FUNDAMENTALS OF STATISTICS

Complementary course for

First

Semester

(2019 Admission)

SHRI VENKATESHWARA UNIVERSITY
SCHOOL OF DISTANCE EDUCATION

STUDY MATERIAL

I Semester

Complementary Course
FUNDAMENTALS OF STATISTICS

Prepared and Scrutinized by:

Dr. Mohit Sharma,
HOD,
Shri Venkateshwara University,
Gajraula (UP)

Layout & Settings
Computer Section, SDE

INDEX

Module 1

INTRODUCTION

The term statistics seems to have been derived from the Latin word *'status'* or Italian word *'statista'* or the German word *'statistic,* each of which means *political state*.

The word 'Statistics' is usually interpreted in two ways. The first sense in which the word is used is a plural noun just refer to a collection of numerical facts. The second is as a singular noun to denote the methods generally adopted in the collection and analysis of numerical facts. In the singular sense the term 'Statistics' is better described as statistical methods.

Different authors have defined statistics in different ways. According to Croxton and Cowden statistics may be defined as *''collection, organisation presentation, analysis and interpretation of numerical data''*

Population and sample

Population

An aggregate of individual items relating to a phenomenon under investigation is technically termed as 'population'. In other words a collection of objects pertaining to a phenomenon of statistical enquiry is referred to as population or universe. Suppose we want to collect data regarding the income of college teachers under University of Calicut,, then, the totality of these teachers is our population.

In a given population, the individual items are referred to as elementary units, elements or members of the population. The population has the statistical characteristic of being finite or infinite. When the number of units under investigation are determinable, it is called finite population. For example, the number of college teachers under Calicut University is a finite population. When the number of units in a phenomenon is indeterminable, eg, the number of stars in the sky, it is called an infinite population.

Sample

When few items are selected for statistical enquiry, from a given population it is called a 'sample'. A sample is the small part or subset of the population. Say, for instance, there may be 3000 workers in a factory. One wants to study their consumption pattern. By selecting only 300 workers from the group of 3000, sample for the study has been taken. This sample is not studied just for its own sake. The motive is to know the true state of the population. From the sample study statistical inference about the population can be done.

Census and sample Method

In any statistical investigation, one is interested in studying the population characteristics. This can be done either by studying the entire items in the population or on a part drawn from it. If we are studying each and every element of the population, the process is called *census method* and if we are studying only a sample, the process is called sample survey, *sample method* or *sampling*. For example, the Indian population census or a socio economic survey of a whole village by a college planning forum are examples of census studies. The national sample survey enquiries are examples of sample studies.

Advantages of Sampling

The sample method is comparatively more economical.

The sample method ensures completeness and a high degree of accuracy due to the small area of operation

It is possible to obtain more detailed information, in a sample survey than complete enumeration.

Sampling is also advocated where census is neither necessary nor desirable.

In some cases sampling is the only feasible method. For example, we have to test the sharpness of blades-if we test each blade, perhaps the whole of the product will be wasted; in such circumstances the census method will not be suitable. Under these circumstances sampling techniques will be more useful.

A sample survey is much more scientific than census because in it the extent of the reliability of the results can be known where as this is not always possible in census.

Variables and Attributes

A quantity which varies from one person to another or one time to another or one place to another is called a variable. It is actually a numerical value possessed by an item. For example, price of a given commodity, wages of workers, production and weights of students etc.

Attribute means a qualitative characteristic possessed by each individual in a group. It can't assume numerical values. For example, sex, honesty, colour etc.

This means that a variable will always be a quantitative characteristic. Data concerned with a quantitative variable is called *quantitative data* and the data corresponding to a qualitative variable is called *qualitative data*.

We can divide quantitative variables into two (i) discrete (ii) continuous. Those variables which can assume only distinct or particular values are called *discrete* or *discontinuous* variables. For example, the number of children per family, number rooms in a house etc. Those variables which can take any numerical value with in a certain range are known as *continuous* variables. Height of a boy is a continuous variable, for it changes continuously in a given range of heights of the boys. Similar is the case of weight,: production, price, demand, income, marks etc.

Types of Frequency Distribution

Erricker states "frequency distribution is a classification according to the number possessing the same values of the variables''. It is simply a table in which data are grouped into classes and the number of cases which fall in each class is recorded. Here the numbers are usually termed as 'frequencies'. There are discrete frequency distributions and continuous frequency distributions.

1. Discrete Frequency Distribution

If we have a large number of items in the data it is better to prepare a frequency array and condense the data further. Frequency array is prepared by listing once and consecutively all the values occurring in the series and noting the number of times each such value occurs. This is called discrete frequency distribution or ungrouped frequency distribution.

Illustration: The following data give the number of children per family in each of 25 families 1, 4, 3, 2, 1, 2, 0, 2, 1, 2, 3, 2, 1, 0, 2, 3, 0, 3, 2, 1, 2, 2, 1, 4, 2. Construct a frequency distribution.

No of children	Tally marks	No of families
0	|||	3
1	𝍸 |	6
2	𝍸 𝍸	10
3	||||	4
4	||	2
Total		25

2. Continuous Frequency Distribution

An important method of condensing and presenting data is that of the construction of a continuous frequency distribution or grouped frequency distribution. Here the data are classified according to class intervals.

The following are the rules generally adopted in forming a frequency table for a set of observations.

Note the difference between the largest and smallest value in the given set of observations

Determine the number classes into which the difference can be divided.

The classes should be mutually exclusive. That means they do not overlap.

Arrange a paper with 3 columns, classes, tally marks and frequency.

Write down the classes in the first column.

Go though the observations and put tally marks in the respective classes.

Write the sum of the tally marks of each class in the frequency column.

Note that the sum of the frequencies of all classes should be equal to the total number of observations.

Concepts of a Frequency Table

i. Class limits: The observations which constitute a class are called class limits. The left hand side observations are called lower limits and the right hand side observations are called upper limits.

Working classes: The classes of the form 0-9, 10-19, 20-29,... are called working classes or nominal classes. They are obtained by the inclusive method of classification where both the limits of a class are included in the same class itself.

Actual classes: If we are leaving either the upper limit or the lower limit from each class, it is called exclusive method of classification. The classes so obtained are called 'actual classes' or 'true classes'. The classes 0.5 - 9.5, 9.5 - 19.5, 19.5 - 29.5,... are the actual classes of the above working classes. The classes of the type 0-10, 10 20, 20 - 30,... are also treated as actual classes. There will be no break in the actual classes. We can convert working classes to the corresponding actual classes using the following steps.

Note the difference between one upper limit and the next lower limit.

Divide the difference by 2.

Subtract that value from the lower limits and add the same to the upper limits.

For example

Working Classes	Frequency	Actual Classes
1-2.9	2	0.95-2.95
3-4.9	8	2.95-4.95
5-6.9	10	4.95-6.95
7-8.9	5	6.95-8.95

Class boundaries: The class limits of the actual classes are called actual class limits or class boundaries.

Class mark: The class marks or mid value of classes is the average of the upper limit and lower limit of that class. The mid value of working classes and the corresponding actual classes are the same. For example, the class mark of the classes 0 - 9, 10 - 19, 20 - 29,... are respectively 4.5, 14.5, 24.5,...

Class interval: The class interval or width of a class is the difference between upper limit and lower limit of an actual class. It is better to note that the difference between the class limits of a working class is not the class interval. The class interval is usually denoted by ‘c’ or i or ‘h’.

Example

Construct a frequency distribution for the following data

70	45	33	64	50	25	65	74	30	20
55	60	65	58	52	36	45	42	35	40
51	47	39	61	53	59	49	41	20	55
46	48	52	64	48	45	65	78	53	42

Solution

Classes	Tally marks	Frequency
20-29	|||	3
30-39	𝍸	5
40-49	𝍸 𝍸 ||	12
50-59	𝍸 𝍸	10
60-69	𝍸 ||	7
70-79	|||	3
Total		40

Cumulative Frequency Distribution

An ordinary frequency distribution show the number of observations falling in each class. But there are instances where we want to know how many observations are lying below or above a particular value or in between two specified values. Such type of information is found in cumulative frequency distributions.

Cumulative frequencies are determined on either a less than basis or more than basis. Thus we get less than cumulative frequencies (<CF) and greater than or more than cumulative frequencies (>CF). Less than CF give the number of observations falling below the upper limit of a class and greater than CF give the number of observations lying above the lower limit of the class. Less than CF are obtained by adding successively the frequencies of all the previous classes including the class against which it is written. The cumulation is started from the lowest size of the class to the highest size, (usually from top to bottom). They are based on the upper limit of actual classes.

More than CF distribution is obtained by finding the cumulation or total of frequencies starting from the highest size of the class to the lowest class, (ie., from bottom to top) More than CF are based on the lower limit of the actual classes.

Classes	f	UL	<CF		LL	>CF	
0-10	2	10	2	2	0	3 + 7 + 10 + 8 + 5 + 2	35
10-20	5	20	2 + 5	7	10	3 + 7 + 10 + 8 + 5	33
20-30	8	30	2 + 5 + 8	15	20	3 + 7 + 10 + 8	28
30-40	10	40	2 + 5 + 8 + 10	25	30	3 + 7 + 10	20
40-50	7	50	2 + 5 + 8 + 10 + 7	32	40	3 + 10	13
50-60	3	60	2 + 5 + 8 + 10 + 7 + 3	35	50	3	3

EXERCISES

Multiple Choice Questions

A qualitative characterisic is also known as

a. attribute b. variable

c . variate d. frequency

A variable which assumes only integral values is called

a. continuous b. discrete

c . random d. None of these

An example of an attribute is

a. Height b. weight

c . age d. sex

Number of students having smoking habit is a variable which is

a. Continuous b. discrete

c . neither disrete nor continuous

None of these

A series showing the sets of all district values individually with their frequencies is known as

grouped frequency distribution

simple frequency distribution

cumulative frequency distribution

none of the above

A series showing the sets of all values in classes with their corersponding frequencies is knowsn as

grouped frequency distribution

simple frequency distribution

cumulative frequency distribution

none of the above

If the lower and upper limits of a class are 10 and 40 respectively, the mid points of the class is

a. 25.0 b. 12.5 c. 15.0 d. 30.0

13. In a grouped data, the number of classes preferred are

a. minimum possible b. adequate

c. maximum possible d. any arbitrarily chosen number

Class interval is measured as:

the sum of the upper and lower limit

half of the sum of upper and lower limit

half of the difference between upper and lower limit

the difference between upper and lower limit

A group frequency distribution with uncertain first or last classes is known as:

exclusive class distribution

inclusive class distribution

open end distribution

discrete frequency distribution

Very Short Answer Questions

Define the term ‘statistics’.

Define the term population.

What is sampling

What is a frequency distribution?

21 Distinguish between discrete and continuous variables.

Short Essay Questions

Explain the different steps in the construction of a frequency table for a given set of observations.

Explain the terms (i) class interval (ii) class mark (iii) class frequency.

Distinguish between census and sampling

What are the advantages of sampling over census?

23. State the various stages of statistical investigation.

Long Essay Questions

Present the following data of marks secured in Statistics (out of 100) of 60 students in the form of a frequency table with 10 classes of equal width, the lowest class being 0-9

41	17	83	60	54	91	60	58	70	07
67	82	33	45	57	48	34	73	54	62
36	52	32	72	60	33	07	77	28	30
42	93	43	80	03	34	56	66	23	63
63	11	35	85	62	24	00	42	62	33
72	53	92	87	10	55	60	35	40	57

Following is a cumulative frequency distribution showing the marks secured and the number of students in an examination:

Marks		No. of students (F)
Below	10	12
“	20	30
“	30	60
“	40	100
“	50	150
“	60	190
“	70	220
“	80	240
“	90	250

Obtain the frequency table (simple) from it. Also prepare ‘More than’ cumulative frequency table.

Module 2

MEASURES OF CENTRAL TENDENCY

A measure of central tendency helps to get a single representative value for a set of usually unequal values. This single value is the point of location around which the individual values of the set cluster. Hence the averages are known also as *measures of location*.

The important measures of central tendencies or statistical averages are the following.

Arithmetic Mean

Geometric Mean

Harmonic Mean

Median

Mode

Weighted averages, positional values, viz., quartiles, deciles and percentiles, also are considered in this chapter.

Criteria or Desirable Properties of an Average

It should be rigidly defined: That is, it should have a formula and procedure such that different persons who calculate it for a set of values get the same answer.

It should have sampling stability: A number of samples can be drawn from a population. The average of one sample is likely to be different from that of another. It is desired that the average of any sample is not much different from that of any other.

1. Arithmetic Mean

The arithmetic mean (AM) or simply mean is the most popular and widely used average. It is the value obtained by dividing sum of all given observations by the number of observations. AM is denoted by $\bar{x}$ (x bar).

Definition for a raw data

For a raw data or ungrouped data if $x_1, x_2, x_3, \dots, x_n$ are n observations,

then $$\bar{x} = \frac{x_1 + x_2 + x_3 + \dots + x_n}{n}$$

ie., $\bar{x} = \frac{\sum x}{n}$ where the symbol $\sum$ (sigma) denotes summation.

Example 1

Calculate the AM of 12, 18, 14, 15, 16

Solution

$$\bar{x} = \frac{\sum x}{n} = \frac{12 + 18 + 14 + 15 + 16}{5} = \frac{75}{5} = \mathbf{15}$$

Definition for a frequency data

For a frequency data if $x_1, x_2, x_3, \dots, x_n$ are 'n' observations or middle values of 'n' classes with the corresponding frequencies

$f_1, f_2, \dots, f_n$ then AM is given by

$$\bar{x} = \frac{f_1 \times x_1 + f_2 \times x_2 + \dots + f_n \times x_n}{f_1 + f_2 + \dots. + f_n} = \frac{\sum fx}{\sum f}$$

ie., $\bar{x} = \frac{\sum fx}{N}$ where N = $\sum f$ = Total frequency

Example 2

The following data indicate daily earnings (in rupees) of 40 workers in a factory.

Daily earnings in ₹	:	5	6	7	8	9
No of workers	:	3	8	12	10	7

Calculate the average income per worker.

X d = x 320

Solution

305 15

Daily Earnings in ₹320(x)	No. 0of workers (f)	fx
5 332	12 3	15
6 350	30 8	48
7	12	84
8	10	80
9	7	63
Total	40	290

$$\bar{x} = \frac{\sum fx}{N} = \frac{290}{40} = \mathbf{7.25}$$

Average income per worker is ₹ **7.25**

Example 3

Calculate the AM of the following data

Class	:	04	4-8	8-12	12-16
Frequency	:	1	4	3	2

Solution

Class	f	Mid values (x)	fx
0-4	1	2	2
4-8	4	6	24
8-12	3	10	30
12-16	2	14	28
Total	10		84

$$\bar{x} = \frac{\sum fx}{N} = \frac{84}{10} = \mathbf{8.4}$$

Shortcut Method: Raw data

Suppose the values of a variable under study are large, choose any value in between them. Preferably a value that lies more or less in the middle, called arbitrary origin or assumed mean, denoted by A. Take deviations of every value from the assumed mean A.

Let d = x   A, Taking summation of both sides and dividing by n, we get

$$- = A + \frac{\sum d}{n}$$

Example 4

Calculate the AM of 305, 320, 332, 350

Solution

X	d = x 320
305	15
320	0
332	12
350	30
	27

$$\bar{x} = A + \frac{\sum d}{n}$$

$$320 + \frac{27}{4}$$

320+6.75

326.75

Shortcut Method: Frequency Data

When the frequencies and the values of the variable x are large the calculation of AM is tedious. So a simpler method is adopted. The deviations of the mid values of the classes are taken from a convenient origin. Usually the mid value of the class with the maximum frequency is chosen as the arbitrary origin or assumed mean. Thus change x values to ‘d’ values by the rule,

$$d = \frac{x - A}{c}$$

where A-assumed mean, c-class interval, x-mid values. Then the formula for calculating AM is given by

$$- = \mathbf{A + \frac{\sum fd}{N} \times c}$$

Example 5

Calculate AM from the following data

Weekly wages	:	0-10	10-20	20-30	30-40	40-50
Frequency	:	3	12	20	10	5

Solution

Weekly wages	f	Mid value x	$d = \frac{x - 25}{10}$	fd	
0-10	3	5	2	6	18
10-20	12	15	1	12	
20-30	20	25	0	0	
30-40	10	35	1	10	20
40-50	5	45	2	10	
Total	50			2	

$$- \quad \hat{} = A + \frac{\sum fd}{N} \times c = 25 + \frac{2}{50} \times 10 = 25 + 0.4 = \mathbf{25.4}$$

Properties

The AM is preserved under a linear transformation of scale.

That is, if x_i is changed to y_i by the rule

$y_i = a + b x_i$, then $\bar{y} = a + b\bar{x}$, which is also linear**.**

The mean of a sum of variables is equal to the sum of the means of the variables.

Algebraic sum of the deviations of every observation from the A.M. zero.

If n_1 observations have an A.M $\bar{x}_1$ and n_2 observations have an AM $\bar{x}_2$ then the AM of the combined group of $n_1 + n_2$ observations is given by $\bar{x} = \dfrac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2}$.

Example 6

Let the average mark of 40 students of class A be 38; the average mark of 60 students of another class B is 42. What is the average mark of the combined group of 100 students?

Here $n_1 = 40$, $\bar{x}_1 = 38$, $n_2 = 60$, $\bar{x}_2 = 42$

Here $\bar{x} = \dfrac{n_1 \cdot \bar{x}_1 + n_2 \cdot \bar{x}_2}{n_1 + n_2} = \dfrac{(40 \times 38) + (60 \times 42)}{40 + 60}$

$= \dfrac{1520 + 2520}{100} = \dfrac{4040}{100} = \mathbf{40.4}$

Note

The above property can be extended as follows. When there are three groups, the combined mean is given by

The algebraic sum of the squares of the observations from AM is always minimum. ie., is always minimum.

Merits and Demerits

Merits

The most widely used arithmetic mean has the following merits.

It is rigidly defined. Clear cut mathematical formulae are available.

It is based on all the items. The magnitudes of all the items are considered for its computation.

It lends itself for algebraic manipulations. Total of a set, Combined Mean etc., could be calculated.

It is simple to understand and is not difficult to calculate. Because of its practical use, provisions are made in calculators to find it.

It has sampling stability. It does not vary very much when samples are repeatedly taken from one and the same population.

It is very much useful in day-to-day activities, later chapters in Statistics and many disciplines of knowledge.

Many forms of the formula are available. The form appropriate and easy for the data on hand can be used.

Demerits

It is unduly affected by extreme items. One greatest item may pull up the mean of the set to such an extent that its representative character is questioned. For example, the mean mark is 35 for the 3 students whose individual marks are 0, 5 and 100.

Theoretically, it cannot be calculated for open-end data.

It cannot be found graphically.

It is not defined to deal with qualities.

Weighted Arithmetic Mean

In calculating simple arithmetic mean it was assumed that all items are of equal importance. This may not be true always. When items vary in importance they must be assigned weights in proportion to their relative importance. Thus, a weighted mean is the mean of weighted items. The weighted arithmetic mean is sum of the product of the values and their respective weights divided by the sum of the weights.

Symbolically, if $x_1, x_2, x_3, ...x_n$ are the values of items and $w_1, w_2, ...w_n$ are their respective weights, then

$$\text{WAM} = \frac{w_1 x_1 + w_2 x_2 + w_3 x_3 + ... + w_n x_n}{w_1 + w_2 + w_3 + ... + w_n} = \frac{\sum wx}{\sum w}$$

Weighted AM is preferred in computing the average of percentages, ratios or rates relating to different classes of a group of observations. Also WAM is invariably applied in the computation of birth and death rates and index numbers.

Example 7

A student obtains 60 marks in Statistics, 48 marks in Economics, 55 marks in law, 72 marks in Commerce and 45 marks in taxation in an examination. The weights of marks respectively are 2, 1, 3, 4, 2. Calculate the simple AM and weighted AM of the marks.

Solution

$$\text{Simple AM} = \frac{\sum x}{n} = \frac{60 + 48 + 55 + 72 + 45}{5} = \frac{280}{5} = \mathbf{56}$$

Marks (x)	Weights (w)	wx
60	2	120
48	1	48
55	3	165
72	4	288
45	2	90
	12	711

$$\text{WAM} = \frac{\sum wx}{\sum w} = \frac{711}{12} = \mathbf{59.25}$$

Geometric Mean

Geometric mean (GM) is the appropriate root (corresponding to the number of observations) of the product of observations. If there are n observations GM is the n-th root of the product of n observations.

Definition for a raw data

If $x_1, x_2, x_3, ..., x_n$ are n observations;

$$\text{GM} = \sqrt[n]{x_1, x_2, ...x_n}$$

Using logarithms, we can calculate GM using the formula,

$$\text{GM} = Anti\log\left(\frac{\sum \log x}{n}\right)$$

Definition for a frequency distribution

For a frequency distribution if $x_1, x_2, x_3, ..., x_n$ are n observations with the corresponding frequencies $f_1, f_2, ..., f_n$

$$\text{GM} = \sqrt[N]{x_1^{f1}, x_2^{f2}, ...x_n^{fn}}$$

using logarithm,

$$\textbf{GM} = \textbf{Ant ilog}\left(\frac{\sum \textbf{f log x}}{\textbf{N}}\right) \text{ where } N = \sum f.$$

Note

GM is the appropriate average for calculating index number and average rates of change.

GM can be calculated only for non zero and non negative values.

3. Weighted GM = $Anti\log\left(\frac{\sum w \log x}{\sum w}\right)$

where w's are the weights assigned.

Example 8

Calculate GM of 2, 4, 8

Solution

$$\text{GM} = \sqrt[n]{x_1, x_2, ...x_n} = \sqrt[3]{2 \times 4 \times 8} = \sqrt[3]{64} = \mathbf{4}$$

Example 9

Calculate GM of 4, 6, 9, 1 1 and 15

Solution

x	logx
4	0.6021
6	0.7782
9	0.9542
11	1.0414
15	1.1761
	4.5520

$$\text{GM} = \text{Anti log}\left(\frac{\sum \log x}{n}\right)$$

$$= \text{Anti log}\left(\frac{4.5520}{5}\right)$$

= Antilog0.9104

= **8.136**

Example 10

Calculate GM of the following data

Classes	:	1-3	4-6	7-9	10-12
Frequency	:	8	16	15	3

Solution

Classes	f	X	logx	f.logx
1-3	8	2	0.3010	2.4080
4-6	16	5	0.6990	11.1840
7-9	15	8	0.9031	13.5465
10-12	3	11	1.0414	3.1242
Total	42			30.2627

$$\text{GM} = \text{Ant ilog}\left(\frac{\sum \mathbf{f \log x}}{.}\right)$$

= Antilog(30.2627/42)

= Antilog 0.7205 = 5.254

Merits and Demerits

Merits

It is rigidly defined. It has clear cut mathematical formula.

It is based on all the items. The magnitude of every item is considered for its computation.

It is not as unduly affected by extreme items as A.M. because it gives less weight to large items and more weight to small items.

It can be algebraically manipulated. The G.M. of the combined set can be calculated from the GMs and sizes of the sets.

It is useful in averaging ratios and percentages. It is suitable to find the average rate (not amount) of increase or decrease and to compute index numbers.

Demerits

It is neither simple to understand nor easy to calculate. Usage of logarithm makes the computation easy.

It has less sampling stability than the A.M.

It cannot be calculated for open-end data.

It cannot be found graphically.

It is not defined for qualities. Further, when one item is zero, it is zero and thereby loses its representative character. It cannot be calculated even if one value or one mid value is negative.

Harmonic Mean

The harmonic mean (HM) of a set of observations is defined as the reciprocal of the arithmetic mean of the reciprocals of the observations.

Definition for a raw data

If $x_1, x_2, x_3, ..., x_n$ are 'n' observations

$$HM = \frac{1}{\frac{\left(\frac{1}{x_1}+\frac{1}{x_2}+..+\frac{1}{x_n}\right)}{n}} = \frac{n}{\frac{1}{x_1}+\frac{1}{x_2}+..+\frac{1}{x_n}} = \frac{\mathbf{n}}{\sum\left(\frac{\mathbf{1}}{\mathbf{x}}\right)}$$

Definition for a frequency data

If $x_1, x_2, x_3, ..., x_n$ are 'n' observations with the corresponding

frequencies $f_1, f_2, f_3, ..., f_n$

then $$HM = \frac{N}{f_1\times\frac{1}{x_1}+f_2\times\frac{1}{x_2}+..+f_n\times\frac{1}{x_n}} = \frac{\mathbf{N}}{\sum\left(\frac{}{\mathbf{x}}\right)}$$

where $N = \sum f$

Note 1 HM can be calculated only for non zero and non negative values.

Note 2 HM is appropriate for finding average speed when distance travelled at different speeds are equal. Weighted HM is appropriate when the distances are unequal. HM is suitable to study rates also.

Note 3 Weighted HM = $\frac{N}{\sum\left(\frac{w}{x}\right)}$ where w's are the weighted assigned.

Example 11

Calculate the HM of 2, 3, 4, 5 and 7

Solution

$$HM = \frac{n}{\sum\frac{1}{x}} = \frac{5}{\frac{1}{2}+\frac{1}{3}+\frac{1}{4}+\frac{1}{5}+\frac{1}{7}}$$

$$= \frac{5}{\frac{210+140+105+84+60}{420}} = \frac{5\times420}{599} = \mathbf{3.50}$$

Example 12

Calculate HM of 5, 11, 12,16, 7, 9, 15, 13, 10 and 8

Solution

X	1/x	X	1/x
5	0.2000	9	0.1111
11	0.0909	15	0.0667
12	0.0833	13	0.0769
16	0.0625	10	0.1000
7	0.1429	8	0.1250
		Total	1.0593

$$HM = \frac{\mathbf{n}}{\sum\left(\frac{\mathbf{1}}{\mathbf{x}}\right)} = (10/1.0593) = 9.44$$

Merits and Demerits

Merits

It is rigidly defined. It has clear cut mathematical formula.

It is based on all the items. The magnitude of every item is considered for its computation.

It is affected less by extreme items than A.M. or even G.M.

It gives lesser weight to larger items and greater weight to lesser items.

It can be algebraically manipulated. The H.M. of the combined set can be calculated from the H.M.s and sizes of the sets. For example,

$$HM_{12} = \frac{N_1 + N_2}{\frac{N_1}{HM_1} + \frac{N_2}{HM_2}}$$

It is suitable to find the average speed.

Demerits

It is neither simple to understand nor easy to calculate.

It has less sampling stability than the A.M.

Theoretically, it cannot be calculated for open-end data.

It cannot be found graphically.

It is not defined for qualities. It is not calculated when atleast one item or one mid value is zero or negative.

It gives undue weightage to small items and least weightage to largest items. It is not used for analysing business or economic data.

Median

Median is defined as the middle most observation when the observations are arranged in ascending or descending order of magnitude. That means the number of observations preceding median will be equal to the number of observations succeeding it. Median is denoted by M.

Definition for a raw data

For a raw data if there are odd number of observations, there will be only one middle value and it will be the median. That means, if there are n observations arranged in order of their magnitude, the size of (n+1)/2 th observation will be the median. If there are even number of observations the average of two middle values will be *th*the median. That means, median will be the average of n/2th and $\left(\frac{n}{2} + 1\right)$ observations.

Definition for a frequency data

For a frequency distribution median is defined as the value of the variable

which divides the distribution into two equal parts. The median can be calculated using the following formula.

$$M = l + \frac{\left(\frac{N}{2} - m\right)}{f} \times c$$

where, l - lower limit of median class

Median class - the class in which N/2^{lh} observation falls

N - total frequency

m - cumulative frequency up to median class

c - class interval of the median class

f - frequency of median class

found to lie with in that interval.

Example 13

Find the median height from the following heights (in cms.) of 9 soldiers.

160, 180, 175, 179, 164, 178, 171, 164, 176

Solution

Step 1. Heights are arranged in ascending order:

160, 164, 164, 171, 175, 176, 178, 179, 180.

Step 2. Position of median = $\frac{n+1}{2}$ is calculated. It is $\frac{9+1}{2} = 5$.

Step 3. Median is identified (5th value) M = **175cms.**

It is to be noted that $\frac{+1}{2}$ may be a fraction, in which case, median is found as follows.

Example 14

Find the median weight from the following weights (in Kgs) of 10 soldiers. 75, 71, 73, 70, 74, 80, 85, 81, 86, 79

Solution

Step 1. Weights are arranged in ascending order:

70, 71, 73, 74, 75, 79, 80, 81, 85, 86

Step 2. Position of median $\frac{n+1}{2} = \frac{10+1}{2} = 5\frac{1}{2}$ is calculated

Step 3. Median is found. It is the mean of the values at 5th and 6th positions and so M = $\frac{75+79}{2}$ = **77Kgs.**

Example 15

Find the median for the following data.

Height in cms	:	160	164	170	173	178	180	182
No. of soldiers	:	1	2	10	22	19	14	2

Solution

Step 1. Heights are arranged in ascending order. Cumulative frequencies (c.f) are found. (They help to know the values at different positions)

Height in cms.	No. of Soldiers	C.f.
160	1	1
164	2	3
170	10	13
173	22	35
178	19	54
180	14	68
182	2	70
Total	70	

Step 2. Position of median, $\frac{N+1}{2} = \frac{70+1}{2} = 35\frac{1}{2}$ is calculated.

Step 3. Median is identified as the average of the values at the positions 35 and 36. The values are 173 and 178 respectively.

$$\therefore M = \frac{173+178}{2}$$ = **175.5cm**

Example 16

Calculate median for the following data

Class	:	0-5	5-10	10-15	15-20	20-25
f	:	5	10	15	12	8

Solution

Class	f	CF
0-5	5	5
5-10	10	15
10-15	15	30
15-20	12	42
20-25	8	50
Total	50	

$$M = l + \frac{\left(\frac{N}{2} - m\right)}{f} \times c$$ Median class is 10-15

Here $l = 10, N/2 = 50/2 = 25, c = 5, m = 15, f = 15$

$$\therefore M = 10 + \frac{(25-15)\,5}{15}$$

$$= 10 + \frac{10 \times 5}{15} = 10 + \frac{10}{3} =$$ 10+3.33 = **13.33**

Example 17

Calculate median for the data given below.

Classes	:	0-6	7-13	14-20	21-27	28-34	35-41
f	:	8	17	28	15	9	3

Solution:

Class	f	Actual class	CF
0-6	8	0.5-6.5	8
7-13	17	6.5-13.5	25
14-20	28	13.5-20.5	53
21-27	15	20.5-27.5	68
28-34	9	27.5-34.5	77
35-41	3	34.5-41.5	80
Total	80		

Median class is 13.5-20.5, l = 13.5, N/2 = 80/2 = 40
c = 7, m = 25, f = 28

$$M = l + \frac{\left(\frac{N}{2} - m\right)}{f} \times c = 13.5 + \frac{(40-25)}{28} \times 7$$

$$13.5 + \frac{15 \times 7}{284} = 13.5 + \underline{15}$$

13.5+3.75

17.25

Graphical Determination of Median

Median can be determined graphically using the following

Steps

Draw the less than or more than ogive

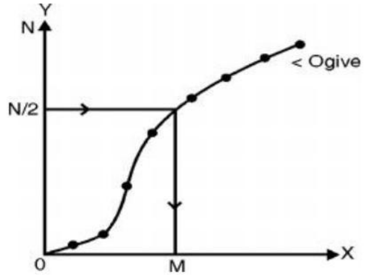
Locate N/2 on the Y axis.

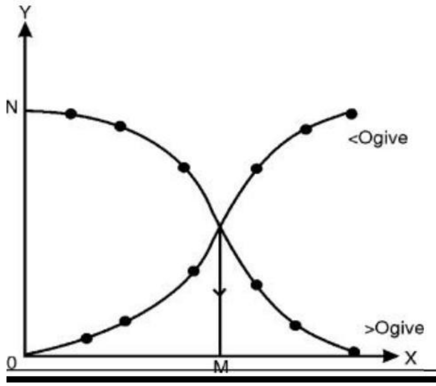
At N/2 draw a perpendicular to the Y axis and extend it to meet the ogive

From the point of intersection drop a perpendicular to the X axis

The point at which the perpendicular meets the X axis will be the median value.

Median can also be determined by drawing the two ogives, simultaneously. Here drop a perpendicular from the point of intersection to the X axis. This perpendicular will meet at the median value.





Merits and Demerits

Merits

It is not unduly affected by extreme items.

It is simple to understand and easy to calculate.

It can be calculated for open end data

It can be determined graphically.

It can be used to deal with qualitative data.

Demerits

It is not rigidly defined. When there are even number of individual observations, median is approximately taken as the mean of the two middle most observations.

It is not based on the magnitude of all the items. It is a positional measure. It is the value of the middle most item.

It cannot be algebraically manipulated. For example, the median of the combined set can not be found from the medians and the sizes of the individual sets alone.

It is difficult to calculate when there are large number of items which are to be arranged in order of magnitude.

It does not have sampling stability. It varies more markedly than A M from sample to sample although all the samples are from one and the same population.

Its use is lesser than that of AM.

Mode

Mode is that value of the variable, which occur maximum number of times in a set of observations. Thus, mode is the value of the variable, which occur most frequently. Usually statements like, ‘average student’, ‘average buyer’, ‘the typical firm’, etc. are referring to mode of the phenomena. Mode is denoted by Z or Mo. For a raw data as well as for a discrete frequency distribution we can locate mode by inspection.

For a frequency distribution mode is defined as the value of the variable having the maximum frequency. For a continuous frequency distribution it can be calculated using the formula given below:

$$Z = l + \frac{\Delta_1}{\Delta_1 + \Delta_2} \times c$$

where l : lower limit of modal class

Modal class : Class having the maximum frequency

Δ_1 : difference between the frequency of modal class and that of the premodal class

Δ_2 : difference between frequency of the modal class and that of the post modal class

c : class interval

For applying this formula, the class intervals should be (i) of equal size (ii) in ascending order and (iii) in exclusive form.

Example 18

Determine the mode of

420, 395, 342, 444, 551, 395, 425, 417, 395, 401, 390

Solution

Mode = **395**

Example 19

Determine the mode

Size of shoes	:	3	4	5	6	7	8	9
No of pairs sold	:	10	25	32	38	61	47	34

Solution

Mode = Z = 7

Example 20

Calculate mode for the following data

Classes	:	0-9	10 - 19	20-29	30-39	40-49	50-59
f	:	5	10	17	33	22	13

Solution

Classes	f	Atual class
0-9	5	0.5-9.5
10-19	10	9.5-19.5
20-29	17	19.5-29.5
30-39	33	29.5-39.5
40-49	22	39.5-49.5
50-59	13	49.5-59.5

$$Z = l + \frac{\Delta_1}{\Delta_1 + \Delta_2} \times c$$

Modal class is 29.5-39.5

$l = 29.5$

$\Delta_1 = 33 - 17 = 16$

$\Delta_2 = 33 - 22 = 11, c = 10$

$$29.5 + \frac{16}{16 + 11} \times 10$$

29.5+5.92 = **35.42**

For a symmetrical or moderately assymmetrical distribution, the empirical relation is

Mean Mode = 3 (Mean Median)

This relation can be used for calculating any one measure, if the remaining two are known.

Example 21

In a moderately assymmetrical distribution Mean is 24.6 and Median 25.1. Find the value of mode.

Solution

We have

$$\begin{aligned} \text{Mean} \quad \text{Mode} &= 3(\text{Mean} \quad \text{Median}) \\ 24.6 \quad Z &= 3(24.6 \quad 25.1) \\ 24.6 \quad Z &= 3(\quad 0.5) = \quad 1.5 \\ Z &= 24.6 + 1.5 = \mathbf{26.1} \end{aligned}$$

Example 22

In a moderately assymmetrical distribution Mode is 48.4 and Median 41.6. Find the value of Mean

Solution

We have,

$$\begin{aligned} \text{MeanMode} &= 3(\text{Mean} \quad \text{Median}) \\ \bar{x} \quad 48.4 &= 3(\bar{x} \quad 41.6) \\ \bar{x} \quad 48.4 &= 3\bar{x} \quad 124.8 \\ 3\bar{x} \quad \bar{x} &= 124.8 \quad 48.4 \\ 2\bar{x} &= 76.4 \\ \bar{x} &= 76.4 \div 2 = \mathbf{38.2} \end{aligned}$$

Merits and Demerits

Merits

Mode is not unduly affected by extreme items.

It is simple to understand and easy to calculate

It is the most typical or representative value in the sense that it has the greatest frequency density.

It can be calculated for open-end data.

It can be determined graphically. It is the x-coordinate of the peak of the frequency curve.

It can be found for qualities also. The quality which is observed more often than any other quality is the modal quality.

Demerits

It is not rigidly defined.

It is not based on all the items. It is a positional value.

It cannot be algebraically manipulated. The mode of the combined set cannot be determined as in the case of AM.

Many a time, it is difficult to calculate. Sometimes grouping table and frequency analysis table are to be formed.

It is less stable than the A.M.

Unlike other measures of central tendency, it may not exist for some data. Sometimes there may be two or more modes and so it is said to be ill defined.

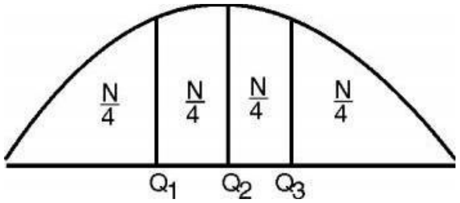
It has very limited use. Modal wage, modal size of shoe, modal size of family, etc., are determined. Consumer preferences are also dealt with.

Partition Values

We have already noted that the total area under a frequency curve is equal to the total frequency. We can divide the distribution or area under a curve into a number of equal parts choosing some points like median. They are generally called *partition values or quantiles*. The important partition values are *quartiles, deciles and percentiles.*

Quartiles

Quartiles are partition values which divide the distribution or area under a frequency curve into 4 equal parts at 3 points namely Q_1, Q_2, and Q_3 . Q_1 is called *first quartile or lower quartile*, Q_2 is called *second quartile, middle quartile or median* and Q_3 is called *third quartile or upper quartile*. In other words Q_1 is the value of the variable such that the number of observations lying below it, is N/4 and above it is 3N/4. Q_2 is the value of the variable such that the number of observations on either side of it is equal to N/2. And Q_3 is the value of the variable such that the number of observations lying below Q_3 is 3N/4 and above Q_3 is N/4.



Deciles and Percentiles

Deciles are partition values which divide the distribution or area under frequency curve into 10 equal parts at 9 points namely D_1, D_2, ..., D_9.

Percentiles are partition values which divide the distribution into 100 equal parts at 99 points namely P_1, P_2, P_3, ... P_{99}. Percentile is a very useful measure in education and psychology. Percentile ranks or scores can also be calculated. Kelly’s measure of skewness is based on percentiles.

Calculation of Quartiles

The method of locating quartiles is similar to that method used for finding median. Q_1 is the value of the item at $(n + 1)/4^{tn}$ position and Q_3 is the value of the item at $3(n + 1) / 4^{th}$ position when actual values are known. In the case of a frequency distribution Q_1 and Q_3 can be calculated as follows.

$$Q_1 = l_1 + \frac{\left(\frac{N}{4} - m\right)}{f} \times c$$

where l_1 - lower limit of Q_1 class

Q_1 class - the class in which N/4th item falls

m - cumulative frequency up to Q_1 class

c - class interval

f - frequency of Q_1 class

$$Q_3 = l_3 + \frac{\left(\frac{3N}{4} - m\right)}{f} \times c$$

where l_3 - lower limit of Q_3 class

Q_3 class - the class in which 3N/4th item falls

m - cumulative frequency up to Q_3 class

c - class interval

f - frequency of Q_3 class

We can combine these three formulae and can be written as

$$Q_i = l_i + \frac{\left(\frac{iN}{4} - m\right)}{f} \times c\,, \quad i = 1, 2, 3$$

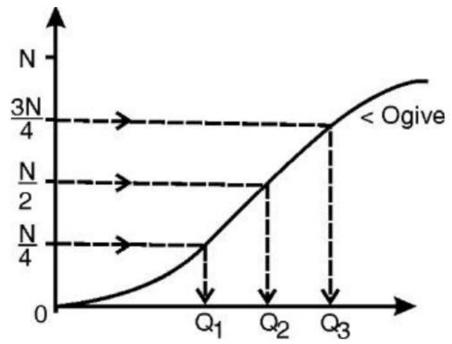
In a similar fashion deciles and percentiles can be calculated as

$$D_i = l_i + \frac{\left(\frac{iN}{10} - m\right)}{f} \times c\,, \quad i = 1, 2, 3, ... 9$$

$$P_i = l_i + \frac{\left(\frac{iN}{100} - m\right)}{f} \times c\,, \; i = 1, 2, 3, ..., 99$$

Graphical Determination of Quartiles

Quartiles can be determined graphically by drawing the ogives of the given frequency distribution. So draw the less than ogive of the given data. On the Y axis locate N/4, N/2 and 3N/4. At these points draw perpendiculars to the Y axis and extend it to meet the ogive. From the points of intersection drop perpen-diculars to the X axis. The point corresponding to the CF, N/4 is Q_1 corresponding to the CF N/2 is Q_2 and corres-ponding to the CF 3N/4 is Q_3.



Example 23

Find , Q_1, Q_3, D_2, D_9, P_{16}, P_{65} for the following data. 282, 754, 125, 765, 875, 645, 985, 235, 175, 895, 905, 112 and 155.

Solution

Step 1. Arrange the values in ascending order

112, 125, 155, 175, 235, 282, 645, 754, 765, 875,

895, 905 and 985.

Step 2. Position of Q_1 is $\frac{n+1}{4} = \frac{13+1}{4} = \frac{14}{4} = 3.5$

Similarly positions of Q_3, D_2, D_9, P_{16} and P_{65} are 10.5, 2.8, 12.6, 2.24 and 9. 1 respectively.

Step 3.

$Q_1 = 155 + 0.5(175 - 155) =$ **165**

$Q_3 = 875 + 0.5(895 - 875) =$ **885**

$_2 = 125 + 0.8(155 - 125) =$ **149.0**

$D_9 = 905 + 0.6(985 - 905) =$ **953**

$P_{16} = 125 + 0.24(155 - 125) =$ **132.20**

$P_{65} = 765 + 0.1(875 - 765) =$ **776.0**

Note

The value of the 12.6-th position (D_9) is obtained as value of 12-th position + 0.6 (value at 13-th position - value at 12-th position)

Example 24

Find Q_1, Q_3, D_4, P_{20} and P_{99} for the data given below.

Mark	:	25	35	40	50	52	53	67	75	80
No of students	:	3	29	32	41	49	54	38	29	27

Solution

Marks	No of students	Cumulative frequency
25	3	3
35	29	32
40	32	64
50	41	105
52	49	154
53	54	208
67	38	246
75	29	275
80	27	302

Step 1. The cumulative frequencies of marks given in ascending order are found

Step 2. The positions of Q_1, Q_3, D_4, P_{20} and P_{99} are found. They are

$$\frac{N+1}{4} = \frac{303}{4} = 75.75$$

$$\frac{3(N+1)}{4} = 3 \times \frac{303}{4} = 227.25$$

$$\frac{4(N+1)}{10} = \frac{40 \times 303}{10} = 121.20$$

$$\frac{20(N+1)}{100} = \frac{20 \times 303}{100} = 60.60$$

$$\frac{99(N+1)}{100} = \frac{99 \times 303}{100} = 299.97$$

Step 3. The marks of students at those positions are found

$$Q_1 = 50 + 0.75(50 - 50) = \textbf{50}\text{ Marks}$$

$$Q_3 = 67 + 0.25(67 - 67) = \textbf{67}\text{ Marks}$$

$$D_4 = 52 + 0.20(52 - 52) = \textbf{52}\text{ Marks}$$

$$P_{20} = 40 + 0.60(40 - 40) = \textbf{40}\text{ Marks}$$

$$P_{99} = 80 + 0.97(80 - 80) = \textbf{80}\text{ Marks}$$

Note

Refer the above example to know the method of finding the values of the items whose positions are fractions.

Example 25

Calculate quartiles for the following data

Classes	:	30-35	35-40	40-45	45-50	50-55	55-60	60-65
Freq.	:	10	16	18	27	18	8	3

Solution

Class	f	CF
30-35	10	10
35-40	16	26
40-45	18	44
45-50	27	71
50-55	18	89
55-60	8	97
60-65	3	100
Total	100	

$$Q_1 = l_1 + \frac{\left(\frac{N}{4} - m\right)c}{f}$$

$$+ \frac{(25 - 10)5}{16}$$

$$35 + \frac{15 \times 5}{16} = 35 + \frac{75}{16}$$

= 35+4.68 = **39.68**

$$Q_2 = l_2 + \frac{\left(\frac{N}{2} - m\right)c}{f}$$

$$45 + \frac{(50 - 44)5}{27}$$

$$45 + \frac{6 \times 5}{27}$$

$$45 + \frac{10}{9}$$ = 45+1.11 = **46.11**

$$Q_3 = l_3 + \frac{\left(\frac{3N}{4} - m\right)c}{f}$$

$$= 50 + \frac{(75 - 71)5}{18}$$

$$= 50 + \frac{4 \times 5}{18} = 50 + \frac{10}{9}$$ = 50+1.11 = **51.11**

EXERCISES

Multiple Choice Questions

1. Mean is a measure of

a. location or central value b. dispersion

c. correlation d. none of the above

If a constant value 50 is subtracted from each observation of a set, the mean of the set is:

a. increased by 50 b. decreased by 50

c. is not affected d. zero

If the grouped data has open end classes, one cannot calculate:

a. median b. mode c.mean d. quartiles

Harmonic mean is better than other means if the data are for:

a. speed or rates b. heights or lengths

c. binary values like 0 & 1 d. ratio or proportions

Extreme value have no effect on:

a. average b. median

c. geometric mean d. harmonic mean

If the A.M. of a set of two observations is 9 and its G.M. is 6. Then the H.M. of the set of observations is:

a. 4 b. $3\sqrt{6}$ c. 3 d. 1.5

The A.M. of two numbers is 6.5 and their G.M. is 6. The two numbers are:

a. 9, 6 b. 9, 5 c. 7, 6 d. 4, 9

8.If the two observations are 10 and 10 then their harmonic mean is:

a. 10 b. 0 c. 5 d. ∞

The median of the variate values 11, 7, 6, 9, 12, 15,, 19 is:

a. 9 b. 12 c. 15 d. 11

10. The second dicile divides the series in the ratio:

a. 1:1 b. 1:2 c. 1:4 d. 2:5

11. For further algebraic treatment, geometric mean is:

a. suitable b. not suitable

c. sometimes suitable d. none of the above

12.The percentage of values of a set which is beyond the third quartile is:

a. 100 percent b. 75 percent

c. 50 percent d. 25 percent

In a distribution, the value around which the items tend to be most heavily concentrated is called:

a. mean b. median

c. third quartile d. mode

14. Sum of the deviations about mean is

a. zero b. minimum c. maximum d. one

15. The suitable measure of central tendency for qualitative data is:

a. mode b. arithmetic mean

c. geometric mean d. median

16. The mean of the squares of first eleven natural numbers is:

a. 46 b. 23 c. 48 d. 42

The percentage of items in a frequency distribution lying between upper and lower quartiles is:

a. 80 percent b. 40 percent

c. 50 percent d. 25 percent

Very Short Answer Questions

What is central tendency?

Define Median and mode.

Define harmonic mean

Define partition values

State the properties of AM.

In a class of boys and girls the mean marks of 10 boys is 38 and the mean marks of 20 girls 45. What is the average mark of the class?

23. Define deciles and percentiles.

24 Find the combined mean from the following data.

	Series x	Series y
Arithmetic mean	12	20
No of items	80	60

Short Essay Questions

25 Define mode. How is it calculated. Point out two

Define AM, median and mode and explain their uses

Give the formulae used to calculate the mean, median and mode of a frequency distribution and explain the symbols used in them.

How will you determine three quartiles graphically from a less than ogive?

Three samples of sizes 80, 40 and 30 having means 12.5, 13 and 11 respectively are combined. Find the mean of the combined sample.

Explain the advantages and disadvantages of arithmetic mean as an average.

For finding out the 'typical' value of a series, what measure of central tendency is appropriate?

32 Explain AM and HM. Which one is better? And Why?

Prove that the weighted arithmetic mean of first n natural numbers whose weights are equal to the corresponding number is equal to $(2n+1)/3$

Show that GM of a set of positive observation lies between AM & AM.

What are the essential requisites of a good measure of central tendency? Compare and contrast the commonly employed measures in terms of these requisites.

Discuss the merits and demerits of the various measures of central tendency. Which particular measure is considered the best and why? Illustrate your answer.

. What is the difference between simple and weighted average? Explain the circumstances under which the latter should be used in preference to the former.

Find the average rate of increase in population which in the first decade has increased 12 percent, in the next by 16 per cent, and in third by 21 percent.

39.. A person travels the first mile at 10 km. per hour, the second mile at 8 km. per hour and the third mile at 6 km. per hour. What is his average speed?

Long Essay Questions

Compute the AM, median and mode from the following data

Age last birth day	:	15-19	20-24	25-29	30-34	35-39	40-44
No of persons	:	4	20	38	24	10	

41. Calculate Arithmetic mean, median and mode for the following data.

Ag e	:	55-60	50-55	45-50	40-45	35-40	30-35	25-30	20-25
No of people	:	7	13	15	20	30	33	28	14

Calculate mean, median and mode from the following data

Class	Frequency
Up to 20	52
20-30	161

30-40	254
40-50	167
50-60	78
60-80	64
Over 80	52

Calculate mean, median and mode

Central wage in Rs. :	15	20	25	30	35	40	45
No. of wage earners:	3	25	19	16	4	5	6

(i)Find the missing frequencies in the following distribution given that N = 100 and median of the distribution is 110.

Calculate the arithmetic mean of the completed frequency distribution.

Class	**:**	**20 - 4 0**	**40 -6 0**	**60 -8 0**	**80- 100**	**100-120**
Frequency	**:**	**6**	**9**	**-**	**1 4**	**2 0**
Class	**:**	**120-140**	**140-160**	**160-180**	**180-200**	
Frequency	**:**	**1 5**	**-**	**8**	**7**	

MODULE 1

PART III

MEASURES OF DISPERSION

By dispersion we mean *spreading* or *scatteredness* or *variation*. It is clear from the above example that dispersion measures the extent to which the items vary from some central value. Since measures of dispersion give an average of the differences of various items from an average, they are also called averages of second order.

Desirable properties of an ideal measure of dispersion

The following are the requisites for an ideal measure of dispersion.

1. It should be rigidly defined and its value should be definite.
2. It should be easy to understand and simple to calculate .
3. It should be based on all observations .
4. It should be capable of further algebraic treatment.
5. It should be least affected by sampling fluctuations .

Methods of Studying Variation

The following measures of variability or dispersion are commonly used.

1. Range
2. Quartile Deviation
3. Mean Deviation
4. Standard Deviation

Here the first two are called positional measures of dispersion. The other two are called calculation measures of deviation.

Absolute and Relative Dispersion

Absolute measures and relative measures are the two kinds of measures of dispersion. The formers are used to assess the variation among a set of values. The latter are used whenever the variability of two or more sets of values are to be compared. Relative measures give pure numbers, which are free from the units of measurements of the data. Even data in different units and with unequal average values can be compared on the basis of relative measures of dispersion. Less is the value of a relative measure, less is the variation of the set and more is the consistency. The terms, stability, homogeneity, uniformity and consistency are used as if they are synonyms.

1. Range

Definition Range is the difference between the greatest (largest) and the smallest of the given values.

In symbols, **Range = L S** where L is the greatest value and S is the smallest value.

The corresponding relative measure of dispersion is defined as

Coefficient of Range = $\frac{-S}{L+S}$

Example 1

The price of a share for a six-day week is fluctuated as follows:

₹156 ₹ 165 ₹ 148 ₹ 151₹ 147 ₹162

Calculate the Range and its coefficient.

Solution

Range = L S = ₹ 165 ₹ 147 = ₹ 18

Coefficient of Range = $\frac{L-S}{L+S} = \frac{165-147}{165+147}$ = **0.0577**

Example 2

Calculate coefficient of range from the following data:

Mark:	10-20	20-30	30-40	40-50	50-60
No. of students:	8	10	12	8	4

Solution

Coefficient of Range = $\frac{L-S}{L+S} = \frac{60-10}{60+10}$ = **0.7143**

Merits and Demerits

Merits

1. It is the simplest to understand and the easiest to calculate.
2. It is used in Statistical Quality Control.

Demerits

1. Its definition does not seem to suit the calculation for data with class intervals. Further, it can not be calculated for open-end data.
2. It is based on the two extreme items and not on any other item.
3. It does not have sampling stability. Further, it is calculated for samples of small sizes only.
4. It could not be mathematically manipulated further.
5. It is a very rarely used measure. Its scope is limited to very few considerations in Quality Control.

2. Quartile Deviation

Definition

Quartile deviation is half of the difference between the first and the third quartiles.

In symbols, Q.D = $\frac{Q_3 - Q_1}{2}$, Q.D is the abbreviation.

Among the quartiles Q_1, Q_2 and Q_3, the range is Q_3 Q_1
.

ie., inter-quartile range is Q_3 Q_1 and Q.D which is

$\frac{Q_3 - Q_1}{2}$ is the semi inter-quartile range.

Coefficient of Quartile Deviation = $\frac{Q_3 - Q_1}{Q_3 + Q_1}$

This is also called quartile coefficient of dispersion.

Example 3

Find the Quartile Deviation for the following:

391, 384, 591, 407, 672, 522, 777, 733, 1490, 2488

Solution

Before finding Q.D., Q_1 and Q_3 are found from the values in ascending order:

384, 391, 407, 522, 591, 672, 733, 777, 1490, 2488

Position of Q_1 is $\frac{n+1}{4} = \frac{10+1}{4}$ 2.75

Q_1 = 391 + 0.75 (407 391) 403

Position of Q_3 is $\frac{3(n+1)}{4} = 3 \times 2.75$ =8.25

Q_3 = 777+0.25 (1490 777) 955.25

QD = $\frac{Q_3 - Q_1}{2} = \frac{955.25 - 403.00}{2}$ **276.125**

Example 5

Calculate Quartile deviation for the following data.
Also calculate quartile coefficient of dispersion.

Class:	20-30	30-40	40-50	50-60	60-70	70-80	80-90	90-100
f :	6	18	25	50	37	30	24	10

Solution

Classes	f	CF
20-30	6	6
30-40	18	24
40-50	25	49
50-60	50	99
60-70	37	136
70-80	30	166
80-90	24	190
90-100	10	200

$Q_1 = l_1 + \frac{\left(\frac{N}{4} - m\right)}{f} c$ $44 = \underline{200} = 50$

$= 50 + \frac{(50 - 49)}{50} 10$ $l_1 = 50, c = 10$

$= 50 + \frac{1 \times 10}{50} = 50 + \frac{1}{5}$ m = 49, f = 50

= 50 + 0.2 = **50.2**

$Q_3 - l_3 + \frac{\left(\frac{3N}{4} - m\right)}{f} c =$ $+ \underline{(150} - \underline{136}$ $)10\ 30$

$70 + \frac{\underline{14} \times \underline{10}}{4.67\ 30} = 70 +$ **74.67**

$$QD = \frac{Q_3 - Q_1}{2} = \frac{74.67 - 50.20}{2} = \frac{24.47}{2} = \mathbf{12.23}$$

Quartile coefficient of dispersion

$$= \frac{Q_3 - Q_1}{Q_3 + Q_1} = \frac{74.67 - 50.20}{74.67 - 50.20} = \frac{24.47}{124.87} = \mathbf{0.196}$$

Merits and Demerits

Merits

1. It is rigidly defined.
2. It is easy to understand and simple to calculate.
3. It is not unduly affected by extreme values.
4. It can be calculated for open-end distributions.

Demerits

1. It is not based on all observations
2. It is not capable of further algebraic treatment
3. It is much affected by fluctuations of sampling.

Mean Deviation

The Mean Deviation is defined as the Arithmetic mean of the absolute value of the deviations of observations from some origin, say mean or median or mode.

Thus for a raw data

$$\textbf{M.D about Mean} = \frac{\sum |x - \bar{x}|}{n}$$

where $|x - \bar{x}|$ stands for the absolute deviation of x from $\bar{x}$ and is read as modulus of $(x - \bar{x})$ or mod $(x - \bar{x})$.

Instead of taking deviation from mean, if we are using median we get the mean deviation about median.

$$\therefore \text{M.D. about Median} = \frac{\sum |x - M|}{n}$$

For a frequency data, MD about Mean is given by

$$(MD)_{\bar{x}} = \frac{\sum f\,|x - \bar{x}|}{N} \quad ; N = \sum f$$

$$\text{MD about Median (MD)} = \frac{\sum f|x - M|}{N}$$

Note

Whenever nothing is mentioned about the measure of Central tendency from which deviations are to be considered, deviations are to be taken from the mean and the required MD is MD about mean.

(i) Coefficient of MD(about mean) = $\dfrac{\textit{MD about mean}}{\textit{Mean}}$

(ii) Coefficient of MD(about median) = $\dfrac{\textit{MD about median}}{\textit{Median}}$

Example 6

Calculate MD about Mean of 8, 24, 12, 16, 10, 20

Solution

x	$x-\bar{x}$	$\lvert x-\bar{x}\rvert$
8	7	7
24	9	9
12	3	3
16	1	1
10	5	5
20	5	5
90		30

Example8

Calculate MD about Mean and the coefficient of MD Classes: 0-10 10-20 20-30 30-40 40-50

f	:	5	15	17	11	2

Solution

Class	f	x	fx	$x-\bar{x}$	$\lvert x-\bar{x}\rvert$	$f\lvert x-\bar{x}\rvert$
0-10	5	5	25	18	18	90
10-20	15	15	225	8	8	120
20-30	17	25	425	2	2	34
30-40	11	35	385	12	12	132
40-50	2	45	90	22	22	44
Total	50		1150			420

$$\bar{x} = \frac{\sum fx}{N} = \frac{1150}{50} = 23$$

$$(ND)_{\bar{x}} = \frac{\sum f\lvert x-\bar{x}\rvert}{N} = \frac{420}{50} = \mathbf{8.4}$$

$$\text{Coefficient of MD} = \frac{\text{MD about mean}}{\text{Mean}} = \frac{8.4}{23} = \mathbf{0.3652}$$

Merits and Demerits

Merits

1. It is rigidly defined
2. It is easy to calculate and simple to understand
3. It is based on all observations.
4. It is not much affected by the extreme values of items.
5. It is stable.

Demerits

1. It is mathematically illogical to ignore the algebraic signs of deviations.
2. No further algebraic manipulation is possible.
3. It gives more weight to large deviations than smaller ones.

Standard Deviation

The standard deviation is the most useful and the most popular measure of dispersion. The deviation of the observations from the AM are considered and then each squared. The sum of squares is divided by the number of observations. The square root of this value is known as the standard deviation. *Thus Standard deviation (SD) is defined as the square root of the AM of the squares of the deviations of observations from AM*. It is denoted by 's' (sigma). We can calculate SD using the following formula. So for a raw data, if $x_1, x_2, x_3 .... x_n$ are n observations

$$\mathbf{SD} = s = \sqrt{\frac{\sum (x_i-\bar{x})^2}{n}}$$

For a frequency data, if $x_1, x_2, x_3 .... x_n$ are n observations or middle values of n classes with the corresponding frequencies f_1, f_2, ... f_n

then,
$$SD = s = \sqrt{\frac{\sum f_i(x_i-\bar{x})^2}{N}}$$

The square of the SD is known as 'Variance' and is denoted as s^2 or SD is the positive square root of variance.

Simplified formula for SD

For a raw data, we have

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

$$= \frac{1}{n}\sum x^2 - 2\bar{x}\frac{1}{n}\sum x + \bar{x}^2\frac{1}{n}\sum 1$$

$$= \frac{\sum x^2}{n} - 2\bar{x}\cdot\bar{x} + \bar{x}^2$$

$$= \frac{\sum x^2}{n} - \bar{x}^2$$

$$\therefore \quad s = \sqrt{\frac{\sum x^2}{n} - \bar{x}^2}$$

In a similar way, for a frequency data

$$s = \sqrt{\frac{\sum fx^2}{N} - \left(\frac{\sum fx}{N}\right)^2}$$

Short Cut Method

For a raw data, $s = \sqrt{\frac{\sum d^2}{n} - \left(\frac{\sum d}{n}\right)^2}$ where $d = x - A$

For a frequency data, $s = c \times \sqrt{\frac{\sum fd^2}{N} - \left(\frac{\sum fd}{N}\right)^2}$

where $d = \frac{x-A}{c}$, A - assumed mean, c – class interval.

The relative measure of dispersion based on SD or coefficient of SD is given by

$$\text{Coefficient of SD} = \frac{SD}{AM} = \frac{\sigma}{x}$$

Importance of Standard Deviation

Standard deviation is always associated with the mean. It gives satisfactory information about the effectiveness of mean as a representative of the data. More is the value of the standard deviation less is the concentration of the observations about the mean and vice versa. Whenever the standard deviation is small mean is accepted as a good average.

According to the definition of standard deviation, it can never be negative. When all the observations are equal standad deviation is zero. Therefore a small value of s suggests that the observations are very close to each other and a big value of s suggests that the observations are widely different from each other.

Properties of Standard Deviation

1. Standard deviation is not affected by change of origin.

Proof

Let $x_1, x_2, \ldots., x_n$ be a set of n observations.

Then $s_x = \sqrt{\frac{1}{n}\sum(x_i - \bar{x})^2}$

Choose $y_i = x_i + c$ for i = 1, 2, 3... n

Then $\bar{y} = \bar{x} + c$

$$\therefore\ y_i - \bar{y} = x_i - \bar{x}$$

$$\sum (y_i - \bar{y})^2 = \sum (x_i - \bar{x})^2$$

$$\frac{1}{n}\sum (y_i - \bar{y})^2 = \frac{1}{n}\sum (x_i - \bar{x})^2$$

$$\text{ie.,} \sqrt{\frac{1}{n}\sum (y_i - \bar{y})^2} = \sqrt{\frac{1}{n}\sum (x_i - \bar{x})^2}$$

ie., $S_y = S_x$

Hence the proof

2. Standard deviation is affected by change of scale.

Proof

Let $x_1, x_2, \ldots, x_n$ be a set of n observations.

Then $s_x = \sqrt{\frac{1}{n}\sum (x_i - \bar{x})^2}$

Choose $y_i = c\,x_i + d$, i = 1, 2, 3... n and c and d are constants. This fulfils the idea of changing the scale of the original values.

Now $\bar{y} \quad c\bar{x} + d$

$$\therefore\ y_i - \bar{y} \quad c(x_i - \bar{x})$$

$$(y_i - \bar{y})^2 \quad c^2 \sum (x_i - \bar{x})^2$$

$$\frac{1}{n}\sum (y_i - \bar{y})^2 \quad c^2 \frac{1}{n}\sum (x_i - \bar{x})^2$$

$$\text{ie.,} \sqrt{\frac{1}{n}\sum (y_i - \bar{y})^2} \quad c\sqrt{\frac{1}{n}\sum (x_i - \bar{x})^2}$$

´ s_x

ie., s_y SD of y values c ´ SD of x values

Hence the proof.

Note

If there are k groups then the S.D. of the k groups combined is given by the formula.

$$(n_1 + n_2 + \ldots + n_k)\sigma^2 = n_1\sigma_1^2 + n_2\sigma_2^2 + \ldots + n_k\sigma_k^2 + n_1 d_1^2 + n_2 d_2^2 + \ldots + n_k d_k^2$$

Coefficient of Variation

Coefficient of variation (CV) is the most important relative measure of dispersion and is defined by the formula.

$$\text{Coefficient of Variation} = \frac{\text{Standard deviation}}{\text{Arithmetic mean}} \times 100$$

$$\text{CV} = \frac{SD}{AM} \times 100 = \frac{\sigma}{\bar{x}} \times 100$$

CV is thus the ratio of the SD to the mean, expressed as a percentage. According to Karl Pearson, Coefficient of variation is the percentage variation in the mean.

Coefficient of Variation is the widely used and most popular relative measure. The group which has less C.V is said to be more consistent or more uniform or more stable. More coefficient of variation indicates greater variability or less consistency or less uniformity or less stability.

Example 9

Calculate SD of 23, 25, 28, 31, 38, 40, 46

Solution

x	$x-\bar{x}$	$(x-\bar{x})^2$
23	10	100
25	8	64
28	5	25
31	2	4
38	5	25
40	7	49
46	13	169
231		436

− = 231/7 = 33

SD = $\sqrt{\frac{\sum(x_i-\bar{x})^2}{n}}$ = **7.89**

Example 10

Calculate SD of 1, 2, 3, 4, 5, 6, 7, 8, 9, 10

Solution

x	1	2	3	4	5	6	7	8	9	10
x^2	1	4	9	16	25	36	49	64	81	100

$x = 55$

$x^2 = 385$

$$\bar{x} = \frac{\sum x}{n} = \frac{55}{10} = 5.5$$

SD $= \sigma = \sqrt{\frac{\sum x^2}{n} - \bar{x}^2} = \sqrt{\frac{385}{10} - 5.5^2}$

$= \sqrt{38.5 - 30.25} = \sqrt{8.25}$ = **2.87**

Example 11

Calculate SD of 42, 48, 50, 62, 65

Solution

x	$d = x - 50$	d^2
42	8	64
48	2	4
50	0	0
62	12	144
65	15	225
Total	17	437

SD = $\sqrt{\frac{\sum d^2}{n} - \left(\frac{\sum d}{n}\right)^2}$ = **8.70**

Example 12

Calculate SD of the following data

Size (x) :	10	12	14	16	18
Frequency:	2	4	10	3	1

Solution

x	f	fx	fx^2
10	2	20	200
12	4	48	576
14	10	140	1960
16	3	48	768
18	1	18	324
Total	20	274	3828

$$\sigma = \sqrt{\frac{\sum fx^2}{N} - \left(\frac{\sum fx}{N}\right)^2} = \sqrt{\frac{3828}{20} - \left(\frac{247}{20}\right)^2}$$

$$\sqrt{191.4 - (13.7)^2} = \sqrt{191.40 - 187.69} = \sqrt{3.71} = \mathbf{1.92}$$

Example 13

Calculate SD of the following data

Classes:	0-4	4-8	8-12	12-16	16-20
f :	3	8	17	10	2

Solution

Classes	f	x	fx	fx^2
0-4	3	2	6	12
4-8	8	6	48	288
8-12	17	10	170	1700
12-16	10	14	140	1960
16-20	2	18	36	648
Total	40		400	4608

$$\sigma = \sqrt{\frac{\sum fx^2}{N} - \left(\frac{\sum fx}{N}\right)^2} = \sqrt{\frac{4608}{40} - \left(\frac{400}{40}\right)^2}$$

$$= \sqrt{115.2 - 100} = \sqrt{15.2} = \mathbf{3.89}$$

Example 14

Calculate mean, SD and CV for the following data

Classes:	0-6	6-12	12-18	18-24	24-30
f :	5	12	30	10	3

Solution

Class	f	x	$d = \frac{x-15}{6}$	fd	fd^2
0-6	5	3	2	10	20
6-12	12	9	1	12	12
12-18	15	15	0	0	0
18-24	10	21	1	10	10
24-30	3	27	2	6	12
Total	60			6	54

$$\bar{x} = A + \frac{\sum fd}{N} \times c = 15 + \frac{-6}{60} \times 6$$

$$= 15 - \frac{6}{10} = 15 - 0.6 = \mathbf{14.4}$$

$$\sigma = c \times \sqrt{\frac{\sum fd^2}{N} - \left(\frac{\sum fd}{N}\right)^2} = 6 \times \sqrt{\frac{54}{60} - \left(\frac{-6}{60}\right)^2}$$

$$6 \times \sqrt{0.90 - 0.01} = 6 \times \sqrt{0.89}$$

$$= 6 \times 0.9434 = \mathbf{5.66}$$

$$\text{CV} = \frac{SD}{AM} \times 100 = \frac{5.66}{14.4} \times 100 = \mathbf{39.30\%}$$

Merits and Demerits

Merits

1. It is rigidly defined and its value is always definite and based on all the observations and the actual signs of deviations are used.
2. As it is based on arithmetic mean, it has all the merits of arithmetic mean.
3. It is the most important and widely used measure of dispersion.

4. It is possible for further algebraic treatment.
5. It is less affected by the fluctuations of sampling, and hence stable.
6. Squaring the deviations make all the deviations positive; as such there is no need to ignore the signs (as in mean deviation).
7. It is the basis for measuring the coefficient of correlation, sampling and statistical inferences.
8. The standard deviation provides the unit of measurement for the normal distribution.
9. It can be used to calculate the combined standard deviation of two or more groups.
10. The coefficient of variation is considered to be the most appropriate method for comparing the variability of two or more distributions, and this is based on mean and standard deviation.

Demerits

1. It is not easy to understand, and it is difficult to calculate.
2. It gives more weight to extreme values, because the values are squared up.
3. It is affected by the value of every item in the series.
4. As it is an absolute measure of variability, it cannot be used for the purpose of comparison.
5. It has not found favour with the economists and businessmen.

EXERCISES

Multiple Choice Questions

1. Which of the following is a unit less measure of dispersion?

   a. standard deviation b. mean deviation

   c. coefficient of variation d. range

2. Formula for coefficient of variation is:

   a. $C.V. = \frac{SD}{mean} \times 100$ b. $C.V. = \frac{mean}{SD} \times 100$

   c. $C.V. = \frac{mean \times SD}{100}$ d. $C.V. = \frac{100}{mean \times SD}$

3. For a symmetrical distribution, $M_d \pm QD$ covers:

   a. 25 percent of the observations

   b. 50 percent of the observations

   c. 75 percent of the observations

   d. 100 percent of the observations

   Md = median and Q.D. = quartile deviation.

4 Sum of squares of the deviations is minimum when deviations are taken from:

   a. mean b. median c. mode d. zero

5 If a constant value 5 is subtracted from each observations of a set, the variance is:

   a. reduced by 5 b. reduced by 25

   c. unaltered d. increased by 25

6. Which measure of dispersion ensures highest degreeof reliability?

   a. range b. mean deviation

   c. quartile deviation d. standard deviation

7. If the mean deviation of a distribution is 20.20, the standard deviation of the distribution is:

   a. 15.15 b. 25.25

   c. 30.30 d. none of the above

8. The mean of a series is 10 and its coefficient of variation is 40 percent, the variance of the series is:

   a. 4 b. 8 c. 12 d. none of the above

9. Which measure of dispersion can be calculated in case of open end intervals?

   a. range b. standard deviation
   c. coefficient of variation d. quartile deviation

Very Short Answer Questions

10. What are the uses of standard deviation?

11. Why measures of dispersion are called averages of second order?

12. For the numbers 3 and 5 show that SD = (1/2) Range.

13. Define CV and state its use.

14. State the desirable properties of a measure of dispersion

15. Define Quartile deviation.

16. Give the empirical relation connecting QD, MD and SD.

Short Essay questions

17. Define coefficient of variation. What is its relevance in economic studies?

18. What is a relative measure of dispersion? Distinguish between absolute and relative measure of dispersion.

19. Calculate coefficient of variation for the following distribution.

x	: 0	1	2	3	4	5	6
f	: 1	4	13	21	16	7	3

20. For the following data compute standard deviation,

x	: 10	20	30	40	50	60
f	: 3	5	7	20	8	7

21. Calculate median and quartile deviation for the following data

x	: 60	62	64	66	68	70	72
f	: 12	16	18	20	15	13	9

22. Calculate SD for the following data

Class interval	: 0-5	5-10	10-15	15-20	20-25	25-30
Frequency	: 4	8	14	6	3	I

Long Essay Questions

23. Compute coefficient of variation from the data given below.

Marks Less than	: 10	20	30	40	50	60	70	80	90	100
No. of students	: 5	13	25	48	65	80	92	97	99	100

24. Calculate the standard deviation of the following series.

More than	: 0	10	20	30	40	50	60	70
Frequency	: 100	90	75	50	25	15	5	0

25. The mean and the standard deviation of a group of 50 observations were calculated to be 70 and 10 respectively, It was later discovered that an observation 17 was wrongly-recorded as 70. Find the mean and the standard deviation (i) if the incorrect observation is omitted (ii) if the incorrect observation is replaced by the correct value.

26. Calculatethe standarddeviationandthecoefficientof variation of a raw data for which n = 50

$$(x_i - \overline{x}) = -10, \sum (x_i - \overline{x})^2 = 400$$

27. For two samples size 10 each, we have the followingvalues

$\sum x = 71;\ \sum x^2 = 555; \sum y = 70; \sum y^2 = 525$ compare the variability of these two samples.

28. Define coefficient of variation. Indicateits use. A factory produces two types of electric lamps A and B. In an experiment relating to their life,. the following results were obtained:

Life in	No. of lamps	
hours	A	B
500-700 ..	5	4
700-900 ..	11	30
900-1100 ..	26	12
1100-1300..	10	8
1300-1500..	8	6

Compare the variability of the two types of lamps using C.V.

MODULE II

CORRELATIONAND REGRESSION

CURVE FITTING

We have already studied the behaviour of a single variable characteristic by analysing a univariate data using the summary measures viz ; measures of central tendency, measures of dispersion measures of skewness and measures of kurtosis.

Very often in practice a relationship is found to exist between two (or more) variables. For example; there may exist some relation between heights and weights of a group of students; the yield of a crop is found to vary with the amount of rainfall over a particular period, the prices of some commodities may depend upon their demands in the market etc.

It is frequently desirable to express this relationship in mathematical form by formulating an equation connecting the variables and to determine the degree and nature of the relationship between the variables. Curve fitting, Correlation and Regression respectively serves these purposes.

Curve fitting

Let x be an independent variable and y be a variable depending on x; Here we say that y is a function of x and write it as y = f(x). If f(x) is a known function, then for any allowable values x_1, x_2,.... x_n of x. we can find the corresponding values y_1, y_2,.... y_n of y and thereby determine the pairs (x_1, y_1), (x_2, y_2) .... (x_n, y_n) which constitute a bivariate data. These pairs of values of x and y give us n points on the curve y = f(x).

Suppose we consider the converse problem. That is, suppose we are given n values x_1, x_2,.... x_n of an independent variable x and corresponding values y_1, y_2,.... y_n of a variable y depending on x. Then the pairs (x_1, y_1), (x_2, y_2) .... (x_n, y_n) give us n points in the xy-plane. Generally, it is not possible to find the actual curve y = f(x) that passes through these points. Hence we try to find a curve that serves as best approximation to the curve y = f(x). Such a curve is referred to as the curve of best fit. The process of determining a curve of best fit is called curve fitting. The method generally employed for curve fitting is known as the *method of least squares* which is explained below.

Method of least squares

This is a method for finding the unknown coefficients in a curve that serves as best approximation to the curve y = (f(x). The basic ideas of this method were created by A.M. Legendire and C.F. Gauss.

"The principle of least squares says that the sum of the squares of the error between the observed values and the corresponding estimated values should be the least."

Suppose it is desired to fit a k-th degree curve given by

$$y = a_0 + a_1 x + a_2 x^2 + \ldots\ldots + a_k x^k \qquad \ldots. (1)$$

to the given pairs of observations (x_1, y_1), (x_2, y_2) .... (x_n, y_n). The curve has k + 1 unknown constants and hence if n = k + 1 we get k + 1 equations on substituting the values of (x_i, y_i) in equation (1). This gives unique solution to the values $a_0\ a_1\ a_2 \ldots.. a_n$. However, if n > k + 1, no unique solution is possible and we use the method of least squares.

Now let

$y_e = a_0 + a_1 x + a_2 x^2 + \ldots\ldots + a_k x^k$ be the estimated value of y when x takes the value x_i. But the corresponding observed value of y is y_i. Hence if e_i is the residual or error for this point,

$$e_i = y_i - y_e = y_i - a_0 - a_1 x_i - a_2 x_i^2 - \ldots\ldots - a_k x_i^k$$

To make the sum of squares minimum, we have to minimise.

$$S = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left(y_i - a_0 - a_1 x_i - a_2 x_i^2 - \ldots.. - a_k x_i^k\right)^2 \ldots. (2)$$

By differential calculus, S will have its minimum value when

$$\frac{\partial S}{\partial a_0} = 0,\ \frac{\partial S}{\partial a_1} = 0, \ldots.\ \frac{\partial S}{\partial a_k} = 0$$

which gives k + 1 equations called *normal equations*. Solving these equations we get the best values of $a_0, a_1, a_2 \ldots.. a_k$. Substituting these values in (1) we get the curve of best fit.

Now we consider the fitting of some curves.

1. Fitting of a Straight Line y = a + bx

Suppose we wish to have a straight line that serves as best approximation to the actual curve y = f(x) passing through n given points (x_1, y_1), (x_2, y_2) .... (x_n, y_n). This line will be referred to as the *line of best fit* and we take its equation as

$$y = a + bx \qquad \ldots. (1)$$

where *a* and b are the parameters to be determined. Let y_e be the value of y corresponding to the value x_i of x as determined by equation (1). The value y_e is called the estimated value of y.

When x = x_i, the observed valued of y is y_i. Then the difference y_i y_e is called residual or error. By the principle of least squares, we have

$$S = \Sigma (y_i - y_e)^2 \qquad \ldots. (2)$$

We determine *a* and b so that S is minimum (least). Two necessary conditions for this are

$$\frac{\partial S}{\partial a} = 0, \qquad \frac{\partial S}{\partial b} = 0$$

Using (2), these conditions yield the following equations:

$$(y_i - a - bx_i) = 0$$

or $$\Sigma y_i = n a + b\Sigma x_i \qquad \ldots. (3)$$

and $\Sigma (y_i - a - bx_i)\, x_i = 0$

$$\Sigma x_i y_i = a \Sigma x_i + b \Sigma x_i^2 \qquad \ldots. (4)$$

These two equations, (3) and (4), called normal equations, serve as two simultaneous equations for determining *a* and b. Putting the values of *a* and b so determined in (1), we get the equation of the line of best fit for the given data.

Example 1

Fit a straight line to the following data:

x	:	1	2	3	4	5
y	:	14	13	4	5	2

Estimate the value of y when x = 3.5

Solution

We note that n = 5, and form the following Table.

x_i	y_i	x_i^2	$x_i y_i$
1	14	1	14
2	13	4	26
3	9	9	27
4	5	16	20
5	2	25	10
$\Sigma x_i = 15$	$\Sigma y_i = 43$	$\Sigma x_i^2 = 55$	$\Sigma x_i y_i = 97$

Hence the normal equations that determine the line of best fit are

$$= 5a + 15b$$

$$= 15a + 55b$$

These give a = 18.2 and b = 3.2. Hence, for the given data, the line of best fit is y = **18.2 3.2x**.

When x = 3.5, the estimated value of y (found from the line of best fit) is y = 18.2 (3.2) × (3.5) = **7**

2. Fitting of a Parabola $y = a + bx + cx^2$

Suppose we wish to have a parabola (second degree curve) as the curve of best fit for a data consisting of n given pairs (x_i, y_i), i = 1, 2.... n.

Let us take the equation of the parabola, called ‘parabola of best fit’ in the form

$$y = a + bx + cx^2 \qquad .... (1)$$

where a, b, c are constants to be determined.

Let y_e be the value of y corresponding to the value x_i of x determined by equation (1). Then the sum of squares of the error between observed value of y and estimated value of y is given by

$$S = \sum_{i=1}^{n} (y_i - y_e)^2$$

Using (1), this becomes, $S = \sum_{i=1}^{n} (y_i - a - bx_i - cx_i^2)^2$ .... (2)

We determine a, b, c so that S is least. Three necessary conditions for this are $\frac{\partial s}{\partial a} = 0$, $\frac{\partial S}{\partial b} = 0$, and $\frac{\partial S}{\partial c} = 0$. Using (2) these conditions yield the following normal equations.

$$\Sigma y_i = na + b\,\Sigma x_i + c\,\Sigma x_i^2 \qquad .... (3)$$

$$\Sigma x_i y_i = a\,\Sigma x_i + b\,\Sigma x_i^2 + c\,\Sigma x_i^3 \qquad .... (4)$$

$$\Sigma x_i^2 y_i = a\,\Sigma x_i^2 + b\,\Sigma x_i^3 + c\,\Sigma x_i^4 \qquad .... (5)$$

Solve (3), (4) and (5) for determining a, b and c. Putting the values of a, b, c so determined in (1) we get the equation of parabola of best fit for the given data.

Example 2

Fit a parabola to the following data:

x	:	1	2	3	4	5	6	7	8	9
y	:	2	6	7	8	10	11	11	10	9

Estimate y when x = 4.5

Solution

Here n = 9, and we form the following Table:

x_i	y_i	x_i^2	x_i^3	x_i^4	$x_i\ y_i$	$x_i^2\ y_i$
1	2	1	1	1	2	2
2	6	4	8	16	12	24
3	7	9	27	81	21	63
4	8	16	64	256	32	128
5	10	25	125	625	50	250
6	11	36	216	1296	66	396
7	11	49	343	2401	77	539
8	10	64	512	4096	80	640
9	9	81	729	6561	81	729
$\Sigma x_i = 45$	$\Sigma y_i = 74$	$\Sigma x_i^2 = 285$	$\Sigma x_i^3 = 2025$	$\Sigma x_i^4 = 15333$	$\Sigma x_i y_i = 421$	$\Sigma x_i^2 y_i = 2771$

The normal equations that determine the parabola of best fit are

$$= 94a + 45b + 285c$$

$$= 45a + 285b + 2025c$$

$$2771 = 285a + 2025b + 15333c$$

Solving these equations, we obtain $a =$ 0.9282,

b= 3.523 and c = 0.2673. Hence the parabola of best fit for the given data is

$$\mathbf{y = -0.9282 + 3.523x \quad 0.2673x^2}$$

For x = 4.5, the parabola of best fit gives the estimated value of y as

y = 0.9282 + 3.523 × 4.5 0.2673 × 4.5^2 = **9.5121**

EXERCISES

Multiple choice questions

The equation Y = ab^{-x} for $\beta < 1$ represents

exponential growth curve

exponentially decay curve

a parabola

none of the above

For fitting of curves, we use

method of moments

method of least squares

method of maximum likelyhood

all the above

While fitting a straight line $y = a + bx$, the value of b measures a. the rate of change in y w.r.t x

b. the proportional variation in y w.r.t the variation in x

c. both (a) and (b)

d. neither (a) nor (b)

Very short answer questions

What is the principle of least squares?

What do you mean by curve fitting?

What are normal equations?

7. Write down the normal equation to fit a straight line y = ax + b.

Short essay questions

How will you fit a straight line to the given data by method of least squares?

What is method of least squares? How will you use it to fit a parabola of second degree?

Explain the principle of least squares method of fitting of a second degree curve of the form Y = a + bx + cx^2 for ‘n’ pairs of values.

Write short notes on

Method of least squares

Curve fitting

Normal equations.

Long essay questions

Fit a straight line by the method of least squares to the following data.

x:	0	1	2	3	4
y:	1	1.8	3.3	4.5	6.3

13. Fit a straight line y = *a* + bx to the following data.

x:	1	2	3	4	6	8
y:	2.4	3	3.6	4	5	6

14. Fit a straight line y = *a*x + b to the following data.

x:	1	2	3	4	5	6	7
y:	80	90	92	83	94	99	92

15. Fit a parabola $y = a + bx + x^2$ to the following data:

x:	0	1	2	3	4
y:	1	1.8	1.3	2.5	6.3

16. Fit a curve of the form $y = ax + bx^2$ for the data given below.

x:	1	2	3	4	5
y:	1.8	5.1	8.9	14.1	19.8

CORRELATIONAND REGRESSION

Introduction

In the earlier chapters we have discussed the characteristics and shapes of distributions of a single variable, eg, mean, S.D. and skewness of the distributions of variables such as income, height, weight, etc. We shall now study two (or more) variables simultaneously and try to find the quantitative relationship between them. For example, the relationship between two variables like (1) income and expenditure (2) height and weight, (3) rainfall and yield of crops, (4) price and demand, etc. will be examined here. The methods of expressing the relationship between two variables are due mainly to Francis Galton and Karl Pearson.

Correlation

Correlation is a statistical measure for finding out degree (or strength) of association between two (or more) variables. By ‘association’ we mean the tendency of the variables to move together. Two variables X and Y are so related that movements (or variations) in one, say X, tend to be accompanied by the corresponding movements (or variations) in the other Y, then X and Y are said to be correlated. The movements may be in the same direction (i.e. either both X, Y increase or both of them decrease) or in the opposite directions (ie., one, say X, increases and the other Y decreases). Correlation is said to be positive or negative according as these movements are in the same or in the opposite directions. If Y is unaffected by any change in X, then X and Y are said to be uncorrelated.

In the words L.R. Conner:

If two or more quantities vary in sympathy so that movements in the one tend to be accompanied by corresponding movements in the other, then they are said to be correlated.”

Correlation may be linear or non-linear. If the amount of variation in X bears a constant ratio to the corresponding amount of variation in Y, then correlation between X and Y is said to be linear. Otherwise it is non-linear. Correlation coefficient (r) measures the degree of linear relationship, (i.e., linear correlation) between two variables.

Determination of Correlation

Correlation between two variables may be determined by any one of the following methods:

Scatter Diagram

Co-variance Method or Karl Pearson's Method

Rank Method

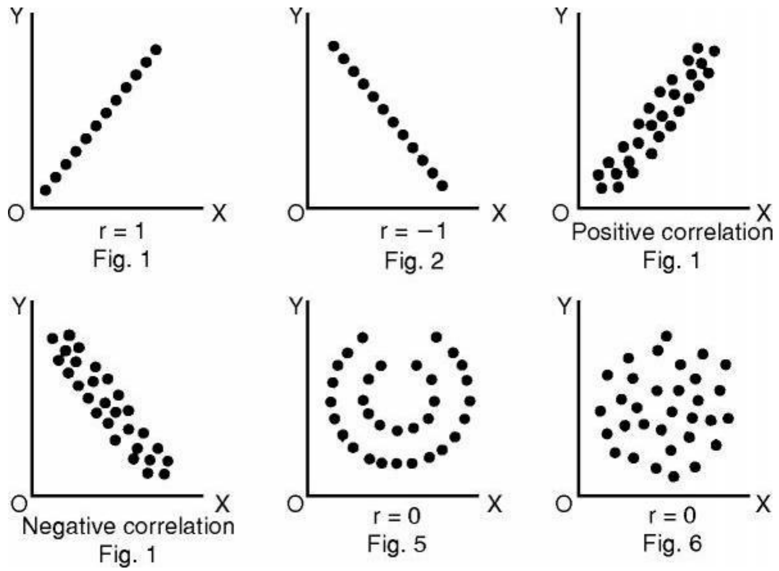
Scatter Diagram

The existence of correlation can be shown graphically by means of a *scatter diagram*. Statistical data relating to simultaneous movements (or variations) of two variables can be graphically represented-by points. One of the two variables, say X, is shown along the horizontal axis OX and the other variable Y along the vertical axis OY. All the pairs of values of X and Y are now shown by points (or dots) on the graph paper. This diagrammatic representation of bivariate data is known as scatter diagram.

The scatter diagram of these points and also the direction of the scatter reveals the nature and strength of correlation between the two variables. The following are some scatter diagrams showing different types of correlation between two variables.

In Fig. 1 and 3, the movements (or variations) of the two variables are in the same direction and the scatter diagram shows a linear path. In this case, correlation is positive or direct.

In Fig. 2 and 4, the movements of the two variables are in opposite directions and the scatter shows a linear path. In this case correlation is negative or indirect.



r = 1
Fig. 1

r = −1
Fig. 2

Positive correlation
Fig. 1

Negative correlation
Fig. 1

r = 0
Fig. 5

r = 0
Fig. 6

In Fig. 5 and 6 points (or dots) instead of showing any linear path lie around a curve or form a swarm. In this case correlation is very small and we can take r = 0.

In Fig. 1 and 2, all the points lie on a straight line. In these cases correlation is perfect and r = +1 or 1 according as the correlation is positive or negative.

Karl Pearson's Correlation Coefficient

We have remarked in the earlier section that a scatter diagram gives us only a rough idea of how the two variables, say x and y, are related. We cannot draw defensible conclusions by merely examining data from the scatter diagram. In other words, we cannot simply look at a scatter diagram variables. On the other hand, neither can we conclude that the correlation at all. We need a quantity (represented by a number), which is a measure of the extent to which x and y are related. The quantity that is used for this purpose is known as the Co-efficient of Correlation, usually denoted by r_{xy} or r. The co-efficient of correlation r_{xy} measures the degree (or extent) of relationship between the two variables x and y and is given by the following formula:

$$r_{xy} = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{n\, \sigma_X\, \sigma_Y} \qquad \text{.... (1)}$$

where X_t and Y_i $(i = 1, 2,.., n)$ are the two sets of values of x and y respectively and $\bar{x}$, $\bar{y}$, σ_x, σ_y are respectively the corresponding means and standard deviations so that

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, \qquad \bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

$$\sigma_x^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \frac{1}{n}\Sigma X_i^2 - \bar{X}^2$$

$$\text{and } \sigma_y^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \frac{1}{n}\Sigma Y_i^2 - \bar{Y}^2$$

The above definition of the correlation co-efficient was given by Karl Pearson in 1890 and is called *Karl Pearson's Correlation Co-efficient* after his name.

Definition

If (X_1, Y_1), (X_2, Y_2) .... (X_n, Y_n) be n pairs of observations on two variables X and Y, then the covariance of X and Y, written as cov (X,Y) is defined by

$$\text{Cov (X,Y)} = \frac{1}{n}\Sigma (X_i - \bar{X})(Y_i - \bar{Y})$$

Covariance indicates the joint variations between the two variables.

So the correlation coefficient or the coefficient of correlation (r) between X and Y is defined by

$$r = \frac{\underline{\text{Cov (X, Y)}}}{\sigma_X\, \sigma_Y}$$

where σ_x, σ_y are standard deviations of X and Y respectively.

The formula for the Correlation Coefficient r may be written in different forms.

i. If $\quad x_i = X - \bar{X}$ and $y_i = Y - \bar{Y}$

then $\quad r = \dfrac{\Sigma x_i y_i}{n\, \sigma_X\, \sigma_Y} \qquad (1)$

$$\therefore \quad \text{from (1), r} = \frac{\frac{1}{n}\Sigma x_i y_i}{\sqrt{\frac{\Sigma x_i^2}{n}} \times \sqrt{\frac{\Sigma y_i^2}{n}}} = \frac{\Sigma x_i y_i}{\sqrt{\Sigma x_i^2} \times \sqrt{\Sigma y_i^2}}$$

ii. We have

$$\text{Cov (X, Y)} = \frac{1}{n}\Sigma (X_i - \bar{X})(Y_i - \bar{Y})$$

$$\frac{1}{n}\Sigma (X_i Y_i - \bar{X}_i \bar{Y} - X\bar{Y}_i + \bar{X}\bar{Y})$$

$$= \frac{\Sigma X_i Y_i}{n} - \bar{Y}\frac{\Sigma X_i}{n} - \bar{X}\frac{\Sigma Y_i}{n} + \frac{n\bar{X}\bar{Y}}{n}$$

$$\frac{\Sigma X_i Y_i}{n} - \bar{X}\bar{Y} - \bar{X}\bar{Y} + \bar{X}\bar{Y}$$

$$= \frac{\Sigma X_i Y_i}{n} - \bar{X}\bar{Y} = \frac{\Sigma X_i Y_i}{n} - \left(\frac{\Sigma X_i}{n}\right)\left(\frac{\Sigma Y_i}{n}\right)$$

and conclude that since more than half of the points appear to be nearly in a straight line, there is a positive or negative correlation between the

Now, $$r = \frac{Cov(X, Y)}{\sigma_X \sigma_Y}$$

$$= \frac{\frac{\Sigma XY}{n} - \left(\frac{\Sigma X_i}{n}\right)\left(\frac{\Sigma Y}{n}\right)}{\sqrt{\frac{\Sigma X^2}{n} - \left(\frac{\Sigma X_i}{n}\right)^2} \times \sqrt{\frac{\Sigma Y^2}{n} - \left(\frac{\Sigma Y}{n}\right)^2}} \qquad ...(2)$$

iii. By multiplying each term of (2) by n^2, we have

$$\mathbf{r} = \frac{n \Sigma X_i Y_i - (\Sigma X_i)(\Sigma Y_i)}{\sqrt{n \Sigma X_i^2 - (\Sigma X_i)^2} \times \sqrt{n \Sigma Y_i^2 - (\Sigma Y_i)^2}}$$

Theorem

The correlation coefficient is independent (not affected by) of the change of origin and scale of measurement.

Proof

Let $(x_1, y_1), (x_2, y_2)$ .... (x_n, y_n) be a set of n pairs of observations.

$$r_{xy} = \frac{\frac{1}{n} \Sigma (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\frac{1}{n} \Sigma (x_i - \bar{x})^2} \sqrt{\frac{1}{n} \Sigma (y_i - \bar{y})^2}} \qquad ...(1)$$

Let us transform x_i to u_i and y_i to v_i by the rules,

$$u_i = \frac{x_i - x_0}{c_1} \quad \text{and } v_i = \frac{y_i - y_0}{c_2} \qquad ...(2)$$

where x_0, y_0, c_1, c_2 are arbitrary constants.

From (2), we have

$$x_i = c_1 u_i + x_0 \quad \text{and } y_i = c_2 v_i + y_0$$

$$\bar{x} = x_0 + c_1 \bar{u} \quad \text{and } \bar{y} = y_0 + c_2 \bar{v}$$

where $\bar{u}$ and $\bar{v}$ are the means u_i^s and v_i^s respectively.

$$x_i - \bar{x} = c_1 (u_i - \bar{u}) \quad \text{and } y_i - \bar{y} = c_2 (v_i - \bar{v})$$

Substituting these values in (1), we get

$$r_{xy} = \frac{\frac{1}{n} \sum_{i=1}^{n} c_1 (u_i - \bar{u})\, c_2 (v_i - \bar{v})}{\sqrt{\frac{1}{n} \sum_{i=1}^{n} c_1^2 (u_i - \bar{u})^2} \sqrt{\frac{1}{n} \sum_{i=1}^{n} c_2^2 (v_i - \bar{v})^2}}$$

$$= \frac{\frac{1}{n} \sum_{i=1}^{n} (u_i - \bar{u})(v_i - \bar{v})}{\sqrt{\frac{1}{n} \sum_{i=1}^{n} (u_i - \bar{u})^2} \sqrt{\frac{1}{n} \sum_{i=1}^{n} (v_i - \bar{v})^2}}$$

$$= \frac{\sum_{i=1}^{n} (u_i - \bar{u})(v_i - \bar{v})}{n \sigma_u \sigma_v} = r_{uv}$$

Here, we observe that if we change the origin and choose a new scale, the correlation co-efficient remains unchanged. Hence the proof.

Here, r_{uv} can be further simplified as

$$r_{xy} = \frac{Cov(u, v)}{\sigma_u \sigma_v}$$

$$= \frac{\frac{1}{n}\sum_{i=1}^{n} u_i v_i - \bar{u}\,\bar{v}}{\sqrt{\frac{1}{n}\Sigma u_i^2 - \bar{u}^2}\sqrt{\frac{1}{n}\Sigma v_i^2 - \bar{v}^2}}$$

$$\frac{n\,\Sigma\, u_i v_i - \Sigma\, u_i \Sigma v_i}{\sqrt{n\,\Sigma\, u_i^2 - (\Sigma\, u_i)^2}\sqrt{n\,\Sigma\, v_i^2 - (\Sigma v_i)^2}}$$

Limits of Correlation Co-efficient

We shall now find the limits of the correlation coefficient between two variables and show that it lies between–1 and + 1 .

ie., $-1 \le r_{xy} < +1$

Proof

Let (x_1, y_1) , (x_2, y_2) .... (x_n, y_n) be the given pairs of observations.

Then $$r_{xy} = \frac{\frac{1}{n}\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\frac{1}{n}\Sigma(x_i - \bar{x})^2}\sqrt{\frac{1}{n}\Sigma(y_i - \bar{y})^2}}$$

We put

$$X_i = x_i - \bar{x}, \qquad Y_i = y_i - \bar{y}$$

$$\sigma_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 \qquad \text{. . . (1)}$$

Similarly $$\sigma_y^2 = \frac{1}{n}\sum_{i=1}^{n} Y_i^2 \qquad \text{...(2)}$$

and $$r_{xy} = \frac{\sum_{i=1}^{n} X_i Y_i}{n\sigma_x \sigma_y} \qquad \text{...(3)}$$

Now we have

$$\sum_{i=1}^{n}\left(\frac{X_i}{\sigma_x} \pm \frac{Y_i}{\sigma_y}\right)^2 = \frac{\sum_{i=1}^{n} X_i^2}{\sigma_x^2} + \frac{\sum_{i=1}^{n} Y_i^2}{\sigma_y^2} + \frac{2\sum_{i=1}^{n} X_i Y_i}{\sigma_x \sigma_y}$$

$$= \frac{n\,\sigma_x^2}{\sigma_x^2} + \frac{n\,\sigma_y^2}{\sigma_y^2} \pm 2nr_{xy} \quad \text{using (1), (2), (3).}$$

$$2n \pm 2n\,r_{xy} = 2n\,(1 \pm r_{xy})$$

Left hand side of the above identity is the sum of the squares of n numbers and hence it is positive or zero.

Hence, $1 \pm r_{xy} \ge 0$or, $r_{xy} \le 1$ and $r_{xy} \ge -1$

or $-1 \le r_{xy} \le +1$

ie., the correlation co-efficient lies between 1 and + 1. Hence the proof.

Note:

If r_{xy} = 1, we say that there is perfect positive correlation between x and y.

If r_{xy} = 1, we say that there is perfect negative correlation between x and y.

If r_{xy} = 0, we say that there is no correlation between the two variables, i.e., the two variables are uncorrelated.

If r_{xy} > 0, we say that the correlation between x and y is positive (direct).

If r_{xy} < 0, we say that the correlation between x and y is negative (indirect).

X	Y	x = X $\overline{X}$	y = Y $\overline{Y}$	x^2	y^2	xy
1	6	3	4	9	16	12
2	8	2	2	4	4	4
3	11	1	1	1	1	1
4	9	0	1	0	1	0
5	12	1	2	1	4	2
6	10	2	0	4	0	0
7	14	3	4	9	16	12
28	70			28	42	29

$$\overline{X} = \frac{\Sigma X}{n} = \frac{28}{7} = 4 \text{ an d } \overline{Y} = \frac{\Sigma Y}{n} = \frac{70}{7} = 10$$

Karl Pearson’s coefficient of correlation (r) is given by

$$r = \frac{xy}{\sqrt{\Sigma x^2}\sqrt{\Sigma y^2}}$$

$$= \frac{29}{\sqrt{28}\sqrt{42}} = \mathbf{0.8457}$$

Example 9

Karl Pearson’s coefficient of correlation between two variables X and Y is 0.28 their covariance is +7.6. If the variance of X is 9, find the standard deviation of Y-series.

Solution

Karl Pearson’s coefficient of correlation r is given by

$$r = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y}$$

Here r =0.28, Cov (X, Y) = 7.6 and $\sigma_X^2 = 9$; $\sigma_X = 3$.

Using (1) $0.28 = \frac{7.6}{\ }$

or, $0.84\,\sigma_Y = 7.6, or$ $\sigma_Y = \frac{7.6}{0.84} = \frac{760}{84}$

$= \mathbf{9.048}$

Example 10

Calculate Pearson’s coefficient of correlation between advertisement cost and sales as per the data given below:

Advt cost in ’000 Rs:	39	65	62	90	82	75	25	98	36	78
Sales in lakh Rs:	47	53	58	86	62	68	60	91	51	84

Solution

Karl Pearson’s coefficient of correlation (r) is given by

$$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \times}\sqrt{\Sigma y^2}} \text{ wh er e } x = X - \overline{X} \text{ an d } y = Y - \overline{Y}$$

X	Y	$x = X - \overline{X}$	$y = Y - \overline{Y}$	x^2	y^2	xy
39	47	26	19	676	361	494
65	53	0	13	0	169	0
62	58	3	8	9	64	24
90	86	25	20	625	400	500
82	62	17	4	289	16	68
75	68	10	2	100	4	20
25	60	40	6	1600	36	240
98	91	33	25	1089	625	825

Example 8

Find the coefficient of correlation from the following data:

X :	1	2	3	4	5	6	7
Y :	6	8	11	9	12	10	14

36	51	29	15	841	225	435
78	84	13	18	169	324	234
650	660	0	0	5398	2224	2704

$$\overline{X} = \frac{\Sigma X}{n} = \frac{650}{10} = 65 \; ; \; \overline{Y} = \frac{\Sigma Y}{n} = \frac{660}{10} = 66$$

$$r = \frac{2704}{\sqrt{5398} \times \sqrt{2224}} = \mathbf{0.78}$$

Example 11

Calculate Pearson’s coefficient of correlation from the following taking 100 and 50 as the assumed average of X and Y respectively:

X:	104	111	104	114	118	117	105	108	106	100	104	105
Y:	57	55	47	45	45	50	64	63	66	62	69	61

Solution

X	Y	u = X 100	v = Y 50	u^2	v^2	uv
104	57	4	7	16	49	28
111	55	11	5	121	25	55
104	47	4	3	16	9	12
114	45	14	5	196	25	70
118	45	18	5	324	25	90
117	50	17	0	289	0	0
105	64	5	14	25	196	70

$$\Sigma X\,Y \;=\; 508 - (6 \times 14 + 8 \times 6) + (8 \times 12 + 6 \times 8)$$

108	63	8	13	64	169	104
106	66	6	16	36	256	96
100	62	0	12	0	144	0
104	69	4	19	16	361	76
105	61	5	11	25	121	55
		96	84	1128	1380	312

$$r = \frac{\Sigma u_i v_i - \Sigma u_i \Sigma v_i}{\sqrt{n \Sigma u_i^2 - (\Sigma u_i)^2}\sqrt{n \Sigma v_i^2 - (\Sigma v_i)^2}}$$

$$= \frac{12 \times 312 - 96 \times 84}{\sqrt{12 \times 1128 - (96)^2}\sqrt{12 \times 1380 - (84)^2}}$$

0.67

Example 12

A computer while calculating the correlation coefficient between two variables X and Y from 25 pairs of observations obtained the following results: $n = 25, \Sigma X = 125, \Sigma Y = 100, \Sigma X^2 = 650, \Sigma Y^2 = 460$ and $X\,Y = 508$. It was, however, discovered at the time of checking that two pairs of observations were not correctly copied. They were taken as (6, 14) and (8, 6), while the correct values were (8, 12) and (6, 8). Prove that the correct value of the correlation coefficient should be 2/3.

Solution

When the two incorrect pairs of observations are replaced by the correctpairs, the revised results for the whole series are:

ΣX = 125 (Sum of two incorrect values of X) +
(Sum of two correct values of X)

= 125 (6 + 8) + (8 + 6) = 125

Similarly

ΣY = 100 (14 + 6) + (12 + 8) = 100

ΣX^2 = 650 $(6^2 + 8^2) + (8^2 + 6^2)$ = 650

ΣY^2 = 460 $(14^2 + 6^2) + (12^2 + 8^2)$

= 460 232 + 208 = 436 and

= 508   132 + 144 = 520 ;

Correct value of the correlation co efficient is

$$r = \frac{n \Sigma XY - (\Sigma X)(\Sigma Y)}{\sqrt{n \Sigma X^2 - (\Sigma X^2)}\sqrt{n \Sigma Y^2 - (\Sigma Y^2)}}$$

$$= \frac{25 \times 520 - 125 \times 100}{\sqrt{25 \times 650 - 125^2}\sqrt{2 \times 5\ 436 - 100^2}}$$

2/3

Rank Correlation Coefficient

Simple correlation coefficient (or product-moment correlation coefficient) is based on the magnitudes of the variables. But in many situations it is not possible to find the magnitude of the variable at all. For example, we cannot measure beauty or intelligence quantitatively. In this case, it is possible to rank the individuals in some order. Rank correlation is based on the rank or the order and not on the magnitude of the variable. It is more suitable if the individuals (or variables) can be arranged in order of merit or proficiency. If the ranks assigned to individuals range from 1 to n, then the Karl Pearson's correlation coefficient between two series of ranks is called Rank correlation coefficient. Edward Spearman's formula for Rank correlation coefficient (R) is given by.

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

where d is the difference between the ranks of the two series and n is the number of individuals in each series.

Derivation of Spearman's Formula for Rank Correlation Coefficient

$$R = 1 - \frac{6 \Sigma d^2}{n(n^2 - 1)}$$

Proof:

Let $(x_1, y_1), (x_2, y_2), \ldots (x_n, y_n)$ be the ranks of n individuals in two characters (or series) Edward Spearman's Rank correlation coefficient R is the product-moment correlation coefficient between these ranks and, therefore, we can write.

$$R = \frac{Cov(x, y)}{\sigma_x \sigma_y} \qquad \ldots(1)$$

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

But the ranks of n individuals are the natural numbers 1, 2,.... n arranged in some order depending on the qualities of the individuals.

$x_1, x_2, \ldots, x_n$ are the numbers 1, 2... n in some order**.**

$$\therefore \Sigma x = 1 + 2 + \ldots + n = \frac{n(n+1)}{2} \text{ and}$$

$$\Sigma x^2 = 1^2 + 2^2 + \ldots + n^2 = \frac{n(n+1)(2n+1)}{6}$$

$$- = \frac{\Sigma x}{n} = \frac{n+1}{2}, \frac{\Sigma x^2}{n} = \frac{(n+1)(2n+1)}{6}$$

$$\therefore \sigma_x^2 = \frac{\Sigma x^2}{n} - \left(\frac{\Sigma x}{n}\right)^2 = \frac{(n+1)(2n+1)}{6} - \frac{(n+1)^2}{4}$$

$$= \left(\frac{n+1}{12}\right)(4n + 2 - 3n - 3) = \frac{n^2 - 1}{12}$$

similarly,

$$\bar{y} = \frac{n+1}{2} \text{ an d } \sigma_y^2 = \frac{n^2 - 1}{12}$$

Let $d_i = x_i - y_i$; then $d_i = (x_i - \bar{x}) - (y_i - \bar{y})$ [$\bar{x} = \bar{y}$]

Calculate the rank correlation coefficient.

$$\therefore \frac{\Sigma d_i^2}{n} = \frac{\Sigma\{(x_i-\bar{x})-(y_i-\bar{y})\}^2}{n}$$

$$= \frac{\Sigma(x_i-\bar{x})^2}{n} + \frac{\Sigma(y_i-\bar{y})^2}{n} - \frac{2\,\Sigma(x_i-\bar{x})(y_i-\bar{y})}{n}$$

$$\sigma_X^2 + \sigma_Y^2 - 2\operatorname{cov}(x,y)$$

or, 2cov (x, y) =

$$\frac{n^2-1}{12} + \frac{n^2-1}{12} - \frac{\Sigma d_i^2}{n} = \frac{2(n^2-1)}{12} - \frac{\Sigma d_i^2}{n}$$

or, cov (x, y) $= \frac{n^2-1}{12} - \frac{\Sigma d_i^2}{2n}$

Hence, from (1), we get

$$R = \left(\frac{n^2-1}{12} - \frac{\Sigma d^2}{2n}\right) \Big/ \left(\frac{n^2-1}{12}\right)$$

$$= \mathbf{1 - \frac{6\,\Sigma d^2}{n(n^2-1)}} \quad \text{[omitting i]}$$

Example 13

Student (Roll No.)	1	2	3	4	5	6	7	8	9	10
Marks in Maths.	78	36	98	25	75	82	90	6 2	6 5	6 9
Marks in Stat.	84	51	91	60	68	62	86	58	53	47

Solution

In Mathematics, Student with Roll No. 3 gets the highest mark 98 and is ranked 1; Roll No. 7 securing 90 marks has rank 2 and so on. Similarly, we can find the ranks of students in statistics.

Roll No.	Mathematics Marks	Rank (x)	Statistics Marks	Rank (y)	Rank Diff. d = x y	2
1	78	4	84	3	1	1
2	36	9	51	9	0	0
3	98	1	91	1	0	0
4	25	10	60	6	4	16
5	75	5	68	4	1	1
6	82	3	62	5	2	4
7	90	2	86	2	0	0
8	62	7	58	7	0	0
9	65	6	53	8	2	4
10	39	8	47	10	2	4
Total						30= Σd^2

Applying Edward Spearman’s formula:

$$R = 1 - \frac{6\,\Sigma d^2}{n(n^2-1)}$$

$$= 1 - \frac{6 \times 30}{10(10^2-1)} = 1 - \frac{18}{99}$$

$$= 1 - \frac{2}{11} = \frac{9}{11} = \mathbf{0.82}$$

Regression

In some situations, one may need to know the probable value of one variable corresponding to certain value of another variable. This is possible using the mathematical relation between the two variables. Scatter diagram, explained above helps to ascertain the nature of relationship such as linear (straight line), second degree polynomial (parabola), etc. Discussion in this book is restricted to linear relation between two variables.

During study of hereditary characteristics, Sir Francis Galton found that the heights of different groups of sons had the tendency to *regress*, that is to *go back* towards the overall average height of all groups of fathers. He called the lines of the average relationship as the lines of the regression. It is also referred to as the estimating equations because based on the value of one variable one can predict or estimate the value of the other variable.

Suppose we are given n pairs of values (x_1, y_1) (x_2, y_2), ... $(x_n\ y_n)$ of two variables x and y. If we fit a straight line to this data by taking x as independent variable and y as dependent variable, then the straight line obtained is called the *regression line of y on x.* Its slope is called the *regression coefficient of y on x.* Similarly, if we fit a straight line to the data by taking y as independent variable and x as dependent variable, the line obtained is the *regression line of x on y;* the reciprocal of its slope is called the *regression coefficient of x on y.*

Equation for regression lines

Let $\quad y = a + bx \quad$ .... (1)

be the equation of the regression line of y on x, where *a* and b are determined by solving the normal equations obtained by the principle of least squares.

$$\Sigma y_i = na + b\,\Sigma x_i \qquad \text{.... (2)}$$

$$\Sigma x_i y_i = a\,\Sigma x_i + b\,\Sigma x_i^2 \qquad \text{.... (3)}$$

Divide the equation (2) by n, we get

$$\frac{1}{n}\Sigma y_i = a + \frac{1}{n} b\,\Sigma x_i$$

or $\quad \bar{y} = a + b\,\bar{x} \quad$ .... (4)

where $\bar{x}$ and $\bar{y}$ are the means of x and y series. Substituting for *a* from (4) in (1), we get the equation,

$$y - \bar{y} = b\,(x - \bar{x}) \qquad \text{.... (5)}$$

Solving the equations (2) and (3) for b after eliminating *'a'* we get the value of b as

$$b = \frac{\Sigma x_i y_i - (\Sigma x_i)(\Sigma y_i)}{n\,\Sigma x_i^2 - (\Sigma x_i)^2}$$

$$= \frac{\dfrac{\Sigma x_i y_i}{n} - \bar{x}\,\bar{y}}{\dfrac{\Sigma x_i^2}{n} - \bar{x}^2} \text{, dividing each term by } n^2$$

$$= \frac{Cov(x, y)}{\sigma_x^2} = \frac{P_{xy}}{\sigma_x^2}$$

Substituting b in (5), we get the regression equation of y on x as

$$y - \bar{y} = \frac{P_{xy}}{\sigma_x^2}(x - \bar{x}) \qquad \text{.... (6)}$$

Similarly, when x is depending on y, the regression equation of x on y is obtained as

$$x - \bar{x} = \frac{P_{xy}}{\sigma_y^2}(y - \bar{y}) \qquad \text{.... (7)}$$

Let us denote $\dfrac{P_{xy}}{\sigma_x^2}$ as b_{yx} and $\dfrac{P_{xy}}{\sigma_x^2}$ as b_{xy}

Thus $b_{yx} = \dfrac{P_{xy}}{\sigma_x^2}$ as $b_{xy} = \dfrac{P_{xy}}{\sigma_y^2}$

Here b_{yx} is called the regression coefficient of y on x and b_{xy} is called the regression coefficient of x on y.

So we can rewrite the regression equation of y on x as

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

and the regression equation of x on y as

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

Some remarks

1.The slope of the regression line of y on x is $b_{xv} = \frac{r\,\sigma_y}{\sigma_x}$ and the slope of the regression line of x on y is the reciprocal of b_{xv} which is $\frac{\sigma_y}{r\sigma_x}$.

2. Since $b_{yx} = r\,(\sigma_y / \sigma_x)$ and σ_x and σ_y are positive, it follows that r has the same sign as that of b_{yx}.

3. Since $b_{xy} = r\,(\sigma_x / \sigma_y)$ we readly find that $(b_{yx})\,(b_{xy}) = r^2$. Since $r^2 > 0$. It follows that b_{xy} has the same sign as that of b_{yx}. Thus, r, b_{xy} and b_{yx} always have the same signs. Also $|r| = \sqrt{(b_{yx})(b_{xy})}$. That is, |r| is the geometric mean of b_{xy} and b_{yx}. Since |r| < 1 it follows that $b_{yx} > 1$ whenever $b_{xy} < 1$ and vice-versa.

4. Since the arithmetic mean is always greater than the geometric mean for any two numbers, we have $\frac{1}{2}(b_{yx} + b_{xy}) > \sqrt{b_{yx} \times b_{xy}} = |r|$.

Thus, the arithmetic mean of b_{xy} and b_{yx} is always greater than the coefficient of correlation.

5. The two lines of regression always pass through the point $(\bar{x}, \bar{y})$.

6. The regression equation of y on x is need for estimating or predicting the value of y for a given value of x and the regression equation of x on y is used for estimating or predicting x for a specified value of y.

SOLVED PROBLEMS

Example 21

Calculate the coefficient of correlation for the following ages of husbands and wives.

Age of husband (x):	23	27	28	29	30	31	33	35	36	39
Age of wife (y):	18	22	23	24	25	26	28	29	30	32

Solution

We have, $\bar{x} = \frac{1}{n}\Sigma x_i = \frac{311}{10} = 31.1$

$$\bar{y} = \frac{1}{n}\Sigma y_i = \frac{257}{10} = 25.7$$

We prepare the following table.

X_i	$x_i = X_i - \bar{x}$	x_i^2	Y_i	$y_i = Y_i - \bar{y}$	y_i^2	$x_i\,y_i$
23	8.1	65.61	18	7.7	52.29	62.37
27	4.1	16.81	22	3.7	13.69	15.17
28	3.1	9.61	23	2.7	7.29	8.37
29	2.1	4.41	24	1.7	2.89	3.57
30	1.1	1.21	25	0.7	0.49	0.77
31	0.1	0.01	26	0.3	0.09	0.03
33	1.9	3.61	28	2.3	5.29	4.37
35	3.9	15.21	29	3.3	10.89	12.87
36	4.9	24.01	30	4.3	18.49	12.07
39	7.9	62.41	32	6.3	39.69	49.77
		202.90			158.10	178.30

Now, $r = \frac{\Sigma x_i y_i}{\sqrt{\Sigma x_i^2}\sqrt{\Sigma y_i^2}} = \frac{178.30}{\sqrt{202.90} \times \sqrt{158.10}} = \mathbf{0.9955}$

Example 22

Calculate the coefficient of correlation for the following data.

x:	6	2	10	4	8
y:	9	11	5	8	7

Solution

Here we prepare the following table

X	Y	X^2	Y^2	XY
6	9	36	81	54
2	11	4	121	22
10	5	100	25	50
4	8	16	64	32
8	7	64	49	56
30	40	220	340	214

$$r = \frac{n\,\Sigma XY - (\Sigma X)(\Sigma Y)}{\sqrt{n\,\Sigma X^2 - (\Sigma X)^2}\sqrt{n\,\Sigma Y^2 - (\Sigma Y)^2}}$$

$$= \frac{5 \times 214 - 30 \times 40}{\sqrt{5 \times 220 - 30^2}\sqrt{5 \times 340 - 40^2}}$$

$$= \frac{-130}{\sqrt{200}\,\sqrt{100}} = \mathbf{0.919}$$

Example 23

Find the correlation coefficient between X and Y given

x:	10	16	13	12	15	17	14
y:	20	33	25	27	26	30	30

Solution

Here we prepare the following table

X	Y	u_i = X14	$v_i = Y - 25$	u_i^2	v_i^2	$u_i v_i$
10	20	4	5	16	25	20
16	33	2	8	4	64	16
13	25	1	0	1	0	0
12	27	2	2	4	4	4
15	26	1	1	1	1	1
17	30	3	5	9	25	15
14	30	0	5	0	25	0
		1	16	35	144	48

$$r_{xy} = r_{uv} = \frac{n\,\Sigma u_i v_i - (\Sigma u_i)(\Sigma v_i)}{\sqrt{n\,\Sigma u_i^2 - (\Sigma u_i)^2}\sqrt{n\,\Sigma v_i^2 - (\Sigma v_i)^2}}$$

$$= \frac{7 \times 48 - (-1) \times 16}{\sqrt{7 \times 35 - (-1)^2}\sqrt{7 \times 144 - 16^2}}$$

$$= \frac{336 + 16}{\sqrt{245 - 1}\sqrt{1008 - 256}}$$

$$= \frac{352}{\sqrt{244}\sqrt{752}} = \mathbf{0.82}$$

Example 24

Calculate the rank correlation coefficient from the following data specifying the ranks of 7 students in two subjects.

Rank in the first subject :	1	2	3	4	5	6	7
Rank in the second subject :	4	3	1	2	6	5	7

Solution

Here n = 7. Let x and y denote respectively the ranks in the first and second subjects. We prepare the following table.

x_i	y_i	$d_i = x_i - y_i$	d_i^2
1	4	3	9
2	3	1	1
3	1	2	4
4	2	2	4
5	6	1	1
6	5	1	1
7	7	0	0
			20

The Spearman’s rank correlation coefficient is

$$R = 1 - \frac{6\,\Sigma\, d_i^2}{n\,(n^2 - 1)} = 1 - \frac{6 \times 20}{7 \times (7^2 - 1)} = 0.643$$

Example 25

Find the rank correlation coefficient between marks in two subjects A and B scored by 10 students

A:	88	72	95	60	35	46	52	58	30	67
B:	65	90	86	72	30	54	38	43	48	75

Solution

The following table is prepared.

A	B	Ranks in A	Ranks in B	d_i	d_i^2
88	65	2	5	3	9
72	90	3	1	2	4
95	86	1	2	1	1
60	72	5	4	1	1
35	30	9	10	1	1
46	54	8	6	2	4
52	38	7	9	2	4
58	43	6	8	2	4
30	48	10	7	3	9
67	75	4	3	1	1

$$= 1 - \frac{6\,\Sigma\, d_i^2}{n\,(n^2 - 1)}$$

$$= 1 - \frac{6 \times 38}{10 \times (10^2 - 1)}$$

$$= 1 \quad 0.2303 = \mathbf{0.7697}$$

Example 26

The coefficient of rank correlation of marks obtained by 10 students in two subjects was computed as 0.5. It was later discovered that the difference in marks in two subjects obtained by one of the students was wrongly taken as 3 instead of 7. Find the correct coefficient of rank correlation.

Solution

Here given R = 0.5, and n = 10.

Then we have,

0.5 =

or = 82.5

Deleting the wrong item from this and adding the correct item to it we obtain corrected

$$82.5 \;\; 3^2 + 72 = 122.5.$$

Consequently, the correct coefficient of rank correlation is

R = = **0.2576**

Example 27

The following are the data on the average height of the plants and weight of yield per plot recorded from 10 plots of rice crop.

Height (X) (cms)	:	28	26	32	31	37	29	36	34	39	40
Yield (Y) (kg)	:	75	74	82	81	90	80	88	85	92	95

Find (i) correlation coefficient between X and Y (ii) the regression coefficient and hence write down regression equation of y on x and that of x on y (iii) probable value of the yield of a plot having an average plant height of 98 cms.

Solution

Here we prepare the following table.

X	Y	$u_i = X - 34$	$v_i = Y - 80$	u_i^2	v_i^2	$u_i v_i$
28	75	6	5	36	25	30
26	74	8	6	64	36	48
32	82	2	2	4	4	4
31	81	3	1	9	1	3
37	90	4	10	16	100	40
29	80	5	0	25	0	0
36	88	2	8	4	64	16
34	85	0	5	0	25	0
39	92	5	12	25	144	60
40	95	6	15	36	225	90
		7	42	219	624	277

i. $$r_{xy} = r_{uv} = \frac{n\sum u_i v_i - (\sum u_i)(\sum v_i)}{\sqrt{n\sum u_i^2 - (\sum u_i)^2}\sqrt{n\sum v_i^2 - (\sum v_i)^2}}$$

$$\frac{10 \times 277 - (-7) \times 42}{\sqrt{10 \times 219 - (-7)^2}\sqrt{10 \times 624 - (42)^2}}$$

$$\frac{3064}{46.271 \times 66.903} = \mathbf{0.989}$$

The regression coefficient of y on x is

$$b_{yx} = \frac{n\sum u_i v_i - (\sum u_i \sum v_i)}{n\sum u_i^2 - (\sum u_i)^2}$$

$$\frac{3064}{2140.99} = 1.431$$

The regression coefficient of x on y is

$$b_{xy} = \frac{n \sum u_i v_i - (\sum u_i \sum v_i)}{n \sum v_i^2 - (\sum v_i)^2}$$

$$= \frac{3064}{4476.01} = 0.684$$

The regression equation of y on x is $\bar{\ } = A + \frac{\sum u_i}{n}$

$$y - \bar{y} = b_{yx}(x - \bar{x}) \qquad 34 + \frac{-7}{10} = 33.3$$

$$- = B + \frac{\sum v_i}{n}$$

ie., $y - 84.2 = 1.431\,(x - 33.3)$

ie., **y = 1.431x 36.55** $\qquad 80 + \frac{42}{10} = 84.2$

The regression equation of x on y is

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

i e . , $x - 33.3 = 0.684\,(y - 84.2)$

ie., **x = 0.684y 24.29**

iii.To estimate the yield (y), the regression equation of y on x is

y = 1.431x 36.55

when x = 98, y = $1.431 \times 98 - 36.55$ = **103.69kg**

Example 27

For the regression lines $4x - 5y + 33 = 0$ and $20x - 9y = 107$, find the mean values of x and y, (b) the coefficient of correlation between x and y, and (c) the variance of y given that the variance of x is 9.

Solution

Since the lines of regression pass through $(\bar{x}, \bar{y})$ we have

$$4\bar{x} - 5\bar{y} + 33 = 0$$

$$20\bar{x} - 9\bar{y} - 107 = 0$$

Solving these equations, we get the mean values of x and y as $\bar{x} = 13$, $\bar{y} = 17$. We rewrite the given equations respectively as

$y = \frac{4}{5}x + \frac{33}{5}$, $x = \frac{9}{20}y + \frac{107}{20}$ so that $b_{yx} = \frac{4}{5}$, $b_{xy} = \frac{9}{20}$

Therefore, the coefficient of correlation between x and y is

$$= \sqrt{b_{xy}\ b_{yx}} = \mathbf{0.6}$$

Here positive sign is taken since both b_{xy} and b_{yx} are positive.

Since $r\frac{\sigma_y}{\sigma_x} = b_{yx} = \frac{4}{5}$, and $\sigma_x^2 = 9$ (given), we get

$$\sigma_y = \frac{4\sigma_x}{5r} = \frac{4 \times 3}{5 \times 0.6} = 4$$

Thus, the variance of y is σ_y^2 = **16**.

EXERCISES

Multiple Choice Questions

The idea of product moment correlation was given by

a) R.A. Fisher b) Sir. Francis Galton

c) Karl Pearson d) Spearman

Correlation coefficient was invented in this year

a) 1910 b) 1890

c) 1908 d) None of the above

The unit of correlation coefficient is

a) kg/cc b) percent

c) non existing d) none of the above

If , the variables X and Y are

a) linearly related b) independent

c) not linearly related d) none of the above

In a scatter diagram if all dots lie on a line falling from left hand top to right hand bottom, then the value of r is

a) +1 b) 0

c) -1 d) 1

The formula for rank correlation coefficient is

a) $1+\frac{6\sum d^2}{n\left(n^2-1\right)}$ b) $1-\frac{6\sum d^2}{n\left(n^2-1\right)}$

c) $1-\frac{\sum d^2}{n\left(n^2-1\right)}$ d) None of the above

Rank correlation coefficient was discovered by

a) Charles Spearman b) Karl Pearson

c) R.A. Fisher d) Francis Galton

In a regression line of y on x, the variable x is known as

a) independent variable b) regressor

c) explanatory variable d) all the above

If b_{yx} and b_{xy} are two regression coefficients, they have

a) same sign b) opposite sign

c) either same or opposite signs d) nothing can be said

If $r_{xy} = -1$, the relation between X and Y is of the type

When Y increases, X also increases

When Y decreases, X also decreases

X is equal to $-Y$

When Y increases, X proportionately decreases.

The correlation between two variables is of order

a) 2 b) 1

c) 0 d) none of the above

If simple correlation coefficient is zero, then the regression coefficient is equal to ....

a. 1 b. 2

c . 0 d. -1

Correlation coefficient is a .... number.

a. imaginary b. unit based

c . pure d. None of these

If $b_{yx} > 1$, then b_{xy} is

a) less than 1 b) greater than 1

c) equal to 1 d) equal to 0

15. Given the regression lines $X + 2Y - 5 = 0$, $2X + 3Y - 8 = 0$ and $\sigma_x^2 = 12$, the value of σ_y^2 is

a) 16 b) 4 c) 3/4 d) 4/3

Very Short Answer Questions

What is Correlation?

Enumerate the different types of Correlation.

What is meant by perfect correlation?

What is meant by spurious correlation?

Give the formula for product moment correlation coefficient.

21. Give the significance of the values r = +1, r = 1 and r = 0.

What is the use of scatter diagram?

What are advantages of rank correlation coefficient?

Why there are two regression lines?

What are regression coefficients.

Short Essay Questions

Distinguish between Correlation and Regression

Define the coefficient of correlation and show that it is free from origin and the unit of measurement.

Explain how coefficient of correlation measures the linear relationship between two variables.

Define (1) Line of regression and (2) Regression coefficient. Show that the coefficient of correlation is the geometric mean of coefficients of regression.

State the important properties of regression coefficient. Prove any one of these properties.

31. What are regression lines? Why there are two regression lines?

32. What is correlation? Enunciate the different types of correlation between two variables.

Long Essay Questions

Compute the coefficient of correlation between X and Y presented in the table below:

X	:	1	3	4	6	8	9	11	14
Y	:	1	2	4	4	5	7	8	9

Find the correlation coefficient between x and y given the following sets of values of x and y:-

x	:	1	2	4	5	8	9
y	:	4	6	7	10	11	15

From the following information, obtain the correlation coefficient: -

N =12; $\sum x = 30$; $\sum y = 5$;

$\sum x^2 = 670$; $\sum y^2 = 285$; $\sum xy = 334$.

For the following pairs of values, obtain the correlation coefficient:-

X	:	4	6	5	9	6	11	8
Y	:	6	14	10	17	12	18	14

37.. Calculate the coefficient of correlation for the following data:

x :	28	45	40	38	35	33	40	32	36	33
y :	23	34	33	34	30	26	28	31	36	35

MODULE III PROBABILITYTHEO RY

CLASSICAL DEFINITION OF PROBABILITY

Introduction

In everyday language, the word probability describes events that do not occur with certainty. When we look at the world around us, we have to conclude that our world functions more on uncertainty than on certainty. Thus we speak of the probability of rain tomorrow, the probability that an electric appliance will be defective, or even the probability of nuclear war. The concept of probability has been an object of debate among philosophers, logicians, mathematicians, statisticians, physicists and psychologists for the last couple of centuries and this debate is not likely to be over in the foreseeable future.

Probability is a number associated with an event, intented to represent its ‘likelihood’, ‘chance of occurring’, ‘degree of uncertainity’ and so on. The probability theory has its origin in ‘*Games of chance*’. Now it has become a fundamental tool of scientific thinking.

Classical Definition of Probability

Some Important Concepts

1. Random experiment

It is a physical phenomenon and at its completion we observe certain results. There are some experiments, called deterministic experiments, whose outcomes can be predicted. But in some cases, we can never predict the outcome before the experiment is performed. An experiment natural, conceptual, physical or hypothetical is called a random experiment if the exact outcome of the trails of the experiment is unpredictable. In other words by a random experiment, we mean

It should be repeatable under uniform conditions.

It should have several possible outcomes.

One should not predict the outcome of a particular trail.

Example: Tossing a coin, rolling a die, life time of a machine, length of tables, weight of a new born baby, weather condition of a certain region etc.

2. Trial and Event

Trial is an attempt to produce an outcome of a random experiment. For example, if we toss a coin or throw a die, we are performing trails.

The outcomes in an experiment are termed as events or cases. For example, getting a head or a tail in tossing a coin is an event. Usually events are denoted by capital letters like A, B, C, etc…

3. Equally likely events

Events or cases are said to be equally likely when we have no reason to expect one rather than the other.

For example, in tossing an unbiased coin the two events head and tail are equally likely because we have no reason to expect head rather than tail. Similarly, when we throw a die the occurrence of the numbers 1 or 2 or 3 or 4 or 5 or 6 are equally likely events.

4. Exhaustive events

The set of all possible outcomes in a trial constitutes the set of exhaustive cases. In other words the totality of all possible outcomes of a random experiment will form the exhaustive cases. For example, in the case of tossing a coin there are two exhaustive cases head or tail. In throwing a die there are six exhaustive cases since any one of the six faces 1, 2, …, 6 may come upper most. In the random experiment of throwing two dice the number of exhaustive cases is $6^2 = 36$. In general, in throwing n dice, the exhaustive number of cases is 6^n.

5. Mutually exclusive events

Events are said to mutually exclusive or incompatible or disjoint if the happening of any one of them precludes or excludes the happening of all the others in a trail. That is, if no two or more of them can happen simultaneously in the same trial.

For example, the events of turning a head or a tail in tossing a coin are mutually exclusive. In throwing a die all the six faces numbered 1 to 6 are mutually exclusive since if any one of these faces comes, the possibility of others in the same trial, is ruled out.

6. Favourable cases

The cases which entail the occurrence of an event are said to be favourable to the events. For example, while throwing a die, the occurrence of 2 or 4 or 6 are the favourable events which entail the occurrence of an even number.

Classical Definition (Mathematical or ‘a priori’)

Classical definition is the oldest and simplest definition of probability. This is sometimes called equally-likely events approach. It is also known by the name Laplace definition. From a practical point of view it is the most useful definition of probability.

Definition

If a trial results in ‘n’ mutually exclusive, equally likely and exhaustive cases and ‘m’ of them are favourable (m < n) to the happening of an event A, then the probability of A, designated as P(A) is defined as

$$P(A)= \frac{m}{n} = \frac{\text{no of favourable cases}}{\text{Total number of cases}} \tag{1}$$

Obviously, 0 £ P(A) £ 1

Note 1

If A is an impossible event, then P(A) = 0
If A a sure event, then P(A) = 1
If A is a random event, then 0 < P(A) < 1

Note 2

We can represent the probability given by (1) by saying that the odds in favour of A are m: (n – m) or the odds against A are (n – m): n.

Limitations of classical definition

The above definition of mathematical probability fails in the following cases.

In the classical or a priori definition of probability only equally likely cases are taken into consideration. If the events cannot be considered equally likely classical definition fails to give a good account of the concept of probability.

When the total number of possible outcomes ‘n’ become infinite or countably infinite, this definition fails to give a measure for probability.

If we are deviating from the games of chances like tossing a coin, throwing a die etc., this definition cannot be applied.

Another limitation is that it does not contribute much to the growth of the probability theory.

Frequency Definition of Probability

Let the trials be repeated over a large number of times under essentially homogeneous conditions. The limit of the ratio of the number of times an event A happens (m) to the total number of trials (n), as the number of trials tends to infinity is called the probability of the event A. It is, however, assumed that the limit is unique as well as finite.

Symbolically, $$P(A) = \lim_{n\to\infty} \frac{m}{n}$$

Remark 1. The application of this definition to many problems cannot be extensive since n is usually finite and the limit of the ratio cannot normally be taken as it leads to mathematical difficulties. Besides, the definition of probability thus proposed by Von Mises would involve a mixture of empirical and theoretical concepts, which is usually avoided in modern axiomatic approach.

Remark 2. The two definitions of probability” are apparently different. The mathematical definition is the relative frequency of favourable cases to the total number of cases while in the statistical definition it is the limit of the relative frequency of the happening of the event.

Set Theory

Set: A Set is a collection of well defined objects. The following arc typical examples of sets.

The students Minu, Jithu, Hari and Devi.

The odd numbers 1, 3. 5, 7, 9

The rivers in South India

The metropolitan cities of India

Sets are usually denoted by capital letters A, B, C, X, Y. Z etc. The items which are included in a set are called the *elements* of the set.

If A = {3, 5, 12, 14, 21} is a set, then ‘3’ is an element of set A, and it is written as ‘3 ∈ A’. This is read as ‘element 3 belongs to set A. Thus the symbol ‘∈’ denotes ‘belongs to’. On the other hand, 8 does not belong to set A in the above case. Then the symbol ‘∉’ is used to indicate ‘does not belong to’. ie., 8 ∉ A implies element 8 is not a member of set A.

There are two methods of representing sets viz.

Roster method. 2. Rule method

I. Roster Method

Here each and every element of the set is listed or mentioned.

Example: i. A = {a. e, i, o, u}   ii. B ={2, 3, 5, 7}

iii. Y = {6, 1, 5, 2, 4, 3}

Note

The flower brackets { } are used for denoting a set. The order in which the elements of a set are listed in the { } brackets is immaterial.

2. Rule Method

Here a rule is stated by which all the elements of the set intended to be identified.

Example: A: {x/x is a vowel among English alphabets}

This is read as set A. Set of all x such that x is a vowel among English alphabets.

Types of Sets

We have different kinds of sets, Consider the following

I. Finite Set

A Set which contains a finite or a fixed number of elements is called a ‘Finite Set’. Example:

Set A has only five elements i,e,, A = {1, 2. 6. 8. 10}

B = {x/x is a composite number between 12 and 18}

i.e., B = {14, 15, 16}

Y = {x/x shows a number on a die}

This is same as Y = {1, 2, 3, 4, 5, 6}

Infinite Set

A set which contains infinite number of elements is called an 'infinite set'.

Example:

X = {x/x is a natural number} i.e.. X = {1, 2, 3. 4...}

Y = {…, –2, –1, 0, +1, +2. …}

3. Singleton Set

A set containing only one element is called a 'Singleton Set'.

Example

A = {0}

B = {x/x is an even number between 3 and 51 i.e., B = {4}

4. Null Set

A set which does not contain any element is called an "empty set or 'void set' or 'Null Set'.

Example:

Set A denotes names of boys in a girls college.

i.e., A = { } since nobody is admitted to a girl's college.

T = {x/x is a perfect square between 10 and 15}

i.e.. T = { } since no number which is a perfect square exists between 10 and 15.

A null a set is denoted by the greek letter ϕ (read as phi)

Example:

= ϕ implies 'set T is a null Set'.

But T = {ϕ} implies 'set T is a singleton set with ϕ as an element'

5. Universal Set

A Universal Set is a set of all elements which are taken into consideration in a discussion. It is usually denoted by the capital letter U or otherwise defined in the context. In this text we shall use S to indicate a universal set since it is more convenient for application to probability.

For instance. Let S = {1, 2, 3, 4. 5, 6} be a universal set, showing possible numbers on a die.

6. Sub-sets and Super-sets

Let A and B be two sets. If every element of B is present in A, then B is a 'Sub Set' of A. ie., $B \subset A$. In other words. A is a 'Super Set' of B i.e., $A \supset B$

Example:

i. If A {a. b, c, d, e} and B {a, d}

then $B \subset A$ or $A \supset B$

If A = {2, 4} and B = {I, 2, 3, 4, 5}

then $A \subset B$ or $B \supset A$

Equal Sets

Two sets A and B are said to be equal if $A \subset B$ and $B \supset A$ and is denoted by A = B

Example:

i. Let A = {3, 2. 5. 6} and B = {2, 5. 6. 3}

Here all the elements of A are elements of B{ie. $A \subset B$} and all the elements of B are elements of A(ie., $B \subset A$). Hence

A = B

8. Equivalent Sets

Two sets A and B are said to be equivalent if they have equal number of elements and is denoted by $A \equiv B$ For example

Let A = {X, Y, Z} and B = {1, 2, 3} Then A and B are said to be equivalent sets and are denoted by A ≡ B

9. Power Set

The power set is defined as the collection of all subsets of a given set. It is also called Master set. Example:

The powerset of a given set {a, b, c} is {ϕ, {a}, {b}, {c}, {a,b}, {b, c}, {c, a}, {a, b, c}}.

The number of elements in a set is called *cardinality* of the set, Thus the cardinality of powerset of a given set having 3 elements is $2^{3^{\cdot}}$ Generally the cardinality of power-set of given set having n elements is 2^n.

Venn Diagrams

Sets can be represented diagrammatically using Venn diagrams. These were introduced by John Venn, an English logician.

Here, the Universal set is represented by a rectangle and all other sub-sets by circles or triangles etc. Venn diagrams are especially useful for representing various set operations. Hence we first learn about set operations and employ Venn diagrammatic approach to represent the same.

Set Operations

The basic set operations are (i) union (ii) intersection [iii) compliment and (iv) difference.

i. Union of sets

If A and B are two sets, then the 'union' of sets A and B is the set of all elements which belong to either A or B or both (i.e., which belongs to at least one). It is denoted by A ∪ B.

That is, **x ∈ A ∪ B implies x ∈ A or x ∈ B**

Example:

If A = {3, 8, 5} and B = {3, 6, 8}

then A ∪ B = {3, 8, 5} ∪ {3, 6, 8} = {3, 8, 5, 6}

2. Intersection of Sets

If A and B are two sets, then the 'intersection" of A and B is the set of all elements which are common to both of them. Intersection of sets A and B is denoted by A ∩ B

That is, **x ∈ A ∩ B implies x ∈ A and x ∈ B**

Example:

If A = {2, 5} and B = {5, 7, 9}

then A ∩ B = {2, 5} ∩ {5, 7, 9} = {5}

Disjoint Sets

Two sets are said to be 'disjoint' or 'mutually exclusive' if they do not have any common element between them

(A ∩ B) = ϕ or (A ∩ B) = { }, a null set

Example:

If A = {1, 2} and B = {a, b, c}. (A ∩ B) = ϕ

3. Difference of Sets

If A and B are two sets, A – B is the difference of two sets A and B which contains all elements which belong to A but not to B,

That is x ∈ A–B implies x ∈ A and x ∉ B

Example:

If A = {0, 1, 2, 3} and B = {2. 3. 5, 7} then

A–B = {0, 1}

4. Complement of Sets

Suppose A is a sub set of some Universal set S. Its complementary set is the set of all elements of the Universal set S which does not belong to the set A. The complementary set of A is denoted by A′ {A dash) or A ($\bar{A}$ bar) or A^C (*A* complement).

That is x ∈ A^C implies x ∉ A but x ∈ S

Note

A complement set A cannot be developed without the elements of the Universal set S being known.

Example

If $S = \{1, 2, 3, 4, 5\}$ and $A = \{2, 4\}$

then $A^C = \{1, 3, 5\}$

Algebra of Sets

The following results are very useful is the context of probability theory. We can easily verify the results by choosing the sets appropriately.

$A \cup B = B \cup A$, $A \cap B = B \cap A$ - Commutative property

$(A \cup B) \cup C = A \cup (B \cup C)$, $(A \cap B) \cap C = A \cap (B \cap C)$

Associative property

$(A \cup B) \cap C = (A \cap C) \cup (B \cap C)$

$(A \cap B) \cup C = (A \cup C) \cap (B \cup C)$ – Distributive property

$(A \cup S) = S$, $A \cap S = A$

$A \cup \phi = A$, $A \cap \phi = \phi$

$A \cup A^C = S$, $A \cap A^C = \phi$

$(A \cup B)^C = A^C \cap B^C$

$(A \cap B)^C = A^C \cup B^C$

More generally,

$$\left(\bigcup_{i=1}^{n} A_i\right)^C = \bigcap_{i=1}^{n} A^C$$

$$\left(\bigcap_{i=1}^{n} A_i\right)^C = \bigcup_{i=1}^{n} A^C$$ – De’ Morgan’s Laws.

Set terminology

The following terminologies are verbally used for calculating the probability of occurrence of events where the events are represented by sets.

1. For one event A

i. Occurrence of an event is represented by – A

ii. No occurrence of an event – A_C

2. For two events A and B

i. Occurrence of none – $A^C \cap B^C$

Occurrence of both A and B – $A \cap B$

Occurrence of exactly one – $(A \cap B_C) \cup (A_C \cap B)$

iv. Occurrence of at least one – $A \cup B$

3. For three events A, B, and C

i. Occurrence of all – $A \cap B \cap C$

ii. Occurrence of None – $A^C \cap B^C \cap C^C$

Occurrence exactly of one

– $(A \cap B^C \cap C^C) \cup (A^C \cap B \cap C^C)(A^C \cap B^C \cap C)$

Occurrence of exactly two

– $(A \cap B \cap C_C)(A \cap B_C \cap C) \cup (A_C \cap B \cap C)$

v. Occurrence of at least one – $A \cup B \cup C$

Permutations and Combinations

Fundamental Principle:

If an event ‘A’ can happen in ‘n ’ ways and another event ‘B’ can happen in ‘n_2’ ways, then the number [1]of ways in which both the events A and B can happen in a specified order is ‘$n_1 \times n_2$’.

If there are three routes from X to Y: two routes from Y to Z then the destination Z can be reached from X in 3 × 2 = 6 ways.

Permutation

Definition: Permutation refers to the *arrangement* which can be made by taking some (say r) of things at a time or all of ‘n’ things at a time with attention given to the order of arrangement of the selected objects.

Mathematicians use a neat notation for permutation (i.e., arrangement) of ‘n’ objects taking ‘r’ objects at a time by writing this statement as ${}_nP_r$ or nPr. Here, letter ‘P’ stands for ‘permutation’ (i.e., a rule for arrangement).

Suppose we want to arrange 3 students A, B and C by choosing 2 of them at a time. This arrangement can be done in the following ways.

AB, BC, CA, BA, CB and AC

The arrangement of 3 things taken 2 at a time is denoted by $3P_2$.

Therefore, $3P_2 = 6 = 3 \times 2$.

In general, suppose there are ‘n’ objects to be permuted in a row taking all at a time. This can be done in ${}^nP_n$ different ways. It is given by

$${}_nP_n = n\,(n-1)\,(n-2)\ldots 3.\,2.\,1$$

Example

$$4P_4 = 4.3.2.1 = 24$$

The permutation of n things taken r at a time (r < n) is given by

$${}_nP_r = n\,(n-1)\ldots(n-r+1)$$

eg: $7P_5 = 7\times6\times5\times4\times3 = 2520$

Factorial notation

We have a compact notation for the full expression given by the product n (n –1) (n – 2) … 3. 2. 1. This is written as n! read as ‘n factorial’.

So, $${}^nP_n = n! = n\,(n-1)\,(n-2)\ldots 3.\,2.\,1.$$

$$6P_6 = 6! = 6.\,5.\,4.\,3.\,2.\,1 = 720$$

By, definition, 0! =1

We have, ${}^nPr = n\,(n-1)\ldots(n-r+1)$

$$= n(n-1)\,(n-2)\ldots(n-r+1)\left[\frac{(n-r)(n-r-1)\ldots3.2.1}{(n-r)(n-r-1)\ldots3.2.1}\right]$$

$$np_r = \frac{n!}{(n-r)!}$$

Results

The number of permutations of n objects when r objects taken at a time when repetition allowed = n^r.

The number of permutations of n objects when all the n taken at a time when repetition allowed = n^n.

The number of permutations of n objects of which, n_1 are of one kind, n_2 are of another kind, n_3 are of another kind etc., taking all the ntogetherisgivenby $\dfrac{n!}{n_1!n_2!n_3!\ldots n_k!}$ where $n_1 + n_2 + \ldots + n_k = n$.

Combination

A combination is a grouping or a selection or a collection of all or a part of a given number of things without reference to their order of arrangement.

If three letters, a, b, c are given, ab, bc, ca are the only combinations of the three things a, b, c taken two at a time and it is denoted as $3C_2$. The other permutations ba, cb and ac are not new combinations. They are obtained by permuting each combination among themselves.

So $3P_2 = 3C_2 \times 2!$

or $$3C_2 = \frac{3P_2}{2!} = \frac{3.2}{1.2} = 3$$

Combination of n different things taken r at time (r < n)

The number of combinations of n different things taken r at a time is denoted as nCr or nC_r or $\binom{n}{r}$. It is given by

$$nC_r = \frac{nP_r}{r!} = \frac{n(n-1)(n-2)\ldots(n-r+1)}{1\cdot 2\cdot 3\quad r}$$

or $$nC_r = \frac{n!}{r!(n-r)!}$$

For example, $^{7}C_{3} = \frac{7!}{3!\,4!} = \frac{7\cdot 6\cdot 5}{1\cdot 2\cdot 3} = 35$

$$^{10}C_{4} = \frac{10!}{4!\,6!} = \frac{10\cdot 9\cdot 8\cdot 7}{1\cdot 2\cdot 3\cdot 4} = 210$$

Important results

$nC_n = \frac{n!}{n!\;0!} = 1$. This is the combination of n things taken all at a time.

$nC_0 = \frac{n!}{n!\;0!} = 1$. This is the combination of n things taken none at a time.

$nC_r = nC_{n-r}$

This says that, $10C_8 = 10C_2 = \frac{10\cdot 9}{1\cdot 2}$

$12C_9 = 12\,C_3 = \frac{12\cdot 11\cdot 10}{1\cdot 2\cdot 3} = 220$

$100C_{98} = 100\,C_2 = \frac{100\cdot 99}{1\cdot 2} = 4950$

4. $nC_r + nC_{r-1} = (n+1)\,C_r$.

SOLVED PROBLEMS

Example 1

What is the probability that a leap year selected at random will contain 53 Sundays?

Solution

In a leap year there are 366 days consisting of 52 weeks plus 2 more days. The following are the possible combinations for these two days. (i) Sunday and Monday (ii) Monday and Tuesday (iii) Tuesday and Wednesday Wednesday and Thursday (v) Thursday and Friday (vi) Friday and Saturday (vii) Saturday and Sunday.

For getting 53 Sundays in a leap year, out of the two days so obtained one should be a Sunday. There are two cases favourable for getting a Sunday out of the 7 cases.

Required probability = **2/7.**

Example 2

Three coins are tossed. What is the probability of getting (i) all heads exactly one head (iii) exactly two heads (iv) atleast one head (v) atleast two heads (vi) at most one head (vii) at most two heads (viii) No head.

Solution

When three coins are tossed, the possible outcomes are given by [HHH, HHT, HTH, THH, HTT, THT, TTH, TTT]

i. P (all heads) = 1/8

ii. P (exactly one head) = 3/8

iii. P (exactly two heads) = 3/8

iv. P (atleast one head) = 7/8

v. P (atleast two heads) = 4/8

vi. P (at most one head) = 4/8

vii. P (at most two heads) = 7/8

viii. P (no head) = 1/8

Example 3

What is the probability of getting a spade or an ace from a pack or cards?

Solution

P (Spade or Ace) = **16/52**

Example 4

A box contains 8 red, 3 white and 9 blue balls. If 3 balls are drawn at random, determine the probability that (a) all three are blue (b) 2 are red and 1 is white (c) atleast one is white and (d) one of each colour is drawn.

Solution

Assume that the balls are dreawn from the urn one by one without replacement.

P(all the three are blue) $= \frac{9C_3}{20C_3} = \frac{7}{95}$

P(2 red and 1 white) $= \frac{8C_2 \times 3C_1}{20C_3} = \frac{4}{95}$

P(at least 1 is white) = 1 – P (None is white)

$$= 1 - \frac{17C_3}{20C_3}$$

$$= 1 - \frac{34}{57} = \frac{23}{57}$$

d) P (one of each colour) $= \frac{8C_1 \times 3C_1 \times 9C_1}{20C_3} = \frac{18}{95}$

Example 5

What is the probability of getting 9 cards of the same suit in one hand at a game of bridge?

Solution

One hand in a game of bridge consists of 13 cards. Total number of possible cases = $52C_{13}$

The number of ways in which a particular player can have 9 cards of one suit are $13C_9$ and the number of ways in which the remaining 4 cards are of some other suit are $39C_4$. Since there are 4 suits in a pack of cards, the total number of favourable cases = 4 × $13C_9 \times 39C_4$.

Required probability = $\frac{4 \times 13C_9 \times 39C_4}{52C_{13}}$

AXIOMATIC DEFINITION OF PROBABILITY

The mathematical and statistical definitions of probability have their own disadvantages. So they do not contribute much to the growth of the probability theory. The axiomatic definition is due to A.N. Kolmogorov (1933), a Russian mathematician, and is mathematically the best definition of probability since it eliminates most of the difficulties that are encountered in using other definitions. This axiomatic approach is based on measure theory. Here we introduce it by means of set operations.

Sample space

A sample space is the set of all conceivable outcome of a random experiment. The sample space is usually denoted by S or W. The notion of a sample space comes from Richard Von Mises.

Every indecomposable outcome of a random experiment is known as a *sample point or elementary outcome.* The number of sample points in the sample space may be finite, countably infinite or noncountably infinite. Sample space with finite or countably infinite number of elements is called discrete sample space. Sample space with continuum of points is called continuous sample space.

Example

The sample space obtained in the throw of a single die is a finite sample space, ie. S = {1, 2, 3, 4. 5, 6}

The sample space obtained in connection with the random experiment of tossing a coin again and again until a head appears is a countably infinite sample space.

ie. S {H, TH, TTH, TTTH .... }

Consider the life time of a machine. The outcomes of this experiment form a continuous sample space.

ie., $S = \{ t : 0 < t < \infty\}$

Event

An event is a subset of the sample space. In other words, “of all the possible outcomes in the sample space of an experiment, some outcomes satisfy a specified description, which we call an event.”

Field of events (F)

Let S be the sample space of a random experiment. Then the collection or class of sets F is called a field or algebra if it satisfies the following conditions.

F is nonempty

the elements of F are subsets of S.

if $A \in F$, then $A^C \in F$

if $A \in F$ and $B \in F$ then $A \cup B \in F$

For example, let S = { 1, 2, 3, 4, 5, 6 }

Choose F as the set with elements ϕ, S, { 5, 6} and {1, 2, 3,4) Then F satisfies all the four conditions. So F is a field.

More generally, when $A \subset S$, $F = \{\phi, A, A^{C'} S\}$ forms a field. Trivially, F with just two elements ϕ and S forms a field.

σ-’field or σ-algebra of events

Let S be a nonempty set and F be a collection of subsets of S. Then F is called a σ-field or σ-algebra if

F is nonempty

Tile elements of F are subsets of S

If $A \in F$,then $A^C \in F$ and

The union of any countable collection of elements of F is an element of F.

i.e., if $A_i \in F$, $i = 1, 2, 3, \dots n$, then $\bigcup_{i=1}^{\infty} A_i \in F$

The σ algebra F is also called *Borel field* and is often denoted by B.

Examples

B = {ϕ, S}

B = {0, A, A^C, S}

B = { ϕ, A, B, S} provided A ∪ B = S and A ∩ B = ϕ

The powerset of S always form a Borel field.

Function and Measure

We know that a function or mapping is a correspondence between the elements of the set X (called domain) and the set Y (called range) by a rule or principle. When the elements of the domain are sets and the elements of the range are real numbers, the function is said to be a 'set function'. A set function is usually denoted by P(A) or μ(A) where A represents an arbitrary set in the domain.

In a set function if $A_1, A_2, \dots, A_n$ are disjoint sets in the domain and if

$\mu(A_1 \cup A_2 \cup A_3 \cup \dots \cup A_n) = \mu(A_1) \cup \mu(A_2) \cup \mu(A_3) \cup \dots \mu(A_n)$ then the set function is said to be *additive.*

If a set S is partitioned into a countable number of disjoint sets $A_1, A_2, \dots$ and if a set function defined on satisfies the property.

$$\mu(A_1 \cup A_2 \cup \dots) = \mu(A_1) + \mu(A_2) + \dots$$

i.e.,
$$\mu\left(\bigcup_{1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mu(A_i)$$

then the set function is said to be *countably additive.*

Measure

A set function which is non negative and totally additive is called a measure. A *measure* will be called a probability measure if

$$\mu(A_1 \cup A_2 \cup A_3 \cup \dots A_n) = \mu(A_1) + \mu(A_2) + \mu(A_3) + \dots + \mu(A_n) = 1$$

where $\bigcup_1 Ai$ = S, Ai ∩ Aj = ϕ, i ≠j

In probability theory, the probability measure is denoted by P instead of μ.

Axiomatic definition

Let S be the sample space. Let B be the class of events constituting the Borel field. Then for each A ∈ B, we can find a real valued set function P (A), known as the probability for the occurrence of A if P(A) satisfies the following three axioms,

Axiom 1. **(Non negativity)**

0 £ P(A) £ 1 for each A ∈ B

Axiom 2. **(Norming)**

P(S) = 1

Axiom 3**. (Countable additivity)**

If $A_1, A_2, \dots, A_n$ is a finite or infinite sequence of elements in B such that Ai ∩ Aj = ϕ, i ≠j.

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i)$$

Probability Space

From the axiomatic definition of probability we can conceive of a probability space constituting the triplet (S, B, P) where S represents the sample space, B is the class of all subsets of S constituting a Borel field, and P is the probability function with domain B and satisfying the axioms 1, 2 and 3 of probability given above.

Probability space is a single term that gives us an expedient way to assume the existence of all three components in its notation. The three components are related; B is a collection of subsets of S and P is a function that has B as its domain. The probability space's main rise is in providing a convenient method of stating background assumptions for future definitions and theorems etc.

Note: The axiomatic definition of probability proposed by Kolmogorov reveals that the numbers in the interval [0, 1] can be assigned as probabilities of events in some initial class of elementary events. Using these probabilities we can determine the probability of any event which may be of interest. The calculus of probability begin after the assignment of probabilities represented by the symbols p_1, p_2, p_3 ...... which are usually determined on the basis of some past experience or on the basis of some empirical study.

Theorems in Probability

The following are some consequences of the axioms of probability, which have got general applications and so they are called theorems. We can make use of Venn diagrams for the better understanding of these theorems.

Theorem I

The probability of an impossible event is ZERO.

ie., $P(\phi) = 0$,

Proof

Let ϕ be the impossible event.

Then $S \in B$ and $\phi \in B$

We have $S \cup \phi = S$

$P(S \cup \phi) = P(S)$

i.e., $P(S) + P(\phi) = P(S)$, since S and ϕ are disjoint.

i.e., $1 + P(\phi) = 1$ – by axiom 2

$\therefore \quad P(\phi) = 0$

Note:

The condition $P(A) = 0$ does not imply that $A = \phi$

Example

Consider an experiment of tossing a coin infinitely many times. The outcomes may be represented as infinite sequences of the form HHTHTTTHHT….. so that the sample space S consists of infinitely many such sequences. The event 'head only' given by the sequence {HHHH….} is not empty. However, the chance of such an outcome is, atleast intuitively, zero. Tails should come up sooner or later.

Theorem 2

Probability is infinitely additive

i.e., $P(A_1 \cup A_2 \cup \ldots \cup A_n) = P(A_1) + P(A_2) + P(A_3) + \ldots + P(A_n)$

where $A_i \cap A_j = \phi$, $i \neq j$.

Proof

Consider an infinite sequence of events $A_1, A_2, A_3, \ldots A_n, \phi, \phi, \phi, \ldots$ which are pairwisely disjoint since Ai's are disjoint.

Then by axiom 3

$P(A_1 \cup A_2 \cup \ldots A_n \cup \phi \cup \phi \cup \ldots) =$

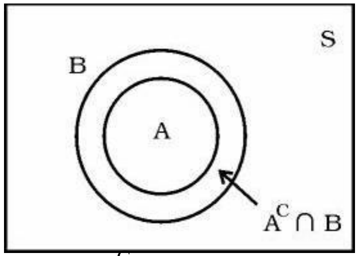
$P(A_1) + P(A_2) + P(A_3) + \ldots + P(A_n) + P(\phi) + P(\phi) \ldots$ i.e.,

$P(A_1 \cup A_2 \cup \ldots \cup A_n) = P(A_1) + P(A_2) + P(A_3) + \ldots + P(A_n)$

$$P\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} P(A_i)$$

Theorem 3 (Monotonicity)

If $A \subset B$, then $P(A) \leq P(B)$



Proof

From the Venn diagram

We have $B = A \cup (A^C \cap B)$

$\therefore P(B) = P[A \cup (A^C \cap B)]$

$= P(A) + P(A^C \cap B)$ since A and $A^C \cap B$ are disjoint

= P(A) + a +ve quantity i.e., P(B) ≥ P(A) or P(A) ≤ P(B)

Note: From the above, we get P(B – A) = P(B) – P(A) since $A_C \cap B$ = B – A

Theorem 4

Probability is countably subadditive. i.e., for every sequence of events A_1, A_2, …

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) \le P(A_1) + P(A_2) + P(A_3) + \ldots$$

Proof

By considering the infinite operations on events, we can write the union of events into union of disjoint events,

$$\text{i.e., } \bigcup_{i=1}^{\infty} A_i = A_1 \cup (A_1^{C} \cap A_2) \cup (A_1^{C} \cap A_2^{C} \cap A_3) \cup \ldots$$

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) \le P(A_1) + P(A_1^{C} \cap A_2) + P(A_1^{C} \cap A_2^{C} \cap A_3) + \ldots$$

$$\le P(A_1) + P(A_2) + P(A_3) + \ldots$$

Since $P(A_1^{C} \cap A_2) \le P(A_2)$ etc., i.e., $P\left(\bigcup_{i=1}^{\infty} A_i\right) \le \sum_{i=1}^{\infty} P(A_i)$

Theorem 5 (Complementation)

$P(A_C) = 1 – P(A)$

Proof

We have $A \cup A^C = S$ $\therefore$ $P(A \cup A^C) = P(S)$

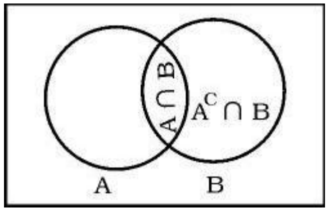
i.e., $P(A) + P(A^C) = 1$, by axiom 2 and 3 $\therefore P(A^C) = 1 – P(A)$

i.e., P [Non occurrence of an event] = 1 – P[Occurrence of that event]

Theorem 6 (Addition theorem of two events)

If A and B are any two events,

$P(A \cup B) = P(A) + P(B) – P(A \cap B)$



Proof

From the Venn diagram,

We can write

$A \cup B = A \cup (A^C \cap B)$

$\therefore P(A \cup B) = P[A \cup (A^C \cap B)]$

$= P(A) + P(A_C \cap B)$ … (1) since $A \cap (A_C \cap B) = \phi$

On the other hand,

$B = (A \cap B) \cup (A^C \cap B)$

$\therefore P(B) = P[(A \cap B) + P(A^C \cap B)]$ since $(A \cap B) \cap (A^C \cap B) = \phi$

$P(A_C \cap B) = P(B) – P(A \cap B)$ .. (2)

On substituting (2) in (1) we get,

$P(A \cup B) = P(A) + P(B) – P(A \cap B)$

Corollary (1) If $A \cap B = \phi$, $P(A \cup B) = P(A) + P(B)$

$P(A \cup B) = 1 – P(A \cup B)_C$

$= 1 – P(A_C \cap B_C)$

i.e., P[the occurrence of atleast one event] = 1 – P[None of them is occurring]

Theorem 7 (Addition theorem for 3 events)

If A, B, C are any three events,

$P(A \cup B \cup C) = P(A) + P(B) + P(C) – P(A \cap B) – P(B \cap C) –$

$P(A \cap C) + P(A \cap B \cap C)$

Proof

Let $B \cup C = D$, Then $P(A \cup B \cup C) = P(A \cup D)$

$= P(A) + P(D) – P(A \cap D)$, by theorem 6

$$P(A) + P(B \cup C) - P[A \cap (B \cup C)]$$

$$P(A) + P(B) + P(C) - P(B \cap C) - P[(A \cap B) \cup (A \cap C)]$$

$$P(A) + P(B) + P(C) - P(B \cap C) - \{P(A \cap B) + P(A \cap C) - P(A \cap B \cap C)\}$$

$$P(A) + P(B) + P(C) - P(A \cap B) - P(B \cap C) - P(A \cap C) + P(A \cap B \cap C)$$

Corollary

If the event A, B, C are mutually exclusive

$$P(A \cup B \cup C) = P(A) + P(B) + P(C)$$

$$P(A \cup B \cup C) = 1 - P(A \cup B \cup C)^C = 1 - P(A^C \cap B^C \cap C^C)$$

Probability in finite sample space with equally likely points

For certain random experiment there is a finite number of outcomes, say n and the probability attached to each outcome is 1/n. The classical definition of probability is generally adopted for these problems. But we can see that the axiomatic definition is applicable as well.

Definition : Let $E_1, E_2, \dots E_n$ be n sample points or simple events in a discrete or finite sample space S. Suppose the set function P with domain the collection of all subsets of S satisfies the following conditions.

(i) $P(E_1) = P(E_2) = P(E_3) = \dots = P(E_n) = \frac{1}{n}$

(ii) $P(S) = P(E_1 \cup E_2 \cup \dots E_n) = P(E_1) + P(E_2) + \dots + P(E_n)$

$$\frac{1}{n} + \frac{1}{n} + \dots \frac{1}{n} \text{ (n terms)} = \frac{n}{n} = 1$$

If A is any event which contains m sample points, say $E_1, E_2, \dots E_m$

then, $P(A) = P(E_1 \cup E_2 \cup \dots E_m)$

$= P(E_1) + P(E_2) + \dots + P(E_m)$, since $Ei \cap Ej = \phi, i \neq j$

$$= \frac{1}{n} + \frac{1}{n} + \dots \frac{1}{n} \text{ (m terms)} = \frac{m}{n}$$

This shows that P(A) satisfies all the axioms of probability. Thus we can see that the classical definition is a particular case of axiomatic definition. In other words, the axiomatic definition can be deduced to classical definition of probability if it is defined on a discrete or finite sample space with equally likely points.

SOLVED PROBLEMS

Example 10

A die is rolled. If x is the number shown on the die. 7x coins are tossed, If y is the number of heads (x , y) is recorded. Write down the sample space of this experiment.

Solution

If x is 1, 7 coins are tossed. If x = 2, 14 coins are tossed and so on. If x = 6, 42 coins are tossed. When y denotes the number of heads obtained, with x = 1, the pair (x, y) takes the values (1, 1) (1 , 2) (1 , 3) … (1, 7}. Thus the required sample space is

S = {(1,1}, (1.2), (1,3)........ (1, 7)}
(2.1), (2,2), (2.3) .... (2, 14)
(3.1), (3.2), (3.3) .... (3, 21)
(4.1), (4.2), (4,3) .... (4,28)
(5.1), (5.2), (5,3) .... (5,35)
(6,1), (6,2), (6.3) .... (6,42)}

Example 11

If A_1, A_2, A_3 are three events which are exhaustive, show that $B_1 = A_1$, $B_2 = A_1^C \cap A_2$, $B_3 = A_1^C \cap A_2^C \cap A_3$ are exhaustive and mutually exclusive.

Solution

Since A_1, A_2, and A_3 are exhaustive, we have

$A_1 \cup A_2 \cup A_3 = S$

We have to show that $B_1 \cup B_2 \cup B_3 = S$, Now $B_1 \cup B_2 \cup B_3 \quad = A_1 \cup (A_1^C \cap A_2) \cup (A_1^C \cap A_2^C \cap A_2)$

$\{(A_1 \cup A_1^C) \cap (A_1 \cup A_2)\} \cup (A_1^C \cap A_2^C \cap A_3)$

$\{S \cap (A_1 \cup A_2)\} \cup (A_1^C \cap A_2^C \cap A_3)$

$(A_1 \cup A_2) \cup \{(A_1 \cup A_2)^C\} \cap A_3$

$(A_1 \cup A_2) \cup \{(A_1 \cup A_2)^C\} \cap (A_1 \cup A_2 \cup A_3)$

$S \cap S = S$

i.e., the events B_1, B_2 and B_3 are exhaustive.

To show that B_1, B_2 and B_3 are mutually exclusive,

$B_1 \cap B_2 \cap B_3 \quad = A_1 \cap (A_1^C \cap A_2) \cap (A_1^C \cap A_2^C \cap A_3)$

$(A_1 \cap A_1^C) \cap A_2 \cap (A_1^C \cap A_2^C \cap A_3)$

$(\phi_1 \cap A_2) \cap (A_1^C \cap A_2^C \cap A_3)$

$\phi \cap (A_1^C \cap A_2^C \cap A_3)$

ϕ $\therefore$ The events B_1, B_2 and B_3 are mutually exclusive.

Example 12

In a swimming race the odds that A will win are 2 to 3 and the odds that B will win are 1 to 4. Find the probability and the odds that A or B wins the race?

Solution

We have P(A) $= \frac{2}{2\,1+3} = \frac{2}{15}$

P(B) $= \overline{1+4} = \overline{5}$

P(A or B) = P(A) + P(B) since A and B are m.e

$2 + 1 \quad = 3$

$= \therefore$ — — —

5 5 5

Odds that A or B wins are 3 to 2.

Example 13

Given P(A) = 0.30, P(B) = 0.78 and $P(A \cap B) = 0.16$. Find

i. $P(A^C \cap B^C)$ ii, $P(A^C \cup B^C)$ iii. $P(A \cap B^C)$

Solution

Given P(A) = 0.30, P(B) = 0.78 and $P(A \cap B) = 0.16$.

(i) $P(A^C \cap B^C) \quad = P\{(A \cup B)^C\} = 1 - P(A \cup B)$

$1 - \{P(A) + P(B) - P(A \cap B)\}$

$1 - \{0.30 + 0.78 - 0.16\}$ = **0.08**

$P(A^C \cup B^C) = P\{(A \cap B)^C\} = 1 - P(A \cap B)$

1 – 0.16 = **0.84**

(iii) $P(A \cap B^C) \quad = P[A - (A \cap B)]$

$P(A) - P(A \cap B)$

0.30 – 0.16 = **0.14**

Example 14

The probability that a student passes statistics test is 2/3 and the probability that he passes both statistics and Mathematics test is 14/45. The probability that he passes at least one test is 4/5. What is the probability that he passes Mathematics test?

Solution

Define, A - the student passes statistics test.

B -he passes the Mathematics test.

Given P(A) = 2/3. $P(A \cap B)$ = 14/45. $P(A \cup B)$ = 4/5

We have to find P(B). By addition theorem,

$P(A \cup B) = P(A) + P(B) - P(A \cap B)$

ie., 4/5 = 2/3 + P(B) – 14/45.

$\therefore$P(B) = 4/5 – 2/3 + 14/45 = $\frac{70}{}$

CONDITIONAL PROBABILITY

Definition

Let A and B be any two events. The probability of the event A given that the event B has already occured or the conditional probability of A given B, denoted by P(A | B) is defined as

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}, \quad P(B) \neq 0$$

Similarly the conditional probability of B given A is defined as

$$P(B \mid A) = \frac{P(A \cap B)}{P(A)}, \quad P(A) \neq 0$$

Remarks:

(i) For P(B) > 0 $\quad P(A \mid B) \leq P(A)$

P(A | B) is not defined if P(B) = 0

P(B | B) = 1

Theorem

For a fixed B with P(B) > 0, P(A | B) is a probability function (or probability measure).

Proof

Here we have to show that conditional probability satisfies all the axioms of probability.

(i) $P(A \mid B) = \frac{P(A \cap B)}{\quad} \geq$ 0, by axiom (1)

(ii) $P(S \mid B) = \frac{P(S \cap B)}{P(B)} = \frac{P(B)}{P(B)} = 1$

(iii) For any two adjoint events A and C

$$P(A \cup C \mid B) = \frac{P[(A \cup C) \cap B]}{P(B)}$$

$$= \frac{P[(A \cap B) \cup (C \cap B)]}{P(B)} \quad \text{by associative property}$$

$$= \frac{P(A \cap B) + P(C \cap B)}{P(B)} \quad \text{since } A \cap B \text{ and } C \cap B \text{ are disjoint}$$

$$= \frac{P(A \cap B)}{P(B)} + \frac{P(C \cap B)}{P(B)} = P(A|B) + P(C|B)$$

That is, conditional probability satisfies all the axioms of probability. Therefore P(A|B) is a probability function or probability measure.

Multiplication law of probability

Theorem

For any two events A and B

$P(A \cap B) = P(A)\, P(B|A),\ P(A) > 0$

$\qquad\quad = P(B)\, .\, P(A|B),\ P(B) > 0$

where P(A|B) and P(B|A) are the conditional probabilities of A and B respectively.

Independent Events

Definition

Two or more events are said to be *independent* if the probability of any one them is not affected by the supplementary knowledge concerning the materialisation of any number of the remaining events. Otherwise they are said to be *dependent.*

Independence of two events A and B

An event A is said to be independent (statistically independent) of event B, if the conditional probability of A given B, i.e., $P(A|B)$ is equal to the unconditional probability of A.

In symbols, $P(A \mid B) = P(A)$

Similarly if the event B is independent of A, we must have

$P(B \mid A) = P(B)$

Since $P(A \cap B) = P(A)\, P(B|A)$ and since $P(B|A) = P(B)$ when B is independent of A, we must have, $P(A \cap B\} = P(A) . P(B)$

Hence, the events A and B are independent if

$P(A \cap B) = P(A)\, P(B)$

Pairwise and Mutual independence

Definition

A set of events $A_1, A_2, \dots , A_n$ are said to be pairwise independent if every pair of different events are independent.

That is, $P(A_i \cap A_j) = P(A_i)\, P(A_j)$ for all i and j, $i \neq j$.

Definition

A set of events $A_1, A_2, \dots , A_n$ are said to be mutually independent if

$P(A_i \cap A_j \cap \dots \cap A_r) = P(A_i)\, P(A_j) \dots P(A_r)$ for every subset $(A_i, A_j, \dots, A_r)$ of $A_1, A_2, \dots , A_n$

That is the probabilities of every two, every three…, every n of the events are the products of the respective probabilities.

For example, three events A, B and C are said to be mutually independent if

$$\begin{aligned} P(A \cap B) &= P(A)\, P(B) \\ P(B \cap C) &= P(B)\, P(C) \\ P(A \cap C) &= P(A)\, P(C) \text{ and} \\ P(A \cap B \cap C) &= P(A)\, P(B)\, P(C) \end{aligned}$$

Note. 1

For the mutal independence of n events, $A_1, A_2, \dots , A_n$ the total number of conditions to be satisfied is $2^n - 1 - n$. In particular, for three events we have $4 = (2_3 - 1 - 3)$ conditions for their mutual independence.

Note. 2

We can note that pairwise or mutual independence of events $A_1, A_2, \dots$ A_n is defined only when $P(A_i) \neq 0$, for $i = 1, 2, \dots, n$.

Note 3

Pairwise independence does not imply mutual independence.

Theorem

Mutual independence of events implies pairwise independence of events. The converse is not true.

Proof

From the definition of mutual independence, it is clear that mutual independence implies pair-wise independence. We shall prove that the converse is not necessarily true. i.e., pair-wise independence does not imply mutual independence. We can illustrate it by means of an example due to S.N. Bernstein.

Let $S = \{\omega_1, \omega_2, \omega_3, \omega_4\}$ where $P(\omega_i) = 1/4$ for for $i = 1, 2, 3, 4$.

Let $A = \{\omega_1, \omega_2\}$, $B = \{\omega_1, \omega_3\}$ and $C = \{\omega_1, \omega_4\}$
Then $P(A)= P(B) = P(C) = 1/2$

and consider the collection of events A.B,C. These events are pairwise independent but not mutually independent.

Since they are pairwise independent we have,

$P(A \cap B) = 1/4 = P(A)P(B)$

$P(B \cap C) = 1/4 = P(B)P(C)$

$P(A \cap C) = 1/4 = P(A)P(C)$

But $P(A \cap B \cap C) = P(\omega_1) = 1/4$

$$P(A).P(B).P(C) = \frac{1}{2} . \frac{1}{2} . \frac{1}{2} = \frac{1}{8}$$

Thus $P(A \cap B \cap C)\ ^{1}\ P(A) P(B) P(C)$

Hence they are not mutually independent.

Multiplication Theorem (independent events)

If A and B are two independent events,

$P(A \cap B) = P(A).P(B)$

Proof

We have, for any two events A and B

$P(A\cap B) = P(A) P(B|A)$

Since A and B are independent, we have $P(B|A) = P(B)$,

$P(A \cap B) = P(A) P(B)$.

Note

If A and B are independent the addition theorem can be stated as $P(A \cup B) = P(A) + P(B) – P(A). P(B)$

Theorem

If A and B are two independent events

then (i) A and B^C are independent

A^C and B are independent

A^C and B^C are independent

Proof

Since A and B are independent, we have

$P(A|B) = P(A).\ P(B|A) = P(B)$ and $P(A\cap B) = P(A).P(B)$

(i) Now, $P(A\cap B^C) = P(A) P(B^C|A)$

$P(A) [1–P(B|A)]$

$P(A) [1– P(B)]$

$P(A) P(B^C)$

ie., A and B^C are independent

(ii) $P(A^C \cap B) = P(B) P(A^C|B)$

$P(B) [1 – P(A | B)]$

$P(B) [1 – P(A)]$

$P(A^C) P(B)$

ie., A^C and B are independent

(iii) $P(A^C \cap B^C) = P(A \cup B)^C = 1 – P(A \cup B)$

$1 – \{P(A) + P(B) – P(A \cap B)\}$

$1 – P(A) – P(B) + P(A) P(B)$

since $P(A\cap B) = P(A) P(B)$

$[1 – P(A)] – P(B) [1 – P(A)]$

$[1 – P(A)] [1 – P(B)] = P(A_C) P(B_C)$

i.e., A^C and B^C are independent

Baye’s Theorem

$$P(B_i / A) = \frac{P(\ B_i \quad)\ P\ \ (A \quad |}{\sum_{i=1}^{n} P(\ B_i \quad)\ P\ (\ A \quad |}$$

Note 1

Here the probabilities $P(B_i | A)$ for i = 1, 2, …, n are the probabilities determined after observing the event A and $P(B_i)$ for i = 1, 2, ..., n are the probabilities given before hand. Hence P(Bi) for i = 1, 2, ..., n are called ‘a priori’ probabilities and $P(B_i | A)$ for i =1, 2, ..., n are called “a posteriori’ probabilities. The probabilities $P(A|B_i)$, i = 1, 2, ..., n are called ‘likely hoods’ because they indicate how likely the event A under consideration is to occur, given each and every, ‘a priori’ probability. Baye’s theorem gives a relationship between $P(B_i | A)$ and $P(A | B_i)$ and thus it involves a type of inverse reasoning. Baye’s theorem plays an important role in applications. This theorem is due to Thomas A Baye’s.

Note 2

In the case of two events A and B satisfying the assumption P(B) > 0 and 0 < P(B) < 1 we have,

$$P(B \mid A) = \frac{P(B)P(A \mid B)}{P(B)P(A \mid B) + P(B^C)P(A \mid B^C)}$$

Example 1

Let A and B be two events associated with an experiment and suppose P(A) = 0.5 while P(A or B) = 0.8. Let P(B) = p. For what values of p are (a) A and B mutually exclusive (b) A and B independent.

Solution

Given P(A) = 0.5, $P(A \cup B) = 0.8$, P(B) = p

(a) If A and B are mutually exclusive

$P(A\cup B) = P(A) + P(B)$

i.e., $0.8 = 0.5 + p$

p = 0.3

If A and B are independent, we have

$P(A\cup B) = P(A) + P(B) – P(A)P(B)$

i.e., $0.8 = 0.5 + p – .5p$

$.5p = 0.3$ ∴ **p = 3/5**

Example 2

If A and B are two events such that P(A) = 1/3, P(B) = 1/4 and P(AÇB) = 1/8. Find P(A|B) and $P(A|B^C)$

Solution

Given P(A) = 1/3, P(B) = 1/4, $P(A\cap B) = 1/8$

$$\therefore P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{1/8}{1/4} = 4/8 = \mathbf{1/2}$$

$$P(A|B^C) = \frac{P(A \cap B^C)}{P(B^C)} = \frac{P(A) - P(A \cap B)}{1 - P(B)}$$

$$= \frac{1/3 - 1/8}{1 - 1/4} = \frac{5/24}{3/4} = \mathbf{5/18}$$

Example 3

The odds that A speaks the truth are 3:2 and the odds that B speaks the truth are 5:3. In what percentage of cases are they likely to contradict each other on an identical point?

Solution

Define the events,

A - A speaks the truth

B - B speaks the truth

$P(A) = 3/5$, $P(A^C) = 2/5$

$P(B) = 5/8$, $P(B^C) = 3/8$

They will contradict each other on an identical point means that when A speaks the truth, B will tell a lie and conversely.

∴ P(They will contradict each other) = $[P(A\cap B^C) \cup (A^C \cap B)]$

$= P(A \cap B^C) + P(A^C \cap B)$, since the events are m.e.

$= P(A)\, P(B^C) + P(A^C)\, P(B)$

$$= \frac{3}{5} \cdot \frac{3}{8} + \frac{2}{5} \cdot \frac{5}{8} = \mathbf{19/40}$$

ie,, In 47.5% of the cases, A and B contradict each other.

Example 4

A husband and wife appear in an interview for two vacancies in a firm. The probability of husbands selection is 1/7 and that of wife’s selection is 1/5. What is the probability that

both of them will be selected.

only one of them will be selected. (c) none of them will be selected.

Solution

Let us define the events as

A - The husband get selection.

B - The wife get selection.

$P(A) = 1/7$, $P(B) = 1/5$; $P(A^C) = 6/7$; $P(B^C) = 4/5$

(a) P(both of them will be selected) = $P(A \cap B)$ =

$P(A) . P(B)$, since A and B are independent

$$= \frac{1}{7} . \frac{1}{5} = \frac{1}{35}$$

P(only one of them will be selected)

$$P[(A \cap B^C) \cup (A^C \cap B)]$$

$$P(A \cap B^C) + P(A^C \cap B)$$

$$= P(A) \quad P(B^C) \quad + \quad P(A^C) \quad P(B)$$

$$\frac{1}{7} \times \frac{4}{5} + \frac{6}{7} \times \frac{1}{5}$$

10/35

(c) P(none of them will be selected) = $P(A^C \cap B^C)$

$$= P(A^C) P(B^C) = \frac{6}{7} . \frac{4}{5} = \mathbf{24/35}$$

Example 5

If A, B and C are independent, show that $A \cup B$ and C are independent.

Solution

Since, A, B and C are independent, we have

$P(A \cap B) = P(A)P(B)$, $P(B \cap C) = P(B)P(C)$

$P(A \cap C) = P(A)P(C)$ and $P(A \cap B \cap C) = P(A)P(B)P(C)$

We have to show that

$$\begin{aligned} P[(A \cup B) \cap C] &= P[(A \cap C) \cup (B \cap C)] \\ &= P(A \cap C) + P(B \cap C) - P(A \cap B \cap C) \\ &= P(A)P(C) + P(B)P(C) - P(A)P(B)P(C) \\ &= P(C)\,[P(A) + P(B) - P(A)\,P(B)] \\ &= P(A \cup B) . P(C) \end{aligned}$$

i.e., $A \cup B$ and C are independent.

Example 6

A problem in statistics is given to 3 students A, B and C whose chances of solving it are 1/2, 3/4 and 1/4 respectively. What is the probability that the problem will be solved?

Solution

Let us define the events as

A – the problem is solved by the student A

B – the problem is solved by the student B

C – the problem is solved by the student C

$P(A) = 1/2$, $P(B) = 3/4$ and $P(C) = ¼$

The problem will be solved if at least one of them solves the problem. That means we have to find $P(A \cup B \cup C)$.

Now $P(A \cup B \cup C)$

$$P(A) + P(B) + P(C) - P(A \cap B)$$
$$- P(B \cap C) - P(A \cap C) + P(A \cap B \cap C)$$

$$P(A) + P(B) + P(C) - P(A)P(B)$$
$$- P(B)P(C) - P(A)P(C) + P(A)P(B)P(C)$$

$$\frac{1}{2} + \frac{3}{4} + \frac{1}{4} - \frac{1}{2} \times \frac{3}{4} - \frac{3}{4} \times \frac{1}{4} - \frac{1}{2} \times \frac{1}{4} + \frac{1}{2} \times \frac{3}{4} \times \frac{1}{4}$$

29/32

Aliter

$P(A \cup B \cup C) \quad = 1 - P(A \cup B \cup C)^c$

$1 - P(A^c \cap B^c \cap C^c)$

$1 - P(A^C)\, P(B^C)\, P(C^C)$

$$= 1 - \left(1 - \frac{1}{2}\right)\left(1 - \frac{3}{4}\right)\left(1 - \frac{1}{4}\right)$$

= **29/32**

Example 7

A purse contains 2 silver coins and 4 copper coins and a second purse contains 4 silver coins and 3 copper coins. If a coin is selected at random from one of the purse. What is the probability that it is a silver coin?

Solution

Define the events

B_1 – selection of 1st purse

B_2 – selection of 2nd purse

– selection of silver coin

$P(B_1) = P(B_2) = 1/2$

$P(A|B_1) = 2/6$, $P(A|B_2) = 4/7$

By theorem on total probabilities

$P(A) \quad = P(A \cap B_1) + P(A \cap B_2)$

$= P(B_1)\, P(A|B_1) + P(B_2)\, P(A|B_2)$

$$= \frac{1}{2} \times \frac{2}{6} + \frac{1}{2} \times \frac{4}{7}$$

$$\frac{1}{6} + \frac{2}{7} = \frac{7+12}{42} = \frac{19}{42}$$

Example 8

Suppose that there is a chance for a newly constructed house to collapse wether the design is faulty or not. The chance that the design is faulty is 10%. The chance that the house collapse if the design is faulty is 95% and otherwise it is 45%. It is seen that the house collapsed. What is the probability that it is due to faulty design?

Solution

Let B_1 and B_2 denote the events that the design is faulty and the design is good respectively. Let A denote the event that the house collapse. Then we are interested in the event $(B_1|A)$, that is, the event that the design is faulty given that the house collapsed. We are given,

$P(B_1) = 0.1$ and $P(B_2) = 0.9$

$P(A|B_1) = 0.95$ and $P(A|B_2) = 0.45$

Hence

$$P(B_1|A) = \frac{P(B_1).P(A \mid B_1)}{P(B_1).P(A \mid B_1) + P(B_2).P(A \mid B_2)}$$

$$\frac{(0.1)(0.95)}{(0.1)(0.95) + (0.9)(0.45)}$$

0.19

Example 9

Two urns I and II contain respectively 3 white and 2 black bails, 2 white and 4 black balls. One ball is transferred from urn I to urn II and then one is drawn from the latter. It happens to be white. What is the probability that the transferred ball was white.

Solution

Define

B_1 - Transfer a white ball from Urn I to Urn II

B_2 - Transfer a black ball from Urn I to Urn II.

A - Select a white ball from Urn II.

Here, $P(B_1) = 3/5$, $P(B_2) = 2/5$

$P(A|B_1) = 3/7$, $P(A|B_2) = 2/7$

We have to find $P(B_1|A)$,

By Baye’s theore,

$$P(B_1|A) = \frac{P(B_1).P(A \mid B_1)}{P(B_1)P(A \mid B_1) + P(B_2)P(A \mid B_2)}$$

$$3/5 \times 3/7 \qquad\qquad 9/359$$

$$= \overline{3/5 \times 3/7 + 2/5 \times 2/7} \quad = \overline{13/35} = \overline{13}$$

EXERCISES

Multiple choice questions

Probability is a measure lying between

a) $-\infty$ to $+\infty$ b) $-\infty$ to $+1$

c) -1 to $+1$ d) 0 to 1

Classical probability is also known as

a) Laplace’s probabilityb) mathematical probability

c) a priori probability d) all the above

Each outcome of a random experiment is called

a) primary event b) compound event

c) derived event d) all the above

If A and B are two events, the probability of occurance of either A or B is given by

a) P(A)+P(B) b) $P(A \cup B)$

c) $P(A \cap B)$ d) P(A)P(B)

The probability of intersection of two disjoint events is always

a) infinity b) zero

c) one d) none of the above

6. If $A \subset B$, the probability P(A|B) is equal to

a) zero b) one

c) P(A)/P(B) d) P(B)/P(A)

The probability of two persons being borned on the same day (ignoring date) is

a) 1/49 b) 1/365

c) 1/7 d) none of the above

8. The probability of throwing an odd sum with two fair dice is

a) 1/4 b) 1/16 c) 1 d) 1/2

9. If P(A|B) = 1/4, P(B|A) = 1/3, then P(A)|P(B) is equal to

a) 3/4 b) 7/12

c) 4/3 d) 1/12

If four whole numbers are taken at random and multiplied, the chance that the first digit is their product is 0, 3, 6 or 9 is

a) $(2/5)^3$ b) $(1/4)^3$ c) $(2/5)^4$ d) $(1/4)^4$

Fill in the blanks

Classical definition of probability was given by ……….

An event consisting of only one point is called ……….

Mathematical probability cannot be calculated if the outcomes are ……….

In statistical probability n is never ……….

If A and B are two events, the $P(A \cap B)$ is ……….

Axiomatic definition of probability is propounded by ……….

Baye’s rule is also known as ……….

If an event is not simple, it is a ……….

Very short answer questions

Define a simple event.

Define random experiment.

Define equally likely cases.

State statistical definition of probability.

Define conditional probability

State Baye’s rule

Short essay questions

Define Sample space and Event When will you say that two events are mutually exclusive?

Define random experiment, sample space and Event. A coin is repeatedly tossed till a head turns up. Write down the sample space.

Give the classical and axiomatic definition of probability, Explain how axiomatic definition is more general than classical.

Define (i) Mutually exclusive events: (ii) Equally likely

events: and (iii) Independent events and give example of each.

Give Von Mises definition of empirical probability, Compare this with the classical definition of probability.

State and prove the addition theorem of probability.

Define Conditional probability.

State and prove addition and multiplication theorem of probability.

Show that

$P(A \cap B) \leq P(A) \leq P(A \cup B) \leq P(A) + P(B)$

State and prove Bayes’ theorem.

Define Conditional probability. Prove that if $P(A) > P(B)$ then

$P(A|B) > P(B|A)$.

Let A, B and C denote events. If $P(A \mid C) \geq P(B \mid C)$ and

$P(A \mid C^{c)} \geq P(B \mid C^{C})$, then show that $P(A) \geq P(B)$

Long essay questions

Two unbiased dice are tossed. What is the probability that the sum of points scored on the two dice is 8?

From a group consisting of 6 men and 4 women a committee of 3 is to be chosen by lot. What is the probability that all 3 are men?

Two events A and B are statistically independent. $P(A) = 0.39$, $P(B) = 0.21$ and $P(A \text{ or } B) = 0.47$. Find the probability that

Neither A nor B will occur

Both A and B will occur

B will occur given that A has occurred

A will occur given that B has occurred

If $P(A) = 0.3$, $P(B) = 0.2$. $P(A \cup B) = 0.4$, find

$P(A \cap B)$. Examine whether A and B are independent.

The probability that A hits a target is 1/4 and the probability that B hits it is 2/5. What is the probability that the target will be hit if A and B each shoot at the target?

A coin is tossed four times. Assuming that the coin is unbiased, find the probability that out of four times, two times result in head,

Two urns each contain balls of different colours are stated below.

urn I : 4 black; 3 red; 3 green.

urn II : 3 black; 6 red: 1 green.

An urn is chosen at random and two balls are drawn from it. What is the probability that one is green and the other is red.

If two dice are rolled, what is the probability that the sum is 7 if we know that at least one die shows 4?

There are three urns containing balls of different colours as stated below:

Urn I : 4 red, 2 black, 4 green.

Urn II : 3 red, 4 black, 5 green.

Urn III: 2 red, 4 black, 2 green.

An urn is chosen at random and two balls are drawn from it. What is the probability that both are red?

Three urns are given each containing red and while chips as indicated.

Urn 1 : 6 red and 4 white.

Urn 2 : 2 red and 6 white.

Urn 3 : 1 red and 8 white.

An urn is chosen at random and a ball is drawn from the urn. The ball is red. Find the probability that the urn chosen was urn 1.

An urn is chosen at random and two balls are drawn without replacement from this urn. If both balls are red, find the probability that urn 1 was chosen. Under these conditions, what is the probability that urn III was chosen.

State Baye's theorem. A box contains 3 blue and 2 red balls while another box contains 2 blue and 5 red balls. A ball drawn at random from one of the boxes turns out to be blue. What is the probability that it came from the first box?

In a factory machines A, B and C produce 2000, 4000 and 5000 items in a month respectively, Out of their output 5%, 3% and 7% are defective. From the factory's products one is selected at random and inspected. What is the probability that it is good? If it is good, what is the probability that it is from machine C?

MODULE IV

RANDOM VARIABLE AND PROBABILITY DISTRIBUTIONS

We have seen that probability theory was generally characterised as a collection of techniques to describe, analyse and predict random phenomena. We then introduced the concept of sample space, identified events with subsets of this space and developed some techniques of evaluating probabilities of events.

The purpose of this chapter is to introduce the concepts of random variables, distribution and density functions and a thorough understanding of these concepts is very essential for the development of this subject.

Random variables, to be introduced now, can be regarded merely as useful tools for describing events. A random variable will be defined as a numerical function on the sample space S.

Definition

A random variable (r.v.) is a real valued function defined over the sample space. So its domain of definition is the sample space S and range is the real line extending from −∞ to +∞. In other words a r.v. is a mapping from sample space to real numbers. Random variables are also called *chance variables or stochastic variables*. It is denoted by X or X(ω).

In symbols, X : S → R (−∞, +∞)

The above definition of a random variable as such is not perfect. Because all functions defined on S cannot be random variables. It has to satisfy some basic requirements.

From the point of view of modern mathematics an acceptable definition of a random variable is given below.

Random variable X is a function whose domain is S and range∈ is set of real values from −∞ to +∞ such that the set {X $\leq$ x} B, the Borel field, for any real number x. That means random variables X are functions on S which are measurable w.r.t. the Borel field.

Here we can note that each set of the form {X $\leq$ x} is an event.

As an illustration, consider a random experiment consisting of two tosses of a coin. Then the sample space is given by S = {HH, HT, TH, TT}

Then to each outcome in the sample space there corresponds a real number X( ). It can be presented in the tabular form as,

Outcome (ω)	:	HH	HT	TH	TT
Values of X(ω)	:	2	1	1	0

Thus we have defined a one dimensional random variable as a real valued function on S which is associated with a random experiment. That is a one dimensional random variable is a measurable function X( ) with domain S and range (−∞, +∞) such that for every real number x,∈

the event { ω: X(ω) $\leq$ x} B.

Example

In coin tossing experiment, we note that

Now define X( ) =

S = { ω_1 , ω_2 } where ω_1 = Head, ω_2 = Tail

$$\omega \quad \begin{cases} 0\ if\ \omega = T\ (T \\ 1\ if\ \omega = \text{Head (H)} \end{cases}$$

Here the random variable X(ω) takes only two values as ω can be either head or tail. Such a random variable is known as Bernoulli random variable.

Remark: If X_1 and X_2 are r.v.s. and C is a constant,

then (i) C X_l is a r.v.

$X_l + X_2$ is a r.v.

$X_l - X_2$ is a r.v.

max[X_l, X_2] is a r.v.

min[X_l, X_2] is a r.v.

Random Variables are of two types (i) discrete (ii) continuous. A random variable X is said to be *discrete* if its range includes finite number of values or countably infinite number of values. The possible values of a discrete random variable can be labelled as x_l , x_2 , x_3... eg. the number of defective articles produced in a factory in a day in a city, number of deaths due to road accidents in a day in a city, number of patients arriving at a doctors clinic in a day etc.

A random variable which is not discrete is said to be *continuous*. That means it can assume infinite number of values from a specified interval of the form [a, b].

A few examples of continuous random variable are given below

(1) A man brushes his teeth every morning. X represents the time taken for brushing, next time (2) X represents the height of a student randomly chosen from a college

(3) X represents the service time of a doctor on his next patient

(4) life time of a tube etc.

Note that r.v.s. are denoted by capital letters X, Y, Z etc. and the corresponding small letters are used to denote the value of a.r.v.

We are not interested in random variables, where as, we will be interested in events defined in terms of random variables. From the definition of the random variable X, we have seen that each set of the form $\{X \leq x\}$ is an event. The basic type of events that we shall consider are the following.

$\{X=a\}$, $\{X=b\}$, $\{a < X < b\}$ $\{a< X \leq b\}$, $\{a \leq X < b\}$, $\{a\leq X \leq b\}$ where $-\infty \leq a \leq b \leq \infty$. The above subsets are being events, it is permissible to speak of its probability. Thus with every random variable we can associate its probability distribution or simply distribution.

Definition

By a distribution of the random variable X we mean the assignment of probabilities to all events defined in terms of this random variable.

Now we shall discuss the probability distributions in the case of discrete as well as continuous random variables.

Probability Distributions

i. Discrete:

The probability distribution or simply distribution of a discrete r.v. is a list of the distinct values of x_i of X with their associated probabilities

$f(x_i) = P(X = x_i)$.

Thus let X be a discrete random variable assuming the values x_1, x_2, ...x_n from the real line. Let the corresponding probabilities be $f(x_1)$, $f(x_2)$....$f(x_n)$. Then $P(X = x_i) = f(x_i)$ is called probability mass function or probability function of X, provided it satisfy the conditions

$f(x_i) \geq 0$ for all i

$\Sigma f(x_i) = 1$

The probability distribution of X may be stated either in the form of a table or in the form of a formula. The formula gives f(x) in terms of x which represents P(X = x). It can also be denoted as p(x). The formula model is always convenient but it need not be available for all random variables. We can sketch the graph of a probability distribution or probability mass function as given below.

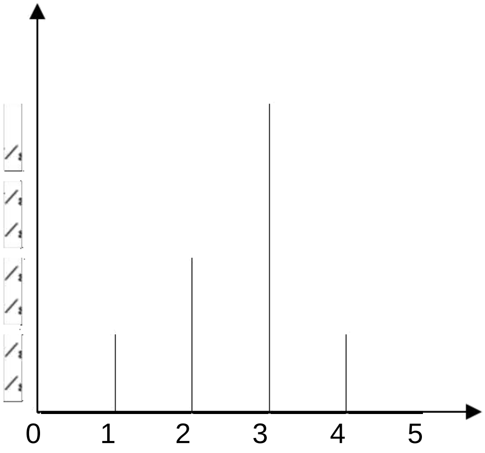
Graphical presentation of a probability distribution

It is usually true that we cannot appreciate the salient features of a Probability distribution by looking at the number in table. The two main ways to graph a discrete probability distribution are the line diagram and Probability histogram.

Line diagram

The distinct values of the r.v. X are marked on the X axis. At each value x, a vertical line is drawn whose height in equal to its probability. f(x) or p(x). for example, the line diagram for the following probability distribution is shown below.

Value x	1	2	3	4
f(x)	8	4	2	8

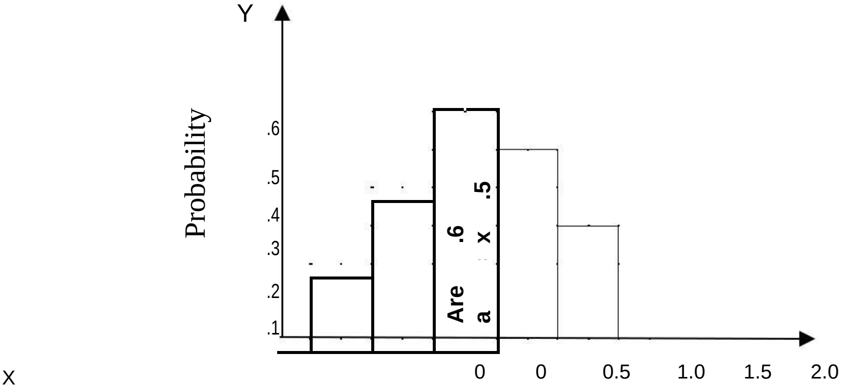


Line Diagram

Probability Histogram

On the X axis we take values of the r.v. With each value x as centre, a vertical rectangle is drawn whose area is equal to the probability f(x). Note that in plotting a probability histogram the area of a rectangle must be equal to the probability of the value at the centre. So that the total area ruler the histogram must be equal to the total probability (i.e. unity). For example the probability histogram of the following probability distribution is given below.

x	0	0.5	1.0	1.5	2.0
f(x)	.1	.2	.3	.25	.15



Probability histogram

The probability histogram is recommended for distribution with equal spaced x values. When the spacing of x values is unequal the line diagram should be used. One advantage of probability histogram is that we can compare two or more probability distribution to determine the nature and extent of their similarities and dissimilarities.

ii. Continuous

We now turn our attention to describing the probability distribution of a random variable that can assume all the values in an interval. The probability distribution of a continuous random variable can be visualised as a smooth form of the relative frequency histogram based on a large number of observations. Because probability is interpreted as long run relative frequency, the curve obtained as the limiting form of the relative frequency histograms represent the manner in which the total probability, is distributed over the range of possible values of the random variable X. The mathematical function denoted by f(x) whose graph produces this curve is called probability density function of the continuous r.v. X.

Definition

If X is a continuous random variable and if $P(x \le X \le x + dx) = f(x)dx$, then f(x) is called probability density function (pdf)∀ of a continuous∫ f(r.xv).dxprovided=1 it satisfy the conditions (i) $f(x) \ge 0$ x and

(ii) ∞ $-\infty$.

We can justify the term ‘density function’ to some extent from the following argument. We have

$x+ x$

x

When Δx is very small, the mean value gives us the approximation. ∫ Δ $f(x)dx = f(x).\Delta x$

$x+ x$

x

$\therefore f(x)\Delta x = P(x \le X \le x + \Delta x)$

$$\frac{P(x \le X \le x + \Delta x)}{x}$$

= Total probability in the interval $\dfrac{(x,\ x+\Delta x}{\text{length of the interval}}$

= Probability per unit length

= Density of probability

Result. 1

$P(a \le x \le b) = P(a \le x < b)$

$$P(a \le x \le b) \int_{a} f(x)dx$$

= the area under the curve y = f(x), enclosed between the ordinates drawn at x = and x = b.

Result. 2

Probability that a continuous r.v. X will assume a particular value is zero ie., P(X =) = 0

Distribution function

Definition

For any random variable X, the function of the real variable x defined as $F_X(x) = P(X £ x)$ is called cumulative probability distribution function or simply cumulative distribution function (cdf) of X. We can note that the probability distribution of a random variable X is determined by its distribution function.

If X is a discrete r.v. with pmf p(x), then the cumulative distribution function is defined as

If X is a

$F_X(x) = P(X \leq x) = \sum_{-\propto}^{x}$.

continuous r.v. with pdf f(x), then the distribution function is defined as

$$\int_{-\propto}^{x}$$

For convenience we will write $F_X(x)$ as F(x) or F.

Properties of distribution function

If F(x) is the distribution function of a r.v. X then it has the following properties.

F(x) is defined for all real values of x.

2. F(−∞) = 0, F(+∞) = 1.

$0 \leq F(x) \leq 1$

$F(a) \leq F(b)$ if $a < b$

That means F(x) is non decreasing

If X is discrete, $F(b) - F(a) = P(a < X \leq b)$

For a discrete r.v. the graph of f(x) indicates that it is a step function or a staircase function.

F(x) is a continuous function of x on the right.

If X is continuous $F(b) - F(a) = P(a \leq X \leq b)$ = Area under the probability curve.

F(x) possesses a continuous graph, if X is continuous. If F(x) possesses a derivative, then $\frac{d\,F(x)}{dx} = f(x)$

The discontinuities of F(x) are at the most countable.

Moments of a continuous probability distribution

Let X be a random variable with the pdf f(x). Then for a continuous r.v., we can calculate the measures of central tendency and the measures of disperson as follows. For a discrete r.v., we have to replace integration by summation.

1. Arithmetic Mean of X = $\int_{-\infty}^{\infty} x f(x)dx$.

Median, M is determined by solving the equation $\int_{-\infty}^{M} f(x)dx = \frac{1}{2}$ or $\int_{M}^{\infty} f(x)dx = \frac{1}{2}$

3. Mode (Z) is determined by solving the equation f'(x) = 0 and verifying the condition f''(x) < 0 at the mode.

4. Geometric Mean is given by the equation $\log G = \int_{-\infty}^{\infty} (\log x) f(x)dx$

5. Harmonic Mean can be calculated by the equation

6. – =$\int$ – f(x)dx

7. f(x)dx = ; $\int$ f(x)dx =

Quartiles are determined by solving the equations

; $\int$ (x − μ) $\frac{1}{4}$, f(x)dx, $\cdot Q3$ $-\infty$ $\frac{3}{4}$

The r_{th} central moment is determined by the equation

= μ_r r = l, 2, 3....

9. where is the mean of X.

8. In $_2$ = μ

MD about Mean is given by

MD = $\int$ |x − mean| f(x)dx,

SOLVED PROBLEMS

Example l

Obtain the probability distribution of the number of heads when three coins are tossed together?

Solution

When three coins are tossed, the sample space is given by S = {HHH, HHT, HTH, THH, HTT, THT, TTH, TTT}. Here the r.v. X defined as the number of heads obtained will takes the values 0, l, 2 and 3 from the real line w.r.t each outcome in S. We can assign probabilities to each value of the r.v. as follows.

P{X = 0} = P{TTT} = $\frac{1}{8}$

P{X = 1} = P{HTT or THT or TTH} = $\frac{3}{8}$

P{X = 2} = P{HHT or HTH or THH} = $\frac{3}{8}$

P{X = 3} = P{HHH} = $\frac{1}{8}$

Thus the probability distribution of X is given by

X = x	0	1	2	3	Total
P(X = x)	$-^{8}$	$-^{8}$	$-^{8}$	$-^{8}$	1

Example 2

From a lot containing 25 items 5 of which are defectives. 4 items are chosen at random. Find the probability distribution of the number of defectives obtained.

Solution

Let X be the number of defectives obtained. Here X takes the values 0, l, 2, 3 and 4.. If the═items are drawn without replacement.

umɔer c .ecti\.

d 4. If th ns ar

P{X = 0} = $\frac{5C_0.20C_4}{25C_4} = \frac{969}{2530}$

= = = ═

P{X = 1} = $\frac{5C_1.20C_3}{25C_4}$ $\frac{1140}{2530}$

P{X = 2} = $\frac{5C_2.20C_2}{25C_4}$ $\frac{380}{2530}$

P{X = 3} = $\frac{5C_3.20C_1}{25C_4}$ $\frac{40}{2530}$

P{X = 4} = $\frac{5C_4.20C_0}{25C_4}$ $\frac{1}{2530}$

Thus the probability distribution of X is

X = x	0	1	2	3	Total
P(X = x)	~~2530~~	~~2530~~	~~2530~~	~~2530~~	

We can write this probability distribution as a probability function as shown below.

$$f(x) = P(X = x) = \frac{5C_x . 20C_{4-x}}{25C_4}, \ x = 0, 1, 2, 3, 4$$

0, elsewhere.

Example 3

Examine whether f(x) as defined below is a pdf.

$$f(x) = 0 \ ; \ x < 2$$

$$\frac{1}{18}(3 + 2x) \ ; \ 2 \leq x < 4$$

$$0 \ ; \ x > 4.$$

Solution Here = $\int$ f(x)dx =

To show that f(x) is a pdf we have to show that

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

$-\infty$ ∞

$$= \int_2^4 \frac{1}{18}(3 + 2x)dx = \frac{1}{18}\left(3x + x^2\right)_2^4$$

$$= \frac{1}{18} |(12 + 16) - (6 + 4)| = \frac{1}{18} \text{ x } 18 = 1$$

So f(x) is a pdf.

Example 4

If the distribution function F(X) is given to be

$$F(x) = \begin{cases} \frac{2x^2}{5} & 0 < x \le 1 \\ -\frac{3}{5} + \frac{2\left(3x-\frac{x^2}{2}\right)}{5} & 1 < x \le 2 \\ 1 & x > 2 \end{cases}$$

find the density function.

Solution

We know (that) $\frac{D\,F(x)}{dx}$ = f(x)

Here f(x) = $\frac{d\,F(x)}{dx} = \frac{4x}{5}$, when 0 < x ≤ 1 = (3 –), when 1 < x ≤ 2

= 0, otherwise.

Example 5

Examine whether the following is a distribution function.

$$F(x) = \begin{cases} 0 & x < -a \\ \frac{1}{2}\left(\frac{x}{a}+1\right) & -a \le x \le a \\ 1 & x > a \end{cases}$$

Solution

a. F(x) is defined for all real values of x.

b. F(-∞) = 0 ;

c. F(∞) = l

d. F(x) is non decreasing.

e. F’(x) = $\begin{cases} \frac{1}{2a} & -a < x < a \\ 0 & x < -a \text{ or } x > a \end{cases}$

∴F'(x) satisfies the pdf properties.

F(x) is a distribution function.

Example 6

Given F(x) = $\begin{cases} 0 & x < 0 \\ x^2 & 0 \leq x \leq 1 \\ 1 & x > 1 \end{cases}$

Determine (a) $P(X \leq 0.5)$ (b) $P(0.5 \leq X \leq 0.8)$ (c) $P(X > 0.9)$

Solution

$P(X \leq 0.5) = F(0.5) = (0.5)^2 =$ **0.25**

$P(0.5 \leq X \leq 0.8) = P(0.5 < X \leq 0.8)$

$F(0.8) - F(0.5)$

$(0.8)^2 - (0.5)^2 =$ **0.39**

$P(X > 0.9) = l - P(X \leq 3.9) = l - F(0.9)$

$l - (0.9)^2 =$ **0.19**

Example 7

A random variable X has the density function

$f(x) = \frac{K}{1+X^2}$

0, elsewhere.

Determine K and the distribution function.

We know that $\int_{-\infty}^{\infty} f(x)\,dx = 1$

ie. $\int_{-\infty}^{\infty} K \frac{1}{1+x^2}\,dx = 1$

$K(\tan^{-1} x)_{-\infty}^{\infty}$

$K\left[\frac{\pi}{2} - \left(-\frac{\pi}{2}\right)\right] = 1$

$K\pi = 1 \quad K = 1/\pi$

The distribution function is = ∫

$= P(X \leq x) = \int_{-\infty}^{x} f(x)\, dx$

F(x) $= \int_{-\alpha}^{x} \frac{1}{\pi} \frac{1}{1+X^2}\, dx$ ( )

$\frac{1}{\pi}$ (t $-\alpha$ x x $-\infty$

(tan)

$= \frac{1}{\pi}\left[t \; \iota^{-1} x \quad \left(-\frac{\pi}{2}\right)\right]$

$= \frac{1}{\pi}\left[t \;\tan^{-1} x - \frac{\pi}{2}\right] -$

tan +

Example 8

Evaluate the distribution function F(x) for the following density function and calculate F(2)

Solution

F(x) $= \begin{cases} \frac{x}{3} & 0 < x \leq 1 \\ \frac{5}{27}\,(4 - x) & 1 < x \leq 4 \\ 0 & elsewhere \end{cases}$

By definition, F(x) = $\int_{-\alpha}^{x} f(y)dy$

Therefore, $\int_{-\alpha}^{x} 0dy = 0$ for any value of x such that $-\infty < x < 0$, F(x) =

For any x in 1 < x ≤4,

since f(x) = 0 in this interval. For any x in 0 < x ≤ l,

F(x) $= \int_{-\infty}^{0} 0\, dy + \int_{0}^{x} \frac{y}{3}\, dy = \frac{x^2}{6}$ $\frac{x^2}{6}$

F(x) $= \int_{-\alpha}^{0} 0\, dy + \int_{0}^{1} \frac{y}{3}\, dy + \int_{1}^{x} \frac{5}{27}(4 - y)\, dy$

$= 0 + \frac{1}{6} + \int_{1}^{x} \frac{5}{17}(4 - y)\, dy$

$\frac{13}{27}$ $\frac{5}{27}\left(4x - \frac{x^2}{2}\right)$

$+ \quad 4 \quad -$

Evidently for x ³ 4, F(x) = 1and hence F(x) can be written as,

$$F(x)=\begin{cases} 0 & -\infty < x < 0 \\ \frac{x^2}{6} & 0 < x \leq 1 \\ \frac{-13}{27}+\frac{5}{27}\left(4x-\frac{x^2}{2}\right) & 1 < x \leq 4 \\ 1 & x \geq 4 \end{cases}$$

Hence, F(2) at x = 2 $= \frac{-13}{27}+\frac{5}{27}\left(4x-\frac{x^2}{2}\right)$

$= \mathbf{\frac{17}{27}}$

Example 9

The probability mass function of a.r.v.X is given as follows.

x	0	1	2	3	4	5
p(x)	k^2	$\frac{k}{4}$	$\frac{5}{2}k$	$\frac{k}{4}$	$2k^2$	2

Final (a) k (b) write down the distribution function of x.

Solution

(a) We know that Σ P(x) = 1

ie., $k^2 + \frac{k}{4} + \frac{5}{2}k + \frac{k}{4} + k^2 + k^2 = 1$

$4k^2 + 3k - 1 = 0$

k = $\frac{-3 \pm \sqrt{9+16}}{8}$ = = $\frac{-3 \pm 5}{8}$ = 1 , $\frac{1}{4}$

Since probability is greater than zero, we have **k = 1/4**

The probability distribution of X is

	012	345	Total				
p(x)	1/16	1/16	10/16	1/16	2/16	1/16	1

So the distribution function of X is given by

$$F(x) = P(X \le x) = \begin{cases} 0 & \text{for } x < 0 \\ 1/16 & \text{for } 0 \le x < 1 \\ 2/16 & \text{for } 1 \le x < 2 \\ 13/16\;12/16 & \text{for } 3 \le x < 4 \\ 15/16 & \text{for } 4 \le x < 5 \\ 1 & \text{for } x \ge 5 \end{cases}$$

Example 10

Evaluate the distribution function for the following probability function.

$$\begin{aligned} f(x) &= \frac{1}{8} \quad \text{for } x = -l \\ &= \frac{2}{8} \quad \text{for } x = 0 \\ &= \frac{3}{8} \quad \text{for } x = 2 \\ &= \frac{2}{8} \quad \text{for } x = 3 \\ &= 0 \quad \text{elsewhere} \end{aligned}$$

Solution

By definition, the distribution function F(x) is, ∑ ∑ ( ) + $_{\infty}$ f f(x) $\frac{1}{8}$ = $\frac{1}{8}$

F(x) = P(X ≤f(x)x)= ∑

F(0) = ∑ f(x) = 0 + $\frac{1}{8}$ $\frac{3}{8}$ $\frac{6}{8}$ x x $-\infty$

F(2) = $\frac{1}{8}$ + $-\infty$ + $\frac{3}{8}$ = $\frac{6}{8}$ $-$ $\frac{1}{8}$ $\frac{2}{8}$ = $\frac{3}{8}$

F(3) = $\frac{1}{8}$ + x $-\infty$ + $\frac{3}{8}$ + $\frac{2}{8}$ = 1

$\frac{2}{8}$

F(x) = 1 for $\frac{}{8}$ x ≥ 3

$\frac{2}{8}$

This may be written in a better way as follows:

F(x) = 00 for x < - l or F(x) = 0,

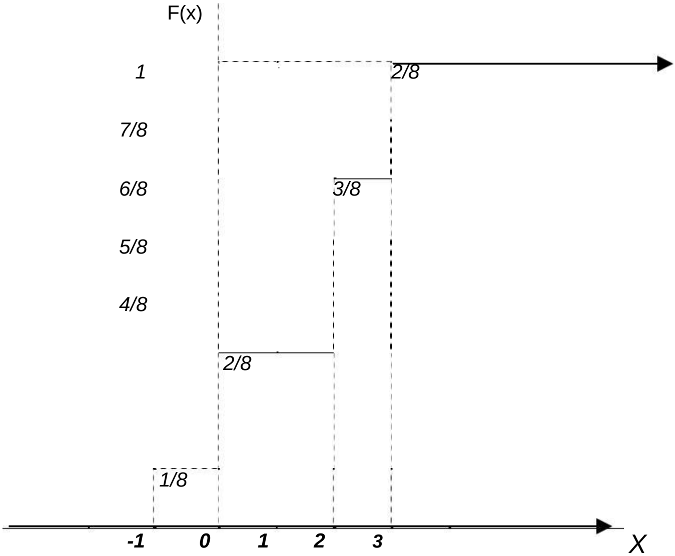
= $\frac{1}{8}$ for x = - l = $\frac{1}{8}$, - l ≤ x < 0

= $\frac{3}{8}$ for x = 0 = $\frac{3}{8}$, 0 ≤ x < 2

= $\frac{6}{8}$ for x = 2 = $\frac{6}{8}$, 2 ≤ x < 3

= 1 for x ≥ 3 = 1, 3 ≤ x < ∞

When F(x) is represented graphically we get a diagram as shown below since F(x) is a step function in this case. In general we can expect a step function for the distribution function when the variate is discrete.



Distribution function of a discrete variable

Example 13

A continuous r.v. X has the pdf……f(x) = $3x^2$; $0 \le x \le 1$.

Find and b such that

(i) $P(X \le a) = P(X > a)$

(ii) $P(X > b) = 0.05$

Solution

(i) $P(X \le a) = \int_0^a f(x)\,dx = \int_0^a 3x^2\,dx = a^3$

$P(X > a) = \int_a^1 f(x)\,dx = \int_a^1 3x^2\,dx = 1 \quad a^3$

$a^3 = 1 \quad a^3$

Statistics - Basic Statistics ¯

ie., $2a^3 = 1$

$a^3 = \frac{1}{2}$, $\therefore a = \left(\frac{1}{2}\right)^{1/3} = 0.79$

(ii) P(X > a) $= \int_b^1 3x^2\,dx = 1 b^3$

$1 \quad b^3 = 0.05$, $b^3 = 0.95$, b = **0.98**

Example 14

Verify that the following is a distribution function:

$$F(x) = 0 \qquad ; x < \quad a$$

$$= \frac{1}{2}\left(\frac{x}{a} + 1\right) \quad ; -a \le x \le a$$

$$= 1 \qquad ; x > a$$

Solution

Obviously the properties (i), (ii), (iii) and (iv) are satisfied. Also we observe that F(x) is continuous at x = a and x = $\quad a$ as well

Now , $\frac{d}{dx}F(x) = \frac{1}{2a}, -a \le x \le a$

0, otherwise

f(x) (say)

In order that F(x) is a distribution function, f(x) must be a p.d.f. Thus we have to show that

$$\int_{-\infty}^{\infty} f(x)dx = 1$$

$$\int_{-\infty}^{\infty} f(x)dx = \int_{-a}^{a} f(x)dx = \frac{1}{2a}\int_{-a}^{a} dx = 1$$

Hence F(x) is a d.f.

Example 15

Let the distribution function of X be

$$F(x) = 0 \qquad \text{if } x < \quad 1$$

$$= \frac{x+2}{4} \quad if - 1 \le x < 1$$

$$= 1 \text{ if } x \ge 1$$

Find P(X = 1)

Solution

We know that P(X = a) = F(a + 0) F(a 0)

$\therefore$P(X = 1) = F(1 + 0) F(1 0)

= 13/4

= **1/4**

EXERCISES

Multiple Choice questions

The outcomes of tossing a coin three time are a variable of the type

a. Continuous r.v b. Discrete r.v.

c. Neither discrete nor continuous

d. Discrete as well as continuous

The weight of persons in a country is a r.v. of the type

a. discrete b. continuous

c. neither a nor b d. both a and b

Let x be a continuous rv with pdf f(x) = kx, 0 ≤ x ≤1 ;

k, 1 ≤ x ≤ 2;

0 otherwise

The value of k is equal to b. 2/3

d. 3/4

a. 1/4 c. 2/5

Fill in the blanks

4. A r.v is a .... function
5. A continuous r.v can assume .... number of values with in the specified range

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

A continuous r.v.X has pdf f(x) = kx, o < x < 1, the value of k = ...

8. In terms of distribution function F(x), $\int_a^b f(x)\,dx$ = ...
9. The distribution function F(x) lies between ....

Very short answer questions

Define a random variable.

What are the two types of r.v.s?

Define probability mass function.

Define probability density function.

What are the axioms of pdf?

Define distribution function of a r.v.

Short essay questions

Define distribution function of a random variable and write down its properties.

What is the relationship between Distribution function and Density function?

State the properties of probability density function

What is random variable? Show by an example that any function defined over the sample space need not be a random variable.

Find the contract C such that the function

$$f(x) = cx^2 \ , 0 < x < 3$$
$$= 0 \ , \text{otherwise}$$

is a pdf and compute P(1 < x < 2)

A c.d.f. F(x) is defined as

$$F(x) = \begin{cases} 0 & , x \le 1 \\ \frac{1}{16}(x-1)^4 & , 1 < x \le 3 \\ 1 & , x > 3 \end{cases}$$

Find the p.d.f.

Long essay questions

A random variable X has pmf given by P(X = 1) = 1/2, P(X = 2) = 1/3 and P(X = 3) = 1/6. Write down the distribution function of X.

Find k and P(12 ≤ X ≤ 20) and P(X > 16) if following is the probability mass function of X.

X	8	12	16	20	24
f(x)	1/8	1/6	k/6	1/4	1/12

An r.v. Z takes the values –2, 0, 3 and 8 with probabilities $\frac{1}{12}, \frac{1}{2}, \frac{1}{6}$ *and* $\frac{1}{4}$ respectively. Write down its d.f. and find

P(Z ≥ 1) and P(1 Z < 8).

Define distribution function of a random variable. If X is a random variable with probability mass function

$$p(x) = \begin{cases} \frac{x^2}{30} & , x = 1, 2, 3, 4 \\ 0 & , otherw\ ise \end{cases}$$

Write down the distribution function of X.

Two coins are tossed. X represents the number of heads produced. Determine the probability distribution and the distribution function of X.

Examine whether the following can be a p.d.f. If so find k and P $(2 \le x \le 3)$

$$f(x) = {}^{\wedge}\left(\frac{x}{3}+\frac{1}{2}\right),\ 2 \le x \le 4 \text{ and 0 elsewhere.}$$

28. Obtain the distribution function F(x) for the following p.d.f.

$$f(x) = \begin{cases} x/3, & 0 < x \le 1 \\ \frac{5}{27}(4-x), & 1 \le x < 4 \\ 0, & otherw\ ise \end{cases}$$

For the p.d.f. $f(x) = 3a\,x^2,\ 0 \le x \le a$ find a and P(X

≤ 1/2 / 1/3 ≤ X ≤ 2/3)

Examine whether

$$f(x) = \begin{cases} 0, & x < 2\ ,\ or\ x > 4 \\ \frac{x}{9}+\frac{1}{6}, & 2 \le x \le 4 \end{cases}$$

is a p.d.f. If so, calculate P(2 < X < 3)

31.If $f(x) = \begin{cases} \frac{3}{4}(1-x^2) & if -1 \le x \le 1 \\ 0, & otherw\ ise \end{cases}$

Show that $F(x) = \begin{cases} 0\ , & x \le -1 \\ \frac{1}{2}+\frac{3}{4}x-\frac{1}{4}x^3\ , & -1 \le x \le 1 \\ 1, & x \ge 1 \end{cases}$

CHANGE OF VARIABLE

Change of variable technique is a method of finding the distribution of a function of a random variable. In many probability problems, the form of the density function or the mass function may be complex so as to make computation difficult. This technique will provide a compact description of a distribution and it will be relatively easy to compute mean, variance etc.

We will illustrate this technique by means of examples separately for discrete and continuous cases. Here we mention only the univariate case.

Suppose that the random variable X take on three values -1, 0 and 1 with probabilities 11/32, 16/32 and 5/32 respectively. Let us transform the random variable X, taking Y = 2X + 1.

The random variable Y can also take on values -1, 1 and 3 respectively, where

$$\begin{aligned} P(Y = -1) &= P(2X + 1 = -1) &&= P(X = -1) &&= 11/32 \\ P(Y = 1) &= P(2X + 1 = 1) &&= P(X = 0) &&= 16/32 \\ P(Y = 3) &= P(2X + 1 = 3) &&= P(X = 1) &&= 5/32 \end{aligned}$$

Thus the probability distribution of Y is

Y	1	1	3	Total
p(y)	11/32	16/32	5/32	1

In the above example, if we transform, the r.v. X as $Y = X^2$, the possible values of Y are 0 and 1.

$$\begin{aligned} \text{Therefore,}P(Y = 0) &= P(X = 0) = 16/32 \\ P(Y = 1) &= P(X = -1 \text{ or } X = 1) \\ &= P(X = -1) + P(X = 1) \\ &= 11/32 + 5/32 = 16/32 \end{aligned}$$

Thus the probability distribution of Y is

Y	0	1	Total
p(y)	16/32	16/32	1

3. A r.v. X has the density $f(x) = \frac{x+2}{6}$, 0 < X < 2

Let (gx) $= \begin{cases} 0 & if \quad 0 < x \le 1 \\ 1 & if \quad 1 < x \le 3/2 \\ 2 & if \quad x > 3/2 \end{cases}$

We can find the probability mass function

P{g(x)= 0} = P(0 < x ≤ 1) = $\int_0^1 f(x)\,dx$

$$= \int_0^1 \frac{x+2}{6}\,dx = \frac{1}{6}\left(\frac{x^2}{2} + 2x\right)_0^1$$

$$= \frac{1}{6}\left(\frac{1}{2} + 2\right)_0^1$$

$$= \frac{1}{6}\left(\frac{1}{2} + 2\right) = \frac{1}{6} \times \frac{5}{2} = \frac{5}{12} = \frac{20}{48}$$

P{g(x) = 1} = P(1 < x ≤ 3/2) = $\int_1^{3/2} f(x)\,dx$

$$= \int_1^{3/2} \frac{x+2}{6}\,dx = \frac{1}{6}\left(\frac{x^2}{2} + 2x\right)_1^{3/2}$$

$$= \frac{1}{6}\left(\frac{9}{8} + 3\right) - \frac{1}{6}\left(\frac{1}{2} + 2\right) = \frac{13}{48}$$

P{g(x) = 2} = P{X > 3/2} = $\int_{3/2}^{2} f(x)\,dx$

$$= \int_{3/2}^{2} \frac{X+2}{6}\,dx = \frac{1}{6}\left(\frac{x^2}{2} + 2x\right)_{3/2}^{2}$$

$$= \frac{1}{6}(2 + 4) - \frac{1}{6}\left(\frac{9}{8} + 3\right) = \frac{15}{48}$$

The probability distribution of g(x) is

g(x)	0	1	2		Total
P[g(x)]	20/48	13/48	15/48		1

If X is a2continuous0<<r.1v. with pdf

f(x) = 0h

Define Y = 3X + 1

Le G(y) be the distribution function of Y.

Then G(y) = P(Y ≤ y) $= \int_0^{\frac{y-1}{3}}$ X

P(3XX≤+ 1 ≤ y)

$\int$ f(x)dx = $\int$ 2x dx = $\left(\frac{y-1}{3}\right.$ 1=if0< 2 0y−1 3

ie., G(y) = (X − 1) if 0 < y < 4

= 0 ; otherwise

= 0,

The density function of y is given by

G(y) = G’(y) = $\frac{2}{9}$ y y

otherwise

$\frac{1}{9}$ y

Remark

From the above examples, we can observe that we can determine the probability distribution of Y from the probability distribution of X directly.

Let X be a r.v. defined on the sample space S. Let Y = g(X) be a single valued and continuous transformation of X. Then g(x) is also a r.v. defined on S. Now we are interested in determining the pdf of the new r.v. Y = g(X) given the pdf of the r.v. X

Result

Let X be a continuous r.v. with pdf f(x). Let Y = g(X) be strictly monotone (increasing or decreasing) function of X. Assume g(x) is differentiable for all x. Then the pdf of the r.v. Y is given by

f(y) = f(x) $\left|\frac{dx}{dy}\right|$, where x is expressed in terms of y.

Example 1

Let X be a continuous r.v. with pdf f(x). Let Y = X2. Find the pdf and the distribution function of Y?

Solution

Let X be a continuous r.v. with pdf f(x) and the distribution function F(x). Let Y = X2. Let G(y) be the distribution function of Y and g(y) its pdf.

Then G(y) = P(Y≤ y)

= P(X2 ≤ y) = P { |X| ≤ $\sqrt{y}$}

= P $(-\sqrt{y} \leq X \leq \sqrt{y})$

Now g(y) = G′(y)

= F $(\sqrt{y}) - F(-\sqrt{y})$ ....(1)

= $\frac{1}{2\sqrt{y}}$ F′($\sqrt{y}$) + $\frac{F'(-\sqrt{y})}{2\sqrt{y}}$

= $\frac{1}{2\sqrt{y}}$ [F′($\sqrt{y}$) + F′($-\sqrt{y}$)]

= $\frac{1}{2\sqrt{y}}$ [f ($\sqrt{y}$) + f (−$\sqrt{y}$)] .....(2)

Note 1

If, however, the random variable X is of the discrete type the distribution function of y is given by

$$G(y) = \begin{cases} F(\sqrt{y}) - F(-\sqrt{y}) - P(X = -\sqrt{y}) & \text{if } y > 0 \\ 0 & \text{if } y \le 0 \end{cases} \quad \ldots\ldots (3)$$

If the point $-\sqrt{y}$ is not a jump point of the r.v.X. then $P(X = -\sqrt{y}) = 0$ and the above result becomes identical with the given above.

Note 2

Let x_1, x_2 ... be the jump points of the r.v. X and y_1, y_2... be the jump points corresponding to the r.v. Y according to the relation $y_i = x_i^2$.

Then $P(Y = y_i) = P(X^2 = y_i) = P(X = \sqrt{y_i}) + P(X = -\sqrt{y_i})$

Example 2

A r.v. X has density $f(x) = Kx^2e^{-x^3}$, x > 0. Determine K and the density of $Y = X^3$

Solution :

Given $f(x) = Kx^2e^{-x^3}$, x > 0.

We know that $\int_{-\infty}^{\infty} f(x)\,dx = 1$

$$\int_0^\infty Kx^2e^{-x^3}dx = 1$$

Put $x^3 = t$

$3x^2dx = dt$

$x^2dx = \frac{dt}{3}$

$$K\int_0^\infty \frac{1}{3}e^{-t}dt = 1$$

$$\frac{K}{3}\left(\frac{e^{-t}}{-1}\right)_0^\infty = 1$$

$$-\frac{K}{3}(0-1) = 1$$

$\therefore K = 3$

The pdf of y is given by $3x \quad e$.

f (y) = $f(\ ) \left|\frac{dx}{dy}\right|$

$^2 \quad -x^3 \quad \frac{1}{3x^2}$

= e^{-x^3}

= e^{-y} , $y > 0$

Example 3

a. If X has a uniform distribution in [0, 1] with pdf.

f(x) = 1, 0 ≤ x ≤ 1

= 0, otherwise

b. If X has a standard, cauchy distribution with p.d.f.f(x)=,+∞<X<∞ Find a p.d.f. for X^2

a d ca

$\frac{1}{\pi}\ \frac{1}{1+x^2}$

Solution

Let Y = -2 log X. Then the distribution function G of Y is

G(y) = P(Y ≤ y) = P(-2 logX ≤ y)

= P(logX ≥ - y/2) = P(X ≥ $e^{-y/2}$)

= 1- P(X ≤ $e^{-y/2}$)

: 0 ∫ $1.\,dx$

∫ f(x)dx

= 1- $\int_0^{e^{-y/2}} f(x)\,dx$ = 1 - $e^{-y/2}$ = 1- $e^{-y/2}$

∵ $= -2$

g(y) = $\frac{d}{dy} G(y)$ = $\frac{1}{2} e^{-y/2}$ < y < ∞ ...(i)

[*as X ranges in* [0, 1], $Y(\ ) =$ log X ranges from 0 to ∞]

Note

is the p.d.f of a chi-square distribution with 2 degrees of freedom

b. Let Y = X2

G(y) = P(Y ≤ y) = P $\left(X^2 \leq y\right) = P\left(-\sqrt{y} \leq X \leq \sqrt{y}\right)$

$= \int_{-\sqrt{y}}^{\sqrt{y}} f(x)\, dx = 2\int_0^{\sqrt{y}} \frac{1}{\pi}\frac{dx}{1+x^2}$

[∵ integrand is an even function of x.]

$= \frac{2}{\pi}\left(\tan^{-1} x\right)_0^{\sqrt{y}} = \frac{2}{\pi}\tan^{-1}\left(\sqrt{y}\right)$

The pdf g(y) of Y is given by

g(y) $= G^1(y) = \frac{2}{\pi}\frac{1}{1+y}\frac{1}{2\sqrt{y}}$

$= \frac{1}{\pi}\frac{y^{-1/2}}{1+y}$; $0 \leq y < \infty$

[which is a beta distribution of second kind]

EXERCISES

Multiple choice question

If a r.v. x has pdf

f(x) = 3x, 0 < x < 1 = 0, otherwise

a. f(y) = $\frac{3}{16}(y-3)$ b. f(y) = $\frac{3}{4}(3-y)$

c. f(y) = $\frac{3}{16}(3-y)$ d. f(y) = $\frac{3}{4}(3-y)$

2. The distribution type of the variable y = $-2\sum_{i=1}^{n}\log X_i$ is same as that of the variable

a. 2 logx_i b. 2 log$\left(\pi_{i=1}^{n} x_i\right)^{-1}$

c. both a and b d. neither a nor b

If X is a r.v with distribution function F(x), then P(X^2 ≤ y) is

a. $P(-\sqrt{y} \le X \le \sqrt{y})$ b. $P(X \le \sqrt{y}) - P(X \le -\sqrt{y})$

c. $F(\sqrt{y}) - F(-\sqrt{y})$ d. All the above.

Very short answer questions:

Define monotone increasing function.

Define monotone decreasing function.

What do you mean by change of variable technique?

Short essay questions:

Let X have the density (fx) = 1, 0 < x < 1. Find the p.d.f. of Y = e^x.

If X has uniform distribution in (0, 1) with p.d.f. f(x) = 1, 0 < x < 1, find the p.d.f. of Y = 2 log X.

Let X be a rv with probability distribution

x	:	1	0	1
p(x)	:	$\frac{11}{32}$	$\frac{1}{2}$	$\frac{5}{32}$

Find the probability function of X^2 and $X^2 + 2$

If the probability mass function (p.m.f.) of X is

$$g(x) = \begin{cases} k(x^2+2), & x = 1, 2, 3 \\ 0, & \text{otherwise} \end{cases}$$

find k. Also obtain the p.m.f. of Y = X^2 + 1

Long essay questions:

If X has the p.d.f. f(x) = e^{-x}, x > 0 find the p.d.f. of Y = e^{-x}.

If the p.d.f. of X is

$$f(x) = \begin{cases} 1/2 & \text{if } -1 < x < 1 \\ 0, & \text{otherwise} \end{cases}$$

What is the p.d.f. of Y = X^2

f(x) = k(x + 2), 1 ≤ x ≤ 5 is the p.d.f. of a r.v. Find k and hence find the p.d.f of Y = X^2.

14. $f(x) = \begin{cases} \frac{1}{2}, & 0.9 < x < 1 \\ 0, & \text{otherwise} \end{cases}$

Determine g(y) if Y = $2X^2$

15. X has the p.d.f. $f(x) = \begin{cases} \frac{kx^3}{(1-2x)^{\upsilon}}, & x > 0 \\ \upsilon & \\ & \text{, elsewhere.} \end{cases}$

Determine k and also the density function of Y = $\frac{2X}{1+2X}$

16. Given $f(x) = \begin{cases} \frac{x^2}{9}, & 0 < x < 3 \\ \upsilon & \\ & \text{, otherwise} \end{cases}$

Find the density of Y = X^3