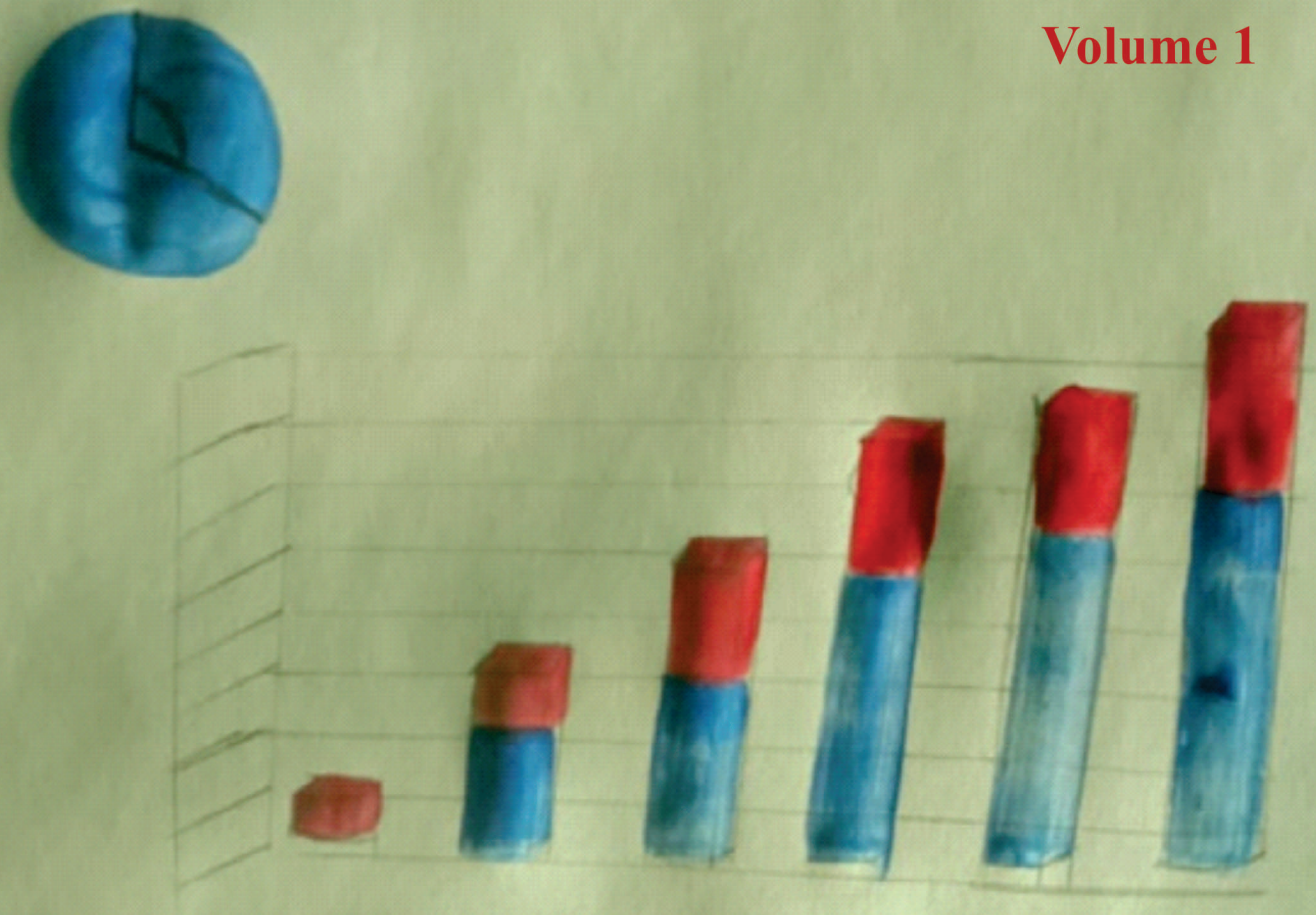
Women University Press
Jayoti Publication Desk
ISBN No. : 978-93-90892-43-3

Theory of Biostatistics

Volume 1



JV'n Dr. Vishal Saxena

JAYOTI VIDYAPEETH WOMEN'S UNIVERSITY, JAIPUR
UGC Approved Under 2(f) & 12(b) | NAAC Accredited | Recognized by Statutory Councils

Printed by :
JAYOTI PUBLICATION DESK

Published by :
Women University Press
Jayoti Vidyapeeth Women's University, Jaipur

Faculty of Agriculture & Veterinary Science

Title: Theory of Biostatistics Vol-1

Author Name Dr. Vishal Saxena

Published By: Women University Press

Publisher's Address: Jayoti Vidyapeeth Women's University, Jaipur
Vedaant Gyan Valley,
Village-Jharna, Mahala Jobner Link Road, NH-8
Jaipur Ajmer Express Way,
Jaipur-303122, Rajasthan (INDIA)

Printer's Detail: Jayoti Publication Desk

Edition Detail: I

ISBN: 978-93-90892-43-3

INDEX

CHAPTER 1. INTRODUCTION AND STATISTICS OF DATA

STATISTICS

Statistics is a form of mathematical analysis that uses quantified models, representations and synopses for a given set of experimental data or real-life studies. Statistics studies methodologies to gather, review, analyze and draw conclusions from data.

Statistics is a term used to summarize a process that an analyst uses to characterize a data set. If the data set depends on a sample of a larger population, then the analyst can develop interpretations about the population primarily based on the statistical outcomes from the sample. Statistical analysis involves the process of gathering and evaluating data and then summarizing the data into a mathematical form.

Statistics is used in various disciplines such as psychology, business, physical and social sciences, humanities, government, and manufacturing. Statistical data is gathered using a sample procedure or other method. Two types of statistical methods are used in analyzing data: descriptive statistics and inferential statistics. Descriptive statistics are used to synopsize data from a sample exercising the mean or standard deviation. Inferential statistics are used when data is viewed as a subclass of a specific population.

BIOSTATISTICS

Biostatistics is the application of statistics to a variety of topics in biology. In this course, we tend to focus on biological topics in the health sciences as we learn about statistics.

In an introductory course such as ours, there is essentially no difference between "biostatistics" and "statistics" and thus you will notice that we focus on learning "statistics" in general but use as many examples from and applications to the health sciences as possible.

DATA

Data can be defined as a systematic record of a particular quantity. It is the different values of that quantity represented together in a set. It is a collection of facts and figures to be used for a specific purpose such as a survey or analysis. When arranged in an organized form, can be called information. The source of data (primary data, secondary data) is also an important factor.

TYPES OF DATA

Data may be qualitative or quantitative. Once you know the difference between them, you can know how to use them.

- Qualitative Data: They represent some characteristics or attributes. They depict descriptions that may be observed but cannot be computed or calculated. For example, data on attributes such as intelligence, honesty, wisdom, cleanliness, and creativity collected using the students of your class a sample would be classified as qualitative. They are more exploratory than conclusive in nature.

- Quantitative Data: These can be measured and not simply observed. They can be numerically represented and calculations can be performed on them. For example, data on the number of students playing different sports from your class gives an estimate of how many of the total students play which sport. This information is numerical and can be classified as quantitative.

DATA COLLECTION

Depending on the source, it can classify as primary data or secondary data. Let us take a look at them both.

PRIMARY DATA

These are the data that are *collected for the first time* by an investigator for a specific purpose. Primary data are 'pure' in the sense that no statistical operations have been performed on them and they are original. An example of primary data is the Census of India.

SECONDARY DATA

They are the data that are *sourced from someplace* that has originally collected it. This means that this kind of data has already been collected by some researchers or investigators in the past and is available either in published or unpublished form. This information is impure as statistical operations may have been performed on them already. An example is an information available on the Government of India, the Department of Finance's website or in other repositories, books, journals, etc.

CLASSIFICATION OF DATA

The process of arranging data into homogenous groups or classes according to some common characteristics present in the data is called classification.

For example: During the process of sorting letters in a post office, the letters are classified according to the cities and further arranged according to streets.

BASES OF CLASSIFICATION

Classification can be divided in following four bases:

(1) QUALITATIVE BASE

When the data are arranged by qualitative characteristics such as sex, literacy and intelligence, etc.

(2) QUANTITATIVE BASE

When the data are classified by quantitative characteristics like height, weight, age, income, etc.

(3) GEOGRAPHICAL BASE

When the data are classified by geographical regions or location, like states, provinces, cities, countries, etc.

(4) CHRONOLOGICAL OR TEMPORAL BASE

When the data are classified or arranged by their time of occurrence, such as years, months, weeks, days, etc.

For example: Time series data.

TABULATION OF DATA

The process of placing classified data into tabular form is known as tabulation. A table is a symmetric arrangement of statistical data in rows and columns. Rows are horizontal arrangements whereas columns are vertical arrangements. It may be simple, double or complex depending upon the type of classification.

TYPES OF TABULATION

(1) SIMPLE TABULATION OR ONE-WAY TABULATION

When the data are tabulated to one characteristic, it is said to be a simple tabulation or one-way tabulation.

For example: Tabulation of data on the population of the world classified by one characteristic like religion is an example of a simple tabulation.

(2) DOUBLE TABULATION OR TWO-WAY TABULATION

When the data are tabulated according to two characteristics at a time, it is said to be a double tabulation or two-way tabulation.

For example: Tabulation of data on the population of the world classified by two characteristics like religion and sex is an example of a double tabulation.

(3) COMPLEX TABULATION

When the data are tabulated according to many characteristics, it is said to be a complex tabulation.

For example: Tabulation of data on the population of the world classified by three or morecharacteristics like religion, sex and literacy, etc. is an example of a complex tabulation.

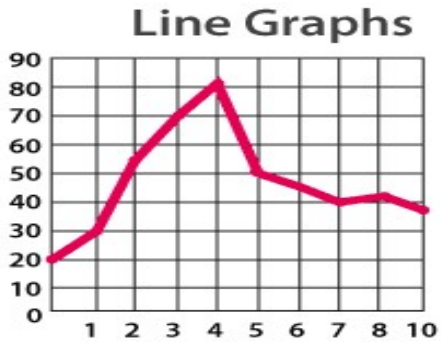
CHAPTER 2. GRAPHICAL REPRESENTATION OF DATA

Graphical Representation is a way of analysing numerical data. It exhibits the relation between data, ideas, information and concepts in a diagram. It is easy to understand and it is one of the most important learning strategies. It always depends on the type of information in a particular domain. There are different types of graphical representation. Some of them are as follows

(1) LINE GRAPHS

Linear graphs are used to display the continuous data and it is useful for predicting the future events over time.

Exp.



(2) BAR GRAPHS

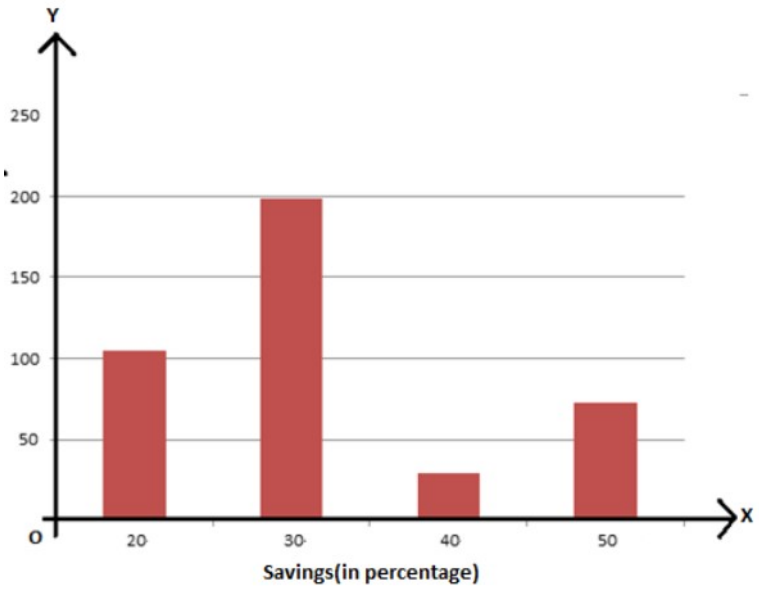
Bar Graph is used to display the category of data and it compares the data using solid bars to represent the quantities.

Exp.

In a firm of 400 employees, the percentage of monthly salary saved by each employee is given in the following table. Represent it through a bar graph.

Savings (in percentage)	20	30	40	50
Number of Employees(Frequency)	105	199	29	73

Sol.



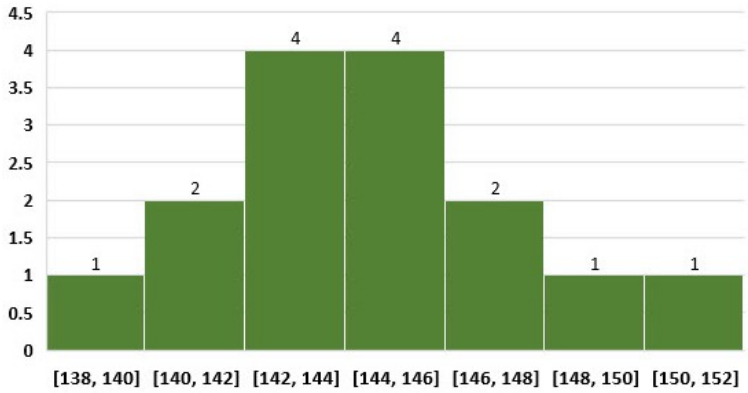
(3) HISTOGRAMS

The graph that uses bars to represent the frequency of numerical data that are organised into intervals. Since all the intervals are equal and continuous, all the bars have the same width.

Exp.

Mr. Larry, a famous doctor, is researching the height of the students studying in the 8^{th} standard. He has gathered a sample of 15 students but wants to know which the maximum category is where they belong.

S.N.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Height	141	143	145	145	147	152	143	144	149	141	138	143	145	148	145



Here we can see the heights of the students on an average is in the range of 142 cm to 146 cm for 8th standard.

(4) LINE PLOT

It shows the frequency of data on a given number line. ‘ x ‘ is placed above a number line each time when that data occurs again.

Exp.

(5) CIRCLE GRAPH

Also known as pie chart that shows the relationships of the parts of the whole. The circle is considered with 100% and the categories occupied is represented with that specific percentage like 15%, 56% , etc.

Exp.

Let's construct a pie chart to visually display the favorite fruits of the students in the class based on the frequency table below.

Mango	Orange	Plum	Pineapple	Melon
45	30	15	30	30

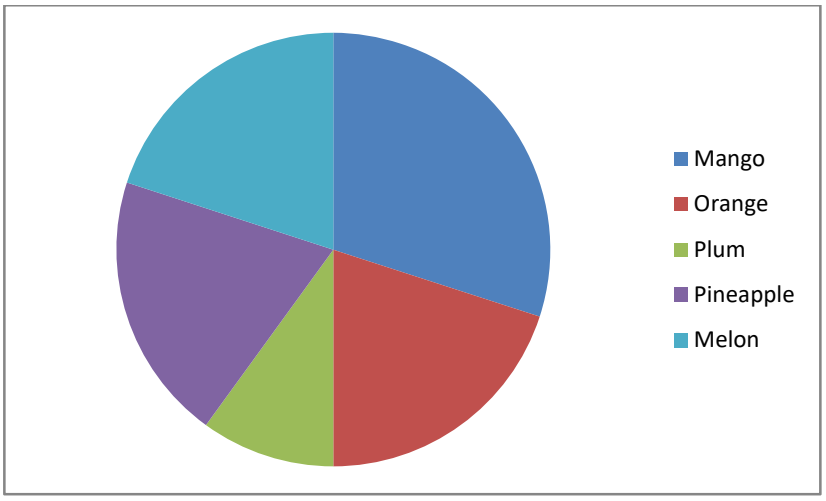
Category	Formula	Degrees
Mango	$45/150 \times 360$	108
Orange	$30/150 \times 360$	72
Plum	$15/150 \times 360$	36
Pineapple	$30/150 \times 360$	72
Melon	$30/150 \times 360$	72

The total frequency sums up to 150.

Draw a circle and draw the radius. With the radius as the base, construct 108^0 using a protractor**.**

Subsequently, construct all other sectors with their respective angles.

Thus we get the pie chart as follows.



CHAPTER 3. MEASURE OF CENTRAL TENDENCY

MEASURE OF CENTRAL TENDENCY

Measures of central tendency are the numbers which indicate the centre of a set of ordered numerical data. The most common measures of central tendency are the mean, median, and mode.

(A) ARITHMETIC MEAN

Arithmetic mean is the simple average of all items in a series. It is the simplest measure of central tendencies.

Basic formula: Arithmetic mean $= X_1 + X_2 + X_3 + \ldots\ldots + Xn / n = \sum X / n$

Exp. For the numbers 5, 6, 7, 8, 9

Arithmetic mean = (5 + 6 + 7 + 8 + 9) / 5 = 35 /5=7

METHODS OF CALCULATING SIMPLE ARITHMETIC MEAN

The three types of statistical series are

1. Individual series
2. Discrete series (Simple frequency distribution)
3. Continuous series (Grouped frequency distribution)

1. CALCULATION OF SIMPLE ARITHMETIC MEAN FOR INDIVIDUAL SERIES

For the individual series, arithmetic mean is calculated by

Mean=$\sum X / N$ = Total value of the items / No. of items

Exp.

Suppose the pocket allowance of 10 students in rupees are

15,20,30,22,25,18,40,50, 55,65.

Find out the average pocket allowance.

Sol.

Arithmetic mean = $\sum X/n$

= 15+20+30+22+25+18+40+50+55+65/10 = 340/10 = 34

Therefore, the average pocket allowance is = Rs 34

2. CALCULATION OF ARITHMETIC MEAN IN DISCRETE SERIES OR SIMPLE FREQUENCY DISTRIBUTION

i. Direct method

ii. Short-cut method

i. Direct method

Formula:- = $\sum fX / \sum f$

Exp. Following are the weekly wage earnings of 19 workers:

Wages (Rs) (X): 10 20 30 40 50

No. of workers (f): 4 5 3 2 5

Sol. Mean = $\sum fX / \sum f$ = 560/19 = 29.47

Therefore mean wage earnings = Rs 29.47

ii. Short-cut method

In short-cut method, the following formula is to be applied to find mean

Mean =A+ $\sum fd / \sum f$

where A is the assumed mean, f denotes frequency and d=X-A (called deviation)

Exp. Find the mean for the following data

x:	900	950	1000	1100	1260	1440	1500
f:	26	22	18	19	15	3	2

Sol.

x	f	d = x-A	fd
900	26	-100	-2600
950	22	-50	-1100
1000=A(Let)	18	0	0
1100	19	100	1900
1260	15	260	3900
1440	3	440	1320
1500	2	500	1000
	$\sum f$ =105		$\sum fd$

Therefore, mean = $A + \frac{\sum fd}{\sum f} = 1000 + \frac{4420}{105} = 1042.1$

3. CALCULATION OF SIMPLE ARITHMETIC MEAN IN CASE OF CONTINUOUS SERIES OR GROUPED FREQUENCY DISTRIBUTION

i. Direct method

ii. Short-cut method

iii. Step-deviation method

i. **Direct method Formula:**

$M = \sum fX / \sum f$

ii. **Short-cut method Formula**:

$M = A + (\sum fd / \sum f)$

where A is the assumed mean, f denotes frequency and d=X-A (called deviation).

iii. **Step-deviation method:** Formula:-

$Mean = A + (\sum f.u / \sum f)\ h$

where u = (X-A) / h=d / h, A is the assumed mean, h is the width of the class interval.

Exp. (By Direct method Formula):

Marks in Statistics of student of Class XI are given below. Find out arithmetic mean.

Sol. :

Marks:	0-10	10-20	20-30	30-40	40-50
No. of students :	5	12	14	10	8
Mid-value:	5	15	25	35	45
fx:	25	180	350	350	360

Mean= $\sum fX / \sum f = 1265/49 = 25.82$

Arithmatic mean=25.82 marks

Exp. (By Step-deviation method):

Find the mean for the following data:

Class-Interval	100-150	150-200	200-250	250-300	300-350	350-400	400-450	450-500
Freq.	24	40	33	28	30	22	16	7

Sol.:

Class-Interval	Freq.	x mid-value	d = x-A	u = d/h	f.u
100-150	24	125	-150	-3	-72
150-200	40	175	-100	-2	-80
200-250	33	225	-50	-1	-33
250-300	28	275=A(let)	0	0	0
300-350	30	325	50	1	30
350-400	22	375	100	2	44
450-450	16	425	150	3	48
450-500	7	475	200	4	28
	$\sum f$ =200				$\sum f.u$ = -35

Therefore, mean = A+ (∑f.u / ∑f) h

$$= 275 + \frac{-35}{200} \times 50 = 266.25$$

(B) MEDIAN:

Median is a centrally located value in a series for which half of the values (or items) of the series are above it and rest of them are below it.

OR

Median is the central value of the variable which divides the series into two equal parts such that half of the values (or items) of the series are above it and rest of them are below it. Median is defined as the value of the middle most term (or the mean of the values of the two middle terms) when the data are arranged in an ascending or descending order of magnitude.

METHODS OF CALCULATING MEDIAN

We know, there are three types of statistical series :

1. Individual series
2. Discrete series (Simple frequency distribution)
3. Continuous series (Grouped frequency distribution)

1. CALCULATION OF MEDIAN FOR INDIVIDUAL SERIES

The "Median" of a individual data set is dependent on whether the number of elements in the data set is odd or even. First reorder the data set from the smallest to the largest (i.e. ascending order).

For Odd series:

Formula:- Median = (n+1)/2th item

For even series:

Formula:- Median=Average of [(n/2)th item+(n/2+1)th item]

Exp. For odd Number of Elements

Data Set= 2, 6, 9, 3, 5, 4, 7

Reordered = 2, 3, 4, 5, 6, 7, 9

Median = (7+1)/2th item = 4^{th} item = 5

Exp. For even Number of Elements

Data Set = 2, 6, 9, 3, 5, 4 Reordered = 2, 3, 4, 5, 6, 9

Median = (3^{rd} item + 4^{th} item) / 2 = (4+5)/2=4.5

2. CALCULATION OF MEDIAN FOR DISCRETE SERIES OR SIMPLE FREQUENCY DISTRIBUTION:

The "Median" of a discrete data set is dependent on whether the number of elements in the data set is odd or even. First reorder the data set from the smallest to the largest (i.e. ascending order).

The median is given by

For Odd series:

Formula:- Median = Size of (N+1)/2th item

For even series:

Formula:- Median=Average of [Size of (N/2)th item + Size of (N/2+1)th item]

Exp. from the following data calculate median

Marks	45	55	25	35	5	15
No. of students	40	30	30	50	10	20

Sol.:

Step I- First we will find out the commutative frequency

Marks in ascending order (x) :	5	15	25	35	45	55
No. of students (f)	10	20	30	50	40	30

Commutative frequency C.f :10 30 60 110 150 180

N = 180

N=180=even

Median = [size of (N/2)th item + size of (N/2+1)th item]/2

Step II – [Size of 90^{th} item + size of 91^{st} item]/2

=(35+35)/2

Median = 35

3. CALCULATION OF MEDIAN FOR CONTINUOUS SERIES OR GROUPED FREQUENCY DISTRIBUTION

In the case of grouped series, the median is calculated with the fallowing formula:

M=L+ [(N/2)-p.c.f.]/f x i

Where L= lower limit of median class interval (MCI)

p.c.f.=previous cumulative frequency of median class

f=frequency of median class

i= size of median class

N=total no of observation

Exp. From the following data, find median

Marks (x)	0-10	10-20	20-30	30-40	40-50	50-60
No. of Students (f)	10	20	30	50	40	30

Sol.:

I Commutative Frequency Table is

Marks (x)	No. of Students (f)	Cumulative frequency (cf)
0-10	10	10
10-20	20	30
20-30	30	60=p.c.f.
30-40 = MCI	50=f	110
40-50	40	150
50-60	30	180
	N=180	

II Size of N /2 item = size of 180/2 item = 90th item

III - Commutative frequency which includes 90^{th} item = 110 Class corresponding to 110 = 30-40, which is the median class

Now we applying the fallowing formula

M=L+ [(N/2)-p.c.f.]/f x i

M=30+[(90-60)/50]x10

Median = 36

(C). MODE

Mode is the value in a series which occurs most frequently or which has the greatest frequency. But it is not exactly true for every case or for every frequency distribution. it is that value around which the terms tend to concentrate most densely.

METHODS OF CALCULATING MEDIAN

Mode can be determined in the following three types of statistical series :

1. Individual series

2. Discrete series (Simple frequency distribution)

3. Continuous series (Grouped frequency distribution)

1. CALCULATION OF MODE FOR INDIVIDUAL SERIES

Exp. Find the mode of the following series: 8, 9, 11, 15, 16, 12, 15, 3, 7, 15

Sol. There are ten observations in the series, where the data 15 occurs maximum number of times (highest frequency). Therefore the mode is 15.

2. CALCULATION OF MODE FOR DISCRETE SERIES

EXP. Find the mode of the following series:

x:	**2**	**5**	**7**	**11**	**18**
f:	**5**	**12**	**19**	**7**	**5**

Sol.: There are five observations in the series, where the data 7 occurs maximum number of times (highest frequency). Therefore the mode is 7.

3. CALCULATION OF MODE FOR CONTINUOUS SERIES

In the case of grouped data, mode is determined by the following formula:

$$\text{Mode} = L + [(f_1-f_0)/(2f_1-f_0-f_2)] \times h$$

where

L = Lower limit of the modal class.

f_1 = modal class frequency

f_0 = frequency of preceding the modal class

f_2 = frequency of succeeding the modal class

h= width of modal class

Example- calculate the modal sales of the 100 companies from the following data

Sales in Rs(lakhs)	58-60	60-62	62-64	64-66	66-68	68-70	70-72
No. of companies	12	18	25	30	10	3	2

Solution-

Sales in Rs(lakhs)	No. of companies
58-60	12
60-62	18
62-64	25
64-66	30
66-68	10
68-70	3
70-72	2

→ **Modal class**

Here the modal class interval is 64-66 (because it has the highest frequency)

Mode = $L + [(f_1-f_0)/(2f_1-f_0-f_2)] X h$

$=64+[(30-25)/60-25-10] X 2= 64.4$

RELATION BETWEEN MEAN, MEDIAN AND MODE

Mode = 3(median) – 2(mean)

CHAPTER 4. MEASURE OF DISPERSION

The measure of dispersion shows the scatterings of the data. It tells the variation of the data from one another and gives a clear idea about the distribution of the data. The measure of dispersion shows the homogeneity or the heterogeneity of the distribution of the observations.

(A) RANGE

A range is the most common and easily understandable measure of dispersion. It is the difference between two extreme observations of the data set. If X_{max} and X_{min} are the two extreme observations then

Range = $X_{max} - X_{min}$

Coefficient of Range = $(X_{max} - X_{min})/(X_{max} + X_{min})$

Exp. For data

88, 89, 89, 89, 90, 91, 91, 91, 92

Range= 92-88 = 4

Exp.

71, 83, 85, 86, 90, 95, 100, 100, 100

Range=100-71 = 29

MERITS OF RANGE

- It is the simplest of the measure of dispersion
- Easy to calculate
- Easy to understand
- Independent of change of origin

DEMERITS OF RANGE

- It is based on two extreme observations. Hence, get affected by fluctuations
- A range is not a reliable measure of dispersion

- Dependent on change of scale

(B) QUARTILE DEVIATION

The values of the variate that divide the total frequency into four parts are called Quartile. The i^{th} Quartile is given by

$$Q_i = l + \left(\frac{iN/4 - C}{f} \right) \times h$$

For i=1, Q_1 is called lower (or first quartile).

For i=2, Q_2 is called median (or second quartile).

For i=3, Q_3 is called upper (or third quartile).

The difference between the upper and lower quartile is called the inter quartile range.

Quartile Deviation defines the absolute measure of dispersion. Whereas the relative measure corresponding to QD, is known as the coefficient of QD.

The Quartile Deviation(QD) is the product of half of the difference between the upper and lower quartiles. Mathematically we can define as:

Quartile Deviation = $(Q_3 - Q_1) / 2$

Coefficient of Quartile Deviation = $(Q_3 - Q_1) / (Q_3 + Q_1)$

A Coefficient of QD is used to study & compare the degree of variation in different situations.

Exp. Find the Quartile Deviation from the given below data:

Marks	0-5	5-10	10-15	15-20	20-25	25-30
No. of students	4	6	8	12	7	2

Sol.: Cumulative frequency table is

Class Interval	No. of students	c.f.
0-5	4	4
5-10	6	10
10-15	8	18
15-20	12	30
20-25	7	37
25-30	2	39

$$N = \sum f = 39$$

$N = 39/4 = 9.75$

Class of Q_1 is 5-10, Now $Q_1 = l + \left(\frac{N/4 - C}{f}\right) \times h = 5 + \left(\frac{9.75 - 4}{6}\right) \times 5 = 9.79$

and 3N/4 =29.25, therefore class of Q_3 is 15-20,

$$Q_3 = l + \left(\frac{3N/4 - C}{f}\right) \times h = 15 + \left(\frac{29.25 - 18}{12}\right) \times 5 = 19.69$$

Quartile Deviation$= \left(\frac{Q_3 - Q_1}{2}\right) = 4.95$

(C) MEAN DEVIATION

Mean deviation is the arithmetic mean of the absolute deviations of the observations from a measure of central tendency. If $x_1, x_2, \ldots, x_n$ are the set of observation, then the mean deviation of x about the average A (mean, median, or mode) is

Mean deviation from average A = $1/n\ [\sum_i |x_i - A|]$

For a grouped frequency, it is calculated as:

Mean deviation from average A = $1/N\ [\sum_i f_i |x_i - A|]$, $N = \sum f_i$

Exp. Find the mean deviation about the mean for 6, 7, 10, 12, 13, 4, 8, 12

Sol. Mean is given by

$$\bar{x} = \frac{\sum x}{n} = \frac{72}{8} = 9$$

x	$x - \bar{x}$	$\left\|x - \bar{x}\right\|$
6	-3	3
7	-2	2
10	1	1
12	3	3
13	4	4
4	-5	5
8	-1	1
12	3	3
		$\sum\left\|x - \bar{x}\right\| = 22$

$$\text{Mean deviation} = \frac{\sum\left|x - \bar{x}\right|}{n} = \frac{22}{8}$$

$$= 2.75$$

MERITS OF MEAN DEVIATION

- Based on all observations
- It provides a minimum value when the deviations are taken from the median
- Independent of change of origin

DEMERITS OF MEAN DEVIATION

- Not easily understandable

- Its calculation is not easy and time
- Dependent on the change of scale
- Ignorance of negative sign creates artificiality and becomes useless for further mathematical treatment

(D) STANDARD DEVIATION

A standard deviation is the positive square root of the arithmetic mean of the squares of the deviations of the given values from their arithmetic mean. It is denoted by a Greek letter sigma, σ. It is also referred to as root mean square deviation.

VARIANCE

The square of the standard deviation is the **variance**. It is also a measure of dispersion.

METHODS TO FIND STANDARD DEVIATION

The standard deviation is given as

$$\sigma = [\Sigma_i (y_i - \bar{y})^2 / n]^{1/2} = [(\Sigma_i y_i^2 / n) - \bar{y}^2]^{1/2}$$

For a grouped frequency distribution, it is

$$\sigma = [\Sigma_i f_i (y_i - \bar{y})^2 / N]^{1/2} = [(\Sigma_i f_i y_i^2 / N) - \bar{y}^2]^{1/2}$$

METHODS TO FIND VARIANCE

For a individual series

$$\sigma^2 = [\Sigma_i (y_i - \bar{y})^2 / n] = [(\Sigma_i y_i^2 / n) - \bar{y}^2]$$

For a grouped frequency distribution, it is

$$\sigma^2 = [\Sigma_i f_i (y_i - \bar{y})^2 / N] = [(\Sigma_i f_i y_i^2 / N) - \bar{y}^2]$$

Exp. Find the Variance and Standard Deviation of the Following Numbers: 1, 3, 5, 5, 6, 7, 9, 10.

Sol.

Step 1: The mean = 46/ 8 = 5.75

Step 2: Deviations from mean: $(y_i - \bar{y})$=

(1 – 5.75), (3 – 5.75), (5 – 5.75), (5 – 5.75), (6 – 5.75), (7 – 5.75), (9 – 5.75), (10 – 5.75)

= -4.75, -2.75, -0.75, -0.75, 0.25, 1.25, 3.25, 4.25

Step 3: Squaring the above values we get $(y_i - \bar{y})^2$ =22.563, 7.563, 0.563, 0.563, 0.063, 1.563, 10.563, 18.063

Step 4:

$\Sigma_i (y_i - \bar{y})^2$ =22.563 + 7.563 + 0.563 + 0.563 + 0.063 + 1.563 + 10.563 + 18.063
= 61.504

Step 4:

n = 8, therefore variance $(\sigma^2) = \Sigma_i (y_i - \bar{y})^2 / n$ = 61.504/ 8 = 7.69

Now, Standard deviation (σ) = 2.77

and variance = (σ^2) = 7.69

MERITS OF STANDARD DEVIATION

- Squaring the deviations overcomes the drawback of ignoring signs in mean deviations
- Suitable for further mathematical treatment
- Least affected by the fluctuation of the observations
- The standard deviation is zero if all the observations are constant
- Independent of change of origin

DEMERITS OF STANDARD DEVIATION

- Not easy to calculate
- Difficult to understand for a layman
- Dependent on the change of scale

CHAPTER 5. CORRELATION AND REGRESSION

The word correlation is used in everyday life to denote some form of association. We might say that we have noticed a correlation between foggy days and attacks of wheeziness. However, in statistical terms we use correlation to denote association between two quantitative variables. We also assume that the association is linear, that one variable increases or decreases a fixed amount for a unit increase or decrease in the other. The other technique that is often used in these circumstances is regression, which involves estimating the best straight line to summarise the association.

CORRELATION ANALYSIS

Correlation analysis is applied in quantifying the association between two continuous variables, for example, an dependent and independent variable or among two independent variables.

CORRELATION COEFFICIENT

The degree of association is measured by a correlation coefficient, denoted by r. It is sometimes called Pearson's correlation coefficient after its originator and is a measure of linear association. If a curved line is needed to express the relationship, other and more complicated measures of the correlation must be used.

TYPES OF CORRELATION:

In a bivariate distribution, the correlation may be:

1. Positive, Negative and Zero Correlation; and
2. Linear or Curvilinear (Non-linear).

1. POSITIVE, NEGATIVE OR ZERO CORRELATION:

When the increase in one variable (X) is followed by a corresponding increase in the other variable (Y); the correlation is said to be positive correlation. The positive correlations range from 0 to +1; the upper limit i.e. +1 is the perfect positive coefficient of correlation.

The perfect positive correlation specifies that, for every unit increase in one variable, there is proportional increase in the other. For example “Heat” and “Temperature” have a

perfect positive correlation. If, on the other hand, the increase in one variable (X) results in a corresponding decrease in the other variable (Y), the correlation is said to be negative correlation.

The negative correlation ranges from 0 to – 1; the lower limit giving the perfect negative correlation. The perfect negative correlation indicates that for every unit increase in one variable, there is proportional unit decrease in the other.

Zero correlation means no relationship between the two variables X and Y; i.e. the change in one variable (X) is not associated with the change in the other variable (Y). For example, body weight and intelligence, shoe size and monthly salary; etc. The zero correlation is the mid-point of the range – 1 to + 1.

2. LINEAR OR CURVILINEAR CORRELATION:

Linear correlation is the ratio of change between the two variables either in the same direction or opposite direction and the graphical representation of the one variable with respect to other variable is straight line.

Consider another situation. First, with increase of one variable, the second variable increases proportionately upto some point; after that with an increase in the first variable the second variable starts decreasing.

METHODS OF COMPUTING COEFFICIENT OF CORRELATION:

1. KARL PEARSON'S CO-EFFICIENT OF CORRELATION:

The Karl Pearson's product-moment correlation coefficient (or simply, the Pearson's correlation coefficient) is a measure of the strength of a linear association between two variables and is denoted by *r* or r*xy*(x and y being the two variables involved). This method of correlation attempts to draw a line of best fit through the data of two variables, and the value of the Pearson correlation coefficient, *r*, indicates how far away all these data points are to this line of best fit.

The value of *r* always lies between +1 and -1. Depending on its exact value, we see the following degrees of association between the variables.

A value greater than 0 indicates a positive association i.e. as the value of one variable increases, so does the value of the other variable. A value less than 0 indicates a negative association i.e. as the value of one variable increases, the value of the other variable decreases.

KARL PEARSON CORRELATION COEFFICIENT FORMULA

The coefficient of correlation rxy between two variables x and y, for the bivariate dataset (xi,yi) where i = 1,2,3…..n; is given by

$$\rho_{X,Y} = \operatorname{corr}(X,Y) = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{\operatorname{E}[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y}$$

Where,

ρXY = Population correlation coefficient between X and Y

μX = Mean of the variable X

μY = Mean of the variable Y

σX = Standard deviation of X

σY = Standard deviation of Y

E = Expected value operator

Cov = Covriance

The above formulas can also be written as:

$$\rho_{X,Y} = \frac{\operatorname{E}(XY) - \operatorname{E}(X)\operatorname{E}(Y)}{\sqrt{\operatorname{E}(X^2) - \operatorname{E}(X)^2} \cdot \sqrt{\operatorname{E}(Y^2) - \operatorname{E}(Y)^2}}$$

The sample correlation coefficient formula is:

$$r_{xy} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n\sum x_i^2 - (\sum x_i)^2}\,\sqrt{n\sum y_i^2 - (\sum y_i)^2}}$$

The above are used to find the correlation coefficient for the given data. Based on the value obtained through these formulas, we can determine how much strong is the association between given two variables

Exp.

Table 5.2 Computation of r from Original Scores

X	Y	X^2	Y^2	XY
13	7	169	49	91
12	11	144	121	132
10	3	100	9	30
8	7	64	49	56
7	2	49	4	14
6	12	36	144	72
6	6	36	36	36
4	2	16	4	8
3	9	9	81	27
1	6	1	36	6
$\Sigma X = 70$	$\Sigma Y = 65$	$\Sigma X^2 = 624$	$\Sigma Y^2 = 533$	$\Sigma XY = 472$

$$\gamma_{xy} = \frac{N\Sigma XY - (\Sigma X)(\Sigma Y)}{\sqrt{[N\Sigma X^2 - (\Sigma X)^2][N\Sigma Y^2 - (\Sigma Y)^2]}}$$

$$= \frac{(10 \times 472) - (70 \times 65)}{\sqrt{(10 \times 624 - 4{,}900)(10 \times 533 - 4{,}225)}} = \frac{170}{\sqrt{1{,}340 \times 1{,}105}}$$

$$= \frac{170}{\sqrt{1{,}480{,}700}} = \frac{170}{1216.84} = +.14$$

2. SPEARMAN'S RANK ORDER CO-EFFICIENT OF CORRELATION

RANK CORRELATION

Sometimes there doesn't exist a marked linear relationship between two random variables but a monotonic relation (if one increases, the other also increases or instead, decreases) is clearly noticed. Pearson's Correlation Coefficient evaluation, in this case, would give us the strength and direction of the linear association only between the variables of interest. Herein comes the advantage of the Spearman Rank Correlation methods, which will instead, give us the strength and direction of the monotonic relation between the connected variables. This can be a good starting point for further evaluation.

THE SPEARMAN RANK-ORDER CORRELATION COEFFICIENT

The Spearman's Correlation Coefficient, represented by ρ or by r_R, is a nonparametric measure of the strength and direction of the association that exists between two ranked variables. It determines the degree to which a relationship is monotonic, i.e., whether there is a monotonic component of the association between two continuous or ordered variables.

Monotonicity is "less restrictive" than that of a linear relationship. Although monotonicity is not actually a requirement of Spearman's correlation, it will not be meaningful to pursue Spearman's correlation to determine the strength and direction of a monotonic relationship if we already know the relationship between the two variables is not monotonic.

SPEARMAN RANKING OF THE DATA

We must rank the data under consideration before proceeding with the Spearman's Rank Correlation evaluation. This is necessary because we need to compare whether on increasing one variable, the other follows a monotonic relation (increases or decreases regularly) with respect to it or not.

Thus, at every level, we need to compare the values of the two variables. The method of ranking assigns such 'levels' to each value in the dataset so that we can easily compare it.

- Assign number 1 to n (the number of data points) corresponding to the variable values in the order highest to lowest.
- In the case of two or more values being identical, assign to them the arithmetic mean of the ranks that they would have otherwise occupied.

The Formula for Spearman Rank Correlation

$$\rho = 1 - 6\frac{\sum_i d_i^2}{n\left(n^2 - 1\right)}$$

where n is the number of data points of the two variables and d_i is the difference in the ranks of the i^{th} element of each random variable considered. The Spearman correlation coefficient, ρ, can take values from +1 to -1.

- A ρ of +1 indicates a perfect association of ranks
- A ρ of zero indicates no association between ranks and
- ρ of -1 indicates a perfect negative association of ranks.
  The closer ρ is to zero, the weaker the association between the ranks.

Exp.

The following table provides data about the percentage of students who have free university meals and their CGPA scores. Calculate the Spearman's Rank Correlation between the two and interpret the result.

x:	14.4	7.2	27.5	33.8	38	15.9	4.9
y:	54	64	44	32	37	68	62

Sol.

d_X = Rank s_X	d_Y = Rank s_Y	$d = (d_X - d_Y)$	d^2
3	4	-1	1
2	6	-4	16
5	3	2	4
6	1	5	25

7	2	5	25
4	7	-3	9
1	5	-4	16
			$\Sigma d^2 = 96$

$$\rho = 1 - 6\frac{\sum_i d_i^2}{n\left(n^2 - 1\right)}$$

$$\rho = 1 - 6\frac{96}{7\left(7^2 - 1\right)}$$

$$\rho = -0.714$$

Therefore there is a strong negative coefficient.

REGRESSION ANALYSIS

Regression analysis refers to assessing the relationship between the outcome variable and one or more variables. The outcome variable is known as the dependent or response variable and the risk elements, and cofounders are known as predictors or independent variables. The dependent variable is shown by "y" and independent variables are shown by "x" in regression analysis.

The sample of a correlation coefficient is estimated in the correlation analysis. It ranges between -1 and +1, denoted by r and quantifies the strength and direction of the linear association among two variables. The correlation among two variables can either be positive, i.e. a higher level of one variable is related to a higher level of another or negative, i.e. a higher level of one variable is related to a lower level of the other.

The sign of the coefficient of correlation shows the direction of the association. The magnitude of the coefficient shows the strength of the association.

For example, a correlation of r = 0.8 indicates a positive and strong association among two variables, while a correlation of r = -0.3 shows a negative and weak association. A correlation near to zero shows the non-existence of linear association among two continuous variables.

LINEAR REGRESSION

Linear regression is a linear approach to modelling the relationship between the scalar components and one or more independent variables. If the regression has one independent variable, then it is known as a simple linear regression. If it has more than one independent variables, then it is known as multiple linear regression. Linear regression only focuses on the conditional probability distribution of the given values rather than the joint probability distribution. In general, all the real world regressions models involve multiple predictors. So, the term linear regression often describes multivariate linear regression.

DIFFERENCES BETWEEN CORRELATION AND REGRESSION:

- Correlation shows the quantity of the degree to which two variables are associated. It does not fix a line through the data points. You compute a correlation that shows how much one variable changes when the other remains constant. When r is 0.0, the relationship does not exist. When r is positive, one variable goes high as the other goes up. When r is negative, one variable goes high as the other goes down.

- Linear regression finds the best line that predicts y from x, but Correlation does not fit a line.

- Correlation is used when you measure both variables, while linear regression is mostly applied when x is a variable that is manipulated.

CHAPTER 6. PROBABILITY AND PROBABILITY DISTRIBUTION

PROBABILITY

Probability is a measure of the likelihood of an event to occur. Many events cannot be predicted with total certainty. We can predict only the chance of an event to occur i.e. how likely they are to happen, using it. Probability can range in from 0 to 1, where 0 means the event to be an impossible one and 1 indicates a certain event. Probability for Class 10 is an important topic for the students which explains all the basic concepts of this topic. The probability of all the events in a sample space adds up to 1.

For example, when we toss a coin, either we get Head OR Tail, only two possible outcomes are possible (H, T). But if we toss two coins in the air, there could be three possibilities of events to occur, such as both the coins show heads or both shows tails or one shows heads and one tail, i.e.(H, H), (H, T),(T, T).

FORMULA

The probability formula is defined as the possibility of an event to happen is equal to the ratio of the number of favourable outcomes and the total number of outcomes.

Probability of event to happen P(E) = Number of favourable outcomes/Total Number of outcomes

Exp.

There are 6 pillows in a bed, 3 are red, 2 are yellow and 1 is blue. What is the probability of picking a yellow pillow?

Ans. The probability is equal to the number of yellow pillows in the bed divided by the total number of pillows, i.e. 2/6 = 1/3.

PROBABILITY OF AN EVENT

Assume an event E can occur in r ways out of a sum of n probable or possible equally likely ways. Then the probability of happening of the event or its success is expressed as;

$P(E) = r/n$

The probability that the event will not occur or known as its failure is expressed as:

$P(E') = (n-r)/n = 1-(r/n)$

E' represents that the event will not occur.

Therefore, now we can say;

$P(E) + P(E') = 1$

This means that the total of all the probabilities in any random test or experiment is equal to 1.

EQUALLY LIKELY EVENTS

When the events have the same theoretical probability of happening, then they are called equally likely events. The results of a sample space are called equally likely if all of them have the same probability of occurring. For example, if you throw a die, then the probability of getting 1 is 1/6. Similarly, the probability of getting all the numbers from 2,3,4,5 and 6, one at a time is 1/6. Hence, the following are some examples of equally likely events when throwing a die:

- Getting 3 and 5 on throwing a die
- Getting an even number and an odd number on a die
- Getting 1, 2 or 3 on rolling a die

are equally likely events, since the probabilities of each event are equal.

COMPLEMENTARY EVENTS

The possibility that there will be only two outcomes which states that an event will occur or not. Like a person will come or not come to your house, getting a job or not getting a job, etc. are examples of complementary events. Basically, the complement of an event occurring in the exact opposite that the probability of it is not occurring. Some more examples are:

- It will rain or not rain today
- The student will pass the exam or not pass.
- You win the lottery or you don't.

PROBABILITY TERMS AND DEFINITION

Term	Definition	Example
Sample Space	The set of all the possible outcomes to occur in any trial	1. Tossing a coin, Sample Space (S) = {H,T} 2. Rolling a die, Sample Space (S) = {1,2,3,4,5,6}
Sample Point	It is one of the possible results	In a deck of Cards: • 4 of hearts is a sample point. • the queen of clubs is a sample point.
Experiment or Trial	A series of actions where the outcomes are always uncertain.	The tossing of a coin, Selecting a card from a deck of cards, throwing a dice.
Event	It is a single outcome of an	Getting a Heads while tossing a

Term	Definition	Example
	experiment.	coin is an event.
Outcome	Possible result of a trial/experiment	T (tail) is a possible outcome when a coin is tossed.
Complimentary event	The non-happening events. The complement of an event A is the event, not A (or A')	Standard 52-card deck, A = Draw a heart, then A' = Don't draw a heart
Impossible Event	The event cannot happen	In tossing a coin, impossible to get both head and tail at the same time

PROBABILITY DENSITY FUNCTION

The Probability Density Function (PDF) is the probability function which is represented for the density of a continuous random variable lying between a certain range of values. Probability Density Function explains the normal distribution and how mean and deviation exists. The standard normal distribution is used to create a database or statistics, which are often used in science to represent the real-valued variables, whose distribution are not known.

Exp. Draw a random card from a pack of cards. What is the probability that the card drawn is a face card?

Sol.

A standard deck has 52 cards.

Total number of outcomes = 52

Number of favourable events = 4 x 3 = 12 (considered Jack, Queen and King only)

Probability, P = Number of Favourable Outcomes/Total Number of Outcomes = 12/52= 3/13.

Example: Find the probability of 'getting 3 on rolling a die'.

Solution:

Sample Space = {1, 2, 3, 4, 5, 6}

Number of favourable event = 1

i.e. {3}

Total number of outcomes = 6

Thus, Probability, P = 1/6

PROBABILITY DISTRIBUTION

Probability distribution yields the possible outcomes for any random event. It is also defined based on the underlying sample space as a set of possible outcomes of any random experiment. These settings could be a set of real numbers or a set of vectors or set of any entities. It is a part of probability and statistics.

Random experiments are defined as the result of an experiment, whose outcome cannot be predicted. Suppose, if we toss a coin, we cannot predict, what outcome it will appear either it will come as Head or as Tail. The possible result of a random experiment is called an outcome. And the set of outcomes is called a sample point. With the help of these experiments or events, we can always create a probability pattern table in terms of variable and probabilities.

PROBABILITY DISTRIBUTION OF RANDOM VARIABLES

A random variable has a probability distribution, which defines the probability of its unknown values. Random variables can be discrete (not constant) or continuous or both. That means it takes any of a designated finite or countable list of values, provided with a probability mass function feature of the random variable's probability distribution or can take any numerical value in an interval or set of intervals. Through a probability density function

that is representative of the random variable's probability distribution or it can be a combination of both discrete and continuous.

Two random variables with equal probability distribution can yet vary with respect to their relationships with other random variables or whether they are independent of these. The recognition of a random variable, which means, the outcomes of randomly choosing values as per the variable's probability distribution function, are called **random variates.**

TYPES OF PROBABILITY DISTRIBUTION

There are two types of probability distribution which are used for different purposes and various types of the data generation process.

1. Discrete Probability Distribution (Binomial and Poisson Distribution)
2. Cumulative Probability Distribution (Normal Distribution)

Let us discuss now both the types along with its definition, formula and examples.

1. DISCRETE PROBABILITY DISTRIBUTION

A distribution is called a discrete probability distribution, where the set of outcomes are discrete in nature.

For example, if a dice is rolled, then all the possible outcomes are discrete and give a mass of outcomes. This is also known as probability mass functions.

So, the outcomes of binomial distribution consist of n repeated trials and the outcome may or may not occur. The formula for the binomial distribution is;

$$P(x) = \frac{n!}{r!(n-r)!} \cdot p^r (1-p)^{n-r}$$

$$P(x) = C(n, r).p^r (1-p)^{n-r}$$

where,

- n = Total number of events
- r = Total number of successful events.

- p = Success on a single trial probability.
- ${}^{n}C_{r} = [n!/r!(n-r)]!$
- 1 – p = Failure Probability

BINOMIAL DISTRIBUTION EXAMPLES

As we already know, binomial distribution gives the possibility of a different set of outcomes. In the real-life, the concept is used for:

- To find the number of used and unused materials while manufacturing a product.
- To take a survey of positive and negative feedback from the people for anything.
- To check if a particular channel is watched by how many viewers by calculating the survey of YES/NO.
- The number of men and women working in a company.
- To count the votes for a candidate in an election and many more.

Exp. If a coin is tossed 5 times, find the probability of:

(a) Exactly 2 heads

(b) At least 4 heads.

Sol.

(a) The repeated tossing of the coin is an example of a Bernoulli trial. According to the problem:

Number of trials: n=5

Probability of head: p= 1/2 and hence the probability of tail, q =1/2

For exactly two heads:

x=2

$P(x=2) = {}^{5}C2\ p^{2}\ q^{5-2} = 5!\ /\ 2!\ 3! \times (½)^{2} \times (½)^{3}$

P(x=2) = 5/16

(b) For at least four heads,

$x \geq 4$, $P(x \geq 4) = P(x = 4) + P(x=5)$

Hence,

$P(x = 4) = {}^{5}C4\ p^{4}\ q^{5-4} = 5!/4!\ 1! \times (½)^{4} \times (½)^{1} = 5/32$

$P(x = 5) = {}^{5}C5\ p^{5}\ q^{5-5} = (½)^{5} = 1/32$

Therefore,

$P(x \geq 4) = 5/32 + 1/32 = 6/32 = 3/16$

POISSON PROBABILITY DISTRIBUTION

The Poisson probability distribution is a discrete probability distribution that represents the probability of a given number of events happening in a fixed time or space if these cases occur with a known steady rate and individually of the time since the last event. It was titled after French mathematician Siméon Denis Poisson. The Poisson distribution can also be practised for the number of events happening in other particularised intervals such as distance, area or volume. Some of the real-life examples are:

- A number of patients arriving at a clinic between 10 to 11 AM.
- The number of emails received by a manager between the office hours.
- The number of apples sold by a shopkeeper in the time period of 12 pm to 4 pm daily.

PROBABILITY DISTRIBUTION FUNCTION

A function which is used to define the distribution of a probability is called a Probability distribution function. Depending upon the types, we can define these functions. Also, these functions are used in terms of probability density functions for any given random variable.

In the case of **Normal distribution**, the function of a real-valued random variable X is the function given by;

$$\mathbf{F_X(x) = P(X \leq x)}$$

Where P shows the probability that the random variable X occurs on less than or equal to the value of x.

2. CUMULATIVE PROBABILITY DISTRIBUTION

The cumulative probability distribution is also known as a continuous probability distribution. In this distribution, the set of possible outcomes can take on values on a continuous range.

For example, a set of real numbers, is a continuous or normal distribution, as it gives all the possible outcomes of real numbers. Similarly, set of complex numbers, set of prime numbers, set of whole numbers etc. are the examples of Normal Probability distribution. Also, in real-life scenarios, the temperature of the day is an example of continuous probability. Based on these outcomes we can create a distribution table. A probability density function describes it.

The formula for the normal distribution is;

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

where,

- μ = Mean Value
- σ = Standard Distribution of probability.
- If mean(μ) = 0 and standard deviation(σ) = 1, then this distribution is known to be normal distribution.
- x = Normal random variable

NORMAL DISTRIBUTION EXAMPLES

Since the normal distribution statistics estimates many natural events so well, it has evolved into a standard of recommendation for many probability queries. Some of the examples are:

- Height of the Population of the world
- Rolling a dice (once or multiple times)
- To judge Intelligent Quotient Level of children in this competitive world
- Tossing a coin

REFERENCES:

1. S. P. Gupta, Statistical Methods, Sultan Chand & Sons, 1976.
2. K. Janardhan, P. Hanmanth Rao Fundamentals of Biostatistics, I K International Publishing House Pvt. Ltd.
3. B Antonisamy, Prasanna S. Premkumar, Solomon Christopher , Principles and Practice of Biostatistics, Elsevier, 2017.
4. P. S. S. Sundar Rao, J. Richard, Introduction to Biostatistics and Research Methods, PHI Learning Pvt. Ltd, 2012.
5. https://www.bmj.com/
6. https://www.yourarticlelibrary.com/
7. https://byjus.com/maths/probability/
8. https://www.thoughtco.com/
9. http://www.brainkart.com/

Contact Us:

University Campus Address:

Jayoti Vidyapeeth Women's University

Vadaant Gyan Valley, Village-Jharna, Mahala Jobner Link Road,

Jaipur Ajmer Express Way, NH-8, Jaipur- 303122, Rajasthan (INDIA)

(Only Speed Post is Received at University Campus Address, No. any Courier Facility is available at Campus Address)

Pages	: 43
Book Price	: ₹ 150/-

JPD Women University Press

Jayoti Publication Desk