 &\quad N-1 = N
 \end{aligned}$$

Q. Compare the simple random sampling and cluster Sampling in terms of variances of sample means such that

$$\begin{aligned}
 Var(\bar{y}_{srs}) &= \frac{N-n}{N} \frac{S^2}{n} \\
 V(\bar{y}_{srs}) &= \left(\frac{NM-nM}{NM} \right) \frac{S^2}{nM}
 \end{aligned}$$

Ans:

➤ **Comparison**

$$\begin{aligned}
 Var(\bar{y}_{srs}) &= \frac{N-n}{N} \frac{S^2}{n} \\
 V(\bar{y}_{srs}) &= \left(\frac{NM-nM}{NM} \right) \frac{S^2}{nM}
 \end{aligned}$$

$$S^2 = \frac{1}{NM-1} \sum_{i=1}^N \sum_{j=1}^M (y_{ij} - \bar{Y})^2 \quad \bar{S}_w^2 = \frac{1}{N} \sum_{i=1}^N S_i^2$$

Mean sum of squares within clusters in the population

$$S^2(NM-1) = \sum_{i=1}^N \sum_{j=1}^M (y_{ij} - \bar{Y})^2 \quad S_i^2 = \frac{1}{M-1} \sum_{j=1}^M (y_{ij} - \bar{y}_i)^2$$

Mean sum of squares for ith cluster

$$\begin{aligned}
 &= \sum_{i=1}^N \sum_{j=1}^M ((y_{ij} - \bar{y}_i) + (\bar{y}_i - \bar{Y}))^2 \\
 &= \sum_{i=1}^N \sum_{j=1}^M (y_{ij} - \bar{y}_i)^2 + M \sum_{i=1}^N (\bar{y}_i - \bar{Y})^2
 \end{aligned}$$

$$\begin{aligned}
 &= N(M-1)S_w^2 + M(N-1)S_b^2 \\
 R.E &= \frac{Var(\bar{y}_{srs})}{Var(\bar{\bar{y}})} \\
 R.E &= \frac{S^2}{MS_b^2} \\
 R.E &= \frac{N(M-1)S_w^2 + M(N-1)S_b^2}{MS_b^2(NM-1)} \\
 &= \frac{1}{(NM-1)} \left[\frac{N(M-1)S_w^2}{MS_b^2} + \frac{M(N-1)S_b^2}{MS_b^2} \right] \\
 &= \frac{1}{(NM-1)} \left[\frac{N(M-1)S_w^2}{MS_b^2} + (N-1) \right]
 \end{aligned}$$

- This value increases when S_w is large and S_b is small. So cluster sampling will be efficient if clusters are so
- Formed that the variation the between cluster means is as small as possible while variation within the clusters is as large as possible.

Q. Compare the simple random sampling and cluster Sampling in terms of interaclass correlation

Ans:

$$\begin{aligned}
 R.E &= \frac{Var(\bar{y}_{srs})}{Var(\bar{\bar{y}})} \\
 V(\bar{\bar{y}}) &= \left(\frac{S^2}{nM} \right) [1 + \rho_w(M-1)] \\
 V(\bar{y}_{srs}) &= \frac{S^2}{nM} \\
 R.E &= \frac{\frac{S^2}{nM}}{\left(\frac{S^2}{nM} \right) [1 + \rho_w(M-1)]} \\
 R.E &= \frac{1}{[1 + \rho_w(M-1)]}
 \end{aligned}$$

Q. Explain the concept of cluster sampling for unequal cluster sizes.

Ans:

➤ **Clusters Settings**

								Sample Mean
Cluster	1,	2	3	...,	j,	...	M_i	
1	y_{11}	y_{12}	y_{13}	...,	y_{1j} ,	...	y_{1M1}	$\frac{y_{1.}}{M} = \bar{y}_1$
2	y_{21}	y_{22}	y_{23}	...,	y_{2j} ,	...	y_{2M2}	
‘	‘			‘		‘	‘	‘
‘	‘			‘		‘	‘	‘
i	y_{i1}	y_{i2}	y_{i3}	...,	y_{ij} ,	...	y_{iM_i}	$\frac{y_{i.}}{M} = \bar{y}_i$
‘	‘			‘		‘	‘	‘
‘	‘			‘		‘	‘	‘
N	y_{N1}	y_{N2}	y_{N3}	...,	y_{Nj} ,	...	y_{NMN}	$\bar{y}_N$
Total	$y_{.1}$	$y_{.1}$	$y_{.1}$		$y_{.1}$			$\bar{\bar{Y}}$

➤ Mean Estimator

- Since the cluster sizes are unequal, the total size is,

$$M' = \sum_{i=1}^N M_i$$

- The mean of i^{th} cluster is,

$$\bar{y}_i = \frac{\sum_{j=1}^{M_i} y_{ij}}{M_i}$$

- The overall mean is

$$\bar{Y} = \frac{\sum_{i=1}^N \sum_{j=1}^{M_i} y_{ij}}{\sum_{i=1}^N M_i}$$

$$\bar{Y} = \frac{\sum_{i=1}^N M_i \bar{y}_i}{M'}$$

➤ Expected Value of Sample Mean

- The mean of clusters

$$\bar{Y} = \frac{\sum_{i=1}^N M_i \bar{y}_i}{M'}$$

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

$$E(\bar{\bar{y}}) = \frac{\sum_{i=1}^N y_i}{N}$$

Q. Find the expression of Bias for the mean estimator in cluster sampling for unequal cluster sizes.

Ans.

The bias is defined as

$$Bias(T) = E(T) - \theta$$

$$Bias(\bar{\bar{y}}) = E(\bar{\bar{y}}) - \bar{Y}$$

$$\begin{aligned} Bias(\bar{\bar{y}}) &= \frac{\sum_{i=1}^N y_i}{N} - \frac{\sum_{i=1}^N M_i \bar{y}_i}{M'} \\ &= \frac{1}{M'} \left(\frac{M' \sum_{i=1}^N y_i}{N} - \sum_{i=1}^N M_i \bar{y}_i \right) \\ &= \frac{1}{M'} \left(\frac{\sum_{i=1}^N M_i \sum_{i=1}^N y_i}{N} - \sum_{i=1}^N M_i \bar{y}_i \right) \\ &= -\frac{1}{M'} \left(\sum_{i=1}^N M_i \bar{y}_i - \frac{\sum_{i=1}^N M_i \sum_{i=1}^N y_i}{N} \right) \\ &= -\frac{1}{M_0} \sum_{i=1}^N (M_i - \bar{M})(\bar{y}_i - \bar{Y}) \\ &= -\left(\frac{N-1}{M_0} \right) S_{\bar{y}\bar{y}} \end{aligned}$$

Q. Find the expression of mean square error for the mean estimator in cluster sampling for unequal cluster sizes

$$\begin{aligned}
 MSE(T) &= Var(T) + (Bias(T))^2 \\
 MSE(\bar{y}) &= Var(\bar{y}) + (Bias(\bar{y}))^2 \\
 MSE(\bar{y}) &= \left(\frac{1-f}{n} \right) S_b^2 \\
 &+ \left(-\frac{(N-1)}{M'} Cov(m, y) \right)^2 \\
 S_b^2 &= \frac{1}{N-1} \sum_{i=1}^N (\bar{y}_i - \bar{Y})^2 \\
 S_{my} &= \frac{1}{N-1} \sum_{i=1}^N (M_i - \bar{M})(\bar{y}_i - \bar{Y}).
 \end{aligned}$$

$$\begin{aligned}
 MSE(T) &= Var(T) + (Bias(T))^2 \\
 MSE(\bar{y}) &= Var(\bar{y}) + (Bias(\bar{y}))^2 \\
 MSE(\bar{y}) &= \left(\frac{1-f}{n} \right) S_b^2 \\
 &+ \left(-\frac{(N-1)}{M'} Cov(m, y) \right)^2
 \end{aligned}$$

An estimate of $Var(\bar{y}_c)$ is

$$\widehat{Var}(\bar{y}_c) = \frac{N-n}{Nn} s_b^2$$

$$\text{where } s_b^2 = \frac{1}{n-1} \sum_{i=1}^n (\bar{y}_c - \bar{\bar{y}}_c)^2.$$

Q. Find the expected value Weighted Mean For Unequal Cluster, where is weighted mean is given by

$$\bar{\bar{y}}_w = \frac{\sum_{i=1}^n M_i \bar{y}_i}{n\bar{M}}$$

Answer:

➤ **Weighted Mean**

- Since the cluster sizes are unequal, the mean of cluster size is,

$$\bar{M} = \frac{M'}{N}$$

- The weighted mean based on the size of i^{th} cluster is,

$$\bar{\bar{y}}_w = \frac{\sum_{i=1}^n M_i \bar{y}_i}{n\bar{M}}$$

➤ **Expectation of Weighted Mean**

- Taking expectation on both sides

$$\begin{aligned}
 E(\bar{y}_w) &= E\left(\frac{\sum_{i=1}^n M_i \bar{y}_i}{n\bar{M}}\right) \\
 E(\bar{y}_w) &= \frac{1}{n} \sum_{i=1}^n \frac{1}{\bar{M}} E(\bar{y}_i M_i) \\
 &= \frac{1}{n} \sum_{i=1}^n \frac{1}{\bar{M}} M_i \bar{y}_i \\
 &= \frac{1}{\bar{M}} \sum_{i=1}^n \sum_{j=1}^{M_i} y_{ij} \\
 &= \bar{Y}
 \end{aligned}$$

Q. Find the variance expression of weighted Mean for unequal cluster, where weighted mean is given by

$$\bar{y}_w = \frac{\sum_{i=1}^n M_i \bar{y}_i}{n\bar{M}}$$

Ans:

- The variance is given by

$$\begin{aligned}
 V(\bar{y}_w) &= V\left(\frac{\sum_{i=1}^n M_i \bar{y}_i}{n\bar{M}}\right) \\
 V(\bar{y}_w) &= \left(\frac{1-f}{n}\right) S_{bw}^2 \\
 S_{bw}^2 &= \frac{1}{N-1} \sum_{i=1}^N \left(\frac{M_i}{\bar{M}} \bar{y}_i - \bar{Y}\right)^2
 \end{aligned}$$

➤ **The Estimator of Variance**

The estimator of variance is given by

$$\begin{aligned}
 \hat{V}(\bar{y}_w) &= \left(\frac{1-f}{n}\right) \hat{S}_{bw}^2 \\
 \hat{S}_{bw}^2 &= \frac{1}{n-1} \sum_{i=1}^n \left(\frac{M_i}{\bar{M}} \bar{y}_i - \bar{y}_w\right)^2
 \end{aligned}$$

Example: Find mean and variance of sample mean in cluster sampling using the following data set.

Ans:

Population of Six Clusters

						Total
Cluster-1	125	115	129	134	111	614
Cluster-2	134	125	142	141	131	673
Cluster-3	144	143	122	134	126	669
Cluster-4	114	111	134	131	146	636
Cluster-5	119	126	122	129	130	626
Cluster-6	140	125	124	124	115	628

➤ **Population Mean**

$$\bar{\bar{Y}} = \frac{\sum_{i=1}^N \sum_{j=1}^M Y_{ij}}{MN}$$

$$\bar{\bar{Y}} = \frac{3846}{30} = 128.2$$

➤ **Variance of Sample Mean Using Population**

$$Var(\bar{y}) = \left(\frac{1-f}{n} \right) S_b^2$$

$$S_b^2 = \frac{\sum_{i=1}^N (\bar{y}_i - \bar{\bar{Y}})^2}{N-1}$$

$$Var(\bar{y}) = \left(\frac{1-f}{n} \right) \frac{\sum_{i=1}^N (\bar{y}_i - \bar{\bar{Y}})^2}{N-1},$$

$$Var(\bar{y}) = \frac{N-n}{Nn} \frac{1}{N-1} \sum_{i=1}^N (\bar{y}_i - \bar{\bar{Y}})^2$$

➤ **Variance of Sample Mean**

$$\frac{N-n}{Nn} \frac{1}{N-1} \sum_{i=1}^N (\bar{y}_i - \bar{\bar{Y}})^2$$

$$(\bar{y}_i - \bar{\bar{Y}})^2$$

y_i	$\bar{y}_i$	$(\bar{y}_i - \bar{\bar{Y}})^2$
614	122.8	29.16
673	134.6	40.96
669	133.8	31.36
636	127.2	1.00
626	125.2	9.00
628	125.6	6.76

$$\bar{\bar{Y}} = 128.2$$

$$\frac{N-n}{Nn} \frac{1}{N-1} \sum_{i=1}^N (\bar{y}_i - \bar{\bar{Y}})^2$$

$$= \frac{6-3}{6 \times 3} \frac{1}{6-1} (118.24)$$

$$= 1.304$$

➤ **Population of Six Clusters**

						Total
Cluster-1	125	115	129	134	111	614
Cluster-2	134	125	142	141	131	673
Cluster-3	144	143	122	134	126	669
Cluster-4	114	111	134	131	146	636
Cluster-5	119	126	122	129	130	626
Cluster-6	140	125	124	124	115	628

➤ **Mean and Variance Using All the Observations**

$$\bar{\bar{Y}} = \frac{3846}{30} = 128.2$$

$$V(\bar{\bar{y}}) = 1.304$$

Example: Find the mean and variance by taking a sample of size 3 from the previous data set under cluster sampling.

Ans:

Sampled clusters

						Total
Cluster-3	144	143	122	134	126	669
Cluster-4	114	111	134	131	146	636
Cluster-6	140	125	124	124	115	628

$$\bar{\bar{y}} = \frac{\sum_{i=1}^n \sum_{j=1}^M y_{ij}}{Mn} = \frac{\sum_{i=1}^3 \sum_{j=1}^5 y_{ij}}{5 \times 3} = \frac{1933}{15} = 128.87$$

➤ **Estimated Variance of Sample mean**

$$\frac{N-n}{Nn} \frac{1}{n-1} \sum_{i=1}^n (\bar{y}_i - \bar{\bar{y}})^2$$

y_i	$\bar{y}_i$	$(\bar{y}_i - \bar{\bar{y}})^2$
669	133.8	31.36
636	127.2	1.00
628	125.6	6.76

$$\bar{\bar{y}} = 128.87$$

➤ **Estimated Variance**

$$\begin{aligned} & var(\bar{y}) \\ &= \frac{N-n}{Nn} \frac{1}{n-1} \sum_{i=1}^3 (\bar{y}_i - \bar{\bar{y}})^2 \\ &= \frac{6-3}{6 \times 3} \times \frac{1}{2} [37.78] \\ &= 3.15 \end{aligned}$$

➤ **Example-2**

- 420 trees is divided into 105 clusters.
- Each cluster of size 4.
- A simple random sample of 15 clusters is selected.
- Estimate the mean yield by using cluster sampling.

➤ **Sample of 15 Clusters**

Cluster	1st Tree	2nd Tree	3rd Tree	4th Tree	Total
1	5.17	26.11	29.54	19.64	80.46
2	3.53	40.76	5.15	1.25	50.69
3	14.23	16.89	28.93	21.70	81.75
4	7.13	34.35	12.18	9.86	63.52
5	27.59	38.10	24.74	6.77	97.20
6	5.58	4.84	0.69	15.69	26.80
7	12.66	32.53	16.92	37.02	99.13
8	6.40	11.68	40.05	5.12	63.25
9	54.21	34.63	52.55	37.96	179.35
10	37.94	47.07	19.64	29.11	133.76
11	45.98	5.17	1.47	6.53	59.15
12	0.87	3.56	4.86	23.56	32.85
13	26.11	10.93	10.08	11.18	58.30
14	1.94	35.97	29.54	25.28	92.73
15	11.08	0.65	4.21	7.56	23.50

$$\bar{\bar{y}} = \frac{\sum_{i=1}^n \sum_{j=1}^M y_{ij}}{Mn}$$

$$\bar{\bar{y}} = \frac{1142.44}{60} = 19.0407$$

➤ **Estimated Variance**

$$V(\bar{\bar{y}}) = \left(\frac{1-f}{n} \right) S_b^2$$

$$S_b^2 = \frac{\sum_{i=1}^N (\bar{y}_i - \bar{\bar{Y}})^2}{N-1}$$

$$V(\bar{\bar{y}}) = \left(\frac{1-f}{n} \right) \frac{\sum_{i=1}^N (\bar{y}_i - \bar{\bar{Y}})^2}{N-1}$$

$$Var(\bar{y}) = \frac{N-n}{Nn} \frac{1}{n-1} \sum_{i=1}^n (\bar{y}_i - \bar{y}_c)^2$$

$$SS = 1495.5596$$

cluster	y_i	$\bar{y}_i$	$(\bar{y}_i - \bar{\bar{Y}})^2$
1	80.4600	20.1150	1.162
2	50.6900	12.6725	41.082
3	81.7500	20.4375	1.961

4	63.5200	15.8800	9.967
5	97.2000	24.3000	27.699
6	26.8000	6.7000	152.202
7	99.1300	24.7825	33.011
8	63.2500	15.8125	10.397
9	179.350	44.8375	665.666
10	133.760	33.4400	207.446
11	59.1500	14.7875	18.058
12	32.8500	8.2125	117.170
13	58.3000	14.5750	19.909
14	92.7300	23.1825	17.185
15	23.5000	5.8750	173.238

$$\begin{aligned}
 &Var(\bar{y}) \\
 &= \frac{N-n}{Nn} \frac{1}{n-1} \sum_{i=1}^n (\bar{y}_i - \bar{\bar{y}})^2 \\
 &= \frac{90}{105 \times 15} \times \frac{1}{14} [1495.5596] \\
 &= 6.104
 \end{aligned}$$

Q. Define the following data in R language in form of clusters and mean and variance.

						Total
Cluster-1	125	115	129	134	111	614
Cluster-2	134	125	142	141	131	673
Cluster-3	144	143	122	134	126	669
Cluster-4	114	111	134	131	146	636

Cluster-5	119	126	122	129	130	626
Cluster-6	140	125	124	124	115	62

Ans:

How to Do This in R?

Cluster-1	125	115	129	134	111
Cluster-2	134	125	142	141	131
Cluster-3	144	143	122	134	126
Cluster-4	114	111	134	131	146
Cluster-5	119	126	122	129	130
Cluster-6	140	125	124	124	115

#--Defining Clusters in R----

Clu1<-c(125,115,129,134,111)

Clu2<-c(134,125,142,141,131)

Clu3<-c(144,143,122,134,126)

Clu4<-c(114,111,134,131,146)

Clu5<-c(119,126,122,129,130)

Clu6<-c(140,125,124,124,115)

pop<-c(Cluc1,Cluc2,Cluc3,Cluc4,Cluc5,Cluc6)

#----Sum of clusters----

sumc1<-sum(Cluc1)

sumc2<-sum(Cluc2)

sumc3<-sum(Cluc3)

sumc4<-sum(Cluc4)

sumc5<-sum(Cluc5)

sumc6<-sum(Cluc6)

#----Grand total----

Yi=c(sumc1,sumc2,sumc3,sumc4,sumc5,sumc6)

t.sum=sum(Yi)

N=6;M=5;

#----Mean of each cluster----

Yibar<-Yi/M

#----Population Mean----

pop.mean=sum(Yi)/(N*M)

#----Sum of Squares----

dv.p<-(Yibar-pop.mean)^2

sdv.p<-sum(dv.p)

vr.p<-sdv.p/(N-1)

#----The Variance----

cvr.p<-((N-n)/(N*n))*vr.p

$$\frac{N-n}{Nn} \frac{1}{N-1} \sum_{i=1}^N (\bar{y}_i - \bar{Y})^2$$

Q. Perform the cluster sampling using R language with following data set. Also find mean and variance.

Cluster-1	125	115	129	134	111
Cluster-2	134	125	142	141	131
Cluster-3	144	143	122	134	126
Cluster-4	114	111	134	131	146
Cluster-5	119	126	122	129	130
Cluster-6	140	125	124	124	115

Answer:

➤ **Defining Clusters**

Cluster-1	125	115	129	134	111
Cluster-2	134	125	142	141	131
Cluster-3	144	143	122	134	126
Cluster-4	114	111	134	131	146
Cluster-5	119	126	122	129	130
Cluster-6	140	125	124	124	115

#--Defining Clusters in R----

Clu1<-c(125,115,129,134,111)

Clu2<-c(134,125,142,141,131)

Clu3<-c(144,143,122,134,126)

Clu4<-c(114,111,134,131,146)

Clu5<-c(119,126,122,129,130)

Clu6<-c(140,125,124,124,115)

➤ **Sum Of Clusters**

```
#----Population----  
pop<-c(Clu1,Clu2,Clu3,Clu4,Clu5,Clu6)  
#----Sum of clusters----  
sumc1<-sum(Clu1)  
sumc2<-sum(Clu2)  
sumc3<-sum(Clu3)  
sumc4<-sum(Clu4)  
sumc5<-sum(Clu5)  
sumc6<-sum(Clu6)
```

➤ **Sampling and Estimated Mean**

```
#----Grand total----  
Yi=c(sumc1,sumc2,sumc3,sumc4,sumc5,sumc6)  
t.sum=sum(Yi)  
#-----Sampling-----  
n=3;N=6;M=5;  
yi=sample(Yi,n)  
#----Estimated Mean----  
yibar<-yi/M  
Est.mean=sum(yi)/(n*M)
```

➤ **Estimated Variance**

```
#----Sum of Squares----  
dv<-(yibar-Est.mean)^2  
sdv<-sum(dv)  
vr<-sdv/(n-1)  
#----The Variance----  
cvr<-((N-n)/(N*n))*vr
```

$$\frac{N-n}{Nn} \frac{1}{n-1} \sum_{i=1}^n (\bar{y}_i - \bar{\bar{y}})^2$$

Q.70 Find the mean and variance for Unequal Clusters in cluster sampling using the following data

Clus-1 125 115 129 134

Clus-2 134 125 142 141 131 151 164 139 141

Clus-3	144	143	122	134	126	157				
Clus-4	114	111	134	131	146	152	131			
Clus-5	119	126	122	129	130					
Clus-6	140	125	124	124	115	111	148	157	143	151

Ans
Population of Six Clusters

Clus-1	125	115	129	134						
Clus-2	134	125	142	141	131	151	164	139	141	
Clus-3	144	143	122	134	126	157				
Clus-4	114	111	134	131	146	152	131			
Clus-5	119	126	122	129	130					
Clus-6	140	125	124	124	115	111	148	157	143	151

➤ **Population Mean in Case of Unequal Cluster Sizes**

$$\bar{\bar{Y}} = \frac{\sum_{i=1}^N M_i \bar{y}_i}{M'} \Rightarrow \bar{\bar{Y}} = 133.66$$

➤ **Estimator-I: Mean of Cluster Mean**

- The mean of cluster means

$$\bar{\bar{y}} = \frac{\sum_{i=1}^n \bar{y}_i}{n}$$

Clus-3	144	143	122	134	126	157	
Clus-4	114	111	134	131	146	152	131
Clus-5	119	126	122	129	130		

$$\bar{\bar{y}} = \frac{\sum_{i=1}^n \bar{y}_i}{n}$$

➤ **Mean Square Error**

$$\begin{aligned} MSE(\bar{\bar{y}}) &= Var(\bar{\bar{y}}) + (Bias(\bar{\bar{y}}))^2 \\ MSE(\bar{\bar{y}}) &= \left(\frac{1-f}{n} \right) S_b^2 \\ &+ \left(-\frac{(N-1)}{M'} Cov(m, \bar{y}) \right)^2 \\ S_b^2 &= \frac{1}{N-1} \sum_{i=1}^N (\bar{y}_i - \bar{\bar{y}})^2 \\ S_{m\bar{y}} &= \frac{1}{N-1} \sum_{i=1}^N (M_i - \bar{M})(\bar{y}_i - \bar{\bar{y}}). \end{aligned}$$

➤ **Population Mean in Case of Unequal Cluster Sizes**

$$\bar{\bar{Y}} = 133.66, S_b^2 = 39.81$$

➤ **Mean Square Error**

$$\begin{aligned} MSE(\bar{\bar{y}}) &= Var(\bar{\bar{y}}) + (Bias(\bar{\bar{y}}))^2 \\ MSE(\bar{\bar{y}}) &= \left(\frac{1-f}{n} \right) S_b^2 \\ &+ \left(-\frac{(N-1)}{M'} Cov(m, \bar{y}) \right)^2 \end{aligned}$$

Q. Find the weighted mean and variance for Unequal Clusters in cluster sampling using the following data

Clus-1	125	115	129	134						
Clus-2	134	125	142	141	131	151	164	139	141	
Clus-3	144	143	122	134	126	157				
Clus-4	114	111	134	131	146	152	131			
Clus-5	119	126	122	129	130					
Clus-6	140	125	124	124	115	111	148	157	143	151

Answer:

➤ **Population of Six Clusters**

Clus-1	125	115	129	134
---------------	-----	-----	-----	-----

Clus-2	134	125	142	141	131	151	164	139	141
Clus-3	144	143	122	134	126	157			
Clus-4	114	111	134	131	146	152	131		
Clus-5	119	126	122	129	130				
Clus-6	140	125	124	124	115	111	148	157	143 151

➤ **Population Mean in Case of Unequal** $\bar{Y} = \frac{\sum_{i=1}^N M_i \bar{y}_i}{M'} \Rightarrow \bar{Y} = 133.66$

➤ **Estimator II: Weighted Mean**

- Since the cluster sizes are unequal, the total size is

$$M' = \sum_{i=1}^N M_i$$

- The weighted mean based on the size of i^{th} cluster is

$$\bar{\bar{y}}_w = \frac{\sum_{i=1}^n M_i \bar{y}_i}{n\bar{M}}$$

➤ **Estimator II: Weighted Mean**

Clus-3							
	144	143	122	134	126	157	
Clus-4							
	114	111	134	131	146	152	131
Clus-5							
	119	126	122	129	130		

$$\bar{\bar{y}}_w = \frac{\sum_{i=1}^n M_i \bar{y}_i}{n\bar{M}}$$

➤ **Variance of Weighted Mean**

- The variance is given by

$$V(\bar{\bar{y}}_w) = V\left(\frac{\sum_{i=1}^n M_i \bar{y}_i}{n\bar{M}}\right)$$

$$V(\bar{\bar{y}}) = \left(\frac{1-f}{n}\right) S_{bw}^2$$

$$S_b^2 = \frac{1}{N-1} \sum_{i=1}^N \left(\frac{M_i}{\bar{M}} \bar{y}_i - \bar{Y} \right)^2$$

➤ **Estimator II: Weighted Mean**

$\bar{y}_i$	Mi	$\frac{\bar{y}_i \times M_i}{\bar{M}}$
125.75	4	73.61
140.89	9	185.56
137.67	6	120.88
131.29	7	134.49
125.20	5	91.61
133.80	10	195.80

$$V(\bar{y}) = \left(\frac{1-f}{n} \right) S_{bw}^2$$

$$\bar{Y} = 133.66$$

Q.72 Perform the cluster sampling using R language with following data set. Also find mean and variance.

Clus-1	125	115	129	134					
Clus-2	134	125	142	141	131	151	164	139	141
Clus-3	144	143	122	134	126	157			
Clus-4	114	111	134	131	146	152	131		
Clus-5	119	126	122	129	130				
Clus-6	140	125	124	124	115	111	148	157	143 151

Answer:

#-----Defining Clusters-----

```

Clu1<-c(125,115,129,134)
Clu2<-c(134,125,142,141,131,151,164,139,141)
Clu3<-c(144,143,122,134,126,157)
Clu4<-c(114,111,134,131,146,152,131)
Clu5<-c(119,126,122,129,130)

```

```
Clu6<-c(140,125,124,124,115,111,148,157,143,151)
```

➤ **Population Mean and Sum Of Clusters**

```
#----Population----  
pop<-c(Clu1,Clu2,Clu3,Clu4,Clu5,Clu6)  
pop.mean=mean(pop)  
#----Sum of clusters----  
sumc1<-sum(Clu1)  
sumc2<-sum(Clu2)  
sumc3<-sum(Clu3)  
sumc4<-sum(Clu4)  
sumc5<-sum(Clu5)  
sumc6<-sum(Clu6)
```

➤ **Mean of Cluster Means**

```
125.7500  
140.8889  
137.6667  
131.2857  
125.2000  
133.8000
```

```
#----Grand total---- 132.431  
n=3;N=6;M=c(4,9,6,7,5,10);  
Yi=c(sumc1,sumc2,sumc3,sumc4,sumc5,sumc6)  
t.sum=sum(Yi)  
# ----- Mean of cluster means using population----  
n=3;N=6;M=c(4,9,6,7,5,10);  
clu.mean.p=Yi/M  
m.c.m.p=mean(clu.mean.p)  
##----For MSE----  
cov=cov(clu.mean.p,M)  
dv.p<-(clu.mean.p-m.c.m.p)^2  
sdv.p<-sum(dv.p)  
vr.p<-sdv.p/(N-1)  
term1<-((N-n)/(N*n))*vr.p  
term2=((-(N-1)/sum(M))*cov)^2  
mse=term1+term2
```

```
####-----From Sample----  
j=sample(1:6,n)  
yi=Yi[j];mi=M[j]  
clu.mean=yi/mi  
m.clu.mean=mean(clu.mean)
```

Q. Perform the Simulation Study using the following data for Equal Cluster Sizes.

						Total
Cluster-1	125	115	129	134	111	614
Cluster-2	134	125	142	141	131	673
Cluster-3	144	143	122	134	126	669
Cluster-4	114	111	134	131	146	636
Cluster-5	119	126	122	129	130	626
Cluster-6	140	125	124	124	115	628

Answer:**➤ How to Do This in R?**

#--Defining Clusters in R---

Clu1<-c(125,115,129,134,111)

Clu2<-c(134,125,142,141,131)

Clu3<-c(144,143,122,134,126)

Clu5<-c(119,126,122,129,130)

Clu6<-c(140,125,124,124,115)

➤ Sum Of Clusters

#---Population---

pop<-c(Clul,Clu2,Clu3,Clu4,Clu5,Clu6)

#---Sum of clusters---

sumc1<-sum(Clul)

sumc2<-sum(Clul2)

sumc3<-sum(Clul3)

sumc4<-sum(Clul4)

sumc5<-sum(Clul5)

sumc6<-sum(Clul6)

➤ Sampling and Estimated Mean

#---Grand total---

Yi=c(sumc1,sumc2,sumc3,sumc4,sumc5,sumc6)

t.sum=sum(Yi)

n=3;N=6;M=5;

for(i in 1:10000)

{ yi=sample(Yi,n)

Est.mean[i]=sum(yi)/(n*M) }

Q. Perform the Simulation Study using the following data for Equal Cluster Sizes.**Answer:****➤ Defining Clusters**

#-----Defining Clusters-----

Clu1<-c(125,115,129,134)

```

Clu2<-c(134,125,142,141,131,151,164,139,141)
Clu3<-c(144,143,122,134,126,157)
Clu4<-c(114,111,134,131,146,152,131)
Clu5<-c(119,126,122,129,130)
Clu6<-c(140,125,124,124,115,111,148,157,143,151)

```

➤ **Population Mean and Sum Of Clusters**

```

#---Population---
pop<-c(Clu1,Clu2,Clu3,Clu4,Clu5,Clu6)
pop.mean=mean(pop)
#---Sum of clusters---
sumc1<-sum(Clu1)
sumc2<-sum(Clu2)
sumc3<-sum(Clu3)
sumc4<-sum(Clu4)
sumc5<-sum(Clu5)
sumc6<-sum(Clu6)

```

➤ **Sum of Clusters**

```

#---Grand total---
n=3;N=6;M=c(4,9,6,7,5,10);
Yi=c(sumc1,sumc2,sumc3,sumc4,sumc5,sumc6)
t.sum=sum(Yi)

```

➤ **With Mean of Cluster Means**

```

for(i in 1:10000)
{
j=sample(1:6,n)
yi=Yi[j];mi=M[j]
clu.mean=yi/mi
m.clu.mean[i]=mean(clu.mean) }
mean(m.clu.mean)
var(m.clu.mean)

```

➤ **Weighted Mean**

Clus-3	144	143	122	134	126	157	
Clus-4	114	111	134	131	146	152	131
Clus-5	119	126	122	129	130		

$$\bar{\bar{y}}_w = \frac{\sum_{i=1}^n M_i \bar{y}_i}{n\bar{M}}$$

➤ **Population Mean and Sum Of Clusters**

```
#----Sum of clusters----
mc1<-mean(Clu1)
mc2<-mean(Clu2)
mc3<-mean(Clu3)
mc4<-mean(Clu4)
mc5<-mean(Clu5)
mc6<-mean(Clu6)
mYi=c(mc1,mc2,mc3,mc4,mc5,mc6)
```

➤ **With Weighted Mean**

```
w.m=c(); Mbar=mean(M);
for(i in 1:10000)
{ j=sample(1:6,n)
myi=mYi[j];mi=M[j];
w.m[i]=sum(mi*myi)/(n*Mbar) }
mean(w.m)
var(w.m)
```

Unequal Probability Sampling

Introduction:

- Hansen and Hurwitz (1943) perhaps the first who discussed the concept of unequal probability theory.
- The most commonly used scheme is probability proportional to size (PPS).
- Sampling with respect to department size.
- Sampling with respect to number of trees.

	Depart-ment	Faculty size		
	1	32		
1-32	2	10		
33-42	3	21		
43-63	4	16		
64-79	5	11		
80-90	Total	90		

➤ PPS With Replacement

- The estimator for population total Y as suggested by Hansen and Hurwitz (1943) is

$$y'_{HH} = \frac{1}{n} \sum_{i=1}^n \frac{y_i}{p_i},$$

➤ Computing HH Estimator

Depart-ment	Faculty size	Pi	Numbers
1	332	32/90	1-32
2	110	10/90	33-42
3	221	21/90	43-63
4	116	16/90	64-79
5	11	11/90	80-90
Total	90		

$$y'_{HH} = \frac{1}{n} \sum_{i=1}^n \frac{y_i}{p_i},$$

Q. Prove that Hansen-Hurwitz Estimator is unbiased to population total. Also find the variance expression.

$$y'_{HH} = \frac{1}{n} \sum_{i=1}^n \frac{y_i}{p_i},$$

Answer:

The estimator for population total Y as suggested by Hansen and Hurwitz (1943) is

$$y'_{HH} = \frac{1}{n} \sum_{i=1}^n \frac{y_i}{p_i},$$

$$E(y'_{HH}) = \frac{1}{n} \sum_{i=1}^n E\left(\frac{y_i}{p_i}\right)$$

$$E\left(\frac{y_i}{p_i}\right) = \sum_{i=1}^N \frac{Y_i}{P_i} P_i = Y$$

Variance of Hansen-Hurwitz Estimator

$$y'_{HH} = \frac{1}{n} \sum_{i=1}^n \frac{y_i}{p_i},$$

$$V(y'_{HH}) = \frac{1}{n} \sum_{i=1}^N p_i \left(\frac{y_i}{p_i} - Y \right)^2 \qquad V\left(\frac{y_i}{p_i}\right) = \sum_{i=1}^N p_i \left(\frac{y_i}{p_i} - Y \right)^2$$

➤ **Another Form of Variance**

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

Q. Find all possible samples of size two using the following data and find Hansen-Hurwitz Estimator. Also find mean and variance of Hansen-Hurwitz Estimator.

Y_i	.5	.2	.1	.2
Z_i	1	2	3	4

Answer:

- The following is a population with four values with respective size.

Y_i	.5	.2	.1	.2
Z_i	1	2	3	4

- We are interested to take all possible samples of size 2.

- HH estimator will be calculated for all the samples. Further mean and variance will be obtained.

Y_i	Z_i	$P_i = Z_i / \sum Z_i$
0.5	1	0.1
1.2	2	0.2
2.1	3	0.3
3.2	4	0.4

➤ All Possible Samples

Small Population Example
Small Population Example
Small Population Example

Possible Samples	Y_i	Y_j	P_i	P_j	Y'_{HH}	$Y'_{HH} P_i P_j$	$Y'^2_{HH} P_i P_j$
1 1	0.5	0.5	.1	.1	5.0	0.05	0.250
1 2	0.5	1.2	.1	.2	5.5	0.11	0.605
1 3	0.5	2.1	.1	.3	6.0	0.18	1.080
1 4	0.5	3.2	.1	.4	6.5	0.26	1.690
2 1	1.2	0.5	.2	.1	5.5	0.11	0.605
2 2	1.2	1.2	.2	.2	6.0	0.24	1.440
2 3	1.2	2.1	.2	.3	6.5	0.39	2.535
2 4	1.2	3.2	.2	.4	7.0	0.56	3.920
3 1	2.1	0.5	.3	.1	6.0	0.18	1.080
3 2	2.1	1.2	.3	.2	6.5	0.39	2.535
3 3	2.1	2.1	.3	.3	7.0	0.63	4.410
3 4	2.1	3.2	.3	.4	7.5	0.90	6.750
4 1	3.2	0.5	.4	.1	6.5	0.26	1.690
4 2	3.2	1.2	.4	.2	7.0	0.56	3.920
4 3	3.2	2.1	.4	.3	7.5	0.90	6.750
4 4	3.2	3.2	.4	.4	8.0	1.28	10.240
						7.00	49.500

➤ Small Population Example

- The expected value of HH estimator is

$$E(y'_{HH}) = \sum y'_{HH} p_i p_j = 7 = Y$$

$$Var(y'_{PPS}) = E(y'^2_{PPS}) - Y^2$$

$$Var(y'_{PPS}) = 49.5 - 49 = 0.5$$

➤ **Using the Formula**

$$Var(y'_{PPS}) = \frac{1}{n} \left[\sum_{i=1}^N \frac{Y_i^2}{P_i} - Y^2 \right] = \frac{1}{2} [50 - 49] = 0.50$$

Y _i	Z _i	$P_i = Z_i / \sum Z_i$
0.5	1	0.1
1.2	2	0.2
2.1	3	0.3
3.2	4	0.4

Q. Describe the Lahiri's Method of selection?

Answer:

➤ **Lahiri's Method of Selection**

- A pair of random numbers is chosen such that one from 1 to N and other 1 to Z_{max} (say R)

$$\circ P_i(ith) = \frac{1}{N} [Z_i / Z_{\max}]$$

- if R exceeds the size of the ith unit; then that unit is rejected otherwise it is accepted.
- probability of selecting the ith unit at the first draw.

Sr.No	Y _i	Z _i
1	0.5	15
2	1.2	20
3	2.1	7

4 3.2 13

Q. Example for Lahiri's Method

S.No. of Village	No. of Trees Y_i	Area in Acres Z_i	P_i	S.No. of Village	No. of Trees Y_i	Area in Acres Z_i	P_i
1	2328	328	0.700	11	2483	537	0.120
2	754	80	0.020	12	128	7	0.002
3	105	6	0.001	13	102	8	0.002
4	949	156	0.030	14	60	22	0.005
5	3091	428	0.100	15	0	4	0.001
6	1736	401	0.090	16	11799	1234	0.280
7	840	94	0.020	17	26	3	0.001
8	311	63	0.010	18	317	30	0.010
9	0	51	0.010	19	190	40	0.010
10	3044	946	0.210	20	180	20	0.004

➤ **Selection with Lahiri's Method**

Pair of Random Numbers	No. of Village	Y_i	P_i	$\frac{Y_i}{P_i}$
(16,37)	16	11799	0.28	42139.286
(5,38)	5	3091	0.10	30910.00
(16,23)	16	11799	0.28	42139.286
(10,54)	10	3044	0.21	14495.238
(10,54)	10	3044	0.21	14495.238
				144179.05

➤ **The Estimated Total**

- The estimated value for population total using Lahiri's method of selection

$$y'_{PPS} = \frac{1}{n} \sum_{i=1}^5 \frac{y_i}{p_i} = \frac{144179.05}{5} = 28836$$

Q. Explain the concept of Unequal Probability Sampling Without Replacement?

Answer:

- let a sample of two units is selected from a population of N units.

- Let the probability of the selection of the i th unit is $P_i = Z_i/Z$.
- Suppose the i th unit is not selected at the first draw but the j th unit is selected ($j \neq i$) then the probability of selecting the j th unit at the first draw is $P_j = Z_j/Z$;
- The conditional probability of selecting the i th unit at the second draw is

$$\frac{Z_i}{(Z - Z_j)} \quad P_i / (1 - P_j)$$

- The probability of inclusion of i th unit at the second draw to be included in the sample is the sum of the product that the j th unit is selected at the first draw and the i th unit is selected at the second draw given the j th unit is selected at the first draw i.e.

$$\sum_{j \neq i}^N P_j \frac{P_i}{1 - P_j}$$

- The total probability π_i , the probability of inclusion of the i th population unit to be in the sample is

$$\begin{aligned} \pi_i &= P_i + \sum_{j \neq i}^N P_j \frac{P_i}{1 - P_j} \\ &= P_i \left[1 + \sum_{j \neq i}^N \frac{P_j}{1 - P_j} - \frac{P_i}{1 - P_i} \right] \end{aligned}$$

- The probability that both i th and j th units are in the sample is denoted by π_{ij} and is defined as

$$\begin{aligned} \pi_{ij} &= P_i P_{j|i} + P_j P_{i|j} = P_i \frac{P_j}{1 - P_i} + P_j \frac{P_i}{1 - P_j} \\ &= P_i P_j \left[\frac{1}{1 - P_i} + \frac{1}{1 - P_j} \right] \end{aligned}$$

Q. The Horvitz Thompson Estimator?

Answer:

➤ HT Estimator

- The general theory of unequal probabilities sampling without replacement was presented firstly by Horvitz and Thompson (1952).
- An unbiased estimator suggested by them for population total Y is

$$y'_{HT} = \sum_{i=1}^n \frac{y_i}{\pi_i} = \sum_{i=1}^N a_i \frac{Y_i}{\pi_i},$$

Q. Prove that HT Estimator is unbiased estimator.

➤ The Horvitz and Thompson in 1952 suggested the estimator total is

$$y'_{HT} = \sum_{i=1}^n \frac{y_i}{\pi_i} = \sum_{i=1}^N a_i \frac{Y_i}{\pi_i},$$

$$E(y'_{HT}) = \sum_{i=1}^N \pi_i \frac{Y_i}{\pi_i} = \sum_{i=1}^N Y_i = Y$$

Q.81 what are the Some Relations of π_i ?

Answer:

➤ **Relations of π_i**

- The following are some relations of

i. $\sum_{i=1}^N \pi_i = n$

ii. $\sum_{j \neq i}^N \pi_{ij} = (n-1) \pi_i$

iii. $\sum_{j \neq i}^N (\pi_{ij} - \pi_i \pi_j) = -\pi_i (1 - \pi_i)$

iv. $\sum_{i=1}^N \sum_{j \neq i}^N \pi_i \pi_j = n \sum_{i=1}^N \pi_i - \sum_{i=1}^N \pi_i^2$

➤ **Relation (i)**

- We know $\sum_{i=1}^N a_i = n$

- Taking expectation $\sum_{i=1}^N \pi_i = n$

➤ **Relation (ii)**

$$\sum_{j \neq i}^N \pi_{ij} = (n-1) \pi_i$$

- is sum of all the probabilities of the samples containing ith and jth units

- is the sum of the probabilities containing first and second units; first and third units; and so on.
- Thus every $P(s)$ containing the first unit occurs $(n-1)$ times in this sum as the sample has $(n-1)$ other members in it and it occurs once for each of these members.

$$\sum_{j \neq i}^N \pi_{ij}, \sum_{j \neq i}^N \pi_{ij} = (n-1)\pi_i$$

➤ **Relation (iii)**

$$\sum_{j \neq i}^N (\pi_{ij} - \pi_i \pi_j) = -\pi_i(1 - \pi_i)$$

Taking L.H.S

$$\sum_{j \neq i}^N (\pi_{ij} - \pi_i \pi_j) = \sum_{j \neq i}^N \pi_{ij} - \pi_i \sum_{j \neq i}^N \pi_j$$

$$\sum_{j \neq i}^N \pi_{ij} - \pi_i \left(\sum_{j=1}^N \pi_j - \pi_i \right)$$

$$\sum_{j \neq i}^N \pi_{ij} - \pi_i \left(\sum_{j=1}^N \pi_j - \pi_i \right)$$

Using (i) and (ii) relation

$$\begin{aligned} &= (n-1)\pi_i - \pi_i(n - \pi_i) \\ &= -\pi_i(1 - \pi_i) \end{aligned}$$

➤ **Relation (iv)**

$$\sum_{i=1}^N \sum_{j \neq i}^N \pi_i \pi_j = n \sum_{i=1}^N \pi_i - \sum_{i=1}^N \pi_i^2$$

$$\sum_{i=1}^N \sum_{j \neq i}^N \pi_i \pi_j = \sum_{i=1}^N \pi_i \sum_{j \neq i}^N \pi_j$$

Q.82 Find the variance expression of Horvitz Thompson estimator?

Answer:

➤ **HT Estimator**

- The Horvitz and Thompson in 1952 suggested the estimator total is

$$y'_{HT} = \sum_{i=1}^n \frac{y_i}{\pi_i} = \sum_{i=1}^N a_i \frac{Y_i}{\pi_i},$$

$$E(y'_{HT}) = \sum_{i=1}^N \pi_i \frac{Y_i}{\pi_i} = \sum_{i=1}^N Y_i = Y$$

➤ **Variance of HT Estimator**

$$Var(y'_{HT}) = E(y'_{HT})^2 - [E(y'_{HT})]^2$$

$$\begin{aligned}
 y'_{HT} &= \sum_{i=1}^n \frac{y_i}{\pi_i} = \sum_{i=1}^N a_i \frac{Y_i}{\pi_i}, \\
 \text{Var}(y'_{HT}) &= E \left(\sum_{i=1}^N a_i \frac{Y_i}{\pi_i} \right)^2 \\
 &\quad - \left[E \left(\sum_{i=1}^N a_i \frac{Y_i}{\pi_i} \right) \right]^2 \\
 \text{Var}(y'_{HT}) &= E \left(\sum_{i=1}^N a_i \frac{Y_i}{\pi_i} \right)^2 \\
 &\quad - \left[\left(\sum_{i=1}^N E(a_i) \frac{Y_i}{\pi_i} \right) \right]^2 \\
 E \left(\sum_{i=1}^N a_i \frac{Y_i}{\pi_i} \right)^2 &= \left[\sum_{i=1}^N E(a_i^2) \frac{Y_i^2}{\pi_i^2} \right] \\
 &\quad + \left[\sum_{\substack{i,j \\ j \neq i}}^N \sum E(a_i a_j) \frac{Y_i}{\pi_i} \frac{Y_j}{\pi_j} \right] \\
 \text{Var}(y'_{HT}) &= E \left(\sum_{i=1}^N a_i \frac{Y_i}{\pi_i} \right)^2 \\
 &\quad - \left[\left(\sum_{i=1}^N E(a_i) \frac{Y_i}{\pi_i} \right) \right]^2 \\
 &\quad \left[\sum_{i=1}^N E(a_i^2) \frac{Y_i^2}{\pi_i^2} \right] + \left[\sum_{\substack{i,j \\ j \neq i}}^N \sum E(a_i a_j) \frac{Y_i}{\pi_i} \frac{Y_j}{\pi_j} \right] \\
 &\quad - \left[\sum_{i=1}^N [E(a_i)]^2 \frac{Y_i^2}{\pi_i^2} \right] - \sum_{\substack{i,j \\ j \neq i}}^N \sum E(a_i) E(a_j) \frac{Y_i}{\pi_i} \frac{Y_j}{\pi_j} \\
 &= \sum_{i=1}^N \frac{Y_i^2}{\pi_i^2} \left[E(a_i^2) - (E(a_i))^2 \right] \\
 &\quad - \sum_{\substack{i,j=1 \\ j \neq i}}^N \sum \frac{Y_i Y_j}{\pi_i \pi_j} \left[E(a_i a_j) - E(a_i) E(a_j) \right] \\
 &= \sum_{i=1}^N \frac{Y_i^2}{\pi_i^2} \left[E(a_i^2) - (E(a_i))^2 \right] \\
 &\quad - \sum_{\substack{i,j=1 \\ j \neq i}}^N \sum \frac{Y_i Y_j}{\pi_i \pi_j} \left[E(a_i a_j) - E(a_i) E(a_j) \right]
 \end{aligned}$$

$$= \sum_{i=1}^N \frac{Y_i^2}{\pi_i^2} \left[E(a_i^2) - (E(a_i))^2 \right] \\ - \sum_{\substack{i,j=1 \\ j \neq i}}^N \frac{Y_i Y_j}{\pi_i \pi_j} \left[E(a_i a_j) - E(a_i) E(a_j) \right]$$

$$\text{Var}(a_i) = \pi_i (1 - \pi_i)$$

$$\text{Cov}(a_i, a_j) = \pi_{ij} - \pi_i \pi_j$$

$$\text{Var}(y'_{HT}) = \sum_{i=1}^N \frac{Y_i^2}{\pi_i^2} \text{Var}(a_i) + \\ \sum_{\substack{i,j=1 \\ j \neq i}}^N \frac{Y_i Y_j}{\pi_i \pi_j} \text{Cov}(a_i, a_j)$$

$$\text{Var}(y'_{HT}) = \sum_{i=1}^N (1 - \pi_i) \pi_i \frac{Y_i^2}{\pi_i^2} \\ + \sum_{\substack{i,j=1 \\ j \neq i}}^N (\pi_{ij} - \pi_i \pi_j) \frac{Y_i Y_j}{\pi_i \pi_j}$$

$$\text{Var}(y'_{HT}) = \sum_{i=1}^N (1 - \pi_i) \pi_i \frac{Y_i^2}{\pi_i^2} \\ + \sum_{\substack{i,j=1 \\ j \neq i}}^N (\pi_{ij} - \pi_i \pi_j) \frac{Y_i Y_j}{\pi_i \pi_j}$$

$$\sum_{j \neq i} (\pi_{ij} - \pi_i \pi_j) = -\pi_i (1 - \pi_i)$$

$$\text{Var}(y'_{HT}) = \sum_{i=1}^N (1 - \pi_i) \pi_i \frac{Y_i^2}{\pi_i^2} \\ + \sum_{\substack{i,j=1 \\ j \neq i}}^N (\pi_{ij} - \pi_i \pi_j) \frac{Y_i Y_j}{\pi_i \pi_j}$$

$$\text{Var}(y'_{HT}) = \sum_{\substack{i=1 \\ j \neq i}}^N \sum (\pi_i \pi_j - \pi_{ij}) \frac{Y_i^2}{\pi_i^2} \\ + \sum_{\substack{i,j=1 \\ j \neq i}}^N \frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j} Y_i Y_j$$

$$\begin{aligned}
 Var(y'_{HT}) &= \sum_{i=1}^N \sum_{j \neq i} (\pi_i \pi_j - \pi_{ij}) \frac{Y_i^2}{\pi_i^2} \\
 &+ \sum_{i,j=1}^N \sum_{j \neq i} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j} Y_i Y_j \\
 &\frac{1}{2} \sum_{i=1}^N \sum_{j \neq i} (\pi_i \pi_j - \pi_{ij}) \left(\frac{Y_i^2}{\pi_i^2} + \frac{Y_j^2}{\pi_j^2} \right) \\
 &- \sum_{i,j=1}^N \sum_{j \neq i} \frac{\pi_i \pi_j - \pi_{ij}}{\pi_i \pi_j} Y_i Y_j \\
 &\frac{1}{2} \sum_{i=1}^N \sum_{j \neq i} (\pi_i \pi_j - \pi_{ij}) \left(\frac{Y_i^2}{\pi_i^2} + \frac{Y_j^2}{\pi_j^2} \right) \\
 &- \sum_{i,j=1}^N \sum_{j \neq i} \frac{\pi_i \pi_j - \pi_{ij}}{\pi_i \pi_j} Y_i Y_j \\
 &\frac{1}{2} \sum_{i=1}^N \sum_{j \neq i} (\pi_i \pi_j - \pi_{ij}) \\
 &\left(\frac{Y_i^2}{\pi_i^2} + \frac{Y_j^2}{\pi_j^2} - 2 \frac{Y_i Y_j}{\pi_i \pi_j} \right) \\
 Var_{SYG}(y'_{HT}) &= \frac{1}{2} \sum_{i,j}^N \sum_{j \neq i} (\pi_i \pi_j - \pi_{ij}) \left(\frac{Y_i}{\pi_i} - \frac{Y_j}{\pi_j} \right)^2
 \end{aligned}$$

Q. Calculate HT estimator by taking all possible samples of size 2 without replacement. Further find mean and variance.

Ans:

Small Population Example

- The following is a population with four values with respective size.

Unit No.	1	2	3	4
Y_i	0.5	1.2	2.1	3.2
Z_i	1	2	3	4
P_i	0.1	0.2	0.3	0.4

- We are interested to take all possible samples of size 2.

- HT estimator will be calculated for all the samples. Further mean and variance will be obtained.

Unit No.	1	2	3	4
Y_i	0.5	1.2	2.1	3.2
Z_i	1	2	3	4
P_i	0.1	0.2	0.3	0.4

$$y'_{HT} = \sum_{i=1}^n \frac{y_i}{\pi_i}$$

$$P_i = Z_i / \sum Z_i$$

$$y'_{HT} = \frac{y_1}{\pi_1} + \frac{y_2}{\pi_2}$$

$$\pi_i = P_i \left[1 + \sum_{j=1}^N \frac{P_j}{1-P_j} - \frac{P_i}{1-P_i} \right]$$

$$\pi_1 = P_1 \left[1 + \sum_{j=1}^N \frac{P_j}{1-P_j} - \frac{P_1}{1-P_1} \right]$$

$$\sum \frac{P_j}{1-P_j} = 1.456$$

$$\pi_1 = 0.1(1 + 1.456 - 0.111)$$

$$= 0.2345$$

Unit No.	1	2	3	4
Y_i	0.5	1.2	2.1	3.2
Z_i	1	2	3	4
P_i	0.1	0.2	0.3	0.4
π_i	0.2345	0.4413	0.6083	0.7159

$$\sum \frac{P_j}{1-P_j} = 1.456$$

$$\pi_1 = 0.1(1 + 1.456 - 0.111)$$

$$= 0.2345$$

$$\pi_{ij} = P_i P_j \left[\frac{1}{1-P_i} + \frac{1}{1-P_j} \right]$$

$$\pi_{ij} = 0.1 * 0.2 \left[\frac{1}{0.9} + \frac{1}{0.8} \right]$$

$$\pi_{ij} = 0.1 * 0.2 [1.111 + 1.25]$$

$$= 0.0472$$

$$y'_{HT} = \sum_{i=1}^n \frac{y_i}{\pi_i}$$

$$E(y'_{HT}) = 7.0000$$

Unit	y _i	y _j	π_i	π_j	π_{ij}	y'_{HT}	$\pi_{ij} y'_{HT}$	$\pi_{ij} (y'_{HT})^2$
1, 2	0.5	1.2	0.2345	0.4413	0.0472	4.8514	0.2291	1.1114
1, 3	0.5	2.1	0.2345	0.6083	0.0762	5.5840	0.4255	2.3757
1, 4	0.5	3.2	0.2345	0.7159	0.1111	6.6020	0.7336	4.8430
2, 3	1.2	2.1	0.4413	0.6083	0.1607	6.1715	0.9918	6.1212
2, 4	1.2	3.2	0.4413	0.7159	0.2333	7.1895	1.6775	12.0607
3, 4	2.1	3.2	0.6083	0.7159	0.3714	7.9221	2.9425	23.3109
					1.00		7.00	49.823

$$Var(y'_{HT}) = 49.8229 - (7.0000)^2$$

$$= 0.8229.$$

Q.85 what are the HT Estimator for with Replacement Sampling?

Answer:

➤ **Sample of Four Units**

Unit No.	1	2	3	4
Y _i	60	60	14	1
P _i	0.05	0.05	0.02	0.01

- Y_i is the count of animals from the sample of four strips.

➤ **HH Estimator**

Unit No.	1	2	3	4
----------	---	---	---	---

Y_i	60	60	14	1
P_i	0.05	0.05	0.02	0.01

$$= \frac{1}{4} \left(\frac{60}{0.05} + \frac{60}{0.05} + \frac{14}{0.02} + \frac{1}{0.01} \right)$$

$$= 800$$

➤ **HT Estimator**

Unit No.	1	2	3	4
Y_i	60	60	14	1
P_i	0.05	0.05	0.02	0.01

$$\pi_i = 1 - (1 - p_i)^n$$

$$y'_{HT} = \sum_{i=1}^n \frac{y_i}{\pi_i}$$

$$P(A \cap B) = P(A) + P(B) - P(A \cup B)$$

$$= P(A) + P(B) - [1 - P(A^c \cap B^c)]$$

$$\pi_{ij} = \pi_i + \pi_j - [1 - (1 - p_i - p_j)^n]$$

Q. Perform PPS Sampling With Replacement using the following data in R?

Y_i 0.5 1.2 2.1 3.2

Z_i 1 2 3 4

Answer:

```
y<-c(0.5,1.2,2.1,3.2)
zi<-c(1,2,3,4)
z<-sum(zi)
pi<-zi/z
N<-4;n=2;means=c();
#----With PPS-----
for(i in 1:10)
{ s <- sample(1:N,n,replace=TRUE,prob=pi)
means[i]<-mean(y[s]) }
```

mean(means)

Q. Select 10000 samples with PPS Sampling With Replacement using the following data in R. Find mean of each sample. Further find mean of means

Y_i	0.5	1.2	2.1	3.2
Z_i	1	2	3	4

```
y<-c(0.5,1.2,2.1,3.2)
zi<-c(1,2,3,4)
z<-sum(zi)
pi<-zi/z
N<-4;n=2
#----With PPS-----
for(i in 1:10000)
{ s <- sample(1:N,n,replace=TRUE,prob=pi)
means[i]<-mean(y[s]) }
mean(means)
```

Q. Select 10000 samples with PPS Sampling With Replacement using the following data in R. Find HH estimator from each sample. Further find mean of HH estimators.

Y_i	0.5	1.2	2.1	3.2
Z_i	1	2	3	4

Answer:

```
y<-c(0.5,1.2,2.1,3.2)
zi<-c(1,2,3,4);z<-sum(zi);pi<-zi/z
N<-4;n=2
for(i in 1:10000)
{ s <- sample(1:N,n,replace=TRUE,prob=pi)
hh[i] <- mean(y[s]/pi[s])/N }
mean(hh)
```

Example:

Answer:

➤ **Population from R**

- Read in the trees data set from R
- let the variable of interest y be tree volume and draw by draw selection probability be proportional to girth.
- The variable of interest is tree volume.

```
# ----Reading the data----
y <- trees$Volume
zi <- trees$Girth
y
10.3 10.3 10.2 16.4 18.8 19.7 15.6 18.2 22.6 19.9 24.2 21.0 21.4 21.3 19.1 22.2 33.8 27.4
25.7 24.9 34.5 31.7 36.3 38.3 42.6 55.4 55.7 58.3 51.5 51.0 77.0
zi
8.3 8.6 8.8 10.5 10.7 10.8 11.0 11.0 11.1 11.2 11.3 11.4 11.4 11.7 12.0 12.9 12.9 13.3
13.7 13.8 14.0 14.2 14.5 16.0 16.3 17.3 17.5 17.9 18.0 18.0 20.6
```

➤ Defining Terms

```
# ----Defining Population----
y <- trees$Volume
zi <- trees$Girth
pi <- zi/sum(zi)
pi
0.02020940 0.02093986 0.02142683 0.02556611 0.02605308 0.02629657 0.02678354
0.02678354 0.02702703 0.02727051 0.02751400 0.02775749 0.02775749 0.02848795
0.02921841 0.03140979 0.03140979 0.03238374 0.03335768 0.03360117 0.03408814
0.03457512 0.03530558 0.0389578 0.03968834 0.04212320 0.04261018 0.04358412
0.04382761 0.04382761 0.05015827
sum(zi)
410.7
```

➤ Sampling With PPS

```
y <- trees$Volume
zi <- trees$Girth
pi <- zi/sum(zi)
N <- 31; n = 10;
#----Sampling with PPS----
s <- sample(1:N, n, replace=TRUE, prob=pi)
y[s]
mean(y[s])
mu <- mean(y)
```

➤ With SRS

```
y <- trees$Volume
zi <- trees$Girth
pi <- zi/sum(zi)
```

```
N<-31;n=10;
#----Sampling with PPS----
s <- sample(1:N,n,replace=TRUE,prob=NULL)
y[s]
mean(y[s])
mu <- mean(y)
> mean(y[s])
[1] 37.96
mu
[1] 30.17
```

➤ **Simulation with PPS**

```
y <- trees$Volume
zi <- trees$Girth
pi<-zi/sum(zi)
N<-31;n=10;
#----With PPS-----
for(i in 1:10000)
{ s <- sample(1:N,n,replace=TRUE,prob=pi)
means[i]<-mean(y[s]) }
mean(means)
mean
33.37
```

➤ **Simulation with SRSWR**

```
y <- trees$Volume
zi <- trees$Girth
pi<-zi/sum(zi)
N<-31;n=10;
#----SRSWR----
for(i in 1:10000)
{ s <- sample(1:N,n,replace=TRUE,prob=NULL)
means[i]<-mean(y[s]) }
mean(means)
mean
30.25
```

➤ **Observations**

- Mean of Population is 30.17
- Mean of means with SRS is 30.25-
- Mean of means with PPS is 33.37

Q. Read the data from the following command and select 10000 samples with PPS Sampling with replacement using the following data in R. Find HH estimator from each sample. Further find mean of HH estimators.

```
y <- trees$Volume
zi <- trees$Girth
```

Answer:

➤ **Population from R**

```
# -----Reading the data-----
y <- trees$Volume
zi <- trees$Girth
y
10.3 10.3 10.2 16.4 18.8 19.7 15.6 18.2 22.6 19.9 24.2 21.0 21.4 21.3 19.1 22.2 33.8 27.4 25.7
24.9 34.5 31.7 36.3 38.3 42.6 55.4 55.7 58.3 51.5 51.0 77.0
```

➤ **Population from R**

```
y
10.3 10.3 10.2 16.4 18.8 19.7 15.6 18.2 22.6 19.9 24.2 21.0 21.4 21.3 19.1 22.2 33.8 27.4 25.7
24.9 34.5 31.7 36.3 38.3 42.6 55.4 55.7 58.3 51.5 51.0 77.0
zi
8.3 8.6 8.8 10.5 10.7 10.8 11.0 11.0 11.1 11.2 11.3 11.4 11.4 11.7 12.0 12.9 12.9 13.3 13.7
13.8 14.0 14.2 14.5 16.0 16.3 17.3 17.5 17.9 18.0 18.0 20.6
```

```
y <- trees$Volume
zi <- trees$Girth
pi <- zi/sum(zi); N <- 31; n = 10;
for(i in 1:10000)
{ s <- sample(1:N, n, replace=TRUE, prob=pi)
  hh[i] <- mean(y[s]/pi[s])/N }
mean(hh)
mean(hh)
[1] 30.18311
```

Q. Select 10000 samples with PPS Sampling with replacement using the following data in R. Find HT estimator from each sample. Further find mean of HT estimators.

Y_i	0.5	1.2	2.1	3.2
Z_i	1	2	3	4

Answer:

```
# -----Defining Population-----
y <- c(0.5, 1.2, 2.1, 3.2)
zi <- c(1, 2, 3, 4)
z <- sum(zi)
pi <- zi/z; N = 4; n = 2
#---Calculation of Pi---
```

```
pii <- 1 - (1-pi)^n

#---Looping-----
for(i in 1:10000)
{ s <- sample(1:N,n,replace=TRUE,prob=pi)
  yu=unique(s)
  ht[i] <- sum(y[yu]/pii[yu])/N}
mean(ht)
```

Q. Select 10000 samples with PPS Sampling without replacement using the following data in R. Find HT estimator from each sample. Further find mean of HT estimators.

Y _i	0.5	1.2	2.1	3.2
Z _i	1	2	3	4

Answer:

```
# -----Defining Population-----
y<-c(0.5,1.2,2.1,3.2)
zi<-c(1,2,3,4)
z<-sum(zi)
pi<-zi/z;N=4;n=2
#---Calculation of Pi---
piinv<-1-pi
sm<-sum(pi/piinv)
pii <- pi*(1+sm-pi/(1-pi))
```

➤ **ith HT Estimator**

```
#---Looping-----
hht=c();
for(i in 1:10000)
{ s <- sample(1:N,n,replace=FALSE,prob=pi)
  hht[i] <- sum(y[s]/pii[s])/N}
mean(hht)
```

Q. Read the data from the following command and select 10000 samples with PPS Sampling with replacement using the following data in R. Find HT estimator from each sample. Further find mean of HT estimators.

```
y <- trees$Volume
zi <- trees$Girth
```

Answer:

```
# -----Reading the data-----
```

```

y <- trees$Volume
zi <- trees$Girth
y
10.3 10.3 10.2 16.4 18.8 19.7 15.6 18.2 22.6 19.9 24.2 21.0 21.4 21.3 19.122.2 33.8 27.4 25.7
24.9 34.5 31.7 36.3 38.3 42.6 55.4 55.7 58.3 51.5 51.0 77.0
y
10.3 10.3 10.2 16.4 18.8 19.7 15.6 18.2 22.6 19.9 24.2 21.0 21.4 21.3 19.122.2 33.8 27.4 25.7
24.9 34.5 31.7 36.3 38.3 42.6 55.4 55.7 58.3 51.5 51.0 77.0
zi
8.3 8.6 8.8 10.5 10.7 10.8 11.0 11.0 11.1 11.2 11.3 11.4 11.4 11.7 12.0 12.9 12.9 13.3 13.7
13.8 14.0 14.2 14.5 16.0 16.3 17.3 17.5 17.9 18.0 18.0 20.6
sum(zi)
410.7

```

```
# -----Defining Population-----
```

```

y <- trees$Volume
zi <- trees$Girth
pi<-zi/sum(zi)
pi
0.02020940 0.02093986 0.02142683 0.02556611 0.02605308 0.02629657 0.02678354
0.02678354 0.02702703 0.02727051 0.02751400 0.02775749 0.02775749 0.02848795
0.02921841 0.03140979 0.03140979 0.03238374 0.03335768 0.03360117 0.03408814
0.03457512 0.03530558 0.0389578 0.03968834 0.04212320 0.04261018 0.04358412
0.04382761 0.04382761 0.05015827

```

```

y <- trees$Volume
zi <- trees$Girth
pi<-zi/sum(zi)
N<-31;n=10;
#---Calculation of Pi---
pii <- 1 - (1-pi)^n
pii

```

➤ Values of pii

```

> pii
0.1846714 0.1907295 0.1947457 0.2281662 0.2320147 0.2339325 0.2377552
0.2377552 0.2396601 0.2415607 0.2434571 0.2453491 0.2453491 0.2509998
0.2566124 0.2732237 0.2732237 0.2804987 0.2877080 0.2895002 0.2930723
0.2966283 0.3019321 0.3279149 0.3330058 0.3497258 0.3530242 0.3595758
0.3612043 0.3612043 0.4022598

```

```

#---Looping-----
hht=c();
for(i in 1:10000)
{s <- sample(1:N,n,replace=TRUE,prob=pi)
yu=unique(s)
hht[i] <- sum(y[yu]/pii[yu])/N}

```

mean(hht)

Q. Describe the Brewer selection procedure

Answer:

➤ **Brewer's Procedure**

- Brewer (1963a) suggested a selection procedure. The procedure for a sample of size 2 is given as
- Select the first unit with probability proportional to

$$\frac{P_i(1 - P_i)}{(1 - 2P_i)}$$

- Select the second unit with probability proportional to size of remaining units.

Q. Describe the Durbin's Procedure

➤ **Durbin's Procedure**

- Durbin (1967) suggested a selection procedure. The procedure for a sample of size 2 is given as
- Select first unit with probability proportional to size.
- Select second unit with probability proportional to size of

$$P_j \left[\frac{1}{1 - 2P_i} + \frac{1}{1 - 2P_j} \right]$$

Q. Describe Shehbaz-Hanif-Samiuddin's Procedure

- They suggested the procedure in 2003.
- Select first unit with probability proportional to

$$\frac{P_i(1 - P_i)}{(1 - 2P_i)}$$

- Select second unit with probability proportional to size of

$$P_j \left[\frac{1}{1 - 2P_i} + \frac{1}{1 - 2P_j} \right]$$

Estimation with Auxiliary Variable

Introduction:

➤ Auxiliary Variable

- We have discussed the estimation of parameters on the basis of single variable.
- Supporting/supplementary variable may be used at design stage or estimation stage.
- Supporting variable is used to enhance the efficiency of estimation.
- The supporting variable must be correlated with main variable.
- The supplementary information usually referred to as benchmark variable or Auxiliary variable.
- Graunt (1662) was the first who used auxiliary information to estimate the population size of England.
- After Graunt(1662), Laplace was the first to introduce the use of auxiliary information for the estimation of population of France.

X = Cell Phones: 2016

Y = Cell Phones: 2017

X = Working hours

Y = Number of Items Processed

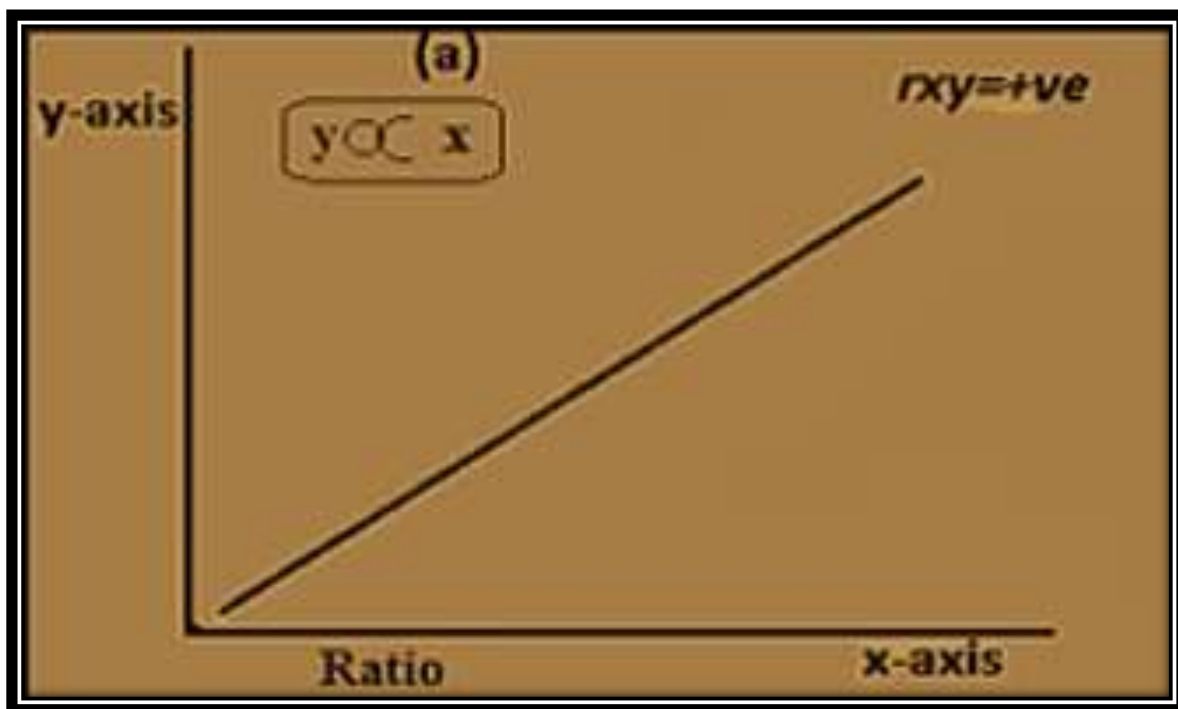
X = the size of unit i

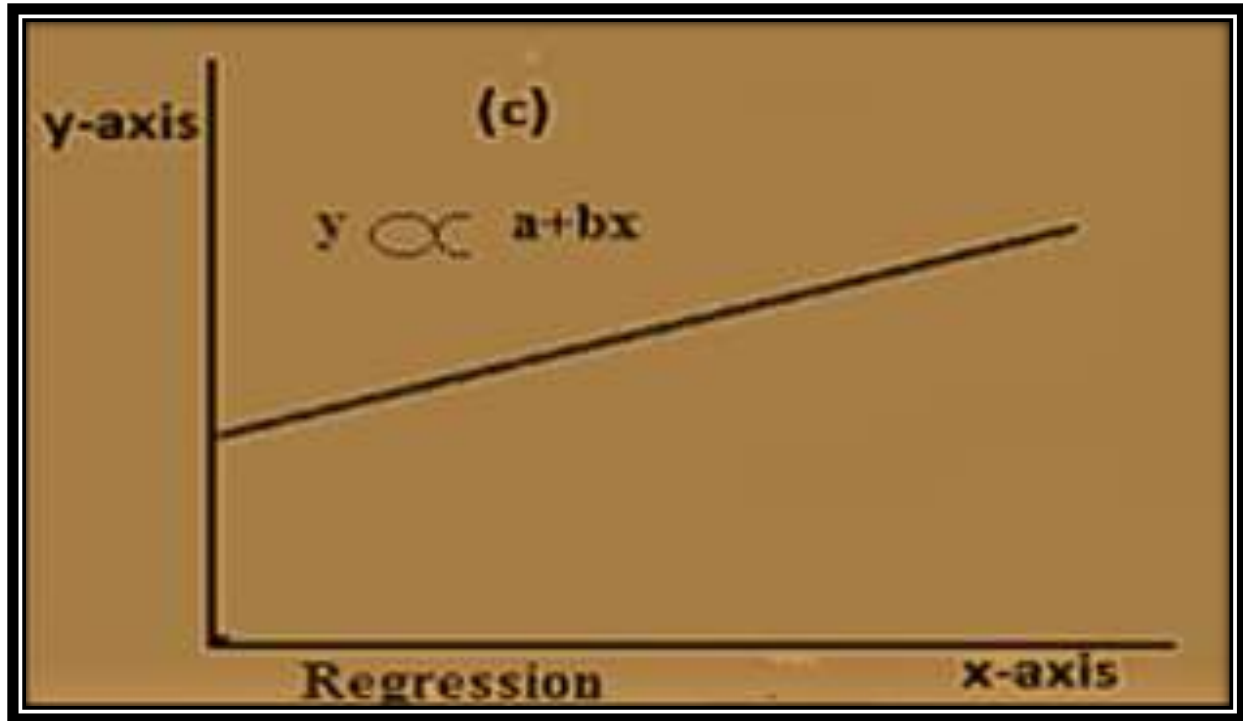
Y = the number of animals in unit i

X = Diameter of the tree

Y = Volume of the tree

➤ Some Types of Relationships





Q. Prove that ratio estimator is almost unbiased for large sample size.

Answer:

➤ **Ratio Estimator**

- The Ratio Estimator is

$$\hat{R} = \frac{y}{x} = \frac{\bar{y}}{\bar{x}}$$

➤ **Ratio Estimator for Mean and Total**

- The Ratio Estimator for population mean

$$\bar{y}_r = \frac{\bar{y}}{\bar{x}} \bar{X}$$

- The Ratio Estimator for population total

$$y_r = \frac{y}{x} X$$

➤ **Expectation of Ratio Estimator**

- For the large sample size the expectation of ratio estimator is approximately equal to population ratio, i.e.

$$E(\hat{R}) \approx R$$

$$\begin{aligned}\hat{R} - R &= \frac{\bar{y}}{\bar{x}} - R \\ &= \frac{\bar{y} - R\bar{x}}{\bar{x}}\end{aligned}$$

When n is large

$$\hat{R} - R \approx \frac{\bar{y} - R\bar{x}}{\bar{X}}$$

$$E(\hat{R} - R) \approx \frac{E(\bar{y}) - RE(\bar{x})}{\bar{X}}$$

$$E(\hat{R} - R) \approx \frac{E(\bar{y}) - RE(\bar{x})}{\bar{X}}$$

$$E(\hat{R} - R) \approx \frac{\bar{Y} - R\bar{X}}{\bar{X}}$$

$$E(\hat{R}) \approx R$$

Q. Find the variance of Ratio Estimator.

Answer:

➤ **Ratio Estimator**

- The Ratio Estimator is

$$\hat{R} = \frac{y}{x} = \frac{\bar{y}}{\bar{x}}$$

➤ **Ratio Estimator for Mean and Total**

- The Ratio Estimator for population mean

$$\bar{y}_r = \frac{\bar{y}}{\bar{x}} \bar{X}$$

- The Ratio Estimator for population total

$$y_r = \frac{y}{x} X$$

➤ **Variance of Ratio Estimator**

$$\hat{R} - R \approx \frac{\bar{y} - R\bar{x}}{\bar{X}}$$

$$E(\hat{R} - R)^2 \approx \frac{1}{\bar{X}^2} E(\bar{y} - R\bar{x})^2$$

$$\bar{d} = \bar{y} - R\bar{x}$$

$$\bar{D} = \bar{Y} - R\bar{X}$$

$$E(\hat{R} - R)^2 \approx \frac{1}{\bar{X}^2} V(\bar{d})$$

$$MSE(\hat{R}) \approx \frac{N-n}{Nn} \frac{1}{\bar{X}^2} \left[\frac{1}{N-1} \sum_{i=1}^N (Y_i - RX_i)^2 \right]$$

$$mse(\hat{R}) \approx \frac{N-n}{Nn} \frac{1}{\bar{X}^2} \left[\frac{1}{n-1} \sum_{i=1}^n (y_i - rx_i)^2 \right]$$

➤ **Alternative Expression of Variance**

$$MSE(\hat{R}) \approx \frac{N-n}{Nn} \frac{1}{\bar{X}^2} \left[\frac{1}{N-1} \sum_{i=1}^N (Y_i - RX_i)^2 \right]$$

$$SE(\hat{R}) \approx \frac{N-n}{Nn} \frac{1}{\bar{X}^2} \frac{1}{N-1} \sum_{i=1}^N [(Y_i - \bar{Y}) - (RX_i - R\bar{X})]^2$$

$$MSE(\hat{R}) \approx \frac{N-n}{Nn} \frac{1}{\bar{X}^2} [S_Y^2 - 2RS_{YX} + R^2S_X^2]$$

➤ **Expression of Variance for Ratio Estimator of Mean**

$$MSE(\hat{R}) \approx \frac{N-n}{Nn} \frac{1}{\bar{X}^2} [S_Y^2 - 2RS_{YX} + R^2S_X^2]$$

$$MSE(\bar{y}_r) \approx \frac{N-n}{Nn} [S_Y^2 - 2R\rho S_Y S_X + R^2S_X^2]$$

$$MSE(\bar{y}_r) \approx \lambda \bar{Y}^2$$

$$[C_Y^2 - 2\rho C_Y C_X + C_X^2]$$

$$C.L.(\hat{R}) : \hat{R} \pm t \sqrt{mse(\hat{R})}$$

$$C.L.(\bar{y}) : \bar{y}_r \pm t \sqrt{mse(\bar{y}_r)}$$

Q. Find the approximate Bias expression of Ratio Estimator?

Answer:

➤ **Ratio Estimator of Mean**

- The Ratio Estimator is

$$\hat{R} = \frac{y}{x} = \frac{\bar{y}}{\bar{x}}$$

➤ **Notations**

$$e_0 = (\bar{y} - \bar{Y}) / \bar{Y}$$

$$e_1 = (\bar{x} - \bar{X}) / \bar{X}$$

Using these notations

$$E(e_0^2) = \lambda C_y^2,$$

$$E(e_1^2) = \lambda C_x^2,$$

$$E(e_0 e_1) = \lambda C_{yx},$$

$$\text{where, } C_{yx} = \rho_{yx} C_y C_x$$

➤ **Using Notations**

$$\bar{y} = \bar{Y}(1 + e_0)$$

$$\bar{x} = \bar{X}(1 + e_1)$$

$$\bar{y}_r = \frac{\bar{y}}{\bar{x}} \bar{X}$$

$$\bar{y}_r = \frac{\bar{Y}(1 + e_0)}{\bar{X}(1 + e_1)} \bar{X}$$

➤ **Bias of Ratio Estimator**

$$= \bar{Y} (1 + e_0) (1 + e_1)^{-1}$$

$$\approx \bar{Y} (1 + e_0) (1 - e_1 + e_1^2)$$

$$\bar{y}_r - \bar{Y} \approx \bar{Y} (e_0 - e_1 - e_0 e_1 + e_1^2)$$

$$\text{Bias}(\bar{y}_r) \approx \frac{1-f}{n} \bar{Y} (C_x^2 - \rho_{yx} C_y C_x)$$

$$\text{Bias}(\bar{y}_r) \approx \frac{1-f}{n} \bar{Y} (C_x^2 - \rho_{yx} C_y C_x)$$

$$\text{Bias}(\hat{R}) \approx \frac{1-f}{n} R (C_x^2 - \rho_{yx} C_y C_x)$$

Q. Comparison of Ratio Estimator with Mean Per Unit Estimator?

Answer:

➤ **Ratio Estimator for Mean**

- The Ratio Estimator for population mean, i.e.

$$\bar{y}_r = \frac{\bar{y}}{\bar{x}} \bar{X}$$

$$\text{Var}(\bar{y}_r) \approx \bar{Y}^2 \lambda$$

$$[C_y^2 - 2\rho C_y C_x + C_x^2]$$

- The mean per unit estimator is

$$V(\bar{y}) = \lambda \bar{Y}^2 C_y^2$$

➤ **Comparison**

$$\begin{aligned}
 &Var(\bar{y}) - Var(\bar{y}_r) > 0 \\
 &\lambda \bar{Y}^2 C_Y^2 - \\
 &\lambda \bar{Y}^2 [C_Y^2 - 2\rho C_Y C_X + C_X^2] > 0 \\
 &C_Y^2 - [C_Y^2 - 2\rho C_Y C_X + C_X^2] > 0 \\
 &2\rho C_Y C_X - C_X^2 > 0 \\
 &\rho > \frac{1}{2} \frac{C_X}{C_Y}
 \end{aligned}$$

Q. Prove that Hartley-Ross is an Unbiased Ratio Estimator

$$r_{HR} = \bar{r} + \frac{n(N-1)}{N(n-1)} \frac{(\bar{y} - \bar{r} \bar{x})}{\bar{X}}$$

Answer:

➤ **Hartley-Ross Estimator**

- The HR Estimator is

$$r_{HR} = \bar{r} + \frac{n(N-1)}{N(n-1)} \frac{(\bar{y} - \bar{r} \bar{x})}{\bar{X}}$$

$$\text{where, } \bar{r} = \frac{1}{n} \sum_{i=1}^n \frac{y_i}{x_i}$$

- The variance of HR Estimator is

$$\begin{aligned}
 Var(r_{HR}) &\approx \frac{1}{n \bar{X}^2} \\
 &\left[S_Y^2 + \bar{R}^2 S_X^2 - 2\bar{R} PS_Y S_X \right]
 \end{aligned}$$

- We know that

$$\begin{aligned}
 E(\bar{r}) &= E\left(\sum_{i=1}^n \frac{r_i}{n}\right) = E(r_i) \\
 E(\bar{r}) &= \frac{E(r_i)E(x_i)}{E(x_i)} \\
 &= \frac{E(r_i)E(x_i)}{\bar{X}} \\
 E(\bar{r}) - R &= \frac{E(r_i)E(x_i)}{\bar{X}} - R \\
 E(\bar{r}) - R &= \\
 &= -\frac{1}{\bar{X}} [\bar{Y} - E(r_i)E(x_i)]
 \end{aligned}$$

$$\begin{aligned}
 E(\bar{r}) - R &= \\
 -\frac{1}{\bar{X}} [\bar{Y} - E(r_i)E(x_i)] \\
 E(\bar{r}) - R &= \\
 -\frac{1}{\bar{X}} [E(r_i x_i) - E(r_i)E(x_i)] \\
 \bar{Y} &= E(y_i) = E\left(\frac{y_i}{x_i} x_i\right) = E(r_i x_i) \\
 E(\bar{r}) - R &= -\frac{1}{\bar{X}} \text{cov}(r_i x_i) \\
 E(\bar{r}) - R &= \\
 -\frac{1}{\bar{X}} \frac{1}{N} \sum_{i=1}^N (r_i - R)(x_i - \bar{X}) \\
 E(\bar{r}) - R &= -\frac{1}{\bar{X}} \frac{N-1}{N} E(s_{rx}) \\
 E(\bar{r}) - R &= \\
 -\frac{1}{\bar{X}} \frac{N-1}{N(N-1)} \sum_{i=1}^N (r_i - R)(x_i - \bar{X}) \\
 E(\bar{r}) - R &= -\frac{1}{\bar{X}} \frac{N-1}{N} E(s_{rx}) \\
 E(\bar{r}) - R &= \\
 -\frac{1}{\bar{X}} \frac{N-1}{N(N-1)} \sum_{i=1}^N (r_i - R)(x_i - \bar{X})
 \end{aligned}$$

➤ **Hartley-Ross Estimator**

$$\begin{aligned}
 s_{rx} &= \frac{1}{n-1} \left(\sum_{i=1}^n r_i x_i - \frac{\sum_{i=1}^n r_i \sum_{i=1}^n x_i}{n} \right) = \frac{1}{n-1} \left(\sum_{i=1}^n r_i x_i - n \bar{r} \bar{x} \right) \\
 s_{rx} &= \frac{1}{n-1} \left(\sum_{i=1}^n y_i - n \bar{x} \bar{r} \right) = \frac{n}{n-1} (\bar{y} - \bar{r} \bar{x}) \\
 E(\bar{r}) - R &= -\frac{1}{\bar{X}} \frac{N-1}{N} E(s_{rx}) = -\frac{1}{\bar{X}} \frac{N-1}{N} E\left(\frac{n}{n-1} \bar{y} - \bar{r} \bar{x}\right) \\
 E(\bar{r}) + E\left[\frac{1}{\bar{X}} \frac{N-1}{N} \frac{n}{n-1} (\bar{y} - \bar{r} \bar{x})\right] &= R \\
 E(\bar{r}) + E\left[\frac{1}{\bar{X}} \frac{N-1}{N} \frac{n}{n-1} (\bar{y} - \bar{r} \bar{x})\right] &= R
 \end{aligned}$$

$$E\left[\bar{r} + \frac{1}{\bar{X}} \frac{N-1}{N} \frac{n}{n-1} (\bar{y} - \bar{r}\bar{x})\right] = R$$

➤ **HR Unbiased Ratio Estimator**

$$E\left[\bar{r} + \frac{1}{\bar{X}} \frac{N-1}{N} \frac{n}{n-1} (\bar{y} - \bar{r}\bar{x})\right] = R$$

$$E[\text{Estimator}] = \text{Parameter}$$

$$r_{HR} = \bar{r} + \frac{n(N-1)}{N(n-1)} \frac{(\bar{y} - \bar{r}\bar{x})}{\bar{X}}$$

Q. Prove that regression estimator is an unbiased estimator.

Ans:

$$\bar{y}_{reg} = \bar{y} + \hat{\beta}_{yx} (\bar{X} - \bar{x})$$

$$E(\bar{y}_{reg}) = E(\bar{y}) + \hat{\beta}_{yx} E(\bar{X} - \bar{x})$$

$$E(\bar{y}_{reg}) = \bar{Y}$$

Q. Find variance of Regression Estimator.

Ans

$$\bar{y}_{reg} = \bar{y} + \hat{\beta}_{yx} (\bar{X} - \bar{x})$$

$$\bar{y}_{reg} = \bar{Y}(1 + e_0) +$$

$$\hat{\beta}_{yx} (\bar{X} - \bar{X}(1 + e_1))$$

$$\bar{y}_{reg} = \bar{Y}(1 + e_0) - \hat{\beta}_{yx} \bar{X}e_1$$

$$\bar{y}_{reg} = \bar{Y}(1 + e_0) - \hat{\beta}_{yx} \bar{X}e_1$$

$$\bar{y}_{reg} - \bar{Y} = \bar{Y}e_0 - \hat{\beta}_{yx} \bar{X}e_1$$

$$E(\bar{y}_{reg} - \bar{Y})^2 = E(\bar{Y}e_0 - \hat{\beta}_{yx} \bar{X}e_1)^2$$

$$V(\bar{y}_{reg}) = E(\bar{Y}e_0)^2 + E(\hat{\beta}_{yx} \bar{X}e_1)^2$$

$$-2E(\bar{Y}e_0 \hat{\beta}_{yx} \bar{X}e_1)$$

$$= E(\bar{Y}e_0)^2 + E(\hat{\beta}_{yx} \bar{X}e_1)^2$$

$$-2E(\bar{Y}e_0 \hat{\beta}_{yx} \bar{X}e_1)$$

$$= \lambda \begin{bmatrix} \bar{Y}^2 C_y^2 + \hat{\beta}_{yx}^2 \bar{X}^2 C_x^2 \\ -2\bar{Y}\bar{X} \hat{\beta}_{yx} \rho C_y C_x \end{bmatrix}$$

$$\begin{aligned}
 &= \lambda \begin{bmatrix} \bar{Y}^2 C_y^2 + \rho^2 \frac{S_y^2}{S_x^2} \bar{X}^2 C_x^2 \\ -2\bar{Y}\bar{X} \rho \frac{S_y}{S_x} \rho C_y C_x \end{bmatrix} \\
 &= \lambda \begin{bmatrix} \bar{Y}^2 C_y^2 + \rho^2 \frac{S_y^2}{S_x^2} \bar{X}^2 C_x^2 \\ -2\bar{Y}\bar{X} \rho \frac{S_y}{S_x} \rho C_y C_x \end{bmatrix} = \lambda \begin{bmatrix} \bar{Y}^2 C_y^2 + \rho^2 \frac{S_y^2}{S_x^2} \bar{X}^2 C_x^2 \\ -2\bar{Y}\bar{X} \rho \frac{S_y}{S_x} \rho C_y C_x \end{bmatrix} \\
 V(\bar{y}_{reg}) &= \lambda S_y^2 [1 - \rho^2]
 \end{aligned}$$

Q.Find expected value and MSE of ratio estimator by taking all possible samples of size 2 from the following population.

Unit No.	1	2	3	4
Y_i	5	12	15	28
X_i	11	21	32	14

Answer:

➤ **Small Population Example**

- The following is an artificial population with four values.

Unit No.	1	2	3	4
Y_i	5	12	15	28
X_i	11	21	32	14

- We are interested to take all possible samples of size 2.
- Calculate expected value for the ratio estimator.

Unit No.	1	2	3	4
Y_i	5	12	15	28
X_i	11	21	32	14

$$\begin{aligned}
 &\frac{\bar{y}}{\bar{x}} \bar{X} \\
 &\bar{X} = 19.5
 \end{aligned}$$

Unit	$\bar{y}$	$\bar{x}$	$\bar{y} / \bar{x}$	$\bar{y}_r$
1 , 2	8.5	16	0.53	10.36
1 , 3	10	21.5	0.47	9.07
1 , 4	16.5	12.5	1.32	25.74
2 , 3	13.5	26.5	0.51	9.93
2 , 4	20	17.5	1.14	22.29
3 , 4	21.5	23	0.93	18.23

$$MSE(\bar{y}_r) = \sum_{i=1}^6 (\bar{y}_r - 15)^2 \frac{1}{6}$$

$$MSE(\bar{y}_r) = 43.54$$

$$MSE(\bar{y}_r) = \sum_{i=1}^6 (\bar{y}_{ri} - \bar{Y})^2 P(s)$$

Q. Read the data from the following command and select 10000 samples by simple random smapling without replacement. Find ratio estimator from each sample. Further find mean of ratio estimators and compare it with mean per unit estimator.

```
y <- trees$Volume
zi <- trees$Girth
```

Answer:

➤ **Example**

- The Ratio Estimator is

$$\bar{y}_r = \frac{\bar{y}}{\bar{x}} \bar{X}$$

➤ **Defining Population**

```
#---Defining Population
y <- trees$Volume
x <- trees$Girth
N <- 31; n <- 4
mux <- mean(x)
r<-c();mratio<-c();
```

➤ **Relative Efficiency**

```
#----repeating 10000 times
for(i in 1:10000)
{ s<-sample(1:N,n)
m[i]<-mean(y[s])
r[i] <- mean(y[s])/mean(x[s])
mratio[i] <- r[i]*mux }
var(m);var(mratio)
re<-var(m)/var(mratio);re
```

Q. Generate a population of size 1000 for the given parameters using bivariate normal distribution, such that

$$[Y, X] \text{ as } \mu = [2 \quad 2].$$

$$\Sigma = \begin{bmatrix} 1 & 0.85 \\ 0.85 & 1 \end{bmatrix}$$

We consider sample sizes: $n = 10$. Select 10,000 random samples considering SRSWOR and calculate

$$t = \bar{y} \frac{\bar{X}}{\bar{x}}$$

Further, calculate the MSEs for above estimators and calculate relative efficiencies with respect to mean per unit estimator ($\bar{y}$)

Answer:

```
library(mvtnorm)
N=1000;ryx=0.85; n=10;
m=c(2,2); # vector of mean.
# variance covariance matrix is given below.
sig=matrix(c(1,0.85,0.85,1),ncol=2);
r=rmvnorm(N,m,sig);
x=r[,2];y=r[,1];
data=data.frame(x,y);
plot(y,x)
```

➤ **Simulation Study**

```
r<-c();mratio<-c();m=c();
for(i in 1:10000)
{ s<-sample(1:N,n)
m[i]<-mean(y[s])
r[i] <- mean(y[s])/mean(x[s])
mratio[i] <- r[i]*mux }
var(m);var(mratio)
re<-var(m)/var(mratio);re
```

Q. Find Bias of Product Estimator?**Answer:****➤ Product Estimator**

$$\bar{y}_r = \frac{\bar{y}}{\bar{x}} \bar{X}$$
$$\bar{y}_p = \frac{\bar{y}\bar{x}}{\bar{X}}$$

➤ Bias of Product Estimator

$$e_0 = (\bar{y} - \bar{Y}) / \bar{Y}$$

$$e_1 = (\bar{x} - \bar{X}) / \bar{X}$$

Using these notations,

$$E(e_0^2) = \lambda C_y^2,$$

$$E(e_1^2) = \lambda C_x^2,$$

$$E(e_0 e_1) = \lambda C_{yx},$$

$$\text{where, } C_{yx} = \rho_{yx} C_y C_x$$

➤ Using Notations

$$\bar{y} = \bar{Y}(1 + e_0)$$

$$\bar{x} = \bar{X}(1 + e_1)$$

$$\bar{y}_p = \frac{\bar{y}\bar{x}}{\bar{X}}$$

$$\bar{y}_p = \frac{\bar{Y}(1 + e_0)\bar{X}(1 + e_1)}{\bar{X}}$$
$$= \bar{Y}(1 + e_0)(1 + e_1)$$

➤ Bias of Product Estimator

$$= \bar{Y}(1 + e_0)(1 + e_1)$$

$$= \bar{Y}(1 + e_0 + e_1 + e_0 e_1)$$

$$\bar{y}_p - \bar{Y} = \bar{Y}(e_0 + e_1 + e_0 e_1)$$

$$E(\bar{y}_p - \bar{Y}) = \bar{Y}E(e_0 + e_1 + e_0 e_1)$$

$$\text{Bias}(\bar{y}_p) = \bar{Y}E(e_0 + e_1 + e_0 e_1)$$

$$\text{Bias}(\bar{y}_p) = \bar{Y}E(e_0 e_1)$$

$$E(e_0 e_1) = \lambda C_{yx},$$

$$\text{where, } C_{yx} = \rho_{yx} C_y C_x$$

$$\text{Bias}(\bar{y}_p) = \bar{Y} \lambda C_{yx}$$

Q. Find Mean Square Error of Product Estimator?

Answer:

➤ **MSE of Product Estimator**

$$\bar{y}_p = \frac{\bar{y}\bar{x}}{\bar{X}}$$

$$e_0 = (\bar{y} - \bar{Y}) / \bar{Y}$$

$$e_1 = (\bar{x} - \bar{X}) / \bar{X}$$

Using these notations,

$$E(e_0^2) = \lambda C_y^2,$$

$$E(e_1^2) = \lambda C_x^2,$$

$$E(e_0 e_1) = \lambda C_{yx},$$

$$\text{where, } C_{yx} = \rho_{yx} C_y C_x$$

➤ **Using Notations**

$$\bar{y}_p = \frac{\bar{y}\bar{x}}{\bar{X}}$$

$$\bar{y}_p = \frac{\bar{Y}(1+e_0)\bar{X}(1+e_1)}{\bar{X}}$$

$$= \bar{Y}(1+e_0)(1+e_1)$$

➤ **MSE of Product Estimator**

$$= \bar{Y}(1+e_0)(1+e_1)$$

$$= \bar{Y}(1+e_0+e_1+e_0e_1)$$

$$\bar{y}_p - \bar{Y} = \bar{Y}(e_0+e_1+e_0e_1)$$

$$E(\bar{y}_p - \bar{Y}) = \bar{Y}E(e_0+e_1+e_0e_1)$$

$$E(\bar{y}_p - \bar{Y})^2 = \bar{Y}^2 E(e_0+e_1+e_0e_1)^2$$

$$MSE(\bar{y}_p) = \bar{Y}^2 E(e_0+e_1+e_0e_1)^2$$

$$MSE(\bar{y}_p) \approx \bar{Y}^2 E(e_0+e_1)^2$$

$$MSE(\bar{y}_p) \approx \bar{Y}^2 E(e_0^2 + e_1^2 + 2e_0e_1)$$

$$MSE(\bar{y}_p) \approx$$

$$\lambda \bar{Y}^2 (C_y^2 + C_x^2 + 2\rho C_y C_x)$$

$$MSE(\bar{y}_r) \approx$$

$$\lambda \bar{Y}^2 (C_y^2 + C_x^2 - 2\rho C_y C_x)$$

Q. Find expected value and MSE of product estimator by taking all possible samples of size 2 from the following population.

Unit No.	1	2	3	4
Y_i	5	12	15	28
X_i	32	21	14	11

Answer:

➤ **Small Population Example**

- The following is an artificial population with four values.

Unit No.	1	2	3	4
Y_i	5	12	15	28
X_i	32	21	14	11

- We are interested to take all possible samples of size 2.
- Calculate expected value for the product estimator.

$$\frac{\overline{y\bar{x}}}{\bar{X}} \quad \bar{X} = 19.5$$

➤ **All Possible Samples**

Unit No.	1	2	3	4
Y_i	5	12	15	28
X_i	32	21	14	11

Unit	$\bar{y}$	$\bar{x}$	$\bar{y} \times \bar{x}$	$\bar{y}_p$
1, 2	8.5	26.5	225.25	11.55
1, 3	10	23	230	11.79
1, 4	16.5	21.5	354.75	18.19
2, 3	13.5	17.5	236.25	12.12
2, 4	20	16	320	16.41
3, 4	21.5	12.5	268.75	13.78

➤ Expected Value

Unit			
1, 2	8.5	26.5	11.55
1, 3	10	23	11.79
1, 4	16.5	21.5	18.19
2, 3	13.5	17.5	12.12
2, 4	20	16	16.41
3, 4	21.5	12.5	13.78

$$popmean = 15 \quad E(\bar{y}_p) = 13.97$$

$$MSE(\bar{y}_p) = \sum_{i=1}^6 (\bar{y}_{pi} - \bar{Y})^2 P(s)$$

$$MSE(\bar{y}_r) = \sum_{i=1}^6 (\bar{y}_r - 15)^2 \frac{1}{6}$$
$$MSE(\bar{y}_p) = 7.36$$

Q.109 Generate a population of size 1000 for the given parameters using bivariate normal distribution, such that

$$[Y, X] \text{ as } \mu = [2 \quad 2].$$

$$\Sigma = \begin{bmatrix} 1 & -0.85 \\ -0.85 & 1 \end{bmatrix}$$

We consider sample sizes: $n = 10$. Select 10,000 random samples considering SRSWOR and calculate

$$t = \bar{y} \frac{\bar{x}}{\bar{X}}$$

Further calculate the MSEs for above estimators and calculate relative efficiencies with respect to mean per unit estimator ($\bar{y}$)

Answer:

➤ **Defining Population**

```
library(mvtnorm)
N=1000;ryx=-0.85; n=10;
m=c(2,2); # vector of mean.
# variance covariance matrix is given below.
sig=matrix(c(1,-0.85,-0.85,1),ncol=2);
r=rmvnorm(N,m,sig);
x=r[,2];y=r[,1];
data=data.frame(x,y);

p<-c();mp<-c();m=c();
for(i in 1:10000)
{ s<-sample(1:N,n)
m[i]<-mean(y[s])
p[i] <- mean(y[s])*mean(x[s])
mp[i] <- p[i]/mux }
var(m);var(mp)
re<-var(m)/var(mp);re
```

Ratio Estimator in Stratified Sampling

Q. Find the expression of Bias of Combined Type Ratio Estimator.

Answer:

➤ **Combined Type Ratio Estimator in Stratified Sampling**

- The combined type ratio estimator is

$$\bar{y}_{rst} = \frac{\bar{y}_{st}}{\bar{x}_{st}} \bar{X}$$

➤ **Notations**

$$e_{0st} = (\bar{y}_{st} - \bar{Y}) / \bar{Y}$$

$$e_{1st} = (\bar{x}_{st} - \bar{X}) / \bar{X}$$

Using these notations,

$$E(e_{0st}^2) = \sum_{h=1}^k W_h^2 \lambda_h C_{yh}^2,$$

$$E(e_{1st}^2) = \sum_{h=1}^k W_h^2 \lambda_h C_{xh}^2,$$

$$E(e_{0st} e_{1st}) = \sum_{h=1}^k W_h^2 \lambda_h C_{yxh}$$

where, $C_{yxh} = \rho_{yxh} C_{yh} C_{xh}$

➤ **Bias of the Estimator**

$$\bar{y}_{st} = \bar{Y}(1 + e_{0st})$$

$$\bar{x}_{st} = \bar{X}(1 + e_{1st})$$

$$\bar{y}_{rst} = \frac{\bar{y}_{st}}{\bar{x}_{st}} \bar{X}$$

$$\bar{y}_{rst} = \frac{\bar{Y}(1 + e_{0st})}{\bar{X}(1 + e_{1st})} \bar{X}$$

➤ **Bias of Ratio Estimator**

$$= \bar{Y}(1 + e_{0st})(1 + e_{1st})^{-1}$$

$$\approx \bar{Y}(1 + e_{0st})(1 - e_{1st} + e_{1st}^2)$$

$$\bar{y}_{rst} - \bar{Y} \approx \bar{Y} \begin{pmatrix} e_{0st} - e_{1st} \\ -e_{0st}e_{1st} + e_{1st}^2 \end{pmatrix}$$

$$E(e_{0st} e_{1st}) = \sum_{h=1}^k W_h^2 \lambda_h C_{yxh}$$

$$\text{Bias}(\bar{y}_{rst}) \approx$$

$$\bar{Y} \sum_{h=1}^k W_h^2 \lambda_h (C_{xh}^2 - \rho_{yxh} C_{yh} C_{xh})$$

Q. Find the expression of MSE of Combined Type Ratio Estimator?

Answer:

➤ **Combined Ratio Estimator in Stratified Sampling**

- The combined type ratio estimator is

$$\bar{y}_{rst} = \frac{\bar{y}_{st}}{\bar{x}_{st}} \bar{X}$$

➤ **Notations**

$$e_{0st} = (\bar{y}_{st} - \bar{Y}) / \bar{Y}$$

$$e_{1st} = (\bar{x}_{st} - \bar{X}) / \bar{X}$$

Using these notations

$$E(e_{0st}^2) = \sum_{h=1}^k W_h^2 \lambda_h C_{yh}^2,$$

$$E(e_{1st}^2) = \sum_{h=1}^k W_h^2 \lambda_h C_{xh}^2,$$

$$E(e_{0st} e_{1st}) = \sum_{h=1}^k W_h^2 \lambda_h C_{yxh}$$

$$\text{where, } C_{yxh} = \rho_{yxh} C_{yh} C_{xh}$$

➤ **MSE of Combined Type Ratio Estimator**

$$\bar{y}_{st} = \bar{Y}(1 + e_{0st})$$

$$\bar{x}_{st} = \bar{X}(1 + e_{1st})$$

$$\bar{y}_{rst} = \frac{\bar{y}_{st}}{\bar{x}_{st}} \bar{X}$$

$$\bar{y}_{rst} = \frac{\bar{Y}(1 + e_{0st})}{\bar{X}(1 + e_{1st})} \bar{X}$$

$$= \bar{Y}(1 + e_{0st})(1 + e_{1st})^{-1}$$

$$\approx \bar{Y}(1 + e_{0st})(1 - e_{1st})$$

$$\bar{y}_{rst} \approx \bar{Y}(1 + e_{0st} - e_{1st})$$

$$\bar{y}_{rst} - \bar{Y} \approx \bar{Y}(e_{0st} - e_{1st})$$

$$E(\bar{y}_{rst} - \bar{Y})^2 \approx \bar{Y}^2 E(e_{0st} - e_{1st})^2$$

$$\begin{aligned}\bar{y}_{rst} &\approx \bar{Y}(1 + e_{0st} - e_{1st} - e_{0st}e_{1st}) \\ E(\bar{y}_{rst} - \bar{Y})^2 &\approx \bar{Y}^2 E(e_{0st} - e_{1st})^2 \\ E(\bar{y}_{rst} - \bar{Y})^2 &\approx \\ \bar{Y}^2 &\left(E(e_{0st}^2) + E(e_{1st}^2) \right) \\ &\quad - 2E(e_{1st}e_{0st}) \\ MSE(\bar{y}_{rst}) &\approx \\ \bar{Y}^2 \sum_{h=1}^L W_h^2 \gamma_h &\begin{bmatrix} C_{yh}^2 + C_{xh}^2 \\ -2C_{yxh} \end{bmatrix} \\ MSE(\bar{y}_{rst}) &\approx \\ \bar{Y}^2 \sum_{h=1}^L W_h^2 \gamma_h &\begin{bmatrix} C_{yh}^2 + C_{xh}^2 \\ -2C_{yxh} \end{bmatrix}\end{aligned}$$

➤ **Example**

- Consider a population of size 700 consisting on three strata such that $N_1=100$, $N_2=250$ and $N_3=350$. The required sample size is 18.
- The population consisting on two variables Y and X.
- The population mean for the Variable Y and X is 15 and 62.14, respectively.
- The sample size from stratum-1, stratum-2 and stratum-3 is arbitrarily decided as 4, 8 and 6, respectively.
- The sample from each stratum is chosen as (y,x)

Stra-1	Stra-2	Stra-3
1,22	7,39	23,92
3,29	12,55	14,65
2,25	8,42	20,84
5,32	5,30	22,89
	11,51	24,96

10,49 17,68

9,45

12,54

➤ **Considering the Variable Y**

	Stra-1	Stra-2	Stra-3
	1	7	23
	3	12	14
	2	8	20
	5	5	22
		11	24
		10	17
		9	
		12	
mean	2.75	9.25	20
N_h	100	250	350
S_h	1.708	2.493	3.847
n_h	4	8	6

➤ **Considering the Variable X**

	Stra-1	Stra-2	Stra-3
	22	39	92
	29	55	65
	25	42	84
	32	30	89
		51	96
		49	68
		45	
		54	
mean	27	45.63	82.33
N_h	100	250	350
S_h	4.397	8.450	12.91

n_h	4	8	6
-------	---	---	---

➤ **Sample Means**

- Sample mean of variable Y

$$\bar{y}_{st} = \sum_{h=1}^k W_h \bar{y}_h = \sum_{h=1}^k N_h \bar{y}_h / N$$

$$\bar{y}_{st} = \frac{1}{N} [N_1 \bar{y}_1 + N_2 \bar{y}_2 + N_3 \bar{y}_3]$$

- Sample mean of variable X

$$\bar{x}_{st} = \sum_{h=1}^k W_h \bar{x}_h = \sum_{h=1}^k N_h \bar{x}_h / N$$

$$\bar{x}_{st} = \frac{1}{N} [N_1 \bar{x}_1 + N_2 \bar{x}_2 + N_3 \bar{x}_3]$$

➤ **Sample mean of Y**

	Stra-1	Stra-2	Stra-3
mean	2.75	9.25	20
N_h	100	250	350
S_h	1.708	2.493	3.847
n_h	4	8	6

$$\bar{y}_{st} = \sum_{h=1}^k W_h \bar{y}_h = \sum_{h=1}^k N_h \bar{y}_h / N$$

$$\bar{y}_{st} = \frac{1}{N} [N_1 \bar{y}_1 + N_2 \bar{y}_2 + N_3 \bar{y}_3]$$

$$\bar{y}_{st} = 13.70$$

➤ **Sample mean of X**

	Stra-1	Stra-2	Stra-3
mean	27	45.63	82.33
N_h	100	250	350
S_h	4.397	8.450	12.91
n_h	4	8	6

$$\bar{x}_{st} = \sum_{h=1}^k W_h \bar{x}_h = \sum_{h=1}^k N_h \bar{x}_h / N$$

$$\bar{x}_{st} = \frac{1}{N} [N_1 \bar{x}_1 + N_2 \bar{x}_2 + N_3 \bar{x}_3]$$

$$\bar{x}_{st} = 61.32$$

➤ **Combined Ratio Estimator**

- The combined type ratio estimator is

$$\begin{aligned}\bar{y}_{rst} &= \frac{\bar{y}_{st}}{\bar{x}_{st}} \bar{X} \\ \bar{y}_{rst} &= \frac{13.70}{61.32} \times 62.14 \\ \bar{y}_{rst} &= 13.89\end{aligned}$$

Q.115 Find the expression of Bias and MSE of Separate Type Ratio Estimator in stratified sampling.

Answer:

➤ **Separate Type Ratio Estimator**

- The separate type ratio estimator is

$$\begin{aligned}\bar{y}_{rsst} &= \sum_{h=1}^k W_h \bar{y}_{rh} \\ \bar{y}_{rsst} &= \sum_{h=1}^k W_h \frac{\bar{y}_h}{\bar{x}_h} \bar{X}_h\end{aligned}$$

➤ **Bias of Ratio Estimator**

$$Bias(\bar{y}_r) \approx$$

$$\frac{1-f}{n} \bar{Y} (C_x^2 - \rho_{yx} C_y C_x)$$

- In case of Stratified Sampling

$$Bias(\bar{y}_{rh}) \approx$$

$$\frac{1-f_h}{n_h} \bar{Y}_h (C_{xh}^2 - \rho_{yxh} C_{yh} C_{xh})$$

$$E(\bar{y}_{rh}) - \bar{Y}_h \approx$$

$$\frac{1-f_h}{n_h} \bar{Y}_h (C_{xh}^2 - \rho_{yxh} C_{yh} C_{xh})$$

$$E(\bar{y}_{rh}) \approx \bar{Y}_h +$$

$$\frac{1-f_h}{n_h} \bar{Y}_h (C_{xh}^2 - \rho_{yxh} C_{yh} C_{xh})$$

➤ **Bias of Separate Type Ratio Estimator**

$$\bar{y}_{rsst} = \sum_{h=1}^k W_h \bar{y}_{rh} \Rightarrow E(\bar{y}_{rsst}) = \sum_{h=1}^k W_h E(\bar{y}_{rh})$$

$$E(\bar{y}_{rsst}) = \sum_{h=1}^k W_h \left(\bar{Y}_h + \frac{1-f_h}{n_h} \bar{Y}_h (C_{xh}^2 - \rho_{yxh} C_{yh} C_{xh}) \right)$$

$$E(\bar{y}_{rsst}) = \bar{Y} + \sum_{h=1}^k W_h \frac{1-f_h}{n_h} \bar{Y}_h (C_{xh}^2 - \rho_{yxh} C_{yh} C_{xh})$$

$$E(\bar{y}_{rsst}) - \bar{Y} = \sum_{h=1}^k W_h \frac{1-f_h}{n_h} \bar{Y}_h (C_{xh}^2 - \rho_{yxh} C_{yh} C_{xh})$$

➤ **Bias of Separate Type Ratio Estimator**

$$Bias(\bar{y}_{rsst}) =$$

$$\sum_{h=1}^k W_h \lambda_h \bar{Y}_h (C_{xh}^2 - \rho_{yxh} C_{yh} C_{xh})$$

➤ **Expression of MSE for Ratio Estimator of Mean**

$$MSE(\bar{y}_r) \approx$$

$$\lambda \bar{Y}^2 (C_y^2 + C_x^2 - 2\rho_{yx} C_y C_x)$$

$$MSE(\bar{y}_{rh}) \approx$$

$$\lambda_h \bar{Y}_h^2 (C_{yh}^2 + C_{xh}^2 - 2\rho_{yxh} C_{yh} C_{xh})$$

➤ **MSE of Separate Type Ratio Estimator**

$$\bar{y}_{rsst} = \sum_{h=1}^k W_h \bar{y}_{rh} \Rightarrow MSE(\bar{y}_{rsst}) = \sum_{h=1}^k W_h^2 MSE(\bar{y}_{rh})$$

$$MSE(\bar{y}_{rh}) \approx \lambda_h \bar{Y}_h^2 (C_{yh}^2 + C_{xh}^2 - 2\rho_{yxh} C_{yh} C_{xh})$$

$$MSE(\bar{y}_{rsst}) \approx \sum_{h=1}^k W_h^2 \lambda_h \bar{Y}_h^2 (C_{yh}^2 + C_{xh}^2 - 2\rho_{yxh} C_{yh} C_{xh})$$

$$MSE(\bar{y}_{rsst}) \approx \sum_{h=1}^k \left[W_h^2 \lambda_h \frac{\sum_{i=1}^{N_h} (Y_{hi} - R_h X_{hi})^2}{N_h - 1} \right]$$

➤ **Bias and MSE of Separate Type Ratio Estimator**

$$Bias(\bar{y}_{rsst}) =$$

$$\sum_{h=1}^k W_h \lambda_h \bar{Y}_h (C_{xh}^2 - \rho_{yxh} C_{yh} C_{xh})$$

$$MSE(\bar{y}_{rsst}) \approx$$

$$\sum_{h=1}^k W_h^2 \lambda_h \bar{Y}_h^2 \begin{pmatrix} C_{yh}^2 + C_{xh}^2 \\ -2\rho_{yxh} C_{yh} C_{xh} \end{pmatrix}$$

Example

- Consider a population of size 700 consisting on three strata such that $N_1=100$, $N_2=250$ and $N_3=350$. The required sample size is 18.
- The population consisting on two variables Y and X.
- The population means for the Variable Y and X is 15 and 62.14, respectively.
- The sample size from stratum-1, stratum-2 and stratum-3 is arbitrarily decided as 4, 8 and 6, respectively.
- The overall mean of stratum-1, stratum-2 and stratum-3 for variable X is 25, 45 and 85, respectively.
- The sample from each stratum is chosen as (y,x)

Stra-1	Stra-2	Stra-3
1,22	7,39	23,92
3,29	12,55	14,65
2,25	8,42	20,84
5,32	5,30	22,89
	11,51	24,96
	10,49	17,68
	9,45	
	12,54	

➤ Considering the Variable Y

	Stra-1	Stra-2	Stra-3
	1	7	23
	3	12	14
	2	8	20
	5	5	22
		11	24
		10	17
		9	
		12	
mean	2.75	9.25	20
N_h	100	250	350
S_h	1.708	2.493	3.847
n_h	4	8	6

➤ Considering the Variable X

	Stra-1	Stra-2	Stra-3
	22	39	92
	29	55	65
	25	42	84
	32	30	89
		51	96
		49	68
		45	
		54	
mean	27	45.63	82.33
N_h	100	250	350
S_h	4.397	8.450	12.91
n_h	4	8	6

➤ Sample Means

- Sample mean of variable Y

$$\bar{y}_{st} = \sum_{h=1}^k W_h \bar{y}_h = \sum_{h=1}^k N_h \bar{y}_h / N$$

$$\bar{y}_{st} = \frac{1}{N} [N_1 \bar{y}_1 + N_2 \bar{y}_2 + N_3 \bar{y}_3]$$

- Sample mean of variable X

$$\bar{x}_{st} = \sum_{h=1}^k W_h \bar{x}_h = \sum_{h=1}^k N_h \bar{x}_h / N$$

$$\bar{x}_{st} = \frac{1}{N} [N_1 \bar{x}_1 + N_2 \bar{x}_2 + N_3 \bar{x}_3]$$

➤ **Sample Information of Y and X**

- Sample mean of variable Y

	Stra-1	Stra-2	Stra-3
mean	2.75	9.25	20
N_h	100	250	350
S_h	1.708	2.493	3.847
n_h	4	8	6

- Sample mean of variable X

	Stra-1	Stra-2	Stra-3
mean	27	45.63	82.33
N_h	100	250	350
S_h	4.397	8.450	12.91
n_h	4	8	6

$$\begin{aligned}
 \bar{y}_{rsst} &= \sum_{h=1}^3 W_h \frac{\bar{y}_h}{\bar{x}_h} \bar{X}_h \\
 &= W_1 \frac{\bar{y}_1}{\bar{x}_1} \bar{X}_1 + W_2 \frac{\bar{y}_2}{\bar{x}_2} \bar{X}_2 + W_3 \frac{\bar{y}_3}{\bar{x}_3} \bar{X}_3 \\
 &= W_1 \bar{y}_{r1} + W_2 \bar{y}_{r2} + W_3 \bar{y}_{r3} \\
 \bar{y}_{r1} &= \frac{\bar{y}_1}{\bar{x}_1} \bar{X}_1
 \end{aligned}$$

➤ **Separate Type Ratio Estimator**

$$\begin{aligned}
 \bar{y}_{rsst} &= \sum_{h=1}^3 W_h \frac{\bar{y}_h}{\bar{x}_h} \bar{X}_h \\
 \bar{y}_{rsst} &= \sum_{h=1}^3 W_h \bar{y}_{rh} \\
 &= W_1 \bar{y}_{r1} + W_2 \bar{y}_{r2} + W_3 \bar{y}_{r3}
 \end{aligned}$$

$N_1=100$,
 $N_2=250$ and $N_3=350$

W1=0.142857	Means
W2=0.357143	2.546296
W3= 0.5	9.122288
	20.64861

➤ **Separate Type Ratio Estimator**

$$\begin{aligned}\bar{y}_{rsst} &= \sum_{h=1}^3 W_h \frac{\bar{y}_h}{\bar{X}_h} \bar{X}_h \\ \bar{y}_{rsst} &= \sum_{h=1}^3 W_h \bar{y}_{rh} \\ &= W_1 \bar{y}_{r1} + W_2 \bar{y}_{r2} + W_3 \bar{y}_{r3}\end{aligned}$$

$$W_1 = \frac{N_1}{N} = \frac{100}{700} = 0.142857$$

$$W_2 = \frac{N_2}{N} = \frac{250}{700} = 0.357143$$

$$W_3 = \frac{N_3}{N} = \frac{350}{700} = 0.5$$

➤ **Separate Type Ratio Estimator**

$$\begin{aligned}\bar{y}_{rsst} &= \sum_{h=1}^3 W_h \frac{\bar{y}_h}{\bar{X}_h} \bar{X}_h \\ \bar{y}_{rsst} &= \sum_{h=1}^3 W_h \bar{y}_{rh} \\ &= W_1 \bar{y}_{r1} + W_2 \bar{y}_{r2} + W_3 \bar{y}_{r3}\end{aligned}$$

$$\begin{aligned}&= 0.143 \times 2.55 + 0.357 \times 9.12 \\ &\quad + 0.5 \times 20.65 \\ &= 13.95\end{aligned}$$

Example

- Consider a population of size 700 consisting on three strata such that $N_1=100$, $N_2=250$ and $N_3=350$. The required sample size is 18.
- The population consisting on two variables Y and X.
- The population means for the Variable Y and X is 15 and 62.14, respectively.
- The sample size from stratum-1, stratum-2 and stratum-3 is arbitrarily decided as 4, 8 and 6, respectively.
- The overall mean of stratum-1, stratum-2 and stratum-3 for variable X is 25, 45 and 85, respectively.
- The sample from each stratum is chosen as (y,x)

Stra-1	Stra-2	Stra-3
1,22	7,39	23,92
3,29	12,55	14,65
2,25	8,42	20,84
5,32	5,30	22,89
	11,51	24,96
	10,49	17,68
	9,45	
	12,54	

➤ Considering the Variable Y

	Stra-1	Stra-2	Stra-3
	1	7	23
	3	12	14
	2	8	20
	5	5	22
		11	24
		10	17
		9	
		12	
mean	2.75	9.25	20
N_h	100	250	350
S_h	1.708	2.493	3.847
n_h	4	8	6

➤ Considering the Variable X

	Stra-1	Stra-2	Stra-3
	22	39	92
	29	55	65
	25	42	84
	32	30	89
		51	96
		49	68
		45	
		54	
mean	27	45.63	82.33
N_h	100	250	350
S_h	4.397	8.450	12.91
n_h	4	8	6

➤ **Sample Means**

- Sample mean of variable Y

$$\bar{y}_{st} = \sum_{h=1}^k W_h \bar{y}_h = \sum_{h=1}^k N_h \bar{y}_h / N$$

$$\bar{y}_{st} = \frac{1}{N} [N_1 \bar{y}_1 + N_2 \bar{y}_2 + N_3 \bar{y}_3]$$

- Sample mean of variable X

$$\bar{x}_{st} = \sum_{h=1}^k W_h \bar{x}_h = \sum_{h=1}^k N_h \bar{x}_h / N$$

$$\bar{x}_{st} = \frac{1}{N} [N_1 \bar{x}_1 + N_2 \bar{x}_2 + N_3 \bar{x}_3]$$

➤ **Sample Information of Y and X**

- Sample mean of variable Y

	Stra-1	Stra-2	Stra-3
mean	2.75	9.25	20
N_h	100	250	350
S_h	1.708	2.493	3.847
n_h	4	8	6

- Sample mean of variable X

	Stra-1	Stra-2	Stra-3
mean	27	45.63	82.33
N_h	100	250	350
S_h	4.397	8.450	12.91
n_h	4	8	6

$$\bar{y}_{rsst} = \sum_{h=1}^3 W_h \frac{\bar{y}_h}{\bar{x}_h} \bar{X}_h$$

$$= W_1 \frac{\bar{y}_1}{\bar{x}_1} \bar{X}_1 + W_2 \frac{\bar{y}_2}{\bar{x}_2} \bar{X}_2 + W_3 \frac{\bar{y}_3}{\bar{x}_3} \bar{X}_3$$

$$= W_1 \bar{y}_{r1} + W_2 \bar{y}_{r2} + W_3 \bar{y}_{r3}$$

$$\bar{y}_{r1} = \frac{\bar{y}_1}{\bar{x}_1} \bar{X}_1$$

➤ **Separate Type Ratio Estimator**

$$\begin{aligned}\bar{y}_{rsst} &= \sum_{h=1}^3 W_h \frac{\bar{y}_h}{\bar{X}_h} \bar{X}_h \\ \bar{y}_{rsst} &= \sum_{h=1}^3 W_h \bar{y}_{rh} \\ &= W_1 \bar{y}_{r1} + W_2 \bar{y}_{r2} + W_3 \bar{y}_{r3}\end{aligned}$$

$$W1=0.142857$$

$$W2=0.357143$$

$$W3= 0.5$$

➤ **Separate Type Ratio Estimator**

$$\begin{aligned}\bar{y}_{rsst} &= \sum_{h=1}^3 W_h \frac{\bar{y}_h}{\bar{X}_h} \bar{X}_h \\ \bar{y}_{rsst} &= \sum_{h=1}^3 W_h \bar{y}_{rh} \\ &= W_1 \bar{y}_{r1} + W_2 \bar{y}_{r2} + W_3 \bar{y}_{r3}\end{aligned}$$

Means
2.546296
9.122288
20.64861

$$W1=0.142857$$

$$W2=0.357143$$

$$W3= 0.5$$

➤ **Separate Type Ratio Estimator**

$$\begin{aligned}\bar{y}_{rsst} &= \sum_{h=1}^3 W_h \frac{\bar{y}_h}{\bar{X}_h} \bar{X}_h \\ \bar{y}_{rsst} &= \sum_{h=1}^3 W_h \bar{y}_{rh} \\ &= W_1 \bar{y}_{r1} + W_2 \bar{y}_{r2} + W_3 \bar{y}_{r3} \\ &= 0.143 \times 2.55 + 0.357 \times 9.12 \\ &\quad + 0.5 \times 20.65 \\ &= 13.95\end{aligned}$$

Q. Find the variance expression of combined Type Regression Estimator?

Answer:

➤ **Regression Estimator**

- The Regression Estimator without stratification

$$\bar{y}_{reg} = \bar{y} + \hat{\beta}_{yx}(\bar{X} - \bar{x})$$

$$\hat{\beta}_{yx} = \frac{S_{yx}}{S_x^2}$$

➤ **Combined Type Regression Estimator**

- The Regression Estimator under stratified sampling

$$\bar{y}_{regc} = \bar{y}_{st} + \hat{\beta}_{yx}(\bar{X} - \bar{x}_{st})$$

$$\bar{y}_{regc} = \sum_{h=1}^k W_h \bar{y}_h + \hat{\beta}_{yx}(\bar{X} - \sum_{h=1}^k W_h \bar{x}_h)$$

➤ **Expected Value of Estimator**

- Taking Expected Value on both sides

$$E(\bar{y}_{regc}) = \sum_{h=1}^k W E(\bar{y}_h) + \hat{\beta}_{yx} \left(\bar{X} - \sum_{h=1}^k W_h E(\bar{x}_h) \right)$$

$$E(\bar{y}_{regc}) = \sum_{h=1}^k W \bar{Y} + \hat{\beta}_{yx} \left(\bar{X} - \sum_{h=1}^k W_h \bar{X} \right)$$

$$E(\bar{y}_{regc}) = \sum_{h=1}^k W \bar{Y} + \hat{\beta}_{yx} \left(\bar{X} - \sum_{h=1}^k W_h \bar{X} \right)$$

➤ **Expectation of Regression Estimator**

$$E(\bar{y}_{regc}) = \bar{Y} + \hat{\beta}_{yx}(\bar{X} - \bar{X})$$

$$E(\bar{y}_{regc}) = \bar{Y}$$

➤ **Variance of Regression Estimator**

$$\bar{y}_{regc} = \bar{y}_{st} + \hat{\beta}_{yx}(\bar{X} - \bar{x}_{st})$$

$$V(\bar{y}_{regc}) = E[\bar{y}_{regc} - \bar{Y}]^2$$

$$= E\left[\bar{y}_{st} + \hat{\beta}_{yx}(\bar{X} - \bar{x}_{st}) - \bar{Y}\right]^2$$

$$V(\bar{y}_{regc}) = E\left[\bar{y}_{st} + \hat{\beta}_{yx}(\bar{X} - \bar{x}_{st}) - \bar{Y}\right]^2$$

$$V(\bar{y}_{regc}) = E\left[\sum_{h=1}^k W_h \bar{y}_h + \hat{\beta}_{yx}(\bar{X} - \sum_{h=1}^k W_h \bar{x}_h) - \bar{Y}\right]^2$$

$$V(\bar{y}_{regc}) = E\left[\sum_{h=1}^k W_h (\bar{y}_h - \bar{Y}) - \hat{\beta}_{yx} \sum_{h=1}^k W_h (\bar{x}_h - \bar{X})\right]^2$$

$$V(\bar{y}_{regc}) = E \left[\sum_{h=1}^k W_h (\bar{y}_h - \bar{Y}) - \hat{\beta}_{yx} \sum_{h=1}^k W_h (\bar{x}_h - \bar{X}) \right]^2$$

➤ **Variance of Combined Type Regression Estimator**

$$V(\bar{y}_{regc}) = \sum_{h=1}^k W_h^2 \lambda_h \begin{bmatrix} S_{hy}^2 + \beta_0^2 S_{hx}^2 \\ -2\beta_0 S_{hxy} \end{bmatrix}$$

$$\beta_0 = \frac{Cov(\bar{x}_st, \bar{y}_st)}{Var(\bar{x}_st)}$$

$$= \frac{\sum_{i=1}^k \frac{w_i^2 f_i}{n_i} S_{iXY}}{\sum_{i=1}^k \frac{w_i^2 f_i}{n_i} S_{iX}^2}$$

ie minimum variance is given by

$$Var_{min}(\hat{\bar{y}}_{reg}) = \sum_{i=1}^k \frac{w_i^2 f_i}{n_i} (S_{iY}^2 - \beta_0^2 S_{iX}^2).$$

Q. Find the variance of Separate Type Regression Estimator?

Answer:

➤ **Regression Estimator**

- The Regression Estimator without stratification

$$\bar{y}_{reg} = \bar{y} + \hat{\beta}_{yx} (\bar{X} - \bar{x})$$

$$\hat{\beta}_{yx} = \frac{S_{yx}}{S_x^2}$$

➤ **Separate Type Regression Estimator**

- The Combined Type Regression Estimator is

$$\bar{y}_{regc} = \bar{y}_{st} + \hat{\beta}_{yx} (\bar{X} - \bar{x}_{st})$$

- The Separate Type Regression Estimator is

$$\bar{y}_{regs} = \sum_{h=1}^k W_h \left[\bar{y}_h + \hat{\beta}_{yxh} (\bar{X}_h - \bar{x}_h) \right]$$

➤ **Expectation of Separate Type Regression Estimator**

$$\begin{aligned}
 E(\bar{y}_{regs}) &= \sum_{h=1}^k W_h \left[E(\bar{y}_h) + \hat{\beta}_{yxh} (\bar{X}_h - E(\bar{x}_h)) \right] \\
 E(\bar{y}_{regs}) &= \sum_{h=1}^k W_h \left[\bar{Y}_h + \hat{\beta}_{yxh} (\bar{X}_h - \bar{X}_h) \right] \\
 E(\bar{y}_{regs}) &= \sum_{h=1}^k W_h \bar{Y}_h = \bar{Y}
 \end{aligned}$$

➤ **Variance of Regression Estimator**

$$\begin{aligned}
 \bar{y}_{regs} &= \sum_{h=1}^k W_h \bar{y}_h + \hat{\beta}_{yxh} \sum_{h=1}^k W_h (\bar{X}_h - \bar{x}_h) \\
 V(\bar{y}_{regs}) &= E[\bar{y}_{regs} - \bar{Y}]^2 \\
 V(\bar{y}_{regs}) &= E \left[\sum_{h=1}^k W_h \bar{y}_h + \hat{\beta}_{yxh} \sum_{h=1}^k W_h (\bar{X}_h - \bar{x}_h) - \bar{Y} \right]^2 \\
 V(\bar{y}_{regs}) &= E \left[\sum_{h=1}^k W_h (\bar{y}_h - \bar{Y}) - \sum_{h=1}^k W_h \hat{\beta}_{yxh} (\bar{x}_h - \bar{X}_h) \right]^2 \\
 V(\bar{y}_{regs}) &= \left[\sum_{h=1}^k W_h^2 E(\bar{y}_h - \bar{Y})^2 + \sum_{h=1}^k W_h^2 \hat{\beta}_{yxh}^2 E(\bar{x}_h - \bar{X}_h)^2 \right. \\
 &\quad \left. - 2 \sum_{h=1}^k W_h^2 \hat{\beta}_{yxh} E(\bar{y}_h - \bar{Y})(\bar{x}_h - \bar{X}_h) \right] \\
 V(\bar{y}_{regc}) &= \sum_{h=1}^k W_h^2 \lambda_h \begin{bmatrix} S_{hy}^2 + \beta_{yxh}^2 S_{hx}^2 \\ -2\beta_{yxh} S_{hxy} \end{bmatrix} \\
 \beta_{0i} &= \frac{S_{ixy}}{S_{ix}^2}
 \end{aligned}$$

Q. Simulation Study for Combined Type Ratio Estimator.

Answer:

➤ **Simulation Study**

```

library(mvtnorm)
#population and sample size.
N=300;N1=100;N2=100;N3=100;n1=15;n2=15;n3=15;n=45;

```

```
w1=N1/N;
```

```
w2=N2/N;
```

```
w3=N3/N;
```

➤ **Defining Parameters**

```
# mean vectors for stratum 1, 2 and 3.
```

```
m1=c(10,5);
```

```
m2=c(20,10);
```

```
m3=c(30,15)
```

```
#variance covariance matrix for #stratum 1, 2 and 3 given #below.
```

```
sig1=matrix(c(1,0.85,0.85,1),ncol=2);
```

```
sig2=matrix(c(1,0.80,0.80,1),ncol=2); sig3=matrix(c(1,0.75,0.75,1),ncol=2);
```

➤ **Generating Populations**

```
r1=rmvnorm(N1,m1,sig1);
```

```
r2=rmvnorm(N2,m2,sig2);
```

```
r3=rmvnorm(N3,m3,sig3);
```

```
# generate random variable for #stratum 1.
```

```
y1=r1[,1];x1=r1[,2]
```

```
# generate random variable for #stratum 2.
```

```
y2=r2[,1];x2=r2[,2]
```

```
# generate random variable for #stratum 3.
```

```
y3=r3[,1];x3=r3[,2]
```

```
x<-c(x1,x2,x3)
```

➤ **Looping**

```
for(i in 1:5000){  
  sa1=sample(1:N1,n1)          yst[i]=w1*mean(ys1)+w2*mean(ys2)+w3*mean(ys3);  
  xs1=x1[sa1];ys1=y1[sa1];    xst[i]=w1*mean(xs1)+w2*mean(xs2)+w3*mean(xs3);  
                                rst[i]=yst[i]*(mean(x)/(xst[i]))  
                                }  
  sa2=sample(1:N2,n2)          mean(yst);mean(xst);  
  xs2=x2[sa2];ys2=y2[sa2];    mean(rst);  
                                var(yst);var(rst);  
  sa3=sample(1:N3,n3)          var(yst)/var(rst)  
  xs3=x3[sa3];ys3=y3[sa3];
```

➤ **Means and Variances**

```
> mean(yst);mean(xst);mean(rst);
```

```
[1] 20.01764
```

```
[1] 10.02838
```

```
[1] 20.02706
```

```
> var(yst);var(rst);var(yst)/var(rst)
```

Q. Define the process of simulation study for combined type ratio estimator also find mean and variance of estimator using following data

```
N=300;N1=100;N2=100;N3=100;n1=15;n2=15;n3=15;n=45;
```

```
# mean vectors for stratum 1, 2 and 3.
```

```
m1=c(10,5);
```

```
m2=c(20,10);
```

```
m3=c(30,15)
```

```
#variance covariance matrix for #stratum 1, 2 and 3 given below.
```

```
sig1=matrix(c(1,0.85,0.85,1),ncol=2)
```

```
sig2=matrix(c(1,0.80,0.80,1),ncol=2); sig3=matrix(c(1,0.75,0.75,1),ncol=2);
```

Ans.

```
library(mvtnorm)
```

```
#population and sample size.
```

```
N=300;N1=100;N2=100;N3=100;n1=15;n2=15;n3=15;n=45;
```

```
w1=N1/N;
```

```
w2=N2/N;
```

```
w3=N3/N;
```

```
# mean vectors for stratum 1, 2 and 3.
```

```
m1=c(10,5);
```

```
m2=c(20,10);
```

```
m3=c(30,15)
```

```
#variance covariance matrix for #stratum 1, 2 and 3 given below.
```

```
sig1=matrix(c(1,0.85,0.85,1),ncol=2)
```

```
sig2=matrix(c(1,0.80,0.80,1),ncol=2); sig3=matrix(c(1,0.75,0.75,1),ncol=2);
```

```
r1=rmvnorm(N1,m1,sig1);
```

```
r2=rmvnorm(N2,m2,sig2);
```

```
r3=rmvnorm(N3,m3,sig3);
```

generate random variable for #stratum 1.

```
y1=r1[,1];x1=r1[,2]
```

generate random variable for #stratum 2

```
y2=r2[,1];x2=r2[,2]
```

generate random variable for #stratum 3.

```
y3=r3[,1];x3=r3[,2]
```

```
x<-c(x1,x2,x3)
```

```
for(i in 1:5000){
```

```
sa1=sample(1:N1,n1)
```

```
xs1=x1[sa1];ys1=y1[sa1];
```

```
sa2=sample(1:N2,n2)
```

```
xs2=x2[sa2];ys2=y2[sa2];
sa3=sample(1:N3,n3)
xs3=x3[sa3];ys3=y3[sa3];
yst[i]=w1*mean(ys1)+w2*mean(ys2)+w3*mean(ys3);
xst[i]=w1*mean(xs1)+w2*mean(xs2)+w3*mean(xs3);
rst[i]=yst[i]*(mean(x)/(xst[i]))
}
mean(yst);mean(xst);
mean(rst);
var(yst);var(rst);
var(yst)/var(rst)
```

#Means and Variances

```
> mean(yst);mean(xst);mean(rst);
> var(yst);var(rst);var(yst)/var(rst)
```

Q. Define the process of simulation study for separate type ratio estimator also find mean and variance of estimator using following data

```
library(mvtnorm)
#population and sample size.
N=300;N1=100;N2=100;N3=100;n1=15;n2=15;n3=15;n=45;
# mean vectors for stratum 1, 2 and 3.
m1=c(10,5);
m2=c(20,10);
m3=c(30,15)
#variance covariance matrix for #stratum 1, 2 and 3 given #below.
sig1=matrix(c(1,0.85,0.85,1),ncol=2);
sig2=matrix(c(1,0.80,0.80,1),ncol=2); sig3=matrix(c(1,0.75,0.75,1),ncol=2);
```

Ans:

```
library(mvtnorm)
#population and sample size.
N=300;N1=100;N2=100;N3=100;n1=15;n2=15;n3=15;n=45;
w1=N1/N;
w2=N2/N;
w3=N3/N;

# mean vectors for stratum 1, 2 and 3.
m1=c(10,5);
m2=c(20,10);
m3=c(30,15)

#variance covariance matrix for #stratum 1, 2 and 3 given #below.
sig1=matrix(c(1,0.85,0.85,1),ncol=2);
sig2=matrix(c(1,0.80,0.80,1),ncol=2); sig3=matrix(c(1,0.75,0.75,1),ncol=2);
r1=rmvnorm(N1,m1,sig1);
```

```
r2=rmvnorm(N2,m2,sig2);
r3=rmvnorm(N3,m3,sig3);

# generate random variable for #stratum 1.
y1=r1[,1];x1=r1[,2]

# generate random variable for #stratum 2.
y2=r2[,1];x2=r2[,2]

# generate random variable for #stratum 3.
y3=r3[,1];x3=r3[,2]
x<-c(x1,x2,x3)
for(i in 1:5000){
sa1=sample(1:N1,n1)
xs1=x1[sa1];ys1=y1[sa1];
r1[i]=mean(ys1)*(mean(x1)/mean(xs1))
sa2=sample(1:N2,n2)
xs2=x2[sa2];ys2=y2[sa2];
r2[i]=mean(ys2)*(mean(x2)/mean(xs2))
sa3=sample(1:N3,n3)
xs3=x3[sa3];ys3=y3[sa3];
r3[i]=mean(ys3)*(mean(x3)/mean(xs3))
rsst[i]=w1*r1[i]+w2*r2[i]+w3*r3[i];
yst[i]=w1*mean(ys1)+w2*mean(ys2)+w3*mean(ys3);
}
mean(rsst);mean(yst);
var(yst);var(rsst);var(yst)/var(rsst)
```

Means and Variances

```
> mean(rsst);mean(yst);
```

Double Sampling

Introduction:

Ratio and regression methods of estimation require the knowledge of population mean of auxiliary variable in advance. An estimate of mean of auxiliary variable from a large sample may be used. This procedure of selecting a large sample for collecting information on auxiliary variable x and then selecting a subsample from it for collecting the information on the study variable y is called double sampling or two-phase sampling.

The estimate is calculated by taking the sample in two phases. The expected value of the statistic is

$$E(t) = E_1(E_2(t))$$

The variance of the statistic is

$$Var(t) = E(t - E(t))^2$$

$$Var(t) = E[(t - E_2(t)) + (E_2(t) - E(t))]^2$$

$$Var(t) = E(t - E_2(t))^2 + E(E_2(t) - E(t))^2$$

$$Var(t) = E_1 E_2 (t - E_2(t))^2 + E_1 E_2 (E_2(t) - E(t))^2$$

$$Var(t) = E_1 [V_2(t)] + E_1 (E_2(t) - E_1(E_2(t)))^2$$

$$Var(t) = E_1 [V_2(t)] + V_1 [E_2(t)]$$

The Ratio Estimator for population mean

$$\bar{y}_r = \frac{\bar{y}}{\bar{x}} \bar{X}$$

The Ratio Estimator for Double Sampling

$$\bar{y}_{rd} = \frac{\bar{y}_2}{\bar{x}_2} \bar{x}_1$$

Notations

$$e_0 = (\bar{y}_2 - \bar{Y}) / \bar{Y}$$

$$e_1 = (\bar{x}_1 - \bar{X}) / \bar{X}$$

$$e_2 = (\bar{x}_2 - \bar{X}) / \bar{X}$$

Using these notations

$$E(e_0) = 0, E(e_0^2) = \lambda_2 C_y^2,$$

$$E(e_1) = 0, E(e_1^2) = \lambda_1 C_x^2,$$

$$E(e_2) = 0, E(e_2^2) = \lambda_2 C_x^2,$$

$$E(e_1 e_2) = \lambda_1 C_x^2,$$

$$E(e_0 e_1) = \lambda_1 \rho_{yx} C_y C_x, E(e_0 e_2) = \lambda_2 \rho_{yx} C_y C_x,$$

Prove that $E(e_1 e_2) = \lambda_1 C_x^2$

Proof:

$$e_1 e_2 = \frac{(\bar{x}_1 - \bar{X})(\bar{x}_2 - \bar{X})}{\bar{X}^2} \Rightarrow E(e_1 e_2) = \frac{E(\bar{x}_1 - \bar{X})(\bar{x}_2 - \bar{X})}{\bar{X}^2}$$

$$E(e_1 e_2) = \frac{E_1 E_2 (\bar{x}_1 - \bar{X})(\bar{x}_2 - \bar{X})}{\bar{X}^2}$$

$$E(e_1 e_2) = \frac{E_1 (\bar{x}_1 - \bar{X})(\bar{x}_1 - \bar{X})}{\bar{X}^2}$$

$$E(e_1 e_2) = \frac{E_1 (\bar{x}_1 - \bar{X})^2}{\bar{X}^2}$$

$$E(e_1 e_2) = \frac{\lambda_1 S_x^2}{\bar{X}^2}$$

$$E(e_1 e_2) = \lambda_1 C_x^2$$

Q. Find Bias of Ratio Estimator in Double Sampling.

The Ratio Estimator for Double Sampling

$$\bar{y}_{rd} = \frac{\bar{y}_2}{\bar{x}_2} \bar{x}_1$$

$$e_0 = (\bar{y}_2 - \bar{Y}) / \bar{Y}$$

$$e_1 = (\bar{x}_1 - \bar{X}) / \bar{X}$$

$$e_2 = (\bar{x}_2 - \bar{X}) / \bar{X}$$

$$\bar{y}_{rd} = \frac{\bar{y}_2}{\bar{x}_2} \bar{x}_1$$

$$\bar{y}_{rd} = \frac{\bar{Y}(1+e_0)}{\bar{X}(1+e_2)} \bar{X}(1+e_1)$$

$$\bar{y}_{rd} = \frac{\bar{Y}(1+e_0)}{(1+e_2)} (1+e_1)$$

The Bias of Ratio Estimator in Double Sampling

$$\bar{y}_{rd} = \bar{Y}(1+e_0)(1+e_1)(1+e_2)^{-1}$$

$$\bar{y}_{rd} = \bar{Y}(1+e_0)(1+e_1)(1-e_2+e_2^2-....)$$

$$\bar{y}_{rd} = \bar{Y}(1+e_0)(1+e_1)(1-e_2+e_2^2)$$

$$\bar{y}_{rd} = \bar{Y}(1+e_0)(1+e_1-e_2-e_1e_2+e_2^2+e_1e_2^2)$$

$$\bar{y}_{rd} = \bar{Y}(1+e_0)(1+e_1-e_2-e_1e_2+e_2^2)$$

$$\bar{y}_{rd} - \bar{Y} = \bar{Y}(e_1 - e_2 - e_1e_2 + e_2^2 + e_0 + e_0e_1 - e_0e_2)$$

$$E(\bar{y}_{rd} - \bar{Y}) = \bar{Y}E(e_1 - e_2 - e_1e_2 + e_2^2 + e_0 + e_0e_1 - e_0e_2)$$

$$E(\bar{y}_{rd} - \bar{Y}) = \bar{Y}(\lambda_2 C_x^2 - \lambda_1 C_x^2 + \lambda_1 \rho C_y C_x - \lambda_2 \rho C_y C_x)$$

$$E(\bar{y}_{rd} - \bar{Y}) = \bar{Y} [(\lambda_2 - \lambda_1)C_x^2 - (\lambda_2 - \lambda_1)\rho C_y C_x]$$

$$E(\bar{y}_{rd} - \bar{Y}) = \bar{Y} (\lambda_2 - \lambda_1) [C_x^2 - \rho C_y C_x]$$

$$Bias(\bar{y}_{rd}) = \bar{Y} (\lambda_2 - \lambda_1) [C_x^2 - \rho C_y C_x]$$

Q.Find MSE of Ratio Estimator in Double Sampling.

The Ratio Estimator for Double Sampling

$$\bar{y}_{rd} = \frac{\bar{y}_2}{\bar{x}_2} \bar{x}_1$$

$$e_0 = (\bar{y}_2 - \bar{Y}) / \bar{Y}$$

$$e_1 = (\bar{x}_1 - \bar{X}) / \bar{X}$$

$$e_2 = (\bar{x}_2 - \bar{X}) / \bar{X}$$

$$\bar{y}_{rd} = \frac{\bar{y}_2}{\bar{x}_2} \bar{x}_1$$

$$\bar{y}_{rd} = \frac{\bar{Y}(1+e_0)}{\bar{X}(1+e_2)} \bar{X}(1+e_1)$$

$$\bar{y}_{rd} = \frac{\bar{Y}(1+e_0)}{(1+e_2)} (1+e_1)$$

MSE of Ratio Estimator in Double Sampling

$$\bar{y}_{rd} = \bar{Y}(1+e_0)(1+e_1)(1+e_2)^{-1}$$

$$\bar{y}_{rd} = \bar{Y}(1+e_0)(1+e_1)(1-e_2+e_2^2-....)$$

$$\bar{y}_{rd} - \bar{Y} = \bar{Y}(e_1 - e_2 + e_0)$$

$$E(\bar{y}_{rd} - \bar{Y})^2 = \bar{Y}^2 E(e_1 - e_2 + e_0)^2$$

$$E(\bar{y}_{rd} - \bar{Y})^2 = \bar{Y}^2 E(e_1^2 + e_2^2 + e_0^2 + e_0e_1 - e_0e_2 - e_1e_2)$$

$$MSE(\bar{y}_{rd}) = \bar{Y}^2 [\lambda_2 C_y^2 + (\lambda_2 - \lambda_1)(C_x^2 - 2C_x C_y \rho_{xy})]$$

$$MSE(\bar{y}_{rd}) = \lambda_2 \bar{Y}^2 (C_y^2 + C_x^2 - 2C_x C_y \rho_{xy}) - \lambda_1 \bar{Y}^2 (C_x^2 - 2C_x C_y \rho_{xy})$$

Q. Find Bias of Product Estimator in Double Sampling.

The Product Estimator is

$$\bar{y}_p = \bar{y} \frac{\bar{x}}{\bar{X}}$$

The Product Estimator in Double Sampling is

$$\bar{y}_{pd} = \frac{\bar{y}_2 \bar{x}_2}{\bar{x}_1}$$

$$e_0 = (\bar{y}_2 - \bar{Y}) / \bar{Y}, e_1 = (\bar{x}_1 - \bar{X}) / \bar{X}, e_2 = (\bar{x}_2 - \bar{X}) / \bar{X}$$

$$\bar{y}_2 = \bar{Y}(1 + e_0), \bar{x}_1 = \bar{X}(1 + e_1), \bar{x}_2 = \bar{X}(1 + e_2)$$

Notations

$$E(e_0) = 0, E(e_0^2) = \lambda_2 C_y^2, E(e_1) = 0$$

$$E(e_1^2) = \lambda_1 C_x^2, E(e_1) = 0, E(e_2^2) = \lambda_2 C_x^2$$

$$E(e_1 e_2) = \lambda_1 C_x^2, E(e_0 e_1) = \lambda_1 \rho_{yx} C_y C_x, E(e_0 e_2) = \lambda_2 \rho_{yx} C_y C_x$$

Bias of Product Estimator in Double Sampling

$$\bar{y}_{pd} = \frac{\bar{y}_2}{\bar{x}_1} \bar{x}_2$$

$$\bar{y}_{pd} = \frac{\bar{Y}(1 + e_0)}{\bar{X}(1 + e_1)} \bar{X}(1 + e_2)$$

$$\bar{y}_{pd} = \frac{\bar{Y}(1 + e_0)}{(1 + e_1)} (1 + e_2)$$

$$\bar{y}_{pd} = \bar{Y}(1 + e_0)(1 + e_2)(1 + e_1)^{-1}$$

$$\bar{y}_{pd} = \bar{Y}(1+e_0)(1+e_2)(1-e_1+e_1^2-....)$$

$$\bar{y}_{pd} = \bar{Y}(1+e_0)(1+e_2)(1-e_1+e_1^2)$$

$$\bar{y}_{pd} = \bar{Y}(1+e_0)(1+e_2-e_1-e_1e_2+e_1^2+e_2e_1^2)$$

$$\bar{y}_{pd} = \bar{Y}(1+e_0)(1+e_2-e_1-e_1e_2+e_1^2)$$

$$\bar{y}_{pd} - \bar{Y} = \bar{Y}(e_2 - e_1 - e_1e_2 + e_1^2 + e_0 + e_0e_2 - e_0e_1)$$

$$E(\bar{y}_{pd} - \bar{Y}) = \bar{Y}E(e_2 - e_1 - e_1e_2 + e_1^2 + e_0 + e_0e_2 - e_0e_1)$$

$$E(\bar{y}_{pd} - \bar{Y}) = \bar{Y}(\lambda_1 C_x^2 - \lambda_1 C_x^2 + \lambda_2 \rho C_y C_x - \lambda_1 \rho C_y C_x)$$

$$Bias(\bar{y}_{pd}) = \bar{Y}[(\lambda_2 - \lambda_1)\rho C_y C_x]$$

Q. Find MSE of Product Estimator in Double Sampling.

The Product Estimator for Double Sampling

$$\bar{y}_{pd} = \frac{\bar{y}_2}{\bar{x}_1} \bar{x}_2$$

$$e_0 = (\bar{y}_2 - \bar{Y})/\bar{Y}, e_1 = (\bar{x}_1 - \bar{X})/\bar{X}, e_2 = (\bar{x}_2 - \bar{X})/\bar{X}$$

$$\bar{y}_{pd} = \frac{\bar{y}_2}{\bar{x}_1} \bar{x}_2, \bar{y}_{pd} = \frac{\bar{Y}(1+e_0)}{\bar{X}(1+e_1)} \bar{X}(1+e_2), \bar{y}_{pd} = \frac{\bar{Y}(1+e_0)}{(1+e_1)}(1+e_2)$$

The MSE of Product Estimator in Double Sampling

$$\bar{y}_{pd} = \bar{Y}(1+e_0)(1+e_2)(1+e_1)^{-1}$$

$$\bar{y}_{pd} = \bar{Y}(1+e_0)(1+e_2)(1-e_1+....)$$

$$\bar{y}_{pd} - \bar{Y} = \bar{Y}(e_2 - e_1 + e_0)$$

$$E(\bar{y}_{pd} - \bar{Y})^2 = \bar{Y}^2 E(e_2 - e_1 + e_0)^2$$

$$E(\bar{y}_{pd} - \bar{Y})^2 = \bar{Y}^2 E(e_1^2 + e_2^2 + e_0^2 + 2e_0e_2 - 2e_0e_1 - 2e_1e_2)$$

$$MSE(\bar{y}_{rd}) = \bar{Y}^2 \left[\lambda_2 C_y^2 + (\lambda_2 - \lambda_1)(C_x^2 + 2C_x C_y \rho_{xy}) \right]$$

$$MSE(\bar{y}_{rd}) = \lambda_2 \bar{Y}^2 (C_y^2 + C_x^2 + 2C_x C_y \rho_{xy}) - \lambda_1 \bar{Y}^2 (C_x^2 + 2C_x C_y \rho_{xy})$$

Q. Generate a population of size 1000 for the given parameters using bivariate normal distribution, such that

$$[Y, X] \text{ as } \mu = [10 \ 2].$$

$$\Sigma = \begin{bmatrix} 1 & 0.85 \\ 0.85 & 1 \end{bmatrix}$$

Consider sample sizes n1=100;n2=20. Select 10,000 random samples considering double and calculate

$$t = \bar{y}_2 \frac{\bar{x}_1}{\bar{x}_2}$$

Further, calculate the MSEs for above estimator.

Ans:

library(mvtnorm)

N=1000;ryx=0.85; n1=100;n2=20;

m=c(10,2); # vector of mean.

variance covariance matrix is given below.

sig=matrix(c(1,0.85,0.85,1),ncol=2);

r=rmvnorm(N,m,sig);

x=r[,2];y=r[,1];

r<-c();mratio<-c();m=c();

for(i in 1:10000)

{

s1<-sample(1:N,n1)

s2<-sample(s1,n2)

m[i]<-mean(x[s1])

r[i] <- mean(y[s2])/mean(x[s2])

mratio[i] <- r[i]*m[i]

}

mean(mratio)

var(mratio)

Two Stage Sampling

Q. Prove that sample mean is unbiased estimator of population mean in two stage sampling

$$\bar{\bar{y}} = \sum_{i=1}^n \sum_{j=1}^m \frac{y_{ij}}{nm}$$

Ans:

The estimate is calculated by taking the sample in two stages

The expected value and variance of the statistic is

$$E(t) = E_1(E_2(t))$$

$$Var(t) = E_1[V_2(t)] + V_1[E_2(t)]$$

Expected Value of Mean

$$\bar{\bar{y}} = \sum_{i=1}^n \sum_{j=1}^m \frac{y_{ij}}{nm}$$

$$E(\bar{\bar{y}}) = E_1(E_2(\bar{\bar{y}}))$$

$$E(\bar{\bar{y}}) = E_1\left(E_2\left(\sum_{i=1}^n \sum_{j=1}^m \frac{y_{ij}}{nm}\right)\right)$$

$$E(\bar{\bar{y}}) = E_1\left(\sum_{i=1}^n \frac{1}{n} E_2\left(\sum_{j=1}^m \frac{y_{ij}}{m}\right)\right)$$

$$E(\bar{\bar{y}}) = E_1\left(\sum_{i=1}^n \frac{1}{n} \left(\sum_{j=1}^m \frac{E_2(y_{ij})}{m}\right)\right)$$

$$E(\bar{\bar{y}}) = E_1\left(\sum_{i=1}^n \frac{1}{n} \left(\sum_{j=1}^M \frac{y_{ij}}{M}\right)\right)$$

$$E(\bar{\bar{y}}) = E_1\left(\sum_{i=1}^n \frac{1}{n} \bar{Y}_i\right)$$

$$E(\bar{\bar{y}}) = \left(\sum_{i=1}^n \frac{1}{n} E_1(\bar{Y}_i)\right)$$

$$E(\bar{\bar{y}}) = E_1(\bar{Y}_i)$$

$$E(\bar{y}) = \left(\sum_{i=1}^N \frac{\bar{Y}_i}{N} \right) \Rightarrow E(\bar{y}) = \bar{\bar{Y}}$$

Q. Prove that sample mean is unbiased estimator of population mean in two stage sampling.

$$\bar{\bar{y}} = \sum_{i=1}^n \sum_{j=1}^m \frac{y_{ij}}{nm}$$

Ans:

$$\bar{\bar{y}} = \sum_{i=1}^n \sum_{j=1}^m \frac{y_{ij}}{nm} = \sum_{i=1}^n \frac{1}{n} \bar{y}_i$$

$$Var(t) = E_1[V_2(t)] + V_1[E_2(t)]$$

$$Var(\bar{\bar{y}}) = E_1[V_2(\bar{\bar{y}})] + V_1[E_2(\bar{\bar{y}})]$$

$$E_2(\bar{\bar{y}}) = E_2 \left[\sum_{i=1}^n \sum_{j=1}^m \frac{y_{ij}}{nm} \right]$$

$$E_2(\bar{\bar{y}}) = \left[\sum_{i=1}^n \frac{1}{n} \sum_{j=1}^m \frac{E_2(y_{ij})}{m} \right]$$

$$E_2(\bar{\bar{y}}) = \left[\sum_{i=1}^n \frac{1}{n} \sum_{j=1}^M \frac{y_{ij}}{M} \right]$$

$$E_2(\bar{\bar{y}}) = \left[\sum_{i=1}^n \frac{1}{n} \sum_{j=1}^M \frac{y_{ij}}{M} \right]$$

$$E_2(\bar{\bar{y}}) = \left[\sum_{i=1}^n \frac{1}{n} \bar{Y}_i \right]$$

$$E_2(\bar{\bar{y}}) = \bar{\bar{Y}}_n$$

$\bar{\bar{Y}}_n$ is the estimator based on the 1st stage sample of size n, by SRS, we have

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

$$V_1[E_2(\bar{y})] = V_1[\bar{Y}_n] \Rightarrow V_1[E_2(\bar{y})] = \left(\frac{N-n}{Nn} \right) S_1^2$$

$$E_1[V_2(\bar{y})] = E_1 \left[V_2 \left(\sum_{i=1}^n \frac{1}{n} \bar{y}_i \right) \right]$$

$$E_1[V_2(\bar{y})] = E_1 \left[\frac{1}{n^2} \sum_{i=1}^n V_2(\bar{y}_i) + \frac{1}{n^2} \sum_{i \neq j=1}^n \sum Cov(\bar{y}_i, \bar{y}_j) \right]$$

$$E_1[V_2(\bar{y})] = E_1 \left[\frac{1}{n^2} \sum_{i=1}^n V_2(\bar{y}_i) \right] = \frac{1}{n^2} E_1 \left[\sum_{i=1}^n \frac{M-m}{Mm} S_{2i}^2 \right]$$

$$E_1[V_2(\bar{y})] = \frac{1}{n^2} \left[\sum_{i=1}^n \frac{M-m}{Mm} E_1(S_{2i}^2) \right] = \frac{1}{n} \left[\frac{M-m}{Mm} E_1(S_{2i}^2) \right]$$

$$E_1[V_2(\bar{y})] = \frac{1}{n} \left[\frac{M-m}{Mm} \sum_{i=1}^N \frac{S_{2i}^2}{N} \right]$$

$$E_1[V_2(\bar{y})] = \frac{1}{n} \left[\frac{M-m}{Mm} S_2^2 \right]$$

$$Var(\bar{y}) = E_1[V_2(\bar{y})] + V_1[E_2(\bar{y})]$$

$$E_1[V_2(\bar{y})] = \frac{1}{n} \left[\frac{M-m}{Mm} S_2^2 \right]$$

$$V_1[E_2(\bar{y})] = \left(\frac{N-n}{Nn} \right) S_1^2$$

$$Var(t) = E_1[V_2(t)] + V_1[E_2(t)]$$

$$Var(t) = \frac{1}{n} \left[\frac{M-m}{Mm} S_2^2 \right] + \left(\frac{N-n}{Nn} \right) S_1^2$$

$$Var(t) = \frac{1}{n} \left[\frac{M-m}{Mm} s_2^2 \right] + \left(\frac{N-n}{Nn} \right) s_1^2$$

Q. Explain two stage sampling for unequal first stage units, also find expected value of following estimator.

$$\bar{\bar{y}} = \sum_{i=1}^n \frac{M_i \bar{y}_i}{n\bar{M}} = \sum_{i=1}^n \frac{u_i \bar{y}_i}{n}$$

Let the population has N clusters with M_i as size of i^{th} cluster. Let a sample of n first stage units be selected from this population. Let m_i units will be selected at second stage.
 M_i = Size of i^{th} cluster

$$M_0 = \sum_{i=1}^N M_i = N\bar{M}$$

Mean of i^{th} first stage unit

$$\bar{Y}_i = \frac{\sum_{j=1}^{M_i} y_{ij}}{M_i}$$

Overall population mean

$$\bar{Y} = \frac{\sum_{i=1}^N M_i \bar{Y}_i}{N\bar{M}}$$

The Expected Value of Mean

$$\bar{\bar{y}} = \sum_{i=1}^n \frac{M_i \bar{y}_i}{n\bar{M}} = \sum_{i=1}^n \frac{u_i \bar{y}_i}{n}$$

$$E(\bar{\bar{y}}) = E_1(E_2(\bar{\bar{y}}))$$

$$E(\bar{\bar{y}}) = E_1\left(E_2\left(\sum_{i=1}^n \frac{u_i \bar{y}_i}{n}\right)\right)$$

$$E(\bar{\bar{y}}) = E_1\left(E_2\left(\sum_{i=1}^n \frac{u_i \bar{y}_i}{n}\right)\right)$$

$$E(\bar{\bar{y}}) = E_1\left(\sum_{i=1}^n \frac{u_i \bar{Y}_i}{n}\right)$$

$$E(\bar{\bar{y}}) = \left(\sum_{i=1}^n \frac{M_i E_1(\bar{Y}_i)}{n} \right)$$

$$E(\bar{\bar{y}}) = \left(\sum_{i=1}^n \frac{E_1(u_i \bar{Y}_i)}{n} \right)$$

$$E(\bar{\bar{y}}) = \left(\sum_{i=1}^N \frac{u_i \bar{Y}_i}{N} \right)$$

$$E(\bar{\bar{y}}) = \left(\sum_{i=1}^N \frac{M_i \bar{Y}_i}{MN} \right)$$

$$E(\bar{\bar{y}}) = \bar{Y}$$

Q. Find the variance expression of following estimator in two stage sampling when first stage units are not equal in sizes.

$$\bar{\bar{y}} = \sum_{i=1}^n \frac{M_i \bar{y}_i}{nM} = \sum_{i=1}^n \frac{u_i \bar{y}_i}{n}$$

Ans

$$Var(t) = E_1[V_2(t)] + V_1[E_2(t)]$$

$$Var(\bar{\bar{y}}) = E_1[V_2(\bar{\bar{y}})] + V_1[E_2(\bar{\bar{y}})]$$

$$V_1[E_2(\bar{\bar{y}})] = V_1\left[E_2\left(\sum_{i=1}^n \frac{u_i \bar{y}_i}{n}\right)\right]$$

$$V_1[E_2(\bar{\bar{y}})] = V_1\left[\sum_{i=1}^n \frac{u_i \bar{Y}_i}{n}\right]$$

$$V_1[E_2(\bar{\bar{y}})] = V_1[\bar{y}]$$

$$V_1[E_2(\bar{\bar{y}})] = \left(\frac{1}{n} - \frac{1}{N}\right) S_1^2$$

$$E_1[V_2(\bar{\bar{y}})] = E_1\left[V_2\left(\sum_{i=1}^n \frac{u_i \bar{y}_i}{n}\right)\right]$$

$$E_1[V_2(\bar{y})] = E_1\left[\sum_{i=1}^n \frac{u_i^2}{n^2} V_2(\bar{y}_i)\right] = E_1\left[\sum_{i=1}^n \frac{u_i^2}{n^2} \frac{M_i - m_i}{M_i m_i} S_{2i}^2\right]$$

$$E_1[V_2(\bar{y})] = \sum_{i=1}^n \frac{u_i^2}{n^2} \frac{M_i - m_i}{M_i m_i} = \sum_{i=1}^N \frac{u_i^2}{n} \frac{M_i - m_i}{M_i m_i} \frac{S_{2i}^2}{N}$$

$$Var(t) = E_1[V_2(t)] + V_1[E_2(t)]$$

$$Var(t) = \sum_{i=1}^N \frac{u_i^2}{n} \frac{M_i - m_i}{M_i m_i} \frac{S_{2i}^2}{N} + \left(\frac{N-n}{Nn}\right) S_1^2$$

Q. Write the program for Two Stage Sampling using R language considering following data taking n1=3;n2=2. Calculate the sample mean

Cluster-1	125	115	129	134	111
Cluster-2	134	125	142	141	131
Cluster-3	144	143	122	134	126
Cluster-4	114	111	134	131	146
Cluster-5	119	126	122	129	130
Cluster-6	140	125	124	124	115

Ans:

#--Defining Clusters in R----

```
Clu1<-c(125,115,129,134,111)
```

```
Clu2<-c(134,125,142,141,131)
```

```
Clu3<-c(144,143,122,134,126)
```

```
Clu4<-c(114,111,134,131,146)
```

```
Clu5<-c(119,126,122,129,130)
```

```
Clu6<-c(140,125,124,124,115)
```

```
Clus<-data.frame(Clus1,Clu2,Clu3,Clu4,Clu5,Clu6)
```

```
pop<-t(Clus)
```

1st Stage Sample

```
n1=3;n2=2;
```

```
w<-pop[sample(nrow(pop),n1),]
```

```
sc1<-w[1,]
```

```
sc2<-w[2,]
```

```
sc3<-w[3,]
s2_sc1<-sample(sc1,n2)
s2_sc2<-sample(sc2,n2)
s2_sc3<-sample(sc3,n2)
f_s=c(s2_sc1,s2_sc2,s2_sc3)
est=mean(f_s)
```

Q2: Generate the three clusters such that

Cluster-1 from normal distribution with mean zero and variance 1.

Cluster-2 from normal distribution with mean 2 and variance 5.

Cluster-3 from normal distribution with mean 4 and variance 9.

Calculate the sample mean using two stage cluster sampling taking $n_1=2; n_2=15$.

Ans:

```
Clu1<-rnorm(100,0,1)
Clu2<-rnorm(100,2,5)
Clu3<-rnorm(100,4,9)
Clus<-data.frame(Clu1,Clu2,Clu3)
pop<-t(Clus)
#1st Stage Sampling
n1=2;n2=15;
w<-pop[sample(nrow(pop),n1),]
sc1<-w[1,]
sc2<-w[2,]
#2nd Stage Sampling
s2_sc1<-sample(sc1,n2)
s2_sc2<-sample(sc2,n2)
f_s=c(s2_sc1,s2_sc2)
est=mean(f_s)
```

Q.Find Bias of Ratio Estimator in Two Stage Sampling.

Ans:

The Ratio Estimator of mean under two stage sampling when the first stage units are equal.

$$\bar{y}_{r2s} = \frac{\bar{\bar{y}}}{\bar{\bar{x}}} \bar{X}$$

The estimator of mean in two stage sampling is

$$\bar{\bar{y}} = \sum_{i=1}^n \sum_{j=1}^m \frac{y_{ij}}{nm} = \sum_{i=1}^n \frac{1}{n} \bar{y}_i$$

$$V(\bar{\bar{y}}) = \frac{1}{n} \left[\frac{M-m}{Mm} S_2^2 \right] + \left(\frac{N-n}{Nn} \right) S_1^2$$

Some important notations

$$e_0 = (\bar{\bar{y}} - \bar{Y}) / \bar{Y}, e_1 = (\bar{\bar{x}} - \bar{X}) / \bar{X}$$

$$E(e_0) = E(e_1) = 0, E(e_0^2) = V_{02}, E(e_1^2) = V_{20}, E(e_0 e_1) = V_{11}$$

$$V_{02} = \frac{V(\bar{\bar{y}})}{\bar{\bar{Y}}^2} = \frac{1}{n\bar{Y}^2} \left[\frac{M-m}{Mm} S_{2y}^2 \right] + \left(\frac{N-n}{Nn} \right) S_{1y}^2$$

$$V_{20} = \frac{V(\bar{\bar{x}})}{\bar{\bar{X}}^2} = \frac{1}{n\bar{X}^2} \left[\frac{M-m}{Mm} S_{2x}^2 \right] + \left(\frac{N-n}{Nn} \right) S_{1x}^2$$

$$V_{11} = \frac{1}{\bar{Y}\bar{X}} \left\{ \frac{1}{nN} \sum_{i=1}^N \left(\frac{M-m}{Mm} \right) (\rho_{xy2i} S_{y2i} S_{x2i}) + \left(\frac{N-n}{Nn} \right) (\rho_{xy1} S_{y1} S_{x1}) \right\}$$

$$S_{y2i}^2 = \frac{1}{M_i - 1} \sum_{j=1}^{M_i} (y_{ij} - \bar{Y}_i)^2$$

$$\rho_{xy2i} = \frac{S_{xy2i}}{S_{x2i} S_{y2i}}$$

$$\bar{\bar{y}} = \bar{Y}(1 + e_0), \bar{\bar{x}} = \bar{X}(1 + e_1)$$

$$\bar{y}_{r2s} = \frac{\bar{\bar{y}}}{\bar{\bar{x}}} \bar{X}, \bar{y}_{r2s} = \frac{\bar{Y}(1 + e_0)}{\bar{X}(1 + e_1)} \bar{X}$$

Bias of Ratio Estimator Under Two Stage Sampling

$$\bar{y}_{r2s} = \bar{Y}(1 + e_0)(1 + e_1)^{-1} \Rightarrow \bar{y}_{r2s} \approx \bar{Y}(1 + e_0)(1 - e_1 + e_1^2)$$

$$\bar{y}_{r2s} \approx \bar{Y}(1 - e_1 + e_1^2 + e_0 - e_0 e_1 + e_0 e_1^2)$$

$$\bar{y}_{r2s} \approx \bar{Y} + \bar{Y}(e_0 - e_1 + e_1^2 - e_0 e_1 + e_0 e_1^2)$$

$$\bar{y}_{r2s} - \bar{Y} \approx \bar{Y}(e_0 - e_1 + e_1^2 - e_0 e_1)$$

$$E(\bar{y}_{r2s} - \bar{Y}) \approx \bar{Y}E(e_0 - e_1 + e_1^2 - e_0 e_1)$$

$$Bias(\bar{y}_{r2s}) \approx \bar{Y}(V_{20} - V_{11})$$

Q.Find MSE of Ratio Estimator in Two Stage Sampling.

Ans:

The Ratio Estimator of mean under two stage sampling when the first stage units are equal

$$\bar{y}_{r2s} = \frac{\bar{\bar{y}}}{\bar{\bar{x}}} \bar{X}$$

Notations

$$E(e_0) = E(e_1) = 0, E(e_0^2) = V_{02}, E(e_1^2) = V_{20}, E(e_0 e_1) = V_{11}$$

$$e_0 = (\bar{y} - \bar{Y}) / \bar{Y}, e_1 = (\bar{x} - \bar{X}) / \bar{X}$$

$$\bar{\bar{y}} = \bar{Y}(1 + e_0), \bar{\bar{x}} = \bar{X}(1 + e_1)$$

$$\bar{y}_{r2s} = \frac{\bar{\bar{y}}}{\bar{\bar{x}}} \bar{X}$$

$$\bar{y}_{r2s} = \frac{\bar{Y}(1 + e_0)}{\bar{X}(1 + e_1)} \bar{X}$$

MSE of Ratio Estimator Under Two Stage Sampling

$$\bar{y}_{r2s} = \bar{Y}(1 + e_0)(1 + e_1)^{-1} \Rightarrow \bar{y}_{r2s} \approx \bar{Y}(1 + e_0)(1 - e_1)$$

$$\bar{y}_{r2s} \approx \bar{Y}(1 + e_0 - e_1 - e_0 e_1) \Rightarrow \bar{y}_{r2s} \approx \bar{Y} + \bar{Y}(e_0 - e_1)$$

$$E(\bar{y}_{r2s} - \bar{Y})^2 \approx \bar{Y}^2 E(e_0 - e_1)^2$$

$$E(\bar{y}_{r2s} - \bar{Y})^2 \approx \bar{Y}^2 E(e_0^2 + e_1^2 - 2e_0 e_1)$$

$$MSE(\bar{y}_{r2s}) \approx \bar{Y}^2 (V_{02} + V_{20} - 2V_{11})$$

Q.Find Bias of Product Estimator in Two Stage Sampling.

Ans:

The Product Estimator of mean under two stage sampling when the first stage units are equal

$$\bar{y}_{p2s} = \frac{\bar{\bar{y}} \bar{\bar{x}}}{\bar{X}}$$

Some important notations

$$E(e_0) = E(e_1) = 0, E(e_0^2) = V_{02}$$

$$E(e_1^2) = V_{20}, E(e_0 e_1) = V_{11}$$

$$e_0 = (\bar{y} - \bar{Y}) / \bar{Y}, e_1 = (\bar{x} - \bar{X}) / \bar{X}$$

$$\bar{\bar{y}} = \bar{Y}(1 + e_0), \bar{\bar{x}} = \bar{X}(1 + e_1)$$

$$\bar{y}_{p2s} = \frac{\bar{\bar{y}} \bar{\bar{x}}}{\bar{X}}$$

$$\bar{y}_{p2s} = \frac{\bar{Y}(1 + e_0)(1 + e_1)}{\bar{X}} \bar{X}$$

Bias of Ratio Estimator Under Two Stage Sampling

$$\bar{y}_{p2s} = \bar{Y}(1+e_0)(1+e_1) \Rightarrow \bar{y}_{p2s} = \bar{Y}(1+e_0+e_1+e_0e_1)$$

$$\bar{y}_{p2s} = \bar{Y} + \bar{Y}(e_0+e_1+e_0e_1)$$

$$E(\bar{y}_{p2s} - \bar{Y}) = \bar{Y}E(e_0+e_1+e_0e_1)$$

$$\text{Bias}(\bar{y}_{p2s}) = \bar{Y}(V_{11})$$

Q.Find MSE of Product Estimator in Two Stage Sampling

Ans:

The Product Estimator of mean under two stage sampling when the first stage units are equal

$$\bar{y}_{p2s} = \frac{\bar{\bar{y}}\bar{\bar{x}}}{\bar{X}}$$

Notations

$$E(e_0) = E(e_1) = 0,$$

$$E(e_0^2) = V_{02}, E(e_1^2) = V_{20}$$

$$E(e_0e_1) = V_{11}$$

$$e_0 = (\bar{\bar{y}} - \bar{Y})/\bar{Y}$$

$$e_1 = (\bar{\bar{x}} - \bar{X})/\bar{X}$$

$$\bar{\bar{y}} = \bar{Y}(1+e_0), \bar{\bar{x}} = \bar{X}(1+e_1)$$

$$\bar{y}_{p2s} = \frac{\bar{\bar{y}}\bar{\bar{x}}}{\bar{X}}, \bar{y}_{p2s} = \frac{\bar{Y}(1+e_0)(1+e_1)}{\bar{X}} \bar{X}$$

MSE of Product Estimator

$$\bar{y}_{p2s} = \bar{Y}(1+e_0)(1+e_1)$$

$$\bar{y}_{p2s} = \bar{Y}(1+e_0+e_1+e_0e_1)$$

$$\bar{y}_{p2s} \approx \bar{Y} + \bar{Y}(e_0+e_1)$$

$$E(\bar{y}_{p2s} - \bar{Y})^2 \approx \bar{Y}^2 E(e_0+e_1)^2 \approx \bar{Y}^2 E(e_0^2+e_1^2+2e_0e_1)$$

$$MSE(\bar{y}_{p2s}) \approx \bar{Y}^2 (V_{02} + V_{20} + 2V_{11})$$

Ranked set sampling

Introduction:

Ranked set sampling (RSS) is an alternative to simple random sampling that can sometimes offer large improvements in precision. McIntyre (1952) introduced the basic concept of ranked set sampling in order to estimate the population means of pasture and forage yields. The RSS procedure is elaborated as follows

First, a simple random sample of size k is drawn from the population and the k sampling units are ranked with respect to the variable of interest, say X , by judgment without actual measurement. Then the unit with rank 1 is identified and taken for the measurement of X . Next, another simple random sample of size k is drawn and the units of the sample are ranked by judgment, the unit with rank 2 is taken for the measurement of X and the remaining units are discarded. Then the unit with rank 2 is identified and taken for the measurement of X . Then another sample of size 3.

$x_{1(1)1}$	$x_{1(2)1}$	$x_{1(3)1}$	$x_{1(4)1}$	$x_{1(5)1}$	$x_{1(6)1}$
$x_{2(1)1}$	$x_{2(2)1}$	$x_{2(3)1}$	$x_{2(4)1}$	$x_{2(5)1}$	$x_{2(6)1}$
$x_{3(1)1}$	$x_{3(2)1}$	$x_{3(3)1}$	$x_{3(4)1}$	$x_{3(5)1}$	$x_{3(6)1}$
$x_{4(1)1}$	$x_{4(2)1}$	$x_{4(3)1}$	$x_{4(4)1}$	$x_{4(5)1}$	$x_{4(6)1}$
$x_{5(1)1}$	$x_{5(2)1}$	$x_{5(3)1}$	$x_{5(4)1}$	$x_{5(5)1}$	$x_{5(6)1}$
$x_{6(1)1}$	$x_{6(2)1}$	$x_{6(3)1}$	$x_{6(4)1}$	$x_{6(5)1}$	$x_{6(6)1}$

Ranking with Auxiliary Variable

In a case when we are using the auxiliary information (or variable) for ranking the variable of interest, then ranking error depend on the correlation. The performance of the RSS estimator is actually based on the ranking of the sample units in variable of interest

Q. Prove that sample mean is an unbiased estimator of mean under ranked set sampling

Ans:

The mean Estimator in RSS is

$$\bar{X}^* = \frac{\sum_{i=1}^k X_{(i)}^*}{k}$$

The Expected Value is

$$E(\bar{X}^*) = \frac{\sum_{i=1}^k E(X_{(i)}^*)}{k}$$

We have assumed perfect rankings. $X_{(i)}^*$ is distributed like the i^{th} order statistic from a continuous distribution with p.d.f. $f(x)$ and c.d.f. $F(x)$.

$$E\left(X_{(i)}^*\right)=\int_{-\infty}^{\infty} x \frac{k!}{(i-1)(k-i)!}\left[F(x)\right]^{i-1}\left[1-F(x)\right]^{k-i} f(x) dx$$

$$E\left(\bar{X}^*\right)=\frac{\sum_{i=1}^k E\left(X_{(i)}^*\right)}{k}$$

$$E\left(\bar{X}^*\right)=\frac{1}{k} \sum_{i=1}^k\left\{\int_{-\infty}^{\infty} k x \frac{(k-1)!}{(i-1)(k-i)!}\left[F(x)\right]^{i-1}\left[1-F(x)\right]^{k-i} f(x) dx\right\}$$

$$E\left(\bar{X}^*\right)=\frac{1}{k} \sum_{i=1}^k\left\{\int_{-\infty}^{\infty} k x\binom{k-1}{i-1}\left[F(x)\right]^{i-1}\left[1-F(x)\right]^{k-i} f(x) dx\right\}$$

$$E\left(\bar{X}^*\right)=\int_{-\infty}^{\infty} x f(x) \sum_{i=1}^k\left[\binom{k-1}{i-1}\left[F(x)\right]^{i-1}\left[1-F(x)\right]^{k-i}\right] dx$$

$$E\left(\bar{X}^*\right)=\int_{-\infty}^{\infty} x f(x) dx=\bar{X}$$

$$E\left(\bar{X}^*\right)=\bar{X}$$

Q. Find variance expression for mean estimator under ranked set sampling.

Ans:

The mean Estimator in RSS is

$$\bar{X}^*=\frac{\sum_{i=1}^k X_{(i)}^*}{k}$$

By definition

$$E\left(X_{(i)}^*-\bar{X}\right)^2=E\left(X_{(i)}^*-\bar{X}_{(i)}^*+\bar{X}_{(i)}^*-\bar{X}\right)^2$$

$$E\left(X_{(i)}^*-\bar{X}\right)^2=E\left(X_{(i)}^*-\bar{X}_{(i)}^*\right)^2+\left(\bar{X}_{(i)}^*-\bar{X}\right)^2$$

$$E\left(X_{(i)}^*-\bar{X}_{(i)}^*\right)^2=V\left(X_{(i)}^*\right)=E\left(X_{(i)}^*-\bar{X}\right)^2-\left(\bar{X}_{(i)}^*-\bar{X}\right)^2$$

$$V(\bar{X}^*) = \frac{1}{k^2} \sum_{i=1}^k E(X_{(i)}^* - \bar{X}_{(i)}^*)^2 - \frac{1}{k^2} \sum_{i=1}^k (\bar{X}_{(i)}^* - \bar{X})^2$$

Taking,

$$\begin{aligned} \sum_{i=1}^k E(X_{(i)}^* - \bar{X}_{(i)}^*)^2 &= \sum_{i=1}^k \left\{ \int_{-\infty}^{\infty} k(x - \bar{X})^2 \binom{k-1}{i-1} [F(x)]^{i-1} [1-F(x)]^{k-i} f(x) dx \right\} \\ &= k \int_{-\infty}^{\infty} (x - \bar{X})^2 f(x) \sum_{i=1}^k \left[\binom{k-1}{i-1} [F(x)]^{i-1} [1-F(x)]^{k-i} \right] dx \\ V(\bar{X}^*) &= \frac{1}{k^2} \left[k\sigma^2 - \sum_{i=1}^k (\bar{X}_{(i)}^* - \bar{X})^2 \right] \end{aligned}$$

$$V(\bar{X}^*) = \frac{\sigma^2}{k} - \frac{1}{k^2} \sum_{i=1}^k (\bar{X}_{(i)}^* - \bar{X})^2$$

143. Generate a population of size 1000 following normal distribution with mean 0 and variance 1. Further calculate the sample mean using ranked set sampling taking sample size as 6.

Ans:

```
k=6;rssx=matrix(k,k);
pop=rnorm(1000,0,1)
for(i in 1:k)
{
s=sample(pop,k)
xs=sort(s,decreasing = FALSE)
rssx[i,]=xs
}
rssx_s=diag(rssx)
est_srs=mean(x)
est_rss=mean(rssx_s)
rssx
```

Q. Find Bias of Ratio Estimator in Ranked Set Sampling such that

$$E(e_0) = E(e_1) = 0$$

$$E(e_1^2) = \frac{1}{rk} \frac{\sigma_x^2}{\bar{X}^2} - \frac{1}{\bar{X}^2 rk^2} \sum_{i=1}^k (\bar{X}_{(i)}^* - \bar{X})^2 = \eta C_x^2 - D_{x[i]}^2$$

$$E(e_0^2) = \frac{1}{rk} \frac{\sigma_y^2}{\bar{Y}^2} - \frac{1}{\bar{Y}^2 rk^2} \sum_{i=1}^k (\bar{Y}_{(i)}^* - \bar{Y})^2 = \eta C_y^2 - D_{y[i]}^2$$

$$E(e_0 e_1) = \frac{1}{rk} \frac{\sigma_{yx}}{\bar{Y}\bar{X}} - \frac{1}{\bar{Y}\bar{X} rk^2} \sum_{i=1}^k (\bar{X}_{(i)}^* - \bar{X})(\bar{Y}_{(i)}^* - \bar{Y}) = \eta C_{xy} - D_{xy[i]}^2$$

Ans:

The ratio estimator of mean is

$$\bar{y}_{[rss]} = \frac{\bar{Y}^*}{\bar{X}^*} \bar{X}$$

The Variance of Sample Mean Under RSS

$$V(\bar{X}^*) = \frac{\sigma_x^2}{k} - \frac{1}{k^2} \sum_{i=1}^k (\bar{X}_{(i)}^* - \bar{X})^2$$

Notations

$$e_0 = (\bar{Y}^* - \bar{Y}) / \bar{Y}, e_1 = (\bar{X}^* - \bar{X}) / \bar{X}$$

$$E(e_0) = E(e_1) = 0$$

$$E(e_1^2) = \frac{1}{rk} \frac{\sigma_x^2}{\bar{X}^2} - \frac{1}{\bar{X}^2 rk^2} \sum_{i=1}^k (\bar{X}_{(i)}^* - \bar{X})^2 = \eta C_x^2 - D_{x[i]}^2$$

$$E(e_0^2) = \frac{1}{rk} \frac{\sigma_y^2}{\bar{Y}^2} - \frac{1}{\bar{Y}^2 rk^2} \sum_{i=1}^k (\bar{Y}_{(i)}^* - \bar{Y})^2 = \eta C_y^2 - D_{y[i]}^2$$

$$E(e_0 e_1) = \frac{1}{rk} \frac{\sigma_{yx}}{\bar{Y}\bar{X}} - \frac{1}{\bar{Y}\bar{X}rk^2} \sum_{i=1}^k (\bar{X}_{(i)}^* - \bar{X})(\bar{Y}_{(i)}^* - \bar{Y}) = \eta C_{xy} - D_{xy[i]}^2$$

$$\bar{Y}^* = \bar{Y}(1 + e_0)$$

$$\bar{X}^* = \bar{X}(1 + e_1)$$

$$\bar{y}_{[rss]} = \frac{\bar{Y}^*}{\bar{X}^*} \bar{X}$$

$$\bar{y}_{[rss]} = \frac{\bar{Y}(1 + e_0)}{\bar{X}(1 + e_1)} \bar{X}$$

Bias of Ratio Estimator Under RS Sampling

$$\bar{y}_{[rss]} = \bar{Y}(1+e_0)(1+e_1)^{-1} \Rightarrow \bar{y}_{[rss]} \approx \bar{Y}(1+e_0)(1-e_1+e_1^2)$$

$$\bar{y}_{[rss]} \approx \bar{Y}(1-e_1+e_1^2+e_0-e_0e_1+e_0e_1^2)$$

$$\bar{y}_{[rss]} \approx \bar{Y} + \bar{Y}(e_0-e_1+e_1^2-e_0e_1+e_0e_1^2)$$

$$\bar{y}_{[rss]} - \bar{Y} \approx \bar{Y}(e_0-e_1+e_1^2-e_0e_1)$$

$$E(\bar{y}_{[rss]} - \bar{Y}) \approx \bar{Y}E(e_0-e_1+e_1^2-e_0e_1)$$

$$Bias(\bar{y}_{[rss]}) \approx \bar{Y} \begin{bmatrix} \eta C_x^2 - D_{x[i]}^2 \\ -(\eta C_{xy} - D_{xy[i]}) \end{bmatrix}$$

Q.Find MSE of Ratio Estimator in Ranked Set Sampling such that

$$E(e_0) = E(e_1) = 0$$

$$E(e_1^2) = \frac{1}{rk} \frac{\sigma_x^2}{\bar{X}^2} - \frac{1}{\bar{X}^2 rk^2} \sum_{i=1}^k (\bar{X}_{(i)}^* - \bar{X})^2 = \eta C_x^2 - D_{x[i]}^2$$

$$E(e_0^2) = \frac{1}{rk} \frac{\sigma_y^2}{\bar{Y}^2} - \frac{1}{\bar{Y}^2 rk^2} \sum_{i=1}^k (\bar{Y}_{(i)}^* - \bar{Y})^2 = \eta C_y^2 - D_{y[i]}^2$$

$$E(e_0e_1) = \frac{1}{rk} \frac{\sigma_{yx}}{\bar{Y}\bar{X}} - \frac{1}{\bar{Y}\bar{X}rk^2} \sum_{i=1}^k (\bar{X}_{(i)}^* - \bar{X})(\bar{Y}_{(i)}^* - \bar{Y}) = \eta C_{xy} - D_{xy[i]}$$

Ans:

The Ratio Estimator of mean for ranked set sampling is

$$\bar{y}_{[rss]} = \frac{\bar{Y}^*}{\bar{X}^*} \bar{X}$$

Notations

$$e_0 = (\bar{Y}^* - \bar{Y})/\bar{Y}, e_1 = (\bar{X}^* - \bar{X})/\bar{X}$$

$$E(e_0) = E(e_1) = 0,$$

$$E(e_1^2) = \frac{1}{rk} \frac{\sigma_x^2}{\bar{X}^2} - \frac{1}{\bar{X}^2 rk^2} \sum_{i=1}^k (\bar{X}_{(i)}^* - \bar{X})^2 = \eta C_x^2 - D_{x[i]}^2$$

$$E(e_0^2) = \frac{1}{rk} \frac{\sigma_y^2}{\bar{Y}^2} - \frac{1}{\bar{Y}^2 rk^2} \sum_{i=1}^k (\bar{Y}_{(i)}^* - \bar{Y})^2 = \eta C_y^2 - D_{y[i]}^2$$

$$E(e_0 e_1) = \frac{1}{rk} \frac{\sigma_{yx}}{\bar{Y}\bar{X}} - \frac{1}{\bar{Y}\bar{X}rk^2} \sum_{i=1}^k (\bar{X}_{(i)}^* - \bar{X})(\bar{Y}_{(i)}^* - \bar{Y}) = \eta C_{xy} - D_{xy[i]}^2$$

$$\bar{Y}^* = \bar{Y}(1 + e_0)$$

$$\bar{X}^* = \bar{X}(1 + e_1)$$

$$\bar{y}_{[rss]} = \frac{\bar{Y}^*}{\bar{X}^*} \bar{X}$$

$$\bar{y}_{[rss]} = \frac{\bar{Y}(1 + e_0)}{\bar{X}(1 + e_1)} \bar{X}$$

MSE of Ratio Estimator Under RS Sampling

$$\bar{y}_{[rss]} = \bar{Y}(1 + e_0)(1 + e_1)^{-1} \Rightarrow \bar{y}_{[rss]} \approx \bar{Y}(1 + e_0)(1 - e_1)$$

$$\bar{y}_{[rss]} \approx \bar{Y}(1 - e_1 + e_0) \Rightarrow \bar{y}_{[rss]} \approx \bar{Y} + \bar{Y}(e_0 - e_1)$$

$$E(\bar{y}_{[rss]} - \bar{Y})^2 \approx \bar{Y}^2 E(e_0 - e_1)^2$$

$$MSE(\bar{y}_{[rss]}) \approx \bar{Y}^2 E(e_0^2 + e_1^2 - 2e_0 e_1)$$

$$MSE(\bar{y}_{[rss]}) \approx \bar{Y}^2 (\eta C_y^2 - D_{y[i]}^2 + \eta C_x^2 - D_{x[i]}^2 - 2(\eta C_{xy} - D_{xy[i]}))$$

$$MSE(\bar{y}_{[rss]}) \approx \bar{Y}^2 \left[\eta (C_y^2 + C_x^2 - 2C_{xy}) - D_{y[i]}^2 - D_{x[i]}^2 + 2D_{xy[i]} \right]$$

Similarly, the Bias and MSE of Product Estimator Under RSS

The Product Estimator of mean is

$$\bar{y}_{[pss]} = \frac{\bar{Y}^*}{\bar{X}} \bar{X}^*$$

Bias of Product estimator is

$$Bias(\bar{y}_{[pss]}) \approx \bar{Y} [\eta C_{xy} - D_{xy[i]}]$$

MSE of product estimator

$$MSE(\bar{y}_{[pss]}) \approx \bar{Y}^2 \left[\eta(C_y^2 + C_x^2 + 2C_{xy}) - D_{y[i]}^2 - D_{x[i]}^2 - 2D_{xy[i]} \right]$$

Q. Write some types of sampling schemes using ranked set sampling.

Ranked Set Sampling

$x_{1(1)1}$	$x_{1(2)1}$	$x_{1(3)1}$	$x_{1(4)1}$	$x_{1(5)1}$	$x_{1(6)1}$
$x_{2(1)1}$	$x_{2(2)1}$	$x_{2(3)1}$	$x_{2(4)1}$	$x_{2(5)1}$	$x_{2(6)1}$
$x_{3(1)1}$	$x_{3(2)1}$	$x_{3(3)1}$	$x_{3(4)1}$	$x_{3(5)1}$	$x_{3(6)1}$
$x_{4(1)1}$	$x_{4(2)1}$	$x_{4(3)1}$	$x_{4(4)1}$	$x_{4(5)1}$	$x_{4(6)1}$
$x_{5(1)1}$	$x_{5(2)1}$	$x_{5(3)1}$	$x_{5(4)1}$	$x_{5(5)1}$	$x_{5(6)1}$
$x_{6(1)1}$	$x_{6(2)1}$	$x_{6(3)1}$	$x_{6(4)1}$	$x_{6(5)1}$	$x_{6(6)1}$

Extreme Ranked Set Sampling

Samawi et al. (1996) suggested a modified RSS design named as Extreme ranked set sampling (ERSS).

$x_{1(1)1}$	$x_{1(2)1}$	$x_{1(3)1}$	$x_{1(4)1}$	$x_{1(5)1}$	$x_{1(6)1}$
$x_{2(1)1}$	$x_{2(2)1}$	$x_{2(3)1}$	$x_{2(4)1}$	$x_{2(5)1}$	$x_{2(6)1}$
$x_{3(1)1}$	$x_{3(2)1}$	$x_{3(3)1}$	$x_{3(4)1}$	$x_{3(5)1}$	$x_{3(6)1}$
$x_{4(1)1}$	$x_{4(2)1}$	$x_{4(3)1}$	$x_{4(4)1}$	$x_{4(5)1}$	$x_{4(6)1}$
$x_{5(1)1}$	$x_{5(2)1}$	$x_{5(3)1}$	$x_{5(4)1}$	$x_{5(5)1}$	$x_{5(6)1}$
$x_{6(1)1}$	$x_{6(2)1}$	$x_{6(3)1}$	$x_{6(4)1}$	$x_{6(5)1}$	$x_{6(6)1}$

Median Ranked set sampling

$X_{1(1)1}$	$X_{1(2)1}$	$X_{1(3)1}$	$X_{1(4)1}$	$X_{1(5)1}$	$X_{1(6)1}$
$X_{2(1)1}$	$X_{2(2)1}$	$X_{2(3)1}$	$X_{2(4)1}$	$X_{2(5)1}$	$X_{2(6)1}$
$X_{3(1)1}$	$X_{3(2)1}$	$X_{3(3)1}$	$X_{3(4)1}$	$X_{3(5)1}$	$X_{3(6)1}$
$X_{4(1)1}$	$X_{4(2)1}$	$X_{4(3)1}$	$X_{4(4)1}$	$X_{4(5)1}$	$X_{4(6)1}$
$X_{5(1)1}$	$X_{5(2)1}$	$X_{5(3)1}$	$X_{5(4)1}$	$X_{5(5)1}$	$X_{5(6)1}$
$X_{6(1)1}$	$X_{6(2)1}$	$X_{6(3)1}$	$X_{6(4)1}$	$X_{6(5)1}$	$X_{6(6)1}$

$X_{1(1)1}$	$X_{1(2)1}$	$X_{1(3)1}$	$X_{1(4)1}$	$X_{1(5)1}$
$X_{2(1)1}$	$X_{2(2)1}$	$X_{2(3)1}$	$X_{2(4)1}$	$X_{2(5)1}$

$X_{3(1)1}$	$X_{3(2)1}$	$X_{3(3)1}$	$X_{3(4)1}$	$X_{3(5)1}$
$X_{4(1)1}$	$X_{4(2)1}$	$X_{4(3)1}$	$X_{4(4)1}$	$X_{4(5)1}$
$X_{5(1)1}$	$X_{5(2)1}$	$X_{5(3)1}$	$X_{5(4)1}$	$X_{5(5)1}$

Mean in Ranked Set Sampling (RSS)

The mean Estimator in RSS is $\bar{X}^* = \frac{\sum_{i=1}^k X_{(i)}^*}{k}$

In case of r cycles $\bar{X}^* = \frac{\sum_{c=1}^r \sum_{i=1}^k X_{(i)c}^*}{rk}$

Mean in Extreme Ranked Set Sampling (ERSS)

$$\bar{X}_{(ERSS)e}^* = \frac{1}{rk} \sum_{c=1}^r \left[\sum_{i=1}^{k/2} X_{i,(1),c}^* + \sum_{i=1}^{k/2} X_{\frac{k}{2}+i,(k),c}^* \right]$$

$$\bar{X}_{(ERSS)o}^* = \frac{1}{rk} \sum_{c=1}^r \left[\sum_{i=1}^{(k-1)/2} X_{i,(1),c}^* + \sum_{i=1}^{(k-1)/2} X_{\frac{k-1}{2}+i,(k),c}^* + X_{k,\left(\frac{k+1}{2}\right),c}^* \right]$$

Mean in Median Ranked Set Sampling (MRSS)

The mean Estimator in Case of even set size

$$\bar{X}_{(MRSS)e}^* = \frac{1}{rk} \sum_{c=1}^r \left[\sum_{i=1}^{k/2} X_{i,\left(\frac{k}{2}\right),c}^* + \sum_{i=1}^{k/2} X_{\left(\frac{k}{2}+i\right),\frac{k+2}{2},c}^* \right]$$

The mean Estimator in Case of odd set size

$$\bar{X}_{(MRSS)o}^* = \frac{1}{rk} \sum_{c=1}^r \left[\sum_{i=1}^k X_{i,\left(\frac{k+1}{2}\right),c}^* \right]$$

Pair Ranked Set Sampling and Double Ranked Set Sampling are further developments in Ranked set sampling.

Dealing with Non-response

Introduction:

The individuals chosen for the sample are not ready to participate in the survey. This is a type of selection bias Unit/Item nonresponse. The problem of non-response can be dealt with using following methods.

Sub-sampling of non-respondents

Randomized response technique.

Hansen and Hurwitz (1946) technique

Taking a sub sample of non-respondents after the first mail attempt and then enumerating the sub sample of non-respondents by personal interviews

Hansen & Hurwitz (1946) Estimator $\bar{y} = \frac{n_1}{n} \bar{y}_1 + \frac{n_2}{n} \bar{y}_2'$

Q. Prove that following Hansen and Hurwitz estimator is unbiased to population mean.

$$\bar{y} = \frac{n_1}{n} \bar{y}_1 + \frac{n_2}{n} \bar{y}_2'$$

Ans:

Hansen & Hurwitz (1946) Estimator

$$\bar{y}^* = \frac{n_1}{n} \bar{y}_1 + \frac{n_2}{n} \bar{y}_2'$$

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

$$E(\bar{y}^*) = E_1 E_2 \left(\frac{n_1}{n} \bar{y}_1 + \frac{n_2}{n} \bar{y}_2'\right)$$

Hansen & Hurwitz (1946) is Unbiased

$$E(\bar{y}^*) = \left(\frac{n_1}{n} \bar{Y} + \frac{n_2}{n} \bar{Y}\right)$$

$$E(\bar{y}^*) = \bar{Y} \left(\frac{n_1}{n} + \frac{n_2}{n}\right)$$

$$E(\bar{y}^*) = \bar{Y}$$

Q. Find Variance of following Hansen & Hurwitz (1946) Estimator.

$$\bar{y} = \frac{n_1}{n} \bar{y}_1 + \frac{n_2}{n} \bar{y}_2'$$

Ans:

The Hansen & Hurwitz (1946) Estimator is

$$\bar{y}^* = \frac{n_1}{n} \bar{y}_1 + \frac{n_2}{n} \bar{y}_2'$$

Variance

$$\bar{y}^* = \frac{1}{n} (n_1 \bar{y}_1 + n_2 \bar{y}_2) + \frac{n_2}{n} (\bar{y}_2' - \bar{y}_2)$$

$$V(\bar{y}^*) = V\left[\frac{1}{n} (n_1 \bar{y}_1 + n_2 \bar{y}_2) + \frac{n_2}{n} (\bar{y}_2' - \bar{y}_2)\right]$$

$$V\left[\frac{1}{n}(n_1\bar{y}_1 + n_2\bar{y}_2)\right] = \theta S_y^2$$

$$\frac{n_2^2}{n^2} V(\bar{y}_2' - \bar{y}_2) = \frac{n_2^2}{n^2} E(\bar{y}_2' - \bar{y}_2)^2$$

$$\frac{n_2^2}{n^2} E(\bar{y}_2' - \bar{y}_2)^2 = \frac{n_2^2}{n^2} E\left[(\bar{y}_2' - \bar{Y}_2) - (\bar{y}_2 - \bar{Y}_2)\right]^2$$

$$= \frac{n_2^2}{n^2} \left[E(\bar{y}_2' - \bar{Y}_2)^2 + E(\bar{y}_2 - \bar{Y}_2)^2 - 2E(\bar{y}_2' - \bar{Y}_2)(\bar{y}_2 - \bar{Y}_2) \right]$$

$$= \frac{n_2^2}{n^2} \left[E(\bar{y}_2' - \bar{Y}_2)^2 - E(\bar{y}_2 - \bar{Y}_2)^2 \right]$$

$$= \frac{n_2^2}{n^2} \left(\frac{1}{\frac{n_2}{k}} - \frac{1}{n_2} \right) S_{y(2)}^2$$

$$= \frac{n_2}{n^2} (k-1) S_{y(2)}^2$$

$$= \lambda S_{y(2)}^2$$

$$V(\bar{y}^*) = \theta S_y^2 + \lambda S_{y(2)}^2$$

150. Generate a population consisting on 1000 values in R program and calculate Hansen & Hurwitz (1946) Estimator such that $n_1=80; n_2=20; r=10$.

```
N=1000; n=100; n1=80; n2=20; r=10;
pop <- rnorm(N,0,1)
s <- sample(pop,n)
s_r<-pop[1:80]
s_nr<-pop[81:100]
s2<-sample(s_nr,r)
```

```
m<-(n1/n)*mean(s_r)+(n2/n)*mean(s2)
```

Q. Find the expression of Bias of ratio estimator in case of nonresponse on study variable only, where

$$E(e_0^*) = E(e_1) = 0,$$

$$E(e_1^2) = \left(\frac{N-n}{nN} \right) \frac{\sigma_x^2}{\bar{X}^2} = \theta C_x^2$$

$$E(e_0^{*2}) = \frac{1}{\bar{Y}^2} [\theta S_y^2 + \lambda S_{y(2)}^2] = C_y^{*2}$$

$$E(e_0^* e_1) = \theta \frac{S_{yx}}{\bar{Y}\bar{X}} = \theta C_{yx}$$

Ans:

Ratio estimator when non-response occurs in study variable y is as follow

$$\bar{y}_r^* = \frac{\bar{y}^*}{\bar{x}} \bar{X}$$

$$V(\bar{y}^*) = \theta S_y^2 + \lambda S_{y(2)}^2$$

Notations

$$e_0^* = (\bar{y}^* - \bar{Y}) / \bar{Y}, e_1 = (\bar{x} - \bar{X}) / \bar{X}$$

$$E(e_0^*) = E(e_1) = 0,$$

$$E(e_1^2) = \left(\frac{N-n}{nN} \right) \frac{\sigma_x^2}{\bar{X}^2} = \theta C_x^2$$

$$E(e_0^{*2}) = \frac{1}{\bar{Y}^2} [\theta S_y^2 + \lambda S_{y(2)}^2] = C_y^{*2}$$

$$E(e_0^* e_1) = \theta \frac{S_{yx}}{\bar{Y}\bar{X}} = \theta C_{yx}$$

$$\bar{y}^* = \bar{Y}(1 + e_0^*)$$

$$\bar{x} = \bar{X}(1 + e_1)$$

$$\bar{y}_r^* = \frac{\bar{y}^*}{\bar{x}} \bar{X} \longrightarrow \bar{y}_r^* = \frac{\bar{Y}(1 + e_0^*)}{\bar{X}(1 + e_1)} \bar{X}$$

Bias of Ratio Estimator Under Non-response

$$\bar{y}_r^* = \bar{Y}(1+e_0^*)(1+e_1)^{-1} \Rightarrow \bar{y}_r^* \approx \bar{Y}(1+e_0^*)(1-e_1+e_1^2)$$

$$\bar{y}_r^* \approx \bar{Y}(1-e_1+e_1^2+e_0^*-e_0^*e_1+e_0^*e_1^2)$$

$$\bar{y}_r^* - \bar{Y} \approx \bar{Y}(e_0^*-e_1+e_1^2-e_0^*e_1+e_0^*e_1^2)$$

$$\bar{y}_r^* - \bar{Y} \approx \bar{Y}(e_0^*-e_1+e_1^2-e_0^*e_1)$$

$$E(\bar{y}_r^* - \bar{Y}) \approx \bar{Y}E(e_0^*-e_1+e_1^2-e_0^*e_1)$$

$$Bias(\bar{y}_r^*) \approx \theta \bar{Y} [C_x^2 - C_{yx}]$$

Q. Find the expression of MSE of ratio estimator in case of nonresponse on study variable only, where

$$E(e_0^*) = E(e_1) = 0,$$

$$E(e_1^2) = \left(\frac{N-n}{nN} \right) \frac{\sigma_x^2}{\bar{X}^2} = \theta C_x^2$$

$$E(e_0^{*2}) = \frac{1}{\bar{Y}^2} [\theta S_y^2 + \lambda S_{y(2)}^2] = C_y^{*2}$$

$$E(e_0^*e_1) = \theta \frac{S_{yx}}{\bar{Y}\bar{X}} = \theta C_{yx}$$

Ans.

$$\bar{y}_r^* = \frac{\bar{y}^*}{\bar{x}}$$

$$\bar{y}^* = \bar{Y}(1+e_0^*)$$

$$\bar{x} = \bar{X}(1+e_1)$$

$$\bar{y}_r^* = \frac{\bar{y}^*}{\bar{x}}$$

$$\bar{y}_r^* = \frac{\bar{Y}(1+e_0^*)}{\bar{X}(1+e_1)} \bar{X}$$

$$\bar{y}_r^* = \bar{Y}(1+e_0^*)(1+e_1)^{-1} \Rightarrow \bar{y}_r^* \approx \bar{Y}(1+e_0^*)(1-e_1)$$

$$\bar{y}_r^* - \bar{Y} \approx \bar{Y}(e_0^*-e_1)$$

$$E(\bar{y}_r^* - \bar{Y})^2 \approx \bar{Y}^2 E(e_0^* - e_1)^2$$

$$MSE(\bar{y}_r^*) \approx \bar{Y}^2 E(e_0^{*2} + e_1^2 - 2e_0^* e_1)$$

$$MSE(\bar{y}_r^*) \approx \bar{Y}^2 (C_y^{*2} + \theta C_x^2 - 2\theta C_{yx})$$

Similary, Bias and MSE of Product Estimator

Product estimator when non-response occurs in study variable y is follow as

$$\bar{y}_p^* = \frac{\bar{y}^* \bar{X}}{\bar{X}}$$

Bias and MSE of Product Estimator Under Non-response

$$Bias(\bar{y}_p^*) = \theta \bar{Y} [C_{yx}]$$

$$MSE(\bar{y}_p^*) \approx \bar{Y}^2 (C_y^{*2} + \theta C_x^2 + 2\theta C_{yx})$$

Q. Find Bias and MSE of Ratio Estimator in Case of Non-response on Both Variable, where

$$E(e_0^*) = E(e_1^*) = 0,$$

$$E(e_0^{*2}) = \frac{1}{\bar{Y}^2} [\theta S_y^2 + \lambda S_{y(2)}^2] = C_y^{*2}$$

$$E(e_1^{*2}) = \frac{1}{\bar{X}^2} [\theta S_x^2 + \lambda S_{x(2)}^2] = C_x^{*2}$$

$$E(e_0^* e_1^*) = \frac{1}{\bar{Y}\bar{X}} [\theta S_{xy} + \lambda S_{xy(2)}^2] = C_{xy}^*$$

Ans

Ratio estimator when non-response occurs in study variable y is follow as

$$\bar{y}_r^{**} = \frac{\bar{y}^*}{\bar{x}^*} \bar{X}$$

Variance of Hansen & Hurwitz (1946) Estimator

$$V(\bar{y}^*) = \theta S_y^2 + \lambda S_{y(2)}^2$$

$$V(\bar{x}^*) = \theta S_x^2 + \lambda S_{x(2)}^2$$

$$e_0^* = (\bar{y}^* - \bar{Y}) / \bar{Y}$$

$$e_1^* = (\bar{x}^* - \bar{X}) / \bar{X}$$

$$E(e_0^*) = E(e_1^*) = 0,$$

$$E(e_0^{*2}) = \frac{1}{\bar{Y}^2} [\theta S_y^2 + \lambda S_{y(2)}^2] = C_y^{*2}$$

$$E(e_1^{*2}) = \frac{1}{\bar{X}^2} [\theta S_x^2 + \lambda S_{x(2)}^2] = C_x^{*2}$$

$$E(e_0^* e_1^*) = \frac{1}{\bar{Y}\bar{X}} [\theta S_{xy} + \lambda S_{xy(2)}^2] = C_{xy}^*$$

$$\bar{y}^* = \bar{Y}(1 + e_0^*)$$

$$\bar{x}^* = \bar{X}(1 + e_1^*)$$

$$\bar{y}_r^{**} = \frac{\bar{y}^*}{\bar{x}^*} \bar{X}$$

$$\bar{y}_r^{**} = \frac{\bar{Y}(1 + e_0^*)}{\bar{X}(1 + e_1^*)} \bar{X}$$

Bias of Ratio Estimator Under Non-response

$$\bar{y}_r^{**} = \bar{Y}(1 + e_0^*)(1 + e_1^*)^{-1} \Rightarrow \bar{y}_r^{**} \approx \bar{Y}(1 + e_0^*)(1 - e_1^* + e_1^{*2})$$

$$\bar{y}_r^{**} \approx \bar{Y}(1 - e_1^* + e_1^{*2} + e_0^* - e_0^* e_1^* + e_0^* e_1^{*2})$$

$$\bar{y}_r^{**} - \bar{Y} \approx \bar{Y}(e_0^* - e_1^* + e_1^{*2} - e_0^* e_1^* + e_0^* e_1^{*2})$$

$$\bar{y}_r^{**} - \bar{Y} \approx \bar{Y}(e_0^* - e_1^* + e_1^{*2} - e_0^* e_1^*)$$

$$E(\bar{y}_r^{**} - \bar{Y}) \approx \bar{Y}E(e_0^* - e_1^* + e_1^{*2} - e_0^* e_1^*)$$

$$Bias(\bar{y}_r^{**}) \approx \theta \bar{Y} [C_x^{*2} - C_{yx}^*]$$

Notations

$$\bar{y}^* = \bar{Y}(1 + e_0^*)$$

$$\bar{x}^* = \bar{X}(1 + e_1^*)$$

$$\bar{y}_r^{**} = \frac{\bar{y}^*}{\bar{x}^*} \bar{X}$$

$$\bar{y}_r^{**} = \frac{\bar{Y}(1 + e_0^*)}{\bar{X}(1 + e_1^*)} \bar{X}$$

MSE of Ratio Estimator Under RS Sampling

$$\bar{y}_r^{**} = \bar{Y}(1 + e_0^*)(1 + e_1^*)^{-1} \Rightarrow \bar{y}_r^{**} \approx \bar{Y}(1 + e_0^*)(1 - e_1^*)$$

$$\bar{y}_r^{**} \approx \bar{Y}(1 - e_1^* + e_0^* - e_0^* e_1^*) \Rightarrow \bar{y}_r^{**} - \bar{Y} \approx \bar{Y}(e_0^* - e_1^*)$$

$$E(\bar{y}_r^{**} - \bar{Y})^2 \approx \bar{Y}^2 E(e_0^* - e_1^*)^2$$

$$MSE(\bar{y}_r^{**}) \approx \bar{Y}^2 E(e_0^{*2} + e_1^{*2} - 2e_0^* e_1^*)$$

$$MSE(\bar{y}_r^{**}) \approx \bar{Y}^2 (C_y^{*2} + C_x^{*2} - 2C_{yx}^*)$$

product estimator when non-response occurs in study variable y is follow as

$$\bar{y}_p^{**} = \frac{\bar{y}^* \bar{x}^*}{\bar{X}}$$

Bias and ME of Product Estimator Under Non-response

$$Bias(\bar{y}_p^{**}) = \bar{Y} [C_{yx}^*]$$

$$MSE(\bar{y}_p^{**}) \approx \bar{Y}^2 (C_y^{*2} + C_x^{*2} + 2C_{yx}^*)$$

Randomized Response Technique

Introduction:

This technique is useful to estimate the sensitive characteristics in the population. It was first proposed by S. L. Warner in 1965. Qualitative and quantitative response models. Qualitative response models are used to estimate the proportion of some behavior or occurrence in a population.

For example, to estimate the “proportion of people who smoke cigarette today”

The Warner introduced the following model

$$Z = Yp + (1 - p)(1 - Y)$$

p be the probability to answer the sensitive question and Y the true proportion of those interviewed bearing the sensitive property. Z is the proportion of YES answers.

The Warner model

$$Z = Yp + (1 - p)(1 - Y)$$

It can be transformed as

$$Y = \frac{Z + p - 1}{2p - 1}$$

For Example

Statement 1: " I smoke cigarettes."

Statement 2: "I never smoke cigarettes."

$$Y = \frac{\frac{3}{4} + \frac{1}{6} - 1}{2\left(\frac{1}{6}\right) - 1} = \frac{1}{8}$$

Example 1.

Statement 1: " I have falsified my tax return.“

Statement 2: " I have never falsified my tax return."

The Warner model

$$Z = Yp + (1 - p)(1 - Y)$$

$$\frac{40}{50} = Y\left(\frac{1}{6}\right) + \left(1 - \frac{1}{6}\right)(1 - Y)$$

$$Y = \frac{\frac{4}{5} + \frac{1}{6} - 1}{2\left(\frac{1}{6}\right) - 1}$$

Example 2.

Statement 1: " Have you ever used a sick day leave when you weren't really sick? “

Statement 2: " Have you never used a sick day leave when you weren't really sick?”

The Warner model

$$Z = Yp + (1 - p)(1 - Y)$$

$$\frac{350}{500} = Y\left(\frac{1}{6}\right) + \left(1 - \frac{1}{6}\right)(1 - Y)$$

$$Y = \frac{\frac{7}{10} + \frac{1}{6} - 1}{2\left(\frac{1}{6}\right) - 1}$$

$$Y = 0.2$$

Quantitative Randomized Response Technique.

Y may be the monthly income of the head of a household. Y may be the total value of purchase orders in a year for a company. Y may be the number of times you bunk the class during a semester.

Let S be a scrambling variable independent of Y. The respondent is asked to report a scrambled response for Y given by

$$Z = Y + S$$

Mean of Response S variable is introduced such that

$$E(S) = 0$$

$$E(Z) = E(Y)$$

Variance of Response

$$V(Z) = V(Y) + V(S)$$

Q. Generate a population of size 1000 using normal distribution with mean 10 and standard deviation 2. Further consider sample size n=100. Select 10,000 random samples using the model $Z=S+Y$. The variable S is taken to be normal variate with mean equal to zero and standard deviation equal 1.

- (i) Calculate the following estimator considering SRSWOR

$$t = \bar{z}$$

- (ii) Calculate the variance for above estimators

Ans:

Let S be a scrambling variable independent of Y. The respondent is asked to report a scrambled response for Y given by

$$Z = Y + S$$

Estimation of Mean

N=1000; n=100;

y <- rnorm(N, 10, 2)

```
s <- rnorm(N, 0, 1)
sa <- sample(1:N, n)
sy<-y[sa]; ss=s[sa];
z<-sy+ss
mz=mean(z)
N=1000; n=100;mz=c();
y <- rnorm(N, 10, 2)
s <- rnorm(N, 0, 1)
for(i in 1:10000)
{
sa <- sample(1:N, n)
sy<-y[sa]; ss=s[sa];
mz[i]=mean(sy+ss);
}
mean(mz)
```

Quantitative RRT with Non-Sensitive Auxiliary Variable.

Y may be the monthly income of the head of a household and X may be her current age. Y may be the total value of purchase orders in a year for a company and X may be the total turnover for that company in that year.

Scrambled Response

Let Y be the study variable, a sensitive variable which cannot be observed directly due to respondent bias. Let X be a non sensitive auxiliary variable which has a positive correlation with Y. Let S be a scrambling variable independent of Y and X. The respondent is asked to report a scrambled response for Y given by

$$Z = Y + S$$

Respondent is asked to provide a true response for X.

Mean & Variance of Z

S variable is introduced such that

$$E(S) = 0$$

$$E(Z) = E(Y)$$

$$V(Z) = V(Y) + V(S)$$

$$C_z^2 = C_y^2 + \frac{V(S)}{\bar{Y}^2}$$

Ratio Estimator of Mean for Sensitive Variable

$$\bar{y}_{rs} = \frac{\bar{z}}{\bar{x}} \bar{X}$$

Q. Find Bias & MSE of Ratio Estimator Using RRT.

The Ratio Estimator is

$$\bar{y}_{rrt} = \frac{\bar{z}}{\bar{x}} \bar{X}$$

Notations

$$e_0 = (\bar{z} - \bar{Z}) / \bar{Z}, e_1 = (\bar{x} - \bar{X}) / \bar{X}$$

$$E(e_0^2) = \lambda C_z^2, E(e_1^2) = \lambda C_x^2, E(e_0 e_1) = \lambda C_{zx},$$

$$\text{where, } C_{zx} = \rho_{zx} C_z C_x$$

$$\bar{z} = \bar{Z}(1 + e_0), \bar{x} = \bar{X}(1 + e_1)$$

$$\bar{y}_{rrt} = \frac{\bar{z}}{\bar{x}} \bar{X}$$

$$\bar{y}_{rrt} = \frac{\bar{Z}(1 + e_0)}{\bar{X}(1 + e_1)} \bar{X}$$

Bias of Ratio Estimator Using RRT

$$\text{Bias of Ratio Estimator} = \bar{Y}(1 + e_0)(1 + e_1)^{-1} \approx \bar{Y}(1 + e_0)(1 - e_1 + e_1^2)$$

$$\bar{y}_{rrt} = \bar{Y}(1 + e_0)(1 + e_1)^{-1}$$

$$\text{Bias}(\bar{y}_r) \approx \frac{1-f}{n} \bar{Y} (C_x^2 - \rho_{zx} C_z C_x)$$

For MSE of Ratio Estimator

$$\bar{z} = \bar{Z}(1 + e_0)$$

$$\bar{x} = \bar{X}(1 + e_1)$$

$$\bar{y}_{rrt} = \frac{\bar{z}}{\bar{x}} \bar{X}$$

$$\bar{y}_{rrt} = \frac{\bar{Z}(1 + e_0)}{\bar{X}(1 + e_1)} \bar{X}$$

$$\text{MSE of Ratio Estimator} = \bar{Y}(1 + e_0)(1 + e_1)^{-1} \approx \bar{Y}(1 + e_0)(1 - e_1)$$

$$\bar{y}_{rrt} - \bar{Y} \approx \bar{Y}(e_0 - e_1)$$

$$E(\bar{y}_{rrt} - \bar{Y})^2 \approx \bar{Y}^2 E(e_0^2 + e_1^2 - 2e_0 e_1)$$

$$\text{MSE}(\bar{y}_{rrt}) \approx \bar{Y}^2 \lambda (C_z^2 + C_x^2 - 2\rho_{zx} C_z C_x)$$

Q. Find MSE of Ratio Estimator with Two Auxiliary Variables

ANs;

$$\bar{y}_{rr} = \bar{y} \frac{\bar{X}}{\bar{x}} \frac{\bar{Z}}{\bar{z}}$$

Notations

$$e_0 = (\bar{y} - \bar{Y})/\bar{Y}, e_1 = (\bar{x} - \bar{X})/\bar{X}, e_2 = (\bar{z} - \bar{Z})/\bar{Z}$$

$$E(e_0) = E(e_1) = E(e_2) = 0$$

$$E(e_0^2) = \lambda C_y^2, E(e_1^2) = \lambda C_x^2, E(e_2^2) = \lambda C_z^2, E(e_0 e_1) = \lambda C_{yx},$$

$$\text{where, } C_{yx} = \rho_{yx} C_y C_x$$

$$\bar{y} = \bar{Y}(1 + e_0), \bar{x} = \bar{X}(1 + e_1), \bar{z} = \bar{Z}(1 + e_2)$$

$$\bar{y}_{rr} = \bar{y} \frac{\bar{X}}{\bar{x}} \frac{\bar{Z}}{\bar{z}}$$

$$\bar{y}_{rr} = \frac{\bar{Y}(1 + e_0)}{\bar{X}(1 + e_1)\bar{Z}(1 + e_2)} \bar{X}\bar{Z}$$

MSE of the Estimator

$$\bar{y}_{rr} = \bar{Y}(1 + e_0)(1 + e_1)^{-1}(1 + e_2)^{-1} \approx \bar{Y}(1 + e_0)(1 - e_1)(1 - e_2)$$

$$\bar{y}_{rr} \approx \bar{Y}(1 + e_0)(1 - e_1 - e_2 + e_1 e_2) \approx \bar{Y}(1 + e_0)(1 - e_1 - e_2)$$

$$\bar{y}_{rr} \approx \bar{Y}(1 + e_0 - e_1 - e_2) \Rightarrow \bar{y}_{rr} \approx \bar{Y} + \bar{Y}(e_0 - e_1 - e_2)$$

$$\bar{y}_{rr} - \bar{Y} \approx \bar{Y}(e_0 - e_1 - e_2) \Rightarrow E(\bar{y}_{rr} - \bar{Y})^2 \approx \bar{Y}^2 E(e_0 - e_1 - e_2)^2$$

$$MSE(\bar{y}_{rr}) \approx \bar{Y}^2 E(e_0^2 + e_1^2 + e_2^2 - 2e_0 e_1 - 2e_0 e_2 + 2e_1 e_2)$$

$$MSE(\bar{y}_{rr}) \approx \lambda \bar{Y}^2 (C_y^2 + C_x^2 + C_z^2 - 2C_{xy} - 2C_{yz} + 2C_{xz})$$

$$MSE(\bar{y}_{rr}) \approx \lambda \bar{Y}^2 (C_y^2 + C_x^2 + C_z^2 - 2(C_{xy} + C_{yz} - C_{xz}))$$

Q. Define Regression Estimator with Two Auxiliary Variables, also find variance expression of the estimator.

ANs:

Regression Estimator with Two Auxiliary Variables for population mean

$$\bar{y}_{reg} = \bar{y} + \hat{\beta}_{yx}(\bar{X} - \bar{x}) + \hat{\beta}_{yz}(\bar{Z} - \bar{z})$$

$$\hat{\beta}_{yx} = \frac{S_{yx}}{S_x^2}, \hat{\beta}_{yz} = \frac{S_{yz}}{S_z^2}$$

Notations

$$e_0 = (\bar{y} - \bar{Y})/\bar{Y}, e_1 = (\bar{x} - \bar{X})/\bar{X}, e_2 = (\bar{z} - \bar{Z})/\bar{Z}$$

$$E(e_0) = E(e_1) = E(e_2) = 0$$

$$E(e_0^2) = \lambda C_y^2, E(e_1^2) = \lambda C_x^2, E(e_2^2) = \lambda C_z^2, E(e_0 e_1) = \lambda C_{yx}$$

$$\text{where, } C_{yx} = \rho_{yx} C_y C_x$$

Variance of Regression Estimator

$$\bar{y}_{reg} = \bar{y} + \hat{\beta}_{yx}(\bar{X} - \bar{x}) + \hat{\beta}_{yz}(\bar{Z} - \bar{z})$$

$$\bar{y}_{reg} = \bar{Y}(1 + e_0) + \hat{\beta}_{yx}(\bar{X} - \bar{X}(1 + e_1)) + \hat{\beta}_{yz}(\bar{Z} - \bar{Z}(1 + e_2))$$

$$\bar{y}_{reg} = \bar{Y}(1 + e_0) - \hat{\beta}_{yx}(\bar{X}e_1) - \hat{\beta}_{yz}(\bar{Z}e_2)$$

$$E(\bar{y}_{reg} - \bar{Y})^2 = E[\bar{Y}e_0 - \hat{\beta}_{yx}(\bar{X}e_1) - \hat{\beta}_{yz}(\bar{Z}e_2)]^2$$

$$V(\bar{y}_{reg}) = E[\bar{Y}^2 e_0^2 + \hat{\beta}_{yx}^2 \bar{X}^2 (e_1)^2 + \hat{\beta}_{yz}^2 \bar{Z}^2 (e_2)^2 - 2\bar{Y}\bar{X}\hat{\beta}_{yx}(e_0e_1) - 2\bar{Y}\bar{Z}\hat{\beta}_{yz}(e_0e_2) + 2\bar{X}\bar{Z}\hat{\beta}_{yz}\hat{\beta}_{yx}(e_1e_2)]$$

$$V(\bar{y}_{reg}) = E[\bar{Y}^2 e_0^2 + \hat{\beta}_{yx}^2 \bar{X}^2 (e_1)^2 + \hat{\beta}_{yz}^2 \bar{Z}^2 (e_2)^2 - 2\bar{Y}\bar{X}\hat{\beta}_{yx}(e_0e_1) - 2\bar{Y}\bar{Z}\hat{\beta}_{yz}(e_0e_2) + 2\bar{X}\bar{Z}\hat{\beta}_{yz}\hat{\beta}_{yx}(e_1e_2)]$$

$$V(\bar{y}_{reg}) = \lambda[\bar{Y}^2 C_y^2 + \hat{\beta}_{yx}^2 \bar{X}^2 C_x^2 + \hat{\beta}_{yz}^2 \bar{Z}^2 C_z^2 - 2\bar{Y}\bar{X}\hat{\beta}_{yx} C_{yx} - 2\bar{Y}\bar{Z}\hat{\beta}_{yz} C_{yz} + 2\bar{X}\bar{Z}\hat{\beta}_{yz}\hat{\beta}_{yx} C_{xz}]$$

$$V(\bar{y}_{reg}) = \lambda S_y^2[1 - \rho_{yx}^2 - \rho_{yz}^2 - 2\rho_{yx}\rho_{yz}\rho_{xz}]$$

Designs for Hard To Detect Populations

Q. Define Capture Recapture Sampling.

Select a sample, mark the selected units and released back to the population.

Select second sample independently from the population.

Estimate the proportion of the marked units from the second sample.

Suppose

T = Number of animals in the population

K = Number of animals marked on the first visit

n = Number of animals captured on the second visit

k = Number of recaptured animals that were marked

Since the proportion of the marked subjects in the recaptured sample is likely to be about the same as the first sample in the whole population

$$\frac{k}{n} = \frac{K}{T}$$

The estimated population size is

$$T = \frac{nK}{k}$$

Example 1.

When you are interested to estimate the homeless people in your city. An initial Sample of 100 homeless people is selected, marked and released. A sample of size 200 is selected on the second visit and 50 of them were selected in the first visit.

$T=??$, $K=100$, $n=200$, $k=50$

$$T = \frac{nK}{k} = \frac{200 \times 100}{50} = 400$$

Example 2.

In a field study $K=300$ mice are caught in traps, tagged, and released. A few days later the researchers return to the study area and independently capture $n=200$ mice, of which they find that $k=50$ have tags. $T=??$, $K=300$, $n=200$, $k=50$

$$T = \frac{nK}{k} = \frac{200 \times 300}{50}$$

Example 3.

We are interested to estimate the size of a population of turtles in a wildlife preserve. An initial Sample of 20 turtles is selected, marked and released. A sample of size 30 is selected on the second visit and 10 marked turtles found among them.

$$T = \frac{nK}{k} = \frac{20 \times 30}{10} = 60$$

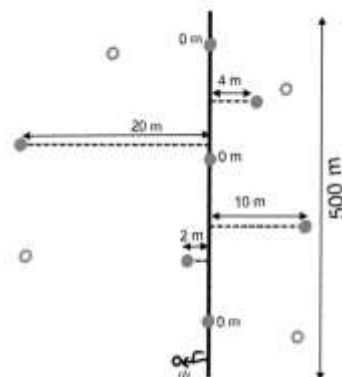
Q. Define Line and Point Transects

Mostly used for animal or plant species. Select the sample of n lines of size L . Observer moves along a selected line and notes the location relative to the line of every individual of the species detected.

Narrow-Strip Method

$$D = \frac{y_0}{2w_0L}$$

$$Est.Total = AD = \frac{Ay_0}{2w_0L}$$



Example

On a line transect of length $L = 100$ meters, a total of $y = 18$ birds were detected at the following distances (in meters) from the transect line

0, 0, 1, 3, 7, 11, 11, 12, 15, 15, 18, 19, 21, 23, 28, 33, 34, 44.

It is desired to estimate the density of birds in the study region. $w_0 = 20$

$$D = \frac{y_0}{2w_0L}$$

Sites are selected as the lines are selected in line transects.

Observations are obtained on selected sites.

Estimation is same as in line transects

163. Explain Adaptive Cluster Sampling

When the selection procedure depends on the observations during the survey

Select an initial sample of size n with a suitable design.

Observe the selected units for a specified condition.

If any of the initially selected unit satisfied the pre-defined condition, its adjacent neighboring units will be sampled and investigated.

Adaptive Cluster Sampling:

- When the selection procedure depends on the observations during the survey.
- Select an initial sample of size n with a suitable design.
- Observe the selected units for a specified condition.
- If any of the initially selected unit satisfied the pre-defined condition, its adjacent neighboring units will be sampled and investigated.

The sample mean is

$$\bar{w}_y = \frac{1}{n} \sum_{i=1}^n w_{yi}$$

