BHARATHIDASAN UNIVERSITY

Tiruchirappalli- 620024

Tamil Nadu, India.

Programme: M.Sc. Statistics

Course Title: R Programming

Course Code: 23ST05CC

Unit-III

Tables and Charts

Dr. T. Jai Sankar
Associate Professor and Head
Department of Statistics

Ms. I. Angel Agnes Mary
Guest Faculty
Department of Statistics

UNIT – III

Descriptive Statistics

1. Measures of Central Tendency (Mean, Median and Mode)

We have registered the speed of 13 cars are given below.

Speed: 99, 86, 87, 88, 111, 86, 103, 87, 94, 78, 77, 85, 86

Calculate Mean, Median and Mode.

Aim

To calculate the Mean, Median and Mode for the give speed of car data.

Procedure

Mean

- Compute the sum of all values.
- Compute divide the sum by the number of values
- Mean formula is given by.

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Median

- To sorted all the values.
- To identify the value in the middle is given by the formula.
- Median formula is given by.

$$Median = \left(\frac{n+1}{2}\right)^{th} Value$$

Mode

- To identify the value that appears the most number of times.

Calculation

Mean

To sum the all values,

99+86+87+88+111+86+103+87+94+78+77+85+86 = 1167

To divide the sum by the number of values

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\overline{X} = \frac{1167}{13}$$

=89.77

Median

To sort all the values,

77, 78, 85, 86, 86, 86, 87, 87, 88, 94, 99, 103, 111

To identify the value in the middle is given by the formula.

$$Median = \left(\frac{n+1}{2}\right)^{th} Value$$

$$Median = \left(\frac{13+1}{2}\right)^{th} value$$

=7th Value is 87.

The median is 87.

Mode

To identify the value that appears the most number of times,

99, **86**, 87, 88, 111, **86**, 103, 87, 94, 78, 77, 85, **86** = 86

Mode Value is 86.

Conclusion

Mean = 89.77

Median = 87

Mode = 86

R Coding

 # Import Data

```
speed = c(99, 86, 87, 88, 111, 86, 103, 87, 94, 78, 77, 85, 86)
```

 # Mean, Median and Mode

```
mean = mean(speed)

median = median(speed)

x = table(speed)

Mode = names(x)[which(x==max(x))]
```

R Output

```
> mean

[1] 89.7692

> median

[1] 87

> mode

[1] 86
```

2. Measures of Variability (Range, Variance and Standard Deviation)

We have registered the speed of 7 cars are given below.

Speed: 32, 111, 138, 28, 59, 77, 97

Calculate Range, Variance and Standard Deviation.

Aim

To calculate the Range, Variance and Standard Deviation for the give speed of car data.

Procedure

- To find the range, all you need to do is subtract the smallest number in the set from the largest number.

  Range = large - small

- To calculate the variance you have to do as follows:

$$s^2 = \frac{1}{n-1}\left(\sum (x-\bar{x}\right)^2$$

  - Find the mean.
  - For each value: find the difference from the mean.
  - For each difference: find the square value.
  - The variance is the average number of these squared differences.

- The formula to find the standard deviation is the square root of the variance.

$$s = \sqrt{s^2}$$

Calculation

Range

Speed: 32, 111, 138, 28, 59, 77, 97

Large Value = 138

Small Value = 28

Range = large – small

Range = 138 – 28

= 110

Variance

To find the mean,

(32+111+138+28+59+77+97) / 7 = 77.4

For each value: find the difference from the mean:

32 - 77.4 = -45.4
111 - 77.4 = 33.6
138 - 77.4 = 60.6
28 - 77.4 = -49.4
59 - 77.4 = -18.4
77 - 77.4 = - 0.4
97 - 77.4 = 19.6

For each difference: find the square value:

$(-45.4)^2 = 2061.16$
$(33.6)^2 = 1128.96$
$(60.6)^2 = 3672.36$
$(-49.4)^2 = 2440.36$
$(-18.4)^2 = 338.56$
$(- 0.4)^2 = 0.16$
$(19.6)^2 = 384.16$

The variance is the average number of these squared differences:

(2061.16 + 1128.96 + 3672.36 + 2440.36 + 338.56 + 0.16 + 384.16) / (7 - 1) = (10025.72/6)

Variance is 1670.95

Standard Deviation

The formula to find the standard deviation is the square root of the variance:

$\sqrt{1670.953} = 40.88$

Conclusion

Range = 110

Variance = 1670.95

Standard Deviation = 40.88

<u>R Coding</u>

 # Import Data

speed = c(32, 111, 138, 28, 59, 77, 97)

 # Range, Variance and SD

Large_Value = max(speed)

Small_Value = min(speed)

range = Large_Value - Small_Value

r = range(speed)

variance = var(speed)

sd = sd(speed)

R Output

> range

[1] 110

> r

[1] 28 138

> variance

[1] 1670.95

> sd

[1] 40.88

Observation Problem

20 students marks is given below.

Mark : 75, 55, 45, 38, 29, 99, 85, 56, 31, 35, 55, 75, 75, 55, 43, 75, 75, 61, 60, 45.

To find (i) Mean, (ii) Median, (iii) Mode, (iv) Range, (v) Variance and (vi) Standard Deviation.

Statistical Graphs in R

Data visualization with different Charts in R

Data Visualization is the presentation of data in graphical format. It helps people understand the significance of data by summarizing and presenting huge amount of data in a simple and easy-to-understand format and helps communicate information clearly and effectively.

Consider this given Data-set for which we will be plotting different charts:

EMPID	Gender	Age	Sales	BMI	Income
E001	M	34	123	Normal	350
E002	F	40	114	Overweight	450
E003	F	37	135	Obesity	169
E004	M	30	139	Underweight	189
E005	F	44	117	Underweight	183
E006	M	36	121	Normal	80
E007	M	32	133	Obesity	166
E008	F	26	140	Normal	120
E009	M	32	133	Normal	75
E010	M	36	133	Underweight	40

Different Types of Charts for Analyzing & Presenting Data

3. Histogram

Histogram is a graphical representation used to create a graph with bars representing the frequency of grouped data in vector. Histogram is same as bar chart but only difference between them is histogram represents frequency of grouped data rather than data itself.

Aim

To draw a histogram is plotted for Age, Income, and Sales. So these plots in the output shows frequency of each unique value for each attribute.

Procedure

- To compute frequency distribution for Age, Income and Sales for the given data.
- To draw a histogram for Age, Income and Sales.

Calculation

To compute frequency distribution for Age,

Age	Number of Persons
26 – 28	1
28 – 30	1
30 – 32	2
32 – 34	1
34 – 36	2
36 – 38	1
38 – 40	1
40 – 42	0
42 – 44	1

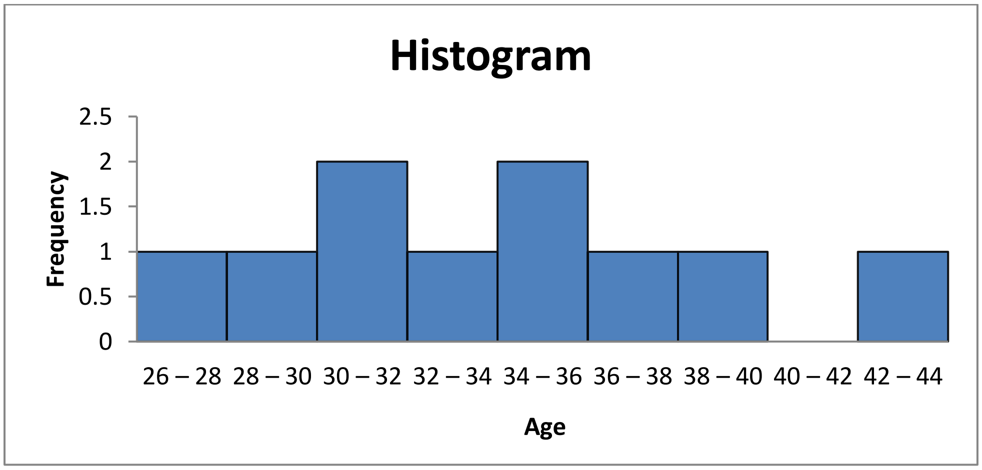
To compute frequency distribution for Sales,

Sales	Number of Persons
below 115	1
115 – 117	1
117 – 119	0
119 – 121	1
121 – 123	1
123 – 125	0
125 – 127	0
127 – 129	0
129 – 131	0
131 – 133	3
133 – 135	1
135 – 137	0
137 – 139	2

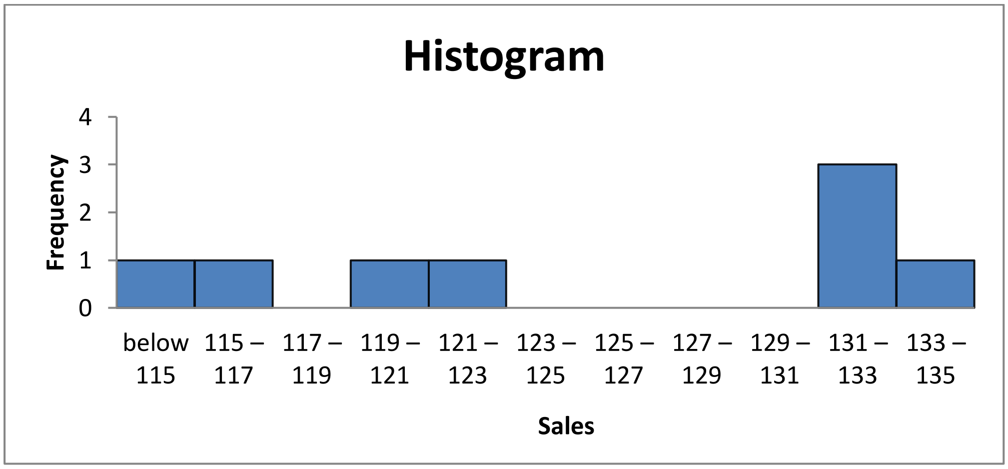
To compute frequency distribution for Income,

Income	Number of Persons
Below 50	1
50 – 100	2
100 – 150	1
150 – 200	3
200 – 250	0
250 – 300	0
300 – 350	1
350 – 400	0
400 – 450	1
Below 50	1

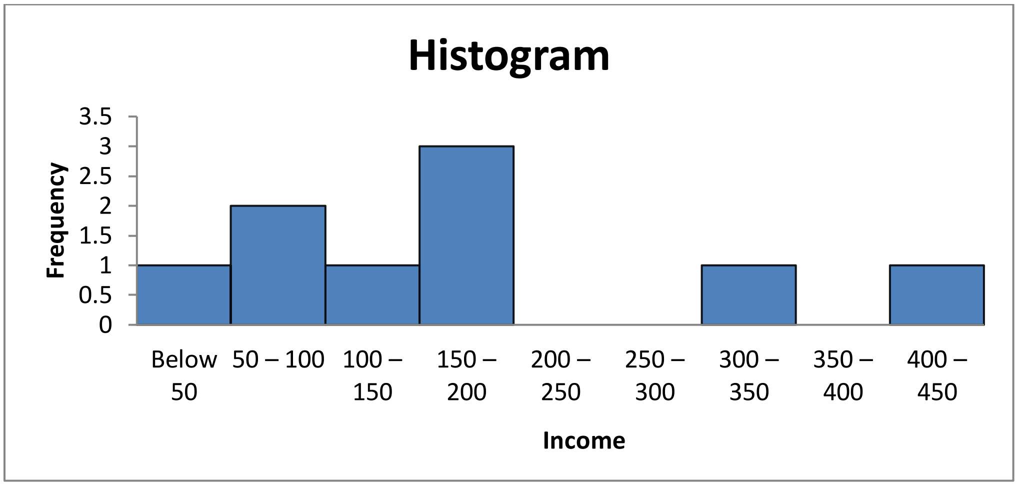
Draw a Histogram for Age Group,



Draw a Histogram for Sales,



Draw a Histogram for Income,



R Coding

 # Given Data

Emp_Id<-c('E001', 'E002', 'E003', 'E004', 'E005', 'E006', 'E007', 'E008', 'E009', 'E010')

Gender<-c('M', 'F', 'F', 'M', 'F', 'M', 'M', 'F', 'M', 'M')

Age<-c(34, 40, 37, 30, 44, 36, 32, 26, 32, 36)

Sales<-c(123, 114, 135, 139, 117, 121, 133, 140, 133, 133)

BMI<-c('Normal', 'Overweight', 'Obesity', 'Underweight', 'Underweight', 'Normal', 'Obesity', 'Normal', 'Normal', 'Underweight')

Income<-c(350, 450, 169, 189, 183, 80, 166, 120, 75, 40)

Data <- data.frame(Emp_Id, Gender, Age, Sales, BMI, Income)

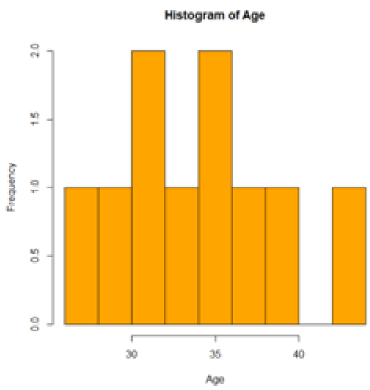
 # creates Histogram for numeric data and Show plot

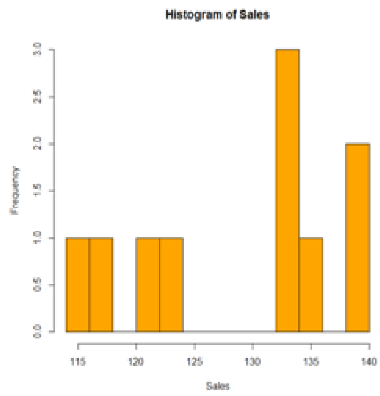
hist(Age, breaks = 10, col = "orange", main = "Histogram of Age", xlab = "Age")

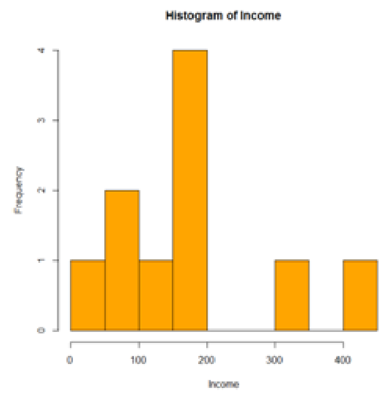
hist(Sales, breaks = 10, col = "orange", main = "Histogram of Sales", xlab = "Sales")

hist(Income, breaks = 10, col = "orange", main = "Histogram of Income", xlab = "Income")

R Output







4. Column Chart

Aim

A column chart is used to show a comparison among different attributes, or it can show a comparison of items over time.

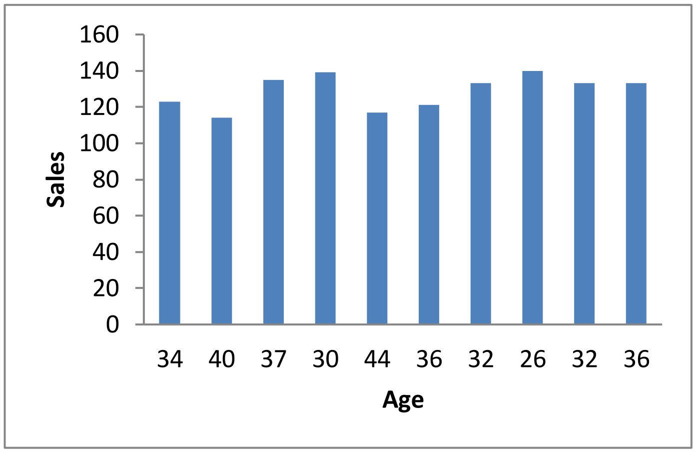
Procedure

- To draw a Column Chart for Age, Income and Sales Individually.
- To draw a Column Chart for Age and Sales.

Calculation

Draw a Column Chart for Age and Sales Comparatively,

Age	*Sales*
34	123
40	114
37	135
30	139
44	117
36	121
32	133
26	140
32	133
36	133



R Coding

 # Bar Plot between 2 attributes

barplot(Sales, main = "Maximum Sales

in a Age",

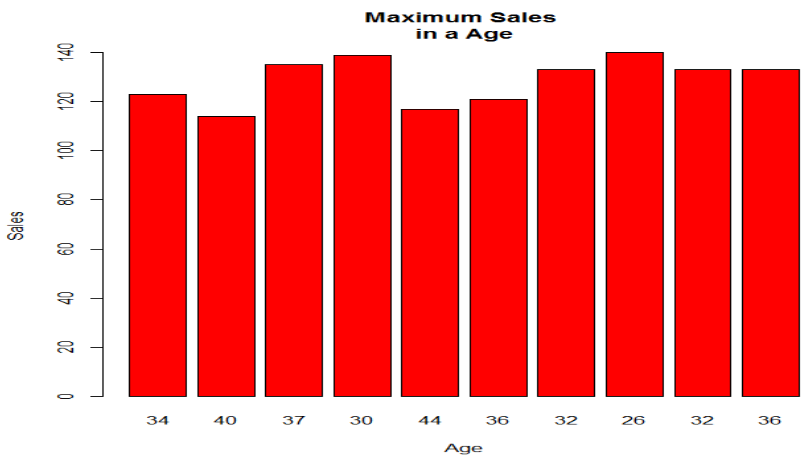
xlab = "Age",

ylab = "Sales",

names.arg= Age,

col = "red")

R Output



5. Box plot chart

A box plot is a graphical representation of statistical data based on the minimum, first quartile, median, third quartile, and maximum. The term "box plot" comes from the fact that the graph looks like a rectangle with lines extending from the top and bottom. Because of the extending lines, this type of graph is sometimes called a box-and-whisker plot. For quantile and median refer to this Quantile and median.

R Coding

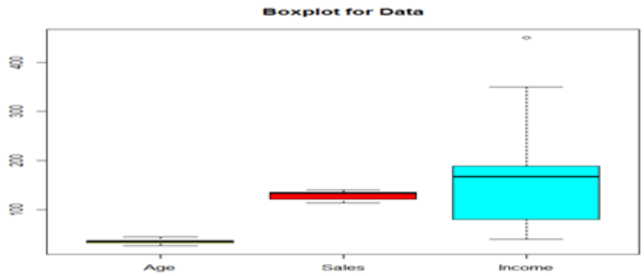
 # For each numeric attribute of dataframe

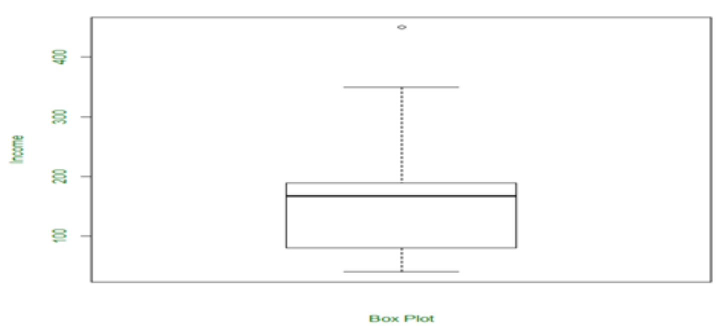
```
Data <- data.frame(Age, Sales, Income)

boxplot(Data, col = c("yellow", "red", "cyan"), main = "Boxplot for Data")
```

 # individual attribute box plot

```
boxplot(Income, xlab = "Box Plot", ylab = "Income",

col.axis = "darkgreen", col.lab = "darkgreen")
```

R Output





6. Pie Chart

A pie chart shows a static number and how categories represent part of a whole the composition of something. A pie chart represents numbers in percentages, and the total sum of all segments needs to equal 100%.

R Coding

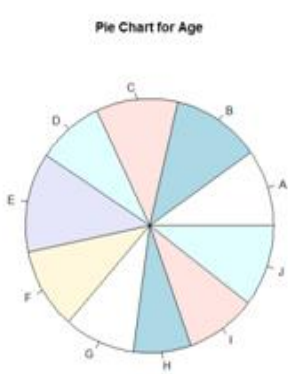
 # For each numeric attribute of dataframe

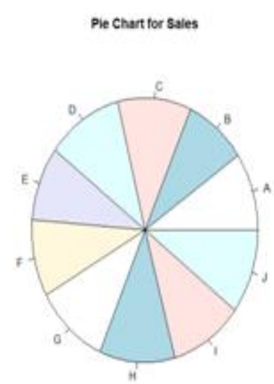
```
Data <- data.frame(Age, Sales, Income)

Names_Age<- c('A', 'B', 'C', 'D', 'E', 'F', 'G', 'H', 'I', 'J')

Names_Sales<-c('A', 'B', 'C', 'D', 'E', 'F', 'G', 'H', 'I', 'J')

Names_Income<- c('A', 'B', 'C', 'D', 'E', 'F', 'G', 'H', 'I', 'J')
```

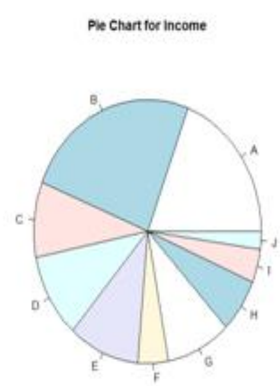
 # individual attribute Pie Chart

```
pie(Data$Age, labels = Names_Age, main = "Pie Chart for Age")

pie(Data$Sales, labels = Names_Sales, main = "Pie Chart for Sales")

pie(Data$Income, labels = Names_Income, main = "Pie Chart for Income")
```

R Output







7. Scatter plot

A scatter chart shows the relationship between two different variables and it can reveal the distribution trends. It should be used when there are many different data points, and you want to highlight similarities in the data set. This is useful when looking for outliers and for understanding the distribution of your data.

R Coding

 # scatter plot between income and age

```
plot(Age, Income, main = "Scatterplot of  Age vs Income", xlab = "Income", ylab = "Age", col= "blue")
```

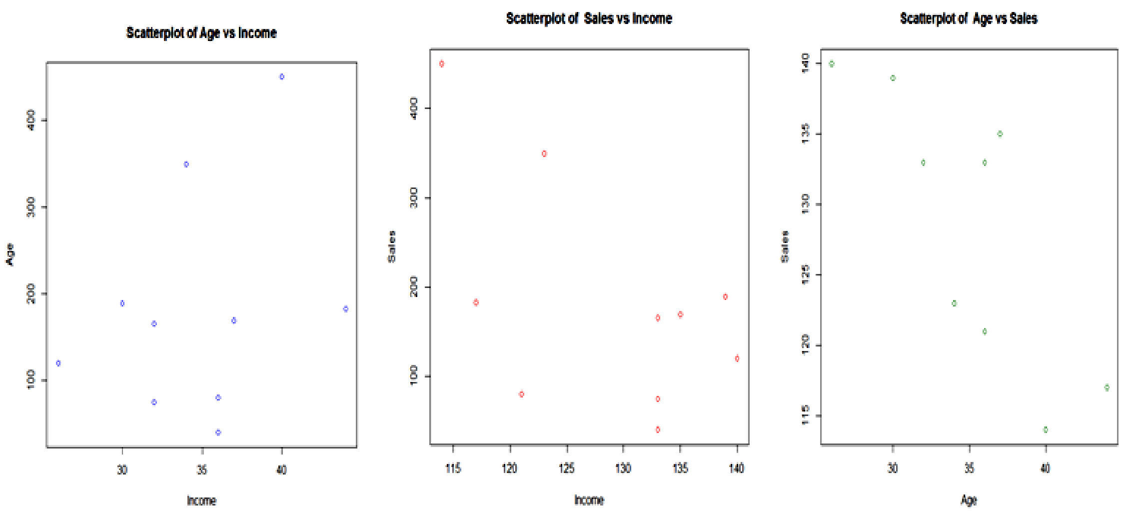
 # scatter plot between income and sales

```
plot(Sales, Income, main = "Scatterplot of  Sales vs Income", xlab = "Income", ylab = "Sales", col= "Red")
```

 # scatter plot between sales and age

```
plot(Age, Sales, main = "Scatterplot of  Age vs Sales", xlab = "Age", ylab = "Sales", col= "darkgreen")
```

R Output



Sampling Theory

8. Simple Random Sampling

Consider a population of 6 units with values 1, 2, 3, 4, 5 and 6. Write down all possible sample of two (without replacement) from this population and verified that the sample mean if an unbiased estimate of the population mean, also calculate its sampling variance and verified that,

1. It agree with the formula for the variance of the sample mean and
2. This variance is less than the variance from sampling with replacement.

Aim

To find all possible sample of two (without replacement) from this population and verified that the sample mean if an unbiased estimate of the population mean, also calculate its sampling variance.

Procedure

- For a given Simple Random Sampling Calculate Population Mean and Variance.

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

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

- For a given Simple Random Sampling Calculate Sample Mean and Variance.
- Calculate Mean and Variance of Simple Random Sampling without Replacement and also drawn from sampling units.

Calculation

Let us assume that the given Population upto units be y and this size of the Population N = 6.

Y	1	2	3	4	5	6	$\Sigma Y_i=21$
Y^2	1	4	9	16	25	36	$\Sigma Y_i^2=91$

We have,

$$\bar{Y}_n = \frac{21}{6}$$

$$= 3.5$$

And

$$\sum_{i=1}^{N}\left(Y_i - \bar{Y}_N\right)^2 = (1 - 3.5)^2 + (2 - 3.5)^2 + (3 - 3.5)^2 + (4 - 3.5)^2 + (5 - 3.5)^2 + (6 - 3.5)^2$$

$$= 117.5$$

Also,

$$\sigma^2 = \frac{N-1}{N}.\bar{Y}_N$$

$$\sigma^2 = \frac{6-1}{6}\times(3.5)$$

$$= 2.917$$

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

$$Var(\bar{Y})_{SRSWR} = \frac{2.917}{2}$$

$$= 1.459$$

$$N_{c_n} = \frac{N!}{n!(N-n)!}$$

$$= \frac{6\times5\times4\times3\times2\times1}{2\times4\times3\times2\times1}$$

$$= 15$$

The sample units table is given by,

Sample (n)	Sample Units	Sample Mean	Sample Variance
1	(1, 2)	1.5	4
2	(1, 3)	2	2.25
3	(1, 4)	2.5	1
4	(1, 5)	3	0.25
5	(1, 6)	3.5	0
6	(2, 3)	2.5	1
7	(2, 4)	3	0.25
8	(2, 5)	3.5	0
9	(2, 6)	4	0.25
10	(3, 4)	3.5	0
11	(3, 5)	4	0.25
12	(3, 6)	4.5	1
13	(4, 5)	4.5	1
14	(4, 6)	5	2.25
15	(5, 6)	5.5	4
Total		52.50	17.5

Sample Mean = $E(\bar{y}) = \frac{52.50}{15}$

$= 3.5$

Sample Variance = $V(\bar{y}) = \frac{17.5}{15}$

$= 1.167$

Result

- Sample Mean (3.5) is Unbiased Estimate of the Population Mean(3.5).
- Sampling Variance:

$$V(\bar{y}) = 1.167$$

$$Var(\bar{Y})_{SRSWR} = 1.459$$

Conclusion

- Sample mean is an unbiased estimate of the population mean.
- Sample variance agrees with the formula for the variance of the sample mean.
- Variance of SRSWOR is less than the Variance of SRSWR.

R coding

#Population Mean - Population Variance

```
yi<-c(1,2,3,4,5,6)

populationmean<-mean(yi)

mean_diff<-(yi-mean(yi))^2

populationvar<-sum((yi-mean(yi))^2)/length(yi)

populationmean

populationvar
```

#All Possible Samples in SRSWOR

```
sample_srswor<-function(N,n)

{

factorial(N)/(factorial(n)*factorial(N-n))

}

sample_srswor(6,2)
```

#SRSWOR_All possible samples of size 2 SRSWR

```
n<-seq(1,15,1)

sample1<-rep(1,15)

sample2<-rep(1,15)

u<-1

for(i in 1:length(yi))

{

for(j in 1:length(yi))

{
```

```
if(i < j)
{
sample1[u]<-yi[i]
sample2[u]<-yi[j]
u<-u+1
}
}
}
samples_mean<-(sample1+sample2)/2
mean_samplemean<-mean(samples_mean)
samples_sample_diff<-(samples_mean-mean_samplemean)^2
samples<-data.frame(cbind(n,sample1,sample2, samples_mean, samples_sample_diff))
samples

variance_srswor<-sum(samples_sample_diff)/15
mean_samplemean
variance_srswor
```

#Variance of SRSWR

```
variance_srswr<-populationvar/2
variance_srswr
```

R Output

> populationmean

[1] 3.5

> populationvar

[1] 2.916667

> **sample_srswor(6,2)**

[1] 15

> **samples**

n	sample1	sample2	Samples_mean	samples_sample_diff
1	1	2	1.5	4.00
2	1	3	2.0	2.25
3	1	4	2.5	1.00
4	1	5	3.0	0.25
5	1	6	3.5	0.00
6	2	3	2.5	1.00
7	2	4	3.0	0.25
8	2	5	3.5	0.00
9	2	6	4.0	0.25
10	3	4	3.5	0.00
11	3	5	4.0	0.25
12	3	6	4.5	1.00
13	4	5	4.5	1.00
14	4	6	5.0	2.25
15	5	6	5.5	4.00

> **mean_samplemean**

[1] 3.5

> **variance_srswor**

[1] 1.166667

> **variance_srswr**

[1] 1.458333

Result

- ➢ Sample mean is an unbiased estimate of the population mean.
- ➢ Sample variance agrees with the formula for the variance of the sample mean.
- ➢ Variance of SRSWOR is less than the Variance of SRSWR.

Observation Problem

Simple Random Sample in selecting 3 units with SRSWOR from a payment having 6 units value 1, 5, 8, 12, 15 and 19. Show that the sample mean is an unbiased estimate of the population mean.

9. Stratified Random Sample

In a survey on the area under a crop total of 186 village in the distributes was divide into 4 strata according to the area of the village from each stratum SRS under proportional allocation where selected village was noted the following is the data from the survey.

Stratum	Stratum Size	Sample	Area under the crop in the village
1	72	8	14, 12, 8, 11, 12, 10, 13, 10
2	53	5	27, 20, 21, 22, 30
3	35	4	36, 47, 52, 61
4	26	3	92, 105, 82

Obtain the estimate of the total area under the crop in the district.

Aim

To find the area of the village from each stratum SRS under proportional allocation, obtain the estimate of the total area under the crop in the district.

Calculation

To fine Mean and Variance of each Stratum,

$$\bar{y}_{n_1} = \frac{14+12+8+11+12+10+13+10}{8}$$

$= 11.25$

$$\bar{y}_{n_2} = \frac{27+20+21+22+30}{5}$$

$= 24$

$$\bar{y}_{n_3} = \frac{36+47+52+61}{4}$$

$= 49$

$$\bar{y}_{n_4} = \frac{92+105+82}{3}$$

$= 93$

$$S_1^2 = \frac{(14^2+12^2+8^2+11^2+12^2+10^2+13^2+10^2)-8(11.25)^2}{8-1}$$

$= 3.64$

$$S_2^2 = \frac{(27^2+20^2+21^2+22^2+30^2)-5(24)^2}{5-1}$$

$= 18.5$

$$S_3^2 = \frac{(36^2+47^2+52^2+61^2)-4(49)^2}{4-1}$$

$= 108.67$

$$S_4^2 = \frac{(92^2+105^2+82^2)-3(93)^2}{3-1}$$

$= 133$

To find Mean and Variance under Proportional Allocation,

Stratum	Stratum Size (N)	Sample Size (n)	Sample Mean	Sample Variance	N* Sample Mean	N* Sample Variance
1	72	8	11.25	3.64	810	262.08
2	53	5	24	18.5	1272	980.5
3	35	4	49	108.67	1715	3803.45
4	26	3	93	133	2418	3458
Total	186	20	177.25	263.81	6215	8504.03

The total area under the crop in the distribution

$$\hat{y} = N \times \frac{\sum \bar{y}}{N}$$

$$\hat{y} = 186 \times \frac{6215}{186}$$

$$= 6215$$

Conclusion

The total area under the crop in the distribution is 6215.

R Coding

#Given Data

stratum1<-c(14,12,8,11,12,10,13,10)

stratum2<-c(27,20,21,22,30)

stratum3<-c(36,47,52,61)

stratum4<-c(92,105,82)

#Stratum Size

population <-c(72,53,35,26)

N<-sum(population)

#Sample Size from each Stratum

sample<-c(length(stratum1),length(stratum2),length(stratum3),length(stratum4))

sample

#Sample Mean and Variance Drawn from each Stratum

mean<-c(mean(stratum1), mean(stratum2), mean(stratum3), mean(stratum4))

variance<-c(var(stratum1), var(stratum2), var(stratum3), var(stratum4))

#Mean of Proportional Allocation Drawn from each Stratum

```
mean_prop<-seq(1,4)
for(i in 1:4)
{
mean_prop [i]<- population[i]* sample_mean[i]
 }
```

#Variance of Proportional Allocation Drawn from each Stratum

```
variance_prop <-seq(1,4)
for(i in 1:4)
{
variance_prop [i]<- population[i]* sample_variance  [i]
 }

output<-data.frame(cbind(population,sample,mean,variance, mean_prop, variance_prop))
output

y_hat<-N*sum(mean_prop /N)
y_hat
```

<u>R Output</u>

Output

stratum_size	sample	mean	variance	mean_prop	variance_prop
72	8	11.25	3.642857	810	262.2857
53	5	24.00	18.500000	1272	980.5000
35	4	49.00	108.666667	1715	3803.3333
26	3	93.00	133.000000	2418	3458.0000

> y_hat

[1] 6215

Result:

The total area under the crop in the distribution is 6215.

Observation Problem

Four stratified sample of units gives that following estimated stratum Mean and Variance.

Stratum	Population	Sample	Stratum Mean	Stratum Variance
1	30	5	35	40
2	50	10	40	55
3	60	15	40	80
4	60	20	55	144