SATHYABAMA
INSTITUTE OF SCIENCE AND TECHNOLOGY
(DEEMED TO BE UNIVERSITY)
Accredited "A" Grade by NAAC | 12B Status by UGC | Approved by AICTE
www.sathyabama.ac.in

SCHOOL OF SCIENCE AND HUMANITIES

DEPARTMENT OF MATHEMATICS

UNIT – I – Introduction to R – SMTA5204

R is a programming language and software environment for statistical analysis, graphics representation and reporting. R was created by Ross Ihaka and Robert Gentleman at the University of Auckland, New Zealand, and is currently developed by the R Development Core Team.

The core of R is an interpreted computer language which allows branching and looping as well as modular programming using functions. R allows integration with the procedures written in the C, C++, .Net, Python or FORTRAN languages for efficiency.

R is freely available under the GNU General Public License, and pre-compiled binary versions are provided for various operating systems like Linux, Windows and Mac.

R is free software distributed under a GNU-style copy left, and an official part of the GNU project called **GNU S**.

Features of R

- R is a well-developed, simple and effective programming language which includes conditionals, loops, user defined recursive functions and input and output facilities.
-
- R has an effective data handling and storage facility,
-
- R provides a suite of operators for calculations on arrays, lists, vectors and matrices.
-
- R provides a large, coherent and integrated collection of tools for data analysis.
-
- R provides graphical facilities for data analysis and display either directly at the computer or printing at the papers.

The variables are assigned with R-Objects and the data type of the R-object becomes the data type of the variable. There are many types of R-objects. The frequently used ones are:

- **Vectors**
- **Lists**
- **Matrices**
- **Arrays**
- **Factors**
- **Data Frames**

1. Vectors

When you want to create vector with more than one element, you should use **c()** function which means to combine the elements into a vector.

```
# Create a vector.
apple <- c('red','green',"yellow")
print(apple)


# Get the class of the vector.
print(class(apple))
```

2. List

A list is an R-object which can contain many different types of elements inside it like vectors, functions and even another list inside it.

```
# Create a list.
list1 <- list(c(2,5,3),21.3,sin)

# Print the list.
print(list1)
```

```
[[1]]
[1] 2 5 3

[[2]]
[1] 21.3

[[3]]
function (x)  .Primitive("sin")
```

Matrices

A matrix is a two-dimensional rectangular data set. It can be created using a vector input to the matrix function.

```
# Create a matrix.
M = matrix( c('a','a','b','c','b','a'), nrow=2,ncol=3,byrow = TRUE)
print(M)
```

```
     [,1] [,2] [,3]
[1,] "a"  "a"  "b"
[2,] "c"  "b"  "a"
```

Arrays

While matrices are confined to two dimensions, arrays can be of any number of dimensions. The array function takes a dim attribute which creates the required number of dimension. In the below example we create an array with two elements which are 3x3 matrices each.

```
# Create an array.
a <- array(c('green','yellow'),dim=c(3,3,2))
print(a)
```

```
, , 1

     [,1]     [,2]     [,3]
[1,] "green"  "yellow" "green"
[2,] "yellow" "green"  "yellow"
[3,] "green"  "yellow" "green"

, , 2

     [,1]     [,2]     [,3]
[1,] "yellow" "green"  "yellow"
[2,] "green"  "yellow" "green"
```

Data Frames

Data frames are tabular data objects. Unlike a matrix in data frame each column can contain different modes of data. The first column can be numeric while the second column can be character and third column can be logical. It is a list of vectors of equal length.

Data Frames are created using the **data.frame()** function.

```
# Create the data frame.
BMI <-      data.frame(
                 gender = c("Male", "Male","Female"),
                height = c(152, 171.5, 165),
                weight = c(81,93, 78),
                Age =c(42,38,26)
                )
print(BMI)

  gender height weight Age
1   Male  152.0     81  42
2   Male  171.5     93  38
3 Female  165.0     78  26
```

Summary

In R, quartiles, minimum and maximum values can be easily obtained by the `summary` command

It gives information on

- minimum,
- maximum
- first quartile
- second quartile (median) and
- third quartile.

Importing Excel files

Spreadsheet (Excel) file data

For reading older Excel files in .xls format, use `gdata` package and function `read.xls()`

Different formats of files can be read in R

- comma-separated values (CSV) data file,
- table file (TXT),
- Spreadsheet (e.g., MS Excel) file,
- files from other software like SPSS, Minitab etc.

Reading Tabular Data Files

Tabular data files are text files with a simple format:

- Each line contains one record.
- Within each record, fields (items) are separated by a one-character delimiter, such as a space, tab, colon, or comma.
- Each record contains the same number of fields.

We want to read a text file that contains a table of data.

`read.table` **function is used and it returns a data frame.**

```
read.table("FileName")
```

UNIT – II – Data Standardizing – SMTA5204

Normal distribution:

A random variable X is said to have a Normal distribution with parameters μ (mean) and σ^2 (variance) if its probability density function is given by the probability law

$$f(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2},-\infty<x<\infty,-\infty<\mu<\infty,\sigma>0$$

Notation: $X \sim N\left(\mu,\sigma^2\right)$ read as X is following normal distribution with mean μ and variance σ^2 are called parameter.

Prove that "For standard normal distribution $N(0,1)$, $M_X(t)=e^{\frac{t^2}{2}}$.

Solution:

Moment generating function of Normal distribution

$$=M_X(t)=E\left[e^{tx}\right]$$

$$=\frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{\infty}e^{tx}e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}dx$$

Put $z=\frac{x-\mu}{\sigma}$ then $\sigma dz=dx$, $-\infty<Z<\infty$

$$\therefore M_X(t)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}e^{t(\sigma z+\mu)-\frac{z^2}{2}}dz$$

$$=\frac{e^{\mu t}}{\sqrt{2\pi}}\int_{-\infty}^{\infty}e^{-\left(\frac{z^2}{2}-t\sigma z\right)}dz$$

$$=\frac{e^{\mu t}}{\sqrt{2\pi}}\int_{-\infty}^{\infty}e^{-\frac{1}{2}(z-t\sigma)^2+\left(\frac{\sigma^2t^2}{2}\right)}dz$$

$$=\frac{e^{\mu t}e^{\frac{\sigma^2t^2}{2}}}{\sqrt{2\pi}}\int_{-\infty}^{\infty}e^{-\frac{1}{2}(z-t\sigma)^2}dz$$

$\because$ the total area under normal curve is unity, we have $\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}e^{-\frac{1}{2}(z-t\sigma)^2}dz=1$

Hence $M_X(t)=e^{\mu t+\frac{\sigma^2t^2}{2}}$ $\therefore$ For standard normal variable $N(0,1)$

$$M_X(t)=e^{\frac{t^2}{2}}$$

X is a normal variate with $mean = 30$ and $S.D = 5$ Find the following $P[26 \le X \le 40]$

Solution:

$$X \sim N(30, 5^2)$$

$$\therefore \mu = 30 \ \& \ \sigma = 5$$

Let $Z = \dfrac{X-\mu}{\sigma}$ be the standard normal variate

$$P[26 \le X \le 40] = P\left[\frac{26-30}{5} \le Z \le \frac{40-30}{5}\right]$$

$$= P[-0.8 \le Z \le 2] = P[-0.8 \le Z \le 0] + P[0 \le Z \le 2]$$

$$= P[0 \le Z\ 0.8] + [0 \le z \le 2]$$

$$= 0.2881 + 0.4772 = 0.7653.$$

The average percentage of marks of candidates in an examination is 45 will a standard deviation of 10 the minimum for a pass is 50%.If 1000 candidates appear for the examination, how many can be expected marks. If it is required, that double that number should pass, what should be the average percentage of marks?

Solution:

Let X be marks of the candidates

Then $X \sim N(42, 10^2)$

Let $z = \dfrac{X-42}{10}$

$$P[X > 50] = P[Z > 0.8]$$

$$= 0.5 - P[0 < z < 0.8]$$

$$= 0.5 - 0.2881 = 0.2119$$

Since 1000 students write the test, nearly 212 students would pass the examination.

If double that number should pass, then the no of passes should be 424.

We have to find z_1, such that $P[Z > z_1] = 0.424$

$$\therefore P[0 < z < z_1] = 0.5 - 0.424 = 0.076$$

From tables, $z = 0.19$

$$\therefore z_1 = \frac{50 - x_1}{10} \Rightarrow x_1 = 50 - 10z_1$$

$$= 50 - 1.9 = 48.1$$

The average mark should be 48 nearly.

In a normal distribution 31% of the items are under 45 and 8% are over 64.Find the mean and standard deviation of the distribution.
Solution:
Let μ be the mean and σ be the standard deviation.
Then $P[X \le 45] = 0.31$ and $P[X \ge 64] = 0.08$

When $X = 45$, $Z = \dfrac{45-\mu}{\sigma} = -z_1$

$\therefore z_1$ is the value of z corresponding to the area $\int_0^{z_1} \phi(z)dz = 0.19$

$\therefore z_1 = 0.495$
$45 - \mu = -0.495\sigma$ ---(1)
When $X = 64$, $Z = \dfrac{64-\mu}{\sigma} = z_2$

$\therefore z_2$ is the value of z corresponding to the area $\int_0^{z_2} \phi(z)dz = 0.42$

$\therefore z_2 = 1.405$
$64 - \mu = 1.405\sigma$ ---(2)
Solving (1) & (2) We get $\mu = 10$ (approx) & $\sigma = 50$ (approx)

Standardization and *Z* Scores

I. Standardization

a. Standardizing scores is the process of converting each raw score in a distribution to a *z* score (or standard deviation units)
   i. Raw Score: the individual observed scores on measured variables

II. *z* Scores (also known as a *standard scores*)

a. Helps to understand where a score lies in relation to other scores in the distribution
   i. Indicates how far above or below the mean a given score in the distribution is in standard deviation units.
      1. For example, if you know that an individual in a sample has a *z* score of .75, you would know that the individual's score was .75 standard deviations above the mean for that sample.

b. Calculated using mean and standard deviation
   i. *z* = (raw score - mean) / standard deviation

c. Using z scores to determine probabilities
   i. You can calculate a z score using either sample data OR population data
      1. You can only calculate percentiles using Appendix A when you know...
         a. The population standard deviation, or
         b. The sample data are normally distributed
      2. z scores let you compare performances on two measures with different scales of measurement
         a. e.g. height and weight, grade point average and standardized test scores

d. z scores can also be calculated for the difference between a sample mean and a population mean, in *standard error* units.

Z score formulas

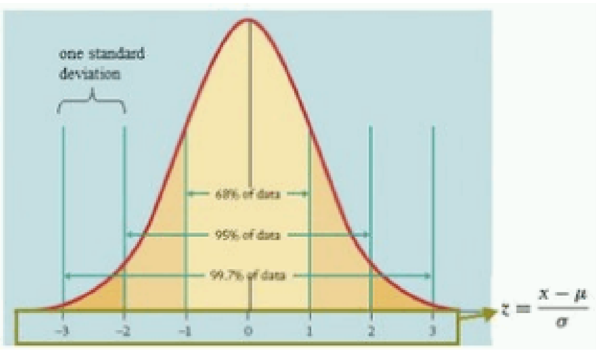
Population data (most common)

$$z = \frac{x - \mu}{\sigma}$$

x = the raw score or the "test score"
μ = population mean
σ = population standard deviation

A z score of 0 indicates that the score is right on the mean.

So, a z score of +1 = 1 SD above the mean. A z score of -1 = 1 SD below the mean. A z score of 0 = the score is right on the mean. See the book for the best pictures, but one is included below.



UNIT – III – Probability Distribution – SMTA5204

Standard Error :
The standard deviation of sampling distribution of a statistic is known as its standard error and is denoted by (S.E)

Test of Significance :
It enable us to decide on the basis of the sample results if the deviation between the observed sample statistic and the hypothetical parameter value is significant or the deviation between two sample statistics is significant.

Null Hypothesis:
A definite statement about the population parameter which is usually a hypothesis of no-difference and is denoted by $H_{o.}$

Alternative Hypothesis:
Any hypothesis which is complementary to the null hypothesis is called an Alternative Hypothesis and is denoted by $H_{1.}$

Errors in Sampling:
Type I and Type II errors.
Type I error : Rejection of H_0 when it is true.
Type II error : Acceptance of H_0 when it is false.

Two types of errors occurs in practice when we decide to accept or reject a lot after examining a sample from it. They are Type 1 error occurs while rejecting H_o when it is true. Type 2 error occurs while accepting H_o when it is wrong.

One tail and two tailed test:
A test of any statistical hyposthesis where the alternate hypothesis is one tailed(right tailed/ left tailed) is called one tailed test.
For the null hypothesis H_0 if $\mu = \mu_0$ then.
$H_1 = \mu > \mu_0$ (Right tail)
$H_1 = \mu < \mu_0$ (Left tail)
$H_1 = \mu \# \mu_0$ (Two tail test)
Large sample (n>30) : Z test.

LARGE SAMPLES

TEST OF SIGNIFICANCE OF LARGE SAMPLES
If the size of the sample n>30 then that sample is called large sample.

<u>Type 1. Test of significance for single proportion</u>

Let p be the sample proportion and P be the population proportion, we use the statistic Z= (p-P) / $\sqrt{(PQ/n)}$

Limits for population proportion P are given by p±3$\sqrt{(PQ/n)}$
Where q = 1-p

1. A manufacture claimed that at least 95% of the equipment which he supplied to a factory conformed to specifications. An examination of a sample of 200 pieces of equipment revealed that 18 were faulty. tEst his claim at 5% level of significance.

Solution:

Calculated Z value = 2.59
Tabulated Value = 1.96 (at 5% level of significance) Calculated value > Tabulated value, Reject Ho (Null hypothesis)

<u>Type II Test of significance for difference of proportions</u>

Let n_1 and n_2 are the two sample sizes and sample proportions are p_1 and p_2

$Z = \frac{(p_1 - p_2)}{\sqrt{pq(1/n_1 + 1/n_2)}}$ where p= ($n_1p_1+n_2p_2/n_1+n_2$ and q=1-p

Proplems

1. Before an increase in excise duty on tea, 800 persons out of a sample of 1000 persons were found to be tea drinkers. After an increase in duty 800 people were tea drinkers in the sample of 1200 people. Using standard error of proportions state whether there is a significant decrease in the consumption of tea after the increase in the excise duty.

Solution:

Calculated Z value = 6.972
Tabulated value at 5% (one tail) = 1.645
Calculated value > Tabulated value, Reject Ho (Null hypothesis)

Type III Test of significance for single Mean

$z = \bar{x} - \mu / (\sigma/\sqrt{n})$ where $\bar{x}$ is the same mean
μ is the population mean, s is the population S.D.
n is the sample size.

The values of $\bar{x} \pm 1.96\ (\sigma/\sqrt{n})$ are called 95% confidence limits for the mean of the population corresponding to the given sample.

The values of $\bar{x} \pm 2.58\ (\sigma/\sqrt{n})$ are called 99% confidence limits for the mean of the population corresponding to the given sample.

PROBLEMS

1. A sample of 900 members has a mean of 3.4 cms and SD 2.61 cms. Is the sample from a large population of mean is 3.25 cm and SD 2.61 cms. If the population is normal and its mean is unknown find the 95% confidence limits of true mean.

Solution:
Calculated Z value = 1.724
Tabulated value at 5% = 1.96
Calculated value < Tabulated value, Accept Ho (Null hypothesis)
Limits (3.57, 3.2295)

Type IV Test of significance for Difference of means

$Z = (\bar{x}_1 - \bar{x}_2) / \sqrt{(\sigma_1^2/n_1) + (\sigma_2^2/n_2)}$

PROBLEMS

1. The means of 2 large samples of 1000 and 2000 members are 67.5 inches and 68 inches respectively. Can the samples be regarded as drawn from the same population of SD 2.5 inches.

Solution:
Calculated Z value = 5.16
Tabulated value at 5% = 1.96
Calculated value > Tabulated value, Reject Ho (Null hypothesis)

Central Limit Theorem (Liapounoff's Form)

If $X_1, X_2, \cdots X_n, \cdots$, be a sequence of independent RVs with $E(X_i) = \mu_i$ and $Var(X_i) = \sigma_i^2, i = 1, 2, \cdots$, and if $S_n = X_1 + X_2 + \cdots X_n$, then under certain general conditions, S_n follows a normal distribution with mean $\mu = \sum_{i=1}^{n} \mu_i$ and variance

$\sigma^2 = \sum_{i=1}^{n} \sigma_i^2$ as n tends to infinity.

Central Limit Theorem (Lindberg-Levy's Form)

If $X_1, X_2, \cdots, X_n, \cdots$, be a sequence of independent identically distributed RVs with $E(X_i) = \mu$ and $\mathrm{Var}(X_i) = \sigma^2, i = 1, 2, \cdots$, and if $S_n = X_1 + X_2 + \cdots + X_n$, then under certain general conditions, S_n follows a normal distribution with mean $n\mu$ and variance $n\sigma^2$ as n tends to infinity.

Corollary

If $\overline{X} = \frac{1}{n}(X_1 + X_2 + \cdots + X_n)$, then $E(\overline{X}) = \mu$ and $\mathrm{Var}(\overline{X}) = \frac{1}{n^2}(n\sigma^2) = \frac{\sigma^2}{n}$

$\therefore \overline{X}$ follows $N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$ as $n \to \infty$

If $X_1, X_2, \cdots X_n$ are Poisson variates with parameter $\lambda = 2$, use the central limit theorem to estimate $P(120 \le S_n \le 160)$, where $S_n = X_1 + X_2 + \cdots X_n$ and $n = 75$.

$E(X_i) = \lambda = 2$ and $\mathrm{Var}(X_i) = \lambda = 2$

By CLT, S_n follows $N(n\mu, \sigma\sqrt{n})$

i.e., S_n follows $N(150, \sqrt{150})$

$$P\{120 \le S_n \le 160\} = P\left\{\frac{-30}{\sqrt{150}} \le \frac{S_n - 150}{\sqrt{150}} \le \frac{10}{\sqrt{150}}\right\}$$
$$= P\{-2.45 \le z \le 0.85\}$$

where z is the standard normal variable.

$= 0.4927 + 0.2939$, (from the normal tables)

$= 0.7866$

UNIT – IV – Introduction to time series – SMTA5204

INTRODUCTION:

A **time series** is a set of observations taken at specified times, usually at equal intervals. In other words, a *series of observations recorded over time is known as a* **time series**. Examples of time series are the data regarding population of a country recorded at the ten-yearly censuses, annual production of a crop, say, wheat over a number of years, the wholesale price index over a number of months, the daily closing price of a share on the stock exchange, the hourly temperature recorded by weather bureau of a city, the total monthly sales receipts in business establishment, and so on. *In fact, data related with business and economic activities, in general, recorded over time give* rise to a **time series.**

One of the most important tasks before the planners and administrators in the field of economic and business activities is to make future estimates based on the past behaviour of a phenomenon under consideration. For example, trade cycles are important to economists and others in business and commerce. The behaviour of the cycles and their causes are of interest to them. Such studies are to be based on the analysis of time series data collected over time. *Thus,* the **analysis of time series** *plays an important role in empirical investigations of economic, commercial, social and even biological phenomena.*

Mathematically, **a time series** is defined by the fractional relationship

$$Y_t = f(t)$$

where Y_t is the value of the variable (or phenomenon) under consideration over time t. Thus, if the values of a variable at time points $t_1, t_2, ..., t_n$ are $Y_1, Y_2, ..., Y_N$ respectively, then the series

t	:	t_1	t_2	$t_3......, t_N$
Y_t	:	Y_1	Y_2	$Y_3......, Y_N$

constitute a time series.

COMPONENTS OF TIME SERIES:

Empirical studies of a number of time series have revealed the presence of certain **characteristic movements or fluctuations** in a time series. *These characteristic movements of a time series may be classified in four different categories called* **components of time series.** In a long time series, generally, we have the following **four components :**

1. Secular Trend or long-term movements
2. Seasonal variations
3. Cyclic variations
4. Random or Irregular movements

SECULAR TREND:

Secular trend *means the general long-term tendency of a series. In fact, secular trend is that characteristic of a time series which extends consistently throughout the entire period of time under consideration. It shows a long-term tendency of an activity to grow or to decline.* For example, a time series on population shows a tendency to increase; time series ot sales of a product shows a tendency to increase, and so on. On the other hand, a downward tendency is observed in the time series on birth and death rates. The factors which remain more or less constant over a long period also produce a trend. *The term* **'long period of time'** *is a relative phenomenon and cannot be defined exactly.* For some cases, a period as small as a week may be fairly long while in other cases, a period as long as 2 years may not be assumed long. For example, an increase in agricultural production over a period of two years would not be termed as secular change, whereas if the count of bacterial population of culture every five minutes, for a week shows an increase, then we would consider it as a secular change.

SEASONAL VARIATION:

The component responsible for the regular rise and fall in the magnitude of the time series is called **seasonal variation**. *In other words* **seasonal movements** or **seasonal variations** *refer to identical, or almost identical, patterns which a time series appears to follow during corresponding months of successive years.* Such variations are due to recurring events which takes place annually, quarterly, monthly, weekly or even daily, depending on the type of data available. *But in no case this period is to exceed one year. In view of their regular nature, seasonal variations are precise and can be foreseen*, as for instance the prices of agricultural commodities fall every year during the harvesting period, the sale of umbrellas pick up very fast in a rainy season, the demand for electric fans goes up during summer. Seasonal variations in general refer to annual periodicity in business and economic activities. These are the effects of seasonal factors like climatic conditions, human habits, fashions, customs and conventions of the people in a particular society.

CYCLICAL VARIATION:

Cyclical movements or **variations** *refer to the long-term oscillations or swings about a trend line. These cycles may or may not be periodic, i.e., they may or may not follow exactly similar patterns after equal intervals of time. Such variations are of longer duration than a year and they do not show the type of regularity as observed in the case of seasonal variations.* An important example of cyclical variations are the so-called **business cycles** representing intervals of **prosperity, recession, depression** and **recovery**. Each phase changes gradually into the phase which follows it in the given order. In a business activity, these phases follow each other with steady regularity and the period from the peak of one boom to the peak of the next boom is called a **complete cycle.** The usual periods of a business cycle may be ranging between 5–11 years. Most of the economic and business series relating to income, investment, wages, production shows this tendency.The study of cyclical fluctuations is therefore very important for predicting the turning phases in a business activity which may greatly help in proper policy formation in the area.

IRREGULAR VARIATION:

Random or Irregular *movements refer to such variations in a time series which do not repeat in a definite pattern.* Irregular movements in a time series may be of two types :

(i) Random or chance variations

(ii) Episodic variations

Random or chance variations *in a real phenomenon are inevitable by nature. It does effect a series in a random way, and as such, the effect of chance or random variations on a series is small.*

On the other hand, **episodic variations** *in a time series arise due to specific events or episodes like epidemic,* fire, strike or natural calamities like flood, earthquake or late monsoon etc. In some cases, irregular variations may not have a significant importance while in others these may be so intense as to result in new cyclical variations.

MEASUREMENT OF TREND:

The main objective behind the study of the trend of a time series are :

1. to describe the long–term growing or declining trend in a phenomenon under study.
2. to eliminate the trend component in order to bring into focus the remaining components in the time series.

In order to meet these objectives, some statistical methods of **estimation** *or* **determination of trend** *are as follows* :

1. Free hand, graphic method
2. Semi-average method
3. Moving average method
4. Method of least squares

GRAPHIC METHOD:

This is the simplest method of trend determination. According to this method, we plot the graph of the series and then draw a free hand curve through the points on the graph. Smoothing of time series data with a free hand curve eliminates the other components, viz., seasonal and irregular. The method does not involve complex mathematical calculations and can be used to describe all types of trend, linear or non-linear. However, the method is very subjective and can be adopted only to have a general idea of the nature of trend.

Example 1 **:** *Using the free hand hand or graphic method, fit a straight line trend to the following time series*

Year	1983	1984	1985	1986	1987	1988	1989	1990
Sales ('000)	80	90	85	92	87	99	93	120

Solution : Choosing a suitable scale, years are marked along the *x*-axis and corresponding sales values are marked along the *y*-axis. The points so obtained are then joined by straight lines which show the behaviour of sale values (actual data) over the given period. Then we draw a free hand straight line through the points of actual data for smoothing the time series data to obtain the trend. The behaviour of actual data and the trend line (dotted) are shown in fig. 1.

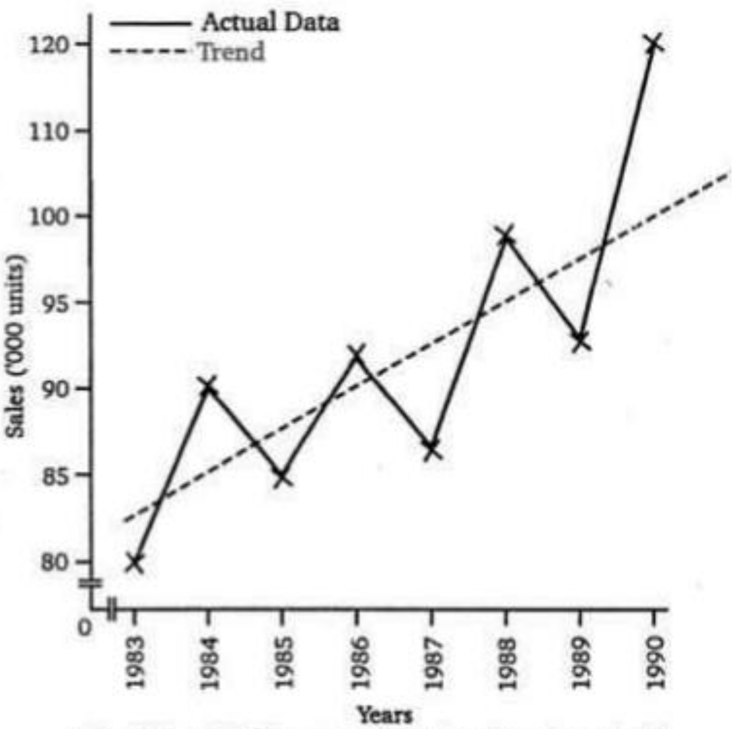


Fig. 1 Straight line trend by free hand method

SEMI- AVERAGE METHOD:

The method of semi-average is also simple. The method consists of dividing the data into two parts, preferably equal, and averaging the data in each part. In this way we obtain two points on the graph of the time series. The line obtained by joining these two points is the required trend line and may be extended in both the directions for estimating the trend values.

As compared with graphic method, the present method is better in view of its objectivity in the sense that every one who applies it would get the same results. However, the method has its limitation as it is appblicable only in a situation when the trend is linear or nearly linear. The following example will clarify the procedure.

Example 2 **:** *Determine straight line trend by semi-average method for the following time series data*

Year	1980	1981	1982	1983	1984	1985	1986	1987	1988	1989	1990
Production ('000 units)	18	25	21	15	26	31	30	20	35	32	23

Solution **:** According to semi-average method, the given time series is divided into two parts. Here,the data about 11 years are given, thus the value corresponding to the middle year, *i. e.*, 1985 is ignored. The averages of first and the last five years are then computed as under :

	Year	Production ('000 units)	Total Production	Semi average	Average year
First five years	1980	18			
	1981	25			
	1982	21	→ 105	105 ÷ 5 = 21	1982
	1983	15			
	1984	26			
Last five years	1986	30			
	1987	20			
	1988	35	→ 140	140 ÷ 5 = 28	1988
	1989	32			
	1990	23			

MOVING AVERAGE METHOD:

The method of moving averages attempts to smooth out the irregularities in a series by a process of averaging. *By using averages of appropriate orders (or extent), cyclical, seasonal and irregular variations may be eliminated, thus leaving only the trend component.* **Moving averages of extent** *m (or period) is a series of successive averages of m terms at a time, starting from 1st, 2nd, 3rd terms and so on until we exhaust the whole time series.* if m is odd, say equal to $(2k + 1)$, then the moving average is put against the mid-value of the period it covers, *i. e.*, against $t = k + 1$. On the other hand, if m is even, say equal to $2k$, it is placed between two middle values of the period it covers. Thus when an even number of years is taken in moving average, the average does not coincide with an original time period. For overcoming this situation, moving average of extent two of these moving averages are taken and the first of such values is put against $t = k + 1$. This procedure of centering puts the moving averages against the time points of the series rather than between these points. Symbolically, the 3-yearly moving averages of a time series can be computed as shown in the following table.

3 YEARLY MOVING AVERAGE			
Col. 1.	**Col. 2.**	**Col. 3.**	**Col. 4 = Col. 3 ÷ 3**
Years (t)	y_t	**3-yearly moving totals**	**3-yearly moving averages**
1	y_1	–	–
2	y_2	$\rightarrow (y_1 + y_2 + y_3)$	$(y_1 + y_2 + y_3)/3$
3	y_3	$\rightarrow (y_2 + y_3 + y_4)$	$(y_2 + y_3 + y_4)/3$
4	y_4	$\rightarrow (y_3 + y_4 + y_5)$	$(y_3 + y_4 + y_5)/3$
5	y_5	$\rightarrow (y_4 + y_5 + y_6)$	$(y_4 + y_5 + y_6)/3$
6	y_6	$\rightarrow (y_5 + y_6 + y_7)$	$(y_5 + y_6 + y_7)/3$
7	y_7	$\rightarrow (y_6 + y_7 + y_8)$	$(y_6 + y_7 + y_8)/3$
.	.	.	.
.	.	.	.
.	.	.	.
.	.	.	.
$N-1$	y_{N-1}	$\rightarrow (y_{N-2} + y_{N-1} + y_N)$	$(y_{N-2} + y_{N-1} + y_N)/3$
N	y_N	–	–

EXAMPLE 1:

Using three year moving averages determine the trend and short term fluctuations.

Year :	1973	1974	1975	1976	1977	1978	1979	1980	1981	1982
Production: ('000 tons)	21	22	23	25	24	22	25	26	27	26

Solution:

year	production	3 year moving total	3 year moving average	Short term fluctuation
1973	21	...	...	...
1974	22	66	22.00	0.00
1975	23	70	23.33	-0.33
1976	25	72	24.00	1.00
1977	24	71	23.67	0.33
1978	22	71	23.67	-1.67
1979	25	73	24.33	0.67
1980	26	78	26.00	0.00
1981	27	79	26.33	0.67
1982	26	...	...	...

Example :2

Obtain trend for four yearly moving averages for the following data.

Year:	1988	1989	1990	1991	1992	1993	1994	1995	1996	1997	1998
Production:	614	615	652	678	681	655	717	719	708	779	757

Solution : **Computation of trend by 4-yearly moving averages**

Year	Production	4-yearly moving totals	4-yearly centred moving totals	4-yearly moving averages (trend)
(1)	(2)	(3)	(4)	Col. (4) ÷ 8
1988	614			–
1989	615			–
		2559		
1990	652		5185	648.125
		2626		
1991	678		5292	661.500
		2666		
1992	681		5397	674.625
		2731		
1993	655		5503	687.875
		2772		
1994	717		5571	696.375

		→	2799			
1995	719			→	5722	715.250
		→	2923			
1996	708			→	5886	735.750
		→	2963			
1997	779					–
1998	757					–

METHOD OF LEAST SQUARES:

The method of least squares has already been explained in the context of regression analysis in chapter 10 of the present book. *As observed, the method is very useful for fitting mathematical functions to a given set of data.* The method is objective, and therefore, gives correct and accurate estimation of trend, once the form of equation representing trend is determined.

An examination of graphical plot of the time series often provides an adequate basis for deciding the functional form of the trend. Some of the common curves used for representing trend are :

(a) $Y = a + bX$, Linear or Straight line trend.

(b) $Y = a + bX + cX^2$, Parabolic or Quadratic trend.

(c) $Y = ab^X$, Exponential trend.

(a) Fitting of Linear or Straight Line Trend

The simplest type of trend equation is the linear equation of the form

$$Y = a + bX \qquad ...(1)$$

where X represents time and Y the value of the variable. Here Y is the dependent and X is an independent variable.

Now for the set of given data $(X_1, Y_1), (X_2, Y_2), \ldots, (X_N, Y_N)$, the constants a and b are determined by solving simultaneously the equations :

$$\Sigma Y = Na + b\,\Sigma X$$
$$\Sigma XY = a\,\Sigma X + b\,\Sigma X^2 \qquad ...(2)$$

The equations in (2), called **normal equations for the least square line in (1),** gives

$$a = \frac{(\Sigma Y)(\Sigma X^2) - (\Sigma X)(\Sigma XY)}{N\,\Sigma X^2 - (\Sigma X)^2} \qquad ..(3)$$

$$b = \frac{N\,\Sigma XY - (\Sigma X)(\Sigma Y)}{N\,\Sigma X^2 - (\Sigma X)^2} \qquad ...(4)$$

If the values of X are equidistant, the calculations involved in the estimation of a and b can be further simplified by shifting the origin to the appropriate mid-point in time, so that $\Sigma X = 0$. Obviously, the normal equations in (2) becomes

$$\left.\begin{aligned}\Sigma Y &= Na \\ \Sigma XY &= b\,\Sigma X^2\end{aligned}\right\} \quad ...(5)$$

Therefore, $a = \dfrac{\Sigma Y}{N}$ and $b = \dfrac{\Sigma XY}{\Sigma X^2}$ ...(6)

Substituting the estimated values of a and b in (1), the fitted linear trend will be

$$Y = a + bX \quad ...(7)$$

we can find the trend values, say Y, by putting different values of X in (7). When writing the trend equation, the origin and unit of time must be clearly specified, as an equation without such specification will be useless.

EXAMPLE:

Below are given the figures of production (in 1000 tons) of a fertilizer factory.

Year	1997	1998	1999	2000	2001	2002	2003
Production	70	75	90	98	84	91	99

Fit a straight line trend by the method os least squares and estimate trend values for 2005.

[U.P.T.U. 2008]

Solution **:** We use the method of least squares to fit a straight line trend. Here, the trend line is

$$Y = a + bX$$

where Y is the production

we make the transformation

$$x = X - 2000 \quad ...(i)$$

Thus, the trend becomes

$$Y = a + bx \quad ...(ii)$$

Computation of trend by least squares method

Year (X)	Number (Y)	$x = X - 2000$	x^2	xY
1997	70	−3	9	−210
1998	75	−2	4	−150
1999	90	−1	1	90
2000	98	0	0	0
2001	84	1	1	84
2002	91	2	4	182
2003	99	3	9	297
$N = 7$	$\Sigma Y = 607$	$\Sigma x = 0$	$\Sigma x^2 = 28$	$\Sigma xY = 113$

The normal equations are

$$\Sigma Y = N a + \Sigma X$$

$$\Sigma xY = a \Sigma X + b \Sigma x^2$$

From the table, these equations becomes

$$607 = 7a + 0 \quad \Rightarrow \quad a = 86.7$$

$$113 = 0 + 28b \quad \Rightarrow \quad b = 4.03$$

Thus, the fitted trend line becomes

$$Y = 86.7 + 4.03x \quad \text{where} \quad x = X - 2000 \qquad \text{...(iii)}$$

Putting $x = -3, -2, -1, 0, 1, 2, 3$ in (iii) we can get trend values as follows :

Year	1997	1998	1999	2000	2001	2002	2003
Trend **Values $Y = 86.7 + 4.03x$**	74.61	78.64	82.67	86.7	90.73	94.76	98.79

Estimate of production for 2005 is

$$\hat{Y} = 86.7 + 4.03(2005 - 2000)$$

$$= 86.7 + 20.15$$

$$= 106.85$$

UNIT – V – Analysis of time series – SMTA5204

SEASONAL VARIATION:

As discussed earlier, *there are certain variations, called seasonal variations, which occur with certain degree of regularity within a definite period.* The period of variations may be a year, a month or even a day. A variety of causes may be listed for such variations. Some times climatic conditions affect production in agriculture and industries. For example, the sale of woollens picks up in every winter; prices of food grains come down in harvesting season; sale of cold drinks goes up during summer, etc. and so on. On the other hand, there are man-made factors which also cause such variations. For instance, the demand for consumer products goes up during the early part of month. The traffic in a city is high during the rush hours. When time series data are given in annual figures, it will not possess the seasonal variations. Thus, such variations are present only when data are given for specific periods of the year *i.e.*, the data are given quarterly, monthly, weekly, daily or hourly.

MEASURES OF SEASONAL VARIATION:

1. Method of averages
2. Moving Average Method
3. Ratio to moving average
4. Ratio to trend.

1. Method of Simple Averages

According to this method the data for each month (if monthly is given) are expressed as percentage of the average for the year. The method involves the following **steps** :

(i) Arrange the data by years and month (or quarters if quarterly data are given).

(ii) The figures for each month are added and averages are obtained by dividing the monthly totals by the number of years. Suppose the averages for the 12 months are denoted by $\bar{X}_1, \bar{X}_2, \ldots, \bar{X}_{12}$.

(iii) Then obtain the overall average of monthly averages as :

$$\bar{\bar{X}} = \frac{\bar{X}_1 + \bar{X}_2 + \ldots + \bar{X}_{12}}{12}$$

(v) Obtain **seasonal indices** for different months by expressing the monthly averages as percentages of the overall average $\bar{\bar{X}}$ in the following way :

$$\text{Seasonal Index for the first month} = \frac{\bar{X}_1}{\bar{\bar{X}}} \times 100$$

$$\text{Seasonal Index for the second month} = \frac{\bar{X}_2}{\bar{\bar{X}}} \times 100$$

... ...

$$\text{Seasonal Index for the twelfth month} = \frac{\bar{X}_{12}}{\bar{\bar{X}}} \times 100$$

It should be noted that the average of the indices will always be 100, *i. e.*, the sum of the indices will be 1200 for 12 monthly data and the sum will be 400 for 4 quarterly data.

Example:

Assuming that the trend is absent, determine if there is any seasonality in the data given below

Year	Ist Quarter	2nd Quarter	3rd Quarter	4th Quarter
2004	3.7	4.1	3.3	3.5
2005	3.7	3.9	3.6	3.6
2006	4.0	4.1	3.3	3.1
2007	3.3	4.4	4.0	4.0

What are the seasonal indices for various quarters ? *(M. Com., M.K. Univ.)*

Solution. COMPUTATION OF SEASONAL INDICES

Year	Ist Quarter	2nd Quarter	3rd Quarter	4th Quarter
2004	3.7	4.1	3.3	3.5
2005	3.7	3.9	3.6	3.6
2006	4.0	4.1	3.3	3.1
2007	3.3	4.4	4.0	4.0
Total	14.7	16.5	14.2	14.2
Average	3.675	4.125	3.55	3.55
Seasonal Index	98.66	110.74	95.30	95.30

Notes for calculating seasonal index

The average of averages $= \frac{3.675 + 4.125 + 3.55 + 3.55}{4} = \frac{14.9}{4} = 3.725$

$$\text{Seasonal Index} = \frac{\text{Quarterly average}}{\text{General average}} \times 100$$

Seasonal Index for the first quarter $= \frac{3.675}{3.725} \times 100 = 98.66$

Seasonal Index for the second quarter $= \frac{4.125}{3.725} \times 100 = 110.74$

Seasonal Index for the third and fourth quarters $= \frac{3.55}{3.725} \times 100 = 95.30$

2. Moving Average Method:

It is a method for computing trend values in a time series which eliminates the shterm and random fluctuations from the time series by means of moving average. Moving average of a period m is a series of successive arithmetic means of m terms at a time starting with 1st, 2nd, 3rd so on. The first average is the mean of first m terms; the second average is the mean of 2nd term to (m+1)th term and 3rd average is the mean of 3rd term to (m+2)th term and so on. If m is odd then the moving average is placed against the mid value of the time interval it covers. But if m is even then the moving average lies between the two middle periods which does not correspond to any time period. So further steps has to be taken to place the moving average to a particular period of time. For that we take 2-yearly moving average of the moving averages which correspond to a particular time period. The resultant moving averages are the trend values.

3. Ratio to Trend Method:

Ratio-to-trend method is also known as **percentage trend method.** The method overcomes the difficulty of the simple average method when trend is present in the time series data. The method involves the following **steps** in measuring the seasonal indices :

(i) Compute the trend values by fitting trend equation to observed data by the method of least squares.

(ii) Express the original time series values as percentages of corresponding trend values.

(iii) Arrange these percentages according to years and months for monthly data (or according to years and quarters for quarterly data).

EXAMPLE:

The main defect of the ratio to trend method is that if there are cyclical swings in the series, the trend whether a straight line or a curve can never follow the actual data as closely as a 12- monthly moving average does. So a seasonal index computed by the ratio to moving average method may b less biased than the one calculated by the ratio to trend method.

Year	1st Quarter	2nd Quarter	3rd Quarter	4th Quarter
2003	30	40	36	34
2004	34	52	50	44
2005	40	58	54	48
2006	54	76	68	62
2007	80	92	86	82

Solution. For determining seasonal variation by ratio-to-trend method, first we will determine the trend for yearly data and then convert it to quarterly data.

CALCULATING TREND BY METHOD OF LEAST SQUARES

Year	Yearly totals	Yearly average Y	Deviations from mid-year X	XY	X^2	Trend values
2003	140	35	−2	−70	4	32
2004	180	45	−1	−45	1	44
2005	200	50	0	0	0	56
2006	260	65	+1	+65	1	68
2007	340	85	+2	+170	4	80
$N = 5$		$\Sigma Y = 280$		$\Sigma XY = 120$	$\Sigma X^2 = 10$	

The equation of the straight line trend is $Y = a + bX$.

$$a = \frac{\Sigma Y}{N} = \frac{280}{5} = 56 \qquad b = \frac{\Sigma XY}{\Sigma X^2} = \frac{120}{10} = 12$$

Quarterly increment $= \frac{12}{4} = 3$.

Calculation of Quarterly Trend Values. Consider 2003, trend value for the middle quarter, *i.e.*, half of 2nd and half of 3rd is 32. Quarterly increment is 3. So the trend value of 2nd quarter is $32 - \frac{3}{2}$, *i.e.*, 30.5 and for 3rd quarter is $32 + \frac{3}{2}$, *i.e.*, 33.5. Trend value for the 1st quarter is $30.5 - 3$, *i.e.*, 27.5 and of 4th quarter is $33.5 + 3$, *i.e.*, 36.5. We thus get quarterly trend values as shown below :

TREND VALUES

Year	*1st Quarter*	*2nd Quarter*	*3rd Quarter*	*4th Quarter*
2003	27.5	30.5	33.5	36.5
2004	39.5	42.5	45.5	48.5
2005	51.5	54.5	57.5	60.5
2006	63.5	66.5	69.5	72.5
2007	75.5	78.5	81.5	84.5

The given values are expressed as percentage of the corresponding trend values.

Thus for 1st Qtr. of 2003, the percentage shall be $(30/27.5) \times 100 = 109.09$, for 2nd Qtr. $(40/30.5) \times 100 = 131.15$, etc.

GIVEN QUARTERLY VALUES AS % OF TREND VALUES

Year	*1st Quarter*	*2nd Quarter*	*3rd Quarter*	*4th Quarter*
2003	109.09	131.15	107.46	93.15
2004	86.08	122.35	109.89	90.72
2005	77.67	106.42	93.91	79.34
2006	85.04	114.29	97.84	85.52
2007	105.96	117.20	105.52	97.04
Total	463.84	591.41	514.62	445.77
Average	92.77	118.28	102.92	89.15
S.I. Adjusted	92.05	117.36	102.12	88.46

Total of averages $= 92.77 + 118.28 + 102.92 + 89.15 = 403.12$.

Since the total is more than 400 an adjustment is made by multiplying each average by $\frac{400}{403.12}$ and final indices are obtained.

4. Ratio to moving average:

Ratio-to-moving average or percentage moving average method consists of expressing the original time series data as percentages of moving averages instead of percentages of trend values as in **'ratio-to-trend method'**, while rest of the steps are essentially the same. The procedure in this method consists of the following steps :

(i) Find the centred 12-monthly-moving averages (if monthly data are given) from the given time series data.

(ii) Express the original time series values as the percentage of the corresponding centred moving average values.

(iii) Average these percentages according to years and months and find averages over the years for all the 12 months.

(iv) Find the overall average of these 12-monthly averages. If the overall average is 100, the 12 monthly averages will be taken as seasonal indices, otherwise the monthly averages expressed as percentages of the overall average will be the required seasonal indices for the 12 months.

Symbolically, the logic behind the process may be explained as under :

The 12-monthly moving averages will eliminate the seasonal and irregular components and give us an estimate of the remaining two components namely trend (T) and cyclic (C). In multiplicative model we thus get an estimate of $T \times C$. Then the second step results in :

$$\frac{Y}{T \times C} \times 100 = \frac{T \times C \times S \times I}{T \times C} \times 100 = (S \times I) \times 100$$

Now on averaging over $S \times I$ in the third step, we are able to eliminate the irregular components with a possible bias. The final step givens us the adjusted seasonal indices.

Example 1:

Obtain seasonal indices by ratio to moving average method:

Year	Quarters			
	I	II	III	IV
2007	68	62	61	63
2008	65	58	66	61
2009	68	63	63	67

Solution : In the 'ratio-to-moving average' method, we first calculate 4 quarterly moving averages and ratios to moving averages as under :

Computation of Ratios to Moving Averages

Year and Quarter	Original data Y	4-quarterly moving totals	4-quarterly centred moving totals 4	4-quarterly centred moving averages (T)	Ratio to moving averages (percentage) = Y/T × 100
2007 I	68				
II	62				
	→	254			
III	61	→	505	63.125	96.63
	→	251			
IV	63	→	498	62.250	101.20
	→	247			
2008 I	65	→	499	62.375	104.21
	→	252			
II	58	→	502	62.750	92.43
	→	250			

	III	66		503	62.875	104.97
			253			
	IV	61		511	63.875	95.50
			258			
2009	I	68		513	64.125	106.04
			255			
	II	63		516	64.500	97.67
			261			
	III	63				
	IV	67				

Again, the percentage of original data to moving averages are arranged according to years and quarters to obtain the seasonal indices as shown in the following table :

Computation of Seasonal Indices

Year	Percentages to moving averages			
	I	II	III	IV
2007	–	–	96.63	101.20
2008	104.21	92.43	104.97	65.50
2009	106.04	97.67	–	–
Totals	210.25	190.10	201.60	196.70
Averages	105.125	95.05	100.80	98.35
Adjusted Quarterly Indices	$\frac{105.125}{99.83}\times 100$ **= 105.30**	$\frac{95.05}{99.83}\times 100$ **= 95.21**	$\frac{100.80}{99.83}\times 100$ **= 100.97**	$\frac{98.35}{99.83}\times 100$ **= 98.52**

$$\textbf{Overall mean} = \bar{X} = \frac{105.125 + 95.05 + 100.80 + 98.35}{4} = \mathbf{99.83}$$

Ex:1) Calculate 3-yearly moving average for the following data.

Years	Production	3-yearly moving avg (trend values)
1971-72	40	
1972-73	45	(40+45+40)/3 = 41.67
1973-74	40	(45+40+42)/3 = 42.33
1974-75	42	(40+42+46)/3 = 42.67
1975-76	46	(42+46+52)/3 = 46.67
1976-77	52	(46+52+56)/3 = 51.33
1977-78	56	(52+56+61)/3 = 56.33
1978-79	61	

Ex:1) Calculate 4-yearly moving average for the following data.

Years	Production	4-yearly moving avg	2-yealry moving avg (trend values)
1971-72	40		
1972-73	45		
		(40+45+40+42)/3 = 41.75	
1973-74	40		42.5
		(45+40+42+46)/3 = 43.15	
1974-75	42		44.12
		(40+42+46+52)/3 = 45	
1975-76	46		47
		(42+46+52+56)/3 = 49	
1976-77	52		51.38
		(46+52+56+61)/3 = 53.75	
1977-78	56		
1978-79	61		