HIGHER SECONDARY COURSE

STATISTICS

CLASS - XII

Government of Kerala

DEPARTMENT OF EDUCATION

State Council of Educational Research and Training (SCERT);
Kerala
2015

THE NATIONAL ANTHEM

Jana-gana-mana adhinayaka, jaya he
Bharatha-bhagya-vidhata.
Punjab-Sindh-Gujarat-Maratha
Dravida-Utkala-Banga
Vindhya-Himachala-Yamuna-Ganga
Uchchala-Jaladhi-taranga
Tava subha name jage,
Tava subha asisa mage,
Gahe tava jaya gatha.
Jana-gana-mangala-dayaka jaya he
Bharatha-bhagya-vidhata.
Jaya he, jaya he, jaya he,
Jaya jaya jaya, jaya he!

PLEDGE

India is my country. All Indians are my brothers and sisters.

I love my country, and I am proud of its rich and varied heritage. I shall always strive to be worthy of it.

I shall give my parents, teachers and all elders respect, and treat everyone with courtesy.

To my country and my people, I pledge my devotion. In their well-being and prosperity alone lies my happiness.

Prepared by :
State Council of Educational Research and Training (SCERT)
Poojappura, Thiruvananthapuram - 695012, Kerala.
Website : www.scertkerala.gov.in *e-mail* : scertkerala@gmail.com
Phone : 0471 - 2341883, Fax : 0471 - 2341869
Typesetting and Layout : SCERT
© Department of Education, Government of Kerala

To be printed in quality paper - 80gsm map litho (snow-white)

Forward

This Textbook of Statistics for Class-XII has been designed to help the students to grasp basic statistical concepts and techniques. Real life opportunities and situations have been provided for applying them in relevant situations.

Two major divisions of Statistics namely Descriptive statistics and Inferential statistics were considered while developing this material. The chapters on Descriptive statistics will provide you with a perspective to look at data and to communicate information effectively and efficiently. The chapters on Inferential statistics will tell you, how statistics supplies powerful concepts to estimate the pervasive effect of chance and will help us to generalise beyond limited sets of actual observations. Scopes for doing practicals have also been provided, wherever possible.

Novel and emerging ideas and concepts in Statistics have been discussed in the Textbook. This will equip you with a global outlook to enter into the field of statistics, as a professional. The learning outcomes envisaged in each chapter will enable you to look for the purpose of learning each chapter. In the end you can assess yourself - your level in achieving the concepts and ideas. Suggestions and comments about this book are most welcome.

Dr. S.Raveendran Nair
Director
SCERT; Kerala

Textbook Development Team

1. **MANOJ. K**
   HSS Panagad , Thrissur.
2. **ANWAR SHAMEEM. Z. A**
   GHSS Kuttiadi, Kozhikode.
3. **SAJISHKUMAR. M**
   MNKM HSS,
   Chittilamchery, Palakkad.
4. **BIJU . G. V**
   GHSS Kulakkada, Kollam.
5. **SREESAN.M.B**
   Karimpuzha HSS, Thottara
   Palakkad.
6. **MOHAMMED ASLAM.K**
   PPM HSS Kottukkara,
   Malappuram .
7. **MARY GEORGE**
   St. Raphael's CGHSS, Ollur, Thrissur.
8. **DEEPA KOSHY**
   PSVPM HSS, Ayravon, Pathanamthitta.
9. **HASIM. M. C**
   GHSS Aroli, Kannur
10. **SAKKEER.M**
    Govt. VHSS Meppayyur,
    Kozhikode.

Experts

P.K. VENUGOPAL

Associate Professor (Rtd.), Department of Statistics

Sree Kerala Varma College, Thrissur.

Dr. K. K. HAMSA

Associate Professor, Department of Statistics

Farook College, Kozhikode.

Dr. C. GOKULADASAN PILLAI

Former Curriculum Head, SCERT, Kerala

Academic Co-ordinator

Dr. Chandini. K. K

Head, Higher Secondary and Teacher Training

State Council of Educational Research and Training (SCERT),

Vidhyabhavan, Poojappura, Thiruvananthapuram-695 012

Contents

1. Correlation Analysis 7

1.1 Meaning of Correlation

1.2 Types of Correlation

1.3 Methods of Studying Correlation

2. Regression Analysis 29

2.1 Meaning of Regression

2.2 Linear regression

2.3 Regression equations

3 Elementary Calculus 45

3.1 Derivative of a Function

3.2 Second Order Derivative

3.3 Applications of second order derivatives

3.4 Integration

3.5 Definite Integrals

4. Random Variables 57

4.1. Random Variable

4.2. Discrete Random Variable

4.3 Probability mass function (pmf)

4.4 Cumulative distribution function (cdf)

4.5 Mathematical Expectation, Mean and Variance

4.6 Continuous Random Variables

4.7 Distribution Function

5. Discrete Probability Distributions 85

5.1 Binomial Probability Distribution

5.2 Poisson Probability Distribution

6. Normal Distribution 111

6.1 Normal Distribution - Concept

6.2 Normal Probability Density Function

6.3 Standard Normal Distribution

7. Sampling Distributions 129

7.1 Parameter and statistic

7.2 Sampling Distribution

7.3 Distribution of Sample Mean

7.4 Central Limit Theorem and its importance

7.5 Chi - square, t and F distributions

7.6 Relation among Z, Chi-square, t and F statistics

8. Estimation of Parameters 149

8.1 Point Estimation

8.2 Method of Moments

8.3 Interval Estimation

8.4 Confidence interval for the population mean

9. Testing of Hypothesis 165

9.1 Statistical Hypothesis

9.2 The Two types of Errors

9.3 Level of Significance and Power of a Test

9.4 Test Statistic and Critical Region

9.5 One - Tailed and Two - Tailed Tests

9.6 Tests of significance of single mean

9.7 Tests for significance for equality of two population means (Z test)

9.8 Chi - square test for independence of attributes

10. Analysis of Variance 207

10.1 Types of Variations

10.2 Causes of Variation

10.3 Assumptions of ANOVA

10.4 One - Way ANOVA

11. Statistical Quality Control 227

11.1 Meaning of Quality

11.2 Quality Control

11.3 Statistical Process Control

11.4 Variation and Causes of Variation

11.5 Control Charts

11.6 Types of Control Charts

11.7 Construction of Control Charts

11.8 Control Charts for Variables

11.9 Control Charts for Attributes

11.10 Uses of Statistical Quality Control

12. Time Series Analysis 255

12.1 Time Series

12.2 Components of Time series

12.3 Uses of Analysis of Time Series

12.4 Trend Analysis

13. Index Numbers 281

13.1 Classification of Index Numbers

13.2 Types of Index Numbers

13.3 Consumer Price index

13.4 Characteristics of Index Numbers

13.5 Uses of index Numbers

Appendix A - Answers 301

Appendix B -Glossary 305

Appendix C - References 309

Appendix D - Statistical Tables 310

Appendix E - R Code 316

Chapter 1

Correlation Analysis

We know that a bivariate data consists of two variables with a certain relationship. The variables in a bivariate data distribution can be both numerical, both categorical or one numerical and one categorical. Scatter plot is the graphical representation of bivariate data. The degree of variation between two variables is covariance. If there exists some relationship between two variables, and if we study it, then that statistical study is called Bivariate Analysis. Consider the examples-

Significant Learning Outcomes

After the completion of this chapter, the learner:

- Identifies the meaning of correlation.
- Recognises different types of correlation.
- Explains the methods of studying correlation.
- Identifies rank correlation coefficient.
- Uses the Karl Pearson's coefficient of correlation.
- Uses rank correlation coefficient in suitable situations.

advertisement cost and sales of a product, price of a commodity and its sale. With increase in the advertisement cost, the quantity sold is bound to increase. Or, with increase in the price of a commodity, the quantity sold is bound to decrease. These relationships may be linear or non linear (curvy linear). Similarly, relationships may exist between two or more variables. In this chapter, we discuss only the linear relationship between two variables.

Correlation analysis is useful in physical and social sciences. It is used to study the relationship between variables. It helps in measuring the degree of relationship and to compare the relationship between variables. Correlation is the basis of the concept of regression which is used for estimation.

1.1 Meaning of Correlation

Correlation refers to the relationship between two variables in a bivariate distribution. We can observe a certain relationship between two variables in the following cases.

- Price of product and its demand.
- Price of product and its supply.
- Wage and price index.
- Height and weight.

In each case, we can statistically analyse the degree or extent to which two variables fluctuate and relate to each other.

Let us consider some cases in detail.

Look at the following cases and corresponding scatter plots.

Case (i): Consider a certain brand of television. The amount utilised for advertisements and quantity sold in different years are given below.

Amount of advertisement (in Rs '000) (X)	25	50	70	80	100	130	170
Quantity sold (Y)	100	220	200	340	370	410	450

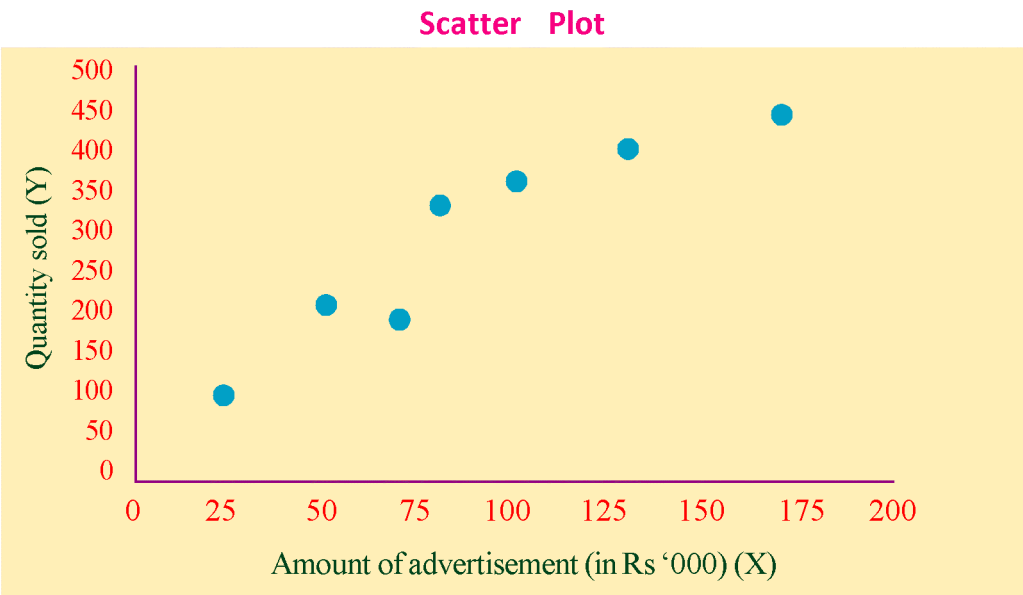


Fig. 1.1

Case (ii): Consider the case of CFL lamps. The price and quantity sold in different months are given below.

Price (X in Rs.)	50	100	120	130	150	200	220	230	240
Quantity sold (Y)	275	234	220	235	160	140	100	130	80

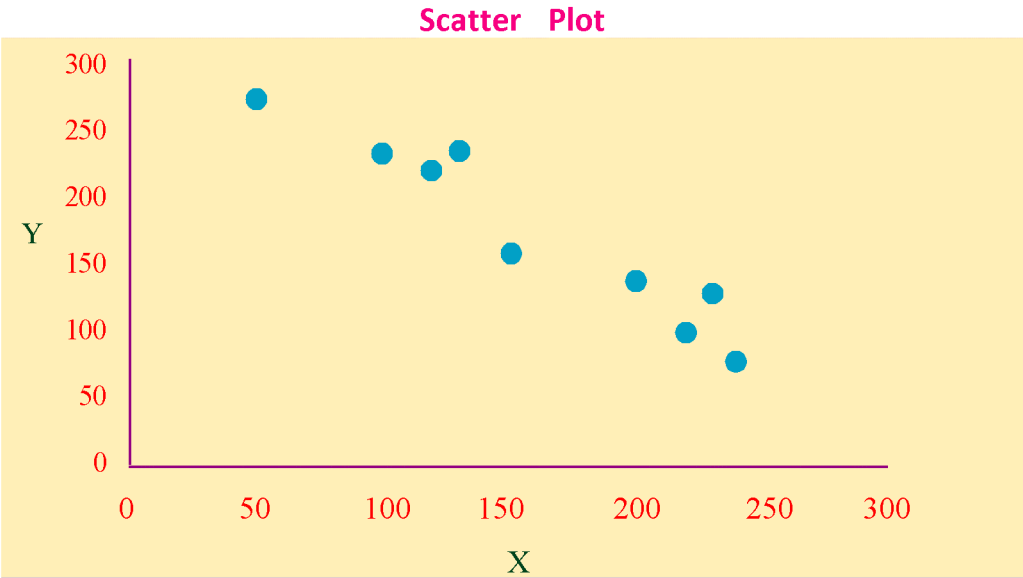


Fig. 1.2

Case (iii): The height in cms. and marks in English out of 50 of 10 students are given as follows.

Height in cms. (X)	150	165	155	156	158	163	158	162	152	167
Marks in English out of 50 (Y)	38	40	35	39	25	27	32	37	28	34

Scatter Plot

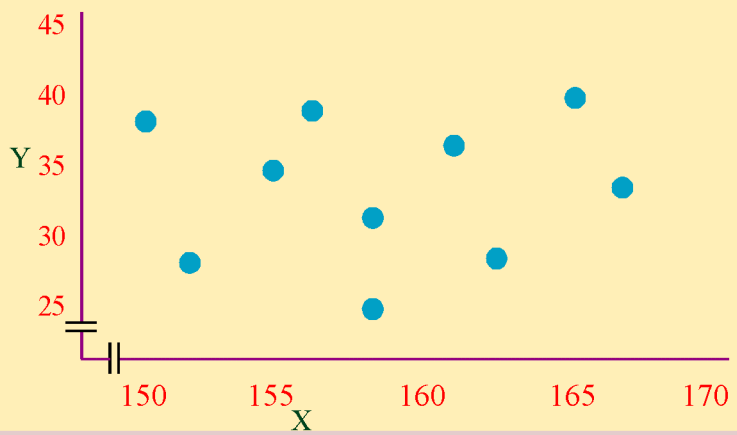


Fig. 1.3

Examine the data and scatter plots in these cases. What is your inference about the relationship between the two variables? We can see that values of X and Y increases together in case (i). while in case (ii), values of Y decreases as the values of X increases and in case (iii), no such relation is seen between X and Y. From this we can conclude that there may be some relation between the variables or there may be no relation between the variables.

Correlation analysis deals with the association or co-variation between two variables and helps to determine the degree of relationship. Correlation is the study of the degree of relationship between two variables. The correlation expresses the relationship or interdependence of two variables upon each other, in such a way that, changes in the values of one variable are sympathetic with the changes in the values of the other. Correlation also shows the degree of co-variation.

> Correlation Analysis is the study of the degree of relationship between two variables in a bivariate distribution.

Know your progress

List any three pairs of related variables which are very familiar to you.

1.2 Types of Correlation

Depending up on the nature of the relationship between the variables, correlation can be classified into:

1. **Positive correlation**
2. **Negative correlation**
3. **No correlation or Zero correlation**

Let us look into some details.

1. Positive Correlation

If the two variables are moving together in the same direction, then the correlation is called positive correlation. That is, increase in the value of one variable is accompanied by an increase in the value of the other variable and decrease in the value of one variable is accompanied by a decrease in the value of the other variable.

Use of fertilizer and yield of crop, price and supply, income and expenditure, etc., are examples for variables with positive correlation.

2. Negative Correlation

If the two variables are moving in opposite direction, then the correlation is called negative correlation. That is, increase in the value of one variable is accompanied by a decrease in the value of the other variable and decrease in the value of one variable is accompanied by an increase in the value of the other variable.

Intensity of light and distance from the source, price and demand, pressure and volume, etc., are examples for variables with negative correlation.

3. No correlation or Zero Correlation

If there is no association between the two variables, we say that there is no correlation or zero correlation. If the change in the value of one variable is not accompanied by any changes in the value of the other variable, then the correlation is zero or the variables have no correlation.

Amount of rainfall and scores in an examination, height and intelligence, etc., are examples for variables with zero correlation.

In some cases the relationship between the two variables may be proportional to each other. This is a case of **perfect correlation.**

Perfect Correlation

If the change in the value of one variable is proportional to the change in the value of the other variable, then the correlation is said to be perfect. If the variables are directely proportional then the correlation is **perfect positive** and if they are inversely proportional, then the correlation is **perfect negative**.

Radius and area of circles, sales and revenue, days of working and income of daily wage workers, hours of working and power consumption of electric appliances are examples for variables with perfect positive correlation.

Examples for variables with perfect negative correlation are pressure and volume (temperature kept constant), speed and time taken for travelling of vehicles, price index and purchasing power of money.

Know your progress

Write examples for the following:

- Positively correlated variables
- Negatively correlated variables
- Perfect Positively correlated variables
- Perfect negatively correlated variables
- Zero correlated variables

1.3 Methods of Studying Correlation

The different methods of studying correlation are discussed below.

1. Scatter Diagram

Scatter diagram is a graphical method of studying correlation. It is the simplest method of finding out whether there is any relationship between the two variables by plotting the values on a chart. It is also known as scatter plot.

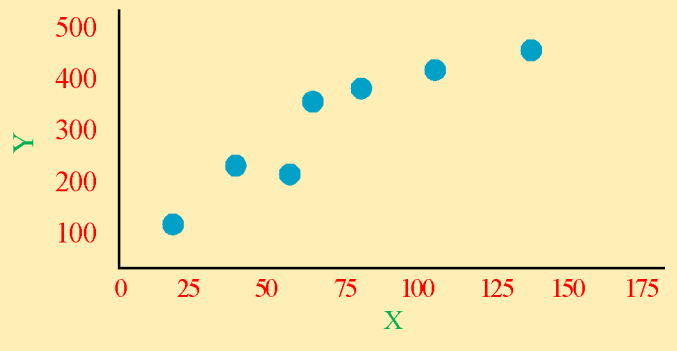


Fig. 1.4

The type of correlation can

be identified by this method.

Look at the scatter plots.

Fig (1.4) reprsents the scatter plot of Price (X) and Supply (Y) of a certain commodity.

The points in the scatter plot are rising from lower left hand corner to upper right hand corner. It shows that, there is positive correlation between the variables.

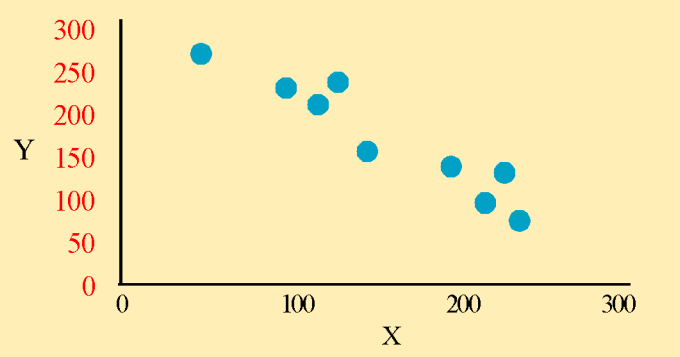


Fig. 1.5

Fig (1.5) represents the scatter plot of the Price (X) and Demand (Y) of a commodity.

The points in the scatter plot are falling from upper left hand corner to lower right hand corner. It shows that, there is negative correlation between the variables.

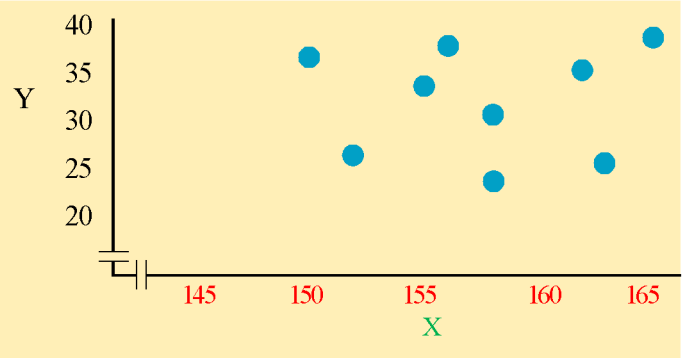


Fig.1.6

Fig (1.6) represents the scatter plot of the Height (X) and Scores in Statistics (Y) of students in a group.

The plotted points are scattered all over the diagram. It shows that there is no correlation between the variables.

Fig (1.7) represents the scatter plot of the length of a side (X) and perimeter (Y) of squares.

The points in the scatter plot are falling in a straight

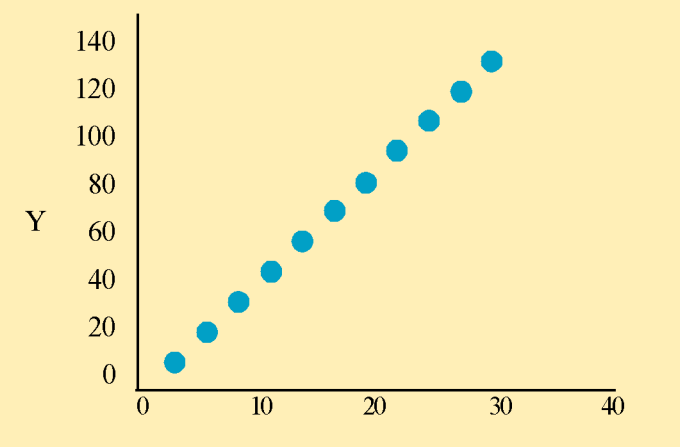


Fig.1.7

line from lower left hand corner to upper right hand corner. It shows that, there is perfect positive correlation between the variables.

Fig (1.8) represents the scatter plot of the Age (X) and remaining years of retirement (Y) with regards to teachers in a school.

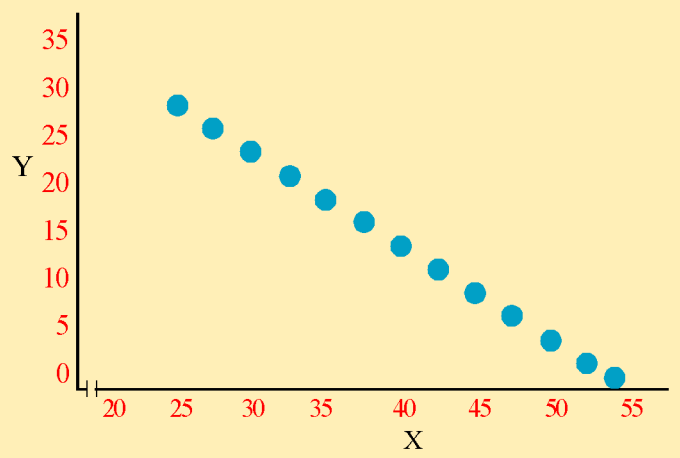


Fig. 1.8

The points in the scatter plot are falling in a stright line from upper left hand corner to lower right hand corner. It shows that, there is perfect negative correlation between the variables.

By observing the scatter diagrams, try to identify the merits and demerits of a scatter diagram. The following are some of them.

Merits

- Simple and attractive
- Easy to understand
- Gives a rough idea at a glance
- Not influenced by extreme items

Demerit

- Does not give the exact degree of correlation

Know your progress

Scores in Economics and Statistics of 10 students in a test are given below. Draw the scatter plot and interpret it.

Scores in Economics out of 100 (X)	85	35	25	14	65	25	78	32	58	45
Scores in Statistics out of 100 (Y)	75	65	32	29	56	18	82	39	62	35

Activity

Collect the scores obtained by 10 students in different subjects in class XI examination and draw the scatter plots for each pair of subjects. Find the subjects which are most correlated, least correlated and not correlated.

Coefficient of Correlation

Coefficient of Correlation is a relative measure showing the degree of relationship between two variables. It is a pure number free from units of measurement which can be used for comparison.

The most commonly used coefficients of correlation are:

- Karl Pearson's Coefficient of Correlation
- Spearman's rank Correlation Coefficient

Karl Pearson's Coefficient of Correlation

Karl Pearson, a great biometrician suggested a mathematical method for measuring the magnitude of the linear relationship between two variables. The most widely used method in practice is Karl Pearson's Coefficient of Correlation. It is usually denoted by 'r'.

Karl Pearson's Coefficient of Correlation between the variables X and Y is given by:

$$r(x, y) = \frac{\text{covariance between x and y}}{(\text{standard deviation of x})(\text{standard deviation of y})}$$

$$= \frac{\text{Cov(x, y)}}{\sigma_x \, \sigma_y}$$

Where

$$\text{cov}(x, y) = \frac{1}{n}\Sigma(x-\bar{x})(y-\bar{y})$$

$$\sigma_x = \sqrt{\frac{1}{n}\Sigma(x-\bar{x})^2}$$

$$\sigma_y = \sqrt{\frac{1}{n}\Sigma(y-\bar{y})^2}$$

$$\therefore\ r = \frac{\frac{1}{n}\Sigma(x-\bar{x})(y-\bar{y})}{\sqrt{\frac{1}{n}\Sigma(x-\bar{x})^2}\sqrt{\frac{1}{n}\Sigma(y-\bar{y})^2}}$$

$$r = \frac{\frac{1}{n}\Sigma xy - \bar{x}\,\bar{y}}{\sqrt{\frac{1}{n}\Sigma x^2 - \bar{x}^2}\sqrt{\frac{1}{n}\Sigma y^2 - \bar{y}^2}}$$, on simplification

or $$r = \frac{n\Sigma xy - \Sigma x\,\Sigma y}{\sqrt{n\Sigma x^2 - (\Sigma x)^2}\sqrt{n\Sigma y^2 - (\Sigma y)^2}}$$

$$r(x, y) = \frac{\text{Cov(x, y)}}{\sigma_x\,\sigma_y}$$

$$= \frac{n\Sigma xy - \Sigma x\,\Sigma y}{\sqrt{n\Sigma x^2 - (\Sigma x)^2}\sqrt{n\Sigma y^2 - (\Sigma y)^2}}$$

It is obvious that the value of coefficient of correlation (r) always takes values from -1 to +1. That is $-1 \le r \le 1$. This implies that r can be +1, -1, 0, between 0 and +1 and between -1 and 0. Look at the interpretations given below.

Interpretation of Karl Pearson's coefficient of correlation

i. If r = +1, then the correlation is perfect positive.

ii. If r = -1, then the correlation is perfect negative.

iii. If r = 0, then the correlation is zero.

iv. If 0 < r < +1, then the correlation is positive.

v. If -1 < r < 0, then the correlation is negative.

Properties of coefficient of correlation

1) Coefficient of correlation takes any value from -1 to +1.

   That is, $-1 \le r \le 1$

2) (i) Magnitude of Coefficient of correlation is unaltered, if a constant is added to (subtract from) the values of one variable (both variables)

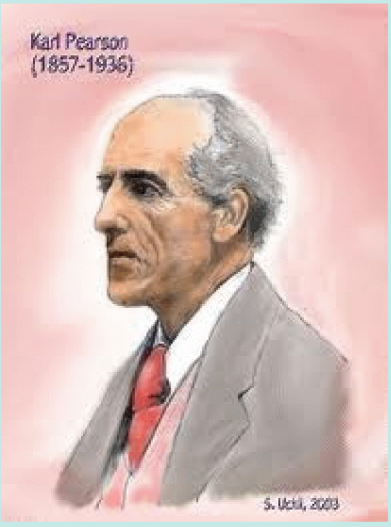
   (ii) Magnitude of Coefficient of correlation is unaltered, if the values of one variable (both variables) are multiplied (divided) by a constant.

   i.e. r (u, v) = $\pm r$ (x ,y) where u = $\frac{x-a}{b}$ and v = $\frac{y-c}{d}$; a, b, c and d are constants.

3) Correlation coefficient is symmetric with respect to variables. i.e. r (x, y) = r (y, x).

4) Correlation coefficient between two independent variables is zero. But the converse need not be true.

Karl Pearson was born in London on 27th March 1857. He worked in the University of London and formed the Department of Applied Statistics. He incorporated the biometric and Galton laboratories to this department. He remained with the department until his retirement in 1933 and continued to work till his death in 1936.



Proof for property 2

Let u = $\frac{x-a}{b}$ and v = $\frac{y-c}{d}$

Then Cov (u, v) = $\frac{Cov(x,y)}{bd}$, $\sigma_u = \frac{\sigma_x}{b}$ and $\sigma_v = \frac{\sigma_y}{d}$

$$r(u, v) = \frac{Cov(u,v)}{\sigma_u \sigma_v} = \frac{\frac{Cov(x,y)}{bd}}{\frac{\sigma_x}{b}\frac{\sigma_y}{d}} = \frac{Cov(x,y)}{\sigma_x \sigma_y} = r(x,y)$$

Proof for property 4

The correlation coefficient between the variables given below is zero but the variables are related by the relation $y = (x-20)^2$

X	10	15	20	25	30
Y	100	25	0	25	100

That is, correlation coefficient is zero implies the absence of linear relationship between them. They may however, be related in some other form. Similarly coefficient of correlation may be calculated mathematically from the given values of two variables even though they are really independent.

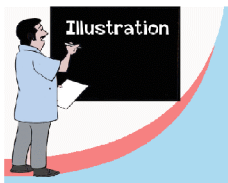


Illustration 1.1

The following gives the scores obtained in Statistics (X) and Economics (Y) out of 50 by 10 students in class tests.

X	18	25	5	31	12	22	28	23	38	45
Y	16	34	12	28	21	25	31	27	29	40

Draw the scatter diagram and find the Karl Pearson's coefficient of correlation.

Solution:

Scatter diagram

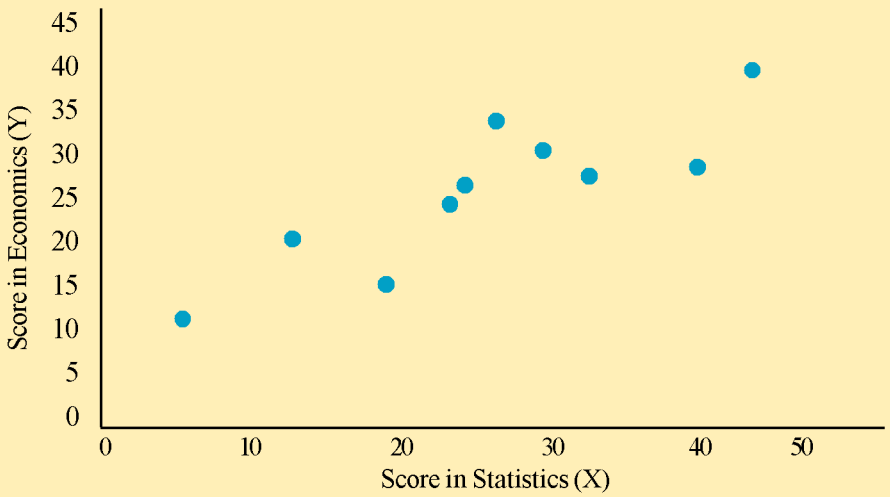


Fig. 1.9

	x	y	x^2	y^2	xy
	18	16	324	256	288
	25	34	625	1156	850
	5	12	25	144	60
	31	28	961	784	868
	12	21	144	441	252
	22	25	484	625	550
	28	31	784	961	868
	23	27	529	729	621
	38	29	1444	841	1102
	45	40	2025	1600	1800
Total	247	263	7345	7537	7259

Karl Pearson's coefficient of correlation, $r = \dfrac{n\sum xy - \sum x \sum y}{\sqrt{n\sum x^2 - (\sum x)^2}\sqrt{n\sum y^2 - (\sum y)^2}}$

$$= \frac{10 \times 7259 - 247 \times 263}{\sqrt{10 \times 7345 - (247)^2}\sqrt{10 \times 7537 - (263)^2}}$$

$$= 0.8686$$

The value of the coefficient of correlation 'r' is between 0 and 1. Therefore the correlation between the scores in Statistics and Economics is positive.

Know your progress

Heights (in inches) of 12 fathers (X) and that of their eldest son (Y)are given. Draw the scatter diagram and find the Karl Pearson's coefficient of correlation.

X	65	63	68	69	62	72	70	65	62	64	62	67
Y	66	60	71	67	65	68	63	61	69	62	65	65

Activity

Collect the data on the number of hours of study and scores in a Statistics examination of 20 students in your class. Find the coefficient of correlation and interpret the result.

Spearman's rank correlation coefficient

Some times in a bivariate data, one or both variables are expressed in terms of ranks instead of being expressed in actual values. Generally qualitative characteristics like honesty, beauty, efficiency, intelligence, etc., are better expressed by allotting ranks such as first, second, etc. The study of correlation of the characteristics expressed by ranks is called rank correlation. The primary purpose of computing a correlation coefficient in such a situation is to determine the extent to which the two sets of ranking of some individuals are in agreement or not.

Charles Edward Spearman,

Charles Edward Spearman, a British psychologist found out the method of ascertaining the coefficient of correlation by ranks. This measure is useful in dealing with qualitative characteristics.

Spearman's rank correlation coefficient is denoted by ρ and is given by:

$$\rho = 1 - \frac{6\Sigma d^2}{n(n^2-1)} \text{ or } \rho = 1 - \frac{6\Sigma d^2}{n^3 - n}$$

Where d = difference of ranks of an individual and n = number of individuals.

$$\rho = 1 - \frac{6\Sigma d^2}{n(n^2-1)} \text{ or } \rho = 1 - \frac{6\Sigma d^2}{n^3 - n}$$

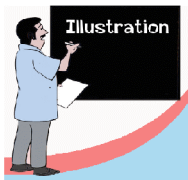


Illustration 1.2

Ranks obtained by 10 students in a Mathematics examination and their ranks in an intelligence test is given below.

Ranks in Mathematics	1	3	6	10	5	9	2	4	7	8
Ranks in intelligence test	4	5	3	8	9	10	1	2	7	6

Find the rank correlation coefficient.

Solution:

Here the ranks of the students are given.

Ranks in Mathematics	1	3	6	10	5	9	2	4	7	8
Ranks in intelligence test	4	5	3	8	9	10	1	2	7	6

X	Y	d	d^2
1	4	-3	9
3	5	-2	4
6	3	3	9
10	8	2	4
5	9	-4	16
9	10	-1	1
2	1	1	1
4	2	2	4
7	7	0	0
8	6	2	4
			Total: 52

$$\rho = 1 - \frac{6\Sigma d^2}{n^3 - n}$$

$$= 1 - \frac{6 \times 52}{10^3 - 10}$$

$$= 1 - \frac{312}{990}$$

$$= 1 - 0.3152$$

$$= 0.6848$$

Know your progress

The ranks given by two judges to 10 competitors in a beauty contest are as follows. Find the rank correlation coefficient.

Judge 1	3	5	8	6	4	9	7	2	1	10
Judge 2	2	6	10	8	3	7	5	1	4	9

Calculation of rank correlation coefficient when the ranks are repeated

Sometimes it may be necessary to assign equal ranks to two or more items. In such cases, it is customary to give an average rank to each item. Thus, if two items are ranked equal say at second place, each of them can be given the rank $\frac{2+3}{2}$, that is 2.5. Similarly if three items are ranked equal say at fifth place, each of them can be given the rank $\frac{5+6+7}{3}$, that is 6. When equal ranks are assigned to some entries a correction factor is to be added to the value of Σd^2 in the above formula for calculating the rank coefficient of correlation. The Correction Factor (C.F) is given by $\frac{\Sigma(m^3 - m)}{12}$, where m stands for the number of items with common rank. If there are more than one such group of items with common rank, this value is added as many times as the number of such group. The formula can be written as: $\rho = 1 - \frac{6\Sigma d^2 + C.F}{n^3 - n}$

Where $\quad C.F = \frac{\Sigma(m^3 - m)}{12}$

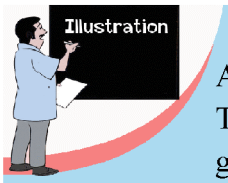


Illustration 1.3

A competitive test includes written test, group discussion and interview. The scores obtained in the written test by 10 top rank holders are given below.

Rank	1	2	3	4	5	6	7	8	9	10
Scores in written test	78	63	65	62	63	58	63	52	50	52

Find the rank correlation coefficient between the final rank and scores in the written test.

Solution:

First we have to rank the individuals according to the scores in the written test. The score 63 is repeated 3 times in the third, fourth and fifth places. Similarly 52 is repeated twice in the places eighth and ninth. Therefore rank 4 is assigned to the persons scored 63 and rank 8.5 is assigned to persons with score 52.

Rank at final	1	2	3	4	5	6	7	8	9	10
Rank in written test	1	4	2	5	4	7	4	8.5	10	8.5

R_1	R_2	d	d^2
1	1	0	0
2	4	-2	4
3	2	1	1
4	5	-1	1
5	4	1	1
6	7	-1	1
7	4	3	9
8	8.5	-0.5	0.25
9	10	-1	1
10	8.5	1.5	2.25
		Total	**20.5**

Rank 4 is repeated three times.

$$\text{There for C. F} = \frac{1}{12}(m^3 - m) = \frac{1}{12}(3^3 - 3) = 2$$

Rank 8.5 is repeated two times.

$$\text{There for C. F} = \frac{1}{12}(m^3 - m) = \frac{1}{12}(2^3 - 2) = 0.5$$

Total C.F = 2+0.5 = 2.5

$$\rho = 1 - \frac{6\Sigma d^2 + C.F}{n^3 - n}$$

$$= 1 - \frac{6 \times 20.5 + 2.5}{10^3 - 10}$$

$$= 1 - 0.1393$$

$$= \mathbf{0.8607}$$

Know your progress

The scores obtained by 12 students in a written test and ranks in performance are given below. Find rank correlation coefficient.

Scores in written test	12	15	18	20	18	16	13	18	17	11	13	18
Rank in performance	8	7	2	1	6	5	11	4	9	12	10	3

Activity

Conduct a quiz competition based on Statistics. Find the rank correlation between the ranks of the top 10 students in the quiz competition and their scores in statistics class test.

Let us conclude

Correlation is the study of relationship between two variables. Correlation can be studied graphically using scatter diagram. Different types of correlation are positive, negative and no correlation. If the variables are directly proportional, then the correlation is perfect positive and if the variables are inversely proportional, then the correlation is perfect negative. Correlation coefficient is the measure of degree of relationship between two variables. Karl Pearson's coefficient of correlation is most widely used. If one of the variables or both are qualitative, then we use Spearman's rank correlation coefficient to find the degree of relationship. The method of finding rank correlation coefficient when the ranks are repeated is also explained in this chapter.

Lab Activity

Verify the results obtained in Illustration 1.3 using spread sheet application.

Let us assess

For Questions 1-6, choose the correct answer from the given choices.

1. The maximum value of coefficient of correlation is:

   a) 0 b) 1 c) -1 d) infinity

2. The value of the coefficient of correlation:

   a) Has no limit b) Can be greater than 1

   c) Can be less than -1 d) Varies from -1 to +1

3. The coefficient of correlation will be:

   a) Positive b) Negative

   c) Either positive or negative d) Positive or negtive or zero

4. If the correlation between the variables X and Y is 0.3, then the correlation between the variables 2X and Y is.....

   a) 0.6 b) 0.9 c) 0.3 d) 0.4

5. If the value of Pearson's correlation coefficient calculated for marks in Statistics and Economics of 100 students is 0.9, there exists ..... type of correlation between the two variables.

   a) High positive b) High negative

   c) Perfect positive d) Perfect negative

6. If the correlation coefficient between the two variables X and Y is 0.4, then the correlation coefficient between X+3 and Y-5 is .....

   a) 0.8 b) 0.4 c) 0.2 d) 0.6

7. Raw cotton imports and cotton manufacture in million tons of a certain state for different years are given. Construct a scatter diagram.

Raw cotton imports (in million tons)	47	64	100	97	126	203	170	115
Cotton manufacture (in million tons)	70	85	100	103	111	139	133	115

8. The price in rupees(X) and supply in quintals (Y) of biriyani rice in a whole sale store is given. Draw a scatter diagram and interpret it.

X	10	22	34	35	69	85	95	98
Y	32	36	25	45	32	56	86	68

9. Calculate Karl Pearson's coefficient of correlation between price and supply of commodity of a retail dealer from the following data.

Price (in Rs.)	25	38	29	32	35	38	40	42
Supply (in Kg.)	38	35	39	45	42	48	39	52

10. Calculate the coefficient of correlation between the height of fathers and sons from the data given below.

Height of father (in inches)	64	65	66	67	68	69	70
Height of son (in inches)	66	67	65	68	70	68	72

11. The following data relate to the income (x) and expenditure (y) of 5 workers. Compute Pearson's correlation coefficient.

$$\Sigma(x-\bar{x})^2 = 1000,\ \Sigma(y-\bar{y})^2 = 40,\ \Sigma(x-\bar{x})(y-\bar{y}) = 100$$

12. The covariance between the variables X and Y is 10 and the variances of X and Y are 16 and 9 respectively. Find the coefficient of correlation.

13. Calculate the coefficient of correlation between age of cars and annual maintenance cost and comment.

Age of cars (years) X	2	4	6	7	8	10	12
Annual maintenance cost (rupees) Y	16000	15000	18000	19000	17000	21000	20000

14. The ranks of 10 students in two subjects of an examination is given as follows

Subject A	1	2	3	4	5	6	7	8	9	10
Subject B	3	4	2	3	5	6	9	10	7	8

Find rank correlation coefficient.

15. The marks in XI and XII examinations for 5 students in Statistics are given below. Compute the rank correlation coefficient.

Marks in XI (x) :	32	49	50	28	30
Marks in XII (y):	40	50	55	25	43

16. The scores given by two judges in an elocution competition for 5 competitors are as follows:

Judge I	70	65	72	64	78
Judge II	91	76	66	48	55

Find the rank correlation coefficient.

17. Find out the coefficient of correlation between X and Y by the method of rank differences

x	:	22	24	27	35	21	20	27	25	27	23
y	:	30	38	40	50	38	25	38	36	41	32

18. Find the spearman's rank coefficient of correlation between sales and profits of the following 10 firms

Firms :	A	B	C	D	E	F	G	H	I	J
Sales :	50	50	55	60	65	65	65	60	60	50
Profit :	11	13	14	16	16	15	15	14	13	13

19. The marks in Class XI and the Class XII exams for 7 higher secondary students in Statistics are given below. Compute the rank correlation.

Marks in Class XI :	15	14	25	14	14	20	22
Marks in Class XII :	25	12	18	25	40	10	7

Chapter 2

Regression Analysis

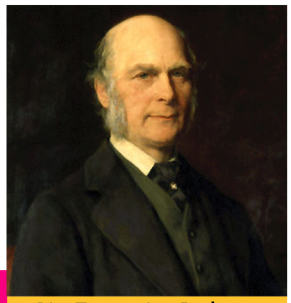
The correlation coefficient we have discussed in the previous chapter simply tells us about the direction and strength of relationship between two variables. In 1889 Sir Francis Galton published a paper on heredity. He reported his findings based on the study of relationship between the heights of fathers and their sons. He observed that the height of offsprings regress towards the mean. While dealing with economic and commerce data, we are required to make prediction and estimation. Prediction is one of the major problems in almost all spheres of human activity. Regression

Significant Learning Outcomes

After the completion of this chapter, the learner:

- Identifies the concept of regression analysis.
- Estimates unknown values for corresponding values given.
- Recognises regression lines and their point of intersection.
- Explains properties of regression coefficients.
- Compares correlation and regression.

analysis is one of the scientific techniques for making such prediction. Regression analysis also measures the percentage variation in dependent variable due to the influence of independent variable. It is one of the most widely used statistical techniques in almost all real life situations. We study more about regression analysis in this chapter.


Sir Francis Galton

2.1 Meaning of regression

Regression is a measure of the functional form of relationship between the variables. Or, in other words, it is a mathematical measure of the nature of relationship between the variables. The word regression means 'going back'. It is the study of cause and effect relationship. In this chapter we will focus only on linear regression, which involves only two variables. They are dependent variable and independent variable. It helps us to estimate the unknown values of dependent variables from the known values of independent variables.

For e.g. : By using the technique of regression, an economist may be able to estimate the demand of a commodity for a given price or an agriculturist can predict production based on rainfall.

> Regression analysis is a mathematical measure of the nature of relationship between two or more variables.

Independent variable and dependent variable

Suppose, a researcher studies the effect of age on a person's blood pressure. Here 'age' is an independent variable and 'blood pressure' is a dependent variable. If expenditure of a person depends on his income, the variable 'income' is independent variable and 'expenditure' is dependent variable. The variable whose value is to be predicted, is called dependent or response variable and the variable used for prediction is called independent or predictor variable.

2.2 Linear regression

When the given bivariate data are plotted on a graph paper we get a scatter diagram. We can construct straight lines through the points in the scatter diagram as shown in the figure given below.

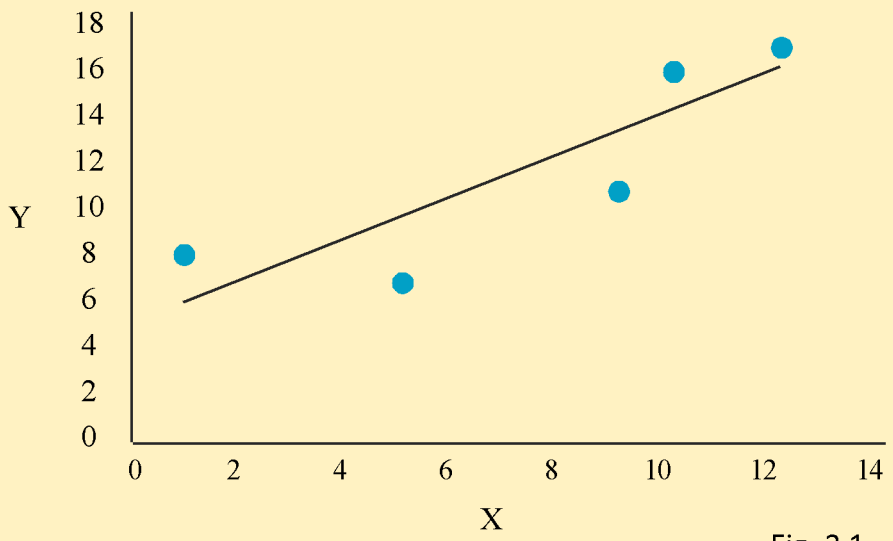


Fig. 2.1

If the points on the scatter diagram concentrate around a straight line that line is called **regression line** or **line of best fit.** The line of best fit is that line which is closer to the points in the scatter diagram. The equation of such a line is the first degree equation in X and Y. Since the relationship between the variables X and Y is not reversible we have two regression lines. One regression line shows regression equation of Y on X and other shows regression equation of X on Y.

Regression line of Y on X is used to predict Y for a given value of X and Regression line of X on Y is used to predict X for a given value of Y.

2.3 Regression equations

Regression equations are the equations of the regression lines. When we have two variables X and Y, we can think of two regression lines. One is regression equation of Y on X and other is regression equation of X on Y. Regression equations can be derived using Legendre's principle of least squares. In regression equation of Y on X, Y is dependent variable and X is independent variable.

Regression equation of Y on X is given by

$$y - \bar{y} = b_{yx}(x - \bar{x})$$

Where $\bar{y} = \frac{\Sigma y}{n}$ = mean value of Y

$\bar{x} = \frac{\Sigma x}{n}$ = mean value of X

b_{yx} = Regression coefficient of Y on X

$= \frac{Cov(X, Y)}{\text{var } X}$ or $b_{yx} = \frac{n\Sigma xy - \Sigma x \Sigma y}{n\Sigma x^2 - (\Sigma x)^2}$

Similarly when X is dependent variable and Y is independent variable we have another equation known as regression equation of X on Y. It is obtained by the formula.

Regression equation of X on Y

$$x - \bar{x} = b_{xy}(y - \bar{y})$$

Where b_{xy} is the regression coefficient of X on Y.

$$b_{xy} = \frac{n\Sigma xy - \Sigma x \Sigma y}{n\Sigma y^2 - (\Sigma y)^2}$$

Principle of least squares states that sum of squares of vertical deviations from the observed values and values obtained by the line of best fit should be minimum. i.e., if $d_1, d_2, d_3 \ldots\ldots$ are the deviations then principle of least squares states that the line of best fit should be drawn so as $d_1^2 + d_2^2 + d_3^2 + \ldots\ldots$ is minimum.

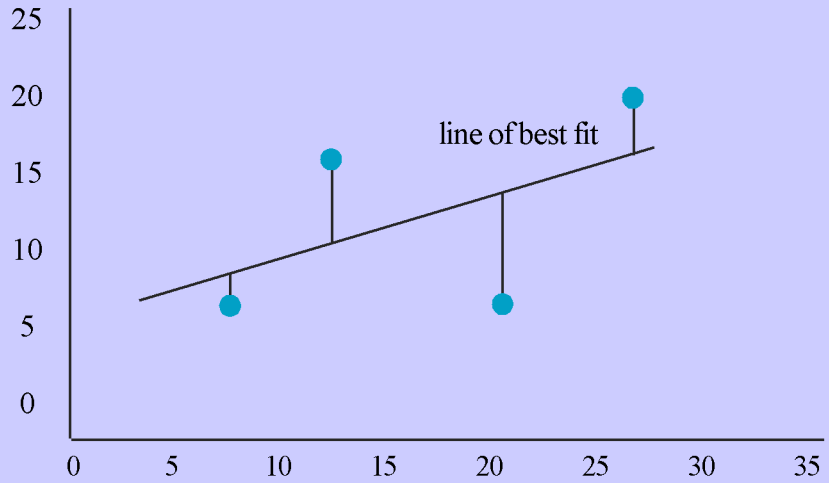


Fig. 2.2

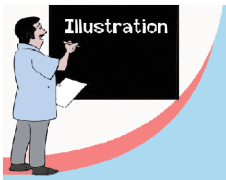


Illustration 2.1

The following data relate to sales and purchase of 10 important shops in a city.

Sales (000's):	4	6	5	9	10	7	2
Purchase(000's):	2	5	3	7	7	3	1

Form the regression equation and also evaluate the amount of sales for a purchase of Rs. 9000.

Solution

Let us take sales as the variable X and purchase as the variable Y

X	Y	XY	Y^2
4	2	8	4
6	5	30	25
5	3	15	9
9	7	63	49
10	7	70	49
7	3	21	9
2	1	2	1
$\Sigma x = 43$	$\Sigma y = 28$	$\Sigma XY = 209$	$\Sigma Y^2 = 146$

The equation of regression line of X on Y is

$$\bar{y} = \frac{\Sigma y}{n} = \frac{28}{7} = 4 \qquad \bar{x} = \frac{\Sigma x}{n} = \frac{43}{7} = 6.14$$

$$b_{xy} = \frac{n\Sigma xy - \Sigma x \Sigma y}{n\Sigma y^2 - (\Sigma y)^2}$$

$$= \frac{7 \times 209 - 43 \times 28}{7 \times 46 - (28)^2}$$

$$= 1.08$$

The equation is $x - \bar{x} = b_{xy}(y - \bar{y})$

i.e., $x - 6.14 = 1.08\,(y - 4)$

To find the amount of sales when purchase is 9000, we put Y = 9 in the above regression equation

$$x - 6.14 = 1.08\,(9 - 4)$$
$$x = 6.14 + 1.08 \times 5$$
$$= 11.54$$

Know your progress

The following data relate to the experience of machine operators and their performance ratings.

Operator experience(X) in years	16	12	18	4	3	10	5	12
Performance ratings(Y)	87	88	89	68	78	80	75	83

Calculate the regression line of performance ratings on experience and estimate performance if an operator has 7 years of experience.

Properties of regression coefficients

The following are important properties of regression coefficients.

- In regression equation of y on x ,b_{yx} is the coefficient of X.
- In regression equation of x on y, b_{xy} is the coefficient of Y.

  E.g.: $y - 4 = 1.2\,(x - 2)$, $b_{yx} = 1.2$

  $x - 6 = 0.7\,(y - 2)$, $b_{xy} = 0.7$
- The signs of both regression coefficients are same. That is, regression coefficients are either both positive or both negative.
- The product of both regression coefficients should be below one. That is, $b_{yx} \cdot b_{xy} \le 1$.
- The geometric mean of the regression coefficient is Coefficient of correlation i.e., $r = \pm\sqrt{b_{yx} \times b_{xy}}$
- $b_{yx} = r \times \dfrac{\sigma_y}{\sigma_x}$ and $b_{xy} = r \times \dfrac{\sigma_x}{\sigma_y}$ where σ_y is standard deviation of Y and σ_x is standard deviation of X.
- $b_{yx} \ne b_{xy}$ in general.

$$b_{yx} \times b_{xy} = r \times \frac{\sigma_y}{\sigma_x} \times r \times \frac{\sigma_x}{\sigma_y} = r^2,$$

i.e., $r = \pm\sqrt{b_{yx} \times b_{xy}}$, where r is correlation coefficient.

When both b_{yx} and b_{xy} are positive, r is positive.

When both b_{yx} and b_{xy} are negative, r is negative.

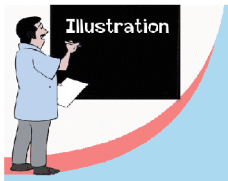


Illustration 2.2

The following data relate to area of cultivation in hectores of land(X) and agricultural output in tonnes(Y).

	X	Y
Arithmetic mean	50	30
Standard deviation	5	2

Coefficient of correlation = 0.7

1) Calculate the regression equation of agricultural output on area of cultivation.
2) Estimate agricultural output when there are 80 hectores of land available.

Solution

Given $\bar{y} = 30$ $\sigma_y = 2$

$\bar{x} = 50$ $\sigma_x = 5$

r = 0.7

1) To find the regression equation of agricultural output on area of cultivation, we need to find the Regression equation of Y on X

Regression coefficient of Y on X, (b_{yx}) $= r \times \frac{\sigma y}{\sigma x}$

$$b_{yx} = 0.7 \times \frac{2}{5} = 0.28$$

Regression equation of Y on X is $y - \bar{y} = b_{yx}(x - \bar{x})$

$$y - 30 = 0.28\,(x - 50)$$

2) To estimate agricultural ouput on Area of cultivation, substitute x = 80

$$y - 30 = 0.28\ (80 - 50)$$

$$y - 30 = 0.28\ (30) = 8.4$$

$$y = 30 + 8.4 = 38.4$$

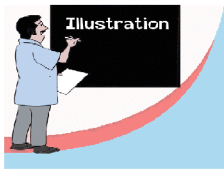


Illustration 2.3

Let $4x + 5y - 10 = 0$ is the regression equation of Y on X. Find b_{yx}.

Solution

$$5y = 4x + 10$$

$$\text{i.e., } y = \frac{4}{5}x + \frac{10}{5}$$

We know that in a regression equation of y on x, b_{yx} is the coefficient of x.

$$\therefore\ b_{yx} = \frac{4}{5}, \text{ the coefficient of x.}$$

Illustration 2.4

Calculate correlation coefficient if $b_{yx} = -0.23$ and $b_{xy} = -0.75$.

Solution

$$r = \pm\sqrt{b_{yx}b_{xy}} = \pm\sqrt{(-0.23)(-0.75)}$$

$$= \pm\sqrt{.1725} = \pm 0.4153$$

$r = -0.4153$ (Since b_{yx} and b_{xy} are negative.)

Know your progress

1. The following data are given for marks in English and Statistics in a certain examination.

	English	Statistics
Mean Marks	39.5	47.5
S.D of Marks	10.8	16.8

Correlation coefficient between Marks in Statistics & English = 0.42.

(a) Find the most probable mark in English if marks in Statistics is 50.

(b) Estimate the marks in Statistics if Marks in English is 35.

2. Regression coefficient between X on Y is $\frac{9}{16}$ Correlation coefficient between the same variables is $\frac{1}{4}$. Find Regression coefficient between X on Y.

Identification of regression lines

Regression lines are not reversible. Therefore when we have two regression lines an important problem is to identify which one is regression equation of Y on X and which one is regression equation of X on Y. By supposing one of the equation as the regression equation of Y on X and other as X on Y, we can obtain regression coefficients. If the product of the these two is numerically less than one then our supposition is true. However if their product is greater than one then our supposition is wrong. The regression lines can be identified as in the following example.

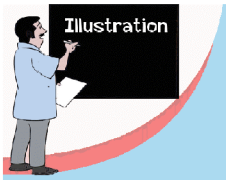


Illustration 2.5

Out of the two lines of regression given by $x + 2y - 5 = 0$ and $2x + 3y - 8 = 0$ which one is the regression line of X on Y and which one is regression line of Y on X.

Solution:

$x + 2y - 5 = 0$.....(1)

$2x + 3y - 8 = 0$.....(2)

Let us assume equation (1) as the regression equation of X on Y and equation (2) as regression equation of Y on X.

Let equation (1) be written as $x = -2y + 5$ therefore $b_{xy} = -2$

Equation (2) as $y = \frac{-2}{3}x + \frac{8}{3}$ therefore $byx = \frac{-2}{3}$

$$b_{xy} \times b_{yx} = -2 \times \frac{-2}{3} = \frac{4}{3} = 1.33 \text{ which is greater than } 1$$

$\therefore$ our supposition is wrong . That means (1) is regression equation of Y on X and (2) is regression equation of X on Y.

Know your progress

For a group of 50 persons, the regression equations of age (X) and the blood pressure (Y) are $3y - 5x + 180 = 0$ and $4x + 10y + 100 = 0$. Find the correlation coefficient.

Point of intersection of two regression lines

When we have two regression lines, they coincide at the point $(\bar{x}, \bar{y})$ as given in the following figure

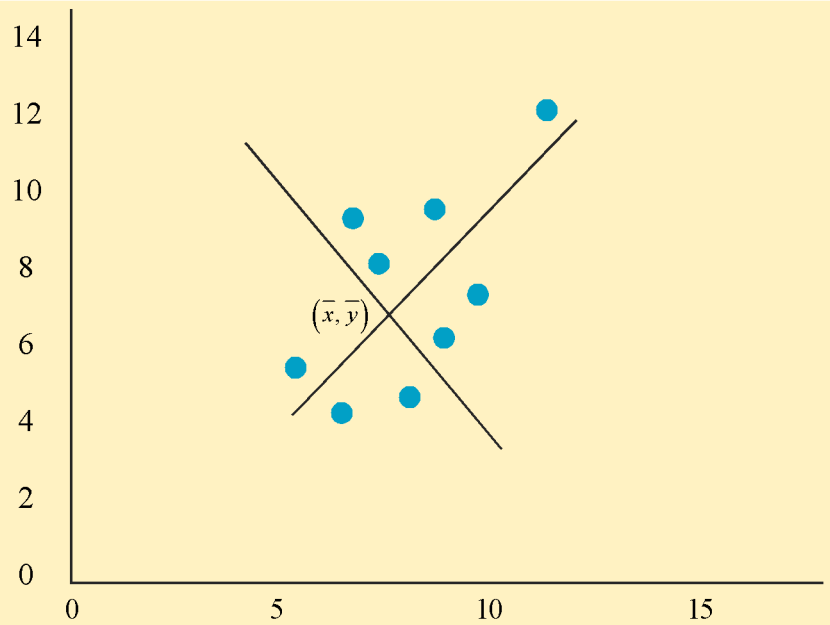


Fig. 2.3

When r = 1, the two regression lines coincide

When r = 0, the two regression lines are perpendicular

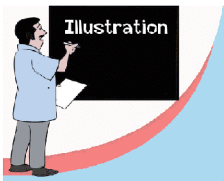


Illustration 2.4

In a linear regression analysis of 20 observations, the two lines of regression are $10y + 7x - 4 = 0$ and $5x + 9y - 1 = 0$.

a) Identify regression lines.

b) Obtain the correlation coefficient.

c) Obtain the mean values of X and Y.

Solution

a) 10y+ 7x – 4 = 0.....(1)

5x+9y – 1= 0.....(2)

Let us assume equation (1) as the regression equation of X on Y and equation (2) as regression equation of of Y on X.

Equation (1) can be written as 7x= – 10y+4 . Therefore $b_{xy} = -\frac{10}{7}$

Equation (2) can be written as 9y= – 5x+1. Therefore $b_{yx} = -\frac{5}{9}$

$$b_{xy} \times b_{yx} = -\frac{10}{7} \times -\frac{5}{9}$$

$= 0.79$ which is less than one.

$\therefore$ Our supposition is correct . That means (1) is regression equation of Y on X and (2) is regression equation of X on Y.

b) $r = \pm\sqrt{b_{xy} \times b_{yx}} = \pm\sqrt{-\frac{10}{7} \times -\frac{5}{9}}$

$= \pm\sqrt{0.79} = \pm 0.89$

$r = -0.89$ (Since b_{yx} and b_{xy} are negative)

c) Since both the lines of regression pass through the mean values, the point $(\overline{X}, \overline{Y})$ will satisfy both the equations. Hence both the equations can be written as

$7\overline{x} + 10\overline{y} = 4$ .....(1)

$5\overline{x} + 9\overline{y} = 1$ .....(2)

$(1) \times 5 \rightarrow \quad 35\overline{x} + 50\overline{y} = 20$ .....(3)

$(2) \times 7 \rightarrow \quad 35\overline{x} + 63\overline{y} = 7$ .....(4)

Substracting equation (3) from (4) we get $13\overline{y} = -13$

$\therefore \overline{y} = -1$

Putting the value of $\overline{y} = -1$ in equation (1) we get $\overline{x} = +2$

Thus mean of X = 2 and mean of Y = –1

Comparison between correlation and regression

Regression	Correlation
1. Regression is asymmetric.	Correlation is symmetric.
2. Regression is the cause and effect relationship between the variables.	Correlation is the association between the variables.
3. It is used for prediction.	It is not used for prediction.
4. Regression is the study of the nature of relationship between the variables.	Correlation is the study of strength of relationship.
5. It is used for further mathematical treatment.	It is not used for further mathematical treatment.
6. Regression is not reversible.	Correlation is reversible.

Let us conclude

In this chapter we have discussed the concept and importance of measuring regression. Regression is widely used for prediction and forecasting . The knowledge of regression helps us to understand how the value of dependent variable changes when the value of independent variable is fixed. We have also discussed about two regression lines and their importance. The comparison between correlation and regression will give us the characteristics of correlation and regression.

Lab Activity

(1) The following are the weights (Kg.) and blood glucose levels (mg./100ml.)of 16 apparently healthy adult males.

Weight	64	75	73	82	76	95	76	82
Glucose	108	109	104	102	105	121	99	100

(a) Obtain the linear regression equations

(b) Predict the glucose level of a person who weighs 95 Kg.

2 The body weight and the Body Mass Index (BMI) of 7 school going children are given in the following table.

Weight (Kg.):	15.0	26.0	27.0	25.0	25.5	27.0	32.0
BMI:	13.35	16.12	16.74	16.00	13.59	15.73	15.65

(a) Find the regression equation of BMI with respect to weight.

(b) Estimate BMI when weight is 40kg.

Let us assess

For Questions 1-10, choose the correct answer from the given choices.

1. If $b_{yx} \geq 1$ then b_{xy} is .....
   a) less than 1 b) greater than 1 c) equal to 1 d) equal to -1
2. The term regression was introduced by .....
   a) R A Fisher b) Sir Francis Galton
   c) Karl Pearson d) none of these
3. If X and Y are two variables, then there can be at the most ..... number of regression lines.
   a) one b) two c) three d) infinite
4. If the correlation coefficient between two variables X and Y is negative , then regression coefficient Y on X is .....
   a) positive b) negative c) zero d) not certain
5. In a regression line of Y on X , the variable X is known as .....
   a) independent variable b) regressor
   c) explanatory variable c) all of the above
6. The geometric mean of two regression coefficients b_{yx} and b_{xy} is equal to .....
   a) r b) r^2 c) 1 d) none of the above
7. Which one of the following can be regression coefficients:
   a) $\left(1, \frac{3}{2}\right)$ b) $\left(\frac{1}{2}, \frac{3}{2}\right)$ c) $\left(2, \frac{3}{2}\right)$ d) (2, 3)
8. Let $2x + 3y - 5 = 0$ is the regression line of X on Y then b_{xy}=.....
   a) $\frac{3}{2}$ b) $\frac{2}{3}$ c) $\frac{-2}{3}$ d) $\frac{-3}{2}$

9. Let $b_{yx} = -0.5$, $b_{xy} = -0.3$ then the value correlation coefficient=.....

a) –0.15 b) 0.15 c) 0.39 d) –0.39

10. To estimate the value of Y for a given value of X ,the regression equation used is .....

a) Y on X b) X on Y c) both of these d) none of these.

11. The following data relate to the age of drivers(X) the and number of motor accidents which occurred in a locality (Y) during the last 6 months.

a) Form the suitable regression equation.

b) Using the above equation calculate the number of accidents caused by a of 20 year old person.

Age of drivers(years)	19	21	30	45	50	54	25
Number of motor accidents	50	52	40	22	10	14	35

12. Given is the data on price and sales of a particular commodity.

Price	20	25	50	15	25	30	20	17
Sales	15	11	10	30	15	17	20	12

Using the technique of regression evaluate sales when the price of the commodity is 40 rupees.

13. The following data relate to time spent for exercising daily in minutes(X) and blood pressure(Y) of a group of patients.

	X	Y
Mean	60	100
Standard Deviation	20	15

Correlation coefficient = –0.81

a. Find the equation of suitable regression line.

b. Calculate the blood pressure of a person who exercised 70 minutes daily.

14. The following data relate to advertising expenditure and sales of 10 major shops in a city.

	Advertising expenditure (lakhs)	Sales (lakhs)
Mean	10	15
SD	5	3

Coefficient of correlation = 0.65

a) Calculate sales when advertising expenditure is 13 lakhs.

b) Calculate advertising expenditure when sales is 20 lakhs.

15. The following calculations have been made for the price of 12 stocks (X) on BSE on a certain day along with volume of sales on shares (Y). From these calculations calculate the regression equation of price of stocks on volume of shares.

$\sum x = 580,\ \sum y = 370,\ \sum xy = 11494,\ \sum x^2 = 41658,\ \sum y^2 = 17206$

16. While studying about the relationship between scores on statistics (X) and scores on accountancy (Y) the following regression equations were obtained.

Regression equation of Y on x: $3y - 2x - 100 = 0$

Regression equation of X on Y: $4y - 3x + 50 = 0$

a) Find the correlation coefficient.

b) Estimate the scores on Statistics if a student got 50 score in Accountancy.

17. Find the mean values of the variables X and Y for the following regression equations.

Regression line of Y on X : $2y - x = 50$

Regression line of X on Y : $3y - 2x = 10$

18. The following regression equations are obtained when studying about the demand (X) and supply(Y) of a group of commodities.

$26 - 3x - 2y = 0$

$31 - 6x - y = 0$

a) Identify regression lines and find the correlation coefficient.

b) Find the mean values of the variables X and Y.

19. In the study of regression lines, regression coefficient of Y on X = 0.75, correlation coefficient =0.5, standard deviation of Y= 4. Find SD of X.

20. In the study of regression lines, regression coefficient of X on Y = $\frac{3}{5}$, variance of Y=30, correlation coefficient = $\frac{5}{6}$. Find variance of X.

21. In a regression analysis of the income tax of government employees in thousands (X) and their annual income in lakhs (Y), the following regression equations have been obtained.

    $25x - 10y + 10 = 0$

    $10y - 7x - 100 = 0$

    a) Identify regression lines and hence find the correlation coefficient.
    b) Find the mean values of the variables X and Y.
    c) If variance of X = 36, find the variance of Y.

22. A regression analysis on the income in thousands (Y) and expenditure in thousands (X) resulted in the following regression equations.

    $x - y - 3 = 0$ and $5x - 8y + 15 = 0$

    a) Identify regression lines.
    b) What is the correlation coefficient between income and expenditure?
    c) Find the values of $\overline{X}$ and $\overline{Y}$.
    d) What is the most probable value of income when expenditure is 2000?

23. If two lines of regression are $4x - 5y + 10 = 0$ and $20x - 9y - 75 = 0$:

    a) Which of these lines is regression equation of X on Y?
    b) Find correlation coefficient.
    c) Find standard deviation of y if standard deviation of X = 5.

24. The equation of two regression lines between two variables are expressed as $2x - 3y = 0$ and $4y - 5x - 7 = 0$.

    a) Identify which of two can be called regression line of Y on X and X on Y.
    b) Find the correlation coefficient.
    c) Find mean value of X and mean value of Y.

Chapter 3

Elementary Calculus

Calculus is one of the most important branches of Mathematics. It has two parts.

1) Differentiation

2) Integration.

In this chapter we study elementary calculus which is usually used as a supplement for statistical analysis. Let X and Y be two variables which are related in such a way that the changes in Y is dependes on the changes in X. Then the relationship between X and Y is denoted by Y = f(X). Here X is called the independent variable and Y is the dependent variable.

Significant Learning Outcomes

After the completion of this chapter, the learner:

- Identifies domain and range of a function.
- Explains the concept of differentiation.
- Uses the concept of first order and second order derivatives.
- Explains the concept of integration.
- Uses definite integrals in suitable situations.
- Uses the concept of calculus in statistics.

The set of values that can be assumed by X is defined as the **domain** of the function. The values of Y that can be assumed in relation to the values of X is called the **range** of the function.

Consider the functional relationship between X and Y as $Y = 2X + 3$. Here X is the independent variable and Y is the dependent variable. That is, value of Y is influenced by value of X. X and Y can take values in the interval $(-\infty, +\infty)$.

Let $S = 5 + 2p$ represents a supply function where 'S' denotes supply and 'p' is the price. The domain is the set of positive real numbers, since price cannot be negative, ie., $p \geq 0$. Then the range of S will take non-negative values. In the case of demand function, $D = 12 - 3p$, where D and p are demand and price. Price is the independent variable and demand is dependent on price. Here price 'p' assumes values in the range of $0 \leq p \leq 4$. When p > 4, demand becomes negative, which is not possible. Therefore domain of the function is $0 \leq p \leq 4$. When p = 0, demand D = 12. When p = 4, Demand D = 0. Therefore the range of D is $0 \leq D \leq 12$.

Know your progress

Identify the dependent and independent variables in the following equations.
(i) X = 3Y + 8 (ii) S = D - 2p (iii) Y = 6X - 9 (iv) K = Z + $\frac{1}{3}$.
Also find the range and domain of the above functions.

3.1 Derivative of a Function

Since both X and Y are variables, we have to calculate a rate of change in the function, when there is a change in the independent variable. The rate of change in a function with respect to a change in the independent variable is called *derivative*. The process of finding the derivative of a function is termed as *differentiation*.

Consider the function $Y = 3X + 5$. The possible values of Y for some values of X may be tabulated as follows.

X	-2	-1	0	1	2	3
Y	-1	2	5	8	11	14

It is observed that when X is increased by 'one' unit, Y is increased by '3' units. The unit changes in X is denoted by 'Δx' and unit changes in Y is 'Δy'. Then the rate of change in Y with respect to a change in X is denoted by $\frac{\Delta y}{\Delta x}$. Here $\frac{\Delta y}{\Delta x} = 3$. In the

case, where the change in independent variable is by discrete units, limiting value of the ratio $\frac{\Delta y}{\Delta x}$ when $\Delta x \to 0$ is called derivative. In the above example the derivative of y = 3x + 5 is 3. Derivative of y with respect to x is denoted by $\frac{dy}{dx}$ or y' or $f'(x)$.

$\frac{dy}{dx} = \lim_{\Delta x \to 0} \frac{\Delta y}{\Delta x}$. It is the derivative of y with respect to indipendent variable x.

Know your progress

Find the derivatives of the following.

1. $y = 5 - 3x$
2. $y = 8x$
3. $y = \frac{x}{5} + 2$
4. $y = 3 - 0.6x$

Standard formulae

Function	Derivative $\frac{dy}{dx}$
K, a constant	0 (Zero)
x	1 (One)
x^n	nx^{n-1}
$kf(x)$	$k\frac{df(x)}{dx}$
$f \pm g$	$\frac{df}{dx} \pm \frac{dg}{dx}$

(Where f and g are functions of x)

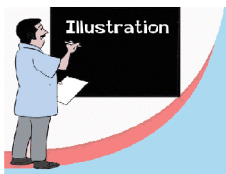


Illustration3.1

Find the derivative of $y = x^{-6}$

Solution

$$\frac{dy}{dx} = -6 \times x^{(-6-1)} = -6x^{-7}$$

Know your progress

Find the derivatives of the following

1. $y = x^{10}$ 2. $y = 2x^5$ 3. $y = x^4 + 3x$

Note: *If R(x) is the revenue function, then* $\frac{dR(x)}{dx}$ *is known as the Marginal Revenue (MR) function. If C(x) is the cost function, then* $\frac{dC(x)}{dx}$ *is known as the Marginal Cost (MC) function. If P(x) is the profit function then* $\frac{dP(x)}{dx}$ *is know as the Marginal Profit (MP) function.*

3.2 Second Order Derivative

Let $y = f(x)$, Then $\frac{dy}{dx}$ is the derivative of y with respect to x. $\frac{dy}{dx}$ is also termed as first order derivative. If we differentiate the function $\frac{dy}{dx}$ again, the result denoted by $\frac{d^2y}{dx^2}$ is called second order derivative.

Let $y = f(x)$. Then the derivative of $f(x)$ is $\frac{dy}{dx}$.

When we differentiate $\frac{dy}{dx}$ again, we get,

$\frac{d}{dx}\left(\frac{dy}{dx}\right) = \frac{d^2y}{dx^2}$ which is known as the second order derivative.

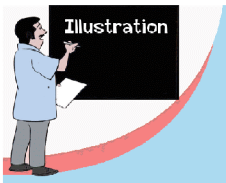


Illustration 3.2

1. $y = x^3 + x + 2$, find $\frac{d^2y}{dx^2}$.

Solution

$$\frac{dy}{dx} = 3x^2 + 1$$

$$\frac{d^2y}{dx^2} = \frac{d}{dx}(3x^2 + 1) = 6x$$

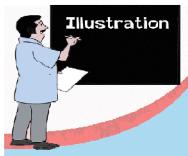


Illustration 3.3

If $y = 4x^3 - 3x^2 + 2x$, find second order derivative.

Solution

$$\frac{dy}{dx} = 12x^2 - 6x + 2$$

$$\frac{d^2y}{dx^2} = \frac{d}{dx}(12x^2 - 6x + 2) = 24x - 6$$

Know your progress

Find $\frac{dy}{dx}$ and $\frac{d^2y}{dx^2}$ of the following functions.

(i) $y = 3x^2 - 26x + 111$ (ii) $y = x^2 - 3x$

(iii) $y = 4x^3 - 3x^{-3} + 6$ (iv) $y = 4x^2 - 24x + 9$

3.3 Applications of second order derivatives

Second order derivatives are helpful in determining maximum or minimum value of a function.

A function attains maximum value at the point where it satisfies the following conditions.

1. $\frac{dy}{dx} = 0$

2. $\frac{d^2y}{dx^2} < 0$

Similarly a function attains the minimum value at a point where the function satisfies the following properties.

1. $\frac{dy}{dx} = 0$

2. $\frac{d^2y}{dx^2} > 0$

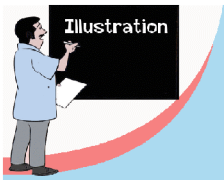


Illustration 3.4

Find the maximum and minimum values of $f(x) = x^3 - 27x + 3$

Solution

Let $y = x^3 - 27x + 3$

$$\frac{dy}{dx} = 0$$

$$\frac{dy}{dx} = 3x^2 - 27 = 0$$

$$3x^2 = 27$$

$$x^2 = 9 \Rightarrow x = -3 \text{ or } x = +3$$

$$\frac{d^2y}{dx^2} = 6x$$

Case (1) When x = -3, then $\frac{d^2y}{dx^2} = 6 \times -3 = -18 < 0$

By conditions of maxima, the function attains maximum at x= –3

The maximum value of function is

$f(x) = x^3 - 27x + 3$

at $x = -3$, $f(x) = (-3)^2 - 27 \times (-3) + 3 = -27 + 81 + 3 = 57$

Case (2) When x=+3, then $\frac{d^2y}{dx^2} = 6 \times +3 = +18 > 0$

By conditions of minima, the function attains minimum at x = +3

The minimum value of function is

$f(x) = x^3 - 27x + 3$

at $x = +3$, $f(x) = (3)^3 - 27 \times 3 + 3 = 27 - 81 + 3 = -51$

Know your progress

Find the maximum and minimum values for the following functions.

i) $f(x) = 2x^3 - 3x + 5$ ii) $f(x) = x^3 - 5x + 14$

3.4 Integration

Integration is the method of determining a function whose derivative is known. It is also used to find the area bounded by the graph of the function under certain conditions. It is also used in mathematical situations and probability.

Integral of a function f (x) is denoted by $\int f(x)dx$. It is function F(x) such that $\frac{d}{dx}F(x) = f(x) + c$, where c is a constant of integration.

Table showing different functions, their derivatives and integral.

Function	Derivative $\frac{dy}{dx}$	Integral of the function
k, constant	0 (zero)	$\int 0dx = k + c$
kx	k	$\int kdx = kx + c$
x	1 (One)	$\int dx = x + c$
$\frac{x^{n+1}}{n+1}$	x^n	$\int x^n dx = \frac{x^{n+1}}{n+1}$
$ax + b$	a	$\int adx = ax + c$
$ax^2 + bx + c$	$2ax + b$	$\int (2ax + b)dx = ax^2 + bx + c$

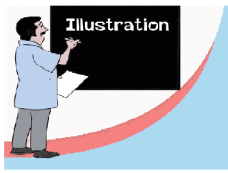


Illustration 3.5

Find $\int x^2 dx$.

Solution: $\int x^2 dx = \frac{x^3}{3} + c$

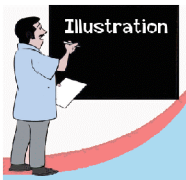


Illustration 3.6

Find $\int (6x - 2x^2)dx$

Solution:

$$\int (6x - 2x^2)dx = \int 6xdx - \int 2x^2 dx = \frac{6x^2}{2} - \frac{2x^3}{3} + c$$

$$= 3x^2 - \frac{2}{3}x^3 + c$$

Know your progress

Find (1) $\int x^6 dx$ (2) $\int (x^3 + 4)dx$

(3) $\int (1 - 3x)$ (4) $\int (x^4 - 2x + 3)dx$

Note: *Revenue function* $R(x) = \int MRdx$ *, where MR is the marginal revenue function.*

Cost function $C(x) = \int MCdx$ *, where MC is the marginal cost function.*

Profit function $P(x) = \int MPdx$ *, where MP is the marginal profit function.*

Illustration 3.7

If marginal cost is $5x^2 - 6x + 8$, where x is the number of output units, find the total cost function.

Solution:

$$\text{Total Cost TC} = \int (\text{Marginal cost})dx$$

$$\text{TC} = \int (5x^2 - 6x + 8)dx$$

$$= \int 5x^2 dx - \int 6xdx + \int 8dx$$

$$= \frac{5x^3}{3} - \frac{6x^2}{2} + 8x + c = \frac{5x^3}{3} + 3x^2 + 8x + c$$

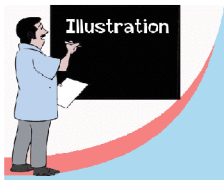


Illustration 3.8

If the Marginal revenue function is $16p+9$ where p is the number of units sold, find total revenue.

Solution:

Total Revenue, TR $= \int MRdp$

$$\text{TR} = \int (16p+9)dp = \frac{16p^2}{2} + 9p + c = 8p^2 + 9p + c$$

3.5 Definite Integrals

Definite Integrals are used to solve the problems of applications of differentiation. They are used to solve problems in Probability, Mathematical Statistics, Economics and Commerce. Definite integrals are used to find the area of a region bounded by the graph of a function under certain conditions. These integrals have a definite value. Consider the area under the curve $y = f(x)$.

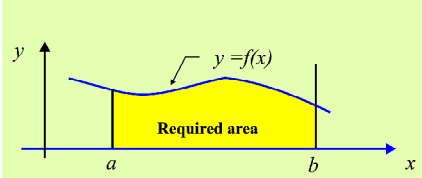


Fig. 3.1

The area under a curve $y = f(x)$ from $x=a$ to $x=b$ is given by the **definite integral.**

Area $= \int_a^b f(x)dx$, where f(x) is the functional form of the curve. 'a' and 'b' are the lower and upper boundaries of the area

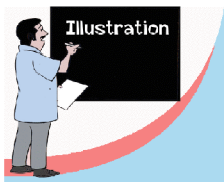


Illustration 3.9

$$\int_1^{10} 3x^2 dx$$

Solution

$$\int_1^{10} 3x^2 dx = \left[3 \times \frac{x^3}{3}\right]_1^{10} \quad \text{(We integrate)}$$

$$= \left[x^3\right]_1^{10}$$

$$= 10^3 - 1^3 = 1000 - 1 = 999$$

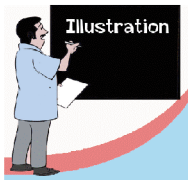


Illustration 3.10

Evaluate $\int_0^1 (x^2 - x)dx$

Solution

$$\int_0^1 (x^2 - x)dx = \left(\frac{x^3}{3} - \frac{x^2}{2}\right)_0^1$$

$$= \left(\frac{1^3}{3} - \frac{1^2}{2}\right) - \left(\frac{0^3}{3} - \frac{0^2}{2}\right)$$

$$= \left(\frac{1}{3} - \frac{1}{2}\right) - 0 = -\frac{1}{6}$$

Know your progress

Evaluate the following

(1) $\int_1^2 x^3 dx$

(2) $\int_{-1}^1 x dx$

(3) $\int_{-2}^2 3x^2 dx$

(4) $\int_1^2 (x^2 - 2x + 1)dx$

Let us conclude

In this chapter, we were familiarized with functional relationship, domain and range of a function. The fundamentals of differentiation and integration were discussed in detail. Differentiation is the process of finding the ratio of change in the dependent variable with respect to the independent variable. Integration can be considered as the reverse process of differentiation. The methods of finding derivatives and integrals and some standard results were discussed. The concept of definite integral was also exemplified.

Let us assess

For Questions 1-10, choose the correct answer from the given choices.

1. $\frac{d(x^8)}{dx} = \ldots\ldots\ldots\ldots$

   a) x^9 b) $8x^7$ c) $\frac{x^9}{9}$ d) $\frac{x^7}{8}$

2. If derivative of x^{10} is kx^9 then the value of k is .....

   a) 10 b) 9 c) 0 d) 1

3. $\frac{d}{dx}(8x+10) = \ldots\ldots\ldots\ldots$

   a) 0 b) 8 c) 18 d) 10

4. Integral of $\frac{1}{x^3}$ is .....

   a) $-3x^{-4}$ b) $3x^2$ c) $-0.5x^{-2}$ d) x^2

5. $\int_0^1 x^2 dx = \ldots\ldots\ldots\ldots$

   a) 1 b) 2 c) 0 d) $\frac{1}{3}$

6. If $\int x^n dx = \frac{x^4}{4}$, then the value of n is .....

   a) 4 b) 3 c) 0 d) 5

7. Differentiate the following functions.

   a) $y = x^5$ b) $y = x^2 + 5x + 8$ c) $y = x^3 + 7x^2 + 10x + 6$

   d) $y = 8x - 10x^2$ e) $y = 10 - 2x$ f) $y = ax + b$

   g) $(x-a)(x+a)$ h) x^{-3}

8. Find second order derivative for the following.

   a) $y = 20x^3$ b) $y = 4x^3 - 20x^2 + 5x - 9$

   c) $y = (x+2)(x^2+3x+5)$

9. Find maximum and minimum value of the following functions, if they exist.

a) $y = x^3 - 12x$ b) $f(x) = x^3 - 6x^2 + 12x - 8$

c) $f(x) = 9x^2 + 12x + 2$ d) $f(x) = (x-1)(x-2)$

e) $f(x) = 2x^3 - 24x + 107$

10. If the total revenue function of a firm is given by $R(x) = 22x - x^2$, where x is the number of units sold, find marginal revenue function.

11. The cost of manufacturing x items is given by $c(x) = 2x^2 - 16x + 10$. To have minium cost, how many items to be manufactured?

12. Find the maximum profit that a company can make, if the profit function $p(x) = 500 - 72x + 4x^2$.

13. Integrate the following.

a) x^8 b) x^{24} c) x^{-12}

d) $\frac{1}{x^6}$ e) $3x^2 + 5x - 2$ f) $25x^4 - 16x^3 + 10$

g) $4x^5 + 3x^{-4} + 7$

14. Evaluate the following defenite integrals.

a) $\int_{1}^{-3} x^2 dx$ b) $\int_{1}^{-2} (4x^3 - 3x^2 + 6x + 9) dx$

c) $\int_{0}^{-8} x^{\frac{5}{3}} dx$ d) $\int_{-1}^{1} (x + 1) dx$

15. If marginal profit is $4 - 6x$, where x is the number of units of production, find the profit function.

16. ABC company find that profit of the company is given by $p(x) = 2x - \frac{x^2}{400} - 75$ where x is quantity of product. Estimate the Maximum profit that company can make.

Chapter 4

Random Variables

We have already learnt about random experiments in the previous year. Consider the random experiment of tossing two coins simultaneously. The sample space associated with this experiment consists of four sample points - HH, HT, TH, and TT. Let X represents the number of tails obtained. Then X takes the values 0, 1 and 2.

X=0: when sample point is HH,

X=1: when sample points are HT and TH. X=2: when the sample point is TT.

Significant Learning Outcomes

After the completion of this chapter, the learner:

- Identifies the importance of random variables.
- Differentiates discrete and continuous random variables.
- Recognises discrete and continuous probability distributions.
- Explains mathematical expectation of random variables.
- Evaluates mean and variance of random variables.

If we conduct the same experiment with fifty or hundred coins, the process is cumbersome. Thus to interpret the sample space more conveniently, we have introduced a variable X, called the Random Variable ,whose values are determined by the results of the random experiment. We can also consider X as a function with sample space S as domain and range as $\{0,1,2\}$. Here in this chapter,we learn about Random variables, its properties and applications.

4.1. Random Variable

A Random Variable is a real valued function defined over the sample space. We can ascertain different probabilities for different values of the Random variable.

> When the value of a variable is determined by the outcome of a random experiment, that variable is called a Random variable. It is a real valued function defined over the sample space.

Examples:

The height of a person, the number of errors in a phone book, today's temperature, the scores of students in an examination, etc.

Usually capital letters of English alphabet is used to represent Random variables.

There are two types of Random variables, they are :

1) Discrete random variable 2) Continuous random variable.

Let us discuss them in detail.

4.2. Discrete Random Variable

Consider the following Random variables.

1. Number of children in a family.
2. Number of telephone calls recieved at a call centre in a month.

In the first case, the number of possible values of the random variable is finite. In the second case, the number of posssible values is countably infinite. In both cases the random varible is said to be discrete random variable.

A random variable is said to be a discrete random varible if it assumes finite or countably infinite number of values.

Discrete random variables are random variables associated with discrete variables.

Examples of discrete random variables are :

- Number of houses in a certain village of a district.
- Number of students in a classroom of a school.
- Number of complaints received at the office of an airlines during a day.
- Number of customers who visit a bank during an hour.

In all these examples, the range of a random variable contains a finite number of values or countable infinite number of values. Hence they are Discrete Random Variable.

Example 1:

Consider the tossing of a coin. Let X represent the number of heads obtained.

Head Tail

X = 1 X = 0

X takes the values **1** and **0**

i.e., with each sample point H and T ,we can associate a number for X as shown in the table given below.

Sample point	T	H
X	0	1

X is a discrete random variable.

Example 2: Suppose you flip a coin twice. This simple statistical experiment can have four possible outcomes: HH, HT, TH, and TT. Let the random variable X represent the number of Heads. The random variable X can take the values 0, 1 or 2; so it is a discrete random variable.

Sample points	TT	HT	TH	HH
X	0	1	1	2

Example 3: Consider tossing of three coins

Sample space ={ HHH,HHT,HTH,THH,TTT,TTH,THT,HTT }

Let X represent the number of heads obtained, then X takes the values 0, 1, 2, or 3.

X = "The number of Heads" is the random variable.

In this case, there could be 0 Head (if all the coins land tails up), 1 Head, 2 Heads or 3 Heads.

The three coins can land in eight possible ways:

Looking at the table given below we can see that 1 case of three Heads, 3 cases of two Heads, 3 cases of one Head, and 1 case of zero Heads.

Sample points:	HHH	HHT	HTH	THH	TTH	THT	HTT	TTT
X :	3	2	2	2	1	1	1	0

Hence X is a discrete random variable

Example 4: A die is rolled two times. Let X represents the sum of face values turns up. Then X can have the values 2, 3, 4, 5, 6, 7, 8, 9, 10 ,11 and 12.

So X is a discrete random variable.

Example 5: Let X = number of tosses of a coin until a head appears. The random variable takes the values X = {1, 2, 3, 4, 5, ...}, here the values of X are countably infinite. X is a discrete random variable.

If X and Y are any two random variables and 'a' and 'b' are constants, then

$X + a$, aX, $X + Y$ and XY, $\frac{1}{X}$, $X - Y$, $aX + bY$

are also random variables.

4.3 Probability mass function (pmf)

Suppose X is a discrete random variable taking distinct values $x_1, x_2, x_3, x_4, \ldots$ with probabilities $p(x_1), p(x_2), p(x_3), p(x_4) \ldots$ respectively, such that:

i) every probability is a non negative number. i.e., $p(x) \geq 0$

ii) the sum of all the probabilities is always unity. i.e., $\sum p(x) = 1$

Then the array of values of X and their probabilities is called probability distribution of the discrete random variable.

X	x_1	x_2	x_3	x_4	.....	Total
P(x)	$p(x_1)$	$p(x_2)$	$p(x_3)$	$p(x_4)$	.....	1

The probability function for a discrete random variable X gives P(X=x) for every value x that X can take. This is known as Probability mass function.

i.e, $P(X = x) = P(x)$

The graphical representation of the above probability distribution is as shown below.

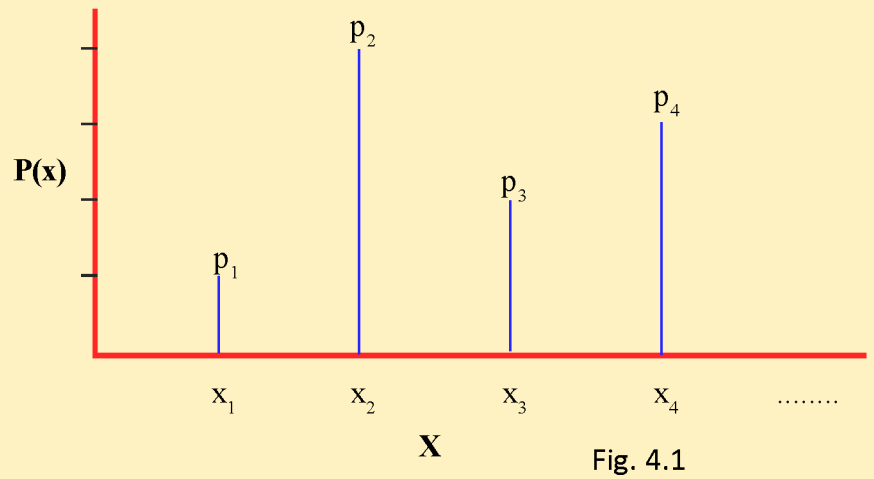


Fig. 4.1

Properties of pmf

1) $p(x) \geq 0$ for all x.

2) $\sum p(x) = 1$.

3) $P(a < X < c) = P(a < X \leq b) + P(b < X < c)$ where b lies between a and c.

4) $P(x_1 + x_2) = P(x_1) + P(x_2)$.

Example 6 : Consider tossing of two coins. Let X represents the number of heads obtained. Sample points for this statistical experiment is given below.

Sample points	TT	HT	TH	HH
X	0	1	1	2

We can find the probabilities of getting heads as follows.

$P(X = 0) = P(TT) = 1/4$

$P(X = 1) = P(HT, TH) = 2/4$

$P(X = 2) = P(HH) = 1/4$

X = x	0	1	2
P(X = x)	$\frac{1}{4}$	$\frac{2}{4}$	$\frac{1}{4}$

x can be considered as the values taken by X.

Here $p(x) \geq 0$ and $\sum p(x) = 1$

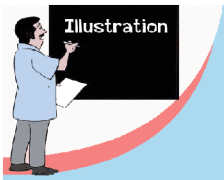


Illustration 4.1

For the following , determine whether the distribution represents a probability distribution, If it does not, state the reason.

x	3	7	9	12	14
P(x)	$\frac{4}{13}$	$\frac{2}{13}$	$\frac{3}{13}$	$\frac{1}{13}$	$\frac{3}{13}$

Solution:

Here $p(x) \geq 0$ for all x and $\sum p(x) = 1$ so the given function is pmf.

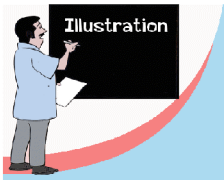


Illustration 4.2

Examine whether the following is a probability distribution of a discrete random variable.

x	1	2	3	4	5
P(x)	$\frac{1}{12}$	$\frac{5}{12}$	$\frac{1}{12}$	$\frac{4}{12}$	$\frac{3}{12}$

Solution:

Here $p(x) \geq 0$ for all x but $\sum p(x) = \frac{14}{12} \neq 1$, so the given function is not a pmf.

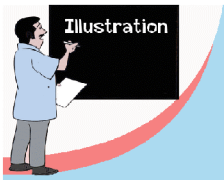


Illustration 4.3

A die is rolled once. Find the probability distribution of the number which turns up.

Solution:

Let X represents the number turns up. X can take the values 1, 2, 3, 4, 5, 6 each with probability $\frac{1}{6}$. The probability distribution of X is,

x	0	2	3	4	5	6
P(x)	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

Here also $p(x) \geq 0$ for all x and $\sum p(x) = 1$. So the given function is pmf.

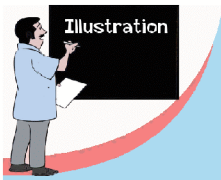


Illustration 4.4

The probability mass function of a random variable X is given by $f(x) = \frac{2x}{k}$; x = 1, 2, 3. Determine k.

Solution:

X	1	2	3
P(X = x)	$\frac{2}{k}$	$\frac{4}{k}$	$\frac{6}{k}$

Given f(x) is a p.m.f. Therefore $p(x) \geq 0$ for all x and $\sum p(x) = 1$.

i.e., $\frac{2}{k} + \frac{4}{k} + \frac{6}{k} = 1 \Rightarrow \frac{12}{k} = 1 \Rightarrow$ k = 12.

Know your progress

1. For the following, determine whether the distribution represents a probability distribution, If it does not, state the reason.

a)

x	4	6	8	10
P(x)	-0.6	0.2	0.7	1.5

b)

x	1	2	3	4
P(x)	$\frac{2}{4}$	$\frac{1}{4}$	0	$\frac{1}{4}$

c)

x	1	3	5	17
P(x)	0.3	0.1	0.2	0.4

d)

x	5	10	15
P(x)	0.3	0.4	0.2

2. If pmf of a discrete random variable X is given by

x	4	8	12	16
P(x)	$\frac{1}{6}$	K	$\frac{1}{2}$	$\frac{1}{12}$

Determine the value of K.

3. If pmf of a discrete random variable X is given by

x	0	1	2	3
P(x)	0.1	0.3	K	0.4

Find a) value of K

b) $P(X < 3)$

c) $P(1 < X < 3)$

4. If $P(x) = \frac{x}{12}$, $x = 2,4,6$

$= 0$, otherwise; is a pmf of X

then find

a) $P(X = 2)$

b) $P(X < 6)$

c) $P(X = 2 \text{ or } 3)$

4.4 Cumulative distribution function (cdf)

The probability of a random variable upto a particular value is called cumulative distribution function of the random variable.

The cumulative distribution function F(x) of a discrete random variable X with p.m.f p(x) is defined as

$$F(x) = P(X \le x) = \sum_{-\infty}^{x} P(x)$$

It is also known as distribution function because it gives cumulative probability of a random variable.

Properties of Distribution functions

1. $F(x) \ge 0$
2. F(x) is non decreasing.
3. F(x) is continuous to right.
4. $F(-\infty) = 0$ and $F(\infty) = 1$.
5. $P(a < x \le b) = F(b) - F(a)$.
6. Graph of F(x) is in the form of steps.

Example 7

Consider flipping of an unbiased coin twice .The probability distribution is:

X= x	0	1	2
P(X=x)	$\frac{1}{4}$	$\frac{2}{4}$	$\frac{1}{4}$

We have, P(X = x) = p(x), and F(x) = $P(X \le x)$

$$F(0) = P(X \le 0) = P(0) = \frac{1}{4}$$

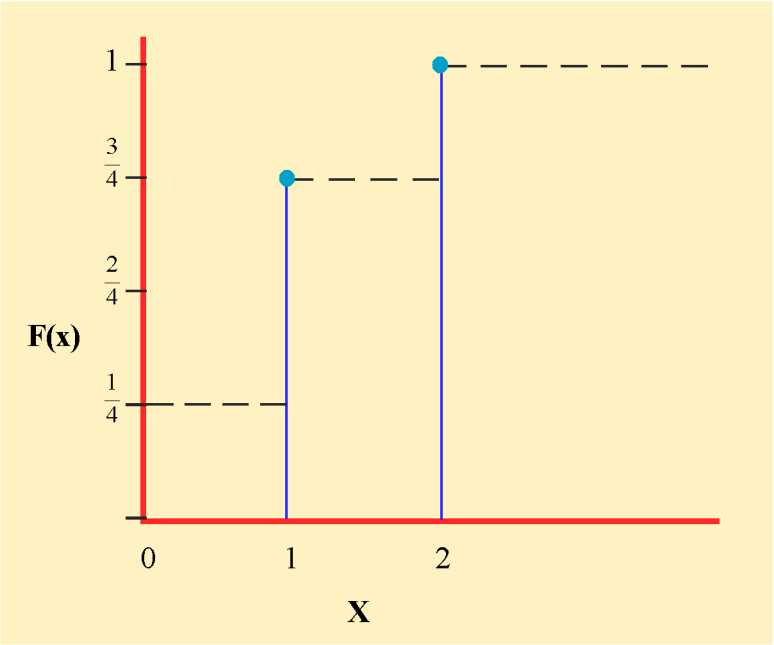
$$F(1) = P(X \le 1) = P(0) + P(1) = \frac{1}{4} + \frac{2}{4} = \frac{3}{4}$$

$$F(2) = P(X \le 2) = P(0) + P(1) + P(2) = \frac{1}{4} + \frac{2}{4} + \frac{1}{4} = 1$$

Distribution function of X is as follows :

X	0	1	2
F(x)	$\frac{1}{4}$	$\frac{3}{4}$	$\frac{4}{4}=1$

Graphical representation is given below.



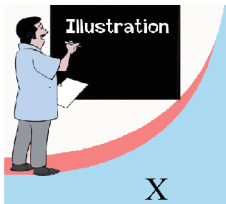


Illustration 4.5

Consider a random variable X which takes the values as given below. Find the distribution function F(x).

X	0	1	2	3	4	5
P(x)	0.1	0.2	0.3	0.2	0.1	0.1

Solution:

X	P(x)	F(x)
0	0.1	0.1
1	0.2	0.3
2	0.3	0.6
3	0.2	0.8
4	0.1	0.9
5	0.1	1

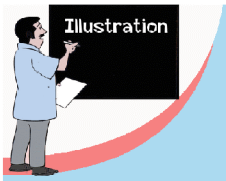


Illustration 4.6

Find the probability distribution of boys and girls in families with 3 children assuming equal probabilities for boys and girls.

Solution:

The probability of child being a boy or a girl is $\frac{1}{2}$. The following probabilities may arise:

(i) All are boys.

$$\text{The } P(BBB) = \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} = \frac{1}{8}$$

(ii) Two are boys and one is a girl.

$$\text{So } P(BBG + BGB + GBB) = \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} + \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} + \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} = \frac{3}{8}$$

iii) Two are girls and one is a boy.

$$\text{So P (GGB + GBG + BGG)} = \frac{3}{8}$$

(iv) All are girls.

$$\text{So P (GGG)} = \frac{1}{8}$$

Thus the probability distribution is as follows :

Number of boys (x)	0	1	2	3
P(x)	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$

4.5 Mathematical Expectation, Mean and Variance

Let X is a discrete random variable taking values $x_1, x_2, x_3, \ldots$ with respective probabilities $p_1, p_2, p_3, \ldots$ Then Mathematical expectation of X, usually denoted by E(X), is defined as

$$E(X) = \Sigma x P(x).$$

E(X) is known as mean of the random variable.

It does not occur for all random variables.

Properties of E(X)

If X represents a random variable, 'a' and 'b' are any two constants, then :

i) $E(a) = a$

ii) $E(aX) = aE(X)$

iii) $E(X + b) = E(X) + b$

iv) $E(aX + b) = aE(X) + b$

v) $E(X \pm Y) = E(X) \pm E(Y)$ where X and Y are random variables.

Variance of a random variable

Variance of a random variable X is the expectation of the square of the difference between X and its expectation. So variance can be computed by.

$$V(X) = E[X - E(X)]^2$$

$$= E(X^2) - [E(X)]^2 \text{ on simplification}$$

Properties of V(X)

If X represents a random variable, 'a' and 'b' are any two constants, then :

i) $V(a) = 0$

ii) $V(aX) = a^2 V(X)$

iii) $V(aX + b) = a^2 V(X)$

iv) $V(X \pm Y) = V(X) + V(Y)$ where X amd Y are independent random variables.

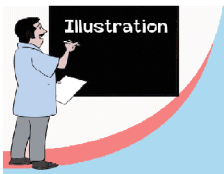


Illustration 4.7

Find the expected value of number of tails appearing,when two coins are tossed.

Solution:

Sample space ={TT, TH, HT, HH }

X represents the number of tails appeared when two coins are tossed.

X takes the values 0,1and 2. P.m.f is given below.

X= x	0	1	2
P(X=x)	$\frac{1}{4}$	$\frac{2}{4}$	$\frac{1}{4}$

$$E(X) = \Sigma x\ P(x) = 0 \times \frac{1}{4} + 1 \times \frac{2}{4} + 2 \times \frac{1}{4} = 1$$

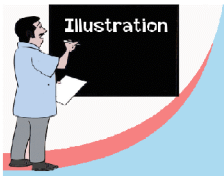


Illustration 4.8

Find the mean of the 'number on the die' when a die is rolled.

Solution:

Let X represents 'the number on the die', when a die is rolled. The probability distribution for the random variable X is as shown below.

X	1	2	3	4	5	6
P(X)	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

Mean $= E(X) = \Sigma x P(x)$

$$= 1 \times \frac{1}{6} + 2 \times \frac{1}{6} + 3 \times \frac{1}{6} + 4 \times \frac{1}{6} + 5 \times \frac{1}{6} + 6 \times \frac{1}{6} = \frac{21}{6} = 3.5$$

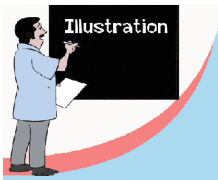


Illustration 4.9

Find the expected value of the number of heads in three tosses of a coin (or a simultaneous toss of three coins)

Solution:

In the tossing of three coins

Sample space={ HHH,HHT,HTH,THH,TTT,TTH,THT,HTT }

Let X represent the number of heads obtained, then X assumes the values 0, 1, 2, or 3.

X	0	1	2	3
P(X)	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$

$$E(X) = \Sigma x P(x) = 0 \times \frac{1}{8} + 1 \times \frac{3}{8} + 2 \times \frac{3}{8} + 3 \times \frac{1}{8} = 1.5$$

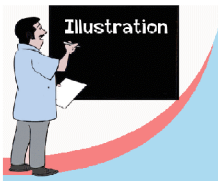


Illustration 4.10

If X assumes the values 1, 2, 3 and the probability mass function is given by

$$p(x) = \frac{x}{6}, \; x = 1, 2, 3$$

$= 0$, otherwise. Determine i)E(X) and ii) V(X)

Solution:

X	1	2	3
P(x)	$\frac{1}{6}$	$\frac{2}{6}$	$\frac{3}{6}$

$$E(X) = \Sigma x P(x) = 1 \times \frac{1}{6} + 2 \times \frac{2}{6} + 3 \times \frac{3}{6} = \frac{14}{6} = 2.33$$

$$E(X^2) = \Sigma x^2 P(x) = 1 \times \frac{1}{6} + 4 \times \frac{2}{6} + 9 \times \frac{3}{6} = \frac{36}{6} = 6$$

$$V(X) = E(X^2) - E(X)^2 = 6 - 5.43 = 0.57$$

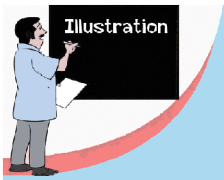


Illustration 4.11

Find the probability distribution of the number of sixes in three tosses of a die .Also find the mean and variance of the distribution.

Solution:

Let X represents the number of sixes obtained in three tosses of a die.

X takes the values 0,1,2,3

Total points in the sample space = $6^3 = 216$

$$P(\text{getting all sixes}) = P(X=3) = \frac{1}{6}\times\frac{1}{6}\times\frac{1}{6} = \frac{1}{216}$$

$$P(\text{getting two sixes}) = P(X=2) = \left(\frac{1}{6}\times\frac{1}{6}\times\frac{5}{6}\right)\times 3 = \frac{15}{216}$$

$$P(\text{getting one six}) = P(X=1) = \left(\frac{1}{6}\times\frac{5}{6}\times\frac{5}{6}\right)\times 3 = \frac{75}{216}$$

$$P(\text{getting no six}) = P(X=0) = \left(\frac{5}{6}\times\frac{5}{6}\times\frac{5}{6}\right) = \frac{125}{216}$$

X	0	1	2	3
P(X)	$\frac{125}{216}$	$\frac{75}{216}$	$\frac{15}{216}$	$\frac{1}{216}$

$$E(X) = \Sigma x P(x) = 0\times\frac{125}{216} + 1\times\frac{75}{216} + 2\times\frac{15}{216} + 3\times\frac{1}{216} = \frac{108}{216} = \frac{1}{2}$$

$$V(X) = E(X^2) - [E(X)]^2 = \frac{5}{12}$$

Know your progress

1.If X assumes the values 1,2,3,4 and the pmf is

$$P(x) = \begin{cases} \frac{x}{10}, x = 1,2,3,4 \\ 0, otherwise \end{cases}$$

Find E(X), V(X) and F(x).

2. Find the probability distribution of Y, the number of sixes in two tosses of a die. (or a simultaneous toss of two dice). Also sketch its graph.
3. An urn contains 4 white and 6 red balls. Four balls are drawn at random from the urn. Find the probability distribution of the number of white balls.

4.6 Continuous Random Variables

Consider the life of a battery. Suppose we can measure it as precisely as we want. It may be 40 hours, or 40.25 hours, or 40.349 hours. Assume that the maximum life of the battery is 300 hours. Let X denote the life of a randomly selected battery of this kind. Then X can assume any value in the interval 0 to 300. Consequently X is a continuous random variable. As shown in the diagram, every point on the line representing the interval 0 to 300 gives a possible value of X.

Every point on this line represents a possible value of X that denotes the life of a battery. There are infinite number of points on this line. The values represented by the points on this line are uncountable.

A random variable that can assume any value contained in one or more intervals is called **Continuous Random Variable.**

Continuous random variables are obtained from data that can be measured rather than counted. They can assume an infinite number of values and may be fractional values .

Examples of continuous random variables are, temperature, lengths, weights, etc. Why temperature is a continuous random variable? Since the variable can assume an infinite number of values between any two given temperatures, it is a continuous random variable.

If a random variable is a continuous variable, its probability distribution is called a continuous probability distribution.

Most often, the equation used to describe a continuous probability distribution is called a **probability density function (pdf)**.

Consider the probability density function shown in the graph. Suppose we want to know the probability that

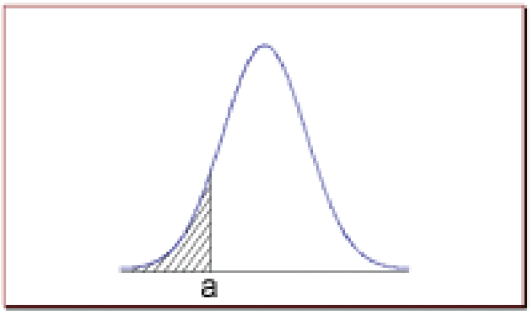


Fig.4.2

the random variable X was less than or equal to a. The probability that X is less than or equal to a is equal to the area under the curve bounded below by a as indicated by the shaded area in the fig. 4.2.

Note: The shaded area in the graph represents the probability that the random variable X is less than or equal to a. This is a cumulative probability. However, the probability that X is *exactly* equal to a would be zero. A continuous random variable can take on an infinite number of values. The probability that it will equal a specific value (such as a) is always zero.

Consider the graph of y = f(x) as given . It can be shown that the area under the curve between the ordinated at x = a and x = b is the probability that X will lie between a and b. This area is the definite integral between a and b which is mathematically denoted by the symbol $\int_a^b f(x)dx$.

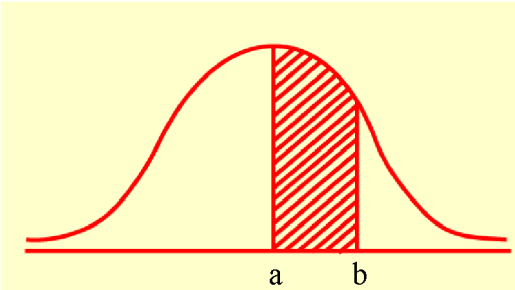


fig. 4.3

That is

$$P(a \le x \le b) = \int_a^b f(x)dx$$

Discrete and Continuous Data

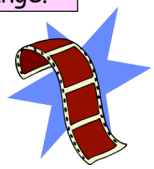
Discrete data can only take on certain individual values.	**Continuous** data can take on any value in a certain range.

Example 1

Number of pages in a book is a **discrete variable.**

Example 2

Length of a film is a **continuous variable.**



Example 3

Shoe size is a **Discrete variable.** E.g. 5, $5\frac{1}{2}$, 6, $6\frac{1}{2}$ etc. Not in between.

Example 4

Temperature is a **continuous variable.**

Example 5

Number of people in a race is a **discrete variable.**

Example 6

Time taken to run a race is a **continuous variable.**

Activity

Prepare a list of discrete and continuous random variables from your daily life. Explain why they are discrete or continuous.

Properties of p.d.f

Let *f(x)* be the p.d.f. of a continuous random variable X. Then it has the following properties.

1. $f(x) \geq 0$ for every x

2. $\int_{-\infty}^{\infty} f(x)dx = 1$, the total probability.

   These two conditions are the sufficient conditions for a function *f(x)* to be a pdf.

3. $P[a \leq X \leq b] = \int_a^b f(x)dx$

4. $P[a \leq X \leq c] = P[a \leq X \leq b] + P[b \leq X \leq c]$

   $= \int_a^b f(x)dx + \int_b^c f(x)dx$ where b lies between a and c

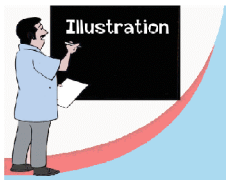


Illustration 4.12

Verify the following is a p.d.f. of a continuous random variable.

$$f(x) = 2x; \text{ for } 0 < x < 1$$
$$= 0; \text{ elsewhere}$$

Solution

Here $f(x) \geq 0$ for every x, condition (1) is verified.

Now $\int_{-\infty}^{\infty} f(x)dx = \int_{-\infty}^{0} f(x)dx + \int_0^1 f(x)dx + \int_1^{\infty} f(x)dx$

$$= 0 + \int_0^1 2x\,dx + 0$$

$$= 2\left(\frac{x^2}{2}\right)_0^1 = 1 - 0 = 1, \text{ condition (2) is verified.}$$

Hence *f(x)* is a p.d.f.

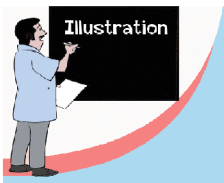


Illustration 4.13

Let X be a continuous random variable with p.d.f

$$f(x) = \frac{1}{8}; \quad 1 \le x \le 9$$

$$= 0; \quad \text{elsewhere.}$$

Find (i) $P(2 < X < 5)$ (ii) $P(X < 3)$

(iii) $P(X \ge 3)$ (iv) $P(|X-2| > 3)$

Solution

(i) $P(2 < X < 5) = \int_2^5 f(x)dx$

$$= \int_2^5 \frac{1}{8}dx$$

$$= \frac{1}{8} \times [x]_2^5 = \frac{1}{8} \times (5-2) = \frac{3}{8}$$

(ii) $P(X < 3) = \int_1^3 f(x)dx$

$$= \int_1^3 \frac{1}{8}dx$$

$$= \frac{1}{8} \times [x]_1^3 = \frac{1}{8} \times (3-1) = \frac{1}{4}$$

(iii) $P(X \ge 3) = 1 - P(X < 3)$

$$= 1 - \frac{1}{4} = \frac{3}{4}$$

(iv) P(|X-2| > 3) $= 1 - P(|X - 2| \le 3)$

$= 1 - P(-3 \le X - 2 \le 3)$

$= 1 - P(-1 \le X \le 5)$

$= 1 - \int_{-1}^{5} f(x)dx = 1 - \frac{4}{8} = \frac{1}{2}$

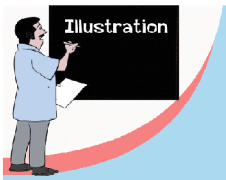


Illustration 4.14

If X has the p.d.f.

$f(x) = kx; \quad 0 \le x \le 2$

$= 0; \quad$ otherwise. Find the value of k.

Solution:

Since $f(x)$ is a p.d.f., we have $\int_{0}^{2} f(x)dx = 1$

i.e., $\int_{0}^{2} kx\,dx = 1$

i.e., $k\left(\frac{x^2}{2}\right)_0^2 = 1$

i.e., $\frac{k}{2}(4-0) = 1$

i.e., $k = \frac{1}{2}$

4.7 Distribution Function

Let X be a continuous random variable defined over $(-\infty, \infty)$ with p.d.f. $f(x)$. Then the distribution function, usually denoted by $F(x)$, is defined as

$$F(x) = P(X \le x)$$

$$= \int_{-\infty}^{x} f(x)dx$$

It is also known as the cumulative distribution function (c.d.f.)

Properties of the Distribution Function

If $F(x)$ is the distribution function of a continuous random variable X, then it has the following properties.

1. $0 \le F(x) \le 1$
2. $F(-\infty) = 0$ and $F(\infty) = 1$
3. F(x) is non decreasing and continuous to the right
4. $P(a \le x \le b) = F(b) - F(a)$ where $a < b$

Remark: If the derivative of $F(x)$ exists, the p.d.f. of X is given by $f(x) = \dfrac{dF(x)}{dx}$

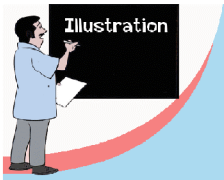


Illustration 4.15

The distribution function of a continuous random variable is given by

$$F(x) = \begin{cases} 0 & \text{for } x \le 0 \\ x^2 & \text{for } 0 < x < 1 \\ 1 & \text{for } x \ge 1 \end{cases}$$

Find the p.d.f. of X

Solution:

We have $f(x) = \dfrac{dF(x)}{dx}$

$$f(x) = \begin{cases} 2x & \text{for } 0 < x < 1 \\ 0 & \text{elsewhere} \end{cases}$$

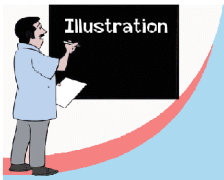


Illustration 4.16

A random variable X has the p.d.f. $f(x) = \dfrac{1}{4}$ for $-2 \le x \le 2$.

Obtain the distribution function of X.

Solution

Distribution function is given by F(X) $= \int_{-\infty}^{x} f(x)dx$

When x < -2 F(X) $= \int_{-\infty}^{x} f(x)dx$

$$= \int_{-\infty}^{x} 0\,dx = 0$$

When $-2 \le x \le 2$ F(X) $= \int_{-\infty}^{x} f(x)dx$

$$= \int_{-\infty}^{-2} f(x)dx + \int_{-2}^{x} f(x)dx$$

$$= \int_{-\infty}^{-2} 0\,dx + \int_{-2}^{x} \frac{1}{4}dx$$

$$= \frac{1}{4}[x]_{-2}^{x} = \frac{1}{4}(x-(-2)) = \frac{x+2}{4}$$

When $x > 2$ F(X) $= \int_{-\infty}^{x} f(x)dx$

$$= \int_{-\infty}^{-2} f(x)dx + \int_{-2}^{2} f(x)dx + \int_{2}^{\infty} f(x)dx$$

$$= \int_{-\infty}^{-2} 0\,dx + \int_{-2}^{2} \frac{1}{4}dx + \int_{2}^{\infty} 0\,dx$$

$$= 0 + \frac{1}{4}[x]_{-2}^{2} + 0$$

$$= \frac{1}{4}[2-(-2)] = \frac{2+2}{4} = 1$$

Therefore, Distribution function is

$$F(x) = \begin{cases} 0 & \text{when } x \le -2 \\ \dfrac{x+2}{4} & \text{when } -2 \le x \le 2 \\ 1 & \text{when } x > 2 \end{cases}$$

Mean and Variance of a continuous random variable

Let $f(x)$ be the p.d.f. of continuous random variable X. Then,

Arithmetic Mean of continuous random variable X = E(X) = $\int_{-\infty}^{\infty} xf(x)dx$, provided it exists.

Variance of continuous random variable X = V(X) $= E[X-E(X)]^2$

$$= \int_{-\infty}^{\infty} [X-E(X)]^2 f(x)dx$$

V(X) $= E(X^2)-[E(X)]^2$ Where $E(X^2) = \int_{-\infty}^{\infty} x^2 f(x)dx$

Standard deviation of X = $\sqrt{V(X)}$

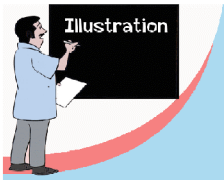


Illustration 4.17

Find the Mean and Variance of X. If p.d.f. of X is given by

$f(x) = \frac{x}{2}$; $0 \le x \le 2$

$= 0$; otherwise

Solution

Mean of X = E(X) $= \int_{-\infty}^{\infty} xf(x)dx$

$$= \int_{-\infty}^{0} xf(x)dx + \int_{0}^{2} xf(x)dx + \int_{2}^{\infty} xf(x)dx$$

$$= \int_{-\infty}^{0} x.0dx + \int_{0}^{2} x.\frac{x}{2}dx + \int_{2}^{\infty} x.0dx$$

$$= 0 + \int_{0}^{2} \frac{x^2}{2}dx + 0$$

$$= \left[\frac{x^3}{6}\right]_0^2 = \frac{1}{6}(8-0) = \frac{4}{3}$$

Now $E(X^2) = \int_{-\infty}^{\infty} x^2 f(x)dx$

$$= \int_{-\infty}^{0} x^2 f(x)dx + \int_{0}^{2} x^2 f(x)dx + \int_{2}^{\infty} x^2 f(x)dx$$

$$= \int_{-\infty}^{0} x^2 .0dx + \int_{0}^{2} x^2 .\frac{x}{2}dx + \int_{2}^{\infty} x^2 .0dx$$

$$= 0 + \int_{0}^{2} \frac{x^3}{2}dx + 0 = 2$$

V(X) $= E(X^2) - [E(X)]^2$

$$= 2 - \left(\frac{4}{3}\right)^2 = 2 - \frac{16}{9} = \frac{2}{9}$$

Know your progress

1. The probability density of the continuous random variable X is given by

$f(x) = \frac{1}{5}$, for $2 < x < 7$

$= 0$, otherwise

a) Draw the graph of f(x)

b) Find $P(3 < x < 5)$

2. The density function of the random variable X is given by

$f(x) = 6x(1 - x)$, for $0 < x < 1$

$= 0$, elsewhere

Find $P\left(X < \frac{1}{4}\right)$ and $P\left(X > \frac{1}{2}\right)$

Let us conclude

We have studied about the variables. Variation is inherent in nature. The variation related to a random experiment or probability is classified into two; discrete and continuous. Based on the range of values of the random variable we can identify them. Thus we get probability mass function and probability density function .Mean and variance of the random variables can be calculated using mathematical expectation.

Let us assess

1. Classify the following variables as discrete or continuous.

   a. Ages of people working in a large factory.

   b. Number of cups of coffee served in a restaurant.

   c. The quantity of drug injected into a guinea pig.

   d. The time taken by a student to reach the school.

   e. The number of gallons of milk sold each day at a grocery store.

   f. Number of pizzas sold through an ice cream parlour each day.

   g. Relative humidity levels in the operation rooms at local hospitals.

   h. Number of bananas in a bunch at several local super markets.

   i. Life (in hours) of 15 iPod batteries.

   j. Weights of the school bags of first standard students in a school bus.

   k. Number of students who make appointments with a statistics teacher in a higher secondary school.

   l. Blood pressures of runners in a marathon.

For Questions 2-17, choose the correct answer from the given choices.

2. The number of children in a family is an example of ——— variable.

   (a) qualitative (b) discrete (c) continuous. (d) normal

3. A real valued function defined over the..... of a random experiment is the random variable.

(a) Event (b) domain (c) range (d) sample space

4. Range of a random variable will be:

(a) always zeros and ones (b) variables

(c) all real numbers (d) all numbers

5. The result of throwing a die thrice is an example of.....

(a) discrete variable (b) continuous variable

(c) both discrete and continuous variable

(d) neither discrete nor continuous variable

6. Mean of a random variable is given by:

(a) E(x) (b) V(x) (c) F(x) (d) P(x)

7. Mean of a random variable X is 24, then E(X+5) =.....

(a) 24 (b) 5 (c)29 (d) 120

8. The mathematical expectation of square of deviation of a random variable from its arithmetic mean is known as....of the variable.

(a) mean (b) variance (c) Mode (d) median

9. V(X) = 4 , then V(2x+4) =.....

(a) 4 (b) 16 (c) 20 (d)36

10. The total area under a probability curve is ...

(a) 1 (b) 0 (c) infinity (d) cant be calculated

11. If in a table all possible values of a random variable are given along with their corresponding probabilities, then this table is called:

(a) Probability density function (b) Distribution function

(c) Probability distribution (d) Continuous distribution

12. A variable that can assume any possible value between two points is called:

(a) Discrete random variable (b) Continuous random variable

(c) Discrete sample space (d) Random variable

13. A formula or equation used to represent the probability distribution of a continuous random variable is called:

(a) Probability distribution (b) Distribution function

(c) Probability density function (d) Mathematical expectation

14. If X is a discrete random variable and f(x) is the probability of X, then the expected value of this random variable is equal to:

(a) $\sum f(x)$ (b) $\sum [x+f(x)]$ (c) $\sum f(x)+x$ (d) $\sum xf(x)$

15. Given E(X) = 6and E(Y) = -4, then E(X - Y) is:

(a) 3 (b) 5 (c) 10 (d) 2

16. If we have $f(x) = 2x, 0 \leq x \leq 1$, then f(x) is a:

(a) Discrete probability distribution (b) Probability density function

(c) Distribution function (d) Discrete random variable

17 The distribution function F(x) is equal to:

(a) $P(X = x)$ (b) $P(X \leq x)$ (c) $P(X \geq x)$ (d) $\sum P(X \geq x)$

18. Given the following probability distribution.

X	1	2	3	4	5	6	7
P(X)	k	2k	2k	3k	K^2	$2k^2$	$7k^2+k$

Determine a) k b) $P(X<3)$ c) $P(X \geq 6)$ d) $P(1<X<4)$

19. A random variable X has the following probability distribution.

X	-1	0	2
P(X)	0.3	0.2	0.5

Determine a) E (x) b) E(3X) c) E(x+5) d) E(2x+8)

20. X has the following probability distribution.

X	1	2	3	4
P(X)	$\frac{2}{12}$	$\frac{2}{12}$	$\frac{5}{12}$	$\frac{3}{12}$

Determine the $P(X \leq 1)$ and $P(1 < X < 4)$.

21. The probability mass function of X is given by $f(x) = \frac{x^2+2}{22}$, $x = 0, 1, 2, 3$.

Prepare the probability distribution in tabular form.

22. Find the distribution functions of X in question numbers 20 and 21.

23. A random variable X has the following probability distribution.

X	0	1	2	3	4	5	6	7
P(X)	a	4a	3a	7a	8a	12a	6a	7a

a) Find the value of a.

b) Find $P(X < 3)$, $P(X \geq 4)$, $P(0 < X < 5)$

24. For the following probability distribution shown below.

Evaluate a) $E(X)$ b) $E(x^2)$ c) $V(X)$ d) $E(X+5)$ e) $V(6X)$ f) $V(2X+7)$

X	8	12	16	20	24
P(X)	$\frac{1}{8}$	$\frac{1}{6}$	$\frac{3}{8}$	$\frac{1}{4}$	$\frac{1}{12}$

25. Find the value of K, if the p.d.f of X is given by

$f(x) = kx^2, 0 < X < 2$

$= 0$, otherwise

26. Examine whether the following is a pdf.

$f(x) = \frac{1}{4(x-1)^3}, 1 < x < 3$

27. Examine whether the following is a pdf.

$f(x) = 3(2 - x)(x - 1)$; $1 < x < 2$

28. Find the cdf if ,the pdf is given by

$$f(x) = \frac{3}{2}x^2, -1<x<1$$

$$= 0 \text{ ,otherwise}$$

29. Calculate mean and variance of a random variable whose pdf is given by

$$f(x) = \frac{1}{2a} \text{ ; } -a < x < a$$

30. Show that $f(x) = 3x^2$ for $0 < x < 1$ represents density function.

31. The probability density of the continuous random variable Y is given by

$$f(y) = \frac{1}{8}(y+1) \quad \text{for} \quad 2 < y < 4$$

$$= 0, \text{ otherwise}$$

Find $P(Y < 3.2)$ and $p(2.9 < y < 3.2)$

32. Find the distribution function of the random variable X whose probability density function is given by

$$f(x) = x \quad \text{for } 0 < x < 1$$

$$= 2 - x \quad \text{for } 1 < x < 2$$

$$= 0 \text{ elsewhere}$$

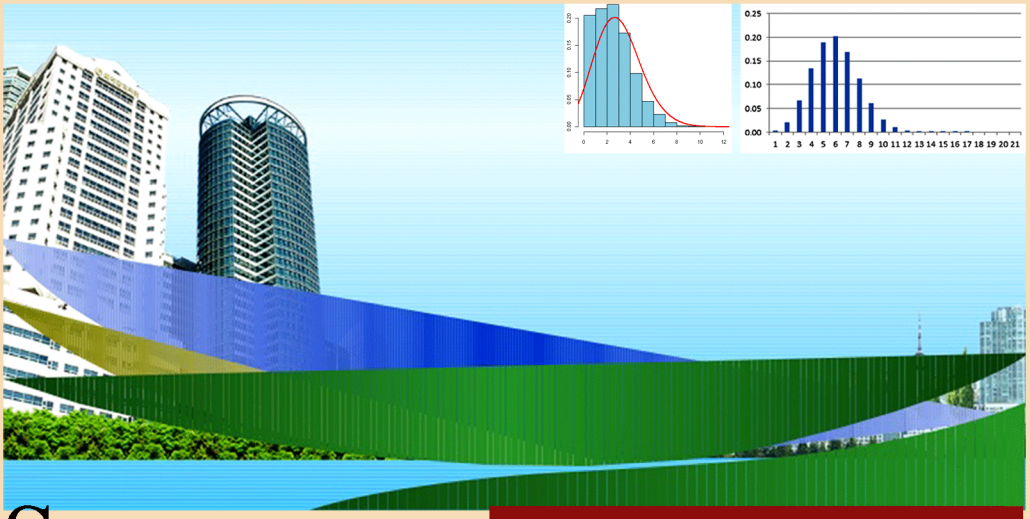
33. The continuous random variable X has cumulative distribution function F(x), where

$$F(x) = \begin{cases} 0, x \le 0 \\ \frac{x^3}{27}, 0 \le x \le 3 \\ 1, x \ge 3 \end{cases}$$

Find the pdf of X.

Chapter 5

Discrete Probability Distributions



So far we have looked at how to calculate and use probability distributions. For simple random experiments, it is not difficult to identify the sample space and to calculate the probability distribution of the random variable under consideration. But, in many practical situations, computation of probabilities after listing the sample points will not be an easy task.

Wouldn't it be nice to have something easier to work with?

Significant Learning Outcomes

After the completion of this chapter, the learner:

- Describes conditions to be satisfied for a Binomial distribution.
- Defines Binomial probability distribution and Poisson probability distribution.
- Evaluates probabilities by applying Binomial distribution.

In this chapter we will see some special probability distributions which follow definite patterns. Once we know these patterns, it will be very easy to calculate probabilities using them. The mean and variance can also be calculated without much time and effort by knowing these definite patterns. Binomial distribution and Poisson distribution are two such popular discrete probability distributions. Let us discuss them in detail and try to compute probabilities using them.

5.1. Binomial probability distribution

Consider the following situations:

- A coin is tossed four times. What is the probability of getting exactly two heads?
- One Yes or No question is asked to ten persons. What is the probability of getting three 'Yes' as response?
- Ten students appear for a test. What is the probability that at least three of them will pass the test?
- A basket contains 80 good oranges and 20 rotten oranges. What is the probability that atleast one good orange is there when three oranges are drawn from the basket?

Binomial distribution was introduced by Jacob Bernoulli (1654-1705) in the book *ArsConjenctendi* published posthumously by his nephew – Nicholas Bernoulli in 1713.

Let us analyse the outcomes of each trial of the above experiments.

- A tossed coin shows **Head** or **Tail**.
- The response to the question might be **Yes** or **No**.
- The student who appeared in a test can either **Pass** or **Fail**.
- The selected orange can be **Good** or **Rotten**.

How many outcomes are there in each case? In each case, there are two outcomes.

Analyse the two outcomes in each case. We can see that, if one outcome is 'success', then the other outcome will be 'failure'. For example, in a coin tossing experiment, if the occurrence of head is considered as a 'success', then occurrence of tail is considered as 'failure'.

If we denote the probability of success by 'p' and probability of failure by 'q', we will have p+q = 1.

In the coin tossing experiment, every time we toss a coin is a trial. When a coin is tossed four times the number of trials is four, each having exactly two outcomes, namely success and failure.

The outcome of a trial is independent of the outcome of any other trial.

In each of the trials, probability of success remains constant.

Such independent trials having only two outcomes, usually referred as success and failure are called Bernoulli trials.

Bernoulli trial

Bernoulli trial is a statistically independent trial having only two possible outcomes. The probability of the outcome at any trial always remains same.

For example, throwing a die 50 times is a case of 50 Bernoulli trials, in which each trial results in success (say, getting 6) or failure (not getting 6) and the probability of success (p) remains same for all 50 throws.

Obviously, the successive throws of the die are independent trials.

If the die is fair with faces numbered 1 to 6, then the probability of success $p = \frac{1}{6}$ and $q = 1 - \frac{1}{6} = \frac{5}{6}$, the probability of failure.

To apply the binomial probability distribution, X must be a dichotomous discrete random variable. In other words, a random variable defined over an experiment results in exactly one of the two possible outcomes.

So, we can apply the binomial distribution to the experiments that satisfy the following four conditions

1. There are n (finite) identical trials. In other words, the given experiment is repeated n times. All these repetitions are performed under identical conditions.
2. Each trial has two and only two outcomes. These outcomes are usually called ***success*** and ***failure***.
3. If the probability of success is denoted by p and probability of failure is denoted by q, then p+q=1. The probabilities p and q remains constant for each trial.

4. The trials are independent. In other words, the outcome of one trial does not affect the outcome of another trial.

Conditions for Binomial Experiments

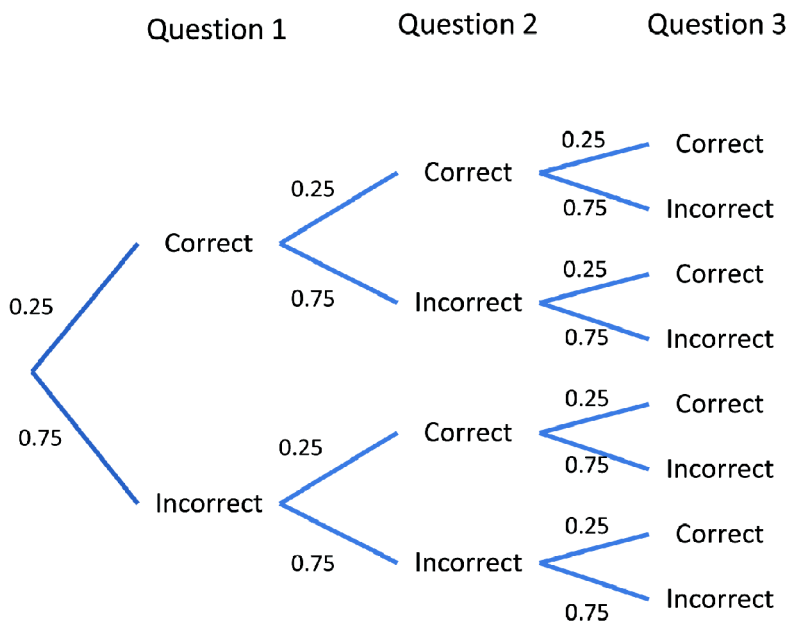
1. There are n (finite) identical trials.
2. Each trial has two possible outcomes.
3. The probabilities of two outcomes remain constant.
4. The trials are independent.

The Binomial formula

We are familiar with multiple choice questions having four choices, where only one choice is the right answer. If the answer is not known, we may answer it randomly. Then the probability of the answer being correct is 0.25 and the probability of incorrect answer is 0.75.

Let us consider a situation where three such questions are to be answered. We can find the probability distribution for the number of questions that have correct answer. That would give direction to us whether to answer at random or not.

Here is the probability tree for the situation:



Let us use X to represent the number of questions we get correct out of three. The probability distribution can be summarised as follows.

X	P(X=x)	Power of 0.25	Power of 0.75
0	$(0.75)^3 = 0.422$	0	3
1	$3x(0.25)^1 (0.75)^2 = 0.422$	1	2
2	$3x(0.25)^2 (0.75)^1 = 0.411$	2	1
3	$(0.25)^3 = 0.016$	3	0

There are three different ways, we can get one question right. All of them have a probability of $(0.25)^1 (0.75)^2$. Similarly, there are three different ways of getting two questions right with a probability $(0.25)^2 (0.75)^1$ of each.

Let us take the experiment made up of three Bernoulli trials with probability of success p and probability of failure q in each trial. Then the probability distribution of X is:

X	0	1	2	3
P(X = x)	q^3	$3\ p^1\ q^2$	$3\ p^2q^1$	p^3

Hence the probability distribution is the binomial expansion of $(q + p)^3$.

We know that $q + p = 1$.

Hence the sum of the probabilities, $q^3 + 3\ p^1q^2 + 3\ p^2q^1 + p^3 = 1$.

Thus we may conclude that in an experiment of n Bernoulli trials, the probabilities of 0,1,2,3.....n successes can be obtained as first, second, third.... $(n+1)^{th}$ terms in the binomial expansion of $(q+p)^n$. Hence the distribution of the number of successes X can be written as:

X	0	1	2	.....	x	.....	n
P(X = x)	${}^nC_0q^n$	${}^nC_1p^1q^{n-1}$	${}^nC_2p^2q^{n-2}$		${}^nC_xp^xq^{n-x}$		${}^nC_np^x$

The above probability distribution is known as binomial distribution with parameters n and p, because for the given values of n and p we can find the complete probability distribution.

Thus, the probability of x successes P(X=x) is given by:

$$P(X = x) = {}^nC_x p^x q^{n-x} \quad ; x = 0,1,2\ldots n;\ q = 1\text{-}p.$$

$$\Sigma {}^nC_x p^x q^{n-x} = (q + p)^n = 1$$

Hence, $P(X = x) = {}^{n}C_{x}\, p^{x} q^{n-x}$; x = 0,1,2…n; q = 1-p. is the probability mass function of binomial distribution.

A random variable X is said to follow binomial distribution, if its probability mass function is given by,

$$P(X = x) = {}^{n}C_{x}\, p^{x} q^{n-x} \text{ ; } x = 0,1,2\ldots n;\ q = 1-p$$
$$= 0, \text{ otherwise.}$$

Binomial distribution with parameters n and p is usually denoted by B(n,p).

Mean and Variance of Binomial distribution

For binomial distribution with parameters n and p,

Mean, E(X) = np

Variance, V(X) = npq

Standard deviation, $\sigma = \sqrt{npq}$

If X ~ B(n,p),

Mean, E(X) = np, Variance, V(X) = npq, Standard deviation, $\sigma = \sqrt{npq}$

Remark:

For binomial distribution, Variance < Mean

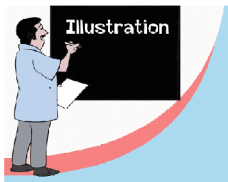


Illustration 5.1

There are five multiple choice questions in a question paper with each question having four choices with only one right answer. If one selects the answers randomly,

a. What is the probability of getting exactly two questions right?

b. What is the probability of getting three questions right?

c. What is the probability of getting two or three questions right?

d. What is the probability of getting no questions right?

e. What is the mean and standard deviation?

Solution

The answers are selected at random, therefore,

The probability of success, $p = \frac{1}{4} = 0.25$

The probability of failure, $q = 1 - p = 1 - \frac{1}{4} = 0.75$

Let X be the number of questions answered correctly,

then X ~ B(n,p); $P(X = x) = {}^{n}C_{x}\, p^{x} q^{n-x}$; x = 0,1,2…n; q = 1-p

a. Probability of getting exactly two questions right = P(X = 2)

$= {}^{5}C_{2}\,(0.25)^{2}(0.75)^{5-2}$

= 10 x 0.0625 x 0.421875

= 0.264

b. Probability of getting three questions right =P(X = 3)

$= {}^{5}C_{3}\,(0.25)^{3}(0.75)^{5-3}$

= 10 x 0.015625 x 0.5625

= 0.0879

c. Probability of getting two or three questions right =P(X = 2 or X = 3)

= P(X = 2) +P (X = 3)

= 0.264 + 0.0879

= 0.3519

d. Probability of getting no questions right = P(X = 0)

$= (0.75)^{5}$

= 0.237

e. Mean, E(X) = np = 5 x 0.25 =1.25

Variance, V(X) = npq = 5 x 0.25 x 0.75 = 0.9375

Standard deviation $\sigma = \sqrt{npq} = \sqrt{0.9375} = 0.9682$

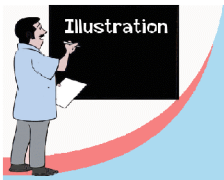


Illustration 5.2

A fair coin is tossed 5 times. What is the probability of getting exactly 2 heads?

Solution:

This is a case of binomial distribution.

Let X be the number of heads,

then X ~ B(n,p)

$$P(X = x) = {}^{n}C_{x}\, p^{x} q^{n-x} \quad ; x = 0,1,2\ldots n;\ q = 1-p$$

$$p = \frac{1}{2},\ q = 1 - p = 1 - \frac{1}{2} = \frac{1}{2}$$

n = 5, x = 2

Probability of getting exactly two heads =P(X = 2)

$$= {}^{5}C_{2} \left(\frac{1}{2}\right)^{2} \left(\frac{1}{2}\right)^{5-2}$$

$$= \frac{5 \times 4}{1 \times 2} \times \left(\frac{1}{2}\right)^{5}$$

$$= 0.3125$$

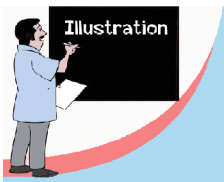


Illustration 5.3

Five cards are drawn from a pack of 52 cards with replacement. Find the probability of getting exactly 3 heart cards.

Solution

Let X represent the number of Hearts

Probability of selecting a Heart card from a pack of 52 cards, $p = \frac{13}{52} = \frac{1}{4}$

$$q = 1 - p = 1 - \frac{1}{4} = \frac{3}{4}, \qquad n = 5$$

Probability of getting exactly 3 Hearts, P(X = 3)

$$= {}^5C_3 \left(\frac{1}{4}\right)^3 \left(\frac{3}{4}\right)^{5-3}$$

$$= \frac{45}{512}$$

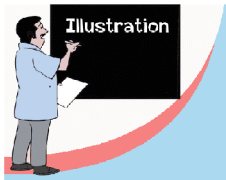


Illustration 5.4

The mean of a binomial distribution is 6 and variance 5.

a. Write the probability mass function.

b. Evaluate P(X= 1).

Solution

Mean, np = 6, Variance = 5, i.e. $npq = 5$

$$\frac{npq}{np} = \frac{5}{6}$$

$$q = \frac{5}{6}$$

$$p = 1 - \frac{5}{6} = \frac{1}{6}$$

$$np = 6$$

$$\therefore n \times \frac{1}{6} = 6 \Rightarrow n = 6 \times 6 = 36$$

a. The probability mass function is given by,

$$f(x) = {}^{36}C_x \left(\frac{1}{6}\right)^x \left(\frac{5}{6}\right)^{36-x} ; x = 0,1,2\ldots36;$$

$$= 0, \text{ otherwise}$$

b. $P(X=1) = {}^{36}C_1 \left(\frac{1}{6}\right)^1 \left(\frac{5}{6}\right)^{36-1} = 6\left(\frac{5}{6}\right)^{35}$

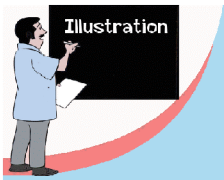


Illustration 5.5

A fair coin is tossed 16 times. Find the mean, variance and standard deviation of the number of heads obtained?

Solution

Let X represents the number of heads obtained.

$$n = 16,\ p = \frac{1}{2}$$

Mean, $np = 16 \times \frac{1}{2} = 8$

Variance, $npq = 8 \times \frac{1}{2} = 4$

Standard deviation, $\sqrt{npq} = \sqrt{4} = 2$

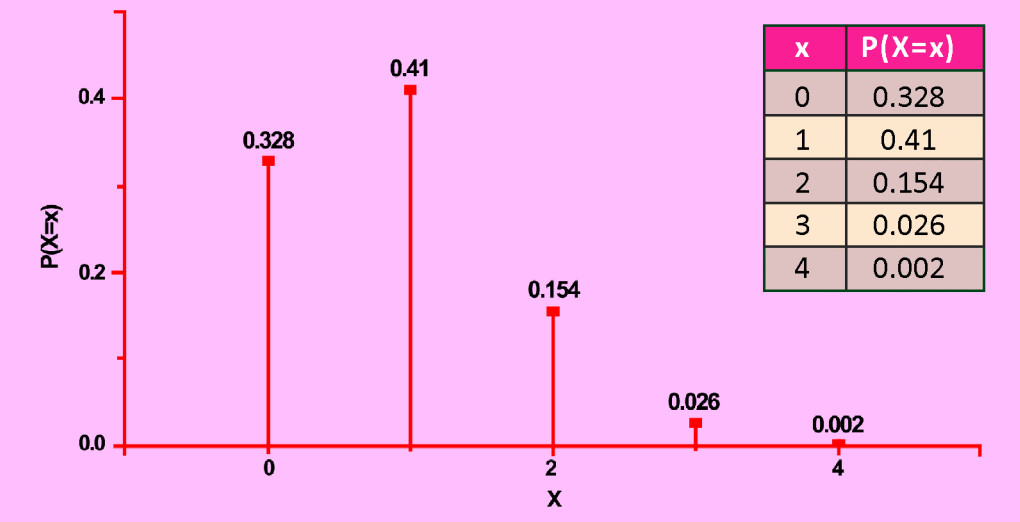
Know your progress

1. Ten percent items produced by a machine are likely to be defective. Five items are selected at random. Find the probability that not more than two items are defective.
2. The normal rate of infection of certain disease in animals is known to be 25%. Six animals were selected at random and inspected. What is the probability that none of the animals are infected?
3. A die is rolled 240 times. Find the mean, variance and standard deviation for the number of sixes that will be rolled?
4. The mean of a binomial distribution is 40 and standard deviation is 2. Obtain the values of n, p and q? Write the probability mass function?

Probability of success and shape of binomial distribution

For any number of trials n,

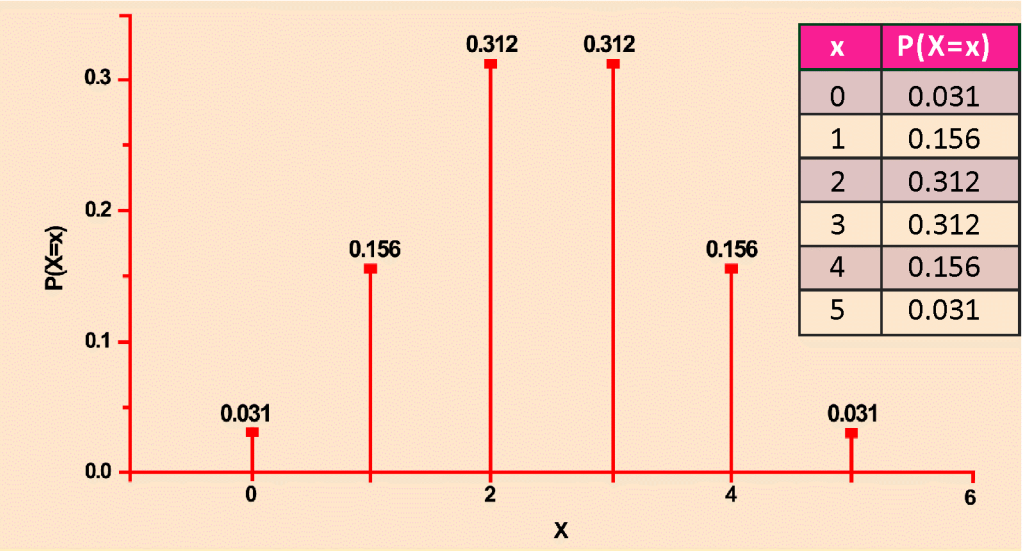
- The binomial probability distribution is symmetric if $p = 0.5$.
- The binomial distribution is positively skewed (skewed to right) if $p < 0.5$.
- The binomial distribution is negatively skewed (skewed to left) if $p > 0.5$.



x	P(X=x)
0	0.328
1	0.41
2	0.154
3	0.026
4	0.002

Figure 5-1. Graph of binomial distribution having n=5 and p=0.3

Here the distribution is positively skewed.



x	P(X=x)
0	0.031
1	0.156
2	0.312
3	0.312
4	0.156
5	0.031

Figure 5-2.Graph of binomial distribution having n=5 and p=0.5

Here the distribution is symmetric.

x	P(X=x)
0	0.002
1	0.028
2	0.132
3	0.309
4	0.36
5	0.168

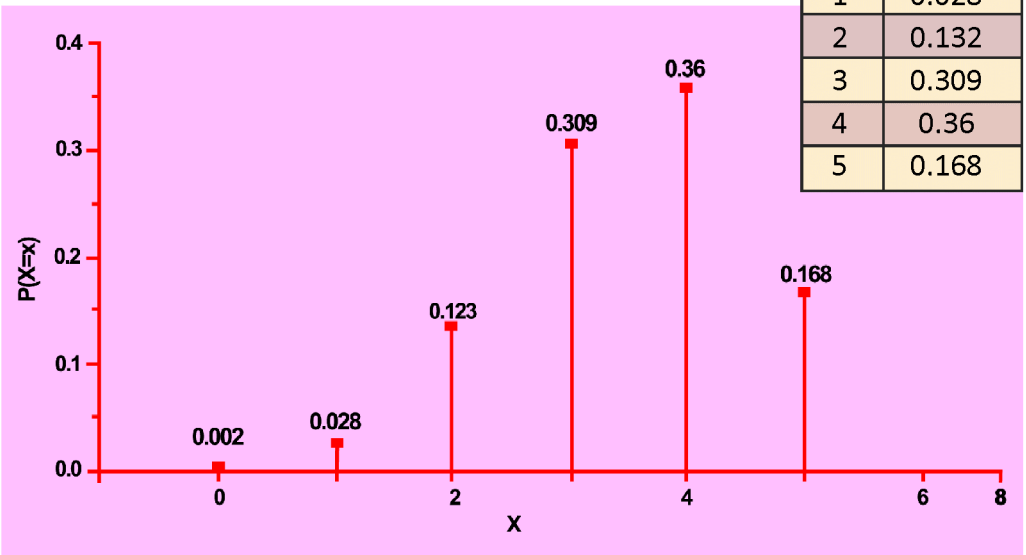


Figure 5-3.Graph of binomial distribution having n=5 and p=0.7

Here the distribution is negatively skewed.

Importance of binomial distribution

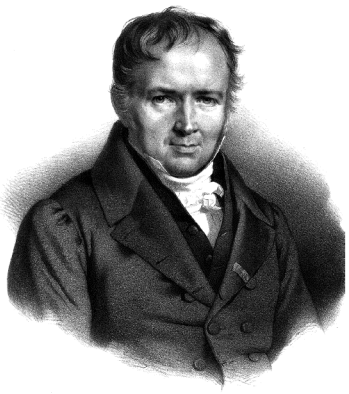
- The binomial distribution is found very useful in decision making situations in business.
- In statistical quality control it has a wider area of application. Binomial distribution describes variety of real life events.
- Binomial distribution is normal (approximate) when n is sufficiently large and p and q do not differ very much.

Additive property of binomial variables.

Let X be a binomial variable with parameters n_1 and p, Y be another binomial variable with parameters n_2 and p then X + Y will again be a binomial variable with parameters $(n_1 + n_2)$, and p.

i.e., If $X \sim B(n_1, p)$ and $Y \sim B(n_2, p)$, then $(X + Y) \sim B(n_1 + n_2, p)$

5.2. Poisson Probability Distribution



The Poisson probability distribution, named after the French mathematician Simon Denis Poisson (1781-1840) is another discrete probability distribution, having a large number of applications. Suppose a pump set of a lift irrigation project breaks down at an average of three times a month. We may want to find the probability of exactly two break downs in the next month. This is an example of Poisson probability problem. Each break down is called an 'occurrence' in Poisson distribution terminology. The Poisson process measures the number of occurrences of a particular outcome of a discrete random variable in a pre-determined time, space or volume interval for which an average number of occurrences of outcomes is known or can be determined.

Examples:

- ISD calls received on a telephone switch board per hour.
- Patients arriving at every hour in a clinic.
- Road accidents occurred in a week in a city.
- Organisms per unit volume in a liquid.
- Cars waiting for service in a service station.
- Printing mistakes found in a page of a good book.

The occurrences are random in the sense that they do not follow any pattern and hence they are unpredictable.

The occurrences are always considered with respect to an interval. The interval may be a time interval, volume interval or space interval.

In the example of the pump set, the interval is one month, a time interval.

In the case of organisms in the liquid, it is a volume interval.

In the case of mistakes per page we have a space interval.

The occurrence within an interval is random and independent.

If the average number of occurrences in a given interval is known, we can compute the probability of certain number of occurrences (x) in that interval.

Conditions to apply the Poisson probability distribution

The following are three conditions to be satisfied to apply Poisson distribution.

1. X is a discrete random variable.
2. The occurrences are random.
3. The occurances are independent.

The following are some examples of discrete random variables, for which Poisson distribution can be applied to compute probabilities.

1. Consider the number of salesmen visiting a household during a particular day. The visit of a salesman is called an occurrence. Here the interval is a day (an interval of time). The occurrences are random. The total number of salesmen who visits on that day can be 0,1,2..... Also the visits are independent.
2. Consider the number of defective items in a lot of 100 items produced in a plant. The occurrences (defective items) are random and independent. They may be 0,1,2... . Here we have a volume (of 100 items) interval.
3. Consider the number of scratches on a six feet PVC pipe. The interval in this example is a space interval. The scratches are random and independent.

We can cite a number of such examples where Poisson distribution has its application.

Look at some of the examples.

1. The number of accidents that occurs on a road during a week.
2. The number of customers entering a shop during an hour.
3. The number of washing machines sold at a Home appliances shop during a week.

In contrast, consider the arrival of patients at a specialist doctor's clinic where appointment is compulsory. These arrivals are not random because the patients have to make appointments to see the doctor. The arrivals of passenger trains in a station or commercial planes in an airport are not random because they have scheduled timings. The Poisson probability distribution cannot be applied in such cases.

The Poisson Probability Formula

In the Poisson probability terminology, the average number of occurrences in an interval is denoted by λ (lambda). The actual number of occurrence in that interval is denoted by x. Then using the Poisson probability distribution, we find the probability of x

occurrences during an interval, provided the mean occurrences during that interval λ is known.

The probability of x occurrences in an interval is given by

$$P(x) = \frac{e^{-\lambda}\lambda^{x}}{x!}, \quad x = 0,1,2,\ldots$$

$$= 0, \text{ otherwise.}$$

Where λ is the mean number of occurrences in that interval and e = 2.71828 (the exponential constant).

A discrete random variable X is said to follow Poisson distribution if its p.m.f is given by $P(X = x) = \frac{e^{-\lambda}\lambda^{x}}{x!}$, $x = 0,1,2,\ldots$

$= 0$, otherwise.

The average number of occurrences, λ is the parameter of Poisson distribution. Poisson distribution is denoted by $P(\lambda)$.

Poisson probability distribution as an approximation to Binomial distribution

The Poisson probability distribution provides a simple, easy to calculate and accurate approximation to binomial distribution when the probability of success (p) is very small and n is large; so that mean, E(X) = np is small.

Poisson distribution may be obtained as a limiting case of binomial distribution under the following conditions.

I. The number of trials is very large. $n \to \infty$.

II. The probability of success is very small. $p \to 0$.

III. np is finite, say λ.

Poisson distribution – The law of improbable (rare) events

Suppose an event is rare. Then it may be possible to count the number of occurrences of it, but may be impossible to count the non-occurrences. This is because the number of trials of the experiment is not known exactly. Poisson distribution describes the behaviour of such rare occurrences. Poisson distribution may be expected in cases

where the chance of occurrence of the event is small. So the distribution is known as law of improbable events.

Mean and Variance of Poisson distribution

For Poisson distribution with parameter λ,

Mean, E(X) = λ

Variance, V(X) = λ

Standard Deviation, SD(X) = $\sqrt{\lambda}$

When λ is not known it is estimated by np, provided n is large and p is very small.

For Poisson distribution with parameter λ,

Mean, E(X) = λ, Variance, V(X) = λ, Standard Deviation, SD(X) = $\sqrt{\lambda}$

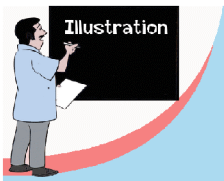


Illustration 5.6

The pump set of a lift irrigation project had an average of 3.4 breakdowns per month during the last year.

a. Find the probability that there will be no breakdowns in the next month.

b. Find the probability of three breakdowns in the next month.

c. Find the mean and variance of the breakdowns.

Solution

Let X be the number of breakdowns in the next month. Then X ~ P(3.4)

$$P(X = x) = \frac{e^{-\lambda}\lambda^{x}}{x\,!} \quad , x = 0,1,2,\ldots.$$

a. Probability that there will be no breakdowns in the next month= P(X=0)

$$= \frac{e^{-3.4}(3.4)^{0}}{0!}$$

$$= e^{-3.4}$$

$$= 0.033$$

b. Probability that there will be three breakdowns in the next month= P(X=3)

$$= \frac{e^{-3.4}(3.4)^3}{3!}$$

$$= \frac{0.033 \times 39.304}{6}$$

$$= 0.216$$

c. Mean, E(X) = λ = 3.4

Variance, V(X) = λ =3.4

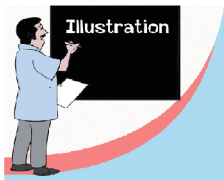


Illustration 5.7

If there are 200 typing errors randomly distributed in a 500 page manuscript, find the probability that a given page will contain exactly three errors.

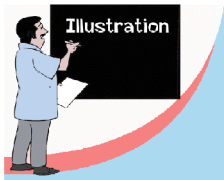
Solution:

The mean number of errors, $\lambda = \frac{200}{500} = 0.4$

Probability that one page will contain exactly 3 errors, P(X=3)

$$= \frac{e^{-0.4}(0.4)^3}{3!}$$

$$= 0.0072$$



Illusration 5.8

In a certain factory, out of 500 items selected for inspection, 0.2% are found to be defective. Find the probability of one defective item per production lot. In 1000 lots produced, how many lots are likely to contain exactly one defective?

Solution

n = 500, p= 0.002

Therefore, λ = np = 500 x 0.002 = 1

Let X be the number of defective items. Then X ~ P(1)

$$P(X=x) = \frac{e^{-\lambda}\lambda^{x}}{x\,!} \quad , x = 0,1,2,\ldots.$$

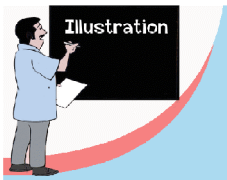
Probabilityof one defective items = P(X = 1)

$$= \frac{e^{-1}1^{1}}{1!}$$

$$= 0.03679$$

Number of production lots having exactly one defective items = 1000 x 0.3679

= 367.9 $\approx$ 368 lots



Illusration 5.9

If a random variable X follows a Poisson distribution such that P(X=1) = P(X=2). Find P(X=0).

Solution

$$P(X=x) = \frac{e^{-\lambda}\lambda^{x}}{x\,!} \quad , x = 0,1,2,\ldots.$$

$$P(X=1) = \frac{e^{-\lambda}\lambda^{1}}{1!} \text{ and } P(X=2) = \frac{e^{-\lambda}\lambda^{2}}{2!}$$

$$\frac{e^{-\lambda}\lambda^{1}}{1!} = \frac{e^{-\lambda}\lambda^{2}}{2!}$$

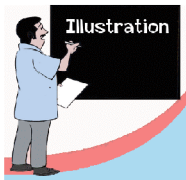
Therefore, $\lambda = \frac{\lambda^{2}}{2}$

Hence, $\lambda = 2$

$$P(X=0) = \frac{e^{-2}2^{0}}{0!}$$

$$= e^{-2}$$

$$= 0.1353$$



Illusration 5.10

If approximately 2% in a group of 200 people are left handed, find the probability that exactly five people are left handed.

Solution

n = 200, p= 0.02

Therefore, λ = np = 200 x 0.02 = 4

Let X is the number left handed. Then X ~ P(4)

$$P(X=x) = \frac{e^{-\lambda}\lambda^{x}}{x\,!} \quad , x = 0,1,2,\ldots.$$

Probability of exactly five are left handed = P(X=5)

$$= \frac{e^{-4}4^{5}}{5!}$$

$$= \frac{0.0183 \times 1024}{120}$$

$$= 0.1563.$$

Know your progress

1. A new automated production process has had an average 1.5 breakdowns per day. Assume that the breakdowns occur randomly. What is the probability of less than three breakdowns per day?
2. A life insurance company insures 5000 persons aged 40. If studies show the probability that any 40 year old person will die in the given year to be 0.001. Find the probability that the insurance company will have to pay atleast two claims during a given year?
3. A factory produces blades in packets of 10. The probability that a blade to be defective is 0.002. Find the number of packets likely to have two defective blades in a consignment of 10,000 packets?

Importance of Poisson distribution

Poisson distribution describes the behaviour of discrete random variables where the probability of occurrence of the event is very small and the total number of possible cases is sufficiently large. As such, Poisson distribution has wide area of applications in queuing theory (waiting line problems) and in the fields of industry, insurance, medicine, economics, physics, biology, etc.

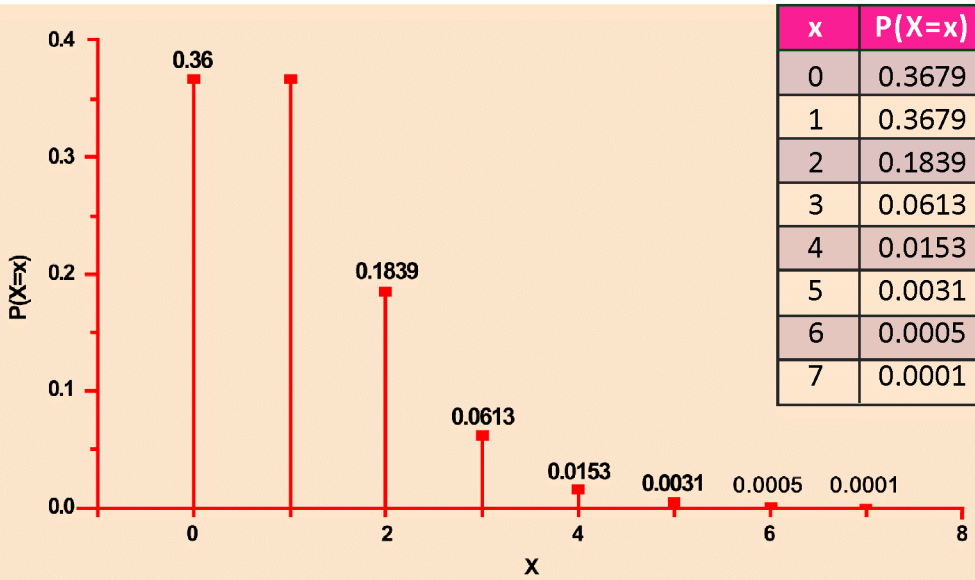
Additive property of Poisson variables

Let X be a Poisson variable with parameter λ_1 and Y be another Poisson variable with parameter λ_2, then X+Y will follow Poisson distribution with parameter $\lambda_1 + \lambda_2$.

If $X \sim P(\lambda_1)$ and $Y \sim P(\lambda_2)$, then $X + Y \sim P(\lambda_1 + \lambda_2)$

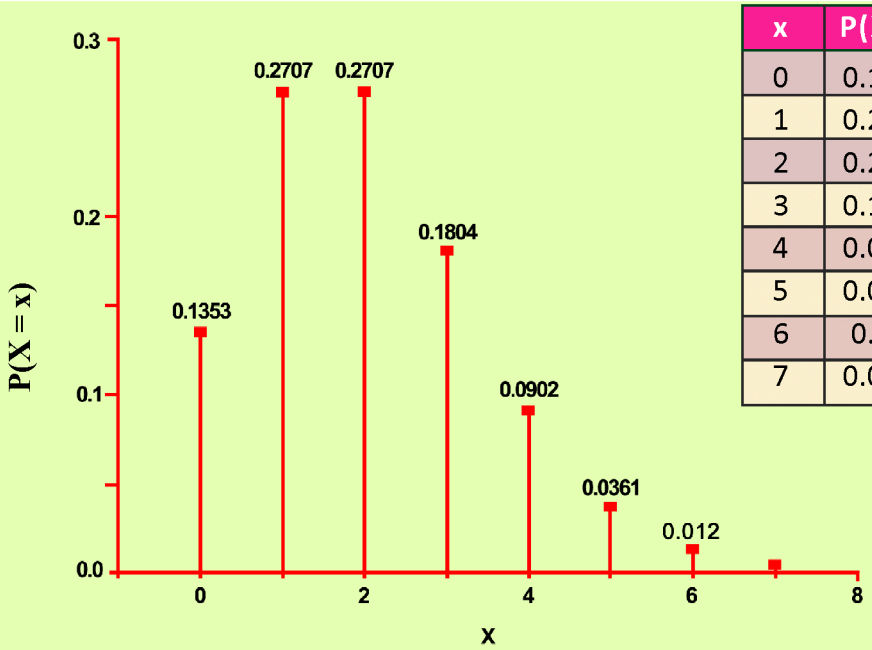
Shape of Poisson distribution

We have seen that binomial distribution is positively skewed (skewed to right) for $p < 0.5$. For Poisson distribution probability of success is very small. Therefore Poisson distribution will always be positively skewed.



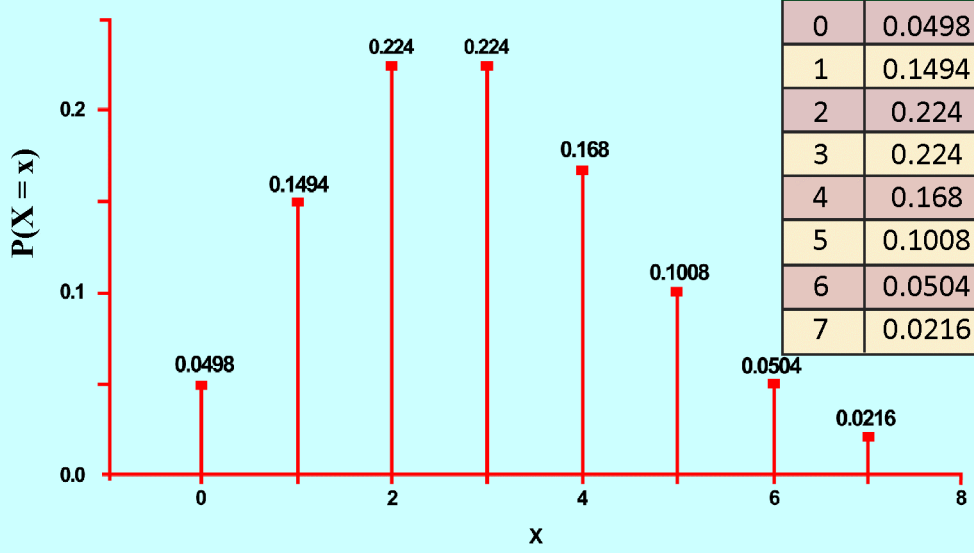
x	P(X=x)
0	0.3679
1	0.3679
2	0.1839
3	0.0613
4	0.0153
5	0.0031
6	0.0005
7	0.0001

Figure 5.4 - Graph of Poisson Distribution with $\lambda = 1$



x	P(X=x)
0	0.1353
1	0.2707
2	0.2707
3	0.1804
4	0.0902
5	0.0361
6	0.012
7	0.0034

Figure 5.5 - Graph of Poisson Distribution with λ = 2



x	P(X=x)
0	0.0498
1	0.1494
2	0.224
3	0.224
4	0.168
5	0.1008
6	0.0504
7	0.0216

Figure 5.6 - Graph of Poisson Distribution with λ = 3

See, the shape of the Poisson distribution. It depends on λ. If λ is small, then the distribution will be skewed to right. It approaches symmetry as λ gets larger.

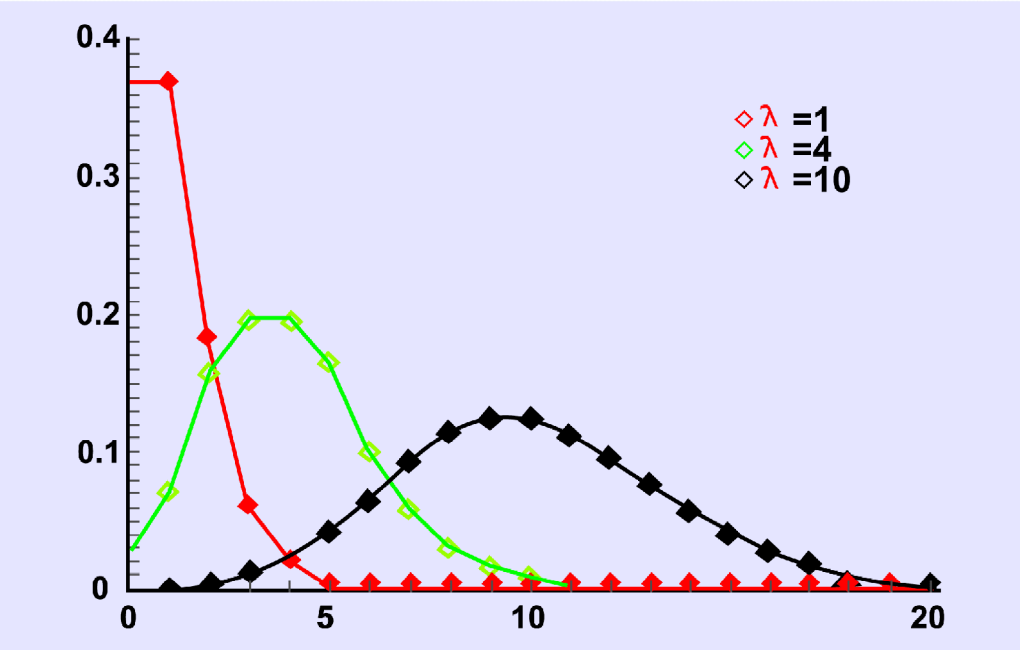


Figure 5.7 -

Let us conclude

In this chapter, Bernoulli process, the Binomial distribution and Poisson distribution are discussed. A process in which each trial has only two possible outcomes, the probability of the outcome at any trial remains fixed over time and the trials are independent is termed as Bernoulli process. The Binomial distribution is used to evaluate the probability of a particular number of successes in the case of finite independent trials, there can be success and failure for each trial and the probability of success is same for each trial.

If X ~ B (n, p), $P(X = x) = {}^{n}C_{x} p^{x} q^{n-x}$; x = 0,1,2…n; q = 1-p. For binomial distribution with parameters n and p,Mean, E(X) = np, Variance, V(X) = npq and Standard deviation, $\sigma = \sqrt{npq}$. For any number of trials n,the binomial probability distribution is symmetric if p = 0.5, positively skewed (skewed to right) if p < 0.5 and negatively skewed (skewed to left) if p > 0.5.

The Poisson distribution is used when individual events occur randomly and independently in a given interval. The mean number of occurrences in the interval, λ is finite. To apply Poisson distribution λ is to be known. When n is large and p is very small, binomial distribution is approximated using poisson distribution by taking λ = np.

If X~ P(λ) , P(X = x) = $\frac{e^{-\lambda}\lambda^{x}}{x\,!}$, x = 0, 1, 2, ... For Poisson distribution with parameter λ , Mean, E(X) = Variance, V(X) = λ . Poisson distribution is always positively skewed.

Let us assess

For Questions 1-15, choose the correct answer from the given choices.

1) If X follows binomial distribution with parameters *n* and *p*, then the values taken by X are.....

   a) infinite b) finite integers c) integers from 0 to n d) infinite integers

2) For binomial distribution mean is.....

   a) less than variance b) greater than variance

   c) equal to variance d) none of these

3) Mean of binomial distribution with parameters n and p is.....

   a) npq b) np c) n/p d) n+p

4) The variance of binomial distribution with n = 6 and p = 0.4 is.....

   a) 0.24 b) 1.44 c) 1.2 d) 1.24

5) If the mean of a binomial distribution is 4 and variance is 3, then the probability of success is.....

   a) 0.25 b) 0.75 c) 0.5 d) 0.35

6) Poisson distribution is also known as.....

   a) law of averages b) law of rare events

   c) law of large numbers d) law of distribution

7. Poisson distribution is a limiting case of ..... distribution.

   a) Binomial b) Normal c) Geometric d) Pearson

8. Parameter of Poisson distribution is also its .....

   a) mean b) standard deviation c) median d) size

9. If X and Y are independent Poisson variables, then X+Y follows.....distribution.

   a) Binomial b) Continuous c) Normal d) Poisson

10. Poisson distribution is .....

    a) positively skewed b) negatively skewed.

    c) symmetric d) both negatively and positively skewed

11. Binomial distribution is symmetric when.....

    a) p =0.5 b) p= 0.1 c) p=0 d) p=0.9

12. The average number of occurrences in 1000 trials is 5. Then the probability of success is.....

    a) 0.05 b) 0.95 c) 0.005 d) 0.995

13. You have to find the probability of exactly 3 male births out of 10 deliveries in a hospital. Which probability distribution you would suggest?

    a) Poisson b) Binomial c) Normal d) Polynomial

14. For a Poisson distribution with $\lambda = 1$, if $P(X = 0) = 0.3679$, $P(X=1)$ is.....

    a) 0.6321 b) 0.3679 c) 0.2231 d) 0.2386

15. Identify the binomial experiments or experiments that can be reduced to binomial experiments from the following:

    a) Surveying 100 people to determine whether they like a particular brand soap.

    b) Tossing a coin 100 times to see how many heads occur.

    c) Drawing a card from a deck and getting a heart.

    d) Asking 1000 people which brand of coffee they use.

    e) Testing different brands of aspirin to see which brands are effective.

f) Asking 100 people whether they smoke?

g) Checking 1000 applicants to see whether they have got admission.

h) Asking 300 participants their age.

i) Surveying 1000 students whether they have a driving licence.

16. Compute the probability using binomial formula.

a) $n = 2, p = 0.30, x = 1$

b) $n = 4, p = 0.60, x = 3$

c) $n = 5, p = 0.10, x = 0$

17. Find the mean, variance and standard deviation for each of the values of n and p when the conditions of binomial distribution are met.

a) $n = 100, p = 0.75$.

b) $n = 300, p = 0.3$.

c) $n = 20, p = 0.5$.

d) $n = 10, p = 0.8$.

e) $n = 1000, p = 0.1$.

f) $n = 36, p = \frac{1}{6}$.

18. If 40% of the workers of a factory uses two wheelers, find the probability that out of 8 workers selected 5 uses two wheelers?

19. If 20% women of Kerala are employed outside the home district, find the probability that in a sample of 9 women,

a) Exactly 3 are employed

b) At most three are employed

20. The probability of passing a driving test in the first chance is 0.8. Find the mean, variance and standard deviation of the number of applicants passed in the first chance from a batch of 300.

21. There are 75 students in a class. If the rate of absence is 12%, find the mean, variance and standard deviation of the number of students absent from each class?

22. Assuming Poisson distribution, compute the probabilities for given values of λ and x.

    a) $\lambda = 0.1$ and $x = 2$.

    b) $\lambda = 4$ and $x = 2$.

    c) $\lambda = 2$ and $x = 3$.

    d) $\lambda = 3$ and $x = 5$.

23. In a town the average number of accidents took place per day in the last year is 0.1. Assuming that number of accidents follows Poisson distribution, find the probability that:

    a) there will be no accidents on a day.

    b) there will be less than 3 accidents per day.

24. It is known that bacteria of a certain kind occur at the rate of two bacteria per cubic centimetre of water. What is the probability that a sample of one cubic centimetre water will contain:

    a) exactly one bacterium?

    b) at least one bacterium?

25. If 3% of electric bulbs produced by a company are defective, find the probability that in a sample of 100 bulbs exactly 5 bulbs are defective?

26. If 2% of all cars fail the pollution control inspection, find the probability that in a sample of 200 cars 5 will fail.

27. A video tape has an average of one defect per every 1000 feet. Find the probability of at least one defect in 3000 feet.

28. If a random variable X has a Poisson distribution such that $P(X=0) = P(X=1)$, find $P(X=4)$.

Chapter 6

Normal Distribution

The concept of random variable is very much familiar to you now. The details about discrete and continuous random variables were discussed in detail in Chapter 4. As a continuation of this, here we discuss distributions of continuous random variables.

The normal distribution is one of the many probability distributions that a continuous random variable can possess. The normal distribution is an important continuous distribution

Significant Learning Outcomes

After the completion of this chapter, the learner:

- Explains the concept of normal distribution.
- Illustrates the characteristics of normal distribution.
- Identifies the importance of normal distribution.
- Uses normal distribution in various situations.

because a good number of random variables occurring in practice can be approximated to it. That is, a large number of phenomena in the real world are normally distributed either exactly or approximately. The continuous random variable representing heights and weights of people, score in an examination, blood sugar levels, behavioural (intelligence, achievements, etc.) sciences, length of pins made by an automatic machine, volume of oil in a particular brand of canned oil, time taken to complete a job are some good examples of normally distributed random variables. All of these are affected by several independent causes and the effect of each cause is small. If a random variable is affected by many independent causes, and the effect of each cause is not overwhelmingly large compared to other effect, then the random variable will closely follow a normal distribution.

Normal distribution is of tremendous importance in the analysis and evaluation of every aspect of experimental data in science and medicine. In fact, the majority of the basic statistical methods that we study in the forthcoming chapters are based on the normal distribution. In this chapter we will discuss the properties of normal distribution and its applications.

6.1 Normal Distribution- Concept

Suppose you have collected the heights of 100 adult people in your city and constructed a histogram of collected data. You will get a graph similar to one as shown below:

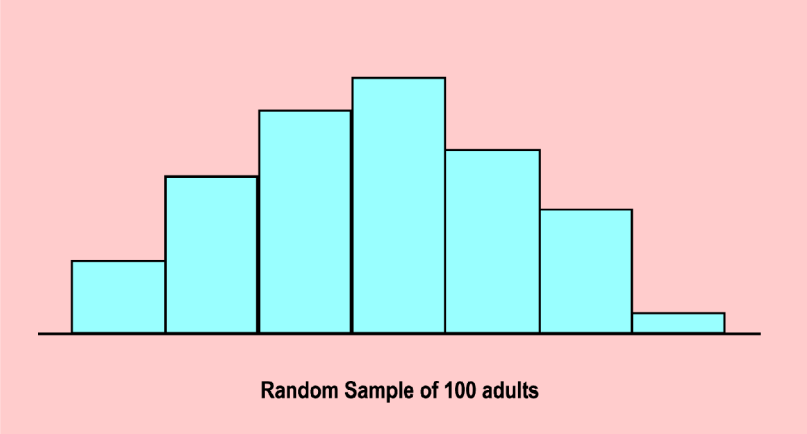


fig.6.1

Now you increase the sample size and decrease the width of the classes, the histogram looks like the picture below.

fig.6.2

Further if you increase sample size and decrease width, the histogram looks like the picture given below:

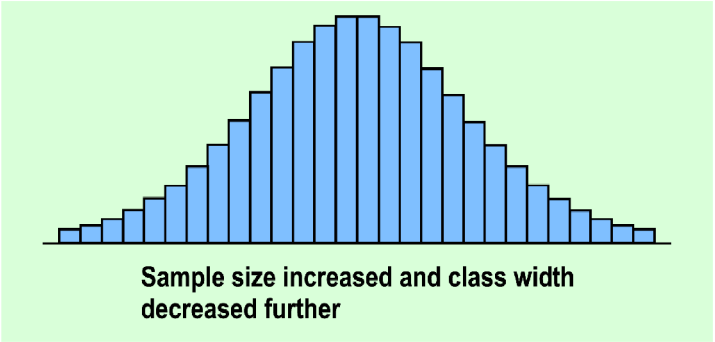


fig.6.3

Finally, if it were possible to measure exactly the heights of all persons of age above 18 in our country and construct a histogram, it would approach what is called normal distribution.

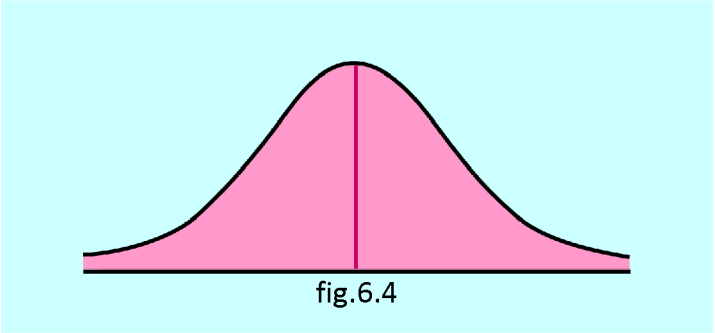


fig.6.4

The Histogram of the normal distribution produces the familiar symmetric curve known as normal curve or bell shaped curve.

Normal distribution can be used to describe many variables such as heights, weights, etc., because the deviations from a normal distribution are very small.

Activity

1) Collect heights of all HSS students in your school.
   (i) Prepare a frequency distribution of the data.
   (ii) Draw a frequency curve for the data.
   Observe the frequency curve and draw your conclusions.

(2) Collect the weights of all HSS students in your school.
   (i) Prepare a frequency distribution of the data.
   (ii) Draw a frequency curve for the data.
   Observe the frequency curve and draw your conclusions.

(3) The scores of 50 students out of 100 in Mathematics for Class XI in a HSS is given below

67	84	80	77	97	59	62	37	33	42
36	54	18	12	19	33	49	24	25	22
24	29	9	21	21	24	31	17	15	21
13	19	19	22	22	30	41	22	18	20
26	33	14	14	16	22	26	10	16	24

a. Construct a frequency distribution for the data.
b. Construct a histogram for the data.
c. Describe the shape of the histogram.
d. Do you feel that the distribution is approximately normal?

6.2 Normal Probability Density Function

A continuous random variable X is said to follow Normal distribution if it has the probability density function

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

μ and σ^2 are known as the parameters of the distribution (where $e \simeq 2.718$, $\pi = 3.14$)

A continuous random variable follows normal distribution with parameters μ and σ can be denoted as $X \sim N(\mu, \sigma^2)$.

A continuous random variable that has a normal distribution is called normal random variable.

Mean and Variance

If $X \sim N(\mu, \sigma^2)$ then Mean $= E(X) = \mu$

Variance $= V(X) = \sigma^2$

Standard deviation $= \sqrt{V(X)} = \sigma$

Normal Curve

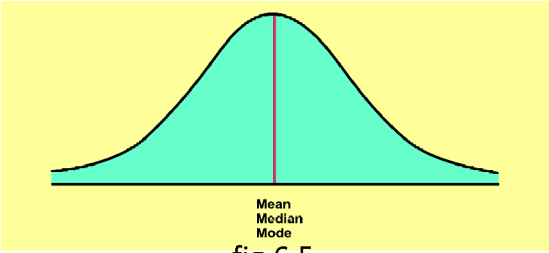
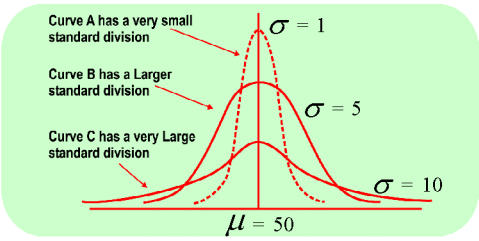
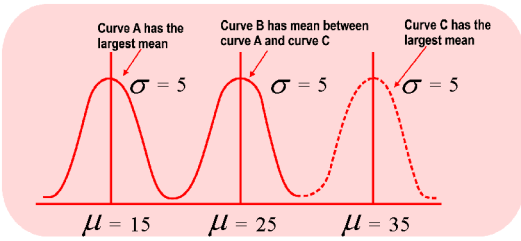


fig.6.5

We can draw the frequency curve for the normal distribution function by giving suitable values for μ and σ. For each different set of values of μ and we get different normal distributions. The value of μ determines the centre of normal distribution curve on the horizontal axis, and the value of σ gives the spread of normal distribution curve. Following figures will explain this point in detail.





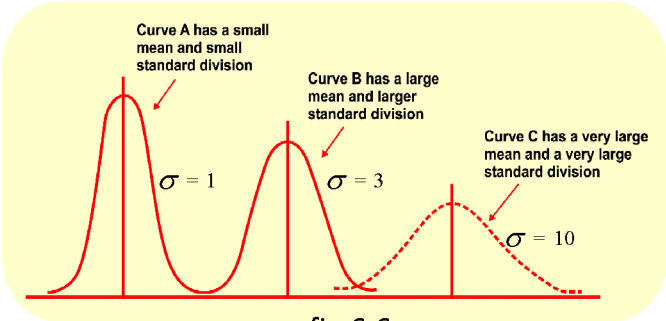


fig.6.6

(μ is location parameter, σ is also known as shape parameter).

Properties of Normal Curve

1. Normal curve is bell shaped and symmetric about X = μ .
2. Mean, Median and Mode coincide. ie, Mean = Median = Mode.
3. The height of normal curve is maximum at mean position i.e., at X = μ . $f(x) = \frac{1}{\sigma\sqrt{2\pi}}$,
4. Quartiles are equidistant from median.
5. Normal curve is unimodal.
6. Coefficient of skewness is zero i.e., $\beta_1 = 0$
7. Normal curve is mesokurtic i.e., $\beta_2 = 3$
8. Normal curve is asymptotic to the x axis. ie, on both ends away from the middle point, normal curve continues to decrease in height, but never touches the X axis
9. Total area under normal curve is 1 square unit.
10. For a normal distribution, Quartile deviation = $\frac{2}{3}\sigma$ and Mean deviation = $\frac{4}{5}\sigma$
11. The normal probability curve between the ordinates $\mu+\sigma$ and $\mu-\sigma$ covers approximately 68% of the total observations. Between the ordinates $\mu+2\sigma$ and $\mu-2\sigma$ covers approximately 95% of the total observations and between $\mu+3\sigma$ and $\mu-3\sigma$ covers 99% (almost all observations).

The discovery of the equation for a normal distribution can be traced to three mathematicians. In 1733, the French mathematician Abraham De Moivre derived an equation for a normal distribution based on the random variation of the number of heads appearing when a large number of coins were tossed. Not realizing any connection with the naturally occurring variables, he showed this formula to only a few friends. About 100 years later, two mathematicians, Pierre Laplace in France and Carl Gauss in Germany, derived the equation of the normal curve independently and without any knowledge of De-Moivre's work. In 1924, Karl Pearson found that De Moivre had discovered the formula before Laplace or Gauss.

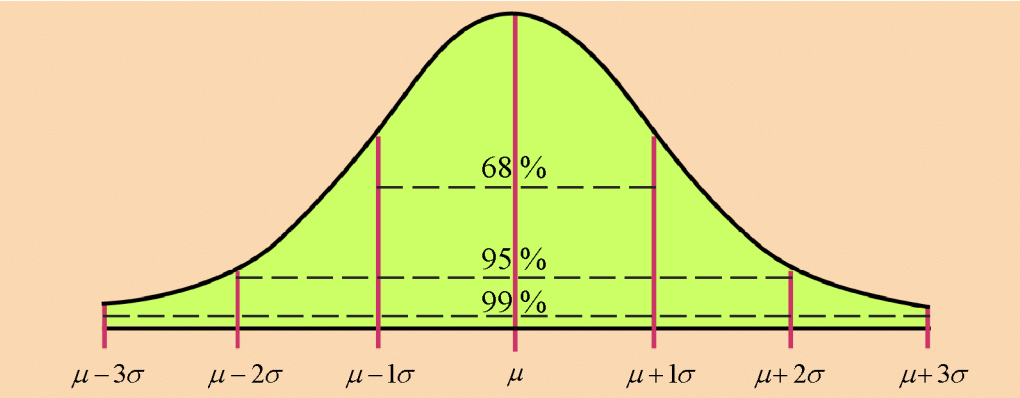


fig.6.7

6.3 Standard Normal distribution

A normal variable with mean 0 and standard deviation 1 is called Standard normal variable and its distribution is called Standard normal distribution. Usually Standard normal variable is denoted by the letter Z. The density function of Standard normal distribution is :

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} \text{ where } -\infty \le z \le \infty$$

The curve for Standard normal distribution is as follows

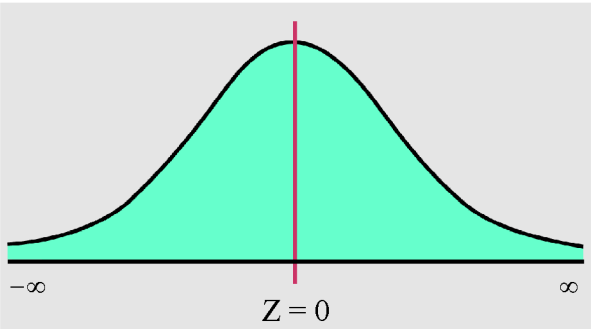


fig.6.8

Z transformation or Z score

Normal variable X with mean μ and standard deviation σ can be transformed in to Standard normal variable by a transformation $Z = \frac{X-\mu}{\sigma}$ is called Z transformation. Once X values are transformed by Z transformation they are called Z score or Z values.

i.e., If $X \sim N(\mu, \sigma^2)$ then $Z = \frac{X-\mu}{\sigma} \sim N(0,1)$.

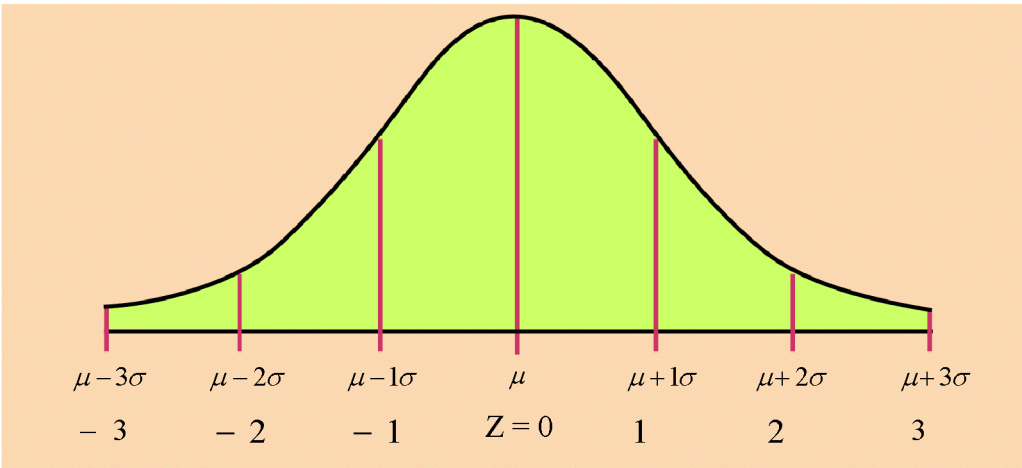


fig.6.9

Activity

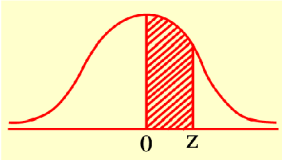
If $X \sim N(\mu, \sigma^2)$ then find E(Z) and V(Z).

Standard Normal Table

Evaluation of probability for a random variable lies in certain range using normal probability function is a difficult job.

i.e., $P(a < X < b) = \int_a^b \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2} dx$, the calculation involved here is not so easy.

To simplify the calculation involved in the evaluation of probability using normal probability function, statisticians use tables of area under the standard normal curve as the probability of the random variable. The tables of area under the standard normal curve is called standard normal table.



Thus, $P(0 < Z < z) = \int_0^z f(z)dz$

= Area under the standard normal curve bounded by the ordinates at Z = 0 and Z = z

If $X \sim N(\mu, \sigma^2)$ then

$$P(a < X < b) = P\left(\frac{a-\mu}{\sigma} < \frac{X-\mu}{\sigma} < \frac{b-\mu}{\sigma}\right)$$

$$= P\left(\frac{a-\mu}{\sigma} < Z < \frac{b-\mu}{\sigma}\right)$$

= Area under standard normal curve between the ordinates at $\frac{a-\mu}{\sigma}$ and $\frac{b-\mu}{\sigma}$

Finding area under the normal distribution curve

For the solution of problem on normal distribution following steps and procedure table will be helpful.

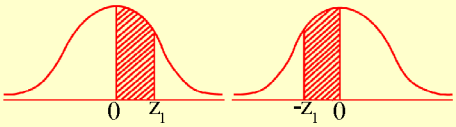
Step 1: Transform normal variable into standard normal variable by Z –transformation.

Step 2: Draw standard normal curve and locate the Z- value. Shade the area desired.

Step3: Select the appropriate block of procedure table given below.

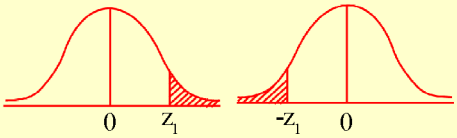
Block I

Between 0 and z_1



Look up the z_1 value in the table to get the area.

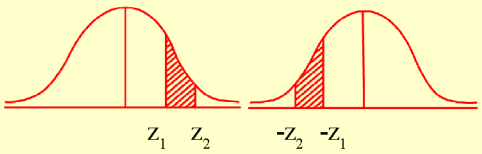
Block II



1. Look up the z_1 value in the table to get the area.
2. Subtract the area from 0.5.

Block III

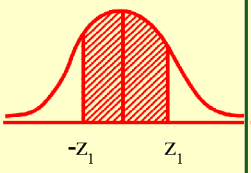
Between two z values on the same side of the mean



1. Look up both z_1 and z_2 in the table to get the area.
2. Subtract the smaller from the larger area.

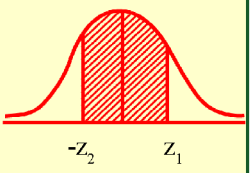
Block IV(a)

Between two z values on the either side of the mean



1. Look up z_1 value to get the area.
2. Add the areas.

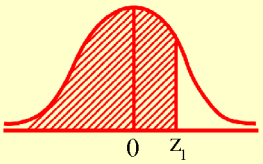
Block IV (b)



1. Look up z_1 and z_2 in the table to get the area.
2. Add both areas.

Block V

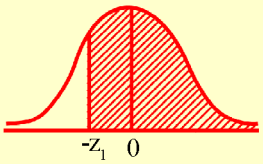
Less than any z value to the right of mean.



1. Look up z_1 value in the table to get the area.
2. Add 0.5 to the area.

Block VI

Greater than any z value to the left of mean.



1. Look up z_1 value in the table to get the area.
2. Add 0.5 to the area.

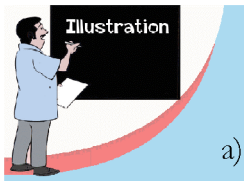


Illustration 6.1

If $Z \sim N(0,1)$ Find the probabilities

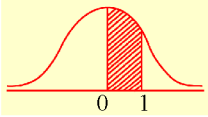
a) $P(0 < Z < 1)$ b) $P(-2 < Z < 0)$ c) $P(-2.5 < Z < 2.72)$

d) $P(2 < Z < 3)$ e) $P(-3 < Z < -2)$ f) $P(Z > 2)$

g) $P(Z < -2)$ h) $P(Z < 1.52)$ i) $P(Z > -1.52)$

Solution:

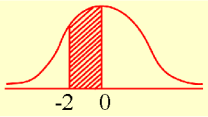
(a) $P(0 < Z < 1)$ = 0.3413 *(using statistical table)*



(Refer block No.1 in procedure table)

(b) $P(-2 < Z < 0)$ = $P(0 < Z < 2)$ *(symmetry)*

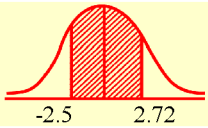
= 0.4772 *(using statistical table)*



(Refer block No.1 in procedure table)

(c) $P(-2.5 < Z < 2.72)$ = $P(-2.5 < Z < 0) + P(0 < Z < 2.72)$

= $P(0 < Z < 2.5) + P(0 < Z < 2.72)$ *(symmetry)*

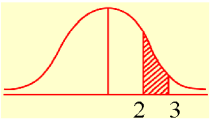


= 0.4938 + 0.4966 *(using statistical table)*

= 0.9904 *(Refer block No.IV(b) in procedure table)*

(d) $P(2 < Z < 3)$ = $P(0 < Z < 3) - P(0 < Z < 2)$

= 0.49865 – 0.4772 *(using statistical table)*

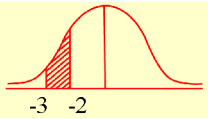


= 0.02145 *(Refer block No.III in procedure tables)*

(e) $P(-3 < Z < -2)$ = $P(2 < Z < 3)$ *(symmetry)*

= $P(0 < Z < 3) - P(0 < Z < 2)$

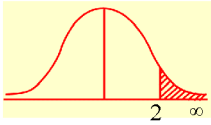
= 0.49865 – 0.4772 *(using statistical table)*



= 0.02145 *(Refer block No.III in procedure tables)*

(f) $P(Z > 2)$ = $P(0 < Z < \infty) - P(0 < Z < 2)$

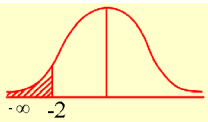
= 0.5 – 0.4772 *(using statistical table)*



= 0.0228 *(Refer block No.II in procedure table)*

(g) $P(Z < -2)$ = $P(Z > 2)$ *(symmetry)*

= $P(0 < Z < \infty) - P(0 < Z < 2)$



= 0.5 – 0.4772 *(using statistical table)*

= 0.0228 *(Refer block No.II in procedure table)*

(h) $P(Z < 1.52)$ $= P(-\infty < Z < 0) + P(0 < Z < 1.52)$

$= 0.5 + 0.4357$ *(using statistical table)*

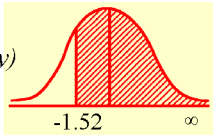
$= 0.9357$ *(Refer block No.V in procedure table)*

(i) $P(Z > -1.52)$ $= P(-1.52 < Z < 0) + P(0 < Z < \infty)$

$= P(0 < Z < 1.52) + P(0 < Z < \infty)$ *(symmetry)*

$= 0.4357 + 0.5$ *(using statistical table)*

$= 0.9357$ *(Refer block No.VI in procedure table)*



Know your progress

If $Z \sim N(0,1)$ then find the following probabilities

(1) $P(Z < 2.3)$ (2) $P(Z > 1.5)$ (3) $P(Z > -1.25)$

(4) $P(Z < -1.02)$ (5) $P(-2.4 < Z < 2.4)$ (6) $P(-1.4 < Z < 3.1)$

(7) $P(1.22 < Z < 2.80)$ (8) $P(-3.03 < Z < 0.72)$ (9) $P(0 < Z < 2.8865)$

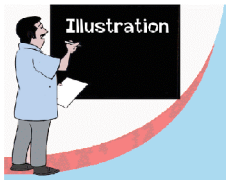


Illustration 6.2

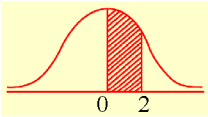
If X follows normal distribution with mean 30 and standard deviation 5. Find the $P(30 < X < 40)$

Solution: Given $\mu = 30$ and $\sigma = 5$

$P(30 < X < 40)$ $= P\left(\frac{30-\mu}{\sigma} < \frac{X-\mu}{\sigma} < \frac{40-\mu}{\sigma}\right)$

$= P\left(\frac{30-30}{5} < Z < \frac{40-30}{5}\right)$

$= P(0 < Z < 2) = 0.4772$ *(using statistical tables)*



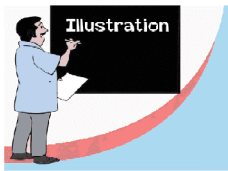


Illustration 6.3

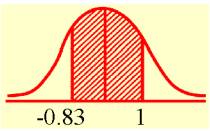
If $X \sim N(50, 6^2)$ then find

(1) $P(45 < X < 56)$ (2) $P(X > 60)$ (3) $P(X < 45)$

Solution: Given $\mu = 50$ and $\sigma = 6$

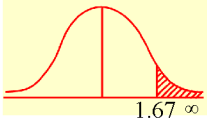
$P(45 < X < 56)$ $= P\left(\frac{45-\mu}{\sigma} < \frac{X-\mu}{\sigma} < \frac{56-\mu}{\sigma}\right) = P\left(\frac{45-50}{6} < Z < \frac{56-50}{6}\right)$

$= P(-0.83 < Z < 1)$



$= 0.2967 + 0.3413$ *(using statistical tables)*

$= 0.6380$



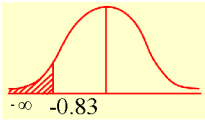
$P(X > 60) = P\left(\frac{X-\mu}{\sigma} > \frac{60-\mu}{\sigma}\right) = P\left(Z > \frac{60-50}{6}\right)$

$= P(Z > 1.67) = 0.5 - 0.4525 = 0.0475$

$P(X < 45) = P\left(\frac{X-\mu}{\sigma} < \frac{45-\mu}{\sigma}\right) = P\left(Z < \frac{45-50}{6}\right)$

$= P(Z < -0.83) = P(Z > 0.83)$ *(symmetry)*

$= 0.5 - 0.2967 = 0.2033$



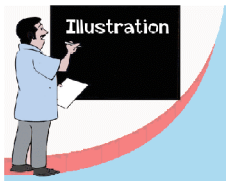


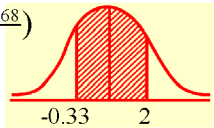
Illustration 6.4

If the weights of 300 students are normally distributed with mean 68 Kg and Standard deviation 3 kg. Find the number of students having weights between 67 kg and 74 kg

Solution: Let X be the normal variable that represents weight of a student.

Given $\mu = 68$ and $\sigma = 3$

$P(67 < X < 74) = P\left(\frac{67-\mu}{\sigma} < \frac{X-\mu}{\sigma} < \frac{74-\mu}{\sigma}\right) = P\left(\frac{67-68}{3} < Z < \frac{74-68}{3}\right)$



$= P(-0.33 < Z < 2)$

$= 0.1293 + 0.4772 = 0.6065$

The number of students having weights between 67 kg and 74 kg $= 300 \times P(67 < X < 74)$

$= 181.95 \simeq 182$

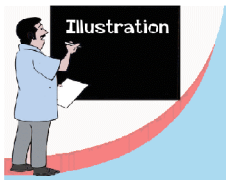


Illustration 6.5

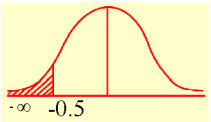
In an examination 500 students have appeared for a paper in Statistics. Their average score is 50 and standard deviation is 10. Estimate the percentage of students who got score less than 45.

Solution:

Let X be the normal variable that represents score

Given $\mu = 50$ and $\sigma = 10$

$P(X < 45) = P\left(\frac{X-\mu}{\sigma} < \frac{45-\mu}{\sigma}\right) = P\left(Z < \frac{45-50}{10}\right)$



$= P(Z < -0.5)$

$= 0.5 - 0.1915 = 0.3085$

Percentage of students who got a score less than 45 $= 100 \times P(X < 45)$

$= 100 \times 0.3085 = 30.85\%$

Know your progress

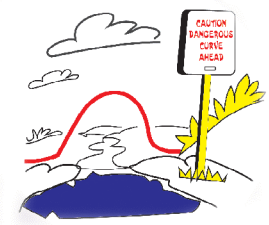
1) A random variable X follows normal distribution with mean 40 and standard deviation 8. Find (a) $P(30 < X < 50)$ (b) $P(X < 45)$

2) If $X \sim N(50, 5^2)$ then find (a) $P(X < 50)$ (b) $P(X < 60)$ (c) $P(30 < X < 70)$

3) The weekly wages of 1000 workers are normally distributed with mean Rs. 70 and Standard deviation Rs.5/-.Estimate the number of workers whose wages will be

(a) between Rs. 70/- and Rs, 72/- (b) more than 75/-

Find the value of 'z' for a given probability



When area between two points under the normal curve is known, we can find the values of scores by simply reversing the process. Following illustrations describe the procedure for finding the z value

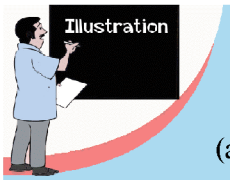


Illustration 6.6

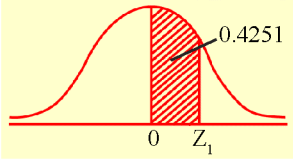
Let $Z \sim N(0,1)$ Find the value of Z_1 if

(a) $P(0 < Z < Z_1)=0.4251$ (b) $P(0 < Z < Z_1)=0.40$

(c) $P(Z < Z_1)=0.90$ (d) $P(Z > Z_1)=0.90$

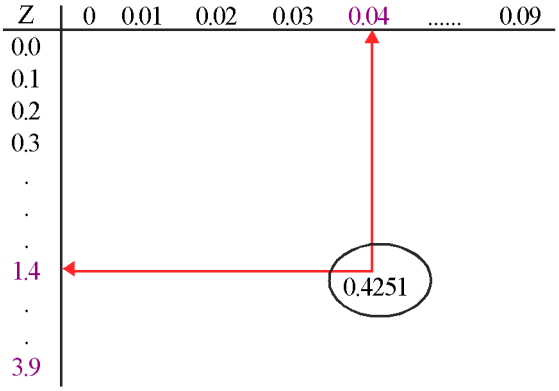
(e) $P(Z < Z_1)=0.15$ (f) $P(Z > Z_1)=0.15$

solution:



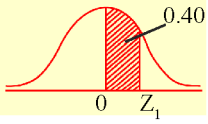
(a)

To find required value of Z_1, we locate 0.4251 in a standard normal table. Now read the numbers in column and row for Z_1 that corresponds to 0.4251, as shown in figure.



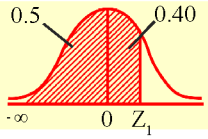
These numbers are 1.4 and 0.04 respectively. Combining these two numbers we get required value of $Z_1 = 1.44$.

(b)



We locate 0.40 or closest value to 0.40 in the standard normal table. We get $Z_1 = 1.28$.

(c)



Given $P(Z < Z_1) = 0.90$.

Since $P(Z < Z_1)$ is greater than 0.5 , so Z_1 must be positive

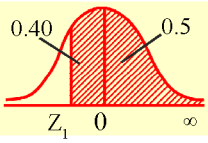
$P(-\infty < Z < 0) + P(0 < Z < Z_1) = 0.90$ *(refer standard normal curve)*

$0.5 + P(0 < Z < Z_1) = 0.90$

$P(0 < Z < Z_1) = 0.40$

$Z_1 = 1.28$

(d)



Given $P(Z > Z_1) = 0.90$

Since $P(Z > Z_1)$ is greater than 0.5 , so Z_1 must be negative

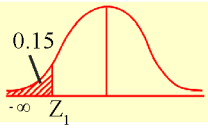
$P(0 < Z < \infty) + P(Z_1 < Z < 0) = 0.90$ *(refer standard normal curve)*

$0.5 + P(0 < Z < Z_1) = 0.90$ *(symmetry)*

$P(0 < Z < Z_1) = 0.40$

$Z_1 = -1.28$

(e)



Given $P(Z < Z_1) = 0.15$

Since $P(Z < Z_1)$ is less than 0.5 , so Z_1 must be negative

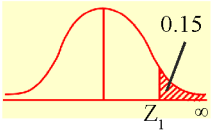
$P(-\infty < Z < 0) - P(Z_1 < Z < 0) = 0.15$ *(refer standard normal curve)*

$0.5 - P(0 < Z < Z_1) = 0.15$ *(symmetry)*

$P(0 < Z < Z_1) = 0.35$

$Z_1 = -1.04$

(f)



Given $P(Z > Z_1) = 0.15$

Since $P(Z > Z_1)$ is less than 0.5 , so Z_1 must be positive

$P(0 < Z < \infty) - P(0 < Z < Z_1) = 0.15$ *(refer standard normal curve)*

$P(0 < Z < Z_1) = 0.35$

$Z_1 = 1.04$

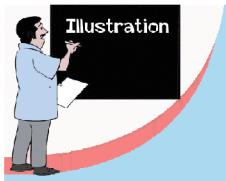
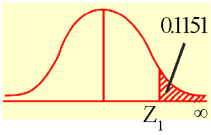


Illustration 6.7

The heights of soldiers are normally distributed. If 11.51% of the soldiers are taller than 70.4 inches and 9.68% are shorter than 65.4 inches, find the mean and standard deviation for the data of heights of soldiers.

Solution:

Since 11.51% are taller than 70.4 inches



We have, $P(X > 70.4) = 0.1151$

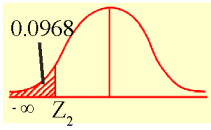
$P(Z > Z_1) = 0.1151$ where $z_1 = \frac{70.4-\mu}{\sigma}$

Now $Z_1 = 1.20$ *(from statistical tables)*

i.e., $\frac{70.4-\mu}{\sigma} = 1.20$

Hence $\mu + 1.20\sigma = 70.4$ - - - - - - - - equation (1)

since 9.68% are shorter than 65.4 inches



we have, $P(X < 65.4) = 0.0968$

$P(Z < Z_2) = 0.0968$ where $Z_2 = \frac{65.4-\mu}{\sigma}$

Now $Z_2 = -1.30$ *(from statistical tables)*

i.e., $\frac{65.4-\mu}{\sigma} = -1.30$

Hence $\mu - 1.30\sigma = 65.4$ - - - - equation (2)

solving equations (1) and (2),

we get standard deviation, $\sigma = 2$ inches, Mean, $\mu = 68$ inches

Know your progress

1) Let $Z \sim N(0,1)$. Find Z_1 if

(a) $P(Z < Z_1)=0.25$ (b) $P(Z > Z_1)=0.38$

(c) $P(Z < Z_1)=0.56$ (f) $P(Z > Z_1)=0.65$

2) If $X \sim N(24,9)$ and $P(X > a) = 0.974$, find the value of 'a'.

Let us conclude

A Normal distribution can be used to describe a variety of variables, such as heights, weights, and temperatures. Normal distribution is a continuous distribution with a single-peaked, bell-shaped curve. Normal curve is symmetric at the mean position. The two tails of normal curve extend indefinitely and never touching the X axis. Normal variable can be transformed in to standard normal variable by Z transformation. Standard normal distribution is a normal distribution with mean zero and variance one. Probability of a normal variable between some interval is obtained by finding area corresponds to that interval under the standard normal curve. We discussed a method of determining the mean and standard deviation of a normal variable from probability statements. A normal distribution can be used to solve a variety of problems in which the variables are approximately normally distributed.

Let us assess

For Questions 1-10, choose the correct answer from the given choices.

1. In Normal distribution, coefficient of skewness β_1 is

   a) one b) zero c) greater than one d) less than one

2. Mode of the normal distribution $N(\mu,\sigma^2)$ is

   a) μ b) 2μ c) σ d) 0

3. Total area under the Normal probability curve is

   a) less than one b) unity c) greater than one d) zero

4. The probability that a random variable x lies in the interval ($\mu - 2\sigma$, $\mu + 2\sigma$) is approximately equal to

   a) 0.95 b) 0.68 c) 0.99 d) 0.0027

5. The z is a standard normal variable, then P(z>4) is .....

   a) 0 b) 1 c) ∞ d) 0.5

6. For the Normal distribution

   a) mean = median =mode b) mean < median < mode

   c) mean > median > mode d) mean > median < mode

7. Probability density function of normal variable X is given by

   $f(x) = \frac{1}{5\sqrt{2\pi}} e^{-\frac{1}{2}\frac{(X-30)^2}{25}}; -\infty < X < \infty$ then mean and variance are:

   a) mean = 30 variance = 5 b) mean = 0, variance = 25

   c) mean = 30 variance = 25 d) mean = 30, variance = 10

8. The mean of a Normal distribution is 60, its mode will be

   a) 60 b) 40 c) 50 d) 30

9. If x is a normal variable with mean =100 and variance = 25 then P(90 < x < 120) is same as

   a) P(-1 < z < 1) b) P(-2 < z < 4) c) P(4 < z < 4.1) d) P(-2 < z < 3)

10. If $X \sim N(6, 1.2^2)$ and P(0 < Z < 1) = 0.3413 then P(4.8 < X < 7.2) is

    a) 0.3413 b) 0.6587 c) 0.6826 d) 0.3174

11. Find the area under the standard normal curve

    a) to the left of Z = 1 b) to the right of Z = - 1.6

    c) between Z= - 0.7 and Z= 1.414

12. A random variable X such that $X \sim N(50, 8)$. Find

    a) P(48 < X < 54) b) P(52 < X < 55)

    c)P(46 < X < 49) d) $P(|X - 50| < \sqrt{8})$

13. If P[Z < Z_1]=0.89 then find the value of Z_1.

14. Suppose the birth weight of a new born baby is a continuous random variable with the p.d.f, $f(x) = \frac{1}{0.5\sqrt{2\pi}} e^{-\frac{(x-3)^2}{2(0.5)^2}}; -\infty < X < \infty$.

i) Find the average birth weight of a body.

ii) Find the standard deviation of weight.

iii) Find the probability that the birth weight of a baby is less than 3 Kg.

15. If $X \sim N(40, 5^2)$ then find $P(45 \leq X \leq 55)$

16. If X follows normal distribution with mean 40 and standard deviation 3 then find $P[X<46]$

17. If X follows normal distribution with mean 62 and variance 9 then find probability of X lies between 60 and 65.

18. The average life of a brand of automobile tyres is 30,000 miles, with a standard deviation of 2000 miles. If a tyre is selected and tested, find the probability that it will have the following life time. (Assume the variable is normally distributed.)

    a) between 25000 and 28000 miles b) between 27000 and 33000 miles

19. The distribution of monthly income of 500 workers assumed to be normal with mean Rs.2000 and standard deviation of Rs. 200. Estimate the number of workers with incomes

    a) exceeding Rs. 2300 b) between Rs.1800 and Rs. 2300

20. As a result of tests on 2000 bulbs manufactured by a company, it was found that the life time of the bulb was normally distributed with an average life of 2040 hours and standard deviation of 60 hours. On the basis of the information estimate the percentage of bulbs that is expected to burn for

    a) more than 2150 hours b) less than 1960 hours.

21. The IQ score of students is normally distributed with mean of 120 and standard deviation of 20. Find the probability of students have an IQ

    a) between 100 and 130 b) above 140 c) Below 150

22. If $X \sim N(100, 6^2)$ and $P(X > a) = 0.1093$, find the value of a.

23. If $X \sim N(70, 25)$, find the value of 'a' such that and $P(|X - 70| < a) = 0.8$

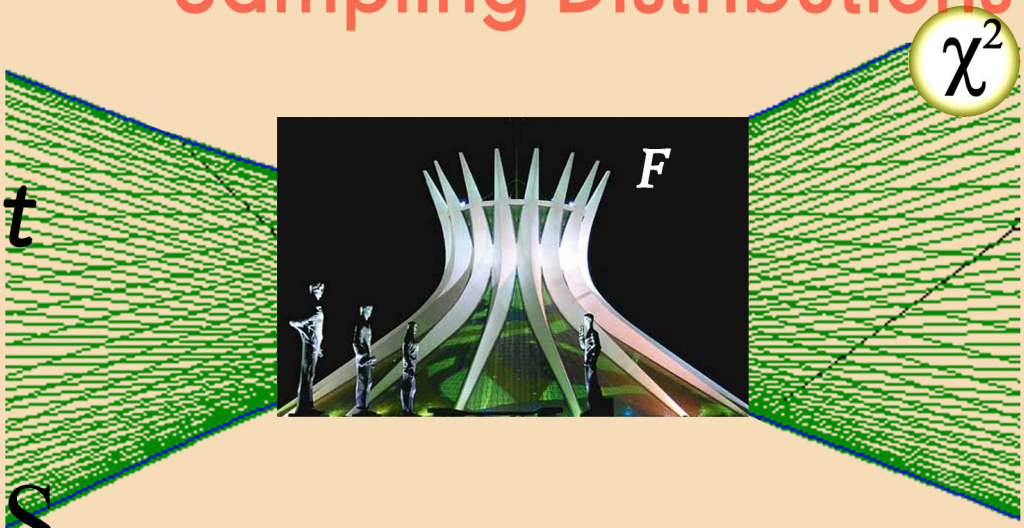
24. The lengths of certain items follow a normal distribution with mean μ cm and standard deviation 6 cm. It is known that 4.78% of the items have a length greater than 82cm. Find the value of the mean μ

25. If $X \sim N(100, \sigma^2)$ and $P(X < 106) = 0.8849$. Find the standard deviation, σ

26. In a distribution exactly normal, 7% of the items are under 35 and 89% are under 63. Find the mean and standard deviation of the distribution.

Chapter 7

Sampling Distributions



Since complete study of population is not always possible, sampling methods are extensively used in almost all of the studies relating to life. In continuation to discrete and continuous probability distributions, our discussion is now progressing to one step forward to Sampling Distributions. In this chapter we familiarize with the terms like statistic, parameter, probability distributions of statistic etc.

It is known that the sample mean, $\bar{x}$ follow normal distribution even when the parent population is non- normal provided that the sample is large. The result follows from a most celebrated theorem- Central Limit Theorem.

Significant Learning Outcomes

After the completion of this chapter, the learner:

- Identifies the meaning and concept of sampling distribution.
- Differentiates between parameter and statistic and to exemplify them.
- Understands the uses and application of different types of sampling distributions.
- Develops the relationship between Chi-Square, Student's t and Snedecor's F statistics.

The sample and its parent population are related to each other. The kind of information we gather from a sample or a series of samples about the population are also very important. The objective is to know the manner in which the function of sample values vary from one sample to another. Sampling distributions explain the way in which a function of sample values behave. Here we get knowledge about some important sample statistics, relation among them, we familiarize with the tables of their distributions and their fields of application.

7.1. Parameter and Statistic

Suppose we are interested in calculating average height of HSS students in Kerala. It is very difficult to measure the height of each and every student. So we will collect the information from a representative number of students from different parts of the state. As you know, totality of all the HSS students in Kerala constitutes the **population** and representative part of it constitutes **sample.**

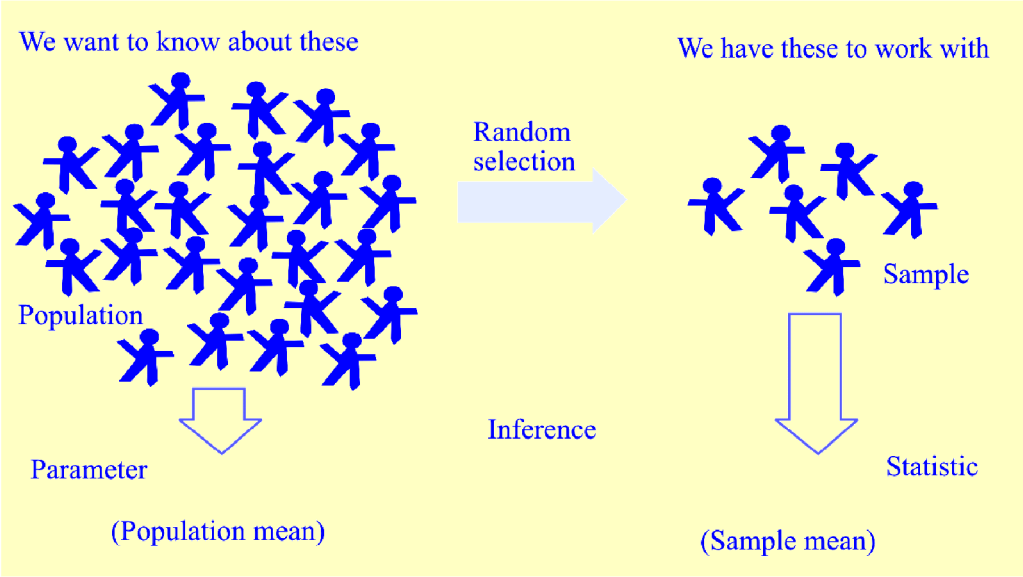
We can describe samples and populations by using measures such as mean, median, mode, range, S.D., etc. When these terms describe the characteristics of a sample, they are called statistics. When they describe the characteristics of a population, they are called parameters.

> A parameter is a measurable characteristic of a population.
>
> A statistic is a measurable characteristic of a sample.

Thus population parameter is a constant computed for the entire population *viz.* population mean, population median, population variance and so on. Population parameter is unknown but fixed, whose value is to be estimated from the sample statistic. Sample Statistic is any constant computed from our sample data viz. sample mean, sample median, sample SD, sample variance and so on.

Usually we denote sample mean as $\bar{x}$, sample standard deviation as 's' and sample variance as 's^2'. Similarly population mean as μ, population standard deviation as 'σ', and population variance as 'σ^2'.

	Statistic	Parameter
Mean	$\bar{x}$	μ
Standard Deviation (SD)	s	σ
Variance	s^2	σ^2



7.2 Sampling distribution

You are familiar with different methods of sampling like Simple Random Sampling With Replacement (SRSWR) and Simple Random Sampling Without Replacement (SRSWOR) and so on.

Let 'N' be the population size, and 'n' be the size of the required sample.

If we are considering Simple Random Sampling Without Replacement (SRSWOR), as we know the number of combinations of samples of size 'n' units can be drawn from 'N' units is NCn ways and hence the probability to select a sample is $\frac{1}{NCn}$.

Consider a population of size '5' having units 3,5,6,8 and 11 and take a sample of size '2'. The number of possible sample is $5C_2 = \frac{5!}{(5-2)!2!} = 10$.

Let us try to write the samples (SRSWOR) and the sample mean.

Sample Numbers	Sample values (x_1, x_2)	Mean $(\bar{x})$	Probability of sample
1	(3,5)	4	$\frac{1}{10}$
2	(3,6)	4.5	$\frac{1}{10}$
3	(3,8)	5.5	$\frac{1}{10}$
4	(3,11)	7	$\frac{1}{10}$
5	(5,6)	5.5	$\frac{1}{10}$
6	(5,8)	6.5	$\frac{1}{10}$
7	(5,11)	8	$\frac{1}{10}$
8	(6,8)	7	$\frac{1}{10}$
9	(6,11)	8.5	$\frac{1}{10}$
10	(8,11)	9.5	$\frac{1}{10}$
	Total $(\Sigma\bar{x})$	**66**	

Table 7.1.

Note that the sample means vary from sample to sample.

When we consider Simple Random Sample With Replacement (SRSWR), the number of possible samples, is N^n and the probability to select a sample is $\frac{1}{N^n}$.

With reference to the above example, the number of combinations of all possible samples by taking simple random sample with replacement is $5^2 = 25$.

In the similar way listed above try to list the samples and sample means.

Sample Numbers	Sample values (x_1, x_2)	Means $(\bar{x})$	Probability of sample	Sample Numbers	Sample values (x_1, x_2)	Means $(\bar{x})$	Probability of sample
1	(3,3)	3	$\frac{1}{25}$	14	(6,8)	7	$\frac{1}{25}$
2	(3,5)	4	$\frac{1}{25}$	15	(6,11)	8.5	$\frac{1}{25}$
3	(3,6)	4.5	$\frac{1}{25}$	16	(8,3)	5.5	$\frac{1}{25}$
4	(3,8)	5.5	$\frac{1}{25}$	17	(8,5)	6.5	$\frac{1}{25}$
5	(3,11)	7	$\frac{1}{25}$	18	(8,6)	7	$\frac{1}{25}$
6	(5,3)	4	$\frac{1}{25}$	19	(8,8)	8	$\frac{1}{25}$
7	(5,5)	5	$\frac{1}{25}$	20	(8,11)	9.5	$\frac{1}{25}$
8	(5,6)	5.5	$\frac{1}{25}$	21	(11,3)	7	$\frac{1}{25}$
9	(5,8)	6.5	$\frac{1}{25}$	22	(11,5)	8	$\frac{1}{25}$
10	(5,11)	8	$\frac{1}{25}$	23	(11,6)	8.5	$\frac{1}{25}$
11	(6,3)	4.5	$\frac{1}{25}$	24	(11,8)	9.5	$\frac{1}{25}$
12	(6,5)	5.5	$\frac{1}{25}$	25	(11,11)	11	$\frac{1}{25}$
13	(6,6)	6	$\frac{1}{25}$	**Total ($\sum\bar{x}$)**		**165**	

Table 7.2.

From Table 7.1 and Table 7.2 we can see that, the value of sample means changes from sample to sample (the sample is taken by SRSWOR in Table 7.1 and by SRSWR in Table 7.2.) .

From the above discussions, the numerical value calculated for mean varies from sample to sample. Therefore it is a random variable. In a similar way we can compute any statistics like median, mean deviation, mode, standard deviations and can be shown that they are random variables.

This means that each statistic has its own distribution and is called the sampling distribution. Sampling distributions are probability distribution of sample statistic.

Probability distribution of a statistic is called **sampling distribution.**

7.3. Distribution of Sample Mean

From Table **7.1** and Table **7.2** , we have following probability distribution for sample mean.

(i) In SRSWOR

$\bar{x}$	4	4.5	5.5	6.5	7	8	8.5	9.5	Total Probability
$P(\bar{x})$	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{2}{10}$	$\frac{1}{10}$	$\frac{2}{10}$	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{1}{10}$	$\Sigma P(\bar{x})=1$

Table 7.3

(ii) In SRSWR, we have following probability distribution for sample mean.

$\bar{x}$	3	4	4.5	5	5.5	6	6.5	7	8	8.5	9.5	11	Total Probability
$P(\bar{x})$	$\frac{1}{25}$	$\frac{2}{25}$	$\frac{2}{25}$	$\frac{1}{25}$	$\frac{4}{25}$	$\frac{1}{25}$	$\frac{2}{25}$	$\frac{4}{25}$	$\frac{3}{25}$	$\frac{2}{25}$	$\frac{2}{25}$	$\frac{1}{25}$	$\Sigma P(\bar{x})=1$

Table 7.4

Standard Error of Sample mean

The standard deviation of a statistic is known as its standard error.

If the samples are taken according to SRSWR

sample variance, $V(\bar{x}) = \frac{\sigma^2}{n}$

Standard Error of sample mean, $SE(\bar{x}) = \sqrt{V(\bar{x})} = \sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}}$

If the samples are taken according to SRSWOR

sample variance, $V(\bar{x}) = \left(\frac{N-n}{N-1}\right)\frac{\sigma^2}{n}$

Standard Error of sample mean, $SE(\bar{x}) = \sqrt{V(\bar{x})} = \sqrt{\left(\frac{N-n}{N-1}\right)\frac{\sigma^2}{n}}$

Let us compute the standard deviation of sampling distribution of means for the example discussed earlier.

$\bar{x}$	$f(\bar{x})$	$\bar{x}f(\bar{x})$	$(\bar{x})^2$	$(\bar{x})^2 f(\bar{x})$
3	$\frac{1}{25}$	0.12	9	0.36
4	$\frac{2}{25}$	0.32	16	1.28
4.5	$\frac{2}{25}$	0.36	20.25	1.62
5	$\frac{1}{25}$	0.2	25	1
5.5	$\frac{4}{25}$	0.88	30.25	4.84
6	$\frac{1}{25}$	0.24	36	1.44
6.5	$\frac{2}{25}$	0.52	42.25	3.38
7	$\frac{4}{25}$	1.12	49	7.84
8	$\frac{3}{25}$	0.96	64	7.68
8.5	$\frac{2}{25}$	0.68	72.25	5.78
9.5	$\frac{2}{25}$	0.76	90.25	7.22
11	$\frac{1}{25}$	0.44	121	4.84
	Total	**6.60**		**47.28**

Table 7.5

$$V(\bar{x}) = E(\bar{x})^2 - \left(E(\bar{x})\right)^2$$

$$= \sum (\bar{x})^2 f(\bar{x}) - \left[\sum (\bar{x}) f(\bar{x})\right]^2$$

$$= 47.28 - 6.6^2 = 3.72. \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (1)$$

Variance of sample mean is 3.72

Now, let us calculate σ^2 of the population values 3,5,6,8,11.

we get $\sigma^2 = \sum \frac{X^2}{N} - \left(\frac{\sum X}{N}\right)^2$ =7.44

We know, in SRSWR variance of sample mean,

$$V(\bar{x}) = \frac{\sigma^2}{n} = \frac{7.44}{2} = 3.72 \ldots\ldots\ldots\ldots\ldots(2)$$

(1) and (2) are same. Second method is much easy to calculate variance of sample mean. Similarly you can verify variance of sample mean in SRSWOR.

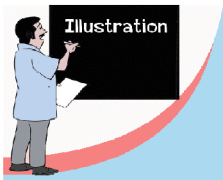


Illustration 7.1:

Draw all possible samples of size 2 from a population consisting of 2,3, 4,5 and find

(i) $E(\bar{x})$ (ii) $V(\bar{x})$ and (iii) Standard Error of sample mean

(a)without replacement (b) with replacement.

Solution:

(a) No of possible samples in SRSWOR= $4C_2$=6

Sample Numbers	Samples (x_1, x_2)	Means $(\bar{x})$
1	(2,3)	2.5
2	(2,4)	3
3	(2,5)	3.5
4	(3,4)	3.5
5	(3,5)	4
6	(4,5)	4.5
	Total	**21**

(i) $E(\bar{x}) = \frac{21}{6} = 3.5$

(ii) $\sigma^2 = \sum \frac{X^2}{N} - \left(\frac{\sum X}{N}\right)^2$

$$= \frac{54}{4} - \left(\frac{14}{4}\right)^2 = 13.5 - 12.25$$

$$= 1.25$$

$$V(\bar{x}) = \left(\frac{N-n}{N-1}\right)\frac{\sigma^2}{n}$$

$$= \left(\frac{4-2}{4-1}\right)\frac{1.25}{2} = 0.42$$

(iii) Standard Error of sample mean,

$$SE(\bar{x}) = \sqrt{V(\bar{x})} = \sqrt{0.42} = 0.648$$

(b) No of possible samples in SRSWR= 4^2=16

Sample Numbers	Samples (x_1, x_2)	Mean $(\bar{x})$	Sample Numbers	Samples (x_1, x_2)	Mean $(\bar{x})$
1	(2,2)	2	9	(4,2)	3
2	(2,3)	2.5	10	(4,3)	3.5
3	(2,4)	3	11	(4,4)	4
4	(2,5)	3.5	12	(4,5)	4.5
5	(3,2)	2.5	13	(5,2)	3.5
6	(3,3)	3	14	(5,3)	4
7	(3,4)	3.5	15	(5,4)	4.5
8	(3,5)	4	16	(5,5)	5
					$\sum \bar{x} = 56$

(i) $E(\bar{x}) = \frac{56}{16} = 3.5$

(ii) $\sigma^2 = 1.25$

$$V(\bar{x}) = \frac{\sigma^2}{n} = \frac{125}{2} = 0.625$$

(iii) Standard Error of sample mean,

$$SE(\bar{x}) = \sqrt{V(\bar{x})} = \sqrt{0.625} = 0.79$$

Activity

Random samples of size 2 with replacement are drawn from the population consisting of four numbers 4, 5, 6, 7. Find sample mean $\bar{x}$ for each sample and make sampling distribution of $\bar{x}$. Calculate the mean and standard deviation of this sampling distribution. Compare your calculations with population mean.

Know your progress

Draw all possible samples of size 2 from a population consisting of 5,8,10,11 and find,

(i) $E(\bar{x})$ (ii) $V(\bar{x})$ and (iii) Standard Error of sample mean

(a) without replacement (b) with replacement

7.4. Central Limit Theorem and its importance

The Central Limit Theorem is stated as follows:

Let $x_1, x_2 \ldots, x_n$ be a sequence of independent random variables each having the same distribution with finite mean μ and finite variance σ^2. The sample mean of $x_1, x_2 \ldots, x_n$, $\bar{x}$ follows approximately normal distribution with mean μ and variance $\frac{\sigma^2}{n}$, when 'n' is large. i.e, standardized variable

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1) \text{ as } n \to \infty$$

Central limit theorem states that regardless of the underlying distribution of x, the distribution of $\bar{x}$ is normal if sample size is sufficiently large.

Usually, samples with size $n < 30$ are considered as small samples. For a given population of size N, we can draw different samples, each of size 'n'. For these samples we can compute a statistic (for example, sample mean, sample variance, etc.) which will vary from sample to sample.

7.5 Chi-square, t and F distributions

Following are some important sampling distributions

- Chi-square distribution
- t-distribution
- F-distribution

Chi–Square distribution

Let $Z_1, Z_2, \ldots Z_n$ are n independent standard normal variables then the sum of squares of these variables is said to follow a Chi-square distribution with n degrees of freedom(d.f) and denoted as $\chi^2_{(n)}$

i.e, $Z_1^2+Z_2^2+\ldots+Z_n^2 = \chi^2_{(n)}$

This Chi-square varibale will follow a distribution which is called Chi-square distribution.

Suppose $X_1, X_2, \ldots, X_n$ be a random sample from a normal distribution with mean μ and variance σ^2 then $Z_i = \frac{x_i - \mu}{\sigma}$, i = 1, 2,..., n are independent standard normal variables.

So $\sum Z_i^2$ follows $\chi^2_{(n)}$. The variable $\sum Z_i^2 = \sum\left(\frac{x_i - \mu}{\sigma}\right)^2$ is also called χ^2 variable.

Degrees of freedom: The degrees of freedom refer to the number of independent observations in a set of data. suppose we are asked to choose 10 numbers. We have then the freedom to choose 10 numbers as we please, and have 10 degrees of freedom. But suppose a condition is imposed on the numbers. The condition is that the sum of all the numbers we choose must be 100. In this case we can't choose all ten numbers as we please. After we have chosen 9th number, let us say the sum of 9 numbers is 94. Our 10th number has to be 6,and we have no choice. Thus we have only 9 degrees of freedom. If we have to choose n numbers, and a condition on their total is imposed, we will have only (n-1) degrees of freedom.

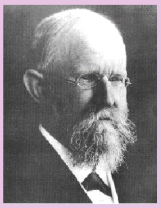
Properties of Chi-square distribution

1. If χ^2 follows χ^2 distribution with 'n' d.f then Mean (χ^2) = n, Variance (χ^2)=2n
2. If Z is a standard normal variable, the Square of Standard Normal Variable, ie. $Z^2 \sim \chi^2_{(1)}$

3. Chi-square distribution is positively skewed.

4. As $n \to \infty$, χ^2 distribution can be approximated by normal distribution.

Note: The ratio $\frac{ns^2}{\sigma^2}$ will be a χ^2 with (n-1) d. f and is denoted as $\chi^2_{(n-1)}$ where s^2 is the sample variance.



Chi-square distribution was first described by the German statistician Friedrich Robert Helmert where he computed the sampling distribution of the sample variance of a normal population. Thus in German this was traditionally known as the *Helmertsche* ("Helmertian") or "Helmert distribution".

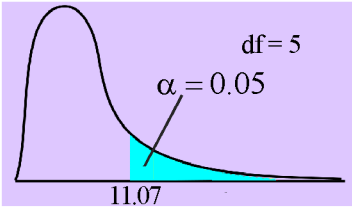
The distribution was independently rediscovered by the English mathematician Karl Pearson in the context of goodness of fit, for which he developed his Pearson's chi-squared test, with computed table of values The name "chi-squared" ultimately derives from Pearson's shorthand with the Greek letter Chi, writing χ.

Applications of Chi-square distribution

Chi-square distribution is used to test

- Goodness of fit.
- The independence of attributes.
- The variance of single population equals a given value.

Table of Chi-square distribution



Chi-square table contains the chi-square values for particular value of significance level (α) (top of the table) for different d.f.(given in the left side column of the table)

For $\alpha = 0.05$ and d.f 5, the value of $\chi^2 = 11.07$

This can be expressed as $P(\chi^2_{(5)} > 11.07) = 0.05$

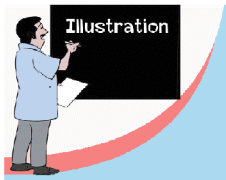


Illustration 7.2

Find the value of χ^2 if

(a) $P(\chi^2_{(10)} > \chi^2) = 0.1$ (b) $P(\chi^2_{(8)} > \chi^2) = 0.01$

(c) $P(\chi^2_{(2)} > \chi^2) = 0.05$

Solution:

(a) $\chi^2 = 15.99$ (b) $\chi^2 = 20.09$ (c) $\chi^2 = 5.99$

Know your progress

Find the value of χ^2 if

(a) $P(\chi^2_{(15)} > \chi^2) = 0.10$ (b) $P(\chi^2_{(7)} > \chi^2) = 0.05$

(c) $P(\chi^2_{(10} > \chi^2) = 0.01$

t distribution

If Z is standard normal variable and Y is an independent χ^2 variable with n degrees of freedom, then $\frac{Z}{\sqrt{\frac{Y}{n}}}$ is known as 't' variable or student's t variable with 'n' d.f.

ie, $$t = \frac{Z}{\sqrt{\frac{Y}{n}}}$$

The t-statistic was introduced in 1908 by William Sealy Gosset, (1876 - 1937) a chemist working for the Guinness brewery in Dublin, Ireland ("Student" was his pen name) Company policy at Guinness forbade its chemists from publishing their findings, so Gosset published his mathematical work under the pseudonym (pen name) "Student".

This t- variable will follow a distribution which is called t- distribution.

Properties of t- distribution

1. If t follows t - distribution with 'n' d.f then E (t) = 0, V(t)= $\frac{n}{n-2}$ *where* $n > 2$.
2. t-distribution is symmetric at t=0.
3. As n $\to \infty$ t-distribution $\to$ normal distribution.

Note: The ratio of Standard normal statistic $\frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}}$ to the square root of χ^2 statistic, $\sqrt{\frac{ns^2}{\sigma^2}}$ dividing with its d.f. will be a t statistic having d.f, (n-1).

ie., $$t = \frac{\bar{x}-\mu}{\frac{s}{\sqrt{n-1}}} \sim t(n-1)$$

Applications of t - Statistic

t statistic is used to test

- The mean of a small sample drawn from a normally distributed population when the population standard deviation is unknown.
- Two independent random samples have the same mean

Tables of t- distribution

t- Table contains the t values for particular value of significance level (α)

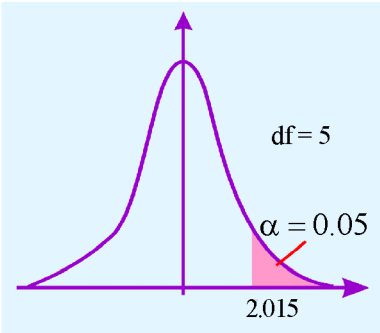
For α =0.05 and d.f 5, the value of t=2.015

Statistically this can be expressed as

$P(t_5 > 2.015) = 0.05$ (one tail)

Also this can be expressed as

$P(|t_5| > 2.015) = 2 \times 0.05 = 0.10$ (two tail)



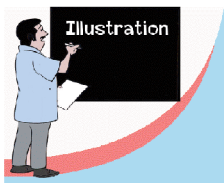


Illustration 7.3

Find the value of t if

(a) $P(t_{(10)} > t) = 0.10$ (b) $P(t_{(8)} > t) = 0.01$ (c) $P(t_{(2)} > t) = 0.05$

(d) $P(|t_5| > t) = 0.05$ (e) $P(|t_{10}| > t) = 0.02$

Solution:

(a) t= 1.372 (b) t= 2.896 (c) t= 2.92 (d) t=2.57 (e) t=2.76

Know your progress

Find the value of t if

(a) $P(t_{(15)} > t) = 0.10$ (b) $P(t_{(7)} > t) = 0.05$ (c) $P(|t_6| > t) = 0.01$

Snedecor's F distribution

Let χ_1^2 and χ_2^2 are two independent chi square random variables with n_1 and n_2 degrees of freedom respectively. Then the ratio of two chi-square variables dividing by their corresponding degrees of freedom will be an F variable with (n_1, n_2) d.f.

i.e., $F = \dfrac{\chi_1^2 / n_1}{\chi_2^2 / n_2}$ F variable will follow a distribution which is called F distribution.

Properties of F distribution

1. If F follows F distribution with (n_1, n_2) d.f then $\frac{1}{F}$ follows F distribution with (n_2, n_1) d.f.
2. F distribution is positively skewed.
3. As $n_1 \to \infty$ and $n_2 \to \infty$, F distribution $\to$ normal distribution
4. If s_1^2 and s_2^2 are the sample variances of two independent random samples of size n_1 and n_2 taken from two normal populations with variances σ_1^2 and σ_2^2, then

$$\frac{\dfrac{n_1 s_1^2}{\sigma_1^2} \Big/ n_1 - 1}{\dfrac{n_2 s_2^2}{\sigma_2^2} \Big/ n_2 - 1}$$

follows F-distribution with (n_1-1, n_2-1) d.f.

The *F*-distribution is also known as Snedecor's *F* distribution or the Fisher–Snedecor distribution named after R. A. Fisher and George W. Snedecor (20 October 1881 to 15 February 1974). George W.Snedecor was an American mathematician and statistician. He contributed to the foundations of analysis of variance, data analysis, experimental design, and statistical methodology.

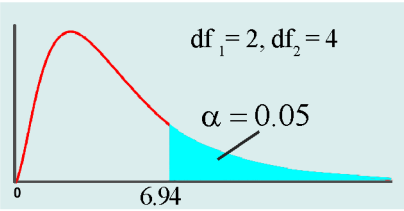
Applications of F distribution

F-distribution is used to test

- Equality of three or more means.
- Equality of two sample variances.

Tables of F distribution

The tables of F statistic are arranged for different pairs of d.fs (n_1, n_2) and for two different levels of significance (0.01 and 0.05). The body of table contains different values of F. First row contains n_1 values and the first column contains n_2.



F- Table contains the F values for particular value of significance level (α).

For α =0.05 and d.f n_1=2, n_2=4, the value of F=6.94.

Statistically this can be expressed as $P(F_{(2,4)} > 6.94) = 0.05$.

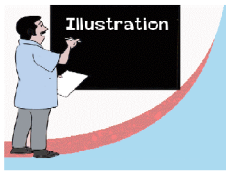


Illustration 7.4

Find the value of F if

(a) $P(F_{(5,12)}>F) = 0.05$ (b) $P(F_{(8,15)}>F) = 0.05$

Solution:

(a) F= 3.11 (b) F= 2.64

These tables are very much useful in solving problems related to statistical inference.

Know your progress

Find the value of F if

(a) $P(F_{(8,19)}>F) = 0.05$ (b) $P(F_{(8,15)}>F) = 0.01$

7.6 Relation among among Z, χ^2, t and F Statistics

Relation between Z and χ^2

If $Z_1, Z_2, \ldots Z_n$ be 'n' independent Standard normal variables, then $Z_1^2 + Z_2^2 + \ldots + Z_n^2$ follows χ^2 with 'n' d.f.

If n = 1, then Z^2 follows χ^2 with 1 d.f.

Relation between χ^2 and t

If Z is standard normal and Y is a $\chi^2_{(n)}$ then the ratio of Z and $\sqrt{\frac{Y}{n}}$ will result a t statistic with 'n' d.f. Hence the square of t will be ratio of two chi- square statistic.

Relation between χ^2 and F

The ratio of two independent chi-square variables which are divided by their respective d.f's n_1 and n_2 will be an F with d.f (n_1, n_2).

Relation between t and F

We have t - variable, $t = \frac{z}{\sqrt{\frac{\chi^2}{n}}}$ where ie, $z \sim N(0,1)\, and\, \chi^2 \sim \chi^2(n)$.

Squaring on both sides, we have $t^2 = \frac{z^2}{\frac{\chi^2}{n}} = \frac{\frac{z^2}{1}}{\frac{\chi^2}{n}} = \frac{\frac{\chi^2}{1}}{\frac{\chi^2}{n}} = F(1,n)$

Thus, square of t statistic will be an F with degrees of freedom (1, n).

Activity

Find different values of t, χ^2 and F statistic values for different degrees of freedom and for different significance level α.

Let us conclude

In this chapter we were familiarized with the meaning of parameter and statistic, their distinction. This chapter went through some supporting mathematical proof of relation between parameter and statistic. In simple random sampling techniques, one can easily realize that the expected value of sample statistic is the population parameter itself. This chapter seriously discussed three important statistic – Chi-square, Student's t and Snedecor's F statistic. The relationships among these statistics were also discussed. Familiarization of the tables of these statistic and their uses were also attempted. The reason for discussing them is that they feature prominently in the testing of hypotheses.

Let us assess

For Questions 1-5, choose the correct answer from the given choices.

1. The Square of Standard Normal variable will be a _____

   a) Normal b) t-statistic c) Chi-square statistic d) F statistic

2. The ratio of two Chi-Square statistic is _____

   a) Normal b) t-statistic c) Chi-square statistic d) F statistic

3. The Square of t Statistic is _____

   a) Normal b) t-statistic c) Chi-square statistic d) F statistic

4. If x ~ N(0,1) and y~ χ^2 (n) then $\frac{x}{\sqrt{\frac{y}{n}}}$ follows.....

a) t distribution with 1d.f b) t distribution n d.f

c) F distribution with (n,1) d.f d) F distribution with (1,n) d.f

5. If Y follows student t distribution with n d.f., then Y^2 follows

a) t distribution with n d.f b) χ^2 with n d.f

c) F distribution with (1,n) d.f d) F distribution with (n,1) d.f

6. The ratio of two independent chi-square variables is .....

7. _____ can be used for conversion of variables even though the population is not normal

8. As per Central Limit Theorem, a distribution with mean μ and variance $ó^2$, the sampling distribution of Sample mean approaches a normal distribution with mean _____ and Variance _____.

9. A sample of size 10 is taken from a population with variance 25 then the variance of sample mean is—

10. Read the following statements and mark them True / false

a) The numerical value calculated for mean, median, mean deviation, mode, Standard Deviations of sample data are called sample statistic.

b) The expected value of the sample mean is Population mean.

c) As sample size increased, the sampling distribution of the mean approaches the normal distribution regardless of the population distribution.

d) The variance of distribution of the random variable $\bar{x}$ is termed as SE of Mean.

e) The CLT is assumed applicable if the sample size is small.

f) Standard Error of mean varies inversely with the SD of population.

11. Explain the concept of statistic and parameter with an example.

12. Explain the concept of sampling distribution and standard error.

13. Establish the relationship among t-statistic, Chi-Square statistic, F statistic.

14. Match the following

A	B
1) $\dfrac{\bar{x}-\mu}{\dfrac{s}{\sqrt{n-1}}}$	a) Snedecor's F variable
2) Z^2	b) Standard normal
3) t^2	c) Chi-square variable
4) $\left[\dfrac{\bar{x}-\mu}{\dfrac{\sigma}{\sqrt{n}}}\right]$	d) t- variable

15. Consider a population of size four, having observations, 58,52,56,63.

    a) Write all possible samples of size 2 by taking simple random sample with replacement.

    b) Find population standard deviation.

16. Using simple random sampling without replacement (SRSWOR), Select samples of size two from the data 2, 3, 6, 8 and 11.

    a) Write all possible samples.

    b) Find the standard error of sample mean.

17. A population consists of values 2,4,6,8,10. Consider all possible random samples of size 3 in SRSWOR.

    a) Write all possible samples. b) $E(\bar{x})$ and $V(\bar{x})$

18. The ages of Badminton players of a school are 18,15,16,19.

    a) Find the average age of a player.

    b) If a team of 2 may be selected according to SRSWOR then how many different possible samples are there? List out the samples.

    c) Find standard error of estimate.

19. Fill up the missing values for the table given below.

Sl No	Value of Variate	α	d.f.
1	χ^2 = ——	0.80	18
2	χ^2 = 22.362	——	13
3	t = ——	0.025	16
4	t = 2.779	0.005	——
5	t = ——	0.025	160

Chapter 8

Estimation of Parameters

Every individual makes estimations in their daily life. When you are ready to cross a road, you estimate the speed of the vehicle that is approaching, the distance between you and the vehicle, and your own speed. Having made these quick estimates, you decide whether to wait, walk or run. Consider another situation, a company which is manufacturing electrical bulbs wants to estimate the average life of a bulb. The company cannot test all the bulbs. Rather it will take a sample and through the sample it will estimate the average life of a bulb in the population. Our real

Significant Learning Outcomes

After the completion of this chapter, the learner:

- Identifies the application of Inferential Statistics.
- Recognises the estimator and estimate.
- Describes point estimation and interval estimation.
- Lists out the desirable properties of a good estimator.
- Examines the properties of estimators.
- Estimates the parameter using method of moments.
- Constructs confidence intervals for population mean.

interest in a statistical study is to draw conclusions about the population. For that we take samples from the population and draw conclusions about the population based on the sample results.

Statistical Inference

The branch of Statistics which use the samples to make inferences about the unknown aspects of the population is known as **Statistical Inference** or **Inferential Statistics**. The unknown aspects of the population may be the value of the parameter, the form of distribution of the population, etc. Inferential statistics basically relates the sample characteristics to population characteristics. Statistical inference is classified into two categories. One is the **Estimation of Parameters**, in which we estimate various unknown parameters of the population and the other is the **Testing of Hypothesis,** in which either we accept or reject the estimated value of the parameter of the population using samples taken from it. In this chapter we discuss some basic concepts of the theory of estimation. Testing of hypothesis will be discussed in the next chapter.

Estimator and Estimate

The value of the parameter is usually estimated with the help of a function of the sample. This function is known as ***Estimator*** of the parameter. Clearly estimator is a statistic and hence it is a random variable. We can see that the value of the estimator will vary from sample to sample. The value obtained for an estimator from a particular sample is called ***Estimate*** of the parameter.

For example, suppose we want to estimate the population mean (μ) of the weights of students in higher secondary schools of Kerala. We can select a sample of 5000 students and find the sample mean ($\bar{x}$). Let it be 54Kg. Then the samples mean ($\bar{x}$) is the estimator and the sample mean obtained from the sample, 54Kg, is the estimate of the population mean.

Estimation of parameter is of two types.

- Point Estimation
- Interval Estimation

If we are selecting a sample from the population and suggesting a single value corresponding to the parameter, which is estimated from the sample, then it is called **Point Estimation.** Instead of suggesting a single value, one can propose an interval, within which the value of the parameter is expected. This is known as **Interval Estimation.**

The characteristics and methods of estimation are discussed below in detail.

8.1 Point Estimation

Many functions of sample observations may be proposed as estimators of the same parameter. For example, we can use the sample mean, sample median, sample mode, etc. as the estimators of the population mean. One can use the average of the minimum and maximum observations of the sample also. In short there will be many estimators for a single parameter. Usually a statistician will choose a best estimator corresponding to a parameter. Some criterions are used to examine the good estimators. An estimator is said to be a good estimator if it is:

- Unbiased
- Consistent
- Efficient
- Sufficient

Unbiased Estimator

Every estimator is a statistic and all the statistics are random variables. Therefore the value of the estimator will have different values for different samples. If the arithmetic mean of the values of a statistic from all the possible samples is equal to the true value of the parameter, then we say it is an unbiased estimator. In other words an estimator is said to be unbiased if its expected value is equal to the true value of the parameter. Several unbiased estimators may exist for a parameter.

> An estimator, 't' is said to be an unbiased estimator for the parameter 'θ', if $E(t) = \theta$.

In the last chapter we have discussed that the sample mean is a statistic with mean μ and standard deviation σ. It means that mean of sample means is equal to population mean. Or the expected value of sample mean is the population mean. That is $E(\bar{x}) = \mu$. Therefore sample mean $\bar{x}$ is an unbiased estimator for the population mean μ.

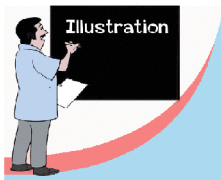


Illustration 8.1

X_1, X_2, X_3 and X_4 is a random sample drawn from a population with mean μ. Show that $T = X_1 + X_2 + X_3 - 2X_4$ is an unbiased estimator for μ.

Solution

Given $E(X_1) = E(X_2) = E(X_3) = E(X_4) = \mu$

Now $E(T) = E(X_1 + X_2 + X_3 - 2X_4)$

$$= E(X_1) + E(X_2) + E(X_3) - 2E(X_4)$$

$$= \mu + \mu + \mu - 2\mu = \mu$$

i.e. $E(T) = \mu$

Therefore T is an unbiased estimator of μ .

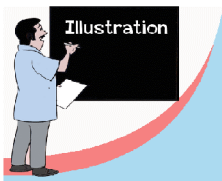


Illustration 8.2

Using simple random sampling without replacement (SRSWOR), select samples of size two from the population consisits of 4, 5, 6, 7 and 8.

(a) Write all possible samples.

(b) Show that sample mean is an unbiased estimate of population means.

Solution

No. of possible samples in SRSWR $=5C_2=10$

Sample No	samples	sample mean ($\bar{X}$)
1	(4,5)	4.5
2	(4,6)	5
3	(4,7)	5.5
4	(4,8)	6
5	(5,6)	5.5
6	(5,7)	6
7	(5,8)	6.5
8	(6,7)	6.5
9	(6,8)	7
10	(7,8)	7.5
	Total	**60**

$$E(\overline{X}) = \frac{60}{10} = 6 \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (1)$$

Population mean, $\mu = \frac{4+5+6+7+8}{10} = 6$ ..... (2)

from (1) and (2) we have $E(\overline{X}) = \mu$

Hence sample, $\overline{X}$ is an unbiased estimator of population mean, μ.

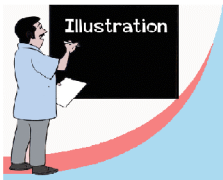


Illustration 8.3

Using simple random sampling with replacement (SRSWR), select samples of size two from the popuation values 4, 6, and 8.

(a) Write all possible samples.

(b) Show that sample mean is an unbiased estimate of population means.

No. of possible samples in SRSWR = $3^2 = 9$.

Sample No	samples	sample mean ($\overline{X}$)
1	(4,4)	4
2	(4,6)	5
3	(4,8)	6
4	(6,4)	5
5	(6,6)	6
6	(6,8)	7
7	(8,4)	6
8	(8,6)	7
9	(8,8)	8
	Total	**54**

$$E(\overline{X}) = \frac{54}{9} = 6 \qquad \text{-- -- -- -- -- -- -- -- (1)}$$

Population mean, $\mu = \frac{4+6+8}{3} = 6$ -- -- -- -- -- -- -- -- (2)

from (1) and (2) we have $E(\overline{X}) = \mu$

Hence sample, $\overline{X}$ is an unbiased estimator of population mean, μ .

Know your progress

1) Let X_1, X_2, and X_3 be sample values drawn from a normal population with mean μ and standard deviation σ. Consider the following estimators for the parameter μ.

$$T_1 = 4X_1 + 3X_2 - 5X_3 \quad T_2 = \frac{X_1 + X_3}{2}, \quad T_3 = X_1 + X_2 - X_3$$

Check the unbiasedness of the estimators T_1, T_2 and T_3.

2) A population consists of values 2, 4, 6, 10. Consider all possible random samples of size 2 in SRSWOR and SRSWR. Show that sample mean is an unbiased estimator of population mean.

Consistent Estimator

A desirable property of a good estimator is that its accuracy should increase when the sample size increases. That is the value of the estimator is expected to come closer to the true value of the parameter and its variance becomes 0, when the sample size is sufficiently large.

An estimator, 't' is said to be a consistent estimator for the parameter 'θ', if $E(t) \to \theta$ and $V(t) \to 0$ as $n \to \infty$

Efficient Estorimator

Efficiency of an estimator is related to its variance. If a parameter has many unbiased estimators, then that estimator with least variance is called the efficient estimator.

Let t_1 and t_2 are two unbiased estimators for the parameter 'θ', then t_1 is said to be more efficient than t_2, if $V(t_1) < V(t_2)$

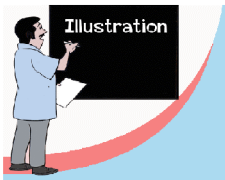


Illustration 8.4

X_1, X_2, X_3 and X_4 is a random sample drawn from a population with mean μ and standard deviation σ.

$T_1 = X_1 + 2X_2 + 2X_3 - 4X_4$ and $T_2 = X_1 + X_2 - X_3 - X_4$ are unbiased estimators for μ. Obtain the efficient estimator among T_1 and T_2.

Solution

Given that T_1 and T_2 are unbiased estimators of the population mean μ.

$$T_1 = X_1 + 2X_2 + 2X_3 - 4X_4$$

$$V(T_1) = V(X_1 + 2X_2 + 2X_3 - 4X_4)$$

$$= V(X_1) + V(2X_2) + V(2X_3) + V(4X_4) \quad [\text{Since } V(X \pm Y) = V(X) + V(Y)]$$

$$= V(X_1) + 4V(X_2) + 4V(X_3) + 16V(X_4) \quad [\text{Since } V(aX) = a^2V(X)]$$

$$= \sigma^2 + 4\sigma^2 + 4\sigma^2 + 16\sigma^2 = 25\sigma^2$$

$$T_2 = X_1 + X_2 - X_3 - X_4$$

$$V(T_2) = V(X_1) + V(X_2) + V(X_3) + V(X_4)$$

$$= \sigma^2 + \sigma^2 + \sigma^2 + \sigma^2$$

$$= 4\sigma^2$$

$V(T_2) < V(T_1)$. Therefore T_2 is efficient than T_1.

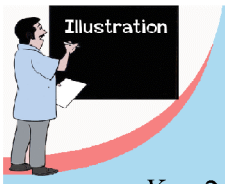


Illustration 8.5

X_1, X_2 and X_3 are values of a random sample drawn from a normal population with mean μ and standard deviation σ.

$T_1 = \dfrac{X_1 + 2X_2 + X_3}{4}$, $T_2 = \dfrac{X_1 - X_2 + X_3}{2}$ and $T_3 = \dfrac{X_1 + X_2 + X_3}{3}$ are estimators for μ. Obtain the efficient estimator.

Solution

Given , $T_1 = \frac{X_1 + 2X_2 + X_3}{4}$, $T_2 = \frac{X_1 - X_2 + X_3}{2}$ and $T_3 = \frac{X_1 + X_2 + X_3}{3}$

$$E(T_1) = E\left(\frac{X_1 + 2X_2 + X_3}{4}\right)$$

$$= \frac{E(X_1) + 2E(X_2) + E(X_3)}{4}$$

$$= \frac{\mu + 2\mu + \mu}{4}$$

$$= \frac{4\mu}{4}$$

$$= \mu$$

$i.e. E(T_1) = \mu$. Therefore T_1 is an unbiased estimator of μ .

$$E(T_2) = E\left(\frac{X_1 - X_2 + X_3}{2}\right)$$

$$= \frac{E(X_1) - E(X_2) + E(X_3)}{2}$$

$$= \frac{\mu - \mu + \mu}{2}$$

$$= \frac{\mu}{2}$$

i. e. $E(T_2)$ is not equal to μ . Therefore T_2 is not an unbiased estimator of μ .

$$E(T_3) = E\left(\frac{X_1 + X_2 + X_3}{3}\right)$$

$$= \frac{E(X_1) + E(X_2) + E(X_3)}{3}$$

$$= \frac{\mu + \mu + \mu}{3}$$

$$= \frac{3\mu}{3}$$

$$= \mu$$

i.e, $E(T_3) = \mu$. Therefore T_3 is an unbiased estimator of μ.

$$V(T_1) = V\left(\frac{X_1 + 2X_2 + X_3}{4}\right)$$

$$= \frac{V(X_1) + 4V(X_2) + V(X_3)}{16}$$

$$= \frac{\sigma^2 + 4\sigma^2 + \sigma^2}{16}$$

$$= \frac{6\sigma^2}{16}$$

$$= \frac{3\sigma^2}{8} = 0.375\sigma^2$$

$$V(T_3) = V\left(\frac{X_1 + X_2 + X_3}{3}\right)$$

$$= \frac{V(X_1) + V(X_2) + V(X_3)}{9}$$

$$= \frac{\sigma^2 + \sigma^2 + \sigma^2}{9}$$

$$= \frac{3\sigma^2}{9}$$

$$= \frac{\sigma^2}{3} = 0.333\sigma^2$$

$V(T_3) < V(T_1)$, Therefore T_3 is the efficient estimator.

Know your progress

If X_1, X_2, X_3, X_4 and X_5 are sample values drawn from a normal population with mean μ and standard deviation σ.

$$T_1 = X_1 + 2X_3 - X_4 - X_5,\ T_2 = \frac{X_1 + X_2 + X_3 + X_4}{4} \text{ and}$$

$$T_3 = \frac{X_1 + X_2 + X_3 + X_4 + X_5}{5}$$

are unbiased estimators for the parameter μ. Find the most efficient estimator.

Sufficient Estimator

Estimators are substitutes for the parameter obtained from the sample. A good estimator should provide the maximum information contained in the sample regarding the parameter. If an estimator is capable to provide all the information contained in the sample about the parameter, then it is sufficient to substitute the parameter and hence it is known as sufficient estimator.

> An estimator is said to be a sufficient estimator of a parameter, if it contains all information in the sample about the parameter.

Note:-

Sample mean is unbiased, consistent and sufficient for the population mean.

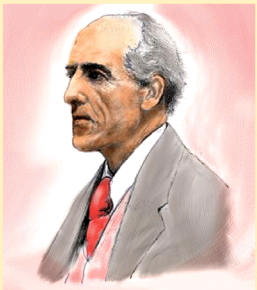
8.2 Method of Moments

We have discussed the concept of moments in our plus one classes. There are two types of moments, raw moments and central moments.

The rth order raw moment of the sample is given by

$$m'_r = \frac{\sum x^r}{n}$$

The rth order raw moment of the population is usually represented by μ'_r and the rth order



Method of moments was discovered and studied in detail by **Karl Pearson**

raw moment of the sample is usually represented by m'_r.

By equating the population moments and sample moments, we get

$$\mu'_r = m'_r$$

Where $\mu'_r = E(X^r)$

In particular if $\mu'_1 = m'_1$,

we have $E(X) = \bar{x}$ (Since $m'_1 = \bar{x}$)

or $\mu = \bar{x}$

That means the moment estimator for the population mean (μ) is the sample mean ($\bar{x}$). We denote it by

$$\hat{\mu} = \bar{x}$$

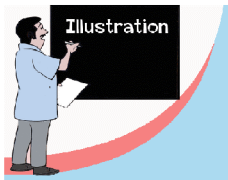


Illustration 8.6

A sample drawn from a population is given below. Obtain the moment estimator for the population mean.

Sample Values: 18, 14, 14, 17, 15, 13, 16, 15, 16, 19, 15, 17, 19, 15 and 17

Solution

$$\hat{\mu} = \bar{x}$$

$$= \frac{\sum x}{n}$$

$$= \frac{1}{15}(18 + 14 + \ldots + 17)$$

$$= \frac{240}{15}$$

$$\hat{\mu} = 16$$

A sample is drawn from a population to analyse the mean

Know your progress

value of the population. The sample values are given below.

45, 40, 42, 46, 48, 43, 58, 52, 43, 44

Obtain the moment estimate for the population mean.

8.3 Interval Estimation

In point estimation we estimate a single quantity for a parameter. This estimated value cannot be expected to coincide with the true value of the population parameter. It can be used as an approximation only. One can expect the actual population value around the estimated value. Sometimes the actual population value may be very close to the estimated value or sometimes very far from the estimated value. So it will be more useful if one can propose an interval which is expected to include the unknown parameter with a specified probability. We can suggest a large interval with more confidence and the confidence to expect the true value within the suggested interval will decrease when the interval becomes small. Such an interval is known as confidence interval or $(1-\alpha) \times 100\%$ confidence interval. Here $(1-\alpha)$ is known as the confidence coefficient, which is the probability that actual value of the parameter to being included in the interval.

8.4 Confidence interval for the population mean

Let a sample of size n be drawn from a normal population with standard deviation σ. Let $\bar{x}$ be the sample mean and s is the sample standard deviation, then the $(1-\alpha) \times 100\%$ confidence interval for the population mean is given by,

$$\left(\bar{x} - \frac{\sigma}{\sqrt{n}} Z_{\alpha/2},\ \bar{x} + \frac{\sigma}{\sqrt{n}} Z_{\alpha/2}\right)$$

The value of $Z_{\alpha/2}$ is determined from standard normal table for different values of α such that $P(-Z_{\alpha/2} < Z < Z_{\alpha/2}) = 1-\alpha$, where Z is the standard normal variable.

When 'σ' is unknown and if the sample size 'n' is large, the confidence interval is

$$\left(\bar{x} - \frac{s}{\sqrt{n}} Z_{\alpha/2},\ \bar{x} + \frac{s}{\sqrt{n}} Z_{\alpha/2}\right)$$

Note: For a 95% confidence interval, the value of $Z_{\alpha/2}$ =1.96 and for a 99% confidence interval $Z_{\alpha/2} = 2.58$

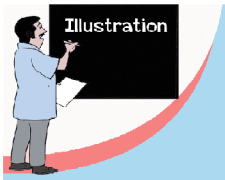


Illustration 8.7

After a wet night, 16 worms were found on the lawn. Their average length measured in cm. was 10.39. Assuming that this sample came from a normal population with variance 4. Calculate the 99% confidence interval mean length of all the worms in the garden.

Solution

Given that

$\bar{x}$ =10.39 , $\sigma = \sqrt{4}$ = 2 and n = 16.

The 99% confidence interval for the population mean is

$$\left(\bar{x} - 2.58 \times \frac{\sigma}{\sqrt{n}},\ \bar{x} + 2.58 \times \frac{\sigma}{\sqrt{n}}\right)$$

$$= \left(10.39 - 2.58 \times \frac{2}{\sqrt{16}},\ 10.39 + 2.58 \times \frac{2}{\sqrt{16}}\right)$$

$$= (9.1,\ 11.68)$$

The 99% confidence interval for the mean length of the worms in the garden is (9.1, 11.68).

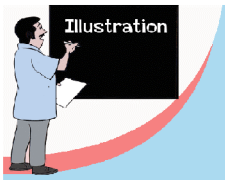


Illustration 8.8

On the basis of the results obtained from a random sample of size 100 men from a particular district, the mean height of the men in the district is found to be 178.2 cm. with a standard deviation 4.8 cm. Calculate a 95% confidence interval of the population mean. Suppose that the population of heights of the men in the district follows a normal distribution.

Solution

We are given the following information about the sample,

$\bar{x}$ =178.2 , s = 4.8 and n = 100.

The 95% confidence interval for the mean of a normal population is,

$$\left(\bar{x}-1.96\times\frac{s}{\sqrt{n}},\ \bar{x}+1.96\times\frac{s}{\sqrt{n}}\right)$$

$$=\left(178.2-1.96\times\frac{4.8}{\sqrt{100}},\ 178.2+1.96\times\frac{4.8}{\sqrt{100}}\right)$$

$$=\ (177.26,\ \ 179.14)$$

The 95% confidence interval for the mean is (17726, 17914)

Know your progress

1. A random sample of 100 is taken from a population with standard deviation 12. The sample mean is 76. Find a 95% confidence interval for the mean of the population.
2. 80 people were asked to measure their pulse rates when they woke up in the morning. The mean was 69 beats and the standard deviation 4 beats. Find a 99% confidence interval for the population mean.

Let us conclude

Statistical inference is the branch of Statistics, deals with the determination of the unknown aspects of a population using samples taken from the population. The unknown aspects may be the parameter of the population, form of distribution of the population, etc. Statistical inference is classified into two categories – estimation of parameters and testing of hypothesis. Estimation of parameters deals with the determination of the value of the unknown parameter of the population using samples taken from it. There are two types of estimation – point estimation and interval estimation. In point estimation we suggest a statistic as the value of the parameter. This statistic is called a point estimate of

the population. We can suggest a lot of statistic as the estimate of a single parameter. So we have to consider certain criteria for selecting a good estimate. The desirable properties of a good estimate are unbiasedness, consistency, efficiency and sufficiency. They are detailed in this chapter. There are different methods for estimating the parameter of a population. One of them is the method of moments. In interval estimation we suggest an interval as an estimate such that this interval contains the true value of the parameter with a specified probability. This probability is called the confidence coefficient and the interval is called the confidence interval. The confidence interval for the mean of a normal population is also discussed in this chapter.

Let us assess

For Questions 1-3, choose the correct answer from the given choices.

1. If X_1, X_2 and X_3 is a random sample of size 3 taken from a population with mean μ and if $T_1 = \dfrac{X_1 + 2X_2 + X_3}{k}$ is unbiased for μ, what is the value of k?

   a) 4 b) 2 c) $\frac{1}{2}$ d) $\frac{1}{4}$

2. Name the property of an estimator which is based on its variance when n is large.

   a) Unbiasedness b) consistency c) sufficiency d) efficiency

3. The estimator t_1 is more efficient than t_2 when:

   a) $V(t_1) = V(t_2)$ b) $V(t_1) > V(t_2)$ c) $V(t_1) < V(t_2)$ d) $V(t_1) = V(t_2) = 0$

4. A teacher asked a group of students about the average time taken by them to reach school. Then, some of them replied – 'about 25 minutes' and some others replied – '20 to 30 minutes'. Name the types of estimation related to each of the above replies.

5. If X_1, X_2 and X_3 is a random sample taken from a population with mean μ and standard deviation σ. Find which of the following estimators for μ are unbiased,

   $$U_1 = \frac{1}{4}X_1 + \frac{1}{2}X_2 + \frac{1}{4}X_3$$

   $$U_2 = \frac{4}{5}X_1 + \frac{1}{10}X_2 + \frac{1}{10}X_3, U_3 = \frac{1}{6}X_1 + \frac{2}{3}X_2 + \frac{1}{2}X_3$$

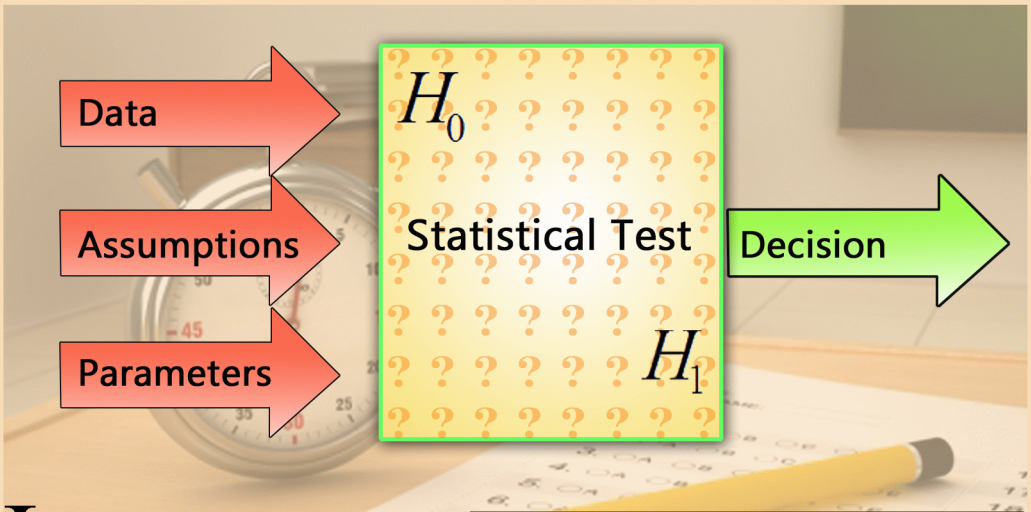
6. If X_1, X_2 and X_3 is a random sample taken from a population with mean μ and standard deviation σ. Find which of the following estimators for μ are unbiased, and which is most efficient.

$$T_1 = \frac{X_1 + X_2 + X_3}{3}, T_2 = \frac{X_1 + 2X_2}{3}, T_3 = \frac{X_1 + 2X_2 + 3X_3}{3}$$

7. Find the estimate of the population mean from which each of the following samples is drawn:
   a) 46, 48, 51, 50, 45, 53, 50, 48.
   b) 35, 42, 38, 55, 70, 69.
   c) 1.684, 1.691, 1.687, 1.688, 1.689, 1.688, 1.690, 1.693, 1.685.
8. A certain type of tennis ball is known to have a height of bounce which is normally distributed with standard deviation 2 cm. A sample of tennis balls is tested and the mean height of bounce of the sample is 140 cm. Find the 95% confidence interval for the mean height.
9. A random sample of 100 is taken from a population. The sample is found to have a mean of 76 and standard deviation 12. Obtain a 99% confidence interval for the mean of the population.
10. 150 bags of flour of a particular brand are weighted and the mean mass is found to be 748 grams with standard deviation 3.6 gms. Find 95% and 99% confidence interval for the mean mass of bags flour of this brand.
11. 80 people were asked to measure their pulse rates when they woke up in the morning. The mean was 69 beats and the standard deviation 4 beats. Find 95% confidence interval and 99% confidence interval of the population mean.
12. Distinguish between estimator and estimate.
13. Suggest 3 point estimates for the mean of a population.
14. The ages of Badminton players of a school are 18,15,16,19. If a team of 2 may be selected according to SRSWOR. Verify the statement "sample mean is an unbiased estimator of population mean".
15. In a random sample of 64 of 600 road crossing in a town, the mean number of automobile accidents per year was found to 4.2 and the sample sd was 0.8. Construct a 95% confidence interval for the mean number of automobile accidents per crossing per year.

Chapter 9

Testing of Hypothesis



In the previous chapter we have discussed how a sample can be used to develop point and interval estimates for assessing population parameters. In Statistics, we know that the inference made is based on estimation and hypothesis testing. In this chapter we will continue our discussion of statistical inference in terms of hypothesis testing. Hypothesis testing is the soul of inferential statistics and is a very important tool for researchers to arrive at a conclusion. It is one of the most important aspects of the theory of

Significant Learning Outcomes

After the completion of this chapter, the learner:

- Recognises and forms statistical hypotheses.
- Identifies Type I and Type II Errors.
- Explains Test statistic, Critical region, Level of significance and Power of a test.
- Illustrates the methods for testing the population means.
- Illustrates the Chi-square test for independence of attributes.
- Constructs suitable statistical test for a situation.

decision making. Testing of hypothesis plays an important role in Industry, Biological sciences, Social sciences, Economics, etc. The purpose of hypothesis testing is to determine whether there is enough statistical evidence in favour of a certain belief or hypothesis about a parameter.

9.1 Statistical Hypothesis

A statistical hypothesis is generally considered as an assumption about the distribution or parameter of a population that has to be proved or disproved. For example, the operation manager of a soft drink manufacturing company has to decide whether the bottling operation is under statistical control or not. The company sells in bottles labeled 1 litre, filled by an automatic bottling machine. The manager wants to examine the claim that on average each bottle contains 1 litre of the soft drink may or may not be true. Here our hypothesis for the bottling process is, "the average quantity of soft drink in the bottle is equal to 1 litre". If the hypothesis is true, the process is said to be under statistical control. If the hypothesis is not true, ie, the average is either less than 1 litre or more than 1 litre, then the process is said to be out of control.

A **statistical hypothesis** is an assumption about an unknown population parameter or distribution. This assumption may or may not be true.

Null and Alternative Hypothesis

The first step of hypothesis testing is to convert the research questions into hypotheses. In any testing of hypothesis problem, we are faced with a pair of hypotheses such that one and only one of them is always true. One of this pair is called the null hypothesis and the other one is the alternative hypothesis.

The hypothesis actually to be tested is referred to as null hypothesis, denoted by H_0. Null hypothesis is the claim of 'no difference'. It is assumed to be true unless there is strong statistical evidence to reject it.

The alternative hypothesis is the negation of the null hypothesis, denoted by H_1. The alternative hypothesis is claim of 'a difference in the population'. It is assumed to be true when the null hypothesis is false.

For example, in the bottling plant problem discussed above, the null and alternative hypotheses are,

H_0 : The average quantity of soft drink in the bottle is equal to 1 litre.

H_1 : The average quantity of soft drink in the bottle is not equal to 1 litre.

Consider another example;

Suppose we want to determine whether a coin is unbiased.

Here, the null hypothesis (H_0) is - half of the flips of the coin will result in Heads and half in Tails.

Symbolically, H_0: $p = \frac{1}{2}$ where p is the proportion of number of heads to total trial.

The alternative hypothesis (H_1) is - the number of Heads and Tails will be different.

Symbolically, H_1: $p \neq \frac{1}{2}$, where p is the proportion of number of Heads.

As we have discussed, the null and alternative hypotheses are opposite statements. So both H_0 and H_1 can not be true and one of them will always be true. Hence rejecting one is equivalent to accepting the other and vice versa. It is to be noted that the roles of null and alternative hypotheses are not symmetrical. That is, they can not be interchanged.

The **Null Hypothesis,** symbolized by H_0, is the hypothesis which is actually to be tested. It is the hypothesis which states that there is no difference between a parameter and a specified value of parameter.

The **Alternative hypothesis,** denoted by H_1, is the negation of the null hypothesis. It is the hypothesis which states that there exist a difference between the parameter and the specified value of the parameter.

For a better understanding about the roles of null and alternative hypotheses, we can consider a non statistical example.

A person who has been accused for committing a crime is on trial in a court. The person presented before the court is assumed to be innocent until he/she proves guilty. Hence in the beginning of the trial it is assumed that the person is innocent. The null hypothesis is usually the hypothesis that is assumed to be true to begin with. Using the language of hypothesis testing, the two hypotheses for this court case can be written as follows:

H_0: The person is innocent.

H_1: The person is guilty.

The outcomes of the trial process may result:

- Accepting H_0 of innocence, when there is not enough evidence to convict the person. It does not prove that the person is truly innocent.
- Rejecting H_0 and accept H_1 of guilty, when there is enough evidence to rule out innocence and to strongly establish the guilt.

In a trial case we do not have to rule out guilt in order to find someone innocent, but we do have to rule out innocence in order to find someone guilty. ie, we do not have to rule out H_1 in order to accept H_0, but we do have to rule out H_0 in order to accept H_1. Thus it is clear that the two hypotheses – null and alternative are not interchangable. So it becomes more important to be clear about what the null and alternative hypothesis should be in a given situation, or else the test is meaningless.

9.2 The Two Types of Errors

After the null and alternative hypotheses are spell out, the next step is to gather evidence from a random sampling of the population. The most important limitation of making inferences from the sample data is that we cannot be 100% confident about it. If the sample is as large as the population the variation will be negligible. But in normal case variation from one sample to another can never be eliminated. This may lead to incorrect drawing of conclusions or errors. In the hypothesis testing, there are four possible outcomes. In reality the null hypothesis may or may not be true and a decision is made to reject or accept it on the basis of the data obtained from a sample. The following table shows the four possible outcomes. Note that there are two possibilities for a correct decision and two possibilities for an incorrect decision.

		States of population	
		H_0 is true	H_0 is false
Decision based on sample	Reject H_0	Error (Type I)	Correct decision (No error)
	Accept H_0 (do not reject H_0)	Correct decision (No error)	Error (Type II)

Table 9.1

The table shows the possible decision making and the states of population. As you can see, there are four possible outcomes, which can be classified as three,

- No error (Correct decision)
- Type I error
- Type II error.

Only one of these outcomes will occur for a hypothesis testing.

Type I Error

In the context of statistical testing, the wrong decision of rejecting a true null hypothesis is called Type I error. ie, Type I error is to reject H_0 when H_0 is true.

Type II Error

The wrong decision of accepting (not rejecting) a false null hypothesis is known as Type II error. ie, type II error is to accept H_0 (not reject H_0) when H_0 is true (H_1 is false).

Both types of errors are undesirable and should be reduced to the minimum. We can not completely avoid either type of errors.

Many Statisticians argue that you should never use the phrase **'accept the null hypothesis'**. Instead you should use, **'do not reject the null hypothesis'**. Thus the only two hypothesis testing decision would be **reject H_0** or **do not reject H_0**.

A **Type I Error** occurs if you **reject the null hypothesis when it is true.**

A **Type II Error** occurs if you **do not reject the null hypothesis when it is false.**

9.3 Level of Significance and Power of a Test

In testing a given hypothesis, the maximum probability which we would be willing to take risk is called level of significance or significance level of the test. In other words significance level is the probability of rejecting H_0 when H_0 is true. The level of significance is denoted by the Greek letter α (alpha). The significance level of a test is generally specified before samples are drawn for the testing purpose.

$$\alpha = P(\text{reject } H_0 / H_0 \text{ is true})$$

$$= P(\text{Type I Error})$$

The level of significance commonly used in testing of hypothesis is 0.05 or 0.01. That is, if the null hypothesis is rejected, the probability of type I error will be 5% or 1%, depending on which level of significance is used. When $\alpha = 0.05$, there is a 5 % chance of rejecting a true null hypothesis. In other words we are about 95% confident that we made the right decision. In such a case we say that the hypothesis has been rejected at 5% level of significance which means that we could be wrong with probability 0.05.

The probability of type II error is denoted by the Greek letter β (Beta).

$$\beta = P(\text{type II error})$$
$$= P(\text{Accept } H_0/H_0 \text{ is false})$$

Type II error is committed when a wrong decision is taken in accepting a false null hypothesis.

The Power of a test is the probability of rejecting a null hypothesis when it is false.

$$\text{Power} = P(\text{Reject } H_0/H_0 \text{ is false})$$
$$= 1 - P(\text{Accept } H_0/H_0 \text{ is false})$$
$$= 1 - \beta$$

That is, power of a test is the probability of not committing a type II error. It is the probability that the test will correctly lead to the rejection of a false null hypothesis. The statistical power is the ability of a test to detect an effect, if the effect actually exists.

The **level of significance** is the maximum probability of committing a Type I error. This is denoted by the Greek letter α.

ie, $\alpha = P(\text{type I error})$

The **power of a test** is the probability of rejecting H_0 when it is false. It is the probability of not committing a type II error.

ie, *Power of the test* $= P(\text{Reject } H_0/ H_0 \text{ is false})$

$= 1 - P(\text{Accept } H_0/H_0 \text{ is false})$

$= 1 - P(\text{Type II Error}) = 1 - \beta$

9.4 Test Statistic and Critical Region

Test Statistic

As we know, a statistic is a function of sample values. Also, a statistical test is conducted using a sample taken from the population. Hence we use the help of a statistic in the testing procedure. This statistic is called the test statistic. It is based on appropriate

probability distribution. With the help of the test statistic we determine whether to accept or reject the null hypothesis. The test statistic compares data with what is expected under the null hypothesis. The commonly used test statistic are Z – statistic, t – statistic, F – statistic and Chi-square statistic.

Critical Region

The test statistic follows some probability distribution.

In a test, the area under the probability curve of the test statistic is divided into two regions,

- the region of rejection
- the region of acceptance

The region of rejection:- The rejection region is the region in which the null hypothesis is rejected. The region of rejection is called the critical region. That is, the critical region is the area or areas of the sampling distribution of the test statistic that will lead to the rejection of the hypothesis tested. The area of the critical region is equal to the level of significance, α.

The region of acceptance:- The acceptance region is the area in which we take decision to accept the null hypothesis. The value which separates the critical region and acceptance region is called the critical value.

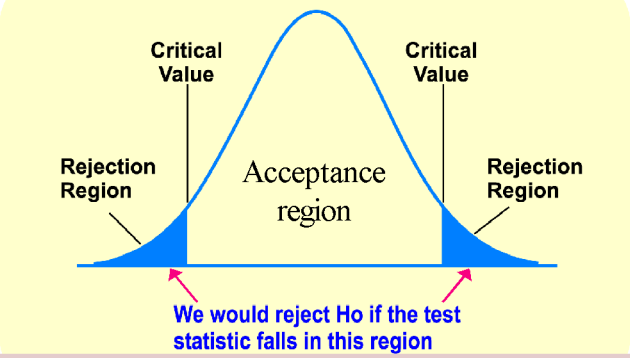


Fig. 9.1

The **Critical region** is the range of values of the test statistic which indicates that there is significant difference and the null hypothesis should be rejected. The value which separates the critical region and acceptance region is called the **Critical value.**

9.5 One - Tailed and Two - Tailed Tests

Let us consider a null hypothesis and alternative hypothesis for testing the mean (μ) of a population as follows,

$$H_0: \ \mu = \mu_0 \ \text{ and } \ H_1: \ \mu < \mu_0$$

In this case we will reject H_0 only when the test statistic is significantly less than μ_0. Thus rejection occurs only when the test statistic is taken a significantly low value in the left tail of its distribution. This is called a left tailed test.

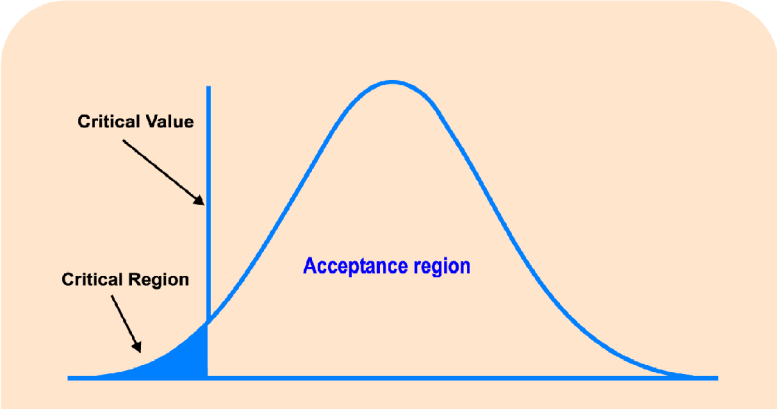


Fig. 9.2 - Left tailed test

Now consider the following null and alternative hypotheses

$$\mathbf{H_0}: \ \mu = \mu_0 \ \text{ and } \ H_1: \ \mu > \mu_0$$

In this case we will reject H_0 only when the value of the test statistic is significantly more than μ_0 . Thus the rejection occurs only when the test statistic is taken a significantly high value in the right tail of the distribution. This is called a right tailed test.

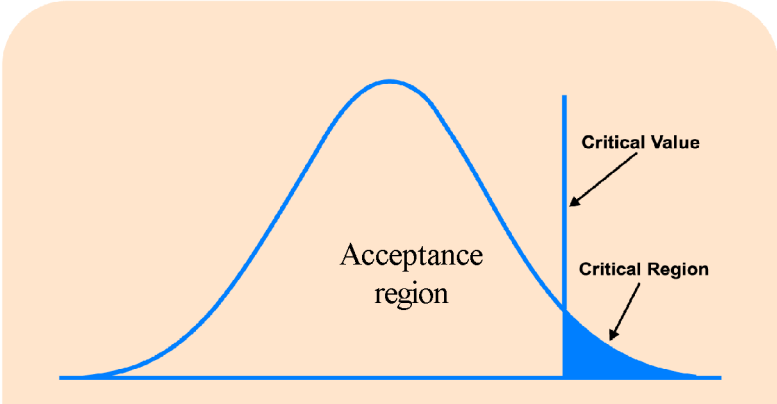


Fig. 9.3 - Right tailed test

In left tailed and right tailed tests, the rejection occurs only on one tail. Hence each of them is called a **one tailed test**.

Finally consider the following hypotheses,

$$\mathbf{H_0}: \ \mu = \mu_0 \text{ and } H_1: \ \mu \neq \mu_0$$

In this case we have to reject H_0 in both cases, whether the value of the test statistic significantly less than or greater than μ_0. Thus the rejection occurs on both tails. Therefore this case is called a two tailed test.

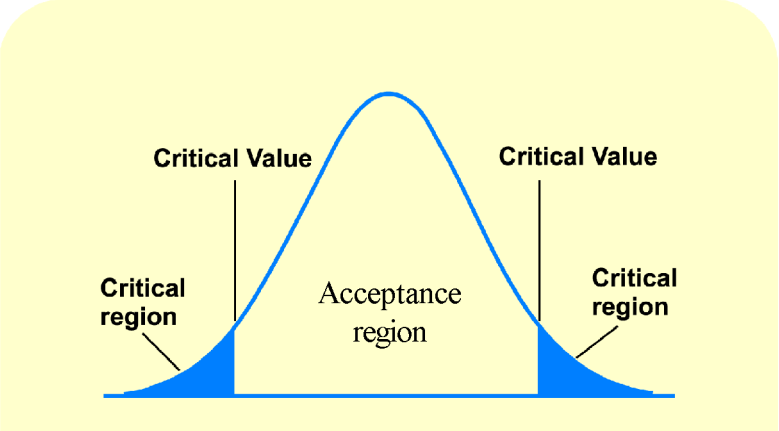


Fig. 9.4 - Two tailed test

A **One – tailed test** indicates that the null hypothesis is rejected when the value of the test statistic is in the critical region on one side of the critical value. A one tailed test is either a right tailed test or a left tailed test.

A **Two– tailed test** indicates that the null hypothesis is rejected when the value of the test statistic is in either of the two critical regions.

The five steps to hypothesis testing

We have already discussed some basic concepts concerning the testing of hypothesis. A sample is selected for estimating the population parameter. Sample statistic is computed from this sample and is used to estimate the population parameter. A systematic procedure needs to be adopted for the testing of the hypothesis concerning the estimated value of the parameter. Here is a description of the five steps of a statistical hypothesis testing.

Step 1: State the hypotheses.

Step 2: Determine the appropriate test statistic.

Step 3: Set the level of significance and the critical region.

Step 4: Compute the test statistic using the sample data provided.

Step 5: Make a decision.

Step 1: State the hypotheses

We begin the test procedure by stating the null and alternative hypotheses. A null hypothesis (H_0) is a hypothesis which is tested for a possible rejection under the assumption that it is true. Theoretically a null hypothesis is set as no difference. It is to be noted that the only reason we are testing the null hypothesis is because we think it is wrong. What is wrong about the null hypothesis is called an alternative hypothesis (H_1). That is, alternative hypothesis is the negation of null hypothesis.

Step 2: Determine the appropriate test statistic

After setting the hypotheses, we have to decide on an appropriate statistical test that will be used for statistical analysis. The statistic used for the analysis is the test statistic. The success of the test depends on the selection of the best test statistic.

Step 3:Set the level of significance and the critical region

To set the criteria for decision making, we state the level of significance (α) for a test. Usually it is taken as 0.05 or 0.01. It is very important that the level of significance must be determined before we draw samples, so that the obtained result is free from the choice bias of a decision maker. Then we establish a critical region. The critical region is the area under which the value of the test statistic falls; here we take decision to reject H_0.

Step 4: Compute the test statistic

In this stage, sample data are collected and the value of the test statistic is calculated.

Step 5: Make a decision

The decision is made using the value of the test statistic. If the test statistic falls in the critical region, we take decision to reject H_0. If the test statistic falls in the acceptance region, we take decision not to reject H_0.

Know your progress

1. What is a statistical hypothesis?
2. Give an example of a null and an alternative hypothesis.
3. Explain the two types of errors occurring in testing a statistical hypothesis.
4. Explain the terms-Significance level, critical region and power of a test.
5. Briefly explain the steps in a statistical test procedure.
6. Mention the difference between one tailed and two tailed tests.

9.6 Tests of significance of single mean

Consider a population having mean μ and standard deviation σ. When the test is about the population mean, the test can either be Z – test or t – test.

Step 1: Formulate the hypotheses

The first step in a statistical testing of hypothesis is to formulate the hypotheses. Let μ_0 denotes the claimed population mean. Then the null hypothesis is,

$$H_0: \ \mu = \mu_0$$

The alternative hypothesis is any one of:

$$H_1: \ \mu \neq \mu_0 \ \text{(two – tailed test)}$$

$$H_1: \ \mu > \mu_0 \ \text{(one – tailed test)}$$

$$H_1: \ \mu < \mu_0 \ \text{(one – tailed test)}$$

Step 2: Determine appropriate test statistic

For testing the mean of a population, we use Z – test or t – test under various conditions.

(a) Z – test

The test is known as Z – test, since we use the Z test statistic. The Z test statistic follows a standard normal distribution. ie, a normal distribution with mean 0 and standard deviation 1.

The following are the cases in which we use the Z – test for testing a single population mean.

Case 1: The population standard deviation is known and the population follows normal distribution.

Case 2: The population standard deviation is known and the sample size n is greater than 30 (large sample). The population need not be normal.

Case 3: The population standard deviation is unknown, the population is normal and the sample size n is large (n is greater than 30).

Let $\bar{x}$ be the mean and s the standard deviation of a sample of size n taken from the population. The Z – test statistic for testing the hypothesis of a single population mean is described below.

In **case 1** and **case 2**, the Z – test statistic is,

$$Z = \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$$

Where Z follows N(0, 1).

In **case 3**, the Z – test statistic is,

$$Z = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}} \quad \text{follows N(0, 1).}$$

Step 3: Deciding the level of significance and critical region

After deciding the appropriate test statistic, the level of significance, α is to be fixed. The commonly used level of significance is either 0.05 or 0.01.

Now we decide the critical region. The critical region is defined in terms of the test statistic. The following table shows the critical region in various situations.

Test	Alternative hypothesis	Critical Rigion
Two – tailed	H_1: $\mu \neq \mu_0$	$\lvert Z \rvert \geq Z_{\frac{\alpha}{2}}$
Right – tailed	H_1: $\mu > \mu_0$	$\lvert Z \rvert \geq Z_{\alpha}$
Left – tailed	H_1: $\mu < \mu_0$	

Table 9.2 - The Critical regions for Z - test

The value of $Z_{\frac{\alpha}{2}}$ and Z_{α} are determined from standard normal table such that the significance level is α .

$$\text{i.e. } P\left(|Z| \geq Z_{\frac{\alpha}{2}}\right) = \alpha \text{ and } P\left(Z \geq Z_{\alpha}\right) = \alpha$$

The critical value of $Z_{\frac{\alpha}{2}}$ and Z_{α} for various significance level are,

Significance level	Critical values of	
	$Z_{\frac{\alpha}{2}}$	Z_{α}
$\alpha = 0.01$	2.58	2.33
$\alpha = 0.05$	1.96	1.645

Table 9 3 - The Critical values for Z - test

Step 4: Compute the test statistic

Now consider the sample data and calculate the value of the test statistic.

Step 5: Make decision

The next step is to make the decision based on the values of the test statistic. If the value of the test statistic lies in the critical region, take decision to reject H_0. If the value of the test statistic does not lie in the critical region, take decision not to reject H_0. If we reject H_0, we say the test value is significant.

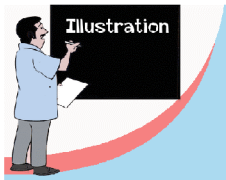


Illustration: 9.1

The lengths of metal bars produced by a particular machine are distributed with mean length 420 cm and standard deviation 12 cm. The machine is serviced, after which a sample of 100 bars are selected. The sample gives a mean length of 423 cm. Is there evidence, at 5% level, of the change in the mean length of the bars produced by the machine?

Solution:

Let X denotes the length of the bars produced by the machine. It is given that X is distributed with mean, $\mu = 420$ cm. and standard deviation $\sigma = 12$ cm.

We can perform the test procedure with the following steps.

1. State the hypotheses.	Here the hypotheses are, H_0: $\mu = 420$ (there is no change in the population mean, μ) H_1: $\mu \neq 420$ (there is a change in the population mean)
2. Determine the appropriate test statistic	Here the sample size is $n = 100$ (large) and the standard deviation σ is known. So the test is Z –test with test statistic, $Z = \dfrac{\bar{x} - \mu_0}{\dfrac{\sigma}{\sqrt{n}}}$ follows N(0, 1)
3. State the level of significance and critical region.	We perform a test with significance level 0.05. Since, the alternative hypothesis is H_1: $\mu \neq 420$, the test is a two – tailed test and the critical region is, $\lvert Z \rvert \geq Z_{\frac{\alpha}{2}}$ We know that for $\alpha = 0.05$, $Z_{\frac{\alpha}{2}} = 1.96$ So the critical region is, $\lvert Z \rvert \geq 1.96$
4. Compute the value of the test statistic	Here, $\bar{x} = 423$, $n = 100$ and $\sigma = 12$ So the test statistic is, $Z = \dfrac{423 - 420}{\dfrac{12}{\sqrt{100}}} = 2.5$
5. Make a decision	As $\lvert Z \rvert = 2.5 \geq 1.96$, we reject H_0 and conclude that there is a change in the mean length of the bars produced by the machine.

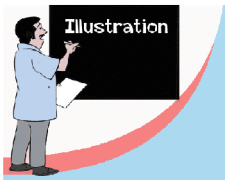


Illustration: 9.2

Experience has shown that the scores obtained in a particular test are normally distributed with mean score 70 and variance 36. When the test is taken by a random sample of 36 students, the mean score is 68.5. Is there sufficient evidence, at 5% level, that these students have not performed as expected?

Solution:

Let X denotes the score of a student. Then X is normally distributed with mean, $\mu = 70$ and variance $\sigma^2 = 36$.

The hypotheses are,

H_0: $\mu = 70$ and H_1: $\mu < 70$.

Since the population is distributed normal and the standard deviation is known, the test is Z – test with test statistic,

$$Z = \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}} \text{ follows N(0, 1).}$$

We perform a test with level of significance $\alpha = 0.05$.

Since the alternative hypothesis is H_1: $\mu < 70$, the test is one – tailed test and the critical region is, $|Z| \geq Z_\alpha$

For $\alpha = 0.05$, $Z_\alpha = 1.645$, so the critical region is, $|Z| > 1.645$,

Here $\bar{x} = 68.5$, $n = 36$ and $\sigma = 6$, so the test statistic is,

$$Z = \frac{68.5 - 70}{\frac{6}{\sqrt{36}}} = -1.5$$

As $|Z| = 1.5 < 1.645$, we can decide: do not reject H_0 (accept H_0) and can conclude that at 5% level the student have performed as expected.

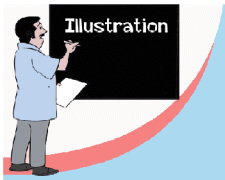


Illustration: 9.3

The mean weight of a sample of 100 students is 52 kgs. with standard deviation 3 kgs. Can it be considered as a sample taken from a normal population having mean greater than 50 kgs. at 1% level of significance?

Solution

Let X denotes the weight of a student. Suppose X is normally distributed with mean μ. The hypotheses are,

H_0: μ = 50 and H_1: μ > 50.

It is given that $\bar{x}$ =52, s = 3 and n = 100.

Here the population is supposed to follow normal distribution and the standard deviation of the population is unknown. So the test is Z – test with test statistic,

$$Z = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}} \quad \text{follows } N(0, 1)$$

The level of significance is $\alpha = 0.01$.

Since H_1: μ > 50, the test is one – tailed with critical region, $|Z| > Z_\alpha$

For $\alpha = 0.01$, $Z_\alpha = 2.33$. So the critical region is, $|Z| > 2.33$.

The value of the test statistic is,

$$Z = \frac{52 - 50}{\frac{3}{\sqrt{100}}} = 6.67 .$$

As |Z| = 6.67 > 2.33, we decide to reject H_0. So the conclusion is that the average weight of the students is greater than 50 kgs at 1% level of significance.

(b) t – test

When the population standard deviation is unknown, the Z – test is not normally used for testing, involving means. A different test, called t – test is used in this situation. The test is known as t – test since we use the t – test statistic, which follows the student's t distribution. The t – test is used for testing the mean of a single population under the following conditions.

1. The population follows normal distribution.
2. The population standard deviation is unknown, but the sample standard deviation s is known.
3. The sample size, n is small (n less than 30).

The t – test statistic is,

$$t = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n-1}}}$$

This test statistic follows t distribution with degrees of freedom $n - 1$.

The following table gives the critical region in various situations.

Test	Alternative hypothesis	Critical Rigion
Two – tailed	$H_1: \mu \neq \mu_0$	$\lvert t \rvert \geq t_{\frac{\alpha}{2}}$
Right – tailed	$H_1: \mu > \mu_0$	$\lvert t \rvert \geq t_{\alpha}$
Left – tailed	$H_1: \mu < \mu_0$	

Table 9.4 - The Critical regions for t - test

The critical value of $t_{\frac{\alpha}{2}}$ and t_{α} are determined from t – table such that the significance level is α.

i.e., $P\left(\lvert t \rvert \geq t_{\frac{\alpha}{2}}\right) = \alpha$ and $P\left(t \geq t_{\alpha}\right) = \alpha$

For one tailed tests, find the α level at the top row of the table and find the degrees of freedom by looking down the left end column.

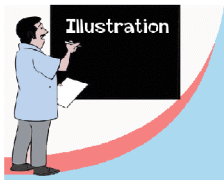


Illustration: 9.4

Sixteen oil tins are taken at random from an automatic filling machine. The mean weight of the tins is 14.2 kg, with a standard deviation of 0.40 kg. Can we conclude that the filling machine is wasting oil by filling more than the intended weight of 14 kg, at a significance level of 5%? (Assuming normality).

Solution:

Let X denotes the weight of oil in a tin. Then X follows a normal distribution with mean μ and standard deviation σ (unknown). The hypotheses for testing are:

H_0: μ = 14 kg and H_1: μ > 14 kg.

Here the population standard deviation is unknown and sample sd, s is known. So the test is *t test* with test statistic,

$$t = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n-1}}} \quad \text{follows} \quad t_{(n-1)}$$

The test is one – tailed with significance level 5%. The critical region is, $|t| > t_\alpha$. From t – table with degrees of freedom, $n - 1 = 15$, $t_\alpha = 1.7530$. So the critical region becomes, $|t| > 1.7530$.

Given that, $\bar{x} = 14.2$ kg, s = 0.40 and n = 16. The test statistic is,

$$t = \frac{14.2 - 14}{\frac{0.40}{\sqrt{15}}} = 1.936$$

Conclusion: As $|t| = 1.936 > 1.7530$, the tabled value, we take decision to reject H_0. So we conclude that the filling machine is wasting oil by filling more than the intended weight of 14 kg.

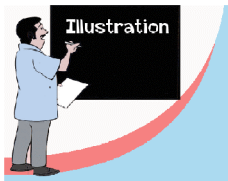


Illustration: 9.5

Five readings of the resistance in ohms, of a piece of wire gave the following results.

1.51, 1.49, 1.54, 1.52, 1.54

If the wire was pure silver, its resistance would be 1.50 ohms. If it was impure, the resistance would be increased. Test at 5% level, the hypothesis that the wire is pure silver. (Assume that the resistance follows a normal distribution)

Solution

Let X denotes the resistance in ohms of a piece of wire. Assume that X is normally distributed with mean μ and standard deviation σ.

The hypotheses for testing are:

H_0: μ = 1.50 ohms and H_1: μ > 1.50 ohms.

The test is one – tailed with significance level, $\alpha = 0.05$.

Here the population is normal and sd, σ is unknown also we are given a sample of

size, $n = 5$ (small sample). So the test is t- test with test statistic,

$$t = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n-1}}} \quad \text{follows} \quad t_{(n-1)},$$

were s is the sample standard deviation.

The critical region is, $|t| > t_\alpha$

From t - table with degrees of freedom, $n - 1 = 4$, the critical values is 2.132. So the critical region becomes $|t| > 2.132$.

From the sample data, we have

$$\bar{x} = \frac{\sum x}{n} = \frac{7.6}{5} = 1.52 \text{ and } s = \sqrt{\frac{\sum x^2}{n} - (\bar{x})^2} = \sqrt{\frac{11.5538}{5} - (1.52)^2} = 0.019$$

Now the value of the test statistic is,

$$t = \frac{1.52 - 1.50}{\frac{0.019}{\sqrt{4}}} = 2.105$$

Conclusion: As $|t| = 2.105 < 2.132$, the tabled value, we take decision to do not reject H_0 (accept H_0). So the conclusion is that, at 5 % level of significance, there is not sufficient evidence that the wire is impure, the wire can be considered as pure silver.

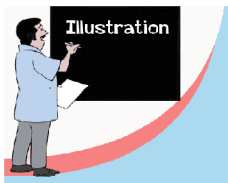


Illustration: 9.6

A random sample of size 16 has 53 as mean and the sum of squares of the deviations taken from the mean is 150. Can the sample be regarded as taken from a normal population with mean 56? (Significance level is 5%)

Solution:

Let μ be the mean of the population. The hypotheses are:

H_0: $\mu = 56$ and H_1: $\mu \neq 56$.

Here the population standard deviation is unknown and sample sd, s is known. So the test is *t test* with test statistic,

$$t = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n-1}}} \quad \text{follows} \quad t_{(n-1)}$$

The test is one – tailed with significance level, $\alpha = 0.05$. The critical region is,

$|t| \geq t_{\frac{\alpha}{2}}$.

From t - table with degrees of freedom, n – 1 = 15, the critical values is 2.132. So the critical region becomes $|t| \geq 2.132$.

From the sample data, we have

$\bar{x} = 53$ and $\sum (x - \bar{x})^2 = 150$

So, $s^2 = \frac{\sum (x - \bar{x})^2}{n} = 9.375$

The test statistic is,

$$t = \frac{53 - 56}{\sqrt{\frac{9.375}{15}}} = -3.79$$

Conclusion: As $|t| = 3.79 > 2.132$, we take decision to reject H_0. So we conclude that the sample can not be regarded as one from a normal population with mean 56.

Activity

Take a simple random sample of size 30 plus two students in your school. Using this sample , estimate the average mark in English for their Plus One examination.Test this average mark using another sample of size 20 from this population.

Know your progress

1. What are the situations in which Z – test is used for testing the mean of a population?
2. Give the conditions, under which t – test is used for testing the mean of a population?

3. A company is engaged in the packing of a superior quality tea in jars of 500gm each. The company is of the view that as long as the jars contains 500gm of tea, the process is under control. The standard deviation of the process is 50gm. A sample of 225 jars is taken at random and the sample average is found to be 510 gm. Has the process gone out of control at 5% level of significance?

4. A sample of size 400 was drawn and the sample mean found to be 99. Test, at 5% level of significance, whether this sample could have come from normal population with mean 100 and variance 64.

5. A manufacturer of a new motorcycle claims for it an average mileage of 60 km/litre under city conditions. However, the average mileage in 16 trials is found to be 57 km, with a standard deviation of 2 km. Is the manufacturer's claim justified at 1% level of significance? (Assume normality).

9.7 Tests for significance for equality of two population means (Z - test)

In this section, we will discuss the difference in mean from two samples taken from two populations. On many occasions an investigator wants to compare two means taken from two different samples from two populations. For example, an investigator is analyzing the difference in consumer satisfaction for a particular product in two cities, Thiruvananthapuram and Chennai. In order to accomplish this, the investigator collects two different samples from the two cities, obtain the sample means and then compare these two means. Finally, he draws a conclusion about the population means based on the inference obtained from the sample means.

Suppose μ_1 and μ_2 are the means of the two populations, the standard deviations are σ_1 and σ_2.

The null hypothesis used for testing is,

$$H_0: \mu_1 = \mu_2$$

The alternative hypothesis is any one of the following

$$H_1: \mu_1 \neq \mu_2 \text{ (two – tailed test)}$$

$$H_1: \mu_1 > \mu_2 \text{ (right – tailed test)}$$

H_1: $\mu_1 < \mu_2$ (left – tailed test)

The formula for Z – test statistic for comparing two population means is,

$$Z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

Where Z follows N(0, 1) and $\bar{x}_1$, $\bar{x}_2$ are the means of samples taken from the two populations.

Under the hypothesis H_0 is true, the test statistic becomes,

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

The critical region and critical values as described in *tables 9.2 and 9.3.*

The following are the assumptions for applying the Z – test for testing the equality of means for two populations.

1. Both samples taken from the populations are random samples.
2. The samples must be independent to each other.
3. The standard deviations of both populations must be known and if the sample sizes are less than 30, the populations must be normally distributed or approximately normally distributed.

If both samples are large and the population standard deviations are unknown, we can approximate the standard deviations by the sample standard deviations. In this case the test statistic becomes,

$$Z = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

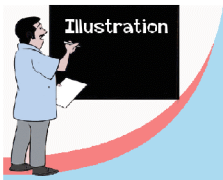


Illustration: 9.7

A random sample of size 100 is taken from a normal population with variance 40. The sample mean is 38.3. Another random sample of size 80 is taken from a normal population with variance 30. The sample mean is 40.1. Test, at 5% level, whether there is a significant difference in the population means.

Solution:

Let μ_1 and μ_2 be the means and σ_1 and σ_2 be the standard deviations of the two normal populations. The hypotheses for testing are,

H_0: $\mu_1 = \mu_2$ and H_1: $\mu_1 \neq \mu_2$

It is given that,

$n_1 = 100$, $\bar{x}_1 = 38.3$, $\sigma_1^2 = 40$ and

$n_2 = 80$, $\bar{x}_2 = 40.1$, $\sigma_2^2 = 30$.

The test statistic is,

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

The test is two tailed with significance level 0.05. The critical region is $|Z| \geq 1.96$

Here the value of the test statistic is,

$$Z = \frac{(38.3 - 40.1)}{\sqrt{\frac{40}{100} + \frac{30}{80}}} = -2.04$$

Conclusion: As $|Z| = 2.04 > 1.96$, we reject H_0 and conclude that there is a difference in population means at 5% level.

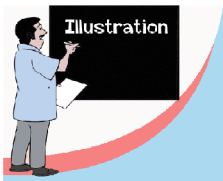


Illustration: 9.8

In order to make a survey of buying habits, two markets A and B are chosen at two different parts of a city from Kerala. 400 women shoppers were chosen at random from Market A. Their average weekly expenditure on food is found to

be Rs.750/- with a standard deviation of Rs. 40/-. These figures are Rs. 660/- and Rs. 55/- respectively in the market B for a sample of 500 women shoppers. Test at 1 % level, whether the average weekly food expenditure for the two population of shoppers are equal.

Solution:

Let μ_1 and μ_2 be the means of the two populations of shoppers. The hypotheses are

H_0: $\mu_1 = \mu_2$ and H_1: $\mu_1 \neq \mu_2$

It is given that,

$n_1 = 400$, $\bar{x}_1 = 750$, $s_1 = 40$ and

$n_2 = 500$, $\bar{x}_2 = 660$, $s_2 = 55$.

Since the samples are large and the population standard deviations are unknown and sample standard deviations are known, the test statistic is,

$$Z = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Here the test is two tailed with significance level, 0.01. So the critical region is $|Z| \geq 2.58$

The value of the test statistic is,

$$Z = \frac{(750 - 660)}{\sqrt{\frac{40^2}{400} + \frac{55^2}{500}}} = 28.39$$

Conclusion: As $|Z| = 28.39 > 2.58$, we take decision to reject H_0. The conclusion is that there is significant difference between the average weekly food expenditure of the two populations at 1% level of significance.

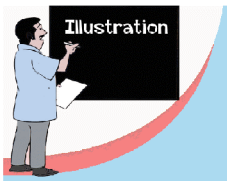


Illustration: 9.9

The same test was given to a group of 100 scouts and to a group of 144 guides. The mean score for the scouts was 27.53 and the mean score for the guides was 26.81. The standard deviations of scores of the two populations of

scouts and guides were 3.48 and 2.52 respectively. Test using a 5% level of significance, whether the scout's performance was better than that of the guide's. Assume that the scores are normally distributed.

Solution:

Let μ_1 and μ_2 be the means the two populations of scouts and guides respectively. The hypotheses for testing are

H_0: $\mu_1 = \mu_2$ and H_1: $\mu_1 > \mu_2$

It is given that

$n_1 = 100$, $\bar{x}_1 = 27.53$, $\sigma_1 = 3.48$ and

$n_2 = 144$, $\bar{x}_2 = 26.81$, $\sigma_2 = 2.52$.

Since the population is normal and standard deviation of the populations are known, the test statistic is,

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

Here the test is one – tailed with significance level 0.05. The critical region is $|Z| > 1.645$.

The value of the test statistic is,

$$Z = \frac{(27.53 - 26.81)}{\sqrt{\frac{(3.48)^2}{100} + \frac{(2.52)^2}{144}}} = 1.77$$

Conclusion: As $|Z| = 1.77 > 1.645$, we take decision to reject H_0. So the conclusion is that at 5% level of significance, the performance of scouts are better than that of guides.

Know your progress

1. Which test statistic is used to test the equality of means of two normal populations?

2 A survey found that the average hotel room rent in New Delhi is Rs. 1015/- and the average room rent in Mumbai

is Rs. 1020/-. Assume that the data were obtained from two samples of 50 hotels each and that the standard deviations of the populations are Rs. 15.62/- and Rs. 14.83/-, respectively. At $\alpha = 0.05$, can it be concluded that there is a significant difference in the rent rates?

9.8 Chi – square test for independence of attributes

When data can be tabulated in tabular form of frequencies, several types of hypothesis can be tested by using chi – square tests. One of them is the test for independence of two attributes.

In many situations, an investigator might be interested in finding the relationship between the two variables or to check whether they are independent of each other. For example, a motor cycle manufacturing company may be interested in knowing whether the purchase of motor cycle is independent of the customer's age or whether it is dependent of the customer's age. Another example, an educationalist may be interested in testing whether the language ability and mathematical ability of school students are independent or not.

When the observations are classified on the basis of two variables and arranged in a table, the resulting table is referred to as a contingency table.

The following is a contingency table showing the observed frequency on two variables X and Y.

Variable X

Variable Y	X_1	X_2	X_3	.	.	.	X_k	Row Total
Y_1	O_{11}	O_{12}	O_{13}				O_{1k}	R_1
Y_2	O_{21}	O_{22}	O_{23}				O_{2k}	R_2
Y_3	O_{31}	O_{32}	O_{33}				O_{3k}	R_3
.								.
.								.
.								.
Y_j	O_{j1}	O_{j2}	O_{j3}				O_{jk}	R_j
Column Total	C_1	C_2	C_3	.	.	.	C_k	N

Table - 9.5 - Contingency table

The observation in each column represents the frequency of observations that are common to the respective row and column. The row totals are denoted by R_1, R_2, ... and column totals are denoted by C_1, C_2, ... When we add the row total or column totals, we get the total frequency N. For applying the chi – square test, it is very important to determine the expected frequencies under the assumption that the two variables are independent.

The calculation of expected frequencies.

The expected frequency of the ij th cell, denoted by E_{ij} is obtained as,

$$E_{ij} = \frac{R_i \times C_j}{N}$$

Where R_i – Row total of the i th row,

C_j – Column total of the j th column,

N – Total frequency

$$\text{Expected frequency of any cell} = \frac{\text{Row total} \times \text{Column total}}{\text{Total frequency}}$$

For example,

$$E_{11} = \frac{R_1 \times C_1}{N},\ E_{12} = \frac{R_1 \times C_2}{N},\ E_{23} = \frac{R_2 \times C_3}{N}, \text{ etc.}$$

Now the expected frequency can be placed in the corresponding cells along with the observed values as shown below,

Variable X

Variable Y	X_1	X_2	X_3	.	.	.	X_k	Row Total
Y_1	O_{11} (E_{11})	O_{12} (E_{12})	O_{13} (E_{13})				O_{1k} (E_{1k})	R_1
Y_2	O_{21} (E_{21})	O_{22} (E_{22})	O_{23} (E_{23})				O_{2k} (E_{2k})	R_2
Y_3	O_{31} (E_{31})	O_{32} (E_{32})	O_{33} (E_{33})				O_{3k} (E_{3k})	R_3
.								.
.								.
.								.
Y_j	O_{j1} (E_{j1})	O_{j2} (E_{j2})	O_{j3} (E_{j3})				O_{jk} (E_{jk})	R_j
Column Total	C_1	C_2	C_3	.	.	.	C_k	N

Table - 9.6 - Contingency table along with expected frequencies

For a better understanding in the determination of expected values, consider the following illustration.

A researcher wishes to determine whether there is a relationship between the hospital and the nature of patient infections. A sample of 400 people is selected from two hospitals A and B, the following data are obtained.

Hospital	No of patients infected by		
	Surgical	Pneumonia	Blood stream
A	100	80	20
B	50	120	30

This is a contingency table with 2 rows and 3 columns; it is called a 2 X 3 contingency table. The observed values are O_{11}=100, O_{12}=80, O_{13}=20, O_{21}=50, O_{22}=120 and O_{23}=30. For finding the expected values, first find the sum of each row and column and find the grand total as shown in the table.

Hospital	No of patients infected by			Row total
	Surgical	Pneumonia	Blood stream	
A	100	80	20	**200 (R_1)**
B	50	120	30	**200 (R_2)**
Column total	**150 (C_1)**	**200 (C_2)**	**50 (C_3)**	**400 (N)**

The expected values of each cell can be calculated as,

$$E_{11} = \frac{R_1 \times C_1}{N} = \frac{200 \times 150}{400} = 75$$

$$E_{12} = \frac{R_1 \times C_2}{N} = \frac{200 \times 200}{400} = 100$$

$$E_{13} = \frac{R_1 \times C_3}{N} = \frac{200 \times 50}{400} = 25$$

$$E_{21} = \frac{R_2 \times C_1}{N} = \frac{200 \times 150}{400} = 75$$

$$E_{22} = \frac{R_2 \times C_2}{N} = \frac{200 \times 200}{400} = 100$$

$$E_{23} = \frac{R_2 \times C_3}{N} = \frac{200 \times 50}{400} = 25$$

Now the completed contingency table along with expected values as shown below table

Hospital	No of patients infected by			Row total
	Surgical	Pneumonia	Blood stream	
A	100 (75)	80 (100)	20 (25)	**200**
B	50 (75)	120 (100)	30 (25)	**200**
Column total	**150**	**200**	**50**	**400**

The Chi - square test statistic

The test statistic for testing the independence of attributes is

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

This test statistic follows a Chi-square distribution with degrees of freedom (No. of rows -1) $\times$ (No. of columns -1). The reason for this formula for degrees of freedom is that all the expected values except one are free to vary in each row and each column. In the above illustration the d.f is $(2-1) \times (3-1) = 2$ as there are 2 rows and 3 columns. The chi – square test statistic is,

$$\chi^2 = \frac{(100-75)^2}{75} + \frac{(80-100)^2}{100} + \ldots + \frac{(30-25)^2}{25} = 26.67$$

The critical region

The critical region for testing the chi – square test of for independence is,

$$\chi^2 \geq \chi_\alpha^{\ 2}$$

Where the value of $\chi_\alpha^{\ 2}$ (known as the tabled value) can be determined from chi – square table with respective degrees of freedom and level of significance α, such that,

$$P(\chi^2 \geq \chi_\alpha^{\ 2}) = \alpha.$$

We are rejecting the test of independence if the χ^2 test value greater than the tabled value.

Now the test procedure can be summarized as follows,

Step 1: Formulate the hypotheses

The hypotheses for testing the independence of attributes can be stated as follows.

H_0: The attributes are independent.

H_1: The attributes are not independent.

Step 2: Determine the test statistic.

The test statistic is,

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

With degrees of freedom $(R-1)\times(C-1)$, where R is the number of rows and C is the number of columns.

Step 3: Decide the level of significance and critical region

The level of significance α is commonly taken as 0.05 or 0.01.

The critical region is $\chi^2 \geq \chi_\alpha^{\ 2}$

Step 4: Calculate the test statistic

For calculating the test statistic, we have to consider the observed values and to compute

the expected values as explained previously. Then compute the contingency table and the value of test statistic can be determined by the formula,

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

Step 5: Make the decision

We make the decision to reject H_0 if the calculated value of test statistic is greater than the tabled value.

i.e. if, $\chi^2 \geq \chi_\alpha^{\ 2}$

Take decision to reject H_0 if $\chi^2 \geq \chi_\alpha^{\ 2}$.

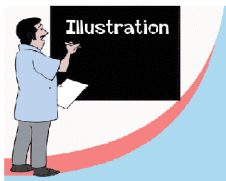


Illustration: 9.10

A driving school examined the result of 100 candidates who were taking their driving test for the first time. They found that, of the 40 men, 28 passed and out of the 60 women, 34 passed. Do these results indicate, at the 5% level of significance, a relationship between the sex of the candidate and the ability to pass first time?

Solution:

The results of the driving test given can be summarized as given in the following table.

Sex	Result		Total
	Pass	Fail	
Male	28	12	**40**
Female	34	26	**60**
Total	**62**	**38**	**100**

This is a 2 X 2 contingency table.

We have to test the hypothesis

H_0: The sex of the candidates and ability to pass first time are independent.

The alternative hypothesis is

H_1: The sex and the ability to pass first time are not independent.

For testing the independence of attributes, we use the χ^2 test. The test statistic is,

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

This follows a χ^2 distribution with degrees of freedom, (C – 1)(R – 1), where, C is the number of columns and R, the number of rows. Here, C = 2 and R = 2. So the degrees of freedom is $1 \times 1 = 1$.

The critical region is $\chi^2 \geq \chi_\alpha{}^2$

For level of significance $\alpha = 0.05$ and degrees of freedom 1, the critical region becomes, $\chi^2 \geq 3.84$, i.e., we reject the hypothesis of independence if the calculated value of $\chi^2 \geq 3.84$.

The expected values are calculated as follows.

$$E_{11} = \frac{40 \times 62}{100} = 24.8, \; E_{12} = \frac{40 \times 38}{100} = 15.2,$$

$$E_{21} = \frac{60 \times 62}{100} = 37.2 \text{ and } E_{22} = \frac{60 \times 38}{100} = 22.8$$

Now the completed contingency table along with the expected values is,

Sex	Result		Total
	Pass	Fail	
Male	28 (24.8)	12 (15.2)	**40**
Female	34 (37.2)	26 (22.8)	**60**
Total	**62**	**38**	**100**

The value of the test statistic is,

$$\chi^2 = \frac{(28-24.8)^2}{24.8} + \frac{(12-15.2)^2}{15.2} + \frac{(34-37.2)^2}{37.2} + \frac{(26-22.8)^2}{22.8} = 1.81$$

As $\chi^2 = 1.81 < 3.84$, the tabled value, we take decision to accept H_0. So the conclusion is that the sex of the candidate and the ability to pass first time are independent.

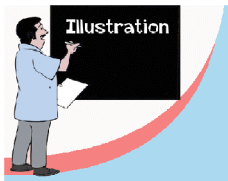


Illustration 9.11

The following table gives the classification of a sample of 150 people according to their eye colour and hair colour. Examine at 5% level of significance, whether the two attributed are associated.

Eye Colour	Hair colour			Total
	Fair	Brown	Black	
Blue	15	5	20	40
Grey	20	10	20	50
Brown	25	15	20	60
Total	60	30	60	150

Solution:

The hypotheses are,

H_0: The eye colour and hair colour are independent.

H_1: The eye colour and hair colour are not independent.

We are given a 3 X 3 contingency table.

For testing the independence of attributes, we use the χ^2 test. The test statistic is,

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

This follows a χ^2 distribution with degrees of freedom, $(R-1)\times(C-1)$, where, R is the number of rows and C, the number of columns. Here, C = 3 and R = 3. So the

degrees of freedom is 2 x 2 = 4.

The critical region is $\chi^2 \geq \chi_\alpha^{\ 2}$

For significance level $\alpha = 0.05$ and degrees of freedom 4, the critical region is, $\chi^2 \geq 9.5$

The observed frequencies are given in the table and the expected frequencies can be calculated as follows.

$$E_{11} = \frac{40 \times 60}{150} = 16,\ E_{12} = \frac{40 \times 30}{150} = 8,\ E_{13} = \frac{40 \times 60}{150} = 16$$

$$E_{21} = \frac{50 \times 60}{150} = 20,\ E_{22} = \frac{50 \times 30}{150} = 10,\ E_{23} = \frac{50 \times 60}{150} = 20$$

$$E_{31} = \frac{60 \times 60}{150} = 24,\ E_{32} = \frac{60 \times 30}{150} = 12,\ E_{33} = \frac{60 \times 60}{150} = 24$$

The completed contingency table along with expected frequency is,

Eye Colour	Hair colour			Total
	Fair	Brown	Black	
Blue	15 (16)	5 (8)	20 (16)	40
Grey	20 (20)	10 (10)	20 (20)	50
Brown	25 (24)	15 (12)	20 (24)	60
Total	*60*	*30*	*60*	*150*

The value of the test statistic is,

$$\chi^2 = \frac{(15-16)^2}{16} + \frac{(5-8)^2}{8} + \ldots + \frac{(20-24)^2}{24} = 3.6$$

As $\chi^2 = 3.6 < 9.5$, the tabled value, we take decision to accept H_0. So the conclusion is that the two attributes, eye colour and hair colour are independent.

Know your progress

1. Which is the null hypothesis used when we are testing the independence of attributes?
2. How can we find the expected frequency of a cell in a contingency table?
3. Give test statistic used in the test for independence of attributes.
4. What is the critical region of the test for independence of attributes?
5. How can we determine the degrees of freedom of chi – square statistic?
6. Give the test procedure for testing the independence of attributes?

Let us conclude

Statistical inference is the branch of Statistics that deals with uncertainty in decision making and provides a basis for making scientific decisions. Statistical inference is based on estimation and hypothesis testing. Testing of hypothesis is the soul of statistical inference. The testing of hypothesis deals with the methods for deciding either to accept or reject a hypothesis based on a sample taken from the population. A statistical hypothesis is an assumption made about the value of a parameter or the form of distribution. The testing process includes two types of hypothesis – null and alternative. Based on the alternative hypothesis, the test is classified as, one – tailed test and two – tailed tests. We select an appropriate test statistic for testing H_0 against H_1 using samples taken from the population. We set a specific significance level and decision making criteria. The area of the test statistic is divided into two – the rejection region and acceptance region. The rejection region is termed as critical region.

In this chapter we discuss the various steps in a statistical test procedure. The various methods for testing the significance of the mean of a single population – Z- test and t - test, the equality of means of two population means using Z – test and the procedure for testing the independence of attributes using chi – square test are explained in this chapter.

Lab Activity

1. Data on the number of tomatoes per plant on two varieties are given below. Test whether the mean of the two varieties are equal. (Significance level is 0.05)

Variety A	6	8	10	12	12	14	11	6	8	9	14	13	7	8	10	12
	14	15	7	8	13	16	9	10	13	14	13	14	14	9	11	
Vareity B	8	10	12	13	15	17	19	9	8	11	13	15	17	21	14	17
	16	14	14	8	9	12	15	19	12	10	13	15	16			

2. The following are two samples taken from two populations with variances 6.76 and 7.34 respectively. Test whether the populations have the same mean at 1% level of significance.

A:	10	12	15	18	13	15	16	6	15	16	14	18	12	14	18
B:	5	8	10	9	9	11	12	16	16	8	8	9	10	11	7

Let us assess

For Questions 1-14, choose the correct answer from the given choices.

1. An assumption made about the value of a population parameter is called a:

   a) hypothesis b) conclusion c) confidence d) significance

2. Null and alternative hypotheses are statements about:

   a) sample statistics

   b) population parameters

   c) sample parameters

   d) sometimes population parameters and sometimes sample parameters

3. The null and alternative hypotheses divide all possibilities into:

a) two sets that overlap b) two non – overlapping sets

c) two sets that may or may not overlap

d) as many sets as necessary to cover all possibilities

4. Which of the following is true of the null and alternative hypotheses?

a) exactly one hypothesis must be true

b) both hypothesis must be true

c) it is possible for both hypotheses to be true

d) it is possible for neither hypothesis to be true

5. The form of alternative hypothesis can be:

a) one -tailed b) two- tailed

c) neither one-tailed nor two-tailed d) one or two – tailed

6. The hypothesis that an analyst is trying to prove is called the:

a) elective hypothesis b) alternative hypothesis

c) optional hypothesis d) null hypothesis

7. Two – tailed alternatives are phrased in terms of:

a) $\neq$ b) $<$ or $>$ c) $\leq$ or $\geq$ d) $=$ or $\approx$

8. A type I error occurs when:

a) the null hypothesis is incorrectly accepted when it is false

b) the null hypothesis is incorrectly rejected when it is true

c) the sample mean differs from the population mean

d) the test is biased

9. In hypothesis testing, a type II error occurs when:

a) the null hypothesis is not rejected when the null hypothesis is true.

b) the null hypothesis is rejected when the null hypothesis is true.

c) the null hypothesis is not rejected when the alternative hypothesis is true.

d) the null hypothesis is rejected when the alternative hypothesis is true.

10. The sum of values of α and β:

    a) always up to 1 b) always up to 0.5

    c) is the probability of Type II error d) None of the above

11. If you wish to test the claim that the mean of the population is 100, the appropriate alternative hypothesis is:

    a) $\bar{x} = 100$ b) $\mu \geq 100$ c) $\mu \neq 100$ d) $\mu \leq 100$

12. Assume the cholesterol levels in a certain population have mean μ = 200 and standard deviation 24. The cholesterol levels for a random sample of n = 9 individuals are measured and the sample mean $\bar{x}$ is determined. What is the z – score for a sample mean $\bar{x} = 180$?

    a) - 3.75 b) - 2.50 c) -0.83 d) 2.50

13. The null hypothesis for the chi – square test for independence is that the variables are:

    a) dependent b) related c) independent d) always 0

14. The degrees of freedom for the chi – square test for independence of attributes is:

    a) n b) $(R-1)\times(C-1)$

    c) $n - 1$ d) none of these

15. Define null and alternative hypotheses. Give an example of each.

16. What is meant by type I and type II errors? What symbols are used to represent their probabilities?

17. When should a one – tailed and a two – tailed tests are used?

18. List the steps in statistical test procedure.

19. A researcher wishes to test the claim that the average cost of tuition and fees at a college is greater than Rs. 5700/-. He selects a random sample of 36 college students and ûnds the mean to be Rs. 5950/-. The population standard deviation is Rs. 659/-. Is there evidence to support the claim at $\alpha = 0.05$?

20. It is claimed by an agency that the average wind speed in a certain city is 8 miles per hour. A sample of 32 days has an average wind speed of 8.2 miles per hour. The standard deviation of the population is 0.6 mile per hour. At $\alpha = 0.05$, is there enough evidence to reject the claim?

21. The distribution of blood pressure of female diabetic patients in a city has an unknown μ and sd, σ = 9 mmHg. It may be useful to doctors to know whether the mean of this population is 77 mmHg. A sample of 10 diabetic women is selected. Their mean blood pressure is 84 mmHg. Using this information, conduct a two tailed test at 5% level of significance.

22. According to the *Digest of Educational Statistics*, a certain group of preschool children under the age of one year each spends an average of 30.9 hours per week in non parental care. A study of state university center-based programs indicated that a random sample of 32 infants spent an average of 32.1 hours per week in their care. The standard deviation of the population is 3.6 hours. At α = 0.01, is there sufficient evidence to conclude that the sample mean differs from the national mean?

23. It is known that, nationally, doctors working for health maintenance organizations (HMOs) average 13.5 years of experience in their specialties, with a standard deviation of 7.6 years. The executive director of an HMO in a state is interested in determining whether or not its doctors have less experience than the national average. A random sample of 150 doctors from HMOs shows a mean of only 10.9 years of experience.

    a. State the null and alternative hypotheses to test whether or not doctors in this HMO have less experience than the national average.

    b. Using an alpha level of .01, make this test.

24. Ten individuals are chosen at random from a normal population and their heights in inches are found to be 63, 63, 66, 67, 68, 69, 70, 70, 71 and 71. In the light of this date examine the claim that the mean height of the population is 66 inches.

25. A medical investigation claims that the average number of infections per week at a hospital is 16.3. A random sample of 10 weeks had a mean number of 17.7 infections. The sample standard deviation is 1.8. Is there enough evidence to reject the investigator's claim at α = 0.05?

26. The life span of a random sample of 10 electric bulbs was examined and the mean and sd of life span were found to be 1190 hours and 10 hours respectively. Do the data support the hypothesis that the life span of the particular brand of bulbs is normally distributed with mean 1200 hours.

27. A physician claims that joggers' maximal volume oxygen uptake is greater than the average of all adults. A sample of 15 joggers has a mean of 40.6 milliliters per

kilogram (ml/kg) and a standard deviation of 6 ml/kg. If the average of all adults is 36.7 ml/kg, is there enough evidence to support the physician's claim at α = 0.05?

28. A researcher estimates that the average height of the buildings of 30 or more stories in a large city is at least 700 feet. A random ample of 10 buildings is selected, and the heights in feet are shown. At α = 0.05, is there enough evidence to reject the claim?

    485 511 841 725 615 520 535 635 616 582

29. The average local cell phone call length was reported to be 2.27 minutes. A random sample of 20 phone calls showed an average of 2.98 minutes in length with a standard deviation of 0.98 minute. At α = 0.05 can it be concluded that the average differs from the population average?

30. A survey found that the average hotel room rate in New Delhi is Rs. 5380/- and the average room rate in Mumbai is Rs. 4840/-. Assume that the data were obtained from two samples of 50 hotels each and that the standard deviations of the populations are Rs. 312/- and Rs. 288/-, respectively. At α = 0.05, can it be concluded that there is a signiûcant difference in the rates?

31. In order to make a survey of buying habits, two markets A and B are chosen at two different parts of a city from Kerala. 400 women shoppers are chosen at random in market A. Their average weekly expenditure on food is found to be Rs. 750/- with a sd of Rs. 40/-. These figures are Rs. 660/- and Rs. 55/- respectively in the market B for a sample of 500 women shoppers. Test at 1% level, whether the average weekly food expenditure of the two population of shoppers are equal.

32. The average length of "short hospital stays" for men is slightly longer than that for women. A random sample of recent hospital stays for both men and women revealed the following. At α = 0.01, is there sufûcient evidence to conclude that the average hospital stay for men is longer than the average hospital stay for women?

	Men	Women
Sample size	32	30
Sample mean	5.5 days	4.2 days
Population standard deviation	1.2 days	1.5 days

33. Electric bulbs manufactured by X and Y companies gave the following result.

	Company X	Company Y
No of bulbs used	100	100
Mean life time	1300	1272
Standard deviation	82	93

Test whether there is any significant difference as the life span of the two markets at 1% level of significance.

34. Some researchers examined internet usage among college students in the United States and in India. Usage was divided into two categories, for personal use and course-related use. We compare average hours between the U.S. and Indian students in the following table.

	Course Work Hours	Personal Hours
U.S. students -	$\bar{x} = 1.76$	$\bar{x} = 2.08$
n = 149	S = 1.52	S = 1.91
Indian students -	$\bar{x} = 0.73$	$\bar{x} = 0.87$
n = 306	S = 0.79	S = 0.78

a) Determine whether U.S. students have significantly higher internet use for course work than Indian students. Test at the 0.05 alpha level.

b) Test whether there is a significant difference in internet use for personal use between the U.S. and Indian students. Test at the 0.01 alpha level.

35. One thousand girls in a college were graded according to their I.Q and the economic conditions of their homes. Find out whether there is any association between economic conditions at home and the I.Q of girls.

	I. Q		
Economic conditions	High	Low	Total
Rich	100	300	400
Poor	350	250	600
Total	450	550	1000

(Given for degrees of freedom 1 and $\alpha = 0.05$, $\chi_{\alpha}^{2} = 3.84$)

36. In an industry, 200 workers, employed for a specific job, were classified according to their performance and training received/not received to test independence of a specific training and performance. The data summarized as followed:

	Performance		Total
	Good	Not good	
Trained	100	50	150
Untrained	20	30	50
Total	120	80	200

Use χ^2 test of independence at 5% level of significance and write conclusions.

37. The result of a survey regarding radio listener's preference for different type of music are given in the table with listeners classified by age groups. Test the hypothesis that age and preference for type of music are independent.

	Age group			
Types of music	19 – 25	26 – 35	Above 36	Total
National music	80	60	9	149
Foreign music	210	325	44	579
Indifferent	16	45	132	193
Total	306	430	185	921

Chapter 10

Analysis of Variance

We discussed various methods to test the significance of single mean or difference of two means in the last chapter-Testing of Hypothesis. We can test whether the population mean is equal to a particular value or two populations have the same mean, by using these methods. But there may be situations where more than two populations are involved and the need to test the significance of differences among their means. For example, if a

Significant Learning Outcomes

After the completion of this chapter, the learner:

- Explains and identifies the concept of ANOVA.
- Categorises the types of variations.
- Identifies causes of variations.
- Constructs ANOVA Table.
- Interprets the ANOVA Table.

company has to study whether the six varieties of products marketed by the company have equal demand in the market or not. Or if a researcher has to study whether four different teaching methods have equal efficiency or not. Or if an agricultural researcher has to study the efficiency of three fertilisers are significantly different or not. Or if a drug manufacturing company has to test whether the five different drugs produced for a particular disease are equally efficient or not etc.

In these situations we cannot use the t-test or z-test to test the hypothesis. In such situations we can use another method, known as Analysis of Variance (abbreviated as ANOVA). ANOVA is a statistical technique used to study the significant difference of several means. In other words, ANOVA tests whether the 'k' samples (k>2) can be considered as having been drawn from the same population. (or equivalently from 'k' populations having the same means.)

> ANOVA is a statistical method to test the significant difference of several means.

The null hypothesis is therefore,

H_0 : The means are equal

Or $H_0 : \mu_1 = \mu_2 = \mu_3 = \ldots\ldots = \mu_k$

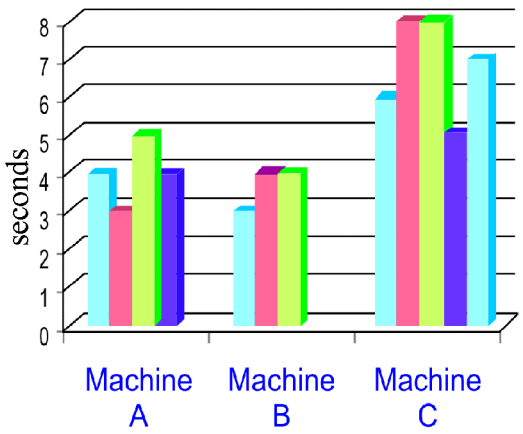
Consider the following example:

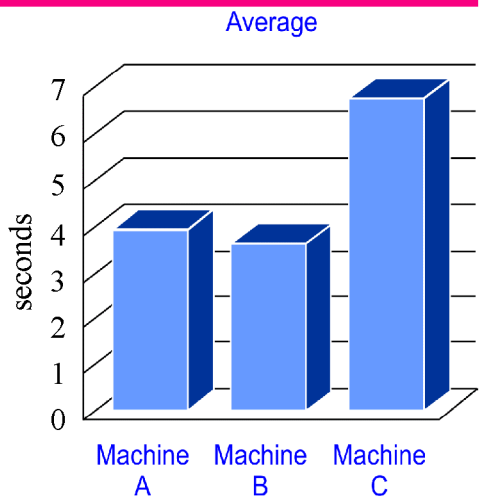
Let a company has three machines for packing. The packing time taken by the machines is observed as follows.

	Time Taken in Seconds					Average
Machine A	4	3	5	4		4
Machine B	3	4	4			3.66
Machine C	6	8	8	5	7	6.8

In the above data you can see the average time taken by the machines is different, or there is a variation between the averages. We can say Machine A and Machine B took almost equal time on the average, but machine C took little more time for packing. From the average it is evident. Here we considered the variation between the sample means to study the significance of difference of several means. In ANOVA also the variations are considered.

10.1 Types of Variations





In the above example we discussed the variation between the sample means. This can be considered due to the variations of the efficiencies of machines. It is known as ***variation between samples.*** One can see that the observations in each sample (here the observations corresponding to each machine is considered as sample) is also different. The minimum time taken by the machine A to complete the packing process is 3 seconds and the maximum time taken is 5 seconds. Similar variations are visible in the other sample sets also. These variations are called ***variation within samples***. The sum of variation between samples and variation within samples is termed as ***total variation.***

> ***Total variation = Variation between samples + Variation within samples***

The null hypothesis is rejected when the variation between samples is high and variations within samples are small. In ANOVA the ratio of these two variations is used as the test statistic. We know that variations follow Chi-Square distribution and hence its ratio will follow F-Distribution.

10.2 Causes of Variations

Variation is inherent in every statistical data. There are various causes for variation. Consider we are studying the effects of a particular fertiliser in an agricultural field. The yield obtained from various plots will be different, even if we apply the same fertiliser in a fixed dose to each plant. The causes of variations are impact of the fertiliser, efficiency of the seed, influence of irrigation, fertility of soil, climate conditions, rainfall, etc. Among these, some components are controllable and measurable by the researcher. The effects

of fertiliser and irrigation are examples. Whereas the effects of fertility of soil, climate conditions, rainfall etc. cannot be controlled and measured by the researcher. Analysis of variance (ANOVA) is developed on the assumption that the variation between samples is due to the assignable causes and variation within the samples is due to chance causes. Variation due to assignable causes is called Treatment Variations and variation due to chance causes are known as Random Variation or Error Variation.

The causes or factors whose effects can be measured and controlled by the researcher are called Assignable Causes and the causes or factors whose effects are beyond the human control are called Chance Causes.

10.3 Assumptions of ANOVA

The following are the underlying assumptions in the use of ANOVA technique.

1. Normality

The population from which the various samples are selected are normally distributed.

2. Homogeneity

The populations from which the samples are drawn have the same variance.

3. Independence

The samples are independently drawn.

4. Additivity

The effects of various components are additive.

10.4 One-Way ANOVA

In a research one can apply any number of treatments in a single plot. For example, suppose we have 12 showrooms for a company. Three different kinds of training programs are conducted to four showrooms each. Four different advertisements are given to three showrooms each as shown below.

	Training – A	Training – B	Training – C
Advertisement – I	X_{11}	X_{12}	X_{13}
Advertisement – II	X_{21}	X_{22}	X_{23}
Advertisement – III	X_{31}	X_{32}	X_{33}
Advertisement – IV	X_{41}	X_{42}	X_{43}

The increase in sales in the showrooms may be due to the effect of advertisements or due to the effect of training programs conducted to the staff or due to the effect of both. We can study whether the advertisement patterns are equally efficient or not and whether the various training programs are equally efficient or not, using ANOVA.

The methods of Analysis of Variance were developed by R. A. Fisher in the early 1920s.

Similarly we can apply three treatments or sometimes only one treatment for a plot. If we apply only one treatment to a plot, then it is called One-way ANOVA. Analysis of Variance with two treatments or factors at a time is called Two-way ANOVA. In our course we are discussing the procedure for One-way ANOVA only.

In our previous example of packing time taken by machines A, B and C, we are considering the efficiency of machines only. Therefore it is an example for One-way ANOVA. The data is repeated below.

	Time Taken in Seconds					Total	
Machine A	4	3	5	4		**16**	**(T_1)**
Machine B	3	4	4			**11**	**(T_2)**
Machine C	6	8	8	5	7	**34**	**(T_3)**
					Total	**61**	**(G)**

The step by step procedure for One-way ANOVA is given below.

Let we have 'k' treatments and i^{th} treatment has applied on n_i plots. Then the total number of plots = $n_1 + n_2 + n_3 + \ldots\ldots\ldots\ldots + n_k = N$. (In the above example: k=3. $n_1 = 4$, $n_2 = 3$ and $n_3 = 5$. Therefore N= 4 + 3 + 5 = 12.)

Step I

The null hypothesis is

H_0 : The means are equal

Or $H_0 : \mu_1 = \mu_2 = \mu_3 = \ldots\ldots = \mu_k$

against the alternative hypothesis

H_1 : The means are not equal

Step II

1) Compute the Correction Factor, $CF = \frac{G^2}{N}$

   Where G is the grand total.

2) Compute Total Sum of Squares, TSS = sum of squares of all observations – CF.

3) Compute Between Sum of Squares (SSB) or Treatment Sum of Squares (SST)

   $SSB = \frac{\sum T_i^2}{n_i} - CF$ where T_i the sum of observations in the ith sample,

   n_i is the number of observations in the ith sample

4) Compute Within Sum of Squares (SSW) or Error Sum of Squares (SSE)

   $SSW = TSS - SSB$

Stept III

Compute Mean Sum of Squares

$$\text{Mean sum of squares} = \frac{Sum\ of\ squares}{Degrees\ of\ freedom}$$

Here degrees of freedom for TSS = N – 1

degrees of freedom for SSB = k – 1

degrees of freedom for SSW = N – k

Therefore, Mean Between Sum of Squares is, $MSB = \frac{SSB}{k-1}$

Mean Within Sum of Squares is, $MSW = \frac{SSW}{N-k}$

Step IV

Compute F ratio

$$F = \frac{MSB}{MSW}$$

Here MSB and MSW are Chi- Square variates with k-1 and N-k degrees of freedom respectively. Therefore F follows F – distribution with (k-1, N-k) degrees of freedom. Find the value of F_α from the table for F – distributions for a given level of significance α .

Step V

Draw the ANOVA Table.

The format of ANOVA Table is shown below.

ANOVA Table

Source	df	Sum of Squares (SS)	Mean Sum of Squares (MSS)	F	F_α
Between Samples (Treatment)	k - 1	SSB	MSB	$F = \frac{MSB}{MSW}$	F(k-1, N-k)
Within Samples (Error)	N - k	SSW	MSW		
Total	N - 1	TSS			

Step VI

If computed F is greater than F_α (i.e. F > F_α) for a given level of significance α, we reject the null hypothesis, otherwise we accept H_0.

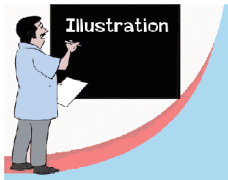


Illustration 9.1

The time taken in seconds by three different packing machines is given below. Test whether the machines are equally efficient or not at 5% level of significance.

	Time Taken (in Seconds)				
Machine A	4	3	5	4	
Machine B	3	4	4		
Machine C	6	8	8	5	7

Solution

Here the null hypothesis is:

$$H_0 : \mu_1 = \mu_2 = \mu_3$$

	Time taken in seconds					Total	
Machine A	4	3	5	4		**16**	**(T_1)**
Machine B	3	4	4			**11**	**(T_2)**
Machine C	6	8	8	5	7	**34**	**(T_3)**
					Total	**61**	**(G)**

Correction Factor, $\text{CF} = \frac{G^2}{N} = \frac{61^2}{12} = 310.08$

Total Sum of Squares,

$$\text{TSS} = \text{sum of squares of all observations} - \text{CF.}$$
$$= (4^2 + 3^2 + 5^2 + 4^2 + 3^2 + 4^2 + 4^2 + 6^2 + 8^2 + 8^2 + 5^2 + 7^2) - 310.08$$
$$= 345 - 310.08$$
$$= 34.92$$

Between Sum of Squares, SSB $= \sum \frac{T_i^2}{n_i} - CF$

$$= \left(\frac{16^2}{4} + \frac{11^2}{3} + \frac{34^2}{5}\right) - 310.833$$
$$= (64 + 40.33 + 231.2) - 310.08$$
$$= 335.53 - 310.08$$
$$= 25.45$$

Within Sum of Squares, SSW $= TSS - SSB$

$$= 34.92 - 25.45$$
$$= 9.47$$

Mean Between Sum of Square, MSB $= \frac{SSB}{K-1}$

$$= \frac{25.45}{2}$$
$$= 12.725$$

Mean Within Sum of Squares, $\text{MSW} = \frac{SSW}{N-K} = \frac{9.47}{9}$

$= 1.05$

$$F = \frac{MSB}{MSW} = \frac{12.725}{1.05}$$

$= 12.12$

ANOVA Table

Source	df	Sum of Squares (SS)	Mean Sum of Squares (MSS)	F	$F_{0.05}$
Between Samples (Treatment)	2	25.45	12.725	12.12	4.26
Within Samples (Error)	9	9.47	1.05		
Total	11	34.92			

Conclusion

Since F > F_{α}, we reject the null hypothesis at 5% level of significance. That is, the treatment effects are significantly different. Or the machines are not equally efficient.

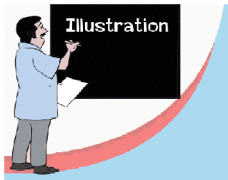


Illustration 9.2

The following table shows the yield obtained when four varieties of fertilisers applied in various identical agricultural plots. Conduct an ANOVA to test whether the efficiencies of the fertilisers are significantly different or not at 1% level of significance.

	Yield						
Fertiliser I	12	14	14	11	13	12	
Fertiliser II	14	17	17	16	18	20	20
Fertiliser III	13	10	12	12			
Fertiliser IV	12	11	11	13	12	12	

Solution

Here the null hypothesis is $H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4$

	Yield							Total
Treatment I	12	14	14	11	13	12		**76**
Treatment II	14	17	17	16	18	20	20	**122**
Treatment III	13	10	12	12				**47**
Treatment IV	12	11	11	13	12	12		**71**
							Total	**316**

Correction Factor, $\text{CF} = \frac{G^2}{N} = \frac{316^2}{23} = 4341.57$

Total Sum of Squares, TSS = sum of squares of all observations – CF.

$$= (12^2 + 14^2 + 14^2 + \ldots + 13^2 + 12^2 + 12^2) - 4341.57$$
$$= 4524 - 4341.57 = 182.43$$

Between Sum of Squares, $\text{SSB} = \sum \frac{T_i^2}{n_i} - CF$

$$= \left(\frac{76^2}{6} + \frac{122^2}{7} + \frac{47^2}{4} + \frac{71^2}{6}\right) - 4341.57$$

$$= 4481.38 - 4341.57 = 139.81$$

Within Sum of Squares, SSW $= TSS - SSB = 182.43 - 139.81 = 42.62$

ean Between Sum of Square, MSB $= \frac{SSB}{K-1} = \frac{139.81}{3} = 46.60$

Mean Within Sum of Squares, $\text{MSW} = \frac{SSW}{N-K} = \frac{42.62}{19} = 2.24$

$$\text{F} = \frac{MSB}{MSW} = \frac{46.60}{2.24}$$

$$= 20.80$$

ANOVA Table

Source	df	Sum of Squares (SS)	Mean Sum of Squares (MSS)	F	$F_{0.01}$
Between Samples (Treatment)	9	139.81	46.60	20.80	3.13
Within Samples (Error)	19	42.62	2.24		
Total	22	182.43			

Conclusion

Since calculated value F > F_{α}, we reject the null hypothesis. We conclude that the treatment effects are significant.

Know your progress

Four different brands of drugs have been developed for the cure of a certain disease. These drugs were tried on 100 patients each at three different Hospitals. The numbers of cases of recovery from the disease are given below. Carry out the analysis of variance to test the effectiveness of drugs and interpret the results.

Hospital	Drugs			
	A	B	C	D
H_1	24	20	24	17
H_2	20	25	30	9
H_3	13	18	31	13

Illustration 9.3

The incomplete ANOVA table of a study is given below. Complete the fields of the table and interpret the result.

Source	df	Sum of Squares (SS)	Mean Sum of Squares (MSS)	F	$F_{0.01}$
Between Samples (Treatment)	5	-	12	-	-
Within Samples (Error)	-	76	-		
Total	24	-			

Solution

Here the null hypothesis is $H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$

Given,

$k - 1 = 5$, $N - 1 = 24$, $SSW = 76$, and $MSB = 12$

Degrees of freedom for between sum of squares $= 5$

$$k - 1 = 5$$

$$k = 5 + 1 = 6$$

Therefore number of treatments = 6

$$N - 1 = 24$$

$$N = 24 + 1 = 25$$

Degrees of freedom within sum of squares $= N - k = 25 - 6 = 19$

$$MSB = \frac{SSB}{k-1}$$

$$12 = \frac{SSB}{5}$$

$$SSB = 12 \times 5 = 60$$

$$TSS = SSB + SSW = 60 + 76 = 136$$

$$MSW = \frac{SSW}{N-k} = \frac{76}{19} = 4$$

$$F = \frac{MSB}{MSW} = \frac{12}{4} = 3$$

The ANOVA Table is

Source	df	Sum of Squares (SS)	Mean Sum of Squares (MSS)	F	$F_{0.01}$
Between Samples (Treatment)	5	60	12	3	4.17
Within Samples (Error)	19	76	4		
Total	24	136			

Conclusion

Here $F < F_{\alpha}$ at 1% level of significance. Therefore we accept the null hypothesis.

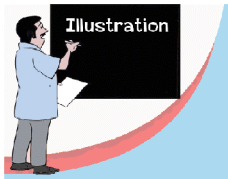


Illustration 9.4

The number of crimes reported per week due to the excess use of alcohol at five major cities were analysed during a month and the following ANOVA table was obtained. Complete the table and write the inference.

Source	df	Sum of Squares (SS)	Mean Sum of Squares (MSS)	F	$F_{0.01}$
Between Samples (Treatment)	-	-	-	-	-
Within Samples (Error)	-	-	4.5		
Total	15	193.5			

Solution

The null hypothesis is

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$$

Given $k = 5$,

Therefore $k - 1 = 5 - 1 = 4$

Form the table $N - 1 = 15$

Therefore $N = 15 + 1 = 16$

$MSW = 4.5$

$$\frac{SSW}{N-k} = 4.5$$

$$\frac{SSW}{16-5} = 4.5$$

$$\frac{SSW}{11} = 4.5$$

$$SSW = 11 \times 4.5$$

$$SSW = 49.5$$

From the Table, $TSS = 193.5$

$$SSB + SSW = 193.5$$

$$SSB = 193.5 - SSW = 193.5 - 49.5 = 144$$

$$MSB = \frac{SSB}{k-1} = \frac{144}{5-1} = \frac{144}{4} = 36$$

$$F = \frac{MSB}{MSW} = \frac{36}{4.5} = 8$$

Source	df	Sum of Squares	Mean Sum of Squares	F	$F_{0.05}$
Between Samples (Treatment)	4	144	36	8	3.36
Within Samples	11	49.5	4.5		
Total	15	193.5			

Since the calculated value of F is 8, which is greater than the table value of F, we reject the null hypothesis. So we can conclude that the number of crimes reported at various cities are significantly different.

Know your progress

An agency studied the rapid increase in the number of orphans at old age homes at different states of the country. Complete the ANOVA table and write your conclusion.

Source	df	Sum of Squares (SS)	Mean Sum of Squares (MSS)	F	$F_{0.05}$
Between Samples (Treatment)	-	-	-	-	-
Within Samples (Error)	19	-	5.2		
Total	26	553.8			

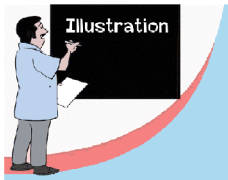


Illustration 9.5

A researcher analysed the efficiency of six management strategies by applying them in 23 fields. The mean sum of squares between samples and mean sum of squares within samples are 35.2 and 19.2 respectively. Compute the F ratio and draw your inference.

Solution

Given

$$MSB = 35.2,\ MSW = 19.2,\ k = 6 \text{ and } N = 23$$

$$F = \frac{MSB}{MSW} = \frac{35.2}{19.2} = 1.83$$

$$F_\alpha = 2.81$$

Here $F < F_\alpha$ so we cannot reject the null hypothesis. Or we cannot consider that the effects of the management strategies are significantly different.

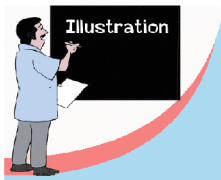


Illustration 9.6

To study the significance of difference of the efficiencies of four verities of seeds, an analysis of variance is conducted at 14 plots. The sum of squares between samples obtained is 173 and the sum of squares within samples is 123. Compute the F ratio and write the inference.

Given $SSB = 173$, $SSW = 123$, $k = 4$ and $N = 14$

$$MSB = \frac{SSB}{k-1} = \frac{173}{4-1} = \frac{173}{3} = 57.67$$

$$MSW = \frac{SSW}{N-k} = \frac{123}{14-4} = \frac{123}{10} = 12.3$$

$$F = \frac{MSB}{MSW} = \frac{57.67}{12.3} = 4.69$$

From the table for F-Distributions $F_\alpha = 3.71$

We can see $F > F_\alpha$. We reject the null hypothesis. That means the efficiencies of the seeds are not same.

Let us conclude

ANOVA is a statistical technique which is used to test the significance of difference of several means. In ANOVA the variations between the samples and variation within the samples are considered to arrive at a conclusion. There may be many causes for the variations. The causes which are measurable and controllable are called assignable causes and others are called chance causes. The ratio of the variations follows Snecors F – Distribution. If the ratio is more than a particular value, we reject the null hypothesis. ANOVA has a wide use in Agriculture, Business, Education, Medical Science, Psychology, Social Science, etc.

Let us assess

For Questions 1-8, choose the correct answer from the given choices.

1. Analysis of variance is a technique used to test the.....

   a) Variance b) Mean c) Covariance d) Correlation

2. The test statistic used in ANOVA is.....

   a) Z b) t c) χ^2 d) F

3. The variations due to assignable causes are called:

   a) Random Variation b) Treatment Variation

   c) Chance Variation d) None of these

4. In an ANOVA table SSB = 190 and MSB = 95, the number of treatments is.....

   a) 2 b) 5 c) 3 d) 4

5. If the grand total of 25 observations arraged in 5 rows and 5 columns is 100. The correction factor will be .....

   a) 100 b) 50 c) 200 d) 400

6. If in an ANOVA, MSB=4, MSE 2, then the variance ratio F=----

   a) 2 b) 1/2 c) 8 d) 4

7. ANOVA is a technique used for testing the.....

   a) Equality of two variances b) Equality of two variances
   c) Equality of two means d) Equality of several means

8. In an ANOVA, treatment sum of squares = 30, No. of treatments= 3, find the mean sum of squares due to treatments?

   a) 90 b)10 c)15 d) 30

9. Write the uses of ANOVA.

10. Distinguish between assignable causes and chance causes.

11. Write the important assumptions underlying ANOVA.

12. In an analysis of variance the mean sum of squares due to the treatment and error are 23.1 and 6.34 respectively. Whether the treatment effects are significant? [Take α = 0.05, df = (4, 17)]

13. Five treatments are applied at 24 plots and the following values are obtained.

    Sum of squares between samples = 42

    Total sum squares = 108.

    Prepare an ANOVA table and write your conclusion.

14. Four medical treatments are examined for its efficiencies at 18 patients. The sum of squares due to treatments and the sum of squares due to error obtained are 67 and 45 respectively. Compute the F ratio and state your conclusion at 5 percent level of significance.

15. An incomplete ANOVA table is given below. Complete the table and draw your inference.

Source	df	Sum of Squares (SS)	Mean Sum of Squares (MSS)	F	$F_{0.01}$
Between Samples (Treatment)	-	63	-	2.4	-
Within Samples (Error)	8	-	-		
Total	11	-			

16. A marketing manager practiced five marketing strategies and analysed the efficiency by using ANOVA. Complete the table and write the inference.

Source	df	Sum of Squares (SS)	Mean Sum of Squares (MSS)	F	$F_{0.05}$
Between Samples (Treatment)	-	54	-	-	-
Within Samples (Error)	13	-	5.4		
Total	-	-			

17. A researcher conducted an experiment to study the efficiencies of six fertilisers. He applied different fertilisers at 22 fields and an ANOVA is conducted. The incomplete ANOVA table is given below. Complete the table and write the inference.

Source	df	Sum of Squares (SS)	Mean Sum of Squares (MSS)	F	$F_{0.05}$
Between Samples (Treatment)	120	-	-	-	-
Within Samples (Error)	-	-	3.25		
Total	-	-			

18. The number of items sold at various outlets of a company is shown below. Conduct an ANOVA and test whether the sales at the outlets are significantly different or not at 5% level of significance.

	Yield						
Outlet I	10	11	11	14			
Outlet II	11	13	10	9	9	11	
Outlet III	14	14	15	14			
Outlet IV	11	13	8	8	12	12	13

19. Four printing machines were installed in a company. The manager inspected the number of copies printed by each machine per ten seconds. It is given below. Can you say whether the machines are equally efficient at 1 % level of significance?

	Yield					
Machine A	21	22	22	24	20	
Machine B	16	17	16	19	15	20
Machine C	18	20	16	14		
Machine D	17	18	20	20	21	

Chapter 11

Statistical Quality Control

All companies, whether they manufacture products or provide services understand that quality is essential for survival in the global economy. Quality has an impact on almost all activities. For example, when we build a house, in the selection process of electrical goods or sanitary item we look for quality products. Some other areas of interest include the design, production, and reliability of our automobiles; services provided by hotels, banks, schools, retailers, and telecommunication companies.

Significant Learning Outcomes

After the completion of this chapter, the learner:

- Identifies the definition of quality, statistical quality control and statistical process control.
- Distinguishes between chance causes and assignable causes.
- Explains the concept of control charts for variables and attributes.
- Constructs control charts for variables and attributes.

11.1 Meaning of Quality

Dr. Walter A. Shewhart (1891 -1967), called the father of quality control analysis, developed the concepts of statistical quality control(SQC).

Quality has become one of the most important consumer decision factors in the selection among competing products and services. A simple answer to questions such as "What is quality?" or "What is quality improvement?" are not easy. The traditional definition of quality is based on the viewpoint that products and services must meet the requirements of those who use them. We now consider two definitions of quality.

> Quality means fitness for use.

There are two general aspects of fitness for use: *quality of design* and quality of conformance. All goods and services are produced in various grades or levels of quality. These variations in grades or levels of quality are intentional, and, consequently, the appropriate technical term is *quality of design*. For example, all automobiles have as their basic objective providing safe transportation for the consumer. However, automobiles differ with respect to size, appearance, and performance. These differences are the result of intentional design differences among the types of automobiles.

The *quality of conformance* is how well the product conforms to the specifications required by the design.

Quality of conformance is influenced by a number of factors, including

- The training and supervision of the workforce
- The types of process controls
- Tests and inspection activities that are employed
- The extent to which these procedures are followed
- The motivation of the workforce to achieve quality

> Quality is inversely proportional to variability.

This definition implies that if variability in the important characteristics of a product decreases the quality of the product increases. Every product possesses a number of elements that jointly describe what the user or consumer thinks of as quality. These parameters are often called quality characteristics. Quality characteristics may be of several types:

- **Physical**: Examples are length, weight, voltage, viscosity
- **Sensory**: Examples are taste, appearance, color
- **Time Orientated**: Examples are reliability, durability, serviceability

11.2 Quality Control

SQC methods can be applied to two phases of manufacturing process.

1. A control to be maintained during the process of manufacturing of articles, which is called *process control*. This is done through the inspection of samples collected at regular intervals of production process.

2. Checking of the quality of the manufactured product in respect of its acceptability. This is carried out through the inspection of sample items, selected randomly from the lot under consideration. Such sampling inspection plans are often termed as *product control* or *acceptance sampling plans*.

For the purpose of process control, Shewhart developed charting techniques and statistical procedures for controlling in-process manufacturing operations. In this chapter we will learn how to develop and analyze control charts, a statistical tool that is widely used for quality improvement.

11.3 Statistical Process Control

If a product is to meet or exceed customer expectations, generally it should be produced by a process that is stable. That is, the process must be capable of operating with little variability around the target or nominal values of the products quality characteristics. Statistical process control (SPC) is a powerful collection of problem-solving tools useful in achieving process stability and improving capability. Shewhart control chart is a graph that show whether a sample of data falls within the chance or normal range of variation.

11.4 Variation and Causes of Variation

A process is the value-added transformation of inputs to outputs. The inputs and outputs of a process can involve machines, materials, methods, measurement, people, and the environment. Each of the inputs is a source of variation. Variability in the output can result in poor service and poor product quality, both of which often decrease customer satisfaction. Variation in the production process leads to quality defects and lack of product consistency. Wise manufacturers understand this. Therefore, they introduce programs those reduces the variability at all manufacturing facilities.

Now let's look at the different types of variation. If you look at bottles of a soft drink in a grocery store, you will notice that no two bottles are filled to exactly the same level. Some are filled slightly higher and some slightly lower.

Chance Causes of Variation

In any production process, regardless of how well designed or carefully maintained it is, a certain amount of inherent or natural variability will always exist. This natural variability is the cumulative effect of many small, inherent and essentially unavoidable causes. In the framework of statistical quality control, this natural variability is due to *chance causes* or *random causes*.

A process that is operating with only chance causes of variation is said to be in *statistical control*.

For example, if the average bottle of a soft drink called Neera contains 300 ml of liquid, we may determine that the amount of natural variation is between 295 and 305 ml. If this were the case, we would monitor the production process to make sure that the amount stays within this range. If production goes out of this range (say, bottles are found to contain on average 290 ml) this would lead us to believe that there is a problem with the process because the variation is greater than the natural random variation.

Assignable Causes of Variation

Other kinds of variability may occasionally be present in the output of a process. This variability in key quality characteristics usually arises from three sources:

1. Improperly adjusted or controlled machines
2. Operator errors
3. Defective raw material

Such variability is generally large when compared to the natural variability, and it usually represents an unacceptable level of process performance. We refer to these sources of variability that are not part of the chance causes, as *assignable causes* of variation. When these types of variations are observed the causes can be identified and eliminated. A process that is operating in the presence of assignable causes is said to be *out of control*.

In the example of the soft drink bottling operation, bottles filled with 290 ml of liquid would signal a problem. The machine may need to be readjusted. This would be an assignable cause of variation. We can assign the variation to a particular cause (machine needs to be readjusted) and we can correct the problem (readjust the machine).

Chance Variation: Variation that is random in nature. This type of variation cannot be completely eliminated unless there is a majour change in the equipment or material used in the process.

Assignable Variation: Variation that is not random. It can be eliminated or reduced by investigating the problem and finding the cause. as improperly adjusted or controlled machines, operator errors, defective raw materials, it is not tolerable and does affect the quality and utility of the product or article. Such causes of variation are called assignable causes of variation.

A brief comparison between chance and assignable causes of variations are given in the following table:

Chance Causes	Assignable Causes
1. Consist of many individual causes	1. Consists of one or just a few individual causes
2. Any one chance cause results in only a minute amount of variation.	2. Any one assignable cause can result in a large amount of variation.
3. Some typical chance causes of variation are a. Slight variation in raw material b. Slight variation of machine c. Lack of human perfection in reading instruments and setting controls	3. Some typical assignable causes of variation are a. Batch of defective raw materials b. Faulty setup of an untrained operator
4. Chance variation cannot be economically eliminated from the process	4. The presence of assignable causes can be detected and action can be taken to eliminate.

Thus statistical quality control (SQC) can be defined as:

A method of monitoring, controlling and improving a process through statistical analysis. Its basic steps include:

1. Measuring the process.
2. Eliminating variances in the process to make it consistent.
3. Monitoring the process.
4. Improving the process to its best target value.

11.5 Control Charts

Control Charts have been in existence for more than 90 years. Walter A. Shewart is credited with developing control chart at Bell Laboratories in the 1920s. Control Charts are used to monitor the production process to make sure that the production stays within the normal range and is functioning without any assignable causes of variations. That is, we want to make sure the process is in a state of control. Different types of control charts are used to monitor different aspects of the production process. We will learn how to develop and use control charts.

A control chart (also called process chart or quality control chart) is a graph that shows whether a sample of data from the process falls within the natural or normal range of variation. A control chart has upper control limit(UCL) and lower control limit(LCL) that separate chance causes from assignable causes of variation. The center line(CL) is reference line in the control chart which represents no variation at all. We say that a process is out of control, when a plot of data reveals that one or more samples fall outside the control limits.

A Specimen of Control Chart is given bellow.

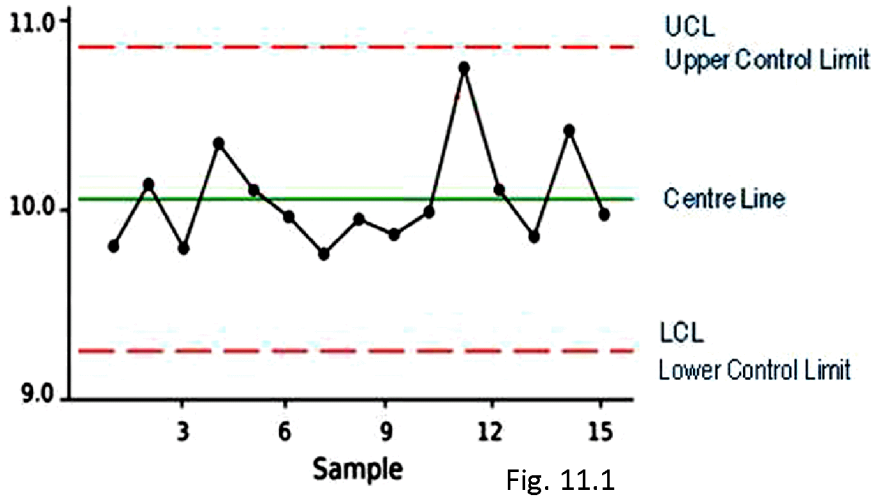


Fig. 11.1

11.6 Types of Control Charts

Control charts can be used to monitor any characteristic of a product, such as the weight of a cereal box, the number of chocolates in a box, or the volume of bottled soft drink. The different characteristics that can be monitored by control charts can be divided into two groups: variables and attributes.

Control Chart for Variables

A *control chart for variables* is used to monitor characteristics that can be measured and have a continuum of values, such as height, weight, or volume. A soft drink bottling operation is an example of a variable measure, since the amount of liquid in the bottles is measured and can take on a number of different values. Other examples are the weight of a bag of sugar, the temperature of a baking oven, or the diameter of ball bearing. Two most commonly used control charts for variables are those that monitor the central tendency of the data (the mean chart or $\bar{x}$ - chart) and the variability of the data (Range chart or *R-chart*).

Control Chart for Attributes

A *control chart for attributes*, on the other hand, is used to monitor characteristics that have discrete values and can be counted. Often they can be evaluated with a simple "yes" or " no" decision. Examples include color, taste, or smell. The monitoring of attributes usually takes less time than that of variables because a variable needs to be measured (e.g., the bottle of soft drink contains 298 ml of liquid).

An attribute requires only a single decision, such as yes or no, good or bad, acceptable or unacceptable. For examples

1. The apple is good or rotten
2. The meat is good or bad
3. The shoes have a defect or do not have a defect
4. The light bulb works or it does not work

We can count the number of defectives in a lot of items. The charts used to monitor these types of attributes are *np*-chart and *p*-chart.

Next we look at how various control charts are developed.

11.7 Construction of Control Charts

To construct a control chart, we collect samples from the output of a process over time. The samples used for constructing control charts are known as subgroups. For each subgroup (i.e., sample), calculate a sample statistic say t. Commonly used statistics include the sample proportion of defective items , the number of defects , and the mean and range of a subgroup. We then plot the values over time and add control limits on both sides of the center line of the chart. The most typical form of a control chart sets control limits that are within 3 standard deviations of the statistical measure of interest.

Such control limits are called Shewhart's control limits.

If t is the charting statistic, then the Center Line(CL), Upper Control Line (UCL) and Lower Control Line(LCL) are constructed by

$$CL = \mu_t$$

$$UCL = \mu_t + 3\sigma_t$$

$$LCL = \mu_t - 3\sigma_t$$

where μ_t and σ_t are the mean and standard deviation of the charting statistic t.

When observed values of the charting statistic t go outside the control limits or when the plots show an identifiable non-random pattern, the process is assumed not to be in control(that is out of control). Production is stopped, and employees attempt to identify the cause of the problem and correct it.

11.8 Control charts for Variables

In this section we discuss two control charts, namely $\bar{x}$ - chart and R - chart, which are used simultaneously.

$\bar{x}$ - chart

A mean control chart is often referred to as an x-bar chart. It is used to monitor changes in the average of a process. To construct a mean chart we first need to construct the center line of the chart. To do this we take multiple samples and compute their means. Usually these samples are small, about four or five observations. Each sample has its own mean. The sample mean is the charting statistic. The center line of the chart is then computed as the mean of all m sample means, where m is the number of samples taken at regular intervals of the process. Therefore the center line (CL) is computed by

$$CL = \bar{\bar{x}} = \frac{\bar{x}_1 + \bar{x}_2 + \cdots + \bar{x}_m}{m} = \frac{\sum \bar{x}}{m}$$

To construct the upper and lower control limits of the charting statistic from m samples we use the following formule:

$$UCL = \bar{\bar{x}} + A_2\bar{R}, \quad LCL = \bar{\bar{x}} - A_2\bar{R}, \text{ where}$$

$\bar{R} = \frac{R_1 + R_2 + \cdots + R_m}{m}$, the mean of m sample ranges and A_2 is a constant that can be read from standard quality control tables. The value of A_2 depends the sample size n which is same for all m samples.

The control limits and centre line are drawn on graph paper. Plot the value for the sample mean (along the vertical axis) against the samples (along the horizontal axis). If any points plot beyond the control limits we say that the process is out-of-control with respect to the process average.

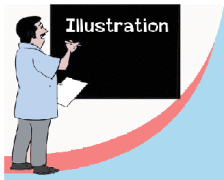


Illustration 11.1

A food company puts mango juice into cans so as to contain 10 ounces of juice. Weights of juice are observed from juice drained from cans immediately after filling. 20 samples are taken by a random method (at an interval of every 30 minutes). Each of the samples includes 4 cans. The sample weights are tabulated in the following table. The weights in the table are given in units of 0.01 ounces in excess of 10 ounces. For example, the juice drained from the first can of the sample is 10.15 ounces which is excess 0.15 is represented as 15 units. Construct x- bar to control the weights of mango juice for the filling.

Sample	Weight of each can (4 cans in each sample, n=4)			
1	15	12	13	20
2	10	8	8	14
3	8	15	17	10
4	12	17	11	12
5	18	13	15	4
6	20	16	14	20
7	15	19	23	17
8	13	23	14	16
9	9	8	18	5
10	6	10	24	20
11	5	12	20	15
12	3	15	18	18
13	6	18	12	10
14	12	9	15	18
15	15	15	6	16
16	18	17	8	15
17	13	16	5	4
18	10	20	8	10
19	5	15	10	12
20	6	14	12	14

Solution

We prepare a computation table for sample means and sample ranges for 20 samples.

Sample	Weight of each can				Total	Sample mean	Sample Range
1	15	12	13	20	60	15.00	8
2	10	8	8	14	40	10.00	6
3	8	15	17	10	50	12.50	9
4	12	17	11	12	52	13.00	6
5	18	13	15	4	50	12.50	14
6	20	16	14	20	70	17.50	6
7	15	19	23	17	74	18.50	8
8	13	23	14	16	66	16.50	10
9	9	8	18	5	40	10.00	13
10	6	10	24	20	60	15.00	18
11	5	12	20	15	52	13.00	15
12	3	15	18	18	54	13.50	15
13	6	18	12	10	46	11.50	12
14	12	9	15	18	54	13.50	9
15	15	15	6	16	52	13.00	10
16	18	17	8	15	58	14.50	10
17	13	16	5	4	38	9.50	12
18	10	20	8	10	48	12.00	12
19	5	15	10	12	42	10.50	10
20	6	14	12	14	46	11.50	8
Total						**263.00**	**211**

$$\bar{\bar{x}} = \frac{263}{20} = 13.15$$

$$\bar{R} = \frac{211}{20} = 10.55$$

The table value of A_2 for n = 4 is 0.729

UCL = 13.15 +(0.729 x 10.55) =13.15 +7.69 = 20.84

LCL = 13.15 – 7.69 = 5.46

X –bar chart

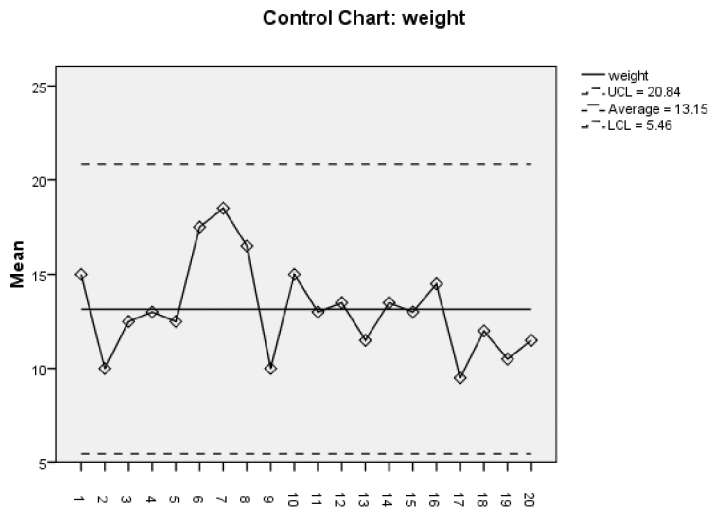


Fig. 11.2

Since all the points are falling within control limits the process is in a state of control.

Range chart

The range chart is also called *R*-chart. It is used to monitor changes inthe dispersion of a process. To construct *R*- chart chart we take multiple samples and compute the range of each sample. The sample range (*R*) is the charting statistic. The center line of the chart is then computed as the range of all *m* sample ranges. When the true process dispersion is not specified and is to be estimated from the sample data, the center line, upper control limit and lower control limit are computed using the following formulae:

$$CL = \overline{R} = \frac{R_1 + R_2 + \cdots + R_m}{m}$$

$$UCL = D_4 \overline{R}$$

$$LCL = D_3 \overline{R}$$

The values of D_3 and D_4 can be read from tables and are constants depend on the sample size *n*.

The control limits and center line are drawn on graph paper. Plot the value for the sample range R_i (along the vertical axis) against the ith sample (along the horizontal axis). If any points plot beyond the control limits we say that the process is out-of-control with respect to the process variability.

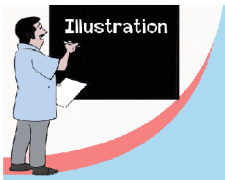


Illustration 11.2

Draw the R chart for the data given in Illustration 11.1.

Solution

From the computation table of illustration 11.1, we compute the mean of the sample ranges,

$$CL = \overline{R} = \frac{R_1 + R_2 + \cdots + R_m}{m} = 10.55$$

$$UCL = D_4\overline{R} = 2.282 \text{ x } 10.55 = 24.08$$

$$LCL = D_3\overline{R} = 0 \text{ x } 10.55 = 0$$

Then R- chart is drawn as follows:

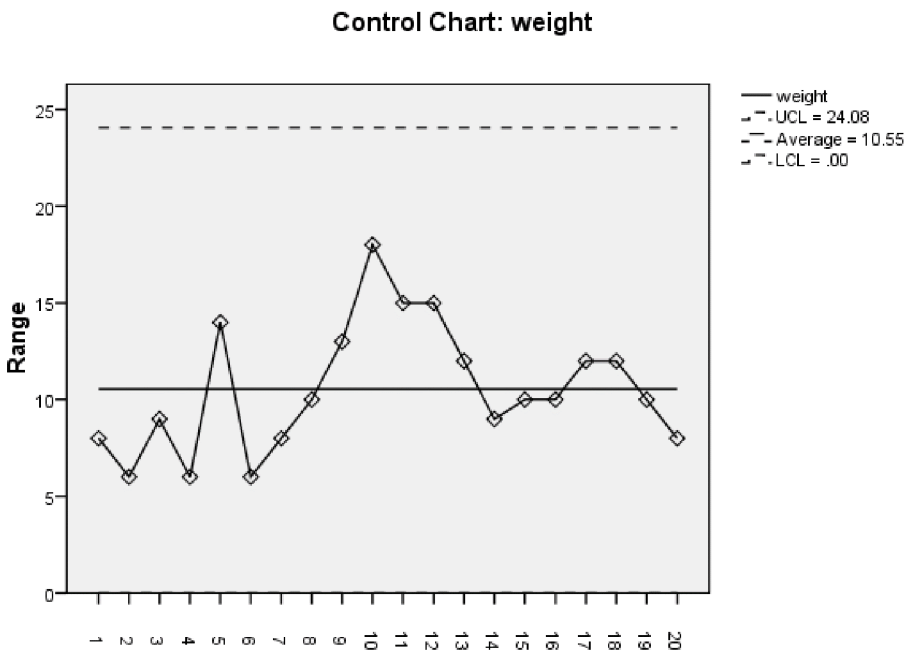


Fig. 11.3

The Chart shows that the process is under control with respect to the process variability.

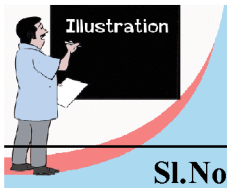


Illustration 11.3

Draw a control chart for x bar and R from the following data relating to 20 samples, each of size 4.

Sl.No	x-bar	R	Sl.No	x-bar	R
1	15.0	8	11	13.00	15
2	10.0	6	12	13.50	15
3	12.5	6	13	11.50	15
4	13.0	6	14	13.50	6
5	12.5	14	15	13.00	9
6	17.5	6	16	14.50	10
7	18.5	8	17	9.50	12
8	16.5	10	18	12.00	12
9	10.0	13	19	10.50	10
10	15.0	18	20	11.50	8

Solution:

Prepare the computation table for sample statistics as follows.

Sl.No	x-bar	R
1	15.0	8
2	10.0	6
3	12.5	6
4	13.0	6
5	12.5	14
6	17.5	6
7	18.5	8
8	16.5	10
9	10.0	13
10	15.0	18
11	13.00	15
12	13.50	15
13	11.50	15
14	13.50	6
15	13.00	9

16	14.50	10
17	9.50	12
18	12.00	12
19	10.50	10
20	11.50	8
	263.0	**207**

Limits for $\bar{x}$ - Chart

Mean of x-bar = $\frac{263}{20}$ = 13.5

R bar = $\frac{207}{20}$ = 10.35

UCL = $\bar{\bar{x}} + A_2\bar{R}$ = 13.5+ 0.729 x 10.35 = 13.5 + 7.545 =21.04

LCL = $\bar{\bar{x}} - A_2\bar{R}$ = 13.5- 7.545 = 5.955

Limits of R chart

R bar = 10.35

UCL = $D_4\bar{R}$ = 2.282 x 10.35 = 23.6187

LCL = $D_3\bar{R}$ = 0 x 10.35 = 0

X bar chart and R chart are drawn bellow:

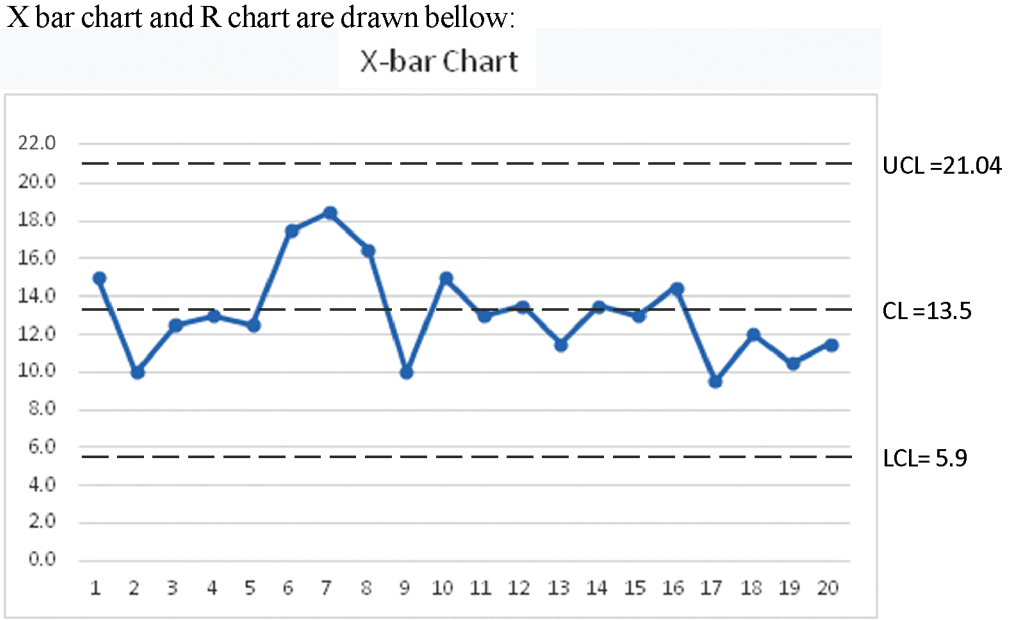


Fig. 11.4

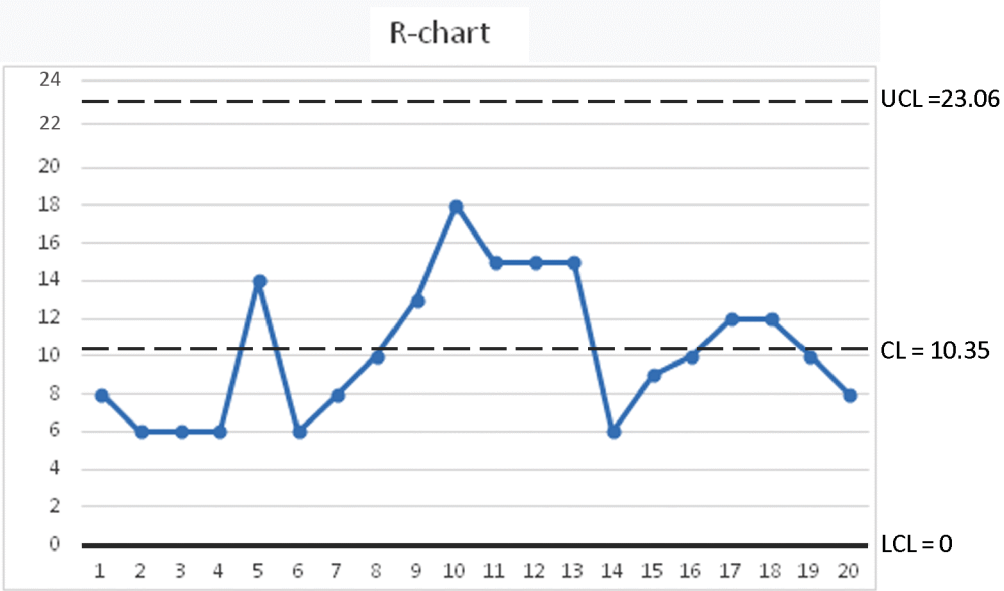


Fig. 11.5

Here the process in control with respect to process average and process variability.

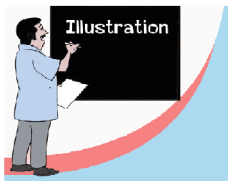


Illustration.11.4

Construct a control chart for mean and range for the following data on the basis of fuses, samples of 5 being taken every hour(each set of 5 has been arranged in ascending order of magnitude. Comment on whether the production seems to be under control.

Values for 12 samples

1	2	3	4	5	6	7	8	9	10	11	12
42	42	19	36	42	51	60	18	15	69	64	61
65	45	24	54	51	74	60	20	30	109	90	78
75	68	80	69	57	75	72	27	39	113	93	94
78	72	81	77	59	78	95	42	62	118	109	109
87	90	81	84	78	132	138	60	84	153	112	136

Solution:

The reader can easily prepare a computation table for sample means and sample ranges for 12 samples.

$$\bar{\bar{x}} = \frac{859.2}{12} = 71.6$$

$$\bar{R} = \frac{716}{12} = 59.67$$

From Tables for n=5 we have $A_2 = 0.58$, $D_3 = 0$ and $D_4 = 2.11$

Also compute the center line, upper control limit and lower control limit for mean chart and range chart as follows.

Mean chart:

CL= 71.6

UCL= 106.21

LCL= 36.99

Range chart:

CL = 59.67

UCL = 125.904

LCL = 0

The mean chart and range chart are drawn bellow:

X-bar chart

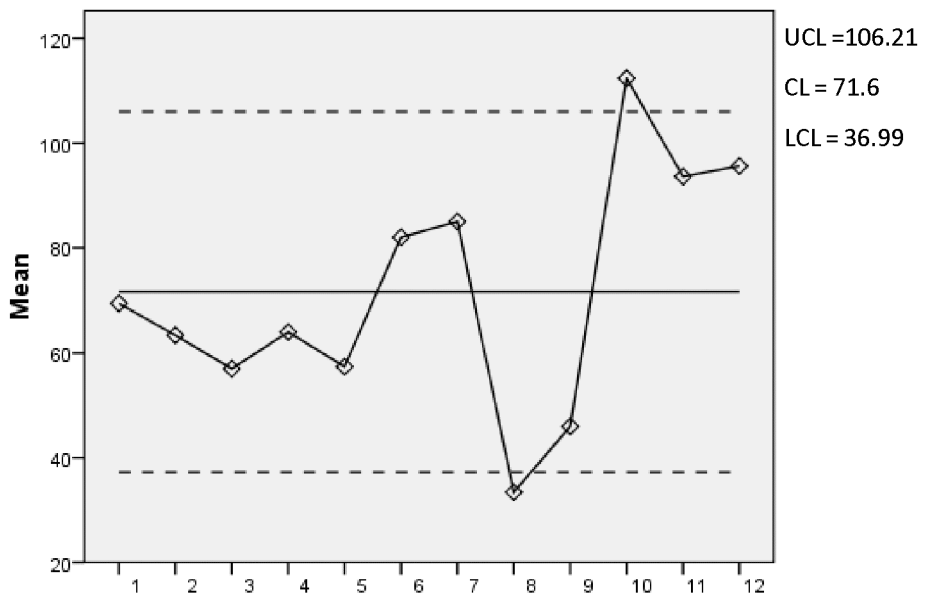


Fig. 11.6

R-chart

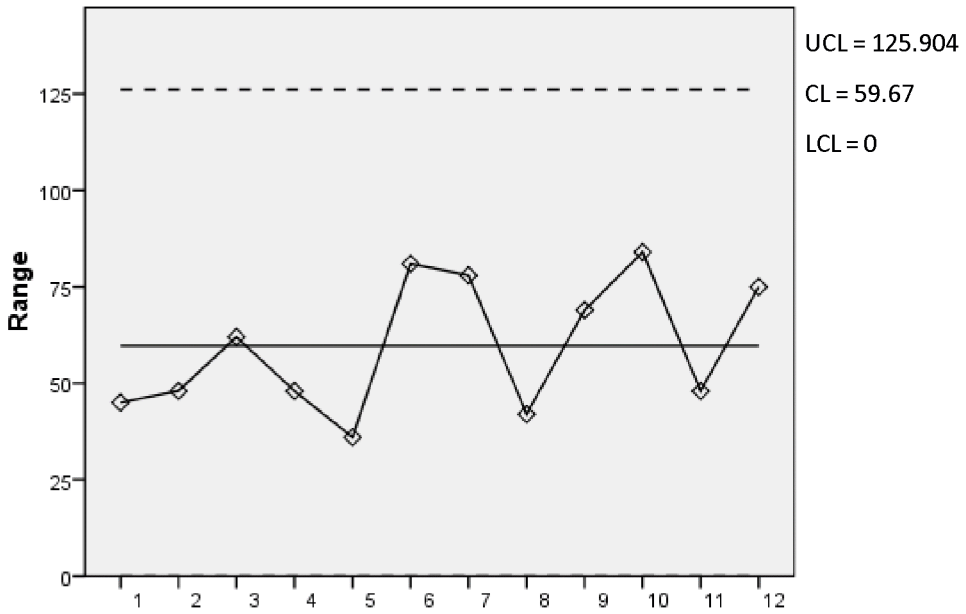


Fig. 11.7

Here the process is out of control regarding process mean since two sample points fall beyond the control limits. However the process variability is in control.

Know your progress

1) Every hour a quality control inspector measures the outside diameter of four parts. The results of the measurements are given below:

	Sample Piece			
Time	1	2	3	4
9 A.M.	1	4	5	2
10 AM	2	3	2	1
11 AM	1	7	3	5

a) Compute the mean outside diameter, mean range and determine the control limits for the mean and the range

b) Are the measurements within the control limits? Interpret the charts.

2) A quality control inspector at a soft drink company has taken twenty-five samples withfour observations each of the volume of bottles filled. Draw x- bar and R chart.

Sample Number	Observations (bottle volume in ounces)			
	1	2	3	4
1	15.85	16.02	15.83	15.93
2	16.12	16.00	15.85	16.01
3	16.00	15.91	15.94	15.83
4	16.2	15.85	15.74	15.93
5	15.74	15.86	16.21	16.10
6	15.94	16.01	16.14	16.03
7	15.75	16.21	16.01	15.86
8	15.82	15.94	16.02	15.94
9	16.04	15.98	15.83	15.98
10	15.64	15.86	15.94	15.89
11	16.11	16	16.01	15.82
12	15.72	15.85	16.12	16.15
13	15.85	15.76	15.74	15.98
14	15.73	15.84	15.96	16.1
15	16.20	16.01	16.10	15.89
16	16.12	16.08	15.83	15.94
17	16.01	15.93	15.81	15.68
18	15.78	16.04	16.11	16.12
19	15.84	15.92	16.05	16.12
20	15.92	16.09	16.12	15.93
21	16.11	16.02	16.00	15.88
22	15.98	15.82	15.89	15.89
23	16.05	15.73	15.73	15.93
24	16.01	16.01	15.89	15.86
25	16.08	15.78	15.92	15.98

11.9 Control Charts for attributes

Attributes are usually judged by the number of defectives or proportion of defectives. A defective item is also called non conforming item. To monitor the process with respect to defectives we use *np*- chart (control chart for number of defectives) or *p*-chart(control chart for proportion of defectives). When attributes are judged by number of defects (or non conformities) per item we use *c*-chart. In the following section we discuss the method of constructing *np*-chart.

Construction of np-chart

We inspect m subgroups(samples)each of n items. Let d_i denotes the number of defective items in the ith subgroup. Then $p_i = \frac{d_i}{n}$ represents the proportion of defectives in the subgroup. The Center Line (CL), Upper Control Limit (UCL) and Lower Control Limit (LCL) are given by

$$CL = n\bar{p}$$

$$UCL = n\bar{p} + 3\sqrt{n\bar{p}\bar{q}}$$

$$LCL = n\bar{p} - 3\sqrt{n\bar{p}\bar{q}}$$

where

$$\bar{p} = \frac{p_1 + p_2 + \cdots + p_m}{m} = \frac{\frac{d_1}{n} + \frac{d_2}{n} + \cdots + \frac{d_i}{n}}{m}$$

$$= \frac{d_1 + d_2 + \cdots + d_m}{mn}$$

The control limits and center lines are drawn on graph paper. Plot the value for the number of defectives (along the vertical axis) against the samples (along the horizontal axis). The decision about the state of process is taken as in the case of $\bar{x}$ - *chart* or *R-chart*.

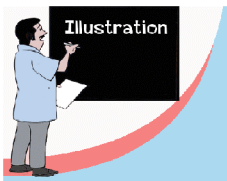


Illustration 11.5

Twenty five boxes, each containing 20 electric switches were randomly selected and inspected for the number of defectives in each box were as follows. Construct control charts for the number of defectives.

Box No	1	2	3	4	5	6	7	8	9	10	11	12	13
Number of Defectives	3	2	1	0	4	2	1	2	3	0	2	1	2
Box No.	14	15	16	17	18	19	20	21	22	23	24	25	
Number of Defectives	0	3	5	4	2	1	3	0	3	1	2	1	

Solution:

Here the sample size of each sample n is 20, and the number of samples inspected m is 25.

There for

$$\bar{p} = \frac{Total\ Number\ of\ Defective}{25 \times 20} = = \frac{48}{500} = 0.096$$

$$\bar{q} = 1 - \bar{p} = 1 - 0.096 = 0.904$$

$$n\bar{p} = 20 \times 0.096 = 1.92$$

$$CL = n\bar{p} = 1.92$$

$$UCL = n\bar{p} + 3\sqrt{n\bar{p}\bar{q}} = 5.87$$

$$LCL = n\bar{p} - 3\sqrt{n\bar{p}\bar{q}} = -2.03 \approx 0$$, since lower limit cannot be negative.

np-chart is drawn below.

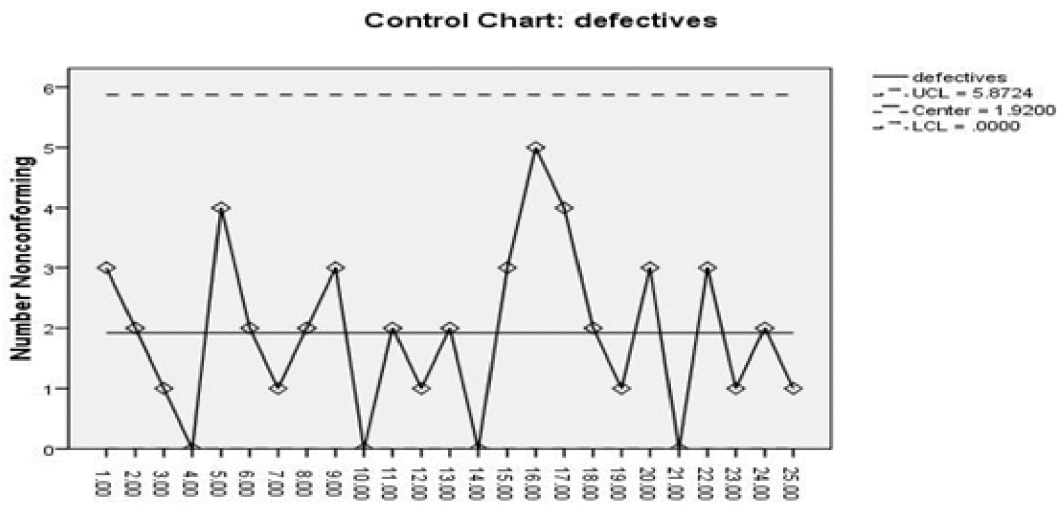


Fig. 11.8

Note that no points plot out of control limits and there for the process is in control with respect to the number of defectives.

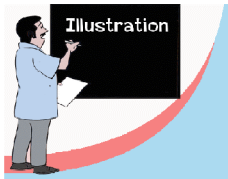


Illustration: 11.6

A company produces bond paper and, at regular intervals, samples of 50 sheets of paper are inspected. Suppose 20 random samples of 50 sheets of paper each are taken during a certain period of time, with the following numbers of sheets in known compliance per sample. Construct np-chart from this data:

Sample	1	2	3	4	5	6	7	8	9	10
Out of compliance	4	3	1	0	5	2	3	1	4	2
Sample	11	12	13	14	15	16	17	18	19	20
Out of compliance	2	6	0	2	1	6	2	3	1	5

Solution:

We plot the number of out of compliances (defectives) against the sample numbers and addind the control limits, we get the np-chart

$CL = n\bar{p}$ = 50 x 0.053 = 2.65

$LCL = n\bar{p} - 3\sqrt{n\bar{p}\bar{q}} = 2.65 - 3\sqrt{50 \times 0.053 \times 0.947} \cong 0$, since the lower limit cannot be negative.

$UCL = n\bar{p} + 3\sqrt{n\bar{p}\bar{q}} = 2.65 + 3\sqrt{50 \times 0.053 \times 0.947} = 7.40$

Then the chart is drawn as follows.

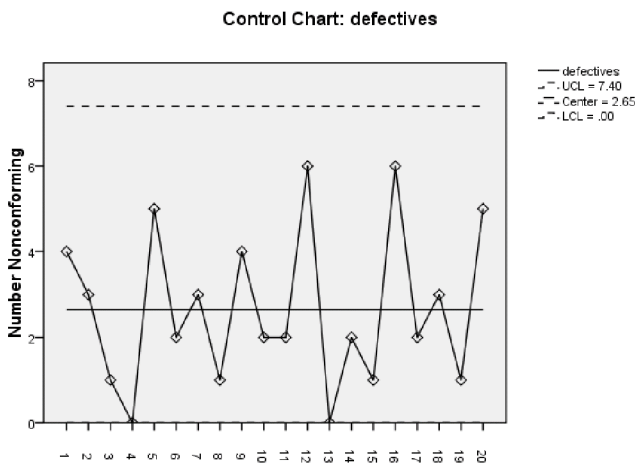


Fig. 11.9

Here all the points fall within the control limits and hence the process is in control.

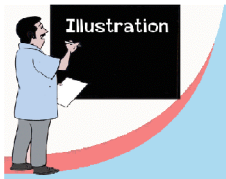


Illustration: 11.7

In a factory producing spark plugs, the number of defectives found in the inspection of 15 lots of 100 each is given below: Draw the control chart for the number of defectives and comment on the state of control.

Sample number (i)	: 1	2	3	4	5	6	7	8	9	10
Number of defective	: 5	10	12	8	6	4	6	3	4	5
Sample number (i)	: 11	12	13	14	15					
Number of defective	: 4	7	9	3	4					

Solution

$$\text{CL} = n\bar{p} = 6$$

$$\text{LCL} = n\bar{p} - 3\sqrt{n\bar{p}(1-\bar{p})}$$

$$= 6 - 3\sqrt{6 \times 0.94} = -1.12 \approx 0$$

Since LCL cannot be negative, LCL = 0

$$\text{LCL} = n\bar{p} + 3\sqrt{n\bar{p}(1-\bar{p})}$$

$$= 6 + 3\sqrt{6 \times 0.94} = -13.12 \approx 0$$

np-Chart

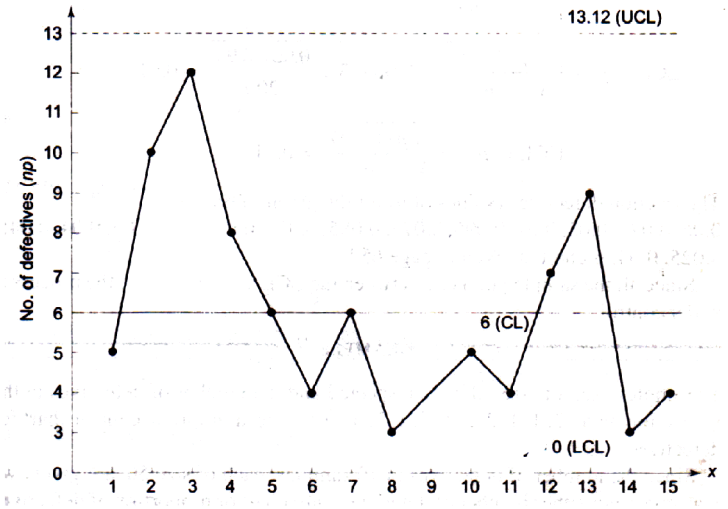


Fig. 11.10

11.10 Uses of Statistical Quality Control

Some uses of statistical Quality Control are listed bellow:

1. It provides means of detecting errors in the process and product at inspection.
2. It leads to more uniform quality of production.
3. It improves the customer relationships.
4. It reduces the cost of inspection.
5. It reduces the number of rejects and saves the cost of materials.
6. It provides a basis for attainable specifications.
7. It estimates the process capability.

Let us conclude

Statistical quality control, also called statistical process control, uses statistics to determine when process or product quality deviate from specifications. It helps to maintain the consistency of a process, which will result in a consistency in the quality as well. The control chart is the tool most often used in statistical process control. The graphic provides a visual representation of data collected through inspections and quality sampling. SQC provides instant feedback when a process goes outside of the process parameters. This allows production to stop and correct the problem before creating a great deal of defective product.

Let us assess

For Questions 1-5, choose the correct answer from the given choices.

1. Which of the following chart types would be used to monitor the average weight of the contents of a box of cereal?

   a) $\bar{x}$ -chart b) R-chart c) p-chart d) c-chart

2. Which of the following are used in the conversion of $\bar{x}$ chart?

   a) A_1 b) A_2 c) D_3 d) D_4

3. Which of the following chart types would be used to monitor the fraction of a production lot of desktops that had scratches on the surface?

   a) $\bar{x}$ -chart b) R-chart c) p-chart d) c-chart

4. The name of control chart used for attributes is _____.

   a) $\bar{x}$ -chart b) R-chart c) np-chart d) c-chart

5. Which of the following is not true?

   a) Quality is the fitness for use.

   b) Quality is not given to services.

   c) Quality is inversely proportional to variability.

   d) Quality is the conformance/ specification of products.

For Questions 6-10, Fill in the blanks

6. The conversion factors used in R chart are _____ and _____

7. The variation in a product which can be tolerable is called _____ variation

8. Slight differences in materials, workers, machines, tools, and other factors are because of _____ variation.

9. Any control chart have _____ number of control limits

10. _____ and _____ charts are used for control chart for variables

11. A Machine is set to deliver the packets of a given weight. Ten samples of size 5 each were examined and the following results were obtained.

Sample No	1	2	3	4	5	6	7	8	9	10
Mean	43	49	37	44	45	37	51	46	43	47
Range	5	6	5	7	7	4	8	6	4	6

   Calculate the values for central line and control limits for the mean chart and range chart. Comment on the state of control.

12. Construct a control chart for mean and range for the following data on the basis of fuses, samples of 5 being taken every hour- each set of 5 has been arranged in ascending order of magnitude.

1	42	65	78	87
2	42	45	72	90
3	10	24	81	81
4	36	54	77	84
5	42	51	59	78
6	51	74	78	132
7	60	60	95	138
8	18	20	42	60
9	15	30	62	84
10	69	109	118	153
11	64	90	109	112
12	61	78	109	136

13. Control charts for $\bar{x}$ and R charts are maintained on the tensile strength in Kg of a certain Yarn. The subgroup size is 5. The values of $\bar{x}$ and R are computed for each subgroup. Average of 25 subgroup is $\Sigma\bar{x} = 514.8$, $\Sigma R = 120$. Compute control limits for $\bar{x}$ and R charts.

14. In order to determine whether or not a production of bronze casting is in control, 20 subgroups of size 6 are taken. The quality characteristic of interest is the weight of the castings and it is found that $\bar{\bar{x}} = 3.126$ gm and $\bar{R} = 0.009$ gm. Setup the control limits for $\bar{x}$ and R charts.

15. The following data shows the values of sample means and Range R for ten sample of size 5 each. Calculate the values for central line and control limits for mean chart and range chart and determine whether the process is in control.

Sample No	1	2	3	4	5	6	7	8	9	10
Mean	11.2	11.8	10.8	11.6	11.0	9.6	10.4	9.6	10.6	10.0
Range R	7	4	8	5	7	4	8	4	7	9

16. A New industrial oven has just been installed at the Chocolate Backery. To develop experience regarding the oven temperature, an inspector reads the temperature at four different places inside the oven each half an hour as given below.

	Reading			
Time	1	2	3	4
8.00 AM	40	50	55	39
8.30 AM	44	42	38	38
9.00 AM	41	45	47	43
9.30 AM	39	39	41	41
10.00 AM	37	42	46	41
10.30 AM	39	40	39	40

a) Based on this initial experience, determine the control limits for the mean tem perature. Draw $\bar{x}$ and R charts.

b) Interpret the chart.

17. The Credit Department at Global National Bank is responsible for entering each tansaction charged to the customer's monthly statement. Each data entry clerk examine a sample of 1500 of their batch and a computer programme checks the numbers match. The results are as follows.

Inspector	Number inspected	Number mismatched
A	1500	4
B	1500	6
C	1500	6
D	1500	2
E	1500	15
F	1500	4
G	1500	4

Construct np chart and interpret the results.

18. The Auto-Lite company manufactures car batteries. At the end of each shift the Quality Assurance Department selects a sample of 8 bateries and test them. The number of defective batteries found over the last 12 shifts is 2, 1, 0, 2, 1, 1, 7, 1, 1, 2, 6 and 1. Construct the control chart for the process and comment on whether the process is in control.

19. A bicycle manufacturer randomly select 10 frames each day and test for defects. The number of defectives frames found over the last 14 days is 3, 2, 1, 3, 2, 2, 8, 2, 0, 3, 5, 2, 0, 4. Construct a control chart for this process and comment on whether the process is in control.

20. The Long Last Tyre company, as part of its inspection process, tests its tyres for tread wear under simulated road conditions. Twenty samples of three tyres each were selected from different shifts over the last month of operation. The tread wear is reported below in hundreadths of an inch.

Sample	TreadWear			Sample	Treadwear		
1	44	41	19	11	11	33	34
2	39	31	21	12	51	34	39
3	38	16	25	13	30	16	30
4	20	33	26	14	22	21	35
5	34	33	36	15	11	28	38
6	28	23	39	16	49	25	36
7	40	15	34	17	20	31	33
8	36	36	34	18	26	18	36
9	32	29	30	19	26	47	26
10	29	38	34	20	34	29	32

a) Determine the control limits for the mean and range.

b) Plot the control charts and interpret it.

21. The Inter Global Moving and Storage Company is setting up a control chart to monitor the proportion of residential moves that result in written complaints due to late delivary, lost items, or damaged items. A sample of 50 moves is selected for each of the last 12 Months. The number of written complaints in each sample is 8, 7, 4, 8, 2, 7, 11, 6, 7, 6, 8 and 12.

a) Determine a chart for number of defectives. Insert the mean percentage defective, UCL and LCL.

b) Interpret the chart. Does it appear that the number of complaints is out of control for any of the months?

22. Draw the mean chart and range chart using the following data relating to 15 samples each of size 5 and comment on the state of control.

X:	65.0	64.6	64.1	68.5	68.4	67.9	65.0	64.6
R:	9.8	9.8	8.4	3.9	7.6	8.7	0.1	9.7
X:	64.1	63.2	62.9	62.4	67.0	66.6	66.1	
R:	7.7	7.5	1.2	9.8	6.4	0.6	6.3	

23. A food company puts mango juice in cans, each of which is advertised to contain 10 ounces of the juice. The weights of the juice drained from cans immediately after filling 20 samples each of 4 cans are taken by random sampling method (at an interval of 30 minutes) and given in the following table in units of 0.01 ounce in excess of 10 ounces. To control the excess weights of mango juice drained while filling, draw the X -chart, R-chart and comment on the nature of control.

Sample No.	1	2	3	4	5	6	7	8	9	10
Weights drained	15	10	8	12	18	20	15	13	9	6
	12	8	15	17	13	16	19	23.	8	10
	13	8	17	11	15	14	23	14	18	24
	20	14	10	12	4	20	17	16	5	20
Sample No.	11	12	13	14	15	16	17	18	19	20
Weights drained	5	3	6	12	15	18	13	10	5	6
	12	15	18	9	15	17	16	20	15	14
	20	18	12	15	6	8	5	8	10	12
	15	18	10	18	16	15	4	10	12	14

24. Fifteen samples each of size 50 were inspected and the number of defectives in the inspection were:

    2, 3,4, 2,3,0, 1,2, 2,3,5,5, 1,2,3

    Draw the control chart for the number of defectives and comment on the state of control.

25. On inspection of 10 samples, each of size 400, the numbers of defective articles were:

    19, 4, 9, 12, 9, 15, 26, 14, 15, 17.

    Draw the np-chart and comment on the state of control.

Chapter 12

Time Series Analysis

One of the most important tasks before economists and businessmen is to identify and estimate the needs of people for the future. For example, a business man may be interested in finding out his likely sales in the year 2020 or as a long term planning in 2030 or 2050. This may help him to adjust his production accordingly and avoid the possibility of either unsold stocks or inadequate production to meet the demand.

Similarly an economist may be interested in estimating the likely

Significant Learning Outcomes

After the completion of this chapter, the learner:

- Identifies Time series.
- Differentiates the components of time series.
- Recognises the uses of time series analysis.
- Estimates future values by using time series analysis.
- Interprets trend lines.

production in the coming year so that the proper planning can be carried out with regard to food supply, jobs for the people, etc. The first step in estimating future needs is gathering information from the past. In this connection one usually looks for statistical data which are collected, observed or recorded at successive intervals of time. Such data are usually referred to as **'Time Series'**. In this chapter we discuss more about time series.

12.1 Time Series

Consider the case of continuous monitoring of a person's heart beat in a hospital. Here time is the most important factor. The heartbeat of the person in regular intervals of time will give a clear picture about the health of the heart. The analysis of time series is a critical factor here. Here we collect quantitative observations in regular intervals of time. So this is an example of a time series.

Time series is the collection of quantitative observations that are evently spaced in time and are measured successively.

The other examples of time series are:

- hourly readings of temperature.
- daily closing price of a company stock.
- monthly rainfall data.

In all these examples, time is the most important factor because the variable is related to time which may be either year, month, week, day, hour or even minutes or seconds.

> A Time Series is a sequence of data observed, collected and arranged in chronological order.

Let us see some more examples:

Example 1

Analyse the graphical representation

- Explain the data given.
- Write the reason for considering this as a time series.

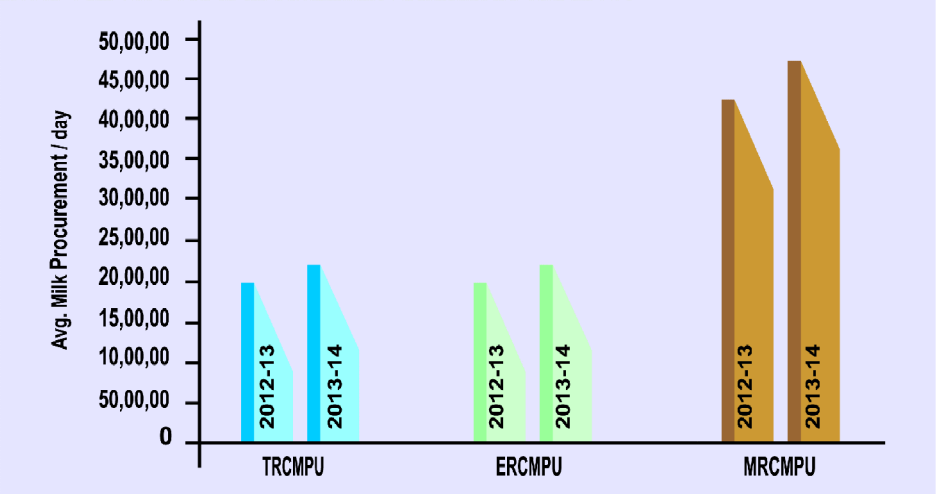


Fig. 12.1 : Average milk procurement in litres per day

Example 2

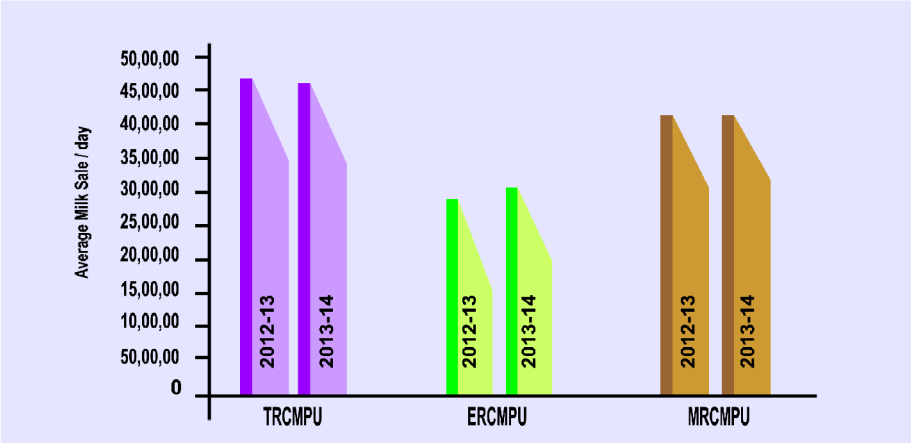


Fig. 12.2 : Average milk sales in litres per day

Regional Unions	2012-13	2013-14	% Diff
TRCMPU	466198	460100	– 1.31
ERCMPU	299385	303664	1.43
MRCMPU	422924	423034	0.03
Total	1188507	1186798	– 0.14

Table 12.1 Milk sales in literes per day

Example 3

Month	Date	Price of 1 Pavan Gold (Rs.)
January	15-Jan-14	21920
February	15-Feb-14	22600
March	15-Mar-14	22680
April	15-Apr-14	22400
May	15-May-14	22320
June	15-Jun-14	20640
July	15-Jul-14	21040
August	15-Aug-14	21480
September	15-Sep-14	20400
October	15-Oct-14	20480
November	15-Nov-14	20000

Table 12.2

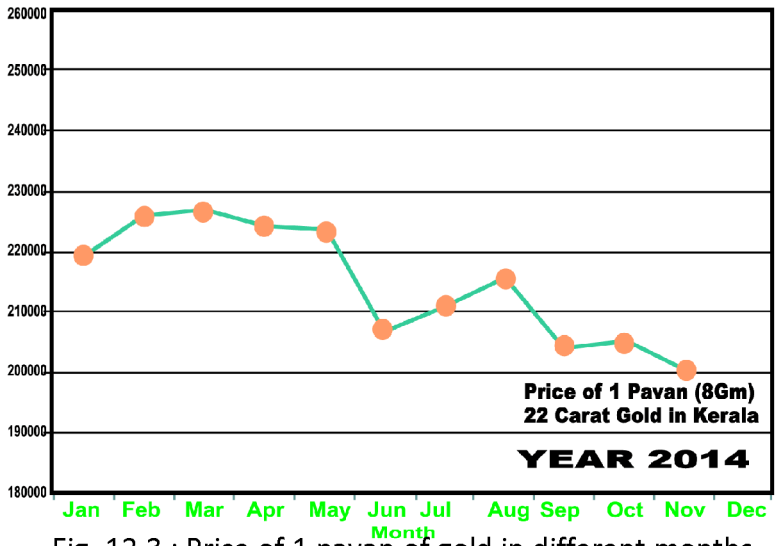


Fig. 12.3 : Price of 1 pavan of gold in different months

Explain why these are considered as time series?

When we scan through these examples, we can see that the data of a time series are bivariate data where time is the independent variable.

Symbolical Representation of Time Series:-

If **'t'** stands for time and y_t represents the value at time **t** then the paired values (t, y_t) represents a time series data.

Symbolically a time series can be expressed as $y = f(t)$

Activity

Collect five Time Series data from available resources in your locality.

12.2 Components of Time Series

The values of a time series may be affected by the number of movements or fluctuations, which are its characteristics. The type of movements characterizing a time series are called components of time series or elements of a time series

Based on types of movements in the time series, there are three types of components

1) **Secular trend**
2) **Periodic movements**
   - **i) cyclical variations**
   - **ii) seasonal variations**
3) **Irregularvariations**

The following figure shows the graphical representation of a time series data on the sale of milk in millions of packets from 1993 to 2007. In this figure we can identify the pattern of secular trend, seasonal variation and cyclical variation inherent in the data.

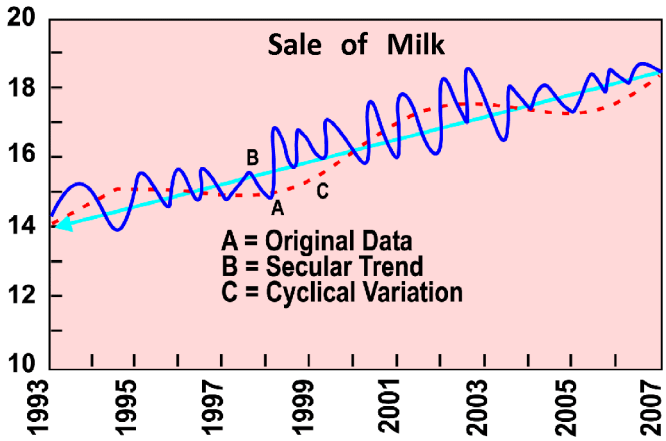


Fig. 12.4

Let us discuss in detail.

Secular Trend

Secular trend is the main component of time series. It is also called long term trend or simply trend.The secular trend refers to the general tendency of data to grow or decline over a long period of time.For example the population of India over years shows a definite rising tendency.The death rate of the country after independence shows a falling tendency because of advancement of literacy and medical facilities.

Mathematically a secular trend may be classified into two types

1) Linear Trend
2) Curvi- Linear Trend or Non- Linear Trend

If one plots the trend values for the time series on a graph paper and if it gives a straight line, then it is called a linear trend. That is in linear trend the rate of change is constant where as in non-linear trend there is varying rate of change.

Periodic Movements

Periodic Movements describes a data which varies in recognizable oscillations. The time required to move from one rise to another is known as the period of oscillation.Periodic movements generally categorized into two:- cyclical variation and seasonal variation.

Cyclical Variations

If the fluctuations or oscillations are not of fixed period then the movements are cyclic.Usually the period of oscillation is more than one year.The period of oscillation is known as a cycle.The time series related to business and economics show some kind of cyclical variations.

In Business Cycle there are four well defined periods or phases.

- Prosperity (Boom)
- Decline
- Depression
- Improvement

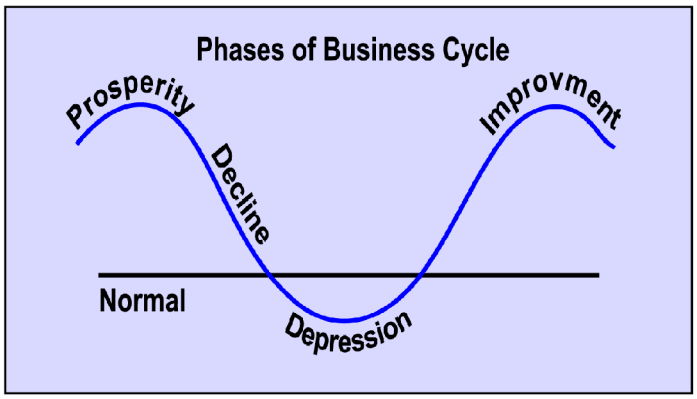


Fig. 12.5

Seasonal variations

In seasonal variations the fluctuations are of a fixed and known period. It occurs due to the rhythmic forces in a regular manner within a period of less than one year. Here the period of time may be monthly, weekly, or hourly.

The seasonal fluctuations in a time series occurs due to two factors

a) Due to natural forces

b) Manmade conventions

The most important factor causing seasonal variations is the climate changes.The climate and weather conditions such as humidity, rainfall, heat, etc., act on different products and industries differently. For example-during winter there is greater demand for woolen clothes, hot drinks, etc. Where as in summer cotton clothes, cold drinks have a greater sale and in rainy season umbrellas and rain coats have greater demand.

Though nature is primarily responsible for seasonal variations in time series, customs, traditions, and habits also have their impact. For example -on occasions like Deepavali, Onam, Christmas, etc., there is a big demand for sweets and clothes. There is a large demand for books and stationaries in the first few months of the opening of schools and colleges.

Irregular variations

The variations so far we have discussed are known as regular variations.But almost all time series includes another variation called random variations. This arises due to some irregular circumstances which are beyond the control of human being such as earth quakes, wars, floods, factory lockouts, etc. These factors are unforeseen and unpredictable but they are so significant like other fluctuations.

Know your progress

Identify the components inherent in each of the following time series

1. Growth of population.
2. Transformation in socio economic setup.

3. Data relating to the sale of banana chips during onam.
4. Increase in the sale of home appliances in Grant Kerala Shopping Festival.
5. Business Cycles.
6. Decrease in consumption due to some epidemic.
7. Lockout in a factory affects the standard of living of a group of people.

The value y_t of a time series at any time **t** can be expressed as a combined effect of resultant of the above mentioned factors. These expressions of time series are called as model for a time series. In most business analysis we use the multiplicative model and finds it more appropriate to business situations. So here also we use the multiplicative model of time series.

There are two types of Time Series models

1. Additive model

$Y = T + S + C + R$

In this model we consider all components as independent. But in almost all cases we cannot consider all components as independent. So multiplicative model is more popular in representing a time series.

2. Multiplicative model

$Y = T \times S \times C \times R$

Where T – Secular Trend

S – Seasonal Variations

C – Cyclical Variations

R – Random Variations

12.3 Uses of Analysis of Time Series

The analysis of time series is of great significance not only to the economists and business men but also to scientists, astronomers, geologists, sociologists, biologists, research workers,etc.

Why the analysis of time series is important?

It helps in understanding the past behavior.

By observing data over a period of time one can easily understand what changes have

taken place in the past. Such analysis will be extremely helpful in predicting the future behavior.

It helps in evaluating the current accomplishments.

The actual performance can be compared with the expected performance and the cause of variation can be analysed. If expected sale for 2014 was 10000 and the actual sale was only 9000. One can easily investigate the cause for the shortfall in achievements.

It helps in planning future operations

The major use of time series analysis is in the theory of forecasting. The analysis of the past behavior enables to forecast the future. Time series forecast are useful in planning and allocating budgets in different sectors of economy.

It facilitates comparison

Different time series are often compared and important conclusions are drawn from there.

12.4 Trend Analysis

Secular trend is a long term movement in a time series. This component represents basic tendency of the series.

The following methods are generally used to determine trend in any given time series.

1. **Free hand curve method**
2. **Semi average method**
3. **Method of Moving Average**
4. **Method of least squares**

Let us discuss them in detail :

Free hand curve Method

This is one of the most preliminary and easy way of fitting a trend line.

The method is as follows:-

First we plot the data on a graph paper, taking time on the X- axis and the corresponding values of the other factor on the Y- axis. After that, we join these points by straight lines.

While drawing a trend line by free hand curve method, one should keep in mind that the sum of perpendicular distances from points above the trend line will be equal to the sum of perpendicular distances from points below the trend line

Now draw straight line on this graph by eliminating the ups and downs by mere observation. This straight line shows the trend of the data.

The disadvantage of this method is that the trend line variesfrom person to person, as it depends on individual judgment. Hence it cannot be used as a basis for exact prediction. But for a rough estimation or to get an idea of change, this method is advisable.

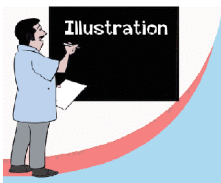


Illustration 12.1

The data below shows the production of cotton (in tones) from the year 2005 to 2013. Determine the trend line by the free hand curve method.

Year	2005	2006	2007	2008	2009	2010	2011	2012	2013
Production(in tones)	91	111	136	412	720	900	1206	1322	1380

Solution:

See the graph drawn with the original data and trend line graph

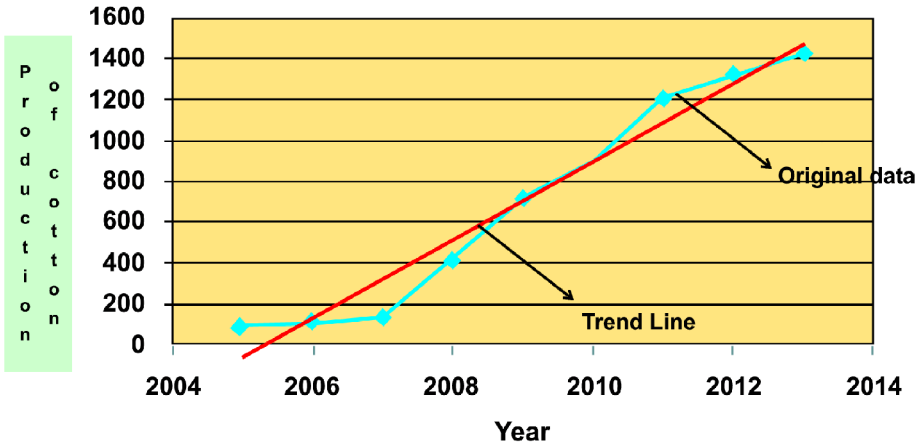


Fig. 12.6

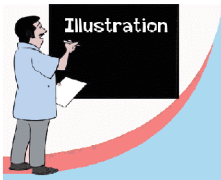


Illustration 12.2

A data related to the production (in crores) of a multinational company from the year 2007 to 2013 are given in the following table. Draw the trend line by free hand curve method

Year	2007	2008	2009	2010	2011	2012	2013
Production	125	128	133	135	140	141	143

Solution:

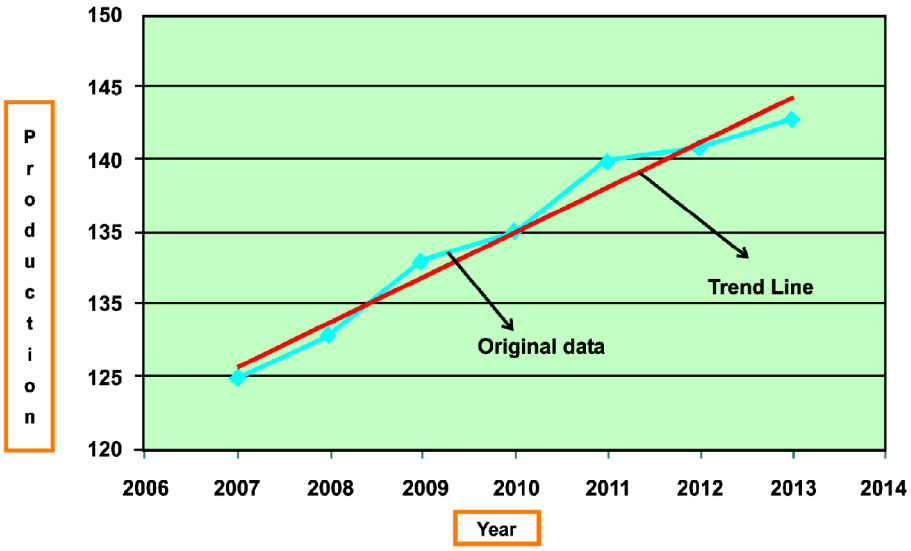


Fig. 12.7

See the graph drawn with the original data and trend line drawn by the free hand curve method.

Know your progress

1. The table below shows the Domestic Tourists visit in Kerala from the year 2008 to 2013. Draw the trend line by free hand curve method.

Year	2008	2009	2010	2011	2012	2013
Domestic Tourists (in million)	70.59	79.13	85.95	93.81	100.76	108.57

2. The following data relate to the production of fish in Orissa (in '000 metric tons) from the year2004 to 2009.

Year	2004	2005	2006	2007	2008	2009
Production	40	45	40	42	46	52

Draw the trend line by free hand curve method

Activity

Using data obtained from activity 1 , draw the trend line by free hand curve method.

Semi average method

In this method the whole data is divided into two equal parts with respect to time. When the time series contains an even number of observations, we divide it into two equal parts. Half of the observations constitute the upper part and the rest will constitute the lower part. When we have an odd number of observations in the series, the series is divided into two parts excluding the middle observation. After the data have been divided into two parts, an average (arithmetic mean) of each part is obtained. We thus get two points. Each point is plotted corresponding to the midyear of each part. Then these points are joined by a straight line which gives us the trend line.

The line can be extended downwards or upwards to get intermediate values or to predict future values. So it can be considered as a scientific method for prediction. But it have all the disadvantages of an arithmetic mean. As the AM of each half is calculated, an extreme value in any half will greatly affect the trend values.

For example - if we are given data from the year 2001 to 2014, i.e, over a period of 14 years, the two equal parts will be the values corresponding to first 7 years, from 2001 to 2007 and the values corresponding to the next seven years, from 2008 to 2014. If data are given from 2001 to 2013, for a period of 13 years, the values corresponding to first 6 years, from 2001 to 2006 constitute the upper part. Exclude the value corresponding to 2007, then the values from 2008 to 2013 constitute the lower part.

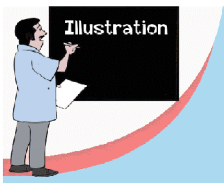


Illustration 12.3

Draw the trend line by Semi average method for the following data on the production of articles of a company from the year 2002 to 2007.

Year	2002	2003	2004	2005	2006	2007
Production	60	75	81	110	106	120

Solution:

Year	Production	Semi averages
2002	60	
2003	75	(60 + 75 + 81)/3 = 72
2004	81	
2005	110	
2006	106	(110 + 106 + 120)/3 = 112
2007	120	

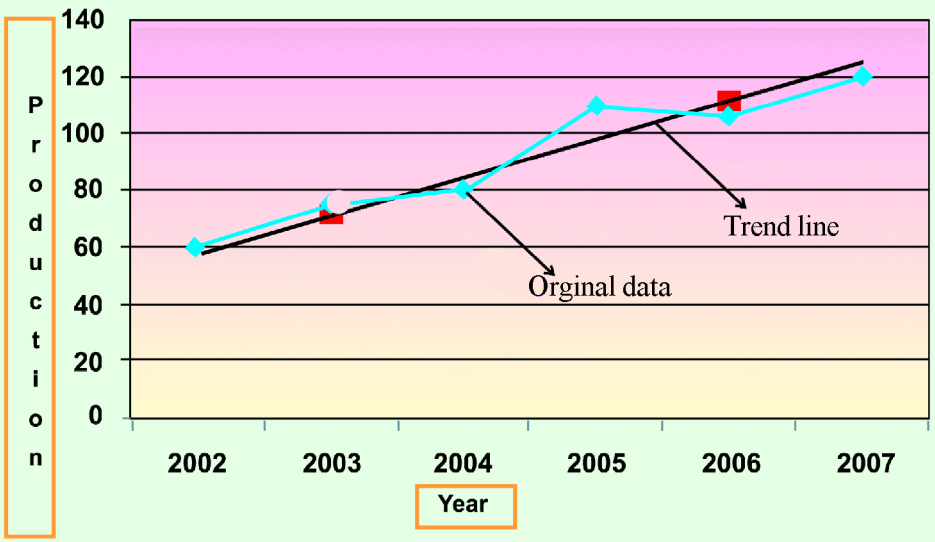


Fig. 12.8

See the graph drawn with the original data and trend line drawn by semi average method

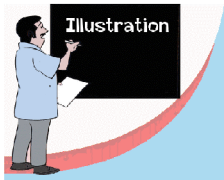


Illustration 12.4

Draw the trend line by the method of semi averages from the year 2006 to 2012 for the following output.

Year	2006	2007	2008	2009	2010	2011	2012
Output(Units)	600	821	1028	800	1280	1024	1400

Solution:

Year	Output (Units)	Semi averages
2006	600	
2007	821	(600 + 821 + 1024)/3 = 816.33
2008	1028	
2009	800	
2010	1280	
2011	1024	(1280 + 1024 + 1400)/3 = 1234.667
2012	1400	

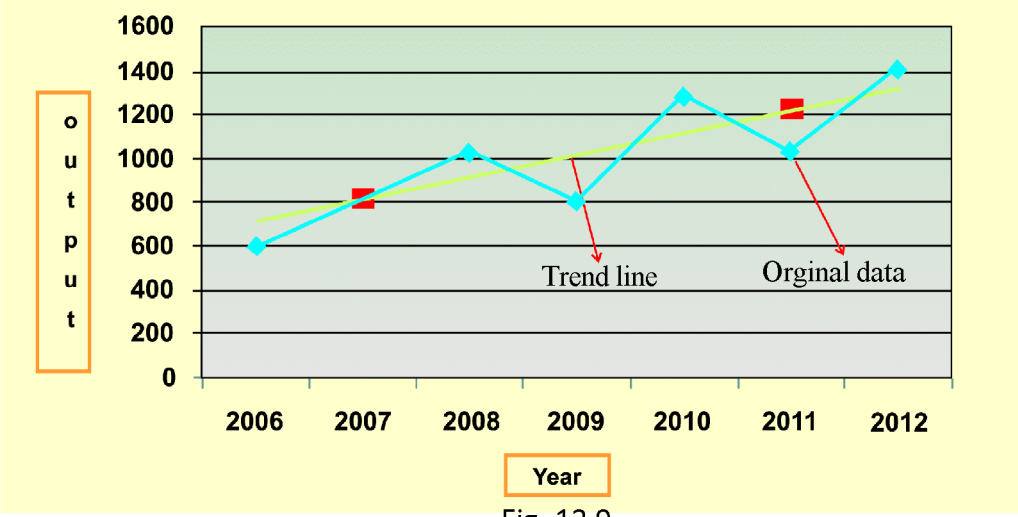


Fig. 12.9

Know your progress

1. Data on the production of bleaching powder ('000 Tonnes) for 12 years 2000 to 2011 is given in the following table. Draw the trend line by Semi-Average method.

Year	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011
Production	7.1	6.7	7.0	7.9	7.4	10.8	9.2	10.5	15.5	13.7	16.7	15.0

2. The following data relate to the export of cashew (in '00 metric tons) from India to foreign countries from the year 2000 to 2009. Draw the trend line by semi average method

Year	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
Quantity	18.6	22.6	38.1	40.9	41.4	40.1	46.6	60.7	57.2	53.4

Method of Moving Average

Moving average is commonly used in time series to smooth-out short term fluctuations and random fluctuations from the time series. Moving average for a period of **k** is a series of successive arithmetic means of **k** terms at a time series starting with 1st, 2nd, and so on. That is, the first average is the mean of first **k** terms, the second average is the mean of **k** terms starting from the second term (discarding the first term), third moving average is the mean of the next **k** terms starting from the third term (discarding the second term) and so on. If **k** is odd then the moving average is placed against the value of the time interval it covers. But if **k** is even then the moving average lies between the two middle periods which does not correspond to any time period. So we take mean of the successive moving averages and place it against the time interval it covers. The resultant moving averages are the trend values.

If n is the number of years and k is the period of the moving average, then:

number of moving average = $n-k+1$, if k is odd

= $n-k$, if k is even

The main disadvantage of this method is that the trend value does not correspond to any mathematical function and hence it cannot be used for predicting future values. Another drawback is that it results in loss of values for some years both in the beginning and at the end of time series, depending on the period of moving average.

The period of Moving Average is determined by the cycle in the data.

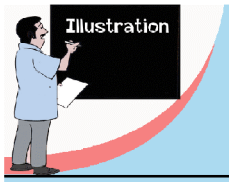


Illustration 12.5

Calculate 3-yearly moving average for the following data from the year 2005 to 2013.

Year	2005	2006	2007	2008	2009	2010	2011	2012	2013
Production (in tones)	45	40	42	46	52	56	61	64	69

Solution

Year	Production(in tones)	3- yearly moving total	3-yearly moving average
2005	45	_____	_____
2006	40	45+40+42 = 127	127/3 = 42.33
2007	42	40+42+46 = 128	128/3= 42.67
2008	46	42+46+52 = 140	140/3= 46.67
2009	52	46+52+56 = 154	154/3= 51.33
2010	56	52+56+61 = 169	169/3= 56.33
2011	61	56+61+64 = 181	181/3= 60.33
2012	64	61+64+69 = 194	194/3= 64.67
2013	69	_____	_____

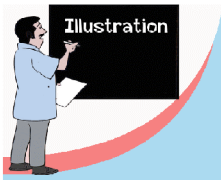


Illustration 12.6

The following data relate to the profit(in '000 Rs.) of a company from the year 2004 to 2012. Construct the trend values by 4-yearly moving average method

Year	2004	2005	2006	2007	2008	2009	2010	2011	2012
Profit	100	110	111	90	99	98	99	87	75

Solution

Year	Profit	4-yearly moving total	4-yeraly moving average	4-yeraly moving average centered
2004	100	_____	_____	_____
2005	110	_____	_____	_____
		411	102.75	
2006	111			(102.75+102.5)/2 = 102.625
		410	102.5	
2007	90			(102.5+99.5)/2 = 101
		398	99.5	
2008	99			(99.5+96.5)/2 = 98
		386	96.5	
2009	98			(96.5+95.75)/2 = 96.25
		383	95.75	
2010	99			(95.75+89.75)/2 = 92.75
		359	89.75	
2011	87	_____	_____	_____
2012	75	_____	_____	_____

Know your progress

1. The following table gives the index numbers of agricultural production of a country (from 2005 to 2010) . Construct the trend values by 5- yearly moving averages

Year	2005	2006	2007	2008	2009	2010
Index Numbers	132.9	125.1	138	117	135.3	142.9

2.5. The following table shows the number of road accidents in a State for the year 2003 – 2013. Construct the trend values by 4- yearly moving average method.

Year	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013
Number of road accidents	103	114	110	115	119	127	125	136	145	151	155

Method of Least squares

It is a mathematical method for finding trend values. Let the trend line equation be of the form **y = ax + b**, where **a** and **b** are unknown constants called parameters. They determine the location of the line. To determine the values of **a** and **b**, we use the method of least squares. By this method a and b can be determined by solving the equations

The method of least square was developed by Adrien Maire Legendre in 1805

$$\Sigma y = a\Sigma x + nb \text{ and}$$

$$\Sigma xy = a\Sigma x^2 + b\Sigma x$$

These equations are known as **normal equations** or **estimating equations.** Solving the normal equations we get the values of **a** and **b.** Substituting these values in the equation **y = ax + b**, we get the **trend equation.**

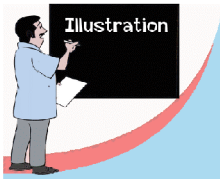


Illustration 12.7

The figures of production(in thousands of quintals) of a sugar factory from the year 2001 to 2007 are given below. Fit a straight line trend.

Year	2001	2002	2003	2004	2005	2006	2007
Production(in'000qtls.)	80	90	92	83	94	99	92

Solution

Year(t)	Production(y)	x = t-2004	x^2	xy
2001	80	-3	9	-240
2002	90	-2	4	-180
2003	92	-1	1	-92
2004	83	0	0	0
2005	94	1	1	94
2006	99	2	4	198
2007	92	3	9	276
n =7	$\Sigma y = 630$	$\Sigma x = 0$	$\Sigma x^2 = 28$	$\Sigma xy = 56$

The equation of the straight line is **y = ax + b**

To find a and b we use the normal equations

$$\Sigma y = a\Sigma x + nb$$

$$630 = a \times 0 + 7 \times b$$

$$630 = 7b$$

$$b = 90$$

$$\Sigma xy = a\Sigma x^2 + b\Sigma x$$

$$56 = a \times 28 + b \times 0$$

$$56 = 28a$$

$$a = 2$$

Then the trend equation is:

$$y = 2x + 90$$

$$y = 2\,(t - 2004) + 90$$

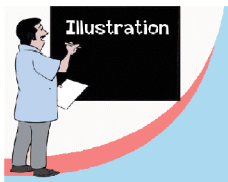


Illustration 12.8

Following is the data related to the production (in crores) of a company in different years. Fit a straight line trend to the data. Estimate the production for the year 2015.

Year	2006	2007	2008	2009	2010	2011	2012	2013
Production	125	128	133	135	140	141	143	145

Solution:

Year (t)	Production (y)	x = t-2009.5	x^2	xy
2006	125	-3.5	12.25	-437.5
2007	128	-2.5	6.25	-320
2008	133	-1.5	2.25	-199.5
2009	135	-0.5	0.25	-67.5
2010	140	0.5	0.25	70
2011	141	1.5	2.25	211.5
2012	143	2.5	6.25	357.5
2013	145	3.5	12.25	507.5
n =8	Σy =1090	Σx = 0	Σx^2 = 42	Σxy =122

The equation of the straight line is **y = ax + b**

To find a and b we use the normal equations

$\Sigma y = a\Sigma x + nb$

$1090 = a\times 0 + 8\times b$

$1090 = 8b$

$b = 136.25$

$\Sigma xy = a\Sigma x^2 + b\Sigma x$

$122 = a\times 42 + b\times 0$

$122 = 42a$

$a = 2.90$

Then the trend equation is:

$y = 2.9x + 136.25$

$y = 2.9\,(t - 2009.5) + 136.25$

When t = 2015

$y = 2.9\,(2015-2009.5) + 136.25$

$y = 152.2$

The estimated production for 2015 is 152.2 Crores.

Know your progress

1) Below are given the production of TV sets (in thousands). Fit a trend equation by the method of least squares. Also estimate likely production for the year 2009

Year	2000	2001	2002	2003	2004	2005	2006	2007
Production	17	20	19	26	24	40	35	55

2) Find trend equation by the method of least squares. Also estimate the sales in 2015.

Year	2008	2009	2010	2011	2012
Sales(in crores)	6.7	6.1	7.9	5.6	6.8

Shifting the trend origin

For simplicity and ease of computation, trends are usually fitted to annual data with the middle of the series as origin. At times it may be necessary to change the origin of the trend equation to some other point in the series. For example, annual trend values must be changed to monthly or quarterly values if we wish to study seasonal or cyclical patterns.

The shifting of the trend origin is a simple process. The procedure of shifting the origin may be generalized by the expression.

$Y = a(X + k) + b$

where k is the number of time units shifted.

k is positive, if the origin is shifted forward in time &

k is negative, if the origin is shifted backward in time

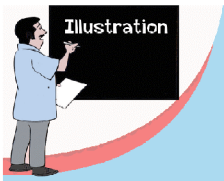


Illustration 12.9

A trend equation is given as $Y = 110 + 2X$ with origin as 2001. Shift the origin to 2005

Solution:

The equation is $Y = a(X + k) + b$. Here $a = 2$, $b = 110$, and $k = 2005-2001 = 4$

The required equation is $Y = 2(X + 4) + 110$

$= 2X + 8 + 110$

$= 2X + 118$

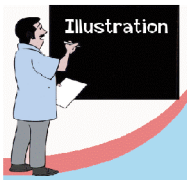


Illustration 12.10

A trend equation is given as Y =210 – 1.5X with origin as 2005. Shift the origin to 2000.

Solution

The equation is Y = a(X + k) + b.Here a = -1.5, b = 210, and k = 2000-2005 = -5

The required equation is Y = -1.5(X -5) + 210

= -1.5X + 7.5 + 210

= -1.5X + 217.5

Know your progress

1. A trend equation is given as Y = 18.04 X +126.55 with origin as 2004. Shift the origin to 2008
2. A trend equation is given as Y = 1.68 X +20.6 with 2011 as origin. Shift the origin to 2009

Let us conclude

A Time series is a sequence of data observed, collected and arranged in chronological order. There are three components for a time series i) secular trend ii)periodic movements iii) random variations. We can model a time series as Y = T × S × C × R. A time series is of great significance in studying a data, about the past behavior, evaluating current accomplishments, planning future operations, and for comparison. Trend analysis can be done through the following four methods. a) Free hand curve method b) Semi average method c) Moving average method and d) Least square method. Also in this chapter we have discussed the method of shifting the origin of a trend line.

Let us assess

For Questions 1-10 : choose the correct answer from the given choices.

1. A time series is a set of data recorded .....
   a) Geographically b) Chronologically
   c) Both geographically and chronologically d) None of these
2. The type of movements characterizing a time series are called .....
   a) fluctuations b) components c)characteristics d)resources
3. Secular trends refers to .....
   a) long term movement b) short term movement
   c) periodic movement d) permanent movement
4. Periodic movements with fixed and known period are known as .....
   a) Cyclic variations b) Business cycle
   c) seasonal variation d) oscillations
5. A business cycle has .....well defined periods.
   a) two b) three c) four d) five
6. The sale of cotton clothes in summer is associated with .....component of time series.
   a) secular trend b) seasonal variation
   c)cyclical variation d)irregular variation
7. Variations in a time series due to the factors which are beyond the control of human hand is known as.....
   a) secular trend b) seasonal variation
   c) cyclical variation d)irregular variation
8. Free hand curve method is not used for exact prediction. Because.....
   a) It is a graphical method b) It depends on individual judgment
   c) It gives only a rough sketch d) It is an easy method

9. Moving average method is used in time series to smooth out:

   a) long term fluctuations b) irregular variations

   c) short term fluctuations d) cyclic variations

10. A time series has 15 observations. Then the number of moving averages of order 5 is.....

    a) 15 b) 10 c) 11 d) 9

11. What is a time series?

12. List out the components of time series

13. Explain periodic movements with example.

14. What is secular trend?

15. What do you mean by irregular variations in time series?

16. Explain business cycle.

17. What are the uses of time series analysis?

18. Give the model of a time series.

19. With which component, the following time series data mainly associate with?

    a) A recession

    b) an increase in the sale of cool drinks in summer season

    c) an epidemic

    d) decline in the death rate due to advancement in science

    e) Continually increasing the sale of vehicle

    f) a factory lockout

    g) sale in a textile in a festive season

20. Fit a trend line to the following information using:

    a) Free hand method

    b) Method of semi average

Year	2005	2006	2007	2008	2009	2010	2011	2012	2013
Sales('000units)	12	15	17	16	14	14	18	20	16

Predict the sales for the year 2015

21. Fit a trend line by a) free hand curve method b) Method of Semi Averages

Year	2007	2008	2009	2010	2011	2012	2013	2014
Sales in(lakhs)	39.2	30.1	38.94	38.6	38.1	38.1	38.1	37.7

22. Construct trend values by using 3- yearly moving average

2,6,1,5,3,7,2,8,9,4.

23. Assuming five yearly cycle, determine trend values of ban clearings (in crores) by moving average methods

Year	1	2	3	4	5	6	7	8	9	10	11	12
Bank clearings	622	680	662	710	690	685	740	724	722	810	800	825

24. The following data refer to the sales of commercial vehicles in Kerala of a leading automobile company in the country from 2004 to 2013. Calculate the trend value by 5 yearly moving average.

Year	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013
Sales	2880	1693	2136	3707	1931	1637	1746	2638	2655	3576

25. The following data are the sales of a company from 2005 to 2013. Calculate the trend value by 4 yearly moving average

Year	2005	2006	2007	2008	2009	2010	2011	2012	2013
Sales(in crores)	26.20	35.40	39.20	45.80	49.00	50.40	54.80	60.00	71.80

26. The following data relate to the sale of mobile phones from a shop in the city fromm 2005 to 2012. Calculate the trend values by 4- yearly moving average.

Year	2005	2006	2007	2008	2009	2010	2011	2012
Sales	128	265	341	412	485	531	578	620

27. Below are given the production of TV sets (in thousands) of a company from 2000 to 2007. Fit a trend equation by the method of least squares.

Year	2000	2001	2002	2003	2004	2005	2006	2007
Production	17	20	19	26	24	40	35	55

28. The following table gives the profits of a concern for 5 Year ending in 2010. Estimate the profit for the year 2014.

Year	1	2	3	4	5
Profit (in lakhs)	2.08	6.74	23.1	45.27	138

29. Growth of merchant shipping fleet from 2008 to 2013 is given in the adjoining table. Estimate the trend value for the years.

Year	2008	2009	2010	2011	2012	2013
Shipping Fleet	2.65	2.89	3.49	3.87	5.09	5.48

30. The following data gives the details of deaths due to cancer reported in a hospital in a locality. Fit a trend line equation. Also estimate the number of deaths in 2014.

Year	2009	2010	2011	2012	2013
Number of deaths	4	7	11	13	17

31. The following table shows the number of students who got placement through campus recruitment conducted by a multinational company in an Engineering college. Fit a straight line trend by the method of least squares. Also estimate the value for the year 2015.

Year	2010	2011	2012	2013	2014
Number of students	16	18	17	24	29

Lab Activity

1. Data of number of students studying in a college from the year 2004 to 2013 are given in the table below. Find 3 yearly moving average.

Year	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013
Number of Students	332	317	357	392	402	405	410	427	430	438

2. Following table relate to price of a commodity from the year 2003 to 2010.

Year	2003	2004	2005	2006	2007	2008	2009	2010
Price in Rs.	380	540	650	720	755	815	870	930

a. Fit a straight line trend to the above data.

b. Estimate the price for the year 2015.

3. Assuming 3- yearly moving averages, calculate trend values for the following data relating to ban deposits (in crores).

Year	1995	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012
Deposits	960	976	984	996	1024	1040	1080	1128	1144	1135	1140	1168	1196

4. Following data gives the sales of a company for the years 2002 to 2012.

Year	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012
Sales('0000)	50.0	46.5	43.0	41.5	38.9	38.1	37.7	35.6	34.9	34.2	33.8

a. Calculate the trend value.

b. Estimate the sales in 2015.

Chapter 13

Index Numbers

Index number is one of the most widely used statistical devices. Consider the price of a commodity in a particular year. This price may be an increased price or a decreased price based on the price in the previous year. Similarly, consider the production of a certain commodity or sales of a product in a particular period. It may be higher or lower when compared with the previous period. Index numbers are used to describe these types of phenomina. Index number gives the net variation in percentages.

Significant Learning Outcomes

After the completion of this chapter, the learner:

- Explains the concept of index numbers.
- Constructs different kinds of simple and weighted index numbers.
- Interprets the value of index numbers.
- Identifies and uses consumer price index number.

Look at the following table. The table indicates the prices of commodities in 2012 and 2013 consumed by a family.

Commodities	Price per unit 2012	2013	Percentage of decrease/increase	Type of variation
Rice	20	24	20%	Increase
Sugar	30	40	33%	Increase
Cooking gas	450	500	11%	Increase
Egg	5	3	40%	Decrease
Tomoto	24	22	8.3%	Decrease

From the table it can be observed that the given prices of some commodities are increased while that of others are decreased. Index number provides a single figure to represent the net variation.

Historically, the first index was constructed in 1764 to compare the Italian price inded in 1750 with the price level in 1500.

Definition:

"Index number is a statistical measure designed to show changes in a variable or a group of related variables with respect to time, geographic location or other characteristics such as income, profession etc" - **Spiegel.**

Index number is a relative measure. It is always computed for a specific purpose for a specific time. The period or year for which index number is calculated is called **current period** or **current year.** Index number measures the relative change with reference to a time period, which is termed as **Base year** or **Base period.**

13.1 Classification of Index Number

Based on the characteristics of index numbers, we can classify index number as:

1) **Price index number**

2) **Quantity index number**

To discuss these in detail, we should be familiar with price relative. Let us discuss that:

Price index number

Price index numbers are used to describe the relative change in the price in a particular period of time. The relative change in the Price is called **Price relative**. If p_0 and p_1

denote the price of a commodity during base year and current year respectively, then the ratio = $\frac{p_1}{p_0}$ is called **price relative**. It is usually expressed in percentage.

Example: The retail price of rice in the year 2012 was Rs. 26 and that in the year 2013 was Rs.30. Find the price relative.

$$\text{Price relative} = \frac{p_1}{p_0} \times 100$$

Here $p_1 = 30$, $p_0 = 26$.

$$\text{Price relative} = \frac{p_1}{p_0} \times 100$$

$$= \frac{30}{26} \times 100$$

$$= 115.39$$

A price index number is a pure number which measures the net variation in the price of a group of commodities for a specific period called current year with respect to some reference period called base year.

If p_0, p_1 denote the prices of a commodity in the base year and current year respectively, then the ratio in percentage $\frac{p_1}{p_0} \times 100$ is called the price index number.

For example: The price index number of a group of commodities for a year 2012 is 250 with respect to the year 2010. This indicates that the variation in price is 150%.

Quantity Index Number

Instead of comparing the price we may be interested in comparing quantity of consumption or of production and so on for a particular period based on the previous period.

If q_0 and q_1 denote the quantity of a commodity during base year and current year respectively, then the ratio in percentage $\frac{q_1}{q_0} \times 100$ is called quantity index.

13.2 Types of Index Numbers

There are two types of index numbers - **Simple index numbers and Weighted index numbers.**

Let us discuss them in detail.

Simple index number

Simple average for relative is termed as simple index number.

Different simple index numbers are

a) Simple arithmetic mean(A.M) index

b) Simple geometric mean(G.M) index

c) Simple harmonic mean(H.M) index

d) Simple aggregate index

If price relative $\frac{p_1}{p_0} \times 100 = x$, and number of items is n then:

a) Simple arithmetic mean (AM) price index number $= \frac{\Sigma x}{n}$

b) Simple Geometric Mean (GM) price index number $= \sqrt[n]{x_1 \times x_2 \times \ldots \times x_n}$

c) Simple Harmonic Mean (HM) price index number $= \frac{n}{\sum \frac{1}{x}}$

Illustration 13.1

Compute a price index for the following by using:

1) Simple arithmetic mean(A.M) index
2) Simple geometric mean(G.M) index
3) Simple harmonic mean(H.M) index

Commodities	Price per kg. in 2012	Price per kg. in 2014
Beens	20	25
Potato	30	30
Tomato	10	15
Onion	25	35

Solution:

Commodities	P_0	P_1	$x = \frac{p_1}{p_0} \times 100$
Beens	20	25	125
Potato	30	30	100
Tomato	10	15	150
Onion	25	35	140
			$\Sigma x = 515$

1) Simple A.M price index $= \frac{\Sigma x}{n} = \frac{515}{4} = 128.75$

2) Simple G.M price index number $= \sqrt[n]{x_1 \times x_2 \times \ldots \times x_n}$

$= \sqrt[4]{125 \times 100 \times 150 \times 140} = 127.29$

3) Simple H.M price index number $= \frac{n}{\sum \frac{1}{x}}$

$= \frac{4}{\frac{1}{125} + \frac{1}{100} + \frac{1}{150} + \frac{1}{140}} = 124.22$

d) Simple Aggregate Method:

This is the simplest method of constructing index numbers.

Simple aggregate index

$$= \frac{\text{The total of current year prices for various commodities}}{\text{The total of base year prices of various commodities}} \times 100$$

$$= \frac{\Sigma p_1}{\Sigma p_0} \times 100$$

Where p_1 = current year prices for various commodities.

p_0 = base year prices for various commodities.

Simple aggregate price index number $= \frac{\sum p_1}{\sum p_0} \times 100$

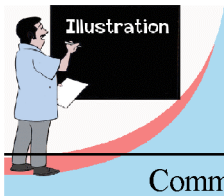


Illustration 13.2

From the following data construct an index number for 2013 taking 2012 as the base.

Commodities	Price in 2013(in Rs)	Price in 2012(in Rs)
Rice	32	28
Oil	88	75
Sugar	40	35
Wheat	22	18

Solution:

Commodities	p_1	p_0
Rice	32	28
Oil	88	75
Sugar	40	35
Wheat	22	18
	$\Sigma p_1 = 182$	$\Sigma p_0 = 156$

Simple aggregate index $= \frac{\Sigma p_1}{\Sigma p_0} \times 100 = \frac{182}{156} \times 100 = 116.67$

This means that in 2013 there is a net increase in the price of commodities in the index to the extent of 16.67% as compared to 2012.

Activity

Collect the prices of some commodities in the previous year and the present year with the help of newspapers, website, etc. Then calculate

a) Simple A.M Index

b) Simple G.M Index

c) Simple H.M Index

d) Simple Aggregate Index

Know your progress

1) From the given table, calculate:

a) Simple A.M. price index

b) Simple G.M price index

c) Simple H.M price index

d) Simple aggregate index

Commodities per kg	Price in 2013	Price in 2014
A	28	25
B	33	30
C	18	15
D	25	35

2, Compute simple aggregate index from the following data

Commodities	Price in 2013(in Rs.)	Price in 2012(in Rs.)
Rice	30	28
Oil	105	95
Sugar	39	35
Wheat	22	18

Weighted Index Number

In simple index numbers equal importance is given to all the items. But in many cases the item may have different importance. In such cases simple index number fails to measure the net variation. Hence we have to consider the relative importance of items in a group. If we consider their relative importance, we get weighted index number. Weighted index numbers are weighted average price relative. Weights are assigned to the price relative according to their relative importance.

In weigted index number, the relative importance of a commodity is measured in terms of money spend on that commodity. That is pq, where p is the price per unit and q is the quantity consumed. Thus we have different weights p_0q_0, p_1q_1, p_0q_1 and p_1q_0.

Commonly used weighted index numbers are:

1) Laspeyre's price index number (L)
2) Paasche's price index number (P)
3) Fisher's price index number (F)
4) Weighted aggregate index number

We can derive the formula for each of the index numbers by using the formula of weighted arithmetic mean. (We know that the weighted arithmetic mean = $\frac{\Sigma wx}{\Sigma w}$)

1) Laspeyre's price index number:

The Laspeyre's price index number is a weighted price index number, where the weights are determined as per the ***Base period.***

Here $x = \frac{p_1}{p_0} \times 100$, $w = p_0 q_0$

Then $$\frac{\Sigma wx}{\Sigma w} = \frac{\Sigma p_0 q_0 \times \frac{p_1}{p_0} \times 100}{\Sigma p_0 q_0}$$

$$= \frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} \times 100$$

Laspeyre's price index number $L = \frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} \times 100$

2) Paasche's price index number:

The Paasche's price index is a weighted price index in which the weights are determined as per the **current (given) year**

Here $w = p_0 q_1$

Then $$\frac{\Sigma wx}{\Sigma w} = \frac{\Sigma p_0 q_1 \times \frac{p_1}{p_0} \times 100}{\Sigma p_0 q_1}$$

$$= \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1} \times 100$$

Paasche's price index $P = \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1} \times 100$

3) Fisher's price index number:

Fisher's price index number is the geometric mean(GM) of the Laspeyer's and Paasche's price index number.

That is, Fisher's price index number, $(F) = \sqrt{L \times P}$

$$F = \sqrt{\frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} \times \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1}} \times 100$$

Fisher's price index number is the Ideal index number.

Why it is known as the Ideal index number ?

The following are the reasons.

1) It is the geometric mean of Laspeyer's and Paasche's index number. Theoretically G.M is the best average for constructing index numbers.

2) It takes into account both current year and base year prices and quantities.

3) It is free from bias.

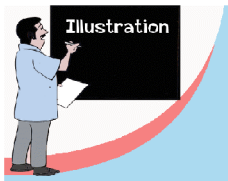


Illustration 13.3

Construct index number of price from the following data by applying

1) Laspeyer's method
2) Paasche's method
3) Fisher's method

Commodities	2013		2014	
	Price (Rs.)	Quantity (Kg.)	Price (Rs.)	Quantity (Kg.)
Ghee	40	1	60	1
Rice	30	20	36	24
Wheat	26	5	30	4

Solution

Commodities	p_0	q_0	p_1	q_1	p_0q_0	p_0q_1	p_1q_0	p_1q_1
Ghee	40	1	60	1	40	40	60	60
Rice	30	20	36	24	600	720	720	864
Wheat	26	5	30	4	130	104	150	120
Total					**770**	**864**	**930**	**1044**

1) Laspeyre's price index $L = \frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} \times 100$

$$= \frac{930}{770} \times 100$$

$$= 120.78$$

2) Paasche's price index P $= \dfrac{\Sigma p_1 q_1}{\Sigma p_0 q_1} \times 100$

$= \dfrac{1044}{864} \times 100$

$= 120.83$

3) Fisher's index, $F = \sqrt{L \times P}$

$= \sqrt{120.78 \times 120.83}$

$= 120.80$

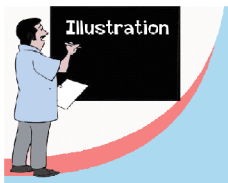


Illustration 13.4

From the following data construct Fisher's index number for the group of four commodities.

Commodities	Base year		Current year	
	Price	Quantity	Price	Quantity
A	2	20	5	15
B	4	4	8	5
C	1	10	2	12
D	5	5	10	6

Solution:

Commodities	p_0	q_0	p_1	q_1	p_1q_0	p_0q_0	p_1q_1	p_0q_1
A	2	20	5	15	100	40	75	30
B	4	4	8	5	32	16	40	20
C	1	10	2	12	20	10	24	12
D	5	5	10	6	50	25	60	30
Total					**202**	**91**	**199**	**92**

Fisher's ideal index, $F = \sqrt{\frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} \times \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1}} \times 100$

$$= \sqrt{\frac{202}{91} \times \frac{199}{92}} \times 100$$

$$= 219.12$$

4. Weighted Aggregate Index Number

$$\textit{Weighted aggregate index number} = \frac{\textit{Total expenditure in the current year}}{\textit{Total expenditure in the base year}} \times 100$$

$$= \frac{\Sigma p_1 q_1}{\Sigma p_0 q_0} \times 100$$

Weighted aggregate index number $= \frac{\Sigma p_1 q_1}{\Sigma p_0 q_0} \times 100$

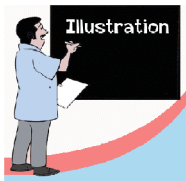


Illustration 13.5

Calculate wieghted aggregate index from the following data

Commodities	Base year		Current year	
	Kg.	Rate	Kg.	Rate
Bread	10	3	8	3.25
Meat	20	15	15	20
Tea	2	25	3	23

Solution

Commodities	p_0	q_0	p_1	q_1	p_0q_0	p_1q_1
Bread	3	10	3.25	8	30	26
Meat	15	20	20	15	300	300
Tea	25	2	23	3	50	69
Total					**380**	**395**

Weighted aggregate index number $= \dfrac{\Sigma p_1 q_1}{\Sigma p_0 q_0} \times 100$

$= \dfrac{395}{380} \times 100$

$= 103.9$

Know your progress

1, Construct the index number by using:

a) Laspeyre's method

b) Paasche's method

c) Fisher's method

Commodities	Base year		Current year	
	Price	Quantity	Price	Quantity
A	22	16	25	15
B	14	9	12	8
C	11	10	15	12
D	15	8	14	10

2) Calculate wieghted aggregate index from the following data:

Commodities	Base year		Current year	
	Kg.	Rate	Kg.	Rate
Bread	12	6	8	8
Meat	22	75	18	95
Tea	5	65	3	73

13.3 Consumer Price Index (Cost of Living Index)

The consumer price index number tells us about the variation in the cost of living of only one group of persons living in a particular region. By region we mean here an area within which retail prices are almost equal. By groups we mean classes distinguished from each other on the basis of income or habits. Thus there cannot be one cost of living index number for any class of workers of the whole country, because retail prices in different places differ and the pattern of consumption is also not alike in different localities. Similarly we cannot have a cost of living index number for the whole population of a particular town because groups on various commodities vary in different ways and the relative importance of various commodities to all persons is not identical.

Still consumer price index number in some situations give better information than general price index numbers.

The main objectives of consumer price index number is to know the direction (increase or decrease) in the cost of living of the consumer. Since the effect of price changes is not uniform on all class of people, different consumer price index numbers are constructed for different classes of people.

Consumer price index indicate the relative change in a basket of goods and service (same set of commodities and service) consumed at two time periods.

$$\text{Cost of living index} = \frac{\textit{Total expenditure in the current year}}{\textit{Total expenditure in the base year}} \times 100$$

$$\text{i.e, Cost of living index} = \frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} \times 100$$

The consumer price index is probably the best known price index. It is published by the U.S Bureau of Labour Statistics and is based on the price of several hundred items. The base year is 1967.For obtaining the base year quantities used as weights,the Bureau of Labor Statistics interviewed thousands of families to determine their consumption patterns since C.P.I reflects the general price level in the country.

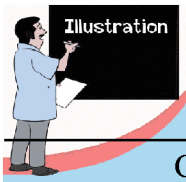


Illustration 13.6

Construct cost of living index number for the year 2014 from the following data

Commodities	Price		Quantity	
	2013	2014	2013	2014
A	5	8	80	100
B	3	4	90	100
C	7	7	60	60
D	11	14	20	25

Solution:

Commodities	p_0	p_1	q_0	q_1	p_0q_0	p_1q_0
A	5	8	80	100	400	640
B	3	4	90	100	270	360
C	7	7	60	60	420	420
D	11	14	20	25	220	280
Total					**1310**	**1700**

i.e, Cost of living index $= \dfrac{\Sigma p_1 q_0}{\Sigma p_0 q_0} \times 100$

$= \dfrac{1700}{1310} \times 100$

$= 129.77$

Know your progress

1, Construct cost of living index number for the year 2014 from the following data

Commodities	Price		Quantity	
	2012	2014	2012	2014
A	15	14	75	98
B	13	20	80	90
C	17	19	60	50
D	11	14	30	25

13.4 Characterstics of Index Numbers

1. Index numbers are expressed in percentages. They are used for comparison.
2. Index numbers are specialised averages.
3. Index number is a relative measure.
4. Index numbers measure change in some quantities which cannot be observed directly.
5. Index numbers are described as economic barometers. Index numbers are used to take the pulse of economy and serve as indicators of inflationary or deflationary tendencies.
6. Index numbers are computed for a specific purpose

13.5 Uses of Index numbers

1) Help in framing suitable policies:

Many of the economic and business policies are guided by index numbers. Fixing of wages and dearness allowances is mainly based on consumer price index. Many economic policies like volume of trade fixing of whole sale and retail prices, wage policy, fixing of price taxation, house rent allowance, etc. are guided by index numbers.

2) Help in studying trend:

Index numbers are most widely used for measuring changes over a period of time. With the help of index numbers it is easy to find out the trend of export, import, balance of payment, price national incomes, etc.

3) Help in forecasting future economic activity:

Index numbers are useful not only in studying the past and present working of our economy, but they are also important in forecasting future economic activity. Index numbers then are often used in time series analysis, the study of long-term trend, seasonal variations and business cycle development.

4) Help in measuring real wages:

$$\text{Real Wages} = \frac{Money\ \ wages}{Cost\ \ of\ \ living\ \ index} \times 100$$

5) Help in measuring purchasing power of money:

$$\text{Purchasing power of money} = \frac{1}{Cost\ of\ living\ index}$$

6) Index number is used to study the inflation and deflation:

Let us conclude

In this chapter we have discussed index numbers and different types of index numbers. Index numbers measure the relative change in the level of a phenomenon with respect to time, geographical locations or some other characterstics. Price index numbers are used to describe the relative change in prices of group of commodities changing with respect to time.

The simple aggregate index is used to construct a price index which is equal to the total of current year prices for the various commodities is divided by the total of base year prices and quotient is multiplied by 100. Simple average of price relatives is used to construct a price index. First of all price relatives are obtained for the various items included in the index and then the average of these relatives is obtained using any one of the measures of central values A.M,G.M or H.M.

In weighted index numbers, certain weights are to be assigned to the prices of the items depending on their relative importance. The important weighted index numbers are a) Laspeyre's index, b) Paasche's index , c) Fisher's index and d) Weighted aggregate method.

Let us assess

For Questions 1-6 : choose the correct answer from the given choices.

1) Index number is:

a) a Measure of relative change b) a special type of average

c) a percentage relative d) all the above

2) Consumer price index number is constructed for:

a) Well defined section of people b) All people

c) Factory workers only d) All the above

3) Most preferred type of averages for index number is:

a) Arithmetic mean b) Geometric mean

c) Harmonic mean d) None of the above

4) Index number helps:

a) In framing of economic policies b) In assessing the purchasing power of money

c) For adjusting national income d) All the above

5) Laspeyer's index formula uses the weights of the:

a) Base year b) Current year

c) Averages of the weights of a number of years d) None of the above

6) The weights used in Paasche's formula belongs to:

a) The base period b) The current period

c) To any arbitrary chosen period d) None of the above

7) Define Index number.

8) Index numbers are called 'economic barometer'. Why?

9) Explain uses of index numbers.

10) Distinguish between simple and weighted index numbers.

11) Define Laspeyer's and paasche's index numbers.

12) Fisher's index number is called ideal index number. Why?

13) Explain the cost of living index.

14) Construct an index number for 2014, taking 2012 as base for the following:

Commodity	Price in 2012	Price in 2014
A	90	95
B	40	50
C	90	110
D	30	35

15) From the following data construct an index for 2014 using simple aggregate method.

Commodity	Price in 2013	Price in 2014
A	50	70
B	40	60
C	80	90
D	110	120
E	20	20

16) The following are the prices of four commodities in 2012 and 2013. Calculate

a) The simple A.M price index

b) The simple G.M price index

c) The simple H.M price index for the year 2013 with the year 2012 as base.

Commodity	A	B	C	D
Price in 2012	16	25	8	20
Price in 2013	32	28	14	30

17) From the following data, construct an index for 2014 taking 2010 as base by the averages of relatives method using a) A.M b) G.M

Commodity	Price in 2010	Price in 2014
A	50	70
B	40	60
C	80	90
D	110	120
E	20	20

18) Calculate Fisher's price index from the following data.

Commodity	Base Year		Current Year	
	Price	Quantity	Price	Quantity
A	9.25	5	15	5
B	8	10	12	11
C	4	6	5	6
D	1	4	1.25	8

19) From the following data, prepare weighted index by using

a) Laspeyre's method

b) Paasche's method

c) Fisher's method

Year	Commodity 1		Commodity 2		Commidity 3	
	Price	Quantity	Price	Quantity	Price	Quantity
2011	5	10	8	6	6	3
2014	4	12	7	7	5	4

20) Using the following data, calculate

a) Laspeyre's index

b) Paasche's index

c) Fisher's price index

Commodities	Base Year		Current Year	
	Price	Quantity	Price	Quantity
A	15	5	25	8
B	20	8	28	7
C	30	3	40	4

21) From the following data costruct a price index number by using Fisher's method

Commodity	Base Year		Current Year	
	Price per unit	Expenditure (Rs)	Price per unit	Expenditure (Rs)
A	2	40	5	75
B	4	16	8	40
C	1	10	2	24
D	5	25	10	60

22) Construct cost of living index number for the year 2013 from the following data

Commodities	2010		2013	
	Price	Quantity	Price	Quantity
A	30	10	35	8
B	18	9	28	12
C	10	7	20	10

23) Construct the consumer price index from the following data

Commodities	2011		2014	
	Price	Quantity	Price	Quantity
Beens	20	8	40	6
Meat	50	10	60	5
Wheat	20	20	20	25

24) Calculate weighted aggregate index from the following data

Commodities	Base year		Current year	
	Price	Quantity	Price	Quantity
A	15	5	25	8
B	20	8	28	7
C	30	3	40	4

Answers

Chapter 1

1. b, 2. d 3. d
4. c 5. a 6.b
9. 0.44 10. 0.81 11. 0.5
12. 0.83 13. 0.836 14. 0.794
15.0.9 16. 0.1 17. 0.81
18. 0.75 19. -0.54

Chapter 2

1.a, 2.b, 3.b,
4.b, 5.d, 6.a,
7.b, 8.d, 9.d,
10.a, 11.a) $y = -1.12x + 70.78$,
b) 48.45, 12.11.67
13.a) $y = -0.61x + 136.6$, b) 93.9
14.a) 16.17, b) 15.4
15. $x = -1.102y + 82.30$ 16.a) 0.94
b) 83.33, 17.(130,90) 18.a) r= -0.5,
b) (4,7) 19. 2.67, 20.15.55,
21.a) r=0.53, 21,b) (5,13.5), 21. c) 7.92
22.b) r=0.79, 22.c) (13,10), 22.d) 1251.875
23.a) $20x - 9y - 75 = 0$ 23.b) 0.6,
23.c) 67 24.b) 0.73, 24.c)(-3,-2)

Chapter 3

1. b 2. a 3. b
4. c 5. d 6. b
7.a) $5x^4$ 7.b) 2x+5 7.c) $3x^2+14x+10$
7.d) 8 - 20x 7.e) -2 7.f) a
7.g) 2x 7.h) -3 x^{-4} 8.a) 120x
8.b) 24x - 40 8.c) 6x+10
9.a) min at x=2, min value= -16
9.b) neither min nor max
9.c) min at x = -2/3, min value = - 2
9.d) min at x = 3/2, min value = - ¼
9.e) minimum at x=2, minimum value=75 and maximum at x=-2, maximum value= 139
10. MC = 22 - 2x 11. $x = \frac{2}{3}$
12. x = 9, Max. profit = 176 13.a) $\frac{x^9}{9} + c$
13.b) $\frac{x^{25}}{25} + c$ 13.c) $\frac{x^{-11}}{-11} + c$ 13.d) $\frac{x^{-5}}{-5} + c$
13.e) $x^3 + \frac{5x^2}{2} - 2x + c$
13.f) $5x^5 - 4x^4 + 10x + c$
14.g) $4\frac{x^6}{6} - 3x^{-3} + 7x + c$ 14.a) $\frac{-28}{3}$
14.b) 6 14.c) 96 14.d) 2
15. P(x) =4x-3x^2+c
16. x=400, profit = 325

Chapter 4

1.a) continuous 1.b) discrete
1.c) continuous 1.d) continuous
1.e) discrete 1.f) discrete
1.g) continuous 1.h) discrete
1. i)continuous 1.j) continuo us
1. k)discrete 1.l) continuous
2. b 3. d 4. c
5. a 6. a 7. c
8. b 9. b 10. a
11. c 12. b 13. c
14. d 15. c 16. b
17. b 18.a) $\frac{1}{10}$ 18.b) $\frac{3}{10}$
18.c) $\frac{19}{100}$ 18.d) $\frac{4}{10}$ 19.a) 0.7
19.b) 2.1 19.c) 5.7, 19.d) 9.4
20. $\frac{2}{12}$, $\frac{7}{12}$

21.

x	0	1	2	3
P(x)	2/22	3/22	6/22	11/22

22.

x	1	2	3	4
F(x)	2/12	4/12	9/12	1

x	1	2	3	4
F(x)	2/22	5/22	11/22	1

23.a) $\frac{1}{48}$ 23.b) 0.167, 0.69, 0.46

24.a)16 24.b) 276 24.c) 20

24.d) 21 24.e) 720 24.f) 80

25. $\frac{3}{8}$ 26). No 27) No

28. $F(x)=\begin{cases}0 & x<-1\\ \frac{x^3+1}{2} & -1<x<1\\ 1 & x>1\end{cases}$

29.a) 0 29.b) $a\frac{2}{3}$

31. 0.454 and 0.1519

32. $F(x)=\begin{cases}0 & x<0\\ \frac{x^2}{2}, & 0\le x\le 1\\ 2x-\frac{x^2}{2}-1, & 1<x\le 2\\ 1 & x>2\end{cases}$

33. $f(x)=\frac{x^2}{9}\ 0\le x\le 3$

Chapter 5

1. c 2. b 3. b
4. b 5. a 6. b
7. a 8. a, 9. a
10. a 11. a 12. c
13. b 14. b 15.a) Yes
b) Yes c) Yes d) No
e) No f) Yes g) Yes
h) No i) Yes 16.a) 0.420
b) 0.346 c) 0.59 d) 0.25
e) 0.25 17.a) 75,18.8,4.3
b) 90,63,7.9 c) 10,5,2.2 d) 8,1.6,1
e)100,90,1.5 f) 6,5,2.2.
18. 0.124 19.a) 0.201 b) 0.878
20.240,48,6.9 21. 9,7.9,2.8 22.a) 0.0045
b) 0.1466 c) 0.1834, d) 0.1008
23.a) 0.9048 b) 0.9998. 24.a) 0.27
b) 865 25.0.101 26. 0.1563
27. 0.9502. 28. 0.0153.

Chapter 6

1. b 2. a 3. b
4. a 5. a 6. a
7. c 8. a 9. b
10. c 11.a) 0.8413 b) 0.9453
c) 0.6814 12.a) 0.6814 b) 0.2014
c) 0.283 d) 0.6826 13. Z1=1.23
14. Avg. birth wt.=3, SD=0.5, Prob.= 0.3
15. 0.15735 16. 0.9772 17. 0.5899
18.a) 0.1525 b) 0.8664 19. 387
20.a) 3.36% b) 9.18% 21.a) 0.5328
b) 0.1587 c) 0.9332 22. a=107.38
23. a=6.41 24. $\mu=72$cm 25. $\sigma=5$
26. mean= 50.29, SD=10.33.

Chapter 7

1. c 2. d 3. d
4. b 5. c 6. F variable
7. Central limit theorem 8. $\mu,\frac{\sigma^2}{n}$
9. 2.5 10.a) False b) True
c) True d) True e) False f) False
14. 1-a, 2-c, 3-d, 4-b

15.a)(58,58), (58,52), (58,56), (58,63), (52,58), (52,52), (52,56), (52,63), (56,58), (56,52), (56,56), (56,63), (63,58), (63,52), (63,56), (63,63)

15.b) 3.96 16.a) (2,3), (2,6), (2,8), (2,11), (3,6), (3,8), (3,11), (6,8), (6,11), (8,11)

b) $\sigma^2 = 10.8$, SE = 2.01

17.a) (2,4,6), (2,4,8), (2,4,10), (2,6,10), (2,8,10), (4,6,8), (4,6,10), (4,8,10), (6,8, 10)

b) 6, 5.33 18.a) 17 b) (18,15), (18,16), (18,19), (15,16), (15,19), (16,19)

c) SE=0.91 $\sigma^2 = 2.5$ 19. 1) 12.587

2) 0.05 3) 2.473 (two tail), 2.120 (one tail)

4) 26 5) 1.96 (two tail), 2.241 (one tail)

Chapter 8

1. a 2. b 3. c

4. Point estimation and interval estimation

5. U_1 and U_2 are unbiased

6. T_1 and T_2 are unbiased, T_1 is most efficient

7.a) 48.875 b) 51.5 c) 1.688

8. (139.45, 140.55)

9. (72.904, 79.096)

10. (747.424, 748.576), (747.242, 748.758)

11. (68.124, 69.876), (67.85, 70.15)

15. (4.004, 4.396)

Chapter 9

1. a 2. b 3. b
4. a 5. d 6. b
7. a 8. b 9. c
10. d 11. c 12. b
13. c 14. b

19. Z = 2.28, reject H_0

20. Z = 1.89, accept H_0

21. Z = 2.46, reject H_0

22. Z = 1.89, accept H_0

23.a) $H_0 : \mu = 13.5, H_1 : \mu > 13.5$

b) Z = - 4.19, reject H_0

24. t = 1.89, accept H_0

25. t = 2.33, reject H_0

26. t = 3, reject H_0

27. |t| = 2.43, reject H_0

28. t = -2.71, reject H_0

29. t = 3.16, reject H_0

30. Z = 8.99, reject H_0

31. Z = 28.39, reject H_0

32. Z = 3.75, reject H_0

33. Z = 2.26, accept H_0

34.a) Z = 7.76, reject H_0,

b) Z = 7.44 reject H_0

35. $\chi^2 = 107.7$, reject H_0

36. $\chi^2 = 11.11$, reject H_0

37. $\chi^2 = 374.96$, reject H_0

Chapter 10

1. b 2. d 3. b
4. a 5. d 6. a
7. d 8. c 12. No

13. F=2.94, the trearment effects are not equal.

14. F=6.95, the teatment effects are not equal.

18. F=4.36, the sales are significantly different.

19. F=6.82, the machines are not equally efficient.

Chapter - 11

1. a. 2. b 3. c

4. c 5.b 6. D_3 and D_4

7. Chance 8.Chance 9. Two

10. $\bar{x}$ and R 11. For $\bar{x}$ chart LCL = 17.546, CL = 44.2, UCL = 40.854, For R chart LCL = 0, CL = 5.8, UCL = 12.263 (out of control)

12. For $\bar{x}$ chart LCL=37.162, CL=71.6, UCL =106.024, For R chart LCL=0, CL=59.66, UCL= 126.1809 (out of control)

13. For $\bar{x}$ chart LCL = 17.79, CL= 20.56, UCL= 23.256, For R chart LCL=0, CL=4.8, UCL = 10.152

14. For $\bar{x}$ chart LCL = 3.12, CL = 3.126, UCL = 3.13, For R chart LCL = 0, CL = .009, UCL = .0019

15. For $\bar{x}$ chart LCL = 6.06, CL = 9.66, UCL = 13.26, For R chart LCL = 0, CL = 6.3, UCL = 13.32

16. For $\bar{x}$ chart LCL = 37.06, CL = 41.92, UCL=46.72, For R chart LCL= 0, CL= 6.67, UCL = 15.22 (out of control)

17. For np chart LCL = 0, CL = 5.85, UCL = 13.05 (out of control)

18. For np chart LCL = 0, CL = 2.08, UCL = 5.803 (out of control)

19. For np chart LCL = 0, CL = 2.6, UCL = 6.8 (out of control)

20. For $\bar{x}$ chart LCL=14.61, CL = 30.5665, UCL = 46.53, For R chart LC = 0, CL = 15.60, UCL = 40.16 (within control)

21. For np chart LCL = 0, CL = 7.165, UCL = 14.5 (within control)

22. For $\bar{x}$ chart LCL = 61.61, UCL = 69.11 For R chart LCL = 0, UCL = 13.75 (within control)

23. For $\bar{x}$ chart LCL = 5.46, UCL = 20.86 For R chart LCL = 0, UCL = 24.84 (within control)

24. LCL = 0, UCL = 7.18 (within control)

25. For np chart LCL = 2.97, UCL = 25.03 (out of control)

Chapter 12

1.b	2.b	3.a
4.c	5.c	6.b
7.d	8.b	9.c
10.c	19. a) secular trend	
b) seasonal	c) irregular d) secular trend	
e) trend	f) irregular g) seasonal	

27. Y = 4.78(t - 2003.5) + 29.5

28. Y = 31.037(t - 2008) + 43.038
Profit for the year 2014 = 229.26

29. Y = 0.603(t - 2010.5) + 3.911

30. Y = 3.2(t - 2011) + 10.4, Number of deaths in 2014 = 20

31. Y = 3.2(t - 2012) + 20.8
Estimated number of students in 2015 = 30 (approximately).

Chapter 13

1. d	2. a	3. b
4. d	5. a	6. b
14. 116	15. 120	
16.a) 159.25	b) 155.72	c)152.04
17.a) 122.30	b) 120.83	18. 148.84
19.a) 83.62	b) 83.57	c) 83.59
20.a) 144.31	b) 146.32	c) 145.31
21. 219.13	22. 139.47	23. 124.53
24. 171.08		

Glossary

Alternative hypothesis	:	The hypothesis which complements null hypothesis
ANOVA	:	Analysis of Variance - Test for significance of several means.
Assignable Causes	:	Causes whose effects can be measured and controlled.
Bernoulli trial	:	A process in which each trial has only two possible outcomes, the probability of the outcome at any trial remains fixed over time, and the trials are statistically independent.
Binomial Distribution	:	Discrete distribution in which there are only two possibilities on any one trial
Central Limit Theorem	:	Sum of large number of independant random variables follow normal distribution.
Chance Causes	:	Causes whose effects are beyond the human control.
Chi Square test	:	Test used to analyse the independance of two attributes.
Chi-square variable	:	Square of a Standard normal variable
Coefficient of Correlation	:	Quantitative measure of linear correlation of two variables.
Components of time series	:	The type of movements characterizing a time series.
Confidence Interval	:	An interval which contains the true value of the parameter
Consistency	:	A property of the estimator to give a stable result.
Contingency Table	:	Observations arranged in a two way table to test independance of attributes.
Continuous Random Variable	:	A random variable, which takes all values within a certain interval.
Control Chart	:	Quality Control Chart is used to evaluate whether a process in a statistical control
Correlation	:	Study of degree of relationship between variables.
Covariance	:	Variance of two variables together.
Critical region	:	The region at which the null hypothesis is rejected.
Critical Value	:	The value that divides the acceptance region from the rejection region
Cyclical variation	:	Periodic movements which are not of fixed period.

Degrees of freedom	:	Number of independent observations
Discrete Random Variables	:	If the values of a random variable are finite or countably infinite
Efficiency	:	A property of the estimator, where the variance is minimum.
Estimate	:	The value of an estimator
Estimation of Parameter	:	Process of estimating the parameter
Estimator	:	Statistic used to estimate a parameter
F	:	Snedcors F variable
Fisher's Index Number	:	Weighted aggrgate price index in which both base year and current year quantity are considered.
Free hand curve method	:	Plot data on the graph and then draw straight line on the graph by eliminating ups and downs by mere observation
Index Numbe	:	Measure used to study the percentage change in price (or Quantity)
Inferential Statistics	:	Branch of statistics used for determining unknown aspects of a population
Interval Estimation	:	Suggests an interval which is expected to contain the value of the parameter
Irregular variations	:	Variations arises due to some irregular circumstances.
Karl Pearsons Correlation	:	Measure of degree of linear relationships
Laspayer's Index Number	:	Weighted aggrgate price index in which base year quantity is considered.
LCL	:	Lower Control Limit
Level of significance	:	Probability of a Type I error
Lower Control Limit	:	The bottom end line of control chart
Mathematical Expectation	:	Arithmetic mean of a random variable
Method of least squares	:	Trend line obtainned by using the principle of least squeare
Moving Average	:	Series of arithmetic means of time series in particular interval of time to study the trend.
Normal Curve	:	Bell shaped mesokurtic symmetriic curve.
Normal Distribution	:	Most widely used continuous probability distribution
np Chart	:	A quality control chart for attributes which shows the actual number of defectives found in each sample.

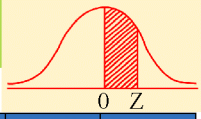
Null Hypothesis	:	The hypothesis which states that there is no difference between a parameter and a specified value of parameter.
Paasche's Index Number	:	Weighted aggrgate price index in which current year quantity is considered.
Parameter	:	Function of population
Perfect correlation	:	Perfect positive (r=1), perfect negative (r= -1).
Periodic movements	:	Data which varies in recognizable oscillations.
Point Estimation	:	Suggests a single value for the parameter
Poisson Distribution	:	The probability distribution of the number of occurances of rare events in a series of trials.
Power of a test	:	The probability of rejecting H0 when it is false.
Price Relative	:	Price index number of a single commodity
Principle of least squares	:	The sum of squares of the actual values and the estimated values should be the minimum.
Qualilty	:	When a produt delivers what is stipulated in its specifications.
Qualilty Control	:	The collection of strategies, Techniques and actions taken by an organisation to ensure the production of quality products.
R - Chart	:	A plot of sample ranges used in quality control
Random Variables	:	A variable whose value is determined by the outcome of a random experiment .
Regression Analysis	:	The process of constructing a mathematical model or function, that can be used to predict or determine one variable by any other variable.
Rejection Region	:	The region at which the null hypothesis is rejected.
Sampling distribution	:	Probability distribution of a statistic
Seasonal variation	:	The fluctuations which are of fixed and known period.
Secular trend	:	Tendency of a data to grow or decline over a long period of time.
Semi average method	:	Divide the data into two equal parts and find the average of each part. Plot it against the midyear of each part and the trend line is obtained by joint these points straight line.
Snedecor's F variable	:	Ratio of two Chi-square variables
Spearmans Rank Correlation	:	Correlation coefficient for qualitative data.

Standard error	:	Standard deviation of a statistic
Standard Normal Variable	:	A normal variable with mean 0 and variance 1
Statistic	:	Function of sample
Statistic	:	Any fuction of a sample
Statistical hypothesis	:	An assumption about an unknown population parameter or distribution.
Students t variable	:	Ratio of standard normal variable to the square root of the ratio of Chi-square to its degrees of freedom.
Sufficiency	:	A property of the estimator, where the estimator processes all the information about the population, contained in the sample.
t	:	Student's t variable
t - test	:	A test in which the test statistic follows a t - distribution. Usually used for testing the mean of normal population when standard deviation of the population is unknown.
Test statistic	:	The statistic used for testing a parameter.
Time series	:	A sequence of data arranged in chronological order.
Trend	:	A long-run general direction of a time series over a period of several years.
Type I error	:	Reject the null hypothesis when it is true.
Type II Error	:	Do not reject the null hypothesis when it is false.
UCL	:	Upper Control Limit
Unbaisedness	:	A property of the estimator, where the expected value of the estimator coincides with the actual value of the parameter.
Upper Control Limit	:	The top line of a control chart
X - Bar Chart	:	A quality control chart which is used to monitor changes in average of a process.
Z	:	Standard Normal variable
Z - Score	:	The value of the normal variable transformed to standard normal.
Z - test	:	A test in which the test statistic follows a standard normal distribution. Usually used for testing mean of a population.
Z Transformation	:	Transformation of a normal variable to standard normal.

References

1. Principles of Statistics, Dr. S M Shukla and Dr.Sahai, Sahitya Bhavan Publications, Delhi
2. Mathematics and Statistics for Economics, G.S Monga, Vikas Publishing House pvt ltd
3. Fundamentals of Mathematical Statistics, S.C Gupta & V K Kapoor, Sultan Chand & sons Educational Publishers.
4. Fundamentals of Statistics, D N Elhance, Veena Elhance & B.L. Agarwal, Kitab Mahal Publishers.
5. Statistical Methods, S. P. Gupta, Sultan Chand & Sons, New Delhi.
6. Applied General Statistics, Frederick E Croxton, Dudley J Cowden, Sidney Klein, Prentice Hall India.
7. Fundamentals of Mathematical Statistics: S C Gupta, V K Kapoor , Sulthan Chand & sons, New Delhi
8. Business Statistics , Naval Bajpai , Pearson Educational Publications
9. Practical Statistics ,R.S.N Pillai & Bagavathi
10. Programmed Statistics, B L Agarwal, New Age Publishers, Delhi
11. Elementary Statistical Methods, S.P. Gupta ,Sultan Chand & sons Publishing co.
12. Introduction to Statistics, R.P.Hooda
13. Elementary Statistics A step by step Approach, Allan G Bluman, McGraw Hill Publishers.
14. Elementary Statistics and Indian Economic development -T.R. Jain,V.K.Ohri
15. Statistics for Management and Economics, Gerald Keller & Brian Warrack, Eastern Economy Edition
16. Statistics for Management, Richard.J.Levin & David S Rubin, Eastern Economy Edition
17. Statistics, David Freedman, Robert Pisani & Roger Purves,w.w. Norton & Company Inc , Viva Books Pvt Ltd, Delhi
18. Probability and Statistics for Engineers,G S S Bhishma Rao, SCITECH Publishers.
19. Schaum's Outlines, Statistics Murray R Spiegal & Larry J Stephens, Metric Editions,Schaum's Publishing Company, New York .
20. Complete Business Statistics,Aniz Daczel & Jayavel Sounderpandian, The MaGraw-Hill Publishing company, Delhi.
21. Introductory statistics, Prem. S. Mann, John Wiley & sons,Inc
22. An introduction to probability and statistics, V,K. Rohatgy, John Wiley & sons

Statistical Tables



Standard Normal Table

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.0000	0.0040	0.0080	0.0120	0.0160	0.0199	0.0239	0.0279	0.0319	0.0359
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753
0.2	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141
0.3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.1480	0.1517
0.4	0.1554	0.1591	0.1628	0.1664	0.1700	0.1736	0.1772	0.1808	0.1844	0.1879
0.5	0.1915	0.1950	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.2190	0.2224
0.6	0.2257	0.2291	0.2324	0.2357	0.2389	0.2422	0.2454	0.2486	0.2517	0.2549
0.7	0.2580	0.2611	0.2642	0.2673	0.2704	0.2734	0.2764	0.2794	0.2823	0.2852
0.8	0.2881	0.2910	0.2939	0.2967	0.2995	0.3023	0.3051	0.3078	0.3106	0.3133
0.9	0.3159	0.3186	0.3212	0.3238	0.3264	0.3289	0.3315	0.3340	0.3365	0.3389
1.0	0.3413	0.3438	0.3461	0.3485	0.3508	0.3531	0.3554	0.3577	0.3599	0.3621
1.1	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.3770	0.3790	0.3810	0.3830
1.2	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3962	0.3980	0.3997	0.4015
1.3	0.4032	0.4049	0.4066	0.4082	0.4099	0.4115	0.4131	0.4147	0.4162	0.4177
1.4	0.4192	0.4207	0.4222	0.4236	0.4251	0.4265	0.4279	0.4292	0.4306	0.4319
1.5	0.4332	0.4345	0.4357	0.4370	0.4382	0.4394	0.4406	0.4418	0.4429	0.4441
1.6	0.4452	0.4463	0.4474	0.4484	0.4495	0.4505	0.4515	0.4525	0.4535	0.4545
1.7	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599	0.4608	0.4616	0.4625	0.4633
1.8	0.4641	0.4649	0.4656	0.4664	0.4671	0.4678	0.4686	0.4693	0.4699	0.4706
1.9	0.4713	0.4719	0.4726	0.4732	0.4738	0.4744	0.4750	0.4756	0.4761	0.4767
2.0	0.4772	0.4778	0.4783	0.4788	0.4793	0.4798	0.4803	0.4808	0.4812	0.4817
2.1	0.4821	0.4826	0.4830	0.4834	0.4838	0.4842	0.4846	0.4850	0.4854	0.4857
2.2	0.4861	0.4864	0.4868	0.4871	0.4875	0.4878	0.4881	0.4884	0.4887	0.4890
2.3	0.4893	0.4896	0.4898	0.4901	0.4904	0.4906	0.4909	0.4911	0.4913	0.4916
2.4	0.4918	0.4920	0.4922	0.4925	0.4927	0.4929	0.4931	0.4932	0.4934	0.4936
2.5	0.4938	0.4940	0.4941	0.4943	0.4945	0.4946	0.4948	0.4949	0.4951	0.4952
2.6	0.4953	0.4955	0.4956	0.4957	0.4959	0.4960	0.4961	0.4962	0.4963	0.4964
2.7	0.4965	0.4966	0.4967	0.4968	0.4969	0.4970	0.4971	0.4972	0.4973	0.4974
2.8	0.4974	0.4975	0.4976	0.4977	0.4977	0.4978	0.4979	0.4979	0.4980	0.4981
2.9	0.4981	0.4982	0.4982	0.4983	0.4984	0.4984	0.4985	0.4985	0.4986	0.4986
3.0	0.4987	0.4987	0.4987	0.4988	0.4988	0.4989	0.4989	0.4989	0.4990	0.4990
3.1	0.4990	0.4991	0.4991	0.4991	0.4992	0.4992	0.4992	0.4992	0.4993	0.4993
3.2	0.4993	0.4993	0.4994	0.4994	0.4994	0.4994	0.4994	0.4995	0.4995	0.4995
3.3	0.4995	0.4995	0.4995	0.4996	0.4996	0.4996	0.4996	0.4996	0.4996	0.4997
3.4	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4998
3.5	0.4998	0.4998	0.4998	0.4998	0.4998	0.4998	0.4998	0.4998	0.4998	0.4998
3.6	0.4998	0.4998	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999
3.7	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999
3.8	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999
3.9	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000
4.0	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000

Table of t - distribution

Degrees of Freedom	Level of Significance (α)								
Two Tailed	**0.5**	**0.25**	**0.2**	**0.1**	**0.05**	**0.025**	**0.02**	**0.01**	**0.005**
One Tailed	**0.25**	**0.125**	**0.1**	**0.05**	**0.025**	**0.0125**	**0.01**	**0.005**	**0.0025**
1	1.000	2.414	3.078	6.314	12.706	25.452	31.821	63.657	127.321
2	0.816	1.604	1.886	2.920	4.303	6.205	6.965	9.925	14.089
3	0.765	1.423	1.638	2.353	3.182	4.177	4.541	5.841	7.453
4	0.741	1.344	1.533	2.132	2.776	3.495	3.747	4.604	5.598
5	0.727	1.301	1.476	2.015	2.571	3.163	3.365	4.032	4.773
6	0.718	1.273	1.440	1.943	2.447	2.969	3.143	3.707	4.317
7	0.711	1.254	1.415	1.895	2.365	2.841	2.998	3.499	4.029
8	0.706	1.240	1.397	1.860	2.306	2.752	2.896	3.355	3.833
9	0.703	1.230	1.383	1.833	2.262	2.685	2.821	3.250	3.690
10	0.700	1.221	1.372	1.812	2.228	2.634	2.764	3.169	3.581
11	0.697	1.214	1.363	1.796	2.201	2.593	2.718	3.106	3.497
12	0.695	1.209	1.356	1.782	2.179	2.560	2.681	3.055	3.428
13	0.694	1.204	1.350	1.771	2.160	2.533	2.650	3.012	3.372
14	0.692	1.200	1.345	1.761	2.145	2.510	2.624	2.977	3.326
15	0.691	1.197	1.341	1.753	2.131	2.490	2.602	2.947	3.286
16	0.690	1.194	1.337	1.746	2.120	2.473	2.583	2.921	3.252
17	0.689	1.191	1.333	1.740	2.110	2.458	2.567	2.898	3.222
18	0.688	1.189	1.330	1.734	2.101	2.445	2.552	2.878	3.197
19	0.688	1.187	1.328	1.729	2.093	2.433	2.539	2.861	3.174
20	0.687	1.185	1.325	1.725	2.086	2.423	2.528	2.845	3.153
21	0.686	1.183	1.323	1.721	2.080	2.414	2.518	2.831	3.135
22	0.686	1.182	1.321	1.717	2.074	2.405	2.508	2.819	3.119
23	0.685	1.180	1.319	1.714	2.069	2.398	2.500	2.807	3.104
24	0.685	1.179	1.318	1.711	2.064	2.391	2.492	2.797	3.091
25	0.684	1.178	1.316	1.708	2.060	2.385	2.485	2.787	3.078
26	0.684	1.177	1.315	1.706	2.056	2.379	2.479	2.779	3.067
27	0.684	1.176	1.314	1.703	2.052	2.373	2.473	2.771	3.057
28	0.683	1.175	1.313	1.701	2.048	2.368	2.467	2.763	3.047
29	0.683	1.174	1.311	1.699	2.045	2.364	2.462	2.756	3.038
30	0.683	1.173	1.310	1.697	2.042	2.360	2.457	2.750	3.030
40	0.681	1.167	1.303	1.684	2.021	2.329	2.423	2.704	2.971
60	0.679	1.162	1.296	1.671	2.000	2.299	2.390	2.660	2.915
120	0.677	1.156	1.289	1.658	1.980	2.270	2.358	2.617	2.860
Infinite	0.674	1.150	1.282	1.645	1.960	2.241	2.326	2.576	2.807

Table of Chi - square distribution

df \ α	0.005	0.01	0.025	0.05	0.1	0.2	0.25	0.5	0.75	0.8	0.9	0.95	0.975	0.99	0.995
1	7.8794	6.6349	5.0239	3.8415	2.7055	1.6424	1.3233	0.4549	0.1015	0.0642	0.0158	0.0039	0.0010	0.0002	0.0000
2	10.5966	9.2103	7.3778	5.9915	4.6052	3.2189	2.7726	1.3863	0.5754	0.4463	0.2107	0.1026	0.0506	0.0201	0.0100
3	12.8382	11.3449	9.3484	7.8147	6.2514	4.6416	4.1083	2.3660	1.2125	1.0052	0.5844	0.3518	0.2158	0.1148	0.0717
4	14.8603	13.2767	11.1433	9.4877	7.7794	5.9886	5.3853	3.3567	1.9226	1.6488	1.0636	0.7107	0.4844	0.2971	0.2070
5	16.7496	15.0863	12.8325	11.0705	9.2364	7.2893	6.6257	4.3515	2.6746	2.3425	1.6103	1.1455	0.8312	0.5543	0.4117
6	18.5476	16.8119	14.4494	12.5916	10.6446	8.5581	7.8408	5.3481	3.4546	3.0701	2.2041	1.6354	1.2373	0.8721	0.6757
7	20.2777	18.4753	16.0128	14.0671	12.0170	9.8032	9.0371	6.3458	4.2549	3.8223	2.8331	2.1673	1.6899	1.2390	0.9893
8	21.9550	20.0902	17.5345	15.5073	13.3616	11.0301	10.2189	7.3441	5.0706	4.5936	3.4895	2.7326	2.1797	1.6465	1.3444
9	23.5894	21.6660	19.0228	16.9190	14.6837	12.2421	11.3888	8.3428	5.8988	5.3801	4.1682	3.3251	2.7004	2.0879	1.7349
10	25.1882	23.2093	20.4832	18.3070	15.9872	13.4420	12.5489	9.3418	6.7372	6.1791	4.8652	3.9403	3.2470	2.5582	2.1559
11	26.7568	24.7250	21.9200	19.6751	17.2750	14.6314	13.7007	10.3410	7.5841	6.9887	5.5778	4.5748	3.8157	3.0535	2.6032
12	28.2995	26.2170	23.3367	21.0261	18.5493	15.8120	14.8454	11.3403	8.4384	7.8073	6.3038	5.2260	4.4038	3.5706	3.0738
13	29.8195	27.6882	24.7356	22.3620	19.8119	16.9848	15.9839	12.3398	9.2991	8.6339	7.0415	5.8919	5.0088	4.1069	3.5650
14	31.3193	29.1412	26.1189	23.6848	21.0641	18.1508	17.1169	13.3393	10.1653	9.4673	7.7895	6.5706	5.6287	4.6604	4.0747
15	32.8013	30.5779	27.4884	24.9958	22.3071	19.3107	18.2451	14.3389	11.0365	10.3070	8.5468	7.2609	6.2621	5.2293	4.6009
16	34.2672	31.9999	28.8454	26.2962	23.5418	20.4651	19.3689	15.3385	11.9122	11.1521	9.3122	7.9616	6.9077	5.8122	5.1422
17	35.7185	33.4087	30.1910	27.5871	24.7690	21.6146	20.4887	16.3382	12.7919	12.0023	10.0852	8.6718	7.5642	6.4078	5.6972
18	37.1565	34.8053	31.5264	28.8693	25.9894	22.7595	21.6049	17.3379	13.6753	12.8570	10.8649	9.3905	8.2307	7.0149	6.2648
19	38.5823	36.1909	32.8523	30.1435	27.2036	23.9004	22.7178	18.3377	14.5620	13.7158	11.6509	10.1170	8.9065	7.6327	6.8440
20	39.9968	37.5662	34.1696	31.4104	28.4120	25.0375	23.8277	19.3374	15.4518	14.5784	12.4426	10.8508	9.5908	8.2604	7.4338
21	41.4011	38.9322	35.4789	32.6706	29.6151	26.1711	24.9348	20.3372	16.3444	15.4446	13.2396	11.5913	10.2829	8.8972	8.0337
22	42.7957	40.2894	36.7807	33.9244	30.8133	27.3015	26.0393	21.3370	17.2396	16.3140	14.0415	12.3380	10.9823	9.5425	8.6427
23	44.1813	41.6384	38.0756	35.1725	32.0069	28.4288	27.1413	22.3369	18.1373	17.1865	14.8480	13.0905	11.6886	10.1957	9.2604
24	45.5585	42.9798	39.3641	36.4150	33.1962	29.5533	28.2412	23.3367	19.0373	18.0618	15.6587	13.8484	12.4012	10.8564	9.8862
25	46.9279	44.3141	40.6465	37.6525	34.3816	30.6752	29.3389	24.3366	19.9393	18.9398	16.4734	14.6114	13.1197	11.5240	10.5197
26	48.2899	45.6417	41.9232	38.8851	35.5632	31.7946	30.4346	25.3365	20.8434	19.8202	17.2919	15.3792	13.8439	12.1981	11.1602
27	49.6449	46.9629	43.1945	40.1133	36.7412	32.9117	31.5284	26.3363	21.7494	20.7030	18.1139	16.1514	14.5734	12.8785	11.8076
28	50.9934	48.2782	44.4608	41.3371	37.9159	34.0266	32.6205	27.3362	22.6572	21.5880	18.9392	16.9279	15.3079	13.5647	12.4613
29	52.3356	49.5879	45.7223	42.5570	39.0875	35.1394	33.7109	28.3361	23.5666	22.4751	19.7677	17.7084	16.0471	14.2565	13.1211
30	53.6720	50.8922	46.9792	43.7730	40.2560	36.2502	34.7997	29.3360	24.4776	23.3641	20.5992	18.4927	16.7908	14.9535	13.7867

F - table (5% level of significance)

n_2 \ n_1	1	2	3	4	5	6	7	8	9	10	12	15	20	30	120	∞
1	161.45	199.50	215.71	224.58	230.16	233.99	236.77	238.88	240.54	241.88	243.91	245.95	248.01	250.10	253.25	254.31
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40	19.41	19.43	19.45	19.46	19.49	19.50
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.74	8.70	8.66	8.62	8.55	8.53
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.91	5.86	5.80	5.75	5.66	5.63
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.68	4.62	4.56	4.50	4.40	4.36
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.00	3.94	3.87	3.81	3.70	3.67
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.57	3.51	3.44	3.38	3.27	3.23
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.28	3.22	3.15	3.08	2.97	2.93
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.07	3.01	2.94	2.86	2.75	2.71
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.91	2.85	2.77	2.70	2.58	2.54
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.79	2.72	2.65	2.57	2.45	2.40
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.69	2.62	2.54	2.47	2.34	2.30
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.60	2.53	2.46	2.38	2.25	2.21
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.53	2.46	2.39	2.31	2.18	2.13
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.48	2.40	2.33	2.25	2.11	2.07
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.42	2.35	2.28	2.19	2.06	2.01
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.38	2.31	2.23	2.15	2.01	1.96
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.34	2.27	2.19	2.11	1.97	1.92
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.31	2.23	2.16	2.07	1.93	1.88
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.28	2.20	2.12	2.04	1.90	1.84
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30	2.23	2.15	2.07	1.98	1.84	1.78
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25	2.18	2.11	2.03	1.94	1.79	1.73
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22	2.15	2.07	1.99	1.90	1.75	1.69
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19	2.12	2.04	1.96	1.87	1.71	1.65
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.09	2.01	1.93	1.84	1.68	1.62
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99	1.92	1.84	1.75	1.65	1.47	1.39
120	3.92	3.07	2.68	2.45	2.29	2.18	2.09	2.02	1.96	1.91	1.83	1.75	1.66	1.55	1.35	1.25
∞	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88	1.83	1.75	1.67	1.57	1.46	1.22	1.00

F - table (1% level of significance)

n_2 \ n_1	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	60	120	Inf
1	4052.18	4999.50	5403.35	5624.58	5763.65	5858.99	5928.36	5981.07	6022.47	6055.85	6083.32	6106.32	6125.86	6142.67	6157.28	6170.10	6181.43	6191.53	6200.58	6208.73	6216.12	6222.84	6228.99	6234.63	6239.83	6244.62	6249.07	6253.20	6257.05	6260.65	6313.03	6339.39	6365.86
2	98.50	99.00	99.17	99.25	99.30	99.33	99.36	99.37	99.39	99.40	99.41	99.42	99.42	99.43	99.43	99.44	99.44	99.44	99.45	99.45	99.45	99.45	99.46	99.46	99.46	99.46	99.46	99.46	99.46	99.47	99.48	99.49	99.50
3	34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49	27.35	27.23	27.13	27.05	26.98	26.92	26.87	26.83	26.79	26.75	26.72	26.69	26.66	26.64	26.62	26.60	26.58	26.56	26.55	26.53	26.52	26.50	26.32	26.22	26.13
4	21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80	14.66	14.55	14.45	14.37	14.31	14.25	14.20	14.15	14.11	14.08	14.05	14.02	13.99	13.97	13.95	13.93	13.91	13.89	13.88	13.86	13.85	13.84	13.65	13.56	13.46
5	16.26	13.27	12.06	11.39	10.97	10.67	10.46	10.29	10.16	10.05	9.96	9.89	9.82	9.77	9.72	9.68	9.64	9.61	9.58	9.55	9.53	9.51	9.49	9.47	9.45	9.43	9.42	9.40	9.39	9.38	9.20	9.11	9.02
6	13.75	10.92	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87	7.79	7.72	7.66	7.60	7.56	7.52	7.48	7.45	7.42	7.40	7.37	7.35	7.33	7.31	7.30	7.28	7.27	7.25	7.24	7.23	7.06	6.97	6.88
7	12.25	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62	6.54	6.47	6.41	6.36	6.31	6.28	6.24	6.21	6.18	6.16	6.13	6.11	6.09	6.07	6.06	6.04	6.03	6.02	6.00	5.99	5.82	5.74	5.65
8	11.26	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81	5.73	5.67	5.61	5.56	5.52	5.48	5.44	5.41	5.38	5.36	5.34	5.32	5.30	5.28	5.26	5.25	5.23	5.22	5.21	5.20	5.03	4.95	4.86
9	10.56	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35	5.26	5.18	5.11	5.05	5.01	4.96	4.92	4.89	4.86	4.83	4.81	4.79	4.77	4.75	4.73	4.71	4.70	4.68	4.67	4.66	4.65	4.48	4.40	4.31
10	10.04	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85	4.77	4.71	4.65	4.60	4.56	4.52	4.49	4.46	4.43	4.41	4.38	4.36	4.34	4.33	4.31	4.30	4.28	4.27	4.26	4.25	4.08	4.00	3.91
11	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63	4.54	4.46	4.40	4.34	4.29	4.25	4.21	4.18	4.15	4.12	4.10	4.08	4.06	4.04	4.02	4.01	3.99	3.98	3.96	3.95	3.94	3.78	3.69	3.60
12	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	4.30	4.22	4.16	4.10	4.05	4.01	3.97	3.94	3.91	3.88	3.86	3.84	3.82	3.80	3.78	3.76	3.75	3.74	3.72	3.71	3.70	3.54	3.45	3.36
13	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19	4.10	4.02	3.96	3.91	3.86	3.82	3.78	3.75	3.72	3.69	3.66	3.64	3.62	3.60	3.59	3.57	3.56	3.54	3.53	3.52	3.51	3.34	3.25	3.17
14	8.86	6.51	5.56	5.04	4.69	4.46	4.28	4.14	4.03	3.94	3.86	3.80	3.75	3.70	3.66	3.62	3.59	3.56	3.53	3.51	3.48	3.46	3.44	3.43	3.41	3.40	3.38	3.37	3.36	3.35	3.18	3.09	3.00
15	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89	3.80	3.73	3.67	3.61	3.56	3.52	3.49	3.45	3.42	3.40	3.37	3.35	3.33	3.31	3.29	3.28	3.26	3.25	3.24	3.23	3.21	3.05	2.96	2.87
16	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78	3.69	3.62	3.55	3.50	3.45	3.41	3.37	3.34	3.31	3.28	3.26	3.24	3.22	3.20	3.18	3.16	3.15	3.14	3.12	3.11	3.10	2.93	2.84	2.75
17	8.40	6.11	5.18	4.67	4.34	4.10	3.93	3.79	3.68	3.59	3.52	3.46	3.40	3.35	3.31	3.27	3.24	3.21	3.19	3.16	3.14	3.12	3.10	3.08	3.07	3.05	3.04	3.03	3.01	3.00	2.83	2.75	2.65
18	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60	3.51	3.43	3.37	3.32	3.27	3.23	3.19	3.16	3.13	3.10	3.08	3.05	3.03	3.02	3.00	2.98	2.97	2.95	2.94	2.93	2.92	2.75	2.66	2.57
19	8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52	3.43	3.36	3.30	3.24	3.19	3.15	3.12	3.08	3.05	3.03	3.00	2.98	2.96	2.94	2.92	2.91	2.89	2.88	2.87	2.86	2.84	2.67	2.58	2.49
20	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46	3.37	3.29	3.23	3.18	3.13	3.09	3.05	3.02	2.99	2.96	2.94	2.92	2.90	2.88	2.86	2.84	2.83	2.81	2.80	2.79	2.78	2.61	2.52	2.42
21	8.02	5.78	4.87	4.37	4.04	3.81	3.64	3.51	3.40	3.31	3.24	3.17	3.12	3.07	3.03	2.99	2.96	2.93	2.90	2.88	2.86	2.84	2.82	2.80	2.79	2.77	2.76	2.74	2.73	2.72	2.55	2.46	2.36
22	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35	3.26	3.18	3.12	3.07	3.02	2.98	2.94	2.91	2.88	2.85	2.83	2.81	2.78	2.77	2.75	2.73	2.72	2.70	2.69	2.68	2.67	2.50	2.40	2.31
23	7.88	5.66	4.76	4.26	3.94	3.71	3.54	3.41	3.30	3.21	3.14	3.07	3.02	2.97	2.93	2.89	2.86	2.83	2.80	2.78	2.76	2.74	2.72	2.70	2.69	2.67	2.66	2.64	2.63	2.62	2.45	2.35	2.26
24	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26	3.17	3.09	3.03	2.98	2.93	2.89	2.85	2.82	2.79	2.76	2.74	2.72	2.70	2.68	2.66	2.64	2.63	2.61	2.60	2.59	2.58	2.40	2.31	2.21
25	7.77	5.57	4.68	4.18	3.85	3.63	3.46	3.32	3.22	3.13	3.06	2.99	2.94	2.89	2.85	2.81	2.78	2.75	2.72	2.70	2.68	2.66	2.64	2.62	2.60	2.59	2.58	2.56	2.55	2.54	2.36	2.27	2.17
26	7.72	5.53	4.64	4.14	3.82	3.59	3.42	3.29	3.18	3.09	3.02	2.96	2.90	2.86	2.81	2.78	2.75	2.72	2.69	2.66	2.64	2.62	2.60	2.58	2.57	2.55	2.54	2.53	2.51	2.50	2.33	2.23	2.13
27	7.68	5.49	4.60	4.11	3.78	3.56	3.39	3.26	3.15	3.06	2.99	2.93	2.87	2.82	2.78	2.75	2.71	2.68	2.66	2.63	2.61	2.59	2.57	2.55	2.54	2.52	2.51	2.49	2.48	2.47	2.29	2.20	2.10
28	7.64	5.45	4.57	4.07	3.75	3.53	3.36	3.23	3.12	3.03	2.96	2.90	2.84	2.79	2.75	2.72	2.68	2.65	2.63	2.60	2.58	2.56	2.54	2.52	2.51	2.49	2.48	2.46	2.45	2.44	2.26	2.17	2.06
29	7.60	5.42	4.54	4.04	3.73	3.50	3.33	3.20	3.09	3.00	2.93	2.87	2.81	2.77	2.73	2.69	2.66	2.63	2.60	2.57	2.55	2.53	2.51	2.49	2.48	2.46	2.45	2.44	2.42	2.41	2.23	2.14	2.03
30	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07	2.98	2.91	2.84	2.79	2.74	2.70	2.66	2.63	2.60	2.57	2.55	2.53	2.51	2.49	2.47	2.45	2.44	2.42	2.41	2.40	2.39	2.21	2.11	2.01
60	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72	2.63	2.56	2.50	2.44	2.39	2.35	2.31	2.28	2.25	2.22	2.20	2.17	2.15	2.13	2.12	2.10	2.08	2.07	2.05	2.04	2.03	1.84	1.73	1.60
120	6.85	4.79	3.95	3.48	3.17	2.96	2.79	2.66	2.56	2.47	2.40	2.34	2.28	2.23	2.19	2.15	2.12	2.09	2.06	2.03	2.01	1.99	1.97	1.95	1.93	1.92	1.90	1.89	1.87	1.86	1.66	1.53	1.38
inf	6.63	4.61	3.78	3.32	3.02	2.80	2.64	2.51	2.41	2.32	2.25	2.18	2.13	2.08	2.04	2.00	1.97	1.93	1.90	1.88	1.85	1.83	1.81	1.79	1.77	1.76	1.74	1.72	1.71	1.70	1.47	1.32	1.00

Constants for control charts (SQC)

Sample Size	Mean Chart		R Chart		σ Chart	
	A_1	A_2	D_3	D_4	B_3	B_4
2	3.760	1.880	0	3.262	0	3.267
3	2.394	1.023	0	2.575	0	2.568
4	1.880	0.729	0	2.282	0	2.266
5	1.596	0.577	00	2.115	0	2.089
6	1.410	0.483	0	2.004	0.030	1.970
7	1/277	0.419	0.076	1.924	0.118	1.882
8	1.175	0.373	0.136	1.864	0.185	1.815
9	1.094	0.337	0.184	1.816	0.239	1.761
10	1.028	0.308	0.223	1.777	0.284	1.761

Poisson distribution – Values of $e^{-\lambda}$

λ	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
0	1.0000	0.9048	0.8187	0.7408	0.6703	0.6065	0.5488	0.4966	0.4493	0.4066
1	0.3679	0.3329	0.3012	0.2725	0.2466	0.2231	0.2019	0.1827	0.1653	0.1496
2	0.1353	0.1220	0.1108	0.1003	0.0907	0.0821	0.0743	0.0672	0.0608	0.0550
3	0.0498	0.0450	0.0408	0.0369	0.0334	0.0302	0.0273	0.0247	0.0224	0.0202
4	0.0183	0.0166	0.0150	0.0136	0.0123	0.0111	0.0101	0.0091	0.0082	0.0074
5	0.0067	0.0061	0.0055	0.0050	0.0045	0.0041	0.0037	0.0033	0.0030	0.0027
6	0.0025	0.0022	0.0020	0.0018	0.0017	0.0015	0.0014	0.0012	0.0011	0.0010
7	0.0009	0.0008	0.0007	0.0007	0.0006	0.0006	0.0005	0.0005	0.0004	0.0004

Example: $e^{-3}=0.0498$, $e^{-0.1}=0.9048$ and $e^{-4.5} = 0.0111$

$e^{-\lambda}$ values for two decimals

λ	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
$e^{-\lambda}$	0.9900	0.9802	0.9704	0.9608	0.9512	0.9418	0.9324	0.9231	0.9139

Example: $e^{-1.23}= e^{-1.2} \times e^{-0.03}=0.3012 \times 0.9704=0.2923$

We already discussed R software in Plus One Class. In continuation with that here also we introduce R codes of some important statistical functions.

Some important Remarks: *In* ***R codes*** *any statement after the character # will be ignored by the computer. Such statements are used for illustrating the respective codes. In this tutorial the* ***R codes*** *are printed in font style* `courier New`.

1. Correlation

The **R codes** plot gives the scatter diagram and cor computes the correlation coefficient between two set of observations.

Example: Following are hours worked and grade received for 10 students taking some class. Find Karl Pearson's correlation coefficient between hours and grade. Also find the Spearman's coefficient.

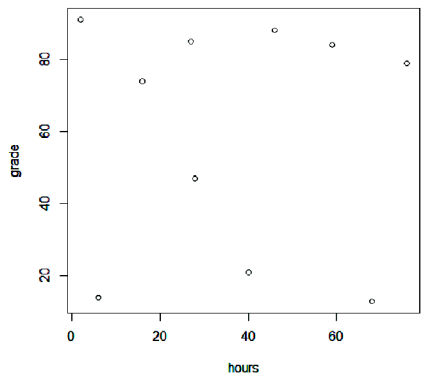
hours : 2,76,40,6,16,28,27,59,46,68

grade : 91,79,21,14,74,47,85,84,88,13

R code

```
hours= c(2,76,40,6,16,28,27,59,46,68) #enter
variable: hours
grade= c(91,79,21,14,74,47,85,84,88,13)   #enter
variable:grade
plot(hours,grade)  #plots the scatter diagram
```

Output



R code(Conts.)

```
r=cor(hours,grade)#compute correlation coefficient
r#prints value of correlation coefficient
```

Output

[1] -0.01871495

R code(Conts.)

```
s=cor(hours,grade,method = "spearman")#compute
spearman's correlation coefficient
s #prints value of spearman's correlation
coefficient
```

Output

[1] -0.1878788

2. Regression

The **R codes** lm fits linear regression model and summary prints the summary of linear egression including regression coefficient and more.

Example: Find the regression equation of "grade" on "hours" in example of section 1.

R code

```
model=lm(grade~hours)# fits linear regression of
grade on hours
summary(model)# gives the summary of regression.
```

Output

```
Call:
```

lm(formula = grade ~ hours)

Residuals:

Min	1Q	Median	3Q	Max
-46.33	-32.09	17.12	25.11	30.57

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	60.47739	19.83815	3.049	0.0159 *
hours	-0.02384	0.45033	-0.053	0.9591

Signif. codes: 0 _S***_S 0.001 _S**_S 0.01 _S*_S 0.05 _S._S 0.1 _S _S1

Residual standard error: 34.48 on 8 degrees of freedom

Multiple R-squared: 0.0003502, Adjusted R-squared: -0.1246

F-statistic: 0.002803 on 1 and 8 DF, p-value: 0.9591

Conclusions: From this one can read the regression coefficient of grade on hours as -0.02384 and regression equation as

grade=60.47739+ (-0.02384)hours

Remark: To find the regression equation of "hours" on "grade" use R code as

```
model=lm(hours~  grade)
summary(model)
```

3. Binomial Distribution

(a) The code dbinom gives the probability of a specified number of successes.

R Code

```
dbinom(3,10,0.5)  # gives the probability of 3
successes in binomial distribution with parameters
n=10 and success probability=0.5
```

Output

[1] 0.1171875

(b) The code pbinom gives the cumulative probability up to a specified number of successes.

R Code

```
pbinom(3,10,0.5)  # gives the probability of 1 or 2
or 3 successes in binomial distribution with
parameters n=10 and success probability=0.5
```

Output

[1] 0.171875

(c) The code rbinom produces 'number of successes'

R Code

```
rbinom(25,10,0.5) # gives number of successes in 25
repetition of binomial trials with parameters n=10
and success probability=0.5
```

Output

[1] 2 5 7 5 2 5 6 5 1 3 7 5 4 9 2 7 7 5 5 4 7 5 6 5 6

4. Poisson Distribution

(a) The code dpois gives the probability of a specified number of events.

R Code

```
dpois(5,3) # gives the probability of 5 events in
poisson distribution with parameter λ=3.
```

Output

[1] 0.1008188

(b) The code ppois gives the cumulative probability up to a specified number of events.

R Code

```
ppois(5,3) # gives the probability of 1 or 2 or 3
or 4 or 5 events in poisson distribution with
parameter λ=3.
```

Output

[1] 0.9160821

(c) The code rpois produces 'number of events'

R Code

```
rpois(25,3) # gives number of events in 25 repetition
of poisson distribution with parameter λ=3.
```

Output

```
[1] 1 3 4 3 6 3 4 4 4 7 3 4 3 5 5 3 1 4 1 4 2 2 6 3 2
```

5. Normal Distribution

(a) The code dnorm gives the density of a specified value of normal random variable.

R Code

```
dnorm(98,100,2.5) # gives the density at x= 98  in
a normal distribution with parameters mean=100 and
standard deviation=2.5.
```

Output

[1] 0.1158766

(b) The code pnorm gives the cumulative probability up to a specified value.

R Code

```
pnorm(98,100,2.5) # gives the cumulative
probability up to 98
```

Output

[1] 0.2118554

(c) The code rnorm produces normal values (that is a random sample from normal distribution)

R Code

```
rnorm(25,100,2.5) # gives 25 normal values in a
normal distribution with parameters mean=100 and
standard deviation=2.5.
```

Output

```
[1]  98.17180  97.14955 102.79323  99.32077 104.61441 101.25060  98.53690
[8]  100.02138  96.73600 100.94530 101.83515  99.10494 101.10174  99.65576
[15] 100.09922 100.42579 101.65153 101.72664  99.52983 101.61313  93.79264
[22] 96.58186 102.11653 101.69954  98.47495
```

Remark: If the parameters are not specified then outputs are for standard normal distribution. For example,

R Code

```
rnorm(25) # gives 25 normal values in a normal
distribution with parameters mean=0 and standard
deviation=1.
```

Output

[1] 0.05959739 -0.02127060 0.67488057 0.38826038 -0.20046064 -0.51904706

[7] 1.96104590 1.09615971 -0.73990959 1.83091371 0.52441002 -0.08380924

[13] -1.22225460 -0.15108502 -0.19495669 1.63919772 -1.20406924 1.79266343

[19] -1.37055720 -1.41156860 -1.29116645 -0.42900737 2.74226453 -0.12517713

[25] 1.03706320

6. Standard Normal Tables

Standard normal table gives

(i) The probability P ($Z \leq z$) for a specified value of z and

(ii) The value of z for the specified vale of P($Z \leq z$)

The R codes are respectively

pnorm and qnorm

Examples

```
1. pnorm(1.96)#gives the cumulative probability up
   to z=1.96
```

Output: [1] 0.9750021

```
2. qnorm(0.9750021) # gives the value of standard
   normal variable for which the cumulative probability
   is 0.9750021.
```

Output: [1] 1.96

7. Testing of Hypothesis

1. One sample t-test

The code t.test tests the hypothesis that the mean of a normal population has a specified value when the standard deviation is unknown.

Example(1): Suppose the extra hours of sleep in 15 patients after administering a new drug are: 0.7 -1.6 -0.2 -1.2 -0.1 3.4 3.7 0.8 0.0 2.0 1.9 0.8 1.1 0.1 -0.1. Can you conclude that average increase in sleeping hours by the drug is 1.5 hours.

R Code

```
xtra=c( 0.7,-1.6,-0.2,-1.2,-0.1, 3.4, 3.7, 0.8,0.0,
2.0,1.9,0.8,1.1,  0.1, -0.1)
t.test(xtra,mu=1.5)
```

Output

One Sample t-test

data: xtra

t = -1.931, df = 14, p-value = 0.07398

alternative hypothesis: true mean is not equal to 1.5

95 percent confidence interval:

-0.07598419 1.58265085

sample estimates: mean of x 0.7533333

Conclusion: Since p-value of the test is 0.07398 and is greater than 0.05, the null hypothesis is accepted at 5% level of significance. Also one can compare the computed value of the test statistic, namely, |t|=1.931 with table value of t with df=14, namely 2.14. Since |t| is less than the table value we accept the null hypothesis at 5% level of significance. Note that p-value of a test is a credibility measure of the null hypothesis.

Remark: In one sided alternative hypothesis the **R codes** are:

`t.test(xtra,mu=1.5,alternative="less")` -for left tailed test and `t.test(xtra,mu=1.5,alternative="greater")` - for right tailed test

2. Two sample t-test

The code t.test also tests the hypothesis that the mean of two normal populations are equal when the standard deviations are not known.

Example(2): Suppose the extra hours of sleep in a group of 10 patients after administering drug A and in another group of 12 patients after administering drug B are respectively as follows

Extra1= 0.7, -1.6, -0.2,-1.2,-0.1, 3.4, 3.7, 0.8, 0.0, 2.0

Extra2= 1.9, 0.8, 1.1, 0.1,-0.1, 4.4, 5.5, 1.6, 4.6, 3.4, 2.5, 0.5

Can you conclude that drug B is more effective than drug A in promoting sleeping time? Why?

R Code

```
xtra1=c(0.7, -1.6, -0.2,-1.2,-0.1, 3.4, 3.7, 0.8,
0.0, 2.0 )

xtra2=c(1.9, 0.8, 1.1, 0.1,-0.1, 4.4, 5.5, 1.6,
4.6, 3.4, 2.5, 0.5 )
t.test(xtra1,xtra2,alternative="less")
```

Output

Welch Two Sample t-test

data: xtra1 and xtra2

t = -1.835, df = 19.628, p-value = 0.04085

alternative hypothesis: true difference in means is less than 0

95 percent confidence interval:

-Inf -0.08539332

sample estimates:

mean of x mean of y

0.750000 2.191667

Conclusion: Since the p-value of the test is 0.04085 and is less than the level of significance 0.05, we reject the null hypothesis.

Remarks

1. In this method we assume that variances of two normal populations are not equal. Hence an approximation of the estimator of the pooled variance(called Welch approximation) is used in computation.
2. If, in case, the true variances are same we use the following R code:

   t.test(xtra1,xtra2,alternative="less",var.equal=TRUE)

3. Paired t-test

   The code t.test also tests the hypothesis that the mean of two normal populations are equal when the samples are depended. The **R code** to this test is

```
t.test(xvlues,yvalues,paired=TRUE)
```

4. Chi-square tests

   The code chisq.test performs the chi-squared test of independence in a contingency table for two attributes.

Example: Following table gives the summary of an opinion survey regarding popularity of candidates of Democrat, Independent and Republican parties among Male and Female voters. Test whether 'gender' and 'party affiliation' are associated.

	party		
gender	Democrat	Independent	Republican
M	762	327	468
F	484	239	477

R code

```
M = as.table(rbind(c(762, 327, 468), c(484, 239,
477)))#prepare the table.
dimnames(M) = list(gender = c("M","F")party =
c("Democrat","Independent", "Republican"))#add row
and column names to the table.
M #print the contingency table.#optional
chisq.test(M)  # Performs and prints the test
summary.
```

Output

Pearson's Chi-squared test

data: M

X-squared = 30.0701, df = 2, p-value = 2.954e-07

Conclusion: Since p-value is 0.0000002954 and is less than 0.05 the null hypothesis that the gender and party affiliation are independent is rejected. So the gender and party affiliation are dependent.