STATISTICS FOR ECONOMICS

COURSE CODE: B21EC04DC

Undergraduate Programme in Economics

Discipline Core Course

Self Learning Material

SREENARAYANAGURU
OPEN UNIVERSITY

SREENARAYANAGURU OPEN UNIVERSITY

The State University for Education, Training and Research in Blended Format, Kerala

SREENARAYANAGURU OPEN UNIVERSITY

Vision

To increase access of potential learners of all categories to higher education, research and training, and ensure equity through delivery of high quality processes and outcomes fostering inclusive educational empowerment for social advancement.

Mission

To be benchmarked as a model for conservation and dissemination of knowledge and skill on blended and virtual mode in education, training and research for normal, continuing, and adult learners.

Pathway

Access and Quality define Equity.

Statistics for Economics

Course Code: B21EC04DC

Semester - IV

Discipline Core Course
Undergraduate Programme in Economics
Self Learning Material

SREENARAYANAGURU
OPEN UNIVERSITY

SREENARAYANAGURU OPEN UNIVERSITY
The State University for Education, Training and Research in Blended Format, Kerala

SREENARAYANAGURU
OPEN UNIVERSITY

STATISTICS FOR ECONOMICS

Course Code: B21EC04DC
Semester- IV
Discipline Core Course
Undergraduate Programme in Economics

Academic Committee

Dr. Sanathanan Velluva
Dr. Priyesh C. A.
Dr. Jisha K. K.
Dr. Muneer Babu M.
Dr. Suchithra Devi S.
Dr. Ratheesh C.
Dr. Resmi C. Panicker
Dr. Suprabha L.
Dr. Rajeev S. R.

Development of the Content

Dr. Sanoop M.S.
Dr. Anitha C.S.
Sreekala M.
Dr. Christabell P.J.

Review and Edit

Dr. Leni Varghese
Dr. Christabell P.J.

Linguistics

Dr. Anitha C.S.

Scrutiny

Dr. Suchithra K.R.
Yedu T. Dharan
Soumya V.D.
Muneer K.
Dr. Smitha K.

Design Control

Azeem Babu T.A.

Cover Design

Jobin J.

Co-ordination

Director, MDDC :
Dr. I.G. Shibi
Asst. Director, MDDC :
Dr. Sajeevkumar G.
Coordinator, Development:
Dr. Anfal M.
Coordinator, Distribution:
Dr. Sanitha K.K.

Scan this QR Code for reading the SLM on a digital device.

Edition:
January 2025

Copyright:
© Sreenarayanaguru Open University

ISBN 978-81-982853-5-5

9 788198 285355

www.sgou.ac.m

YouTube

Visit and Subscribe our Social Media Platforms

MESSAGE FROM VICE CHANCELLOR

Dear learner,

I extend my heartfelt greetings and profound enthusiasm as I warmly welcome you to Sreenarayanaguru Open University. Established in September 2020 as a state-led endeavour to promote higher education through open and distance learning modes, our institution was shaped by the guiding principle that access and quality are the cornerstones of equity. We have firmly resolved to uphold the highest standards of education, setting the benchmark and charting the course.

The courses offered by the Sreenarayanaguru Open University aim to strike a quality balance, ensuring students are equipped for both personal growth and professional excellence. The University embraces the widely acclaimed "blended format," a practical framework that harmoniously integrates Self-Learning Materials, Classroom Counseling, and Virtual modes, fostering a dynamic and enriching experience for both learners and instructors.

The university aims to offer you an engaging and thought-provoking educational journey. The undergraduate programme in Economics is designed to be on par with the high-quality academic programmes offered at state universities throughout the country. The curriculum incorporates the latest methodologies for presenting economic ideas and concepts. It stimulates students' interest in developing a deeper comprehension of the discipline. The curriculum encompasses both theoretical concepts and historical evidence. Suitable emphasis is placed on India's experiences with economic transformation. This would aid learners in preparing for competitive examinations, should they choose to take them. Upon successfully completing the programme, we anticipate that students will be well-equipped to handle key areas within the economics discipline. The Self-Learning Material has been meticulously crafted, incorporating relevant examples to facilitate better comprehension.

Rest assured, the university's student support services will be at your disposal throughout your academic journey, readily available to address any concerns or grievances you may encounter. We encourage you to reach out to us freely regarding any matter about your academic programme. It is our sincere wish that you achieve the utmost success.

Regards,
Dr. Jagathy Raj V. P.

01-01-2025

Contents

1

BLOCK

Measures of Central Tendency and Dispersion

UNIT 1

Mean, Median & Mode

Learning Outcomes

After completing this unit, learners will be able to:

- understand the concept of central tendency
- calculate data using measures such as mean, median, and mode
- apply the measures of central tendency to real-world scenarios

Prerequisites

Statistics involves systematically studying data through processes like collection, classification, tabulation, analysis, interpretation, and presentation, facilitating decision-making across various fields. Techniques are employed to gather and summarise information, drawing meaningful conclusions to understand population characteristics. The collected raw data undergoes classification and tabulation, supporting in analysis. The Government of India conducts an all India census every ten years to record the actual number of persons alive at a given time, along with their age, sex, occupation etc. The Ministry of Labour conducts a family budget survey to determine the expenditure required by an average worker's family to meet essential needs such as food, clothing, education, and medical aid. This survey aims to understand the financial dynamics of households and ensure that policies and support systems align with the realistic economic demands of the workforce.

Averages, a common statistical practice, condense extensive data, offering a representative figure. Averaging proves useful in expressing central values for variables with fluctuations, enabling comparisons, trend identification, and informed decision-making. It is widely applicable to variables like prices, income, age, weight, and more, providing a concise indicator for analysis and decision support.

Keywords

Mean, Median, Mode

Discussion

1.1.1 Measures of Central Tendency

One of the most important objectives of statistical analysis is to get one single value that describe the characterises of the entire mass of the data. Such a value is called central value or average. The word average is very commonly used in day-to-day life. For example, we often talk of average height of a boy, average income etc. But in statistics the term average has a different meaning. It may be defined as the value of the distribution which is considered as the most representative for the group. Since the average represents the entire data, its value lies between two extremes, the largest and the smallest items. So, the average is frequently measured as the measure of central tendency. Here we consider three types of average, Mean, Median and Mode.

1.1.1.1 Mean

In everyday language, what many refer to as an 'average' is formally known to statisticians as the 'arithmetic mean.' This widely employed measure is derived by summing up all the items in a dataset and then dividing this total by the number of items. In simpler terms, it provides a single, representative value that encapsulates the central tendency of the entire set, making it a commonly used and easily understandable measure of the 'average' in statistical analysis.

Computation of Arithmetic Mean

We know that mean is a fundamental measure of central tendency that represent the average value of a data set. Let us now discuss how the mean is calculated in the case of individual, discrete and continuous series.

1. **Individual Series**

If x is the variable which takes the values x_1, x_2, ... x_n over n items then mean of x , denoted by $\bar{x}$ is

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{\sum x}{n}$$

where $\sum$ denote the summation over all values of x.

2. **Discrete Series**

There are two ways for calculating the mean for individual data series.

i. Direct Method

Mean $\bar{x} = \frac{\Sigma f \times x}{\Sigma f}$

where, f - frequency

ii. Short Cut Method

This method uses an assumed mean and deviations taken from that assumed mean to determine the arithmetic average. This method is also known as the deviation method. The assumed mean is chosen as some number midway between the largest and smallest of the observations.

$$\bar{x} = A + \frac{\Sigma f \times d}{\Sigma f}$$

where, A - Assumed mean

d - Deviation of the observations from the assumed mean ie. (x-A)

f - frequency

3. Continuous Series

i. Direct Method

Mean $\bar{x} = \frac{\Sigma f \times x'}{\Sigma f}$

where, x'- mid-point of various class

f - frequency

ii. Short Cut Method

Arithmetic mean $\bar{x} = A + \frac{\Sigma f \times d'}{\Sigma f}$

where, A- Assumed mean

d' - Deviations of mid-points from the assumed mean $ie, (x' - A)$

f - frequency

iii. Step Deviation Method

This method is used when class intervals are equal

Arithmetic mean $\bar{x} = A + \frac{\Sigma f \times d'}{\Sigma f} \times c$

where, A- Assumed mean

d'- Deviations of mid-points from the assumed mean $ie, d' = \frac{x - A}{c}$

f - frequency

c - class interval.

Properties of Arithmetic Mean

The arithmetic mean of a distribution has the following mathematical properties.

1. The sum of the deviations of the observations from the arithmetic mean in a data set is always zero.

   i.e., $\sum(x - \bar{x}) = 0$

2. The sum of squares of the deviations of the observations in a data is least when the deviations are taken from the arithmetic mean.

   i.e, $\sum(x - a)^2$ is least when $a = \bar{x}$

3. If the mean of n observations, x_1, x_2, ...x_n is $\bar{x}$ then the mean of the observation,

   $(x_1 \pm a), (x_2 \pm a), \dots \dots, (x_n \pm a)$ is $(\bar{x} \pm a)$.

4. If the mean of n observation, x_1, x_2, ...x_n is $\bar{x}$ and if each observation is multiplied by p, $p \neq 0$, then the mean of the new observation is $p\bar{x}$.

5. If n_1 and n_2 are the sizes and $\bar{x}_1$ and $\bar{x}_2$ are the respective means of two groups then the mean $\bar{x}$ of the combined group of size $n_1 + n_2$ is given by

   $$\bar{x} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2}$$

Merits of Arithmetic Mean

1. It has a rigid definition.
2. It is simple to comprehend and compute.
3. Based upon all the observations.
4. It is least affected by sampling fluctuations.
5. It can be subjected to further mathematical analysis.

Demerits of Arithmetic Mean

1. Extreme value has a significant impact for calculating mean.
2. It cannot be determined by inspection.
3. It cannot be used to measure qualitative characteristics like honesty, beauty, cleverness, and so on.

4. It is impossible to calculate for open-ended classes.
5. It is not suitable for averaging ratios and percentages.

Illustration : 1.1.1

The monthly income (Rs) of 10 families in a town are 280, 180, 96, 98, 104, 75, 80, 94, 100,75. Find the mean income of the family.

Solution

The mean income is

$$\overline{X} = \frac{\sum x}{n}$$

$$= \frac{280+180+96+98+104+75+80+94+100+75}{10}$$

$$= \frac{1182}{10}$$

$$= 118.2$$

Illustration 1.1.2

From the following data of marks obtained by 60 students of a class, calculate mean mark.

Marks	20	30	40	50	60	70
No of students	8	12	20	10	6	4

Solution

Direct method

Marks x	No of students (f)	$f \times x$
20	8	160
30	12	360
40	20	800
50	10	500
60	6	360
70	4	280
Total	$\sum f = 60$	2460

$$\bar{x} = \frac{\Sigma(f \times x)}{\Sigma f}$$

$$= \frac{2460}{60}$$

$$= 41$$

Mean Mark = 41

Short cut method

Assumed Mean = 40

Marks x	No of students (f)	$d = x - 40$	$f \times d$
20	8	-20	-160
30	12	-10	-120
40	20	0	0
50	10	10	100
60	6	20	120
70	4	30	120
Total	$\Sigma f = 60$		60

$$\bar{x} = A + \frac{\Sigma(f \times d)}{\Sigma f}$$

$$= 40 + \frac{60}{60}$$

$$= 40 + 1$$

$$= 41$$

Mean mark = 41

Illustration : 1.1.3

Find the arithmetic mean of the following data

Age in years	11	12	13	14	15
No of students	8	27	33	20	12

Solution

Direct method

Age in years x	No of students f	$f \times x$
11	8	88
12	27	324
13	33	429
14	20	280
15	12	180
Total	$\Sigma f = 100$	1301

$$\bar{x} = \frac{\Sigma(f \times x)}{\Sigma f}$$

$$= \frac{1301}{100}$$

$$= 13.01$$

Short cut method

Assumed Mean = 13

Age in years x	No of students	$d' = x - 13$	$f \times d$
11	8	-2	-16
12	27	-1	-27
13	33	0	0
14	20	1	20
15	12	2	24
Total	N= 60		1

$$\bar{x} = A + \frac{\Sigma(f \times d)}{\Sigma f}$$

$$= 13 + \frac{1}{100}$$

$$= 13 + 0.01$$

$$= 13.01$$

Illustration : 1.1.4

From the following data of marks obtained by 100 students, determine the mean mark by short cut method and step deviation method.

No. of students	0-10	10-20	20-30	30-40	40-50	50-60
Marks	5	10	25	30	20	10

Solution

Short cut method

Assumed Mean = 35

No. of students	Mid value (x')	F	$d' = x - 35$	$f \times d'$
0 - 10	5	5	-30	-150
10 - 20	15	10	-20	-200
20 - 30	25	25	-10	-250
30 - 40	35	30	0	0
40 - 50	45	20	10	200
50 - 60	55	10	20	200
Total		100		-200

$$\bar{x} = A + \frac{\Sigma(f \times d')}{\Sigma f}$$

$$= 35 - \frac{200}{100}$$

$= 35 - 2$

$= 33$

Step-deviation Method

Assumed Mean = 35

No. of students	Mid value (x')	f	$d' = \frac{x-35}{10}$	$f \times d'$
0-10	5	5	-3	-15
10-20	15	10	-2	-20
20-30	25	25	-1	-25
30-40	35	30	0	0
40-50	45	20	1	20
50-60	55	10	2	20
Total		100		-20

$$\bar{x} = A + \frac{\Sigma f \times d'}{\Sigma f} \times c$$

$$= 35 - \frac{20}{100} \times 10$$

$$= 35 - 2$$

$= 33$

Illustration : 1.1.5

From the following data of profits in a shop, determine the mean profit by short cut method and step deviation method.

Profit per shop	0-20	20-40	40-60	60-80	80-100
No. of shops	6	12	22	7	3

Solution

Short cut method

Assumed Mean = 50

Profit per shop	No. of shops f	Mid value x'	$d' = x' - 50$	$f \times d'$
0-20	6	10	-40	-240
20-40	12	30	-20	-240
40-60	22	50	0	0
60-80	7	70	20	140
80-100	3	90	40	120
Total	50			-220

$$\bar{x} = A + \frac{\Sigma(f \times d')}{\Sigma f}$$

$$= 50 - \frac{220}{50}$$

$$= 50 - 4.4$$

= 45.6

Step-deviation Method

Assumed Mean = 50

Profit per shop	No. of shops f	Mid value x'	$d' = \frac{x - 50}{20}$	$f \times d'$
0-20	6	10	-2	-12
20-40	12	30	-1	-12
40-60	22	50	0	0
60-80	7	70	1	7
80-100	3	90	2	6
Total	50			-11

$$\bar{x} = A + \frac{\Sigma(f \times d')}{\Sigma f} \times c$$

$$= 50 - \frac{11}{50} \times 20$$

$$= 50 - 4.4$$

$$= 45.6$$

Illustration : 1.1.6

200 people were interviewed by a public opinion pooling agency. The frequency distribution gives the age of the people interviewed. Calculate the mean age.

Age group (yrs)	80-89	70-79	60-69	50-59	40-49	30-39	20-29	10-19
Frequency	2	2	6	20	56	40	42	32

Solution

Here the classes are of the inclusive type. Before computing the mean, the inclusive class should be converted into an exclusive class (no difference between the upper limit of one class and the lower limit of the other class interval) to get the actual class limit.

Assumed Mean = 44.5

Age group (yrs)	Frequency f	x	Mid value x'	$d' = \frac{x' - 44.5}{10}$	$f \times d'$
80-89	2	79.5-89.5	84.5	4	8
70-79	2	69.5-79.5	74.5	3	6
60-69	6	59.5-69.5	64.5	2	12
50-59	20	49.5-59.5	54.5	1	20
40-49	56	39.5-49.5	44.5	0	0
30-39	40	29.5-39.5	34.5	-1	-40
20-29	42	19.5-29.5	24.5	-2	-84
10-19	32	9.5-19.5	14.5	-3	-96
Total	200				-174

$$\bar{x} = A + \frac{\Sigma(f \times d')}{\Sigma f} \times c$$

$$= 44.5 - \frac{174}{200} \times 10$$

$$= 44.5 - 8.7$$

$$= 35.8$$

Illustration : 1.1.7

Find the mean mark of students in a class by direct method.

Mark	Below 10	10-20	20-30	30- 40	40-50	50-60
No. of students	4	6	10	15	8	7

Solution

Since the class intervals are uniform, we assume that the lower limit of the first class is zero.

Mark	No. of students f	Mid value x'	$f \times x'$
0-10	4	5	20
10-20	6	15	90
20-30	10	25	250
30-40	15	35	525
40-50	8	45	360
50-60	7	55	385
Total	50		1630

$$\bar{x} = \frac{\Sigma(f \times x')}{\Sigma f}$$

$$= \frac{1630}{50}$$

$$= 32.6$$

Correction in Mean

The process for correction in mean is as follows

i. Find the sum of the values.

ii. Subtract incorrect value from the total.

iii. Add the correct value to the total.

iv. Divide the total by number of items.

Illustration : 1.1.8

The mean weight of a group of 25 boys was calculated to be 78.4 lb. It was later discovered that one value was misread as 69 lb instead of the correct value 96 lb. Calculate the correct mean.

Solution

$$\bar{x} = \frac{\sum x}{n}$$

$$78.4 = \frac{\sum x}{25}$$

$$\sum x = 78.4 \times 25 = 1960$$

Incorrect $\sum x = 1960$

Correct $\sum x$ = incorrect x- incorrect item + correct item

Correct $\sum x = 1960 - 69 + 96$

$= 1987$

$$\text{Correct Mean} = \frac{\text{Correct } \sum x}{n}$$

$$= \frac{1987}{25}$$

$= 79.48$ lb

Illustration : 1.1.9

The mean of 200 items was 50. Later on, it was discovered that two items were misread as 92 and 8, instead of 192 and 88. Find the correct mean

Solution

$$\bar{x} = \frac{\sum x}{n}$$

$$50 = \frac{\sum x}{200}$$

$\sum x = 200 \text{ x } 50 = 10000$

Incorrect $\sum x = 10000$

Correct $\sum x$ = incorrect x- incorrect item + correct item

Correct $\sum x = 10000 – (92+8) + (192+88)$

$= 10000 – 100 + 280$

$= 10{,}180$

$$\text{Correct Mean} = \frac{\text{Correct } \sum x}{n}$$

$$= \frac{10180}{200}$$

$= 50.9$

Illustration : 1.1.10

The arithmetic mean of 50 items were 100.At the time of calculation, two items 180 and 90 were wrongly taken as 100 and 10. Find the correct mean.

Solution

$$\bar{x} = \frac{\sum x}{n}$$

$$100 = \frac{\sum x}{50}$$

$\sum x = 100 \text{ x } 50 = 5000$

Incorrect $\sum x = 5000$

Correct $\sum x$ = incorrect x- incorrect item + correct item

Correct $\sum x = 5000 – (100+10) + (180+90)$

$= 5000 – 110 + 270$

$= \ 5160$

$$\text{Correct Mean} = \frac{\text{Correct } \sum x}{n}$$

$$= \frac{5160}{50}$$

$= 103.2$

Illustration : 1.1.11

The mean of marks in Statistics of 100 students in a class was 72. The mean of marks of boys was 75, while their number was 70. Find out the mean marks of girls in the class.

Solution

$$n_1 = 70, \quad \bar{x}_1 = 75, \quad n_1 + n_2 = 100, \quad \bar{x} = 72$$

$$n_2 = 100 - n_1 = 100 - 70 = 30$$

$$\bar{x} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2}$$

$$72 = \frac{70 \times 75 + 30 \times \bar{x}_2}{100}$$

$$72 \times 100 = 70 \times 75 + 30 \times \bar{x}_2$$

$$7200 = 5250 + 30 \times \bar{x}_2$$

$$30 \times \bar{x}_2 = 7200 - 5250 = 1950$$

$$\bar{x}_2 = \frac{1950}{30}$$

$$\bar{x}_2 = 65$$

Illustration : 1.1.12

100 labourers working in a factory, running two shifts of 60 and 40 workers respectively. The mean hourly wage of 60 labourers working in the morning shift is Rs. 40 and the mean hourly wage of 40 labourers working in the evening shift is Rs. 35. Find the mean hourly wage of 100 labourers working in a factory.

Solution

$$n_1 = 60, \quad n_2 = 40, \quad \bar{x}_1 = 75, \quad \bar{x}_2 = 35$$

$$\bar{x} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2}$$

$$= \frac{60 \times 75 + 40 \times 35}{100}$$

$$= \frac{4500 + 1400}{100}$$

$$= \frac{5900}{100}$$

$$= 59$$

Weighted Arithmetic Mean

Arithmetic computed by considering relative importance of each items is called weighted Arithmetic mean. Here we assign weight to each item in proportion to its relative importance.

Weighted Arithmetic mean $\bar{x} = \frac{\sum(w \times x)}{\sum w}$

Where,

x = values of items and w = weight of the item

Illustration : 1.1.13

Marks of a student in the subjects of Mathematics, Statistics, Physics, Chemistry and Biology are 40, 50, 60, 80 and 45 respectively. Assuming weights 5, 4, 3, 2 and 1 respectively, find weighted Arithmetic mean.

Solution:

Subjects	**Marks** (x)	**Weight** (w)	wx
Mathematics	40	5	200
Statistics	50	4	200
Physics	60	3	180
Chemistry	80	2	160
Biology	45	1	45
TOTAL		$\sum w = 15$	$\sum wx = 785$

$$\bar{x} = \frac{\sum w \times x}{\sum w} = \frac{785}{15} = 52.3$$

Geometric Mean

Geometric mean is the n^{th} positive root of the product of n positive given value. Hence geometric mean for a value x containing and n values for others $x_1, x_2, x_3, x_4, , \dots \dots x_n$ is given as :

G M of $x = \sqrt[n]{x_1 \times x_2 \times x_3 \dots \dots \times x_n}$ (For ungrouped data)

If we have a series of n positive values with repeated values such as $x_1, x_2, x_3, x_4,, \dots \dots x_n$ are repeated

$f_1, f_2 \cdots \cdots f_n$ times respectively then geometric mean will becomes,

G M of $x = \sqrt[n]{x_1^{f_1} \times x_1^{f_1} \times \dots \dots x_n^{f_n}}$ (For grouped data) where $n = f_1 + f_2 \dots . . f_n$

When the no. of items is large the task of multiplication and extraction of root is almost difficult. Hence logarithms are used in the calculation of geometric mean. The processis simple. It is the antilog of the average of the logs of all the items of a series. Symbolicallyit may be expressed as follows

G M of $x = Antilog \frac{((\log x_1 + log\, x_2 + \cdots log\ x_n)}{n}$

$= Antilog \frac{\Sigma \log x_i}{n}$

Illustration 1.1.14

Find the geometric mean of the values 10, 5, 15, 8, 12

Solution:

Here $x_1 = 10, x_2 = 5, x_3 = 15, x_4 = 8, x_5 = 12\ and\ n = 5$

G M of $x = \sqrt[n]{x_1 \times x_2 \times x_3 \dots \dots \times x_n}$

$= \sqrt[5]{10 \times 5 \times 15 \times 8 \times 12}$

$= \sqrt[5]{72000} = 9.36$

Illustration 1.1.15

Find the geometric mean for the monthly income of 10 persons given below1000, 850,500, 250, 200, 50, 100, 50, 40, 30

Solution:

X	Log X
1000	3.0000
850	2.9294
500	2.6990
250	2.3079
150	2.1761
100	2.0000

50	1.6990
40	1.6021
30	1.4771
n = 10	22.2816

G M of $x = Antilog\ \frac{((\log x_1 + log x_2 + \cdots log\ x_n)}{n}$

G M of $x = Antilog\ \frac{22.2816}{10} = Antilog\ 2.22816 = 169.0$

Therefore G M = Rs. 169.

Illustration 1.1.16

Calculate the geometric mean for the following distribution

Marks	1 - 10	10 - 20	20 - 30	30 - 40	40 - 50
No. of Students	8	12	20	6	4

Solution:

C.I	**F**	**X**	**Log X**	**f (log X)**
0-10	8	5	0.6990	5.5920
10-20	12	15	1.1761	14.1132
20-30	20	25	1.3978	27.9560
30-40	6	35	1.5441	9.2646
40-50	4	45	1.6532	6.6128
TOTAL	50			63.5386

$$GM = Antilog \frac{\sum logx_i}{n}$$

$$= Antilog \frac{63.5386}{50}$$

$$= Antilog 1.2708 = 18.65$$

Merits of Geometric Mean

1. It is rigidly defined

2. It gives less weight to large items and large weight to small items. Hence large items have less effect in it than on the arithmetic mean

3. It is amenable to further algebraic calculations

4. It gives equal importance to equal ratio of change

Demerits of Geometric Mean

1. It is not commonly used as it is very difficult to calculate

2. It is impossible to use it when anyone of the item in the series is 0 or negative

3. The value of GM may not be an actual item in the distribution

Harmonic Mean

Harmonic mean is a reciprocal of the Arithmetic mean of the reciprocals of the givenset of observation. Harmonic mean in mathematical terms is defined as follows:

For Ungrouped Data:

HM of $X = \bar{x} = \frac{n}{\Sigma\left(\frac{1}{x}\right)}$

For Grouped Data:

$$\bar{x} = \frac{N}{\Sigma\left(\frac{f}{x}\right)}$$

Illustration 1.1.17

Calculate the harmonic mean of the numbers 13.2, 14.2, 14.8,15.2 and 16.1.

Solution:

The harmonic mean is calculated as below:

X	1/X
13.2	0.0758
14.5	0.0704
14.8	0.0676
15.2	0.0658
16.1	0.0021
TOTAL	∑(1/X)= 0.3417

HM of $X = \bar{x} = \frac{n}{\Sigma\left(\frac{1}{x}\right)}$

$$\bar{x} = \frac{5}{0.3417} = 14.63$$

Illustration 1.1.18

A cyclist pedals from his house to college at a speed of 8 kmph and back from college to his house at 12kmph. Finds the average speed

Solution:

X	$\frac{1}{X}$
8	0.125
12	0.083

Average Speed = HM of X = $\bar{x} = \dfrac{n}{\Sigma\left(\frac{1}{x}\right)}$

$$HM\ of\ X = \frac{2}{0.208} = 9.6\ kmph$$

Merits of Harmonic Mean

1. It is rigidly defined
2. Since it gives larger weight to smaller items, it is a suitable measure where there are wide variations in the series
3. It is amenable to further algebraic manipulations
4. It will be a suitable average for the problems relating to time and rates

Demerits of Harmonic Mean

1. It is not easily understood as it is based on the reciprocals. cannot be calculated if any of the items is missing
2. The value of HM may not be an actual item in the distribution

1.1.1.2 Median

The median is the middle value of a distribution when data are arranged in either ascending or descending order. Unlike the mean, which is calculated by summing all values and dividing by the total number of items, the median represents a positional average. ‘Position’ refers to the place of a value within a series. The median’s position in a series is such that an equal number of items lie on either side of it. For example, if the incomes of 5 persons are 100, 120, 150, 160, and 180, then the median income would be 150. Even if the series is changed to 10, 25, 150, 200, and 300, the median remains 150. In the case of the mean, a change in the value of a single term alters the mean. The median may be defined as the value of the variable that divides the group into two equal parts, with one part comprising all values greater than the median and the other containing all values less than the median. The middle value serves as the median for an odd number of data points, while the average of the two middle values represents the median for an even number of data points. For instance, if there are 10 values, the median is the average of the 5th and 6th values.

Merits

1. Median is easy to understand and easy to calculate for a non-mathematical person.

2. Since median is a positional average, it is not affected at all by extreme observations.

3. Median can be computed while dealing with a distribution with open end classes.

4. Median can sometimes be located by simple inspection and can also be computed graphically.

5. Median is the only average to be used while dealing with qualitative characteristics which cannot be measured quantitatively but can still be arranged in ascending or descending order of magnitude.

Demerits

1. In case of even number of observations for an ungrouped data, median cannot be determined exactly.

2. Median, being a positional average, is not based on each and every item of the distribution.

3. Median is not suitable for further mathematical treatment.

4. Median is relatively less stable than mean, particularly for small samples since it is affected more by fluctuations of sampling as compared with arithmetic mean.

Computation of Median

1. For individual series

i. Sort the data into ascending or descending order.

ii. Use the formula.

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

Illustration 1.1.19

From the following data of the weekly wages of 7 workers compute the median wage.

Wages (Rs)	100	150	80	90	160	200	140

Solution

Arrange the data in ascending order

Wages (Rs)	80	90	100	140	150	160	200

Apply the formula

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

= $\left(\frac{7+1}{2}\right)^{th}$ item

= 4th item

The 4th item in the series is 140.

∴ Median wage is 140

Illustration 1.1.20

Obtain the medium value from the following data

391	384	591	407	672	522	777	2488	1490	753

Solution

Arrange the data in ascending order

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

= $\left(\frac{10+1}{2}\right)^{th}$ item

= 5.5th item

Size of 5.5th term = $\frac{1}{2}(5^{th} term + 6^{th} term)$ = $\frac{1}{2}(591 + 672)$ = 631.5

∴ Median is 631.5

ii. For discrete series

Arrange the data in ascending or descending order

Calculate cumulative frequency (c f). The cumulative frequency is the total of frequencies, in which the frequency of the first-class interval is added to the frequency of the second class and then the sum is added to the frequency of the third class interval and so on.

Determine $\frac{N+1}{2}$ where N is the total frequency

Median is the value for the $\left(\frac{N+1}{2}\right)^{th}$ item of the data

Illustration 1.1.21

From the following data find the median income.

Income:	100	150	80	200	250	180
No of employees:	24	26	16	20	6	30

Solution

Income arranged in ascending order	**f**	**c f**
80	16	16
100	24	40
150	26	66
180	30	96
200	20	116
250	6	122
Total	122	

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

$= \left(\frac{122+1}{2}\right)^{th}$ item

$= \left(\frac{123}{2}\right)^{th}$ item

= 61.5th item

∴ Median is the value in the data which comes in the 61.5th position, which is the value of the item having cumulative frequency 66. Since cumulative frequency of 66 comes under the income 150, median is the value in the data that comes in the 66th position,

∴ Median = 150

Illustration 1.1.22

Compute median for the following distribution.

Height of Women in inches	60	61	62	63	64	65	66
No of employees:	27	146	435	398	210	128	98

Solution

Height of women in inches	**f**	**c f**
60	27	27
61	146	173
62	435	608
63	398	1006
64	210	1216
65	128	1344
66	98	1442
Total	1442	

$$\text{Median} = \left(\frac{N+1}{2}\right)^{th} \text{ item}$$

$$= \left(\frac{1442+1}{2}\right)^{th} \text{ item}$$

$$= \left(\frac{1443}{2}\right)^{th} \text{ item}$$

= 721.5th item

∴ Median = 63

Illustration 1.1.23

Compute median for the following distribution.

No. of persons	28	29	30	31	32	33
Persons wearing	10	20	40	65	50	15

Solution

No. of persons	**f**	**c f**
28	10	10
29	20	30
30	40	70
31	65	135
32	50	185
33	15	200
Total	200	

$$\text{Median} = \left(\frac{N+1}{2}\right)^{th} \text{ item}$$

$$= \left(\frac{200+1}{2}\right)^{th} \text{ item}$$

$$= \left(\frac{201}{2}\right)^{th} \text{ item}$$

$= 100.5^{\text{th}}$ item

∴ Median = 31

For continuous series

i. Convert inclusive classes to the exclusive class (if any). i.e, no difference between the upper limit of one class and the lower limit of the other class interval.

ii. Calculate the cumulative frequencies (c f).

iii. Calculate $\frac{N}{2}$, where N is the total frequency.

iv. Identify the class having cumulative frequency $\frac{N}{2}$ which is the median class.

v. Find median by using this formula;

$$\text{Median} = l + \frac{\frac{N}{2} - m}{f} \times c$$

where,

l – Lower limit of the median class.

m – Cumulative frequency of the class preceding the median class.

f – Frequency of the median class.

c – Class interval of the median class.

Illustration 1.1.24

The following table shows the household income of 80 families.

Income (Rs in 1000's):	0-10	10-20	20-30	30-40	40-50	50-60	60-70
No of household:	18	15	12	22	13	12	8

Find the median income.

Solution

The cumulative distribution table is

Class	**f**	**c f**
0-10	18	18
10-20	15	33
20-30	12	45
30-40	22	67
40-50	13	80
50-60	12	92
60-70	8	100
	N = 100	

$\frac{N}{2} = \frac{100}{2} = 50$

The class having cumulative frequency 50 is 30-40

∴ Median class is 30-40

Median = $l + \frac{\frac{N}{2} - m}{f} \times c$

l – Lower limit of the median class - 30

m – Cumulative frequency of the class preceding the median class - 45

f – Frequency of the median class - 22

c – Class interval of the median class - 10

$$= 30 + \frac{50 \ - 45}{22} \text{ x } 10$$

$$= 30 + \frac{5 \times 10}{22}$$

$$= 30 + \frac{50}{22}$$

$$= 30 + 2.273$$

$$= 32.273$$

Illustration 1.1.25

Find the median mark from the following frequency table giving the distribution of marks of 100 students.

Mark:	0-9	10-19	20-29	30-39	40-49
No of students:	5	15	32	41	7

Solution

Here the classes are of the inclusive type. Before computing the median, the inclusive class should be converted into an exclusive class to get the actual class limit.

Marks	**Actual class**	**f**	**c f**
0-9	-0.5 - 9.5	5	5
10-19	9.5 – 19.5	15	20
20-29	19.5 – 29.5	32	52
30-39	29.5 - 39.5	41	93
40-49	39.5 – 49.5	7	100
		N = 100	

$$\frac{N}{2} = \frac{100}{2} = 50$$

The class having cumulative frequency 40 is 19.5 – 29.5

∴ Median class is 19.5 - 29.5

$$\text{Median} = l + \frac{\frac{N}{2} - m}{f} \times c$$

$$= 19.5 + \frac{(50 - 20)}{32} \text{x}\, 10$$

$$= 19.5 + \frac{300}{32}$$

= 19.5+ 9.375

= 28.875

Illustration 1.1.26

The following table shows the age of persons in an office. Find the median age.

Age :	Less than 10	Less than 20	Less than 30	Less than 40	Less than 50	Less than 60	Less than 70	Less than 80
No of persons:('000)	4	16	40	76	96	112	120	125

Solution

First convert the given distribution into continuous frequency distribution and convert the given cumulative frequency into frequency.

Age	**Cf**	**frequency**
0 - 10	4	4
10 -20	16	16 - 4 = 12
20 - 30	40	40 -16 = 24
30 - 40	76	76 - 40 = 36
40 - 50	96	96 -76 = 20
50 - 60	112	112 - 96 = 16
60 - 70	120	120 -112 = 8
70 - 80	125	125 -120 = 5
		N = 125

$$\frac{N}{2} = \frac{125}{2} = 62.5$$

the class having cumulative frequency 50 is 30-40

∴ median class is 30-40

$$\text{median} = l + \frac{\frac{N}{2} - m}{f} \times c$$

$$= 30 + \frac{62.5 - 40}{36} \times 10$$

$$= 30 + \frac{22.5}{36} \times 10$$

$$= 30 + \frac{225}{36}$$

$$= 30 + 6.25$$

$$= 36.25$$

Illustration 1.1.27

The following table gives the distribution of monthly wages of 600 middle-class families in a certain city. Find the median income of the families.

Monthly income(Rs)	Below 75	75-150	150-225	225-300	300-375	375-450	450 above
No of families	69	167	207	65	58	24	10

Solution

Age	**f**	**c f**
0 -75	69	69
75 - 150	167	236
150 - 225	207	443
225 - 300	65	508
300 - 375	58	566
375 - 450	24	590
450 above	10	600
	600	

$\frac{N}{2} = \frac{600}{2} = 300$

the class having cumulative frequency 300 is 150-225

∴ median class is 150-225

median = $l + \frac{\frac{N}{2} - m}{f} \times c$

$= 150 + \frac{300 - 236}{207} \times 75$

$= 150 + \frac{64}{207} \times 75$

$= 150 + \frac{4800}{207}$

$= 150 + 23.188$

$= 173.188$

Illustration 1.1.28

The frequency distribution of heights of 100 college students is as follows. From the given data, compute the median height.

Height	141-150	151-160	161-170	171-180	181-190
Frequency	5	16	56	19	4

Solution

Marks	**Actual class**	**f**	**c f**
141-150	140.5 - 150.5	5	5
151-160	150.5 – 160.5	16	21
161-170	160.5 – 170.5	56	77
171-180	170.5 - 180.5	19	96
181-190	180.5 – 190.5	4	100
		N = 100	

$\frac{N}{2} = \frac{100}{2} = 50$

The class having cumulative frequency 40 is 19.5 – 29.5

∴ Median class is 160.5 – 170.5

Median = $l + \dfrac{\frac{N}{2} - m}{f} \times c$

$= 160.5 + \dfrac{(50 - 21)}{56} \text{x} 10$

$= 160.5 + \dfrac{290}{56}$

= 160.5+ 5.18

= 165.68

1.1.3 Mode

Mode or Model value is that value in the series which occours or repeat itself with greatest number of times. The mode of a distribution is the value at the point around which the items tend to be most heavily concentrated. For example, if seven men are receiving daily wages of Rs. 5, 6, 7, 7, 7, 8 and 10, it is clear that the model value is Rs. 7 per day. If we have a series such that 2, 3, 5, 6, 7, 10, it is apparent that there is no mode.

Merits

1. Mode is representative value of the distribution.
2. It is not affected by extremely large or small items.
3. It can be determined in open end distribution.
4. It is used to describe quantitative phenomenon.
5. It can be determined graphically.

Demerits

1. The value of mode can not always be determined.
2. It is not capable for algebraic manipulations.
3. It is not based on all items of the distribution.
4. It is not rigidly defined.
5. It is affected to a greater extent by the fluctuations of sampling.

Relation between Mean, Median and Mode

There exist a relationship between Mean, Median and Mode for moderately asymmetric distribution. For asymmetric distribution, Mean, Median, and Mode will have identical value.

The relation is

Mean – Mode = 3 (Mean – Median)

OR

Mode = 3 Median-2 Mean

Computation of mode

i. For individual series

The mode in individual observations is the most occurring value in a series.

Illustrations 1.1.29

Find the mode of the following set of numbers :

10,13,18,9,10,16,11,8,7,10 and 19

Solution

Here 10 occurs maximum number of times, ie, 3 times, mode = 10

i. For Discrete series

Observation with highest frequency is considered as the mode in the discrete series.

ii. For continuous series

Steps

$$Mode = l + \frac{f_1 - f_0}{2f_1 - f_0 - f_2} \times c$$

where,

l – Lower limit of the modal class.

f_1 - Frequency of the modal class.

f_0 – frequency of the preceding class to the modal class.

f_2 – frequency of the succeeding class to the modal class.

c – Class interval of the modal class

Grouping Table and Analysis Table

The item with the highest frequency is referred to as a mode. However, if the maximum frequency is repeating or if the maximum frequency occurs at the beginning or end of the distribution or if there are irregularity in the distribution, it may be impossible to find the mode simply by looking at the distribution. In rare circumstances, the frequency concentration may be more concentrated around a frequency that is lower than the highest frequency. A grouping table and an analysis table should be developed to determine the correct modal value in such circumstances.

Steps for calculation

i. Construct a six-column grouping table.

ii. In column (1), record the frequency in relation to the item.

iii. The frequencies in column (2) are arranged in twos, starting at the top. Their totals are calculated, and the highest total is highlighted.

iv. The frequencies are grouped in twos again in column (3), leaving the first frequencies. The highest total is once again noted.

v. The frequencies in column (4) are arranged in threes, starting at the top. Their totals are calculated, and the highest total is highlighted.

vi. The frequencies are grouped in threes again in column (5), leaving the initial frequency. Their totals are calculated, and the highest total is highlighted.

vii. The frequencies are grouped in threes again in column (6), leaving the first and second frequencies. After totalling the frequencies, the highest total is identified and highlighted again.

viii. Create an analysis table to find the modal value or modal class that the largest frequencies cluster around for the longest periods of time. Place the column number on the left-hand side of the table and the item sizes on the right-hand side. Mark 'X' in the relevant box corresponding to the values they represent to input the values against which the highest frequencies are found. The mode is the set of values with the most 'X' marks against them.

Illustration 1.1.30

The following table shows the frequency distribution of the marks of 100 students. Find the mode.

Marks	0-10	10-20	20-30	30-40	40-50	50-60	60-70	70-80	80-90
No of students	4	12	18	22	21	19	10	3	1

Solution

There is no irregularities in the distribution, hence the model class is 30-40.

c = 10 $f_1 = 22$, $f_2 = 21$, $f_0 = 18$

$$Mode = l + \frac{f_1 - f_0}{2f_1 - f_0 - f_2} \times c$$

$$= 30 + \frac{22 - 18}{2 \times 22 - 18 - 21} \times 10$$

$$= 30 + \frac{4}{5} \times 10$$

$$= 30 + \frac{40}{5}$$

$$= 30 + 8$$

$$= 38$$

Illustration 1.1.31

Find the mode of the following distribution of marks obtained by 125 students in a certain examination.

Marks	0-10	10-20	20-30	30-40	40-50	50-60	60-70	70-80	80-90	90-100
No of students	3	4	8	10	15	35	20	16	8	6

Solution

There is no irregularities in the distribution, hence the model class is 50-60.

c = 10 $f_1 = 35$, $f_2 = 20$, $f_0 = 15$

$$Mode = l + \frac{f_1 - f_0}{2f_1 - f_0 - f_2} \times c$$

$$= 50 + \frac{35 - 15}{2 \times 35 - 15 - 20} \times 10$$

$$= 50 + \frac{20}{35} \times 10$$

$$= 50 + \frac{200}{35}$$

$$= 50 + 5.71$$

$$= 55.71$$

Illustration 1.1.32

Find the mode of the following distribution.

Size	Below 10	Below 20	Below 30	Below 40	Below 50	Below 60	Below 70	Below 80
Frequency	5	20	50	105	190	250	295	320

Solution

Size	**Cf**	**frequency**
0 -10	5	5
10 - 20	20	15
20 - 30	50	30
30 - 40	105	55
40 - 50	190	85
50 - 60	250	60
60 -70	295	45
70 - 80	320	25

There is no irregularities in the distribution, hence the model class is 40-50.

$c = 10 \quad f_1 = 85,\ f_2 = 60,\ f_0 = 55$

$$Mode = l + \frac{f_1 - f_0}{2f_1 - f_0 - f_2} \times c$$

$$= 40 + \frac{85 - 55}{2 \times 85 - 55 - 60} \times 10$$

$$= 40 + \frac{30}{55} \times 10$$

$$= 40 + \frac{300}{55}$$

$$= 40 + 5.45$$

$$= 45.45$$

Illustration 1.1.33

Find mode from the following data

Weight (Kg):	93-97	98-102	103-107	108-112	113-117	118-122	123-127	128-132
No of students	3	5	12	17	14	6	3	1

Solution

Since the formula for mode requires the distribution to be continuous with 'exclusive type' classes we first convert the classes into class boundaries as given in the following table

X	F
92.5-97.5	3
97.5-102.5	5
102.5-107.5	12
107.5-112.5	17
112.5-117.5	14
117.5-122.5	6
122.5-127.5	3
127.5-132.5	1

Here 107.5-112.5 is the model class. c = 5 f_1 =17, f_2=14, f_0 = 12

$$\text{Mode} = l + \frac{(f_1 - f_0)}{2f_1 - f_0 - f_2} \text{ x C}$$

$$= 107.5 + \frac{(17-12)}{2\times17-12-14} \text{ x } 5$$

$$= 107.5 + \frac{25}{8}$$

$$= 107.5 + 3.125$$

$$=110.625$$

Illustration 1.1.34

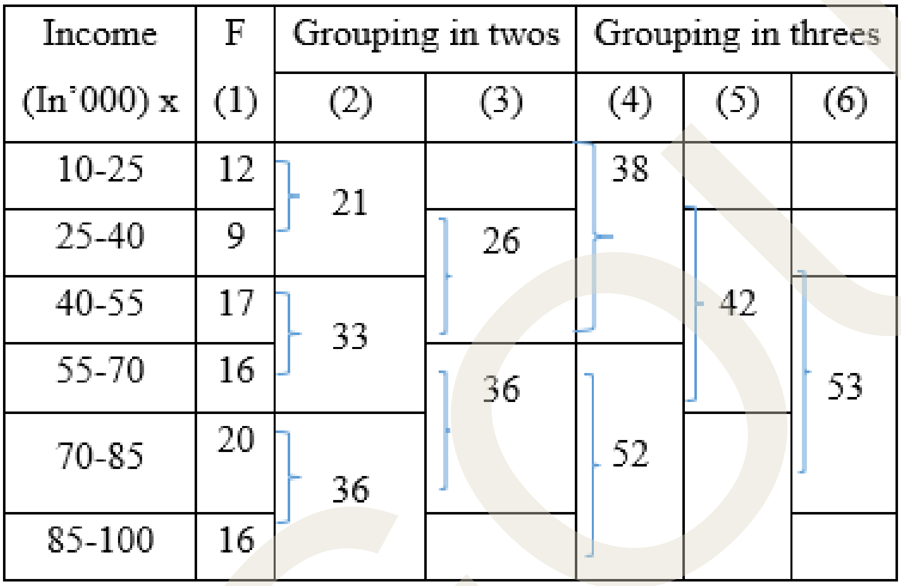
The following table shows the monthly income of 130 families. Calculate the mode value.

Income (in’000):	10-25	25-40	40-55	55-70	70-85	85-100
No of Families:	12	9	17	16	20	16

Solutions

Let us use grouping table to determine mode.

(a) Grouping table



Income	F	Grouping in twos		Grouping in threes		
(In’000) x	(1)	(2)	(3)	(4)	(5)	(6)
10-25	12	21		38		
25-40	9		26		42	
40-55	17	33				53
55-70	16		36	52		
70-85	20	36				
85-100	16					

In column 1 the highest frequency is 20 corresponds to the 70-85. So, we put X mark in 70-85. In column 2 the highest frequency is 36 corresponds to 70-85 and 85-100. So we put X mark in 70-85 and 85-100. In column 3, the highest frequency is 36 corresponds to 55-70 and 70-85. So we put X mark in 55-70 and 70-85. In column 4 the highest frequency is 52 corresponds to 55-70, 70-85, and 85-100. So we put X mark in 55-70, 70-85 and 85-100. In column 5 the highest frequency is 42 corresponds to 25-40, 40-55 and 55-70. So we put X mark in 25-40, 40-55 and 55-70. In column 6 the highest frequency is 53 corresponds to 40-55, 55-70 and 70-85. So we put X mark in 40-55, 55-70 and 70-85.

(b) Analysis table

Variable	10-25	25-40	40-55	55-70	55-70	70-85	85-100
F column							
1						X	
2						X	X

3					X	X	
4					X	X	X
5		X	X	X			
6			X	X	X		
Total	-	1	2	2	3	4	2

The greatest total (4) is noted to be against 70-85.

$$Mode = l + \frac{f_1 - f_0}{2f_1 - f_0 - f_2} \times c$$

$$= 70 + \frac{20 - 16}{2 \times 20 - 16 - 16} \times 15$$

$$= 70 + \frac{4}{8} \times 15$$

$$= 70 + \frac{60}{8}$$

$$= 70 + 7.5$$

$$= 77.5$$

Illustration 1.1.35

The following table shows the marks of students in a class. Calculate the mode value.

Marks	0-10	10-20	20-30	30-40	40-50	50-60	60-70	70-80	80-90
No of students	4	8	18	30	20	10	30	3	2

Solution

Here the maximum frequency 30 is repeating, we use grouping table and an analysis table

Grouping Table

Income (In'000)	F (1)	Grouping in twos		Grouping in threes		
		(2)	(3)	(4)	(5)	(6)
0-10	4	12		30		
10-20	8		26		56	
20-30	18	48				68
30-40	30		50	60		
40-50	20	30			60	
50-60	10		40			43
60-70	30	33		35		
70-80	3		5			
80-90	2					

(b) Analysis table

Variable F column	0-10	10-20	20-30	30-40	40-50	50-60	60-70	70-80
1				X			X	
2			X	X				
3				X	X			
4				X	X	X		
5					X	X	X	
6			X	X	X			
Total	**-**	**-**	**2**	**5**	**4**	**2**	**2**	**-**

The modal class is identified as 30-40. The following formula can be used to calculate mode.

$$Mode = l + \frac{f_1 - f_0}{2f_1 - f_0 - f_2} \times c$$

$$= 30 + \frac{(30 - 18)}{(2 \times 30) - 18 - 20} \times 10$$

$$= 30 + \frac{12}{22} \times 10$$

$$= 30 + \frac{120}{22}$$

$$= 30 + 5.45$$

$$= 35.45$$

Illustration 1.1.36

Calculate the mode of the following frequency distribution.

x	0-10	10-20	20-30	30-40	40-50	50-60	60-70	70-80	80-90
f	4	2	18	22	21	19	10	3	1

Solutions

We may use grouping table to determine the mode.

(a) Grouping table

Income (In'000)	f (1)	Grouping in twos		Grouping in threes		
		(2)	(3)	(4)	(5)	(6)
0-10	4	6		24		
10-20	2		20		42	
20-30	18	40				61
30-40	22		43	62		
40-50	21	40			50	
50-60	19		19			32
60-70	10	13		14		
70-80	3		4			
80-90	1					

(b) Analysis table

Variable F column	0-10	10-20	20-30	30-40	40-50	50-60	60-70	70-80	80-90
1				X					
2			X	X	X	X			
3				X	X				
4				X	X	X			
5					X	X	X		
6			X	X	X				
Total	**-**	**-**	**2**	**5**	**5**	**3**	**1**		

In the above table there are two classes, 30-40 and 40-50 which are repeated maximum number (5) of times and as such we cannot decide about the modal class. Thus, even the method of grouping fails to give the modal class.

We can find mode from the formula Mode = 3 Median – 2 Mean

By Calculation Mean = 42.2 and Median = 41.9.

Therefore Mode = $3 \times 41.9 - 2 \times 42.2 = 41.3$

Illustration 1.1.37

The mean of 10 observations is 20 and median is 15. Find the mode of the observations.

Solution

Given mean = 20, median = 15.

The relation between mean , median and mode is Mode = 3 Median-2 Mean

Mode = $3 \times 15 - 2 \times 20 = 45 - 40 = 5$

Illustration 1.1.38

The mean of 10 observations is 24.6 and mode is 26.1. Find the median of the observations.

Solution

Given mean = 24.6, mode = 26.1

The relation between mean , median and mode is Mode = 3 Median-2 Mean

$$26.1 = 3\ Median - 2 \times 24.6$$

$$3\ Median = 26.1 + 2 \times 24.6$$

$$= 26.1 + 49.2$$

$$= 75.3$$

Median = $\frac{75.3}{3}$ = 25.1

Recap

- The arithmetic mean - the sum of all observations' values divided by the number of observations.
- Weighted Arithmetic mean $\bar{x} = \frac{\sum wx}{\sum w}$
- G M of $x = \sqrt[n]{x_1 \times x_2 \times x_3 \dots \dots \times x_n}$
- HM of $X = \bar{x} = \frac{n}{\sum\left(\frac{1}{x}\right)}$
- The sum of deviations of items from the arithmetic mean is always equal to zero.
- Median is the middle most value in the series.
- Mode is the most frequent value in the series.

Objective Questions

1. Which measure of central tendency is used to find the per capita income in different cities.
2. Represent arithmetic mean for individual series mathematically.
3. What does median represents?
4. What is the sum of deviations of the observations from the arithmetic mean?
5. If the maximum frequency is repeated which method is used to find mode?
6. What is the relation between Mean Median and Mode.

Answers

1. Mean

2. $\frac{\sum x}{n}$

3. Middle value of a distribution

4. Zero

5. Grouping table and analysing table

6. Mean-Mode = 3(Mean-Median)

Self-Assessment Questions

1. Explain the relationship between mean, median and mode.

2. What are the limitations of Mean?

Assignments

1. The heights in inches of 70 employees in an office are given below. Find the mean height of an employee

Height (in inches)	60	62	63	65	67	68
No of employees	5	10	12	18	15	10

2. Given below is the following frequency distribution of weights of 60 oranges.

Weight (in gram):	65-84	85-104	105-124	125-144	145-164	165-184	185-204
Frequency:	9	10	17	10	5	4	5

Find out how much an orange weighs on average.

3. The mean wage of 120 factory workers was found to be ₹17,000. It was then discovered that an amount of ₹18750 wage was misread as ₹17850. Find the right mean.

4. The mean salary paid to 1,000 employees of an establishment was found to be Rs. 180·40. Later on, after disbursement of salary, it was discovered that the salary of two employees was wrongly entered as Rs. 297 and 165. Their correct salaries were Rs. 197 and Rs. 185. Find the correct Arithmetic Mean.

5. The average daily wage of all workers in a factory is Rs. 444. If the average daily wages paid to male and female workers are Rs. 480 and Rs. 360 respectively, find the percentage of male and female workers employed by the factory.

6. Find the median wage of the following distribution

Wages (Rs):	20-30	30-40	40-50	50-60	60-70
No of labourers:	3	5	20	10	5

7. The following table represent the income of 122 families. Calculate Median income.

Income:	1000	1500	3000	2000	2500	1800
No of family:	24	26	16	20	6	30

8. Find mode from the following data.

Mark:	20-24	25-29	30-34	35-39	40-44	45-49
No of student:	20	24	32	28	20	26

9. Calculate mode from the following data

Weight in kg:	25	30	35	40	45	50	55	60
No of person:	50	70	80	180	70	30	20	10

References

1. Gupta, S.P, *Statistical Methods*, Sulthan Chand and Sons, New Delhi.
2. SC Gupta and VK Kapoor, *Fundamentals of Mathematical Stastistics*

Suggested Readings

1. Jan Ubøe, *Introductory Statistics for Business and Economics*: Theory, Exercises and Solutions, Springer International Publishing.
2. T. Rajaretnam, *Statistics for Social Sciences*, Sage India.
3. Gupta S. and V.K. Kapoor, *Fundamentals of Applied Statistics*, S.Chand and Sons, New Delhi.
4. Monga, G.S. *Mathematics and Statistics for Economics*, Vikas Publishing, New Delhi

2

UNIT

Measures of Dispersion

Learning Outcomes

After completing this unit, the learner will be able to:

- understand the role of these measures in describing the spread or dispersion of data.
- apply the appropriate measure based on the characteristics of the data and the goals of the analysis.
- utilize real-world examples to illustrate the practical application of each measure in diverse contexts.

Prerequisites

Understanding measures of central tendency and mastering data collection and analysis methods are important in the study of dispersion. While averages or measures of central tendency offer insights into the concentration of observations around the central part of the dataset, relying solely on the average can be insufficient for a comprehensive understanding of the distribution. To illustrate, consider two series: a) 3, 5, 7, 14, 16, and b) 2, 5, 10, 11, 17. Both series have an average of 9. Knowing only the average, however, does not provide clarity on which series we are dealing with. It could be the first series, the second series, or any other with a sum of 45. This highlights the limitations of measures of central tendency in offering a complete picture of distribution. To address this gap, we turn to another essential concept- measures of dispersion-to gain a more understanding of how data is spread out.

Keywords

Range, Interquartile range, Quartile deviation, Standard deviation

Discussion

The Measures of Central Tendency give us one single figure that represents the entire data.

Consider the Salary of workers in three fields

A	B	C
100	100	1
100	105	489
100	102	2
100	103	3
100	90	5

$$\bar{x} = \frac{500}{5} = 100 \qquad \bar{x} = \frac{500}{5} = 100 \qquad \bar{x} = \frac{500}{5} = 100$$

Since the mean is same in all the cases, one is likely to conclude that these series are alike in nature. But the distribution differs widely from one another. In series A, none of the term is deviated from mean hence there is no dispersion. In series B, all the terms except one item varies from mean, but the variation is very small compared to Series C. In series C, not a single item is represented by the mean and the items vary widely from one another. In C dispersion is very greater compared to series B.

Measures of central tendency are therefore insufficient for the study of observations. Thus, we study measure of dispersion which is one that measures the extent to which there are difference between individual observations and some central value.

1.2.1 Range

The range is the simplest of all the measures of dispersion. It is defined as the difference between the two extreme observations of the distribution. In other words, range is the difference between the greatest (maximum) and the smallest (minimum) observation of the distribution. It indicates the limit with which the values fall.

Range = L - S

where,

L = Largest value, S = Smallest value

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

In case of the grouped frequency distribution (for discrete values) or the continuous frequency distribution, range is defined as the difference between the upper limit of the highest class and the lower limit of the smallest class.

If the average of the two distributions are about the same, a comparison of the range indicate that the distribution with smaller range has less dispersion and the average of that distribution is more representative of the group.

Merits

1. It is simplest to understand and easier to compute as it only involves finding the difference between the maximum and minimum values in a dataset.

2. It provides an intuitive understanding of the spread in data.

3. Its simplicity makes it a quick and accessible measure of dispersion.

4. Calculating the range requires only the knowledge of the two extreme values, making it applicable even in situations where detailed data may not be readily available.

Demerits

1. Since it is based on extreme values it may be changed either of the extreme values happens to drop out.

2. It does not consider the size of the dataset.

3. It does not take into account the frequencies of the distribution.

4. The range can change drastically with the addition or removal of just one data point.

5. Range cannot be found for open end distributions.

Uses

Range is useful in the following situations,

1. In statistical quality control range can be used as a measure of variation.

2. Range is used to describe the difference between a commodity's highest and lowest price. It is the most widely used measure of variability in our daily lives.

3. For weather forecasts, the meteorological department uses a range.

4. Range can be applied in areas where the data have small variations

Illustration : 1.2.1

Below are the prices of 1 kg of sugar for the first six months. Find Range and Coefficient of Range.

Month	January	February	March	April	May	June
Price/kg	125	110	160	130	165	160

Solution

Range = L-S

L =165, S = 110

Range = 165-110

= 55

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

$= \frac{165 - 110}{165 + 110}$

$= \frac{55}{275}$

= 0.2

For discrete series

Illustration :1.2.2

From the following data relating to the monthly income of 60 people, determine the range and coefficient of range.

Income:	210	240	290	360	440	510	500	350	290
No of person:	5	10	15	7	3	10	2	3	5

Solution

Range = L - S

= 510 – 210

= 300

$$\text{Coefficient of Range} = \frac{510-210}{510+210}$$

$$= \frac{300}{720}$$

$$= 0.417$$

For continuous series

Illustration: 1.2.3

The following table gives the age distribution of a group of 50 individuals.

Age (in years) :	16 – 20	21 – 25	26 – 30	31 – 35
No. of persons :	10	15	17	8

Calculate range and the coefficient of range.

Solution

Since age is a continuous variable, we should first convert the given classes into continuous classes.

Age	**x**	**f**
16-20	15.5 - 20.5	10
21-25	20.5 -.25.5	15
26-30	25.5 - 30.5	17
31-35	30.5.5 - 35.5	8

Largest value = 35·5; Smallest value = 15·5 ∴ Range = 35·5 – 15·5 = 20 years

$$\text{Coefficient of Range} = \frac{35.5-15.5}{35.5+15.5}$$

$$= \frac{20}{51}$$

$$= 0.39$$

1.2.2 Inter Quartile Range

The quartile deviation is a measure of dispersion based on quartiles. Quartiles are the points which divide the array in 4 equal parts. The interquartile range is a measure of dispersion based on the upper quartile Q_3 and lower quartile Q_1. The upper quartile (first quartile) is $\left(\frac{n+1}{4}\right)^{th}$ term and lower quartile (third quartile) is $3 \times \left(\frac{n+1}{4}\right)^{th}$ term of the observations.

Inter quartile range = $Q_3 - Q_1$

Inter quartile range divided by two is the Quartile Deviation. It gives the average amount by which the two quartiles differ from the median. In a symmetric distribution the two quartiles Q_1 and Q_3 are equidistant from the median. Small Quartile Deviation means that the variations among the central items are small and high Quartile Deviation means that the variations among the central items are high.

Quartile Deviation = $\dfrac{Q_3 - Q_1}{2}$

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

It can be used to compare the degree of variation in different distributions.

Merits of Quartile Deviation

- It is easy to understand and calculate
- It can be calculated for open-ended classes.
- It is unaffected by extreme values.

Demerits of Quartile Deviation

- It is not based on all observation. It ignores 50% of the observations
- It is not capable of further algebraic treatment.
- It is very much affected by samplying fluctuations.

Computation of inter quartile range

For individual series

Illustration :1.2.4

Compute the inter-quartile range, quartile deviation, and coefficient of quartile deviation from the following data:

X:	20	28	40	12	30	15	50

Solution

Arrange the data in ascending order

X:	12	15	20	28	30	40	50

$n = 7$

Q_1 = value of $\left(\frac{n+1}{4}\right)^{th}$ item

$= \left(\frac{7+1}{4}\right)^{th}$ item

= 2nd item

=15

Q_3 = value of 3 $\left(\frac{n+1}{4}\right)^{th}$ item

= 3 x 2nd item

= 6th item

= 40

Inter-quartile range = $Q_3 - Q_1$

= 40 - 15

= 25

Quartile Deviation = $\frac{Q_3 - Q_1}{2}$

$= \frac{25}{2}$

= 12.5

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

$= \frac{40 - 15}{40 + 15}$

$= \frac{25}{55}$

= 0.46

Illustration : 1.2.5

Compute the inter-quartile range, quartile deviation, and coefficient of quartile deviation from the following data:

X:	14	13	9	7	12	17	8	10	6	15	18	20	21

Solution

Arrange the data in ascending order

X: 6 7 8 9 10 12 13 14 15 17 18 20 21

$n = 13,$

Q_1 = value of $\left(\frac{n+1}{4}\right)^{th}$ item

= $\left(\frac{13+1}{4}\right)^{th}$ item

= 3.5th item

= 3rd item + 0.5 (4th item -3rd item)

= 8 + 0.5 (9-8)

= 8.5

Q_3 = value of $\left(\frac{n+1}{4}\right)^{th}$ item

= 3 x 3.5th item

= 10.5th item

= 10th item + 0.5 (11th item – 10th item)

= 17 + 05 (18-17)

=17.5

Inter-quartile range = $Q_3 - Q_1$

= 17.5 - 8.5

= 9

Quartile Deviation = $\frac{Q_3 - Q_1}{2}$

= $\frac{17.5 - 8.5}{2}$

= $\frac{9}{2}$

= 4.5

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

$$= \frac{17.5 - 8.5}{17.5 + 8.5}$$

$$= \frac{9}{26}$$

$$= 0.346$$

For discrete series

Illustration : 1.2.6

Below are the heights (in inches) of 43 people.

Height (in inches):	12	20	30	40	50	80
No of persons:	4	7	15	8	7	2

Calculate inter-quartile range, quartile deviation, and coefficient of quartile deviation.

Solution

Height (in inches)	**Frequency**	**Cum.f**
12	4	2
20	7	11
30	15	26
40	8	34
50	7	41
80	2	43
	43	

$n = 43$

Q_1 = Series having cum.f $\left(\dfrac{43+1}{4}\right)$

= Series having cf 11

= 20

Q_3 = Series having cf $3\left(\dfrac{43+1}{4}\right)$

= Series having cf 33

= 40

Inter-quartile range = $Q_3 - Q_1$

= 40-20

= 20

Quartile Deviation = $\frac{Q_3 - Q_1}{2}$

$= \frac{40 - 20}{2}$

$= \frac{20}{2}$

= 10

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

$= \frac{40 - 20}{40 + 20}$

$= \frac{20}{60}$

$= \frac{1}{3}$

Illustration : 1.2.7

Calculate inter-quartile range, quartile deviation, and coefficient of quartile deviation for the following data.

Value	58	59	60	61	62	63	64	65	66
Frequency	15	20	32	35	33	22	22	10	8

Solution

Value	Frequency	Cum.f
58	15	15
59	20	35
60	32	67

61	35	102
62	33	135
63	22	157
64	22	179
65	10	189
66	8	197
	197	

$n = 195$

Q_1 = Series having cf $\left(\frac{197+1}{4}\right)$

= Series having cf 49.5th

= 60

Q_3 = Series having cf $3\left(\frac{197+1}{4}\right)$

= Series having cf 148.5

= 63

Inter-quartile range = $Q_3 - Q_1$

= 63-60

= 3

Quartile Deviation = $\frac{Q_3 - Q_1}{2}$

$= \frac{3}{2}$

= 1.5

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

$= \frac{63 - 60}{63 + 60}$

$= \frac{3}{123} = 0.024$

Illustration : 1.2.8

Calculate inter-quartile range, quartile deviation, and coefficient of quartile deviation for the following data.

Size of item	2	4	6	8	10	12
Frequency	3	5	10	12	6	4

Solution

Size of item	**Frequency**	**Cum.f**
2	3	3
4	5	8
6	10	18
8	12	30
10	6	36
12	4	40
	40	

$n = 40$

Q_1 = Series having cf $\left(\frac{40+1}{4}\right)$

= Series having cf 10.25

= 6

Q_3 = Series having cf 3 $\left(\frac{40+1}{4}\right)$

= Series having cf 30.75

= 10

Inter-quartile range = $Q_3 - Q_1$

= 10-6

= 4

Quartile Deviation = $\frac{Q_3 - Q_1}{2}$

$$= \frac{4}{2}$$

$$= 2$$

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

$$= \frac{10 - 6}{10 + 6}$$

$$= \frac{4}{16}$$

$$= \frac{1}{4} = 0.25$$

For continuous series

Illustration : 1.2.9

Calculate the inter-quartile range, quartile deviation, and coefficient of quartile deviation.

Farm Size (acres)	0-40	41-80	81-120	121-160	161-200	200 above
No of farms	13	17	50	60	55	45

Solution

Farm size	**X**	**f**	**cf**
0 - 40	-0.5 - 40.5	13	13
41 - 80	40.5 - 80.5	17	30
81-120	80.5 - 120.5	50	80
121-160	120.5 - 160.5	60	140
161-200	160.5 - 200.5	55	195
Above 200	Above 200.5	45	240

Q_1Class $= \left(\frac{n}{4}\right)^{th}$ Class

$$= \left(\frac{240}{4}\right)^{th} \text{Class}$$

= 60 th Class

Series having cum. frequency 60 is 80.5-120.5 . ∴ Q_1 class is 80.5-120.5

$$Q_1 = l + \frac{(\frac{N}{4} - m_1)}{f_1} \times c_1$$

$$= 80.5 + \frac{(60-30)}{50} \times 40$$

$$= 80.5 + \frac{30 \times 40}{50}$$

$$= 80.5 + \frac{1200}{50}$$

$$= 80.5 + 24$$

$$= 104.5$$

$$Q_3 \text{ Class} = 3\left(\frac{n}{4}\right)^{th} \text{Class}$$

$$= 3\left(\frac{240}{4}\right)^{th} \text{Class}$$

= 180 th Class

160.5-200.5 is Q_3 class

$$Q_3 = l + \frac{(\frac{3N}{4} - m_1)}{f_1} \times c_1$$

$$= 160.5 + \frac{(180-140)}{55} \times 40$$

$$= 160.5 + \frac{40 \times 40}{55}$$

$$= 160.5 + 29.09$$

$$= 189.59$$

$Q_1 = 104.5$, $Q_3 = 189.59$

Inter quartile range $= Q_3 - Q_1$

$= 189.59 - 104.5$

$= 85.09$

Quartile Deviation $= \frac{Q_3 - Q_1}{2}$

$= \frac{85.09}{2}$

$= 42.545$

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

$= \frac{189.59 \ - \ 104.5}{189.59 + \ 104.5}$

$= \frac{85.09}{294.09}$

$= 0.289$

Illustration : 1.2.10

The following are the height of students in a class. Find the Quartile deviation.

Height (inches)	50-53	53-56	56-59	59-62	62-65	65-68
No of students	2	7	24	27	13	3

Solution

Height	**f**	**c f**
50-53	2	2
53-56	7	9
56-59	24	33
59-62	27	60
62-65	13	73
65-68	3	76

Q_1 Class $= \left(\frac{76}{4}\right)^{th}$ Class

$= 19$ th Class

56-59 is Q_1 class

$$Q_1 = l + \frac{(\frac{N}{4} - m_1)}{f_1} \times c_1$$

$$= 56 + \frac{(19-9)}{24} \times 3$$

$$= 56 + \frac{10\times3}{24}$$

$$= 56 + 1.25$$

$$= 57.25$$

Q_3 Class = $3\left(\frac{n}{4}\right)^{th}$ Class

$= 3\left(\frac{76}{4}\right)^{th}$ ^{h}Class

$= 57^{th}$ Class

59-62 is Q_3 class

$$Q_3 = l + \frac{(\frac{3N}{4} - m_1)}{f_1} \times c_1$$

$$= 59 + \frac{(57-33)}{27} \times 3$$

$$= 59 + \frac{24\times3}{27}$$

$$= 59 + \frac{72}{27}$$

$$= 59 + 2.67$$

$$= 61.67$$

$Q_1 = 57.25$, $Q_3 = 61.67$

Inter quartile range $= Q_3 - Q_1$

= 61.67 _ 57.25

= 4.42

Quartile Deviation $= \frac{Q_3 - Q_1}{2}$

$= \frac{4.42}{2}$

= 2.21

Illustration : 1.2.11

Find the inter-quartile range, quartile deviation, and coefficient of quartile deviation.

Marks	10-20	20-30	30-40	40-50	50-60	60-70	70-80	80-90
No of students	60	45	120	25	90	80	120	60

Solution

Mark	**F**	**c f**
10-20	60	60
20-30	45	105
30-40	120	225
40-50	25	250
50-60	90	340
60-70	80	420
70-80	120	540
80-90	60	600

Q_1 Class $= \left(\frac{n}{4}\right)^{th}$ Class

$= \left(\frac{600}{4}\right)^{th}$ Class

= 150 th Class

30-40 is Q_1 class

$$Q_1 = l + \frac{(\frac{N}{4} - m_1)}{f_1} \times c_1$$

$$= 30 + \frac{(150 - 105)}{120} \times 10$$

$$= 30 + \frac{145 \times 10}{120}$$

$$= 30 + 3.75$$

$= 33.75$

Q_3 Class = $3\left(\frac{n}{4}\right)^{th}$ Class

$= 3\left(\frac{600}{4}\right)^{th}$ ^{h}Class

= 450 th Class

70-80 is Q_3 class

$$Q_3 = l + \frac{(\frac{3N}{4} - m_1)}{f_1} \times c_1$$

$$= 70 + \frac{(450 - 420)}{120} \times 10$$

$$= 70 + \frac{30 \times 10}{120}$$

$$= 70 + \frac{300}{120}$$

$$= 70 + 2.5$$

$$= 72.5$$

$Q_1 = 33.75$, $Q_3 = 72.5$

Inter quartile range $= Q_3 - Q_1$

= 72.5 _ 33.75

= 38.75

Quartile Deviation = $\frac{Q_3 - Q_1}{2}$

$= \frac{38.75}{2}$

= 19.37

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

$$= \frac{72.5 - 33.75}{72.5 + 33.75}$$

$$= \frac{38.75}{106.25}$$

= 0.365

1.2.3 Standard deviation

The most important and widely used measure of dispersion is Standard deviation. It is the positive square root of the mean of the squares of deviation from the arithmetic mean. It is denoted by the Greek letter σ (sigma). It cannot be negative. The standard deviation concept was introduced by Karl Pearson in 1893. It is the most used methods of dispersion since it is free from some defects of other measures of dispersion.

The square of the Standard deviation σ^2 is termed as variance and is more often specified than standard deviation. It has the same properties as Standard deviation.

Merits of standard deviation

1. It is rigidly defined.
2. It is based on all observation.
3. It is never disregards the plus or minus sign
4. It can be subjected to more mathematical analysis.
5. The changes in sampling have little effect on it.
6. It allows us to compare and contrast two or more series and determine their consistency or stability.
7. It is used in testing of hypothesis

Demerits of standard deviation

1. A layman would find it difficult to comprehend.
2. It is complex to calculate since it incorporates several mathematical models.
3. It cannot be used to compare the dispersion of two or more series of observations with different units of measurement.

Coefficient of variation

The coefficient of variation is calculated by dividing the standard deviation by the arithmetic mean, which is given as a percentage. It is the most popular way of comparing the consistency or stability of two or more sets of data. The series for which the CV is greater is said to be more variable or less consistent or less stable. On the other hand, the series for which CV is less is said to be less variable or more consistent or more stable.

$$CV = \frac{\text{Standard Deviation}}{\text{Mean}} \times 100$$

$$= \frac{\sigma}{\bar{x}} \times 100$$

Computation of Standard deviation and Coefficient of variation

For individual series

Standard deviation $= \sigma = \sqrt{\frac{\Sigma(x - \bar{x})^2}{n}}$

Variance $= \sigma^2 = \frac{\Sigma(x - \bar{x})^2}{n}$

where,

$\bar{x}$ -Actual mean of the observation

n - Total number of items

Illustration -1.2.12

Calculate the standard deviation for the following data

1,3,5,7,4

Solution

$$\bar{x} = \frac{\Sigma x}{n}$$

$$= \frac{20}{5}$$

$$= 4$$

X	(x - 4)	$(x- 4)^2$
1	-3	9
3	-1	1
5	1	1
7	3	9
4	0	0
		20

Standard deviation $= \sqrt{\frac{\Sigma(x - \bar{x})^2}{N}}$

$$= \sqrt{\frac{20}{5}}$$

$$= \sqrt{4}$$

$$= 2$$

$$\text{CV} = \frac{\text{Standard Deviation}}{\text{Mean}} \times 100$$

$$= \frac{2}{4} \times 100$$

$$= 50\%$$

For discrete series

Standard deviation $= \sqrt{\frac{\Sigma f \times x^2}{\Sigma f} - \bar{x}^2}$

Where, f – Frequency, $\bar{x} = \frac{\Sigma f \times x}{\Sigma f}$

Illustration : 1.2.13

An arithmetic test was given to 100 students. The following is the time in minutes required to finish the test:

Time (in minute):	18	19	20	21	22	23	24	25	26	27
No of students	3	7	11	14	18	17	13	8	5	4

Calculate the standard deviation of their test completion time as well as the coefficient of variation.

Solution

x	f	fx	x^2	fx^2
18	3	54	324	972
19	7	133	361	2527
20	11	220	400	4400

21	14	294	441	6174
22	18	396	484	8712
23	17	391	529	8993
24	13	312	576	7488
25	8	200	625	5000
26	5	130	676	3380
27	4	108	729	2916
	N = 100	**2238**		**50562**

$$\bar{x} = \frac{\sum f\times x}{\sum f}$$

$$= \frac{2238}{100}$$

$$= 22.38$$

$$\text{Standard deviation} = \sqrt{\frac{\sum f\times x^2}{\sum f} - \bar{x}^2}$$

$$= \sqrt{\frac{50562}{100} - 22.38^2}$$

$$= \sqrt{505.62 - 500.8644}$$

$$= \sqrt{4.7556}$$

$$= 2.181$$

$$\text{CV} = \frac{\text{Standard Deviation}}{\text{Mean}} \text{ x } 100$$

$$= \frac{2.181}{22.38} \text{x}100$$

$$= 9.75\%$$

Illustration : 1.2.14

Find the standard deviation for the following data

Size:	6	9	12	15	18
Frequency	7	12	19	10	2

Solution

x	*f*	*fx*	x^2	fx^2
6	7	42	36	252
9	12	108	81	972
12	19	228	144	2736
15	10	150	225	2250
18	2	36	324	648
Total	**50**	564		6858

$$\overline{x} = \frac{\Sigma f \times x}{\Sigma f}$$

$$= \frac{564}{50}$$

$$= 11.28$$

$$\text{Standard deviation} = \sqrt{\frac{\Sigma f \times x^2}{\Sigma f} - \overline{x}^2}$$

$$= \sqrt{\frac{6858}{50} - 11.28^2}$$

$$= \sqrt{137.16 - 127.24}$$

$$= \sqrt{9.92}$$

$$= 3.15$$

Illustration : 1.2.15

Find the standard deviation for the following data. Also find the coefficient of variation.

No. of letters	2	3	4	5	6	7	8	9	10	11
Frequency	9	6	2	2	2	4	3	3	2	3

Solution

x	*f*	*fx*	x^2	fx^2
2	9	18	4	36
3	6	18	9	54
4	2	8	16	32
5	2	10	25	50
6	2	12	36	72
7	4	28	49	196
8	3	24	64	192
9	3	27	81	243
10	2	20	100	200
11	3	33	121	363
	N = 36	198		1438

$$\bar{x} = \frac{\sum fx}{N}$$

$$= \frac{198}{36}$$

$$= 5.5$$

$$\text{Standard deviation} = \sqrt{\frac{\sum f \times x^2}{\sum f} - \bar{x}^2}$$

$$= \sqrt{\frac{1438}{36} - 5.5^2}$$

$$= \sqrt{39.94 - 30.25}$$

$$= \sqrt{9.69}$$

$$= 3.11$$

$$CV = \frac{\text{Standard Deviation}}{\text{Mean}} \times 100$$

$$= \frac{3.11}{5.5} \text{x} 100$$

= 56.5%

Illustration -1.2.16

The score of 2 batsman A and B in 10 innings during a certain match are as under .

A	32	28	47	63	71	39	10	60	96	14
B	19	31	48	53	67	90	10	62	40	80

Who is the better batsman? Who is more consistent?

Solution

In order to decide as to which of the two batsman, A or B, is better player, we should find their average score. The one whose average is higher will be considered as a better batsman.

T determine the consistency we should determine the coefficient of variation. The less this coefficient of variation is more consistent.

A's $\bar{x} = \frac{\sum x}{n}$

$= \frac{460}{10}$

$= 46$

B's $\bar{x} = \frac{\sum x}{n}$

$= \frac{500}{10}$

$= 50$

A			B		
X	**(x - 46)**	**$(x - 46)^2$**	**X**	**(x -50)**	**$(x - 50)^2$**
32	-14	196	19	-31	961
28	-18	324	31	-19	361
47	1	1	48	-2	4
63	17	289	53	3	9
71	25	625	67	17	289

39	-7	49	90	40	1600
10	-36	1296	10	-40	1600
60	14	196	62	12	144
96	50	2500	40	-10	100
14	-32	1024	80	30	900
		6500			5968

A

Standard deviation $= \sqrt{\frac{\Sigma(x-\bar{x})^2}{N}}$

$= \sqrt{\frac{6500}{10}}$

$= \sqrt{650}$

$= 25.5$

$$CV = \frac{\text{Standard Deviation}}{\text{Mean}} \times 100$$

$$= \frac{25.5}{46} \text{x} 100$$

$= 55.4\%$

B

Standard deviation $= \sqrt{\frac{\Sigma(x-\bar{x})^2}{N}}$

$= \sqrt{\frac{5968}{10}}$

$= \sqrt{596.8}$

$= 24.43$

$$CV = \frac{\text{Standard Deviation}}{\text{Mean}} \times 100$$

$$= \frac{24.43}{50} \text{x} 100$$

$= 48.8\ \%$

B is better batsman since his average is 50 as compared to 46 of A.

B is more consistent since the coefficient of variation of B is less than the coefficient of variation of A

For continuous series

Standard deviation = $\sqrt{\frac{\Sigma f(x-\bar{x})^2}{\Sigma f}}$

OR

Standard deviation = $\sqrt{\frac{\Sigma(fx^2)}{\Sigma f} - (\bar{x})^2}$

Where,

f – Frequency

x – Mid value

Illustration : 1.2.17

The marks of 75 students in a class is given below.

Marks	1-3	3-5	5-7	7-9	9-11	11-13	13-15
No of students	1	9	25	35	17	10	3

Find the standard deviation and the coefficient of variation of the data.

Solution

Mark	**Mid value (x)**	**F**	**f x**	**(x- 8)**^{2}	**f(x- 8)**^{2}
1-3	2	1	2	36	36
3-5	4	9	36	16	144
5-7	6	25	150	4	100
7-9	8	35	280	0	0
9-11	10	17	170	4	68
11-13	12	10	120	16	160
13-15	14	3	42	36	108
Total		100	800		616

$$\overline{x} = \frac{\sum fx}{\sum f}$$

$$= \frac{800}{100}$$

$$= 8$$

Standard deviation $= \sqrt{\frac{\sum f(x-\overline{x})^2}{\sum f}}$

$$= \sqrt{\frac{616}{100}} = 2.48$$

$$Coefficient\,of\,Variation = \frac{Standard\,Deviation}{Mean} \times 100$$

Substitute the values:

$$Coefficient\,of\,Variation = \frac{2.48}{8} \times 100 = 31\%$$

Standard Deviation = 2.48

Coefficient of Variation = 31%

Illustration : 1.2.18

Find the standard deviation and the coefficient of variation of the data.

Wages (Rs)	30-32	32-34	34-36	36-38	38-40	40-42	42-44
No of labours	12	18	16	14	12	8	6

Solution

Mark	**F**	**Mid value (x)**	**f x**	x^2	$f\,x^2$
30-32	12	31	372	961	11532
32-34	18	33	594	1089	19602
34-36	16	35	560	1225	19600
36-38	14	37	518	1369	19166
38-40	12	39	468	1521	18252
40-42	8	41	328	1681	13448
42-44	6	43	258	1849	11094
Total	86		3098		112694

$$\bar{x} = \frac{\sum fx}{\sum f} = \frac{3098}{86} = 36.02$$

Standard deviation = $\sqrt{\frac{\Sigma(fx^2)}{\Sigma f} - (\bar{x})^2}$

$$= \sqrt{\frac{112694}{86} - 36.02^2}$$

$$= \sqrt{1310.39 - 1297.68}$$

$$= \sqrt{12.71}$$

$$= 3.57$$

$$Coefficient\, of\, Variation = \frac{3.57}{36.02} \times 100 = 9.91\%$$

Illustration : 1.2.19

Find the standard deviation of the following data.

Age (years)	Less than 10	Less than 20	Less than 30	Less than 40	Less than 50	Less than 60	Less than 70	Less than 80
No of Persons	15	30	53	75	100	110	115	125

Solution

Age	**Cum f**	**f**	**Mid value (x)**	xf	x^2	fx^2
0-10	15	15	5	75	25	375
10-20	30	15	15	225	225	3375
20-30	53	23	25	575	625	14375
30-40	75	22	35	779	1225	26950
40-50	100	25	45	1125	2025	50625
50-60	110	10	55	550	3025	30250
60-70	115	5	65	325	4225	21125
70-80	125	10	75	750	5625	56250
		125		4395		203325

$$\bar{x} = \frac{\sum fx}{\sum f}$$

$$= \frac{4395}{125}$$

$$= 35.16$$

$$\textbf{Standard deviation} = \sqrt{\frac{\Sigma(fx^2)}{\Sigma f} - (\bar{x})^2}$$

$$= \sqrt{\frac{203325}{125} - 35.16^2}$$

$$= \sqrt{1626.6 - 1236.23}$$

$$= \sqrt{390.37}$$

$$= 19.76$$

Illustration : 1.2.20

The standard deviation calculated from a set of 32 observations is 5. If the sum of the observations is 80, what is the sum of the square of the observations?

Solution

Given $n = 32,\ \sigma = 5,\ \sum x = 80$

$$\bar{x} = \frac{\sum x}{n}$$

$$= \frac{80}{32} = 2.5$$

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

$$5 = \sqrt{\frac{\sum x^2}{32} - 2.5^2}$$

Squaring both sides

$$25 = \frac{\sum x^2}{32} - 2.5^2$$

$$25 = \frac{\sum x^2}{32} - 6.25$$

$$\frac{\sum x^2}{32} = 25 + 6.25 = 31.25$$

$$\sum x^2 = 31.25 \times 32 = 1000$$

Illustration : 1.2.21

Suppose you are a teacher and you want to analyze the performance of two students.

Rahul	20	22	17	23	28
Manu	10	20	18	12	15

Determine which of the two students, Rahul or Manu, is the most consistent in terms of scoring.

Solution

Rahul's $\overline{x} = \frac{\sum x}{n}$

$= \frac{110}{5}$

$= 22$

Manu's $\overline{x} = \frac{\sum x}{n}$

$= \frac{75}{5}$

$= 15$

Rahul			**Manu**		
X	**(x -22)**	**$(x-22)^2$**	**X**	**(x -15)**	**$(x-15)^2$**
20	-2	4	10	-5	25
22	0	0	20	5	25
17	-5	25	18	3	9
23	1	1	12	-3	9

28	6	36	15	0	0
$\sum x = 110$		**66**	$\sum x = 75$		**68**

Rahul

Standard deviation $= \sqrt{\frac{\Sigma(x-\bar{x})^2}{N}}$

$= \sqrt{\frac{66}{5}}$

$= \sqrt{13.2}$

$= 3.63$

$$CV = \frac{\text{Standard Deviation}}{\text{Mean}} \times 100$$

$= \frac{3.63}{22} x100$

$= 16.5\%$

Manu

Standard deviation $= \sqrt{\frac{\Sigma(x-\bar{x})^2}{N}}$

$= \sqrt{\frac{68}{5}}$

$= \sqrt{13.6}$

$= 3.69$

$$CV = \frac{\text{Standard Deviation}}{\text{Mean}} \times 100$$

$= \frac{3.69}{15} x100$

$= 24.6\ \%$

In comparison to Manu, Rahul is more consistent in his scoring because his coefficient of variation is lower.

Combined standard deviation

The following formula can be used to calculate the combined standard deviation of two or more groups:

$$\sigma_{1.2} = \sqrt{\frac{n_1\sigma_1^2 + n_2\sigma_2^2 + n_1 d_1^2 + n_2 d_2^2}{n_1 + n_2}}$$

Where,

$\sigma_{1.2}$ - Combined standard deviation

σ_1 – Standard deviation of the first series

σ_2 - Standard deviation of the second series

d_1 – $(\bar{x}_1 - \bar{x}_{1.2})$

d_2 – $(\bar{x}_2 - \bar{x}_{1.2})$

$\bar{x}_{1.2}$ – Combined mean

$$\bar{x}_{1.2} - \frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2}$$

n_1 −Number of items of the first series

n_2 −Number of items of the second series

Illustration : 1.2.22

Calculate the combined standard deviation of the two Factories using the given information.

	Factory A	Factory B
Mean	63	54
SD	8	7
Number of item	50	40

Solution

$$\sigma_{1.2} = \sqrt{\frac{n_1\sigma_1^2 + n_2\sigma_2^2 + n_1 d_1^2 + n_2 d_2^2}{n_1 + n_2}}$$

$$\bar{x}_{1.2} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2}$$

$$= \frac{(50 \times 63)+(40 \times 54)}{50+40}$$

$$= \frac{3150+2160}{90}$$

$$= \frac{5310}{90}$$

$= 59$

$d_1 - (\bar{x}_1 - \bar{x}_{1.2})$

$= (63 - 59)$

$= 4$

$d_2 - (\bar{x}_2 - \bar{x}_{1.2})$

$= (54 - 59)$

$= -5$

$$\sigma_{1.2} = \sqrt{\frac{(50 \times 8^2) + (40 \times 7^2) + (50 \times 4^2) + (40 \times -5^2)}{50 + 40}}$$

$$= \sqrt{\frac{(50 \times 64) + (40 \times 49) + (50 \times 16) + (40 \times 25)}{90}}$$

$$= \sqrt{\frac{3200 + 1960 + 800 + 1000}{90}}$$

$$= \sqrt{\frac{6960}{90}}$$

$$= \sqrt{77.33}$$

= 8.79

Illustration : 1.2.23

Analysis of the monthly wages of two hospitals gave the following information.

	Hospital I	**Hospital II**
No. of staff	550	600
Average wages	60	48.5
Variance	100	144

Obtain the average wage and combined standard deviation of the two hospitals together.

Solution

$$\sigma_{1.2} = \sqrt{\frac{n_1\sigma_1^2 + n_2\sigma_2^2 + n_1 d_1^2 + n_2 d_2^2}{n_1 + n_2}}$$

$$\bar{x}_{1.2} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2}$$

$$= \frac{(550 \times 60)+(600 \times 48.5)}{550+600}$$

$$= \frac{33000+29100}{1150}$$

$$= \frac{62100}{1150}$$

$= 54$

$d_1 – (\bar{x}_1 - \bar{x}_{1.2})$

$= (60 – 54)$

$= 6$

$d_2 – (\bar{x}_2 - \bar{x}_{1.2})$

$= (48.5 \text{ -}54)$

$= \text{-}5.5$

$$\sigma_{1.2} = \sqrt{\frac{(550 \times 100) + (600 \times 144) + (550 \times 6^2) + (600 \times (-5.5)^2)}{550 + 600}}$$

$$= \sqrt{\frac{55000 + 86400 + 19800 + 18150}{1150}}$$

$$= \sqrt{\frac{179350}{1150}}$$

$$= \sqrt{155.96}$$

$= 12.49$

Illustration : 1.2.24

For a group containing 100 observations the mean $\bar{x} = 8$ and $\sigma = \sqrt{10.5}$. For 50 observations selected from these 100 observations, the mean and standard deviation are 10 and 2 respectively. Find the mean and standard deviation of the other half.

Solution

Given $n_1 + n_2 = 100,\ \ \bar{x}_{1.2} = 8,\ \ \sigma_{1.2} = \sqrt{10.5}\,, \bar{x}_1 = 10,\ \sigma_1^2 = 2^2$

$$\bar{x}_{1.2} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2}$$

$$8 = \frac{50 \times 10 + 50 \times \bar{x}_2}{100}$$

$$50 \times 10 + 50 \times \bar{x}_2 = 8 \times 100 = 800$$

$$50 \times \bar{x}_2 = 800 - 500 = 300$$

$$\bar{x}_2 = \frac{300}{50} = 6$$

$$d_1 - (\bar{x}_1 - \bar{x}_{1.2})$$

$$= (10 - 8)$$

$$= 2$$

$$d_2 - (\bar{x}_2 - \bar{x}_{1.2})$$

$$= (6 -8)$$

$$= -2$$

$$\sigma_{1.2} = \sqrt{\frac{n_1\sigma_1^2 + n_2\sigma_2^2 + n_1 d_1^2 + n_2 d_2^2}{n_1 + n_2}}$$

$$\sqrt{10.5} = \sqrt{\frac{(50 \times 4) + (50 \times \sigma_2^2) + (50 \times 4) + (50 \times 4)}{100}}$$

Squaring both sides

$$10.5 = \frac{(50 \times 4) + (50 \times \sigma_2^2) + (50 \times 4) + (50 \times 4)}{100}$$

$$(50 \times 4) + (50 \times \sigma_2^2) + (50 \times 4) + (50 \times 4) = 10.5 \times 100$$

$$(200) + (50 \times \sigma_2^2) + (200) + (200) = 1050$$

$$(50 \times \sigma_2^2) = 1050 - 200 - 200 - 200 = 1050 - 600 = 450$$

$$(\sigma_2^2) = \frac{450}{40} = 9$$

$$\sigma_2 = 3$$

Correction in mean and standard deviation

Because one or more observations in a series are inaccurate, the mean and standard deviation computed from the series may be incorrect. The correct values of those observations may be known only after the calculations are over. As a result, we must rectify the mean and standard deviation by taking the correct values of those observations into account.

Illustration : 1.2.25

The mean and standard deviation of 11 observations were calculated as 5 and 3.67, respectively. But later, it was identified that one item having a value of 2 was misread as 13. Calculate the correct mean and standard deviation.

Solution

Incorrect $\sum x = \bar{x} \times n$

$= 5 \text{ x } 11$

$= 55$

Correct $\sum x$ = Incorrect $\sum x$ – wrong item + correct item

$= 55–13+2$

$= 44$

Correct $\bar{x} = \frac{44}{11}$

$= 4$

Calculation of the correct Standard Deviation

$$= \sqrt{\frac{\sum x^2}{N} - (\bar{x})^2}$$

$$3.67 = \sqrt{\frac{\sum x^2}{11} - (5)^2}$$

Squaring both sides

$$3.67^2 = \frac{\sum x^2}{11} - 25$$

$$13.4689 + 25 = \frac{\sum x^2}{11}$$

$$38.4689 = \frac{\sum x^2}{11}$$

$\sum x^2 = 38.4689 \text{ x } 11$

Incorrect $\sum x^2 = 423.1579$

Correct $\sum x^2$ = Incorrect $\sum x^2$- square of wrong item + square of correct item.

$= 423.1579 – 13^2 + 2^2$

$= 423.1579 - 169 + 4$

$= 258.1579$

Correct SD $= \sqrt{\frac{258.1579}{11} - 4^2}$

$$= \sqrt{\frac{258.1579}{11} - 16}$$

$$= \sqrt{23.4689 - 16}$$

$$= \sqrt{7.4686}$$

$= 2.73$

Illustration : 1.2.26

For a group of 200 candidates the mean and standard deviation of scores were found to be 40 and 15 respectively. Later on it was discovered that the score 43 and 35 were wrongly written as 34 and 53 respectively. Find the corrected mean and standard deviation corresponding to the corrected figure,

Solution

Incorrect $\sum x = \bar{x}$ x n

$= 40$ x 200

$= 8000$

Correct $\sum x$ = Incorrect $\sum x$ – wrong item + correct item

$= 8000 - (34+53) + (43+35)$

$= 8000 - 87 + 78$

$= 7991$

$$\text{Correct } \bar{x} = \frac{7991}{200}$$

$= 39.955$

Calculation of the correct Standard Deviation

$$= \sqrt{\frac{\sum x^2}{N} - (\bar{x})^2}$$

$$15 = \sqrt{\frac{\sum x^2}{200} - (40)^2}$$

Squaring both sides

$$15^2 = \frac{\sum x^2}{200} - 1600$$

$$1600 + 225 = \frac{\sum x^2}{200}$$

$$1825 = \frac{\sum x^2}{200}$$

$\sum x^2 = 1825 \text{ x } 200$

Incorrect $\sum x^2 = 365000$

Correct $\sum x^2$ = Incorrect $\sum x^2$- square of wrong item + square of correct item.

$= 365000 – (34^2 +53^2)+ (43^2 +35^2)$

$= 365000 -3965 + 3074 = 364109$

Correct SD $= \sqrt{\frac{364109}{200} -39.955^2}$

$= \sqrt{224.143}$

$= 14.971$

Recap

- Range - difference between the largest and the least numbers in the set.
- Inter-quartile range- The difference between upper quartile and lower quartile.
- Quartile Deviation - half of the difference between the upper and lower quartile.
- Standard deviation - statistic that calculates the square root of the variance and measures the dispersion of a dataset relative to its mean.
- Standard deviation - calculated as the square root of variance by determining each data point’s deviation relative to the mean and the value of standard deviation cannot be negative.
- Coefficient of variation (CV) - ratio of the standard deviation to the mean.

Objective Questions

1. What average is obtained when a inter quartile range is divided by 2?
2. The standard deviation is calculated using which type of average?
3. What term is used to denote the positive square root of the mean of the squares of deviation from the arithmetic mean?

4. When comparing the consistency of two variables, what statistical tool is used?

5. What term is used to denote the difference between the highest and lowest value in a series?

Answers

1. Quartile deviation
2. Mean
3. standard deviation
4. Coefficient of variation
5. Range

Assignments

1. Calculate Range, Interquartile range and Quartile Deviation from the following data.

X:	4	8	3	9	16	10	14	20	18	15	21

2. A survey of domestic consumption of electricity in a village gave the following distribution of units consumed.

Units:	Below 100	100-200	200-300	300-400	400-500	500-600	600-700	Above 700
No of Consumers:	20	21	30	46	20	25	16	10

Find Quartile deviation and interquartile range.

3. The arithmetic mean and standard deviation of series of 20 items were calculated by a student as 20 cm. and 5 cm. respectively. But while calculating them an item 13 was misread as 30. Find the correct arithmetic mean and standard deviation.

4. The results of ten distinct class tests for two students, Radha and Syama, are shown here.

Radha:	25	50	45	30	70	42	36	48	34	60
Syama:	10	70	50	20	95	55	42	60	48	80

Determine which of the two students, Radha or Syama, is the most consistent in terms of scoring.

5. Below are the profits earned by 100 sole proprietorship businesses.

Profit in '000:	0-10	10-20	20-30	30-40	40-50	50-60
No of companies:	8	12	20	30	20	10

Calculate the standard deviation and the coefficient of variation of the data.

6. Obtain the Standard deviation of the following data

Value	90-99	80-89	70-79	60-69	50-59	40-49	30-39
Frequency	2	12	22	20	14	4	1

7. Two workers on the Same job show the following results over a long period of time.

	Worker A	Worker B
Mean	30	25
SD	6	4

Which worker appears to be more consistent.

References

1. Gupta, S.P, *Statistical Methods*, Sulthan Chand and Sons, New Delhi.

2. SC Gupta and VK Kapoor, *Fundamentals of Mathematical Stastistics*

3. Gupta S. and V.K. Kapoor, *Fundamentals of Applied Statistics*, S.Chand and Sons, New Delhi.

Suggested Readings

1. Jan Ubøe, *Introductory Statistics for Business and Economics*: Theory, *Exercises and Solutions*, Springer International Publishing.

2. T. Rajaretnam, *Statistics for Social Sciences*, Sage India.

3. Monga, G.S. *Mathematics and Statistics for Economics*, Vikas Publishing, New Delhi

2

BLOCK

Correlation

1

UNIT

Correlation Analysis

Learning Outcomes

After going through this unit, the learner will be able to:

- understand scatter diagram as a graphical representation of the relationship between two variables.
- understand different types of correlation.

Prerequisites

In our previous units, we studied about Measures of Central Tendency and Measures of Dispersion, where we primarily focused on understanding the characteristics of individual variables such as height, weight, age, marks, and wages. Now, let us consider a scenario with vast amounts of data, such as the annual income of all Indians. Processing such extensive data poses challenges, as it is impractical to grasp the entire dataset widely. Instead, condensing this mass of data into a single value becomes imperative. This single value should serve as a representative measure for all observations, offering a clear and concise overview of the entire dataset. Referred to as the Measure of Central Tendency, this central value becomes crucial for making informed decisions regarding the entire dataset. Arithmetic mean stands as one of the most widely used and popular measures of central tendency. Meanwhile, measures of dispersion illustrate how data is spread or scattered around the mean. Among these, Standard deviation emerges as the most commonly used measure. Both measures of central tendency and dispersion relate to the study of individual variables, constituting what is known as univariate analysis.

In real-life scenarios, we frequently encounter situations involving the analysis of multiple variables. For instance, consider the relationship between the price (P) of a product and the quantity demanded (Q). According to the law of demand, as the price of a product increases, the quantity demanded typically decreases, and vice versa. Here, price (P) serves as the independent variable, while quantity demanded (Q) is the dependent variable. The relationship between price and quantity demanded can be expressed through the equation $Q = f(P)$, where Q depends on the price level P. This measure of association between price and demand is called correlation.

Keywords

Positive correlation, Negative correlation, Scatter diagram

2.1.1 Correlation

In practice we come across a large number of problems involving the use of two or more than two variables. If two quantities vary in such a way that movement of one are accompanied by movement of other, these quantities are correlated. For example, there exists some relationship between price of a commodity and amount demanded, increase in rain fall up to a point and production of rice etc. The degree of relationship between the variables under consideration is measured through the correlation analysis. The measure of correlation is called correlation coefficient. It helps us in determining the degree of relationship between two or more variables.

If the variables do not have a relationship with each other, then there is no correlation.

2.1.2 Types of Correlation

Correlation is described or classified in different ways.

Three of the most important are:

I. Simple Correlation II. Multiple Correlation III. Partial Correlation

The distinction between simple, partial and multiple correlation is based up on the number of variables studied.

Simple Correlation:

When only two variables are studied, it is a case of simple correlation. For example, when one studies relationship between the marks secured by students and the attendance of students in a class, it is a problem of simple correlation.

The relationship between price of a commodity and amount demanded, increase in rain fall up to a point and production of rice etc. are examples of simple correlation.

Multiple Correlation

When three or more variables are studied, it is a case of multiple correlation. For example, when examining the relationship between student marks and factors such as attendance, teacher effectiveness, and the use of teaching aids, it exemplifies a scenario of multiple correlation. when we study the relationship between the yield of rice per acre and both the amount of rainfall and the number of fertilizers used, it is a problem of multiple correlation.

Partial Correlation

In case of partial correlation, one studies three or more variables but considers only two variables to be influencing each other and the effect of other influencing variables being held constant. For example, in the above example of relationship between student marks and attendance, the other variable influencing such as effective teaching of teacher, use of teaching aid like computer, smart board etc are assumed to be constant. In the study of production of rice, if we consider yield and rainfall only, maintaining the daily temperature as constant, it illustrates an example of partial correlation.

2.1.3 Scatter Diagram Method

The simple and the easiest method for studying correlation is scatter diagram method. The two variables *x* and *y* are plotted on a two-dimensional graph. Variable time is always be represented as *x* variable and the other variable as *y* variable. The graph thus plotted will represent the relationship between the variables. Thus, this study of relationship between two variables based on the graphical representation is called scatter diagram method. By looking to the scatter of the various points we can form an idea as to whether the variables are related or not. The greater the scatter of the plotted points on the chart, the lesser is the relationship between the two variables. The more closely the points come to a straight line, the higher the degree of relationship. If all the points lie on a straight line falling from the lower left-hand corner to the upper right-hand corner, correlation is said to be perfectly positive. On the other hand, if all the points are lying on a straight line falling from the upper left-hand corner to the lower right-hand corner, the correlation is said to be perfectly negative.

This method is unscientific and not popular now with the advanced methods of studying correlation which we will learn as Karl Pearson's method and Spearman's Rank Correlation Methods. This method is influenced by the personal judgement of the investigator and we use the advanced methods for studying correlation between the variables.

Uses of correlations

- Correlation can be used to build predictive models to estimate the relationship between two variables. For example, the correlation between interest rates and consumer spending can be used to predict how changes in interest rates will affect consumer spending.

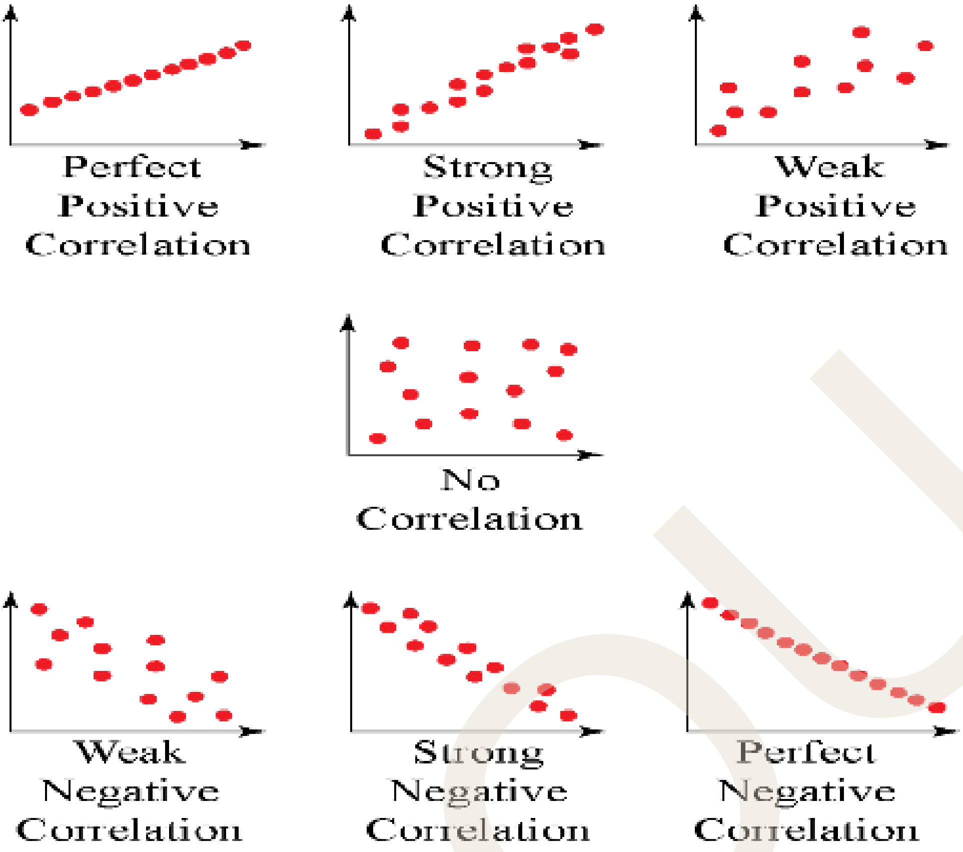


Fig 2.1.1 Types of Correlation

- ♦ It is used in quality control to measure the relationship between two variables that affect product quality. For example, in manufacturing, the correlation between temperature and product quality can be used to control the temperature to ensure that the product meets quality standards.

- ♦ It is used in survey research to measure the relationship between two variables that are measured using survey questions. For example, in social science research, the correlation between income and education level can be used to study the relationship between these two variables.

- ♦ It is used in medical research to study the relationship between two variables, such as a treatment and a disease outcome. For example, the correlation between smoking and lung cancer can be used to study the relationship between smoking and the risk of developing lung cancer.

- ♦ It is used in risk management to estimate the relationship between two risks. For example, in financial risk management, the correlation between two assets can be used to estimate the risk of a portfolio that includes both assets.

Recap

- If two quantities vary in such a way that movements in one are accompanied by movements in the other, then these quantities are said to be correlated.
- If both the variables are varying in the same direction, correlation is positive.
- If one variable is increasing, while the other decreases the correlation is said to be negative.
- When only two variables are studied, it is simple correlation.
- When three or more variables are studied it is multiple correlation.
- If we study only two variables and eliminate other factors, it is called partial correlation.
- The variables do not have a relationship with each other then there is no correlation.
- One of the methods of studying correlation is scatter diagram method.

Objective Questions

1. What is the measure of association between two variables?
2. If two variables move in opposite direction, which type of correlation is it?
3. Which type of correlation, if we consider more than two variables in correlation study?
4. Which type of correlation if the variables do not have a relationship with each other?
5. Which type of correlation, if we consider only two variables and eliminates some other factors or variables.
6. In the scatter diagram, if all the points lie on a straight line falling from the lower left-hand corner to the upper right-hand corner, which type of correlation is it?

7. In the scatter diagram, if all the points lie on a straight line falling from the upper left-hand corner to the lower right-hand corner, which type of correlation is it?

Answers

1. Correlation
2. Negative correlation
3. Multiple correlation
4. No correlation
5. Partial correlation
6. Perfectly positive
7. Perfectly Negative

Assignments

1. What is simple, partial and multiple correlation?
2. What are the uses of correlation?
3. Explain the meaning and significance of the concept of correlation.
4. Critically examine the different methods of ascertaining correlation with suitable example.

References

1. Gupta S. and V.K. Kapoor, (2014) *Fundamentals of Applied Statistics*, S.Chand and Sons, New Delhi.
2. Monga, G.S. (2001) *Mathematics and Statistics for Economics*, Vikas Publishing, New Delhi

Suggested Readings

1. Gupta, S.P, (2014) *Statistical Methods*, Sulthan Chand and Sons, New Delhi.

2. SC Gupta and VK Kapoor, (2020) *Fundamentals of Mathematical Stastistics,* Sulthanchand and Sons, NewDelhi

3. Jan Ubøe, (2018) *Introductory Statistics for Business and Economics: Theory, Exercises and Solutions*, Springer International Publishing.

4. T. Rajaretnam, (2016) *Statistics for Social Sciences*, Sage India.

UNIT 2

Measure of Correlation

Learning Outcomes

After going through this unit, the learner will be able to

- compute correlation coefficient between two variables
- compute Rank correlation coefficient when ranking rather than actual values for variables are known
- know practical applications of correlation

Prerequisites

In our day- to- day lives, we frequently encounter closely connected variables where changes in one variable correspond to changes in another. For instance, higher income often leads to an increased income tax. These interconnected variables, known as correlated variables, indicates that changes in one variable will coincide with changes in the other. Such correlations are evident in scenarios such as price of a commodity and the quantity demanded, fluctuation in rainfall and its impact on production, and the relationship between advertisement expenditure and sales.

Hence, there arises a necessity to study and understand the relationships between interconnected variables. This underscores the significance of grasping these connections assisting governments in crafting effective policies to address issues like unemployment and inflation. Additionally, it guides us in making informed decisions for our daily activities, contributing to better control over various situations and outcomes.

The statistical tool used to examine the relationship between two or more variables are known as correlation. The measure that quantifies this relationship is called the correlation coefficient, which provides us with a single value indicating both direction and strength of the correlation.

For example, a positive correlation indicates that variables are related in same direction, while a negative correlation indicated that the variables are related in opposite direction.

Similarly, the computed values reveal the extend of variation between the variables. As such, correlation coefficient values will have assign either positive or negative assignments. The computed value cannot exceed unity, as the value of correlation coefficient will always lie between plus or minus one. Consequently, correlation analysis refers to the techniques used in measuring the closeness or the relationship between the variables.

Keywords

Coefficient of correlation, Karl Pearson's Coefficient of Correlation, Rank Correlation

2.2.1 Methods of Studying Correlation

To determine the linearity and non- linearity among the variables, and the extent to which they are correlated, various methods are used to define and measure the correlation among the variables. The various methods of studying correlation coefficient are as follows.

1. Karl Pearson's Coefficient of Correlation
2. Spearman's Rank Correlation

2.2.2 Correlation coefficient

Degree of relationship between two variables is called coefficient of correlation. It is an algebraic method of measuring correlation. Coefficient of correlation is denoted by the symbol r and r lies between −1 and +1.

i.e., $-1 \leq r \leq 1$

Properties of Correlation coefficient

i. Coefficient of correlation lies between −1 and +1.

ii. When r lies between 0 and 1, the correlation is positive, when r lies between -1 and 0, the correlation is negative. If $r = 0$ there is no correlation.

iii. If $r = +1$ it is perfect positive correlation. If $r = -1$ it is called perfect negative correlation.

iv. It is a pure number lies between -1 and +1 and has no units.

Correlation coefficient does not change with reference to change of origin and scale. By origin, we mean that there will be no effect on the correlation coefficients if any constant is subtracted from the value of X and Y. By scale, we mean that if the value of X and Y is either multiplied or divided by some constant, then the correlation coefficients will not change.

2.2.3 Karl Pearson's Coefficient of Correlation

Of several mathematical methods of measuring correlation Karl Pearson's coefficient of correlation is the most widely used method for measuring correlation. It is popularly known as Pearson coefficient of correlation. It is denoted by the symbol "*r*". It is also known as Product Moment Method.

Computation of correlation coefficient

$$r(x,y) = \frac{Cov\ (x,y)}{\sigma(x)\sigma(y)}$$

where $Cov\ (x,y)$ = covariance of (x,y). Covariance is a statical measure that quantifies the degree to which two variables change together. It is the sum of the product of the average of the observations from arithmetic mean.

$$Cov\ (x,y) = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{n}$$

$$\sigma(x) = \sqrt{\frac{\Sigma(x-\bar{x})^2}{n}}$$ is the standard deviation of *x*.

$$\sigma(y) = \sqrt{\frac{\Sigma(y-\bar{y})^2}{n}}$$ is the standard deviation of *y*.

$\bar{x} = \frac{\Sigma x}{n}$, $\bar{y} = \frac{\Sigma y}{n}$ are the arithmetic means.

So,

$$r(x,y) = \frac{\frac{\Sigma(x-\bar{x})(y-\bar{y})}{n}}{\sqrt{\frac{\Sigma(x-\bar{x})^2}{n}}\sqrt{\frac{\Sigma(y-\bar{y})^2}{n}}} = \frac{\Sigma(x-\bar{x})\,(y-\bar{y})}{\sqrt{\Sigma(x-\bar{x})^2}\,\sqrt{\Sigma(y-\bar{y})^2}} = \frac{\Sigma\, dx\, dy}{\sqrt{\Sigma\, dx^2\, \Sigma\, dy^2}}$$

where $dx = x - \bar{x}$, $\quad dy = y - \bar{y}$

or

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

Computation of Pearsonian Correlation Coefficient Taking Deviations from an Assumed Mean

If the actual mean is in fractions, computing Pearsonian coefficient of correlation based on actual mean will be difficult. Therefore, unless specified in the question we can compute Pearsonian coefficient of correlation based on assumed mean. Assumed mean can be a value from the series or outside the series. Irrespective of the assumed mean value, the answer will be uniform. Moreover, this will simplify the calculations also.

The formula for computing Karl Pearson's coefficient of correlation using assumed mean

$$r = \frac{\sum \mathrm{dxdy} - \frac{(\sum dx)(\sum dy)}{n}}{\sqrt{\sum dx^2 - \frac{[\sum dx]^2}{n}} \times \sqrt{\sum dy^2 - \frac{[\sum dy]^2}{n}}}$$

where $dx = x - A$, $dy = y - A$, where A is assumed mean, n – no of observations

Illustration : 2.2.1

Calculate the coefficient of correlation for the data

x	6	8	9	14	17	28	24	31	7
y	10	12	15	15	18	25	22	26	28

Solution

x	$dx = x - 16$	dx^2	y	$dy = y - 19$	dy^2	$dx.dy$
6	-10	100	10	-9	81	90
8	-8	64	12	-7	49	56
9	-7	49	15	-4	16	28
14	-2	04	15	-4	16	8
17	1	1	18	-1	1	-1
28	12	144	25	6	36	72

24	8	64	22	3	9	24
31	15	225	26	7	49	105
7	-9	81	28	9	81	-81
Total 144	**0**	**732**	**171**	**0**	**338**	**301**

$$\bar{x} = \frac{\sum x}{n} = \frac{144}{9} = 16$$

$$\bar{y} = \frac{\sum y}{n} = \frac{171}{9} = 19$$

$$r(x, y) = \frac{\sum dx\, dy}{\sqrt{\sum dx^2 \sum dy^2}}$$

$$= \frac{301}{\sqrt{732 \times 338}}$$

$$= \frac{301}{\sqrt{247416}}$$

$$= \frac{301}{497.41}$$

$$= 0.605$$

Illustration : 2.2.2

Find Karl Pearson's coefficient of correlation between sales and expenses of the following ten firms :

Firm	1	2	3	4	5	6	7	8	9	10
Sales ('000 units)	50	50	55	60	65	65	65	60	60	50
Expenses ('000 rupees)	11	13	14	16	16	15	15	14	13	13

Solution:

x	y	$dx = x - 58$	$dy = y - 14$	dx^2	dy^2	$dx.dy$
50	11	-8	-3	64	9	24

50	13	-8	-1	64	1	8
55	14	-3	0	9	0	0
60	16	2	2	4	4	4
65	16	7	2	49	4	14
65	15	7	1	49	1	7
65	15	7	1	49	1	7
60	14	2	0	4	0	0
60	13	2	-1	4	1	-2
50	13	-8	-1	64	1	8
Total-580	**140**			**360**	**22**	**70**

$$\bar{x} = \frac{\sum x}{n} = \frac{580}{10} = 58$$

$$\bar{y} = \frac{\sum y}{n} = \frac{140}{10} = 14$$

$$r(x,y) = \frac{\sum dx\, dy}{\sqrt{\sum dx^2 \sum dy^2}}$$

$$= \frac{70}{\sqrt{360 \times 22}}$$

$$= \frac{70}{\sqrt{7920}}$$

$$= \frac{70}{88.99}$$

$$= 0.7866$$

Illustration : 2.2.3

Find Karl Pearson's coefficient of correlation between capital employed and profit obtained from the following data.

Capital Employed (Rs. In Crore)	10	20	30	40	50	60	70	80	90	100
Profit (Rs. In Crore)	2	4	8	5	10	15	14	20	22	50

Solution

x	y	$dx = x - 55$	$dy = y - 15$	dx^2	dy^2	$dx.dy$
10	2	-45	-13	2025	169	585
20	4	-35	-11	1225	121	385
30	8	-25	-7	625	49	175
40	5	-15	-10	225	100	150
50	10	-5	-5	25	25	25
60	15	5	0	25	0	0
70	14	15	-1	225	1	-15
80	20	25	5	625	25	125
90	22	35	7	1225	49	245
100	50	45	35	2025	1225	1575
Total-550	**150**			**8250**	**1764**	**3250**

$$\bar{x} = \frac{\sum x}{n} = \frac{550}{10} = 55$$

$$\bar{y} = \frac{\sum y}{n} = \frac{150}{10} = 15$$

$$r(x,y) = \frac{\sum dx\, dy}{\sqrt{\sum dx^2 \sum dy^2}}$$

$$= \frac{3250}{\sqrt{8250 \times 1764}}$$

$$= \frac{3250}{\sqrt{14553000}}$$

$$= \frac{3250}{3814.84}$$

$$= 0.8519$$

Illustration : 2.2.4

Using assumed mean, calculate Pearsonian coefficient of correlation for the following X and Y series.

x	45	55	56	58	60	65	68	70	75	80
y	56	50	48	60	62	64	65	70	74	82

Solution

Assumed mean $\bar{x} = 65,\ \bar{y} = 66$

x	y	$dx = x - 65$	$dy = y - 66$	dx^2	dy^2	$dx.dy$
45	56	-20	-10	400	100	200
55	50	-10	-16	100	256	160
56	48	-9	-18	81	324	162
58	60	-7	-6	49	36	42
60	62	-5	-4	25	16	20
65	64	0	-2	0	4	0
68	65	3	-1	9	1	-3
70	70	5	4	25	16	20
75	74	10	8	100	64	80
80	82	15	16	225	256	240
Total-632	**631**	**-18**	**-29**	**1014**	**1073**	**921**

$$r = \frac{\sum dxdy - \frac{(\sum dx)(\sum dy)}{n}}{\sqrt{\sum dx^2 - \frac{[\sum dx]^2}{n}} \times \sqrt{\sum dy^2 - \frac{[\sum dy]^2}{n}}}$$

$$= \frac{921 - \frac{18 \times 29}{10}}{\sqrt{1014 - \frac{(-18)^2}{10}}\sqrt{1073 - \frac{(-29)^2}{10}}}$$

$$= \frac{921 - 52.2}{\sqrt{1014 - 32.4}\,\sqrt{1073 - 84.1}}$$

$$= \frac{921 - 52.2}{\sqrt{981.6}\,\sqrt{988.9}}$$

$$= \frac{868.8}{31.33 \times 31.45}$$

$$= 0.882$$

Illustration : 2.2.5

Taking deviations from an assumed mean, find the correlation coefficient between X and Y variables are given below;

x	20	30	30	20	19	23	35	16	38	40
y	18	35	20	18	25	28	33	18	20	40

Solution

Assumed mean $\bar{x} = 30$, $\bar{y} = 30$

X	$x - 30 = dx$	$(dx)^2$	Y	$y - 30 = dy$	$(dy)^2$	$dxdy$
20	-10	100	18	-12	144	120
30	0	0	35	5	25	0
30	0	0	20	-10	100	0
20	-10	100	18	-12	144	120
19	-11	121	25	-5	25	55

23	-7	49	28	-2	4	14
35	5	25	33	3	9	15
16	-14	196	18	-12	144	168
38	8	64	20	-10	100	-80
40	10	100	40	10	100	100
Total-271	**-29**	**755**	**255**	**-45**	**795**	**512**

$$r = \frac{\sum \text{dxdy} - \frac{(\sum dx)(\sum dy)}{n}}{\sqrt{\sum dx^2 - \frac{[\sum dx]^2}{n}} \times \sqrt{\sum dy^2 - \frac{[\sum dy]^2}{n}}}$$

$$= \frac{512 - \frac{29 \times 45}{10}}{\sqrt{755 - \frac{(-29)^2}{10}}\sqrt{795 - \frac{(-45)^2}{10}}}$$

$$= \frac{512 - 130.5}{\sqrt{755 - 84.1}\,\sqrt{795 - 202.5}}$$

$$= \frac{381.5}{\sqrt{670.9} \times \sqrt{592.5}}$$

$$= \frac{381.5}{25.9017 \times 24.3413}$$

$$= \frac{381.5}{630.48}$$

$$= 0.605$$

Illustration : 2.2.6

Given: $\sum X = 125, \sum Y = 100, \sum X^2 = 650, \sum Y^2 = 436, \sum XY = 520$ *and* $n = 25$, obtain the value of Karl Pearson's correlation coefficient $r(X,Y)$.

Solution

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

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

$$= \frac{13000 - 12500}{\sqrt{16250 - 15625\sqrt{10900 - 10000}}}$$

$$= \frac{500}{\sqrt{625}\sqrt{900}}$$

$$= \frac{500}{25 \times 30}$$

$$= \frac{500}{750}$$

$$= 0.67$$

Illustration 2.2.7

A computer while calculating the correlation coefficient between the variable X and Y obtained the following results:

$n = 30,\ \sum X = 120,\ \sum X^2 = 600,\ \sum Y = 90,\ \sum Y^2 = 250,\ \sum XY = 335.$

It was later discovered at the time of checking that it had copied down two pairs of observations as: $(X, Y):\ (8, 10)\ (12, 7)$ While the correct values were:

$(X, Y):\ (8, 12)\ (10, 8)$ Obtain the correct value of the correlation coefficient between X and Y.

Solution

Correct $\sum X = 120 - 8 - 12 + 8 + 10 = 118$

Correct $\sum X^2 = 600 - 8^2 - 12^2 + 8^2 + 10^2 = 600 - 64 - 114 + 64 + 100 = 556$

Correct $\sum Y = 90 - 10 - 7 + 12 + 8 = 93$

Correct$\sum Y^2 = 250 - 10^2 - 7^2 + 12^2 + 8^2 = 250 - 100 - 49 + 144 + 64 = 309$

Correct

$\sum XY = 335 - (8 \times 10) - (12 \times 7) + (8 \times 12) + (10 \times 8) = 335 - 80 - 84 + 96 + 80 = 347$

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

$$= \frac{30 \times 347 - 118 \times 93}{\sqrt{30 \times 556 - 118^2}\sqrt{30 \times 309 - 93^2}}$$

$$= \frac{10410 - 10974}{\sqrt{16680 - 13924}\sqrt{9270 - 8649}}$$

$$= \frac{-564}{\sqrt{2756}\sqrt{621}}$$

$$= \frac{-564}{52.4976 \times 24.9199}$$

$$= \frac{-564}{1308.23}$$

$$= -0.43$$

2.2.4 Rank Correlation Coefficient

In instances where statistical series involve variables that cannot be quantitatively measured but can be ordered sequentially, such as honesty, beauty, character, morality, and similar attributes, Karl Pearson's coefficient of correlation becomes impractical. These qualitative characteristics lack numerical values but can be organised in a serial manner. To address this challenge, Charles Edward Spearman, a British psychologist, introduced a formula in 1904. This formula calculates the correlation coefficient between the ranks of n individuals concerning the two attributes being studied. By focusing on the ordinal arrangement of the data rather than their specific quantitative values, Spearman's formula provides a method to analyse and understand relationships between attributes that challenge direct numerical measurement.

In other words, the Rank Correlation Coefficient between two variables is a correlation coefficient obtained based on the ranking of the variables.

Spearman's Rank correlation Coefficient,

$$r = 1 - 6\frac{\sum D^2}{n(n^2-1)}$$

Where, r is the Rank correlation coefficient.

D is the difference of the corresponding ranks.

n is the number of items.

In rank correlation coefficient we may have three types of problems

1. When actual ranks are given,

2. When ranks are not given.

3. When ranks are Equal

When Actual Ranks are Given

If the actual ranks are given, the steps required for computing Spearman's Correlation Coefficient are;

Take the differences of the two ranks, that is $(R_1 - R_2)$ and denote these differences by D

Square these differences and obtain the total $\sum D^2$

Apply the formula $r = 1 - 6\frac{\sum D^2}{n(n^2-1)}$

Illustration.2.2.8

Two judges in a beauty contest rank the twelve entries as follows. Compute the Spearman's Rank Correlation and interpret the data.

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

Solution

R_1	R_2	$d = R_1 - R_2$	d^2
1	12	-11	121
2	9	-7	49
3	6	-3	9
4	10	-6	36
5	3	2	4
6	5	1	1
7	4	3	9
8	7	1	1
9	8	1	1
10	2	8	64

11	11	0	0
12	1	11	121
			416

$$r = 1 - 6 \times \frac{\sum D^2}{n(n^2 - 1)}$$

$$= 1 - 6 \times \frac{416}{12(12^2 - 1)}$$

$$= 1 - 6 \times \frac{416}{12 \times 143}$$

$$= 1 - \frac{2496}{1716}$$

$$= 1 - 1.454$$

$$= -0.455$$

Illustration.2.2.9

Ten competitors in a beauty competition were ranked by three judges X, Y and Z in the following order;

Judge I	1	6	5	10	3	2	4	9	7	8
Judge II	3	5	8	4	7	10	2	1	6	9
Judge III	6	4	9	8	1	2	3	10	5	7

Use Rank Correlation Coefficient to determine which pair of judges has the nearest approach to the common taste in beauty?

Solution

Ranks by judge X is denoted as R$_1$, by Judge Y as R$_2$ and by Judge Z as R$_{3.}$ So, we have to calculate the rank correlation coefficient between R$_1$ and R$_{2,}$ R$_2$ and R$_3$ and R$_1$ and R$_{3.}$

R_1	R_2	R_3	$D_1 = R_1 - R_2$	D_1^2	$D_2 = R_2 - R_3$	D_2^2	$D_3 = R_1 - R_3$	D_3^2
1	3	6	-2	4	-3	9	-5	25
6	5	4	1	1	1	1	2	4

5	8	9	-3	9	-1	1	-4	16
10	4	8	6	36	-4	16	2	4
3	7	1	-4	16	6	36	2	4
2	10	2	-8	64	8	64	0	0
4	2	3	2	4	1	1	1	1
9	1	10	8	64	-9	81	-1	1
7	6	5	1	1	1	1	2	4
8	9	7	-1	1	2	4	1	1
Total				**200**		**214**		**60**

Rank correlation coefficient between 1st and 2nd Judge

$$r = 1 - 6\frac{\sum D_1{}^2}{n(n^2-1)}$$

$$= 1 - 6\frac{200}{10(10^2-1)}$$

$$= 1 - 6\frac{200}{10 \times 99}$$

$$= 1 - \frac{1200}{990}$$

$$= 1 - 1.212$$

$$= -0.212$$

Rank correlation coefficient between 2nd and 3rd Judge

$$r = 1 - 6\frac{\sum D_2{}^2}{n(n^2-1)}$$

$$= 1 - 6\frac{214}{10(10^2-1)}$$

$$= 1 - \frac{1284}{990}$$

$$= 1 - 1.297$$

$= -0.297$

Rank correlation coefficient between 1st and 3rd Judge

$$r = 1 - 6\frac{\sum D_3{}^2}{n(n^2-1)}$$

$$= 1- 6\frac{60}{10(10^2-1)}$$

$$= 1 - \frac{360}{990}$$

$= 1 - 0.363$

$= -0.636$

Since coefficient of correlation is maximum in the judgement of the first and third judges we conclude that they have the nearest approach to common tastes in beauty.

When Ranks are Not Given

When we are given the actual data and not the ranks, it will be necessary to assign the ranks. Ranks can be assigned by taking either the highest value as the first rank or the lowest value as the first rank. But whether we start with the highest value or the lowest value we must follow same method for both the variables.

Illustration : 2.2.10

Calculate Spearman's Coefficient of Correlation from the following data;

X	92	89	87	86	83	77	71	63	53	50
Y	86	83	91	77	68	85	52	82	37	57

Solution

We have to assign ranks for the given X and Y values._Here ranks are assigned by taking the highest value as one, therefore from X series 92 got the first rank, 89 second rank and so on. For Y series, 86 got the first rank and 83 the second rank and so on.

X	R_1	Y	R_2	$D = R_1 - R_2$	D^2
92	1	86	2	-1	1
89	2	83	4	-2	4
87	3	91	1	2	4
86	4	77	6	-2	4

83	5	68	7	-2	4
77	6	85	3	3	9
71	7	52	9	-2	4
63	8	82	5	3	9
53	9	37	10	-1	1
50	10	57	8	2	4
Total					**44**

$$r = 1 - 6\frac{\sum D^2}{n(n^2-1)}$$

$$= 1 - 6\ \frac{44}{10(10^2-1)}$$

$$= 1 - \frac{264}{990}$$

$$= 1 - 0.267$$

$$= 0.733$$

Illustration : 2.2.11

Find the Spearman's rank correlation coefficient between marks in economics and Statistics

Marks in Statistics	48	60	72	62	56	40	39	52	30
Marks in Economics	62	78	65	70	38	54	60	32	31

Solution

X	R_1	Y	R_2	$D = R_1 - R_2$	D^2
48	6	62	4	2	4
60	3	78	1	2	4
72	1	65	3	-2	4
62	2	70	2	0	0

56	4	38	7	¯3	9
40	7	54	6	1	1
39	8	60	5	3	9
52	5	32	8	¯3	9
30	9	31	9	0	0
					40

$$r = 1 - 6\frac{\sum D^2}{n(n^2 - 1)}$$

$$= 1 - 6\ \frac{40}{9(9^2 - 1)}$$

$$= 1 - 6\ \frac{40}{9 \times 80}$$

$$= 1 - \frac{240}{720}$$

$$= 1 - 0.3333$$

$$= 0.6667$$

Illustration : 2.2.12

Find the Spearman's rank correlation coefficient between marks in economics and Statistics

X	78	89	69	97	59	57	79	68	83	64
Y	125	137	156	107	112	118	123	138	115	122

Solution

X	**R_1**	**Y**	**R_2**	**$D = R_1\text{-}R_2$**	**D^2**
78	5	125	4	1	1
89	2	137	3	-1	1
69	6	156	1	5	25
97	1	107	10	-9	81

59	9	112	9	0	0
57	10	118	7	3	9
79	4	123	5	-1	1
68	7	138	2	5	25
83	3	115	8	-5	25
64	8	122	6	2	4
					172

$$r = 1 - 6\frac{\sum D^2}{n(n^2-1)}$$

$$= 1 - 6\,\frac{172}{10(10^2-1)}$$

$$= 1 - 6\,\frac{172}{10 \times 99}$$

$$= 1 - \frac{1032}{990}$$

$$= 1 - 1.04$$

$$= -0.042$$

Ranks are Equal

In some cases, we might encounter cases where two or more items share equal ranks. In such cases, it is costumery to give each individual item an average rank. For example, if two values are both ranked equal at the third place and second place, each is given a rank of $\frac{2+3}{2} = 2.5$

However, when three values are ranked equally at the third place, fourth place and fifth place the individual ranks are calculated as $\frac{3+4+5}{3} = 4$. When equivalent ranks are assigned to multiple entries, certain adjustments in the formula become necessary for calculating the Rank Correlation coefficient. This adjustment involves adding $\frac{m^3-m}{12}$ to the sum of squared differences $\sum D^2$,where m stands for the number of items which have the common rank. In case, there are more than one such group of items with same rank, the value is added as many times as the number of such groups. The formula in that case is written as

$$r = 1 - \frac{6[(\Sigma D^2 + \frac{1}{12}(m^3 - m) + \frac{1}{12}(m^3 - m) + \ldots\ldots]}{n(n^2 - 1)} + \cdots$$

Illustration : 2.2.13

A psychologist wanted to compare two methods A and B of teaching. He selected a random sample of 22 students. He grouped them into 11 pairs so that the students in a pair have approximately equal scores on an intelligence test. In each pair one student was taught by method A and the other by method B and examined after the course. The marks obtained by them are tabulated below :

Pair	1	2	3	4	5	6	7	8	9	10	11
I	24	29	19	14	30	19	27	30	20	28	11
II	37	35	16	26	23	27	19	20	16	11	21

Solution

Ranks are assigned as follows for x series;

The highest value is 30 which is repeated twice, therefore average of the first and second ranks will be taken, that is $\frac{1+2}{2}$ = 1.5 is assigned to 30 which is the first and second rank. 29 gets the third rank, 28 gets the fourth rank, 27 gets the fifth rank, 24 gets the sixth rank, 20 gets the seventh rank, Now the next highest value 19 is repeated twice, therefore average of the next two ranks will be taken, that is $\frac{8+9}{2}$ = 8.5 is assigned to 8th and 9th rank. 14 gets the tenth rank, 11 gets the eleventh rank.

Now, let us explain how the ranks for y series are assigned.

The highest value is 37 gets first rank, 35 gets the second rank, 27 gets the third rank, 26 gets the fourth rank, 23 gets the fifth rank, 21 gets the sixth rank, 20 gets the seventh rank, 19 gets eighth rank. Now the next highest value 16 is repeated twice, therefore average of the next two ranks will be taken, that is $\frac{9+10}{2}$ = 9.5 is assigned to 9th and 10th rank. 11 gets the eleventh rank.

X	**R_1**	**Y**	**R_2**	**$D = R_1\text{-}R_2$**	**D^2**
24	6	37	1	5	25
29	3	35	2	1	1
19	8.5	16	9.5	-1	1
14	10	26	4	6	36

30	1.5	23	5	-3.5	12.25
19	8.5	27	3	5.5	30.25
27	5	19	8	-3	9
30	1.5	20	7	-5.5	30.25
20	7	16	9.5	-2.5	6.25
28	4	11	11	-7	49
11	11	21	6	5	25
Total					**225**

Hence, we see that in the X-series the items 19 and 30 are repeated, each occurring twice and in the Y-series the item 16 is repeated. Thus in each of the three cases $m = 2$

Hence on applying the correction factor $\frac{m(m^2-1)}{12}$ for each repeated item, we get

$$r = 1 - \frac{6[(\Sigma D^2 + \frac{1}{12}(m^3 - m) + \frac{1}{12}(m^3 - m) + \frac{1}{12}(m^3 - m) \;\dots\dots]}{n(n^2 - 1)} \;\dots.$$

$$= 1 - \frac{6[225 + \frac{1}{12}(2^3 - 2) + \frac{1}{12}(2^3 - 2) + \frac{1}{12}(2^3 - 2)]}{11(11^2 - 1)}$$

$$= 1 - \frac{6[225 + \frac{1}{12}(6) + \frac{1}{12}(6) + \frac{1}{12}(6)]}{11(121 - 1)}$$

$$= 1 - \frac{6[225 + \frac{1}{2} + \frac{1}{2} + \frac{1}{2}]}{11(121 - 1)}$$

$$= 1 - \frac{6[225 + 1.5]}{11 \times 120}$$

$$= 1 - \frac{6 \times 226.5}{11 \times 120}$$

$$= 1 - \frac{1359}{1320}$$

$$= 1 - 1.0295$$

$= -0.0295$

Illustration : 2.2.14

Calculate the coefficient of rank correlation from the following data,

X	48	33	40	9	16	16	65	24	16	57
Y	13	31	31	6	15	4	20	9	6	19

Solution

Ranks are assigned as follows for x series;

We see that 16 is repeating three times. i.e, 7th , 8th and 9th rank is repeating. So the average $\frac{7+8+9}{3} = \frac{24}{3}$ = 8 is assigned to 16. So $m = 3$.

Ranks assigned for y series are as follows;

31 is repeating two times. i.e, 1st and 2nd rank is repeating. So the average $\frac{1+2}{2} = \frac{3}{2}$ = 1.5 is assigned to 31. So $m = 2$. 6 is repeating two times. i.e, 8th and 9th rank is repeating. So the average $\frac{8+9}{2} = \frac{17}{2}$ = 8.5 is assigned to 6 . So $m = 2$.

X	R_1	Y	R_2	$D=R_1-R_2$	D^2
48	3	13	6	-3	9
33	5	31	1.5	3.5	12.25
40	4	31	1.5	2.5	6.25
9	10	6	8.5	1.5	2.25
16	8	15	5	3	9
16	8	4	10	-2	4
65	1	20	3	-2	4
24	6	9	7	-1	1
16	8	6	8.5	-0.5	0.25
57	2	19	4	-2	4
Total					**52**

$$r = 1 - \frac{6[(\Sigma D^2 + \frac{1}{12}(m^3 - m) + \frac{1}{12}(m^3 - m) + \dots\dots]}{n(n^2 - 1)} \dots.$$

$$r = 1 - \frac{6[52 + \frac{1}{12}(3^3 - 3) + \frac{1}{12}(2^3 - 2) + \frac{1}{12}(2^3 - 2)\,]}{n\,(n^2 - 1)}$$

$$= 1 - \frac{6[52 + \frac{1}{12}(24) + \frac{1}{12}(6) + \frac{1}{12}(6)\,]}{10(10^2 - 1)}$$

$$= 1 - \frac{6[52 + 2 + \frac{1}{2} + \frac{1}{2}]}{10(100 - 1)}$$

$$= 1 - \frac{6[52 + 3\,]}{10 \times 99}$$

$$= 1 - \frac{6 \times 55}{10 \times 99}$$

$$= 1 - \frac{330}{990}$$

$$= 1 - 0.33$$

$$= 0.67$$

llustration : 2.2.15

The coefficient of rank correlation between Micro-Economics and Statistics marks of 10 students was found to be 0·5. It was later discovered that the difference in ranks in two subjects obtained by one of the students was wrongly taken as 3 instead of 7. Find the correct value of coefficient of rank correlation.

Solution

$n = 10, \qquad r = 0.5$

$$0.5 = 1 - \frac{6 \times \sum d^2}{n(n^2 - 1)}$$

$$= 1 - \frac{6 \times \sum d^2}{10(100 - 1)}$$

$$0.5 = 1 - \frac{6 \times \sum d^2}{10 \times 99}$$

$$\frac{6 \times \sum d^2}{10 \times 99} = 1 - 0.5 = 0.5$$

$$\frac{6 \times \sum d^2}{990} = 0.5$$

$$\sum d^2 = \frac{0.5 \times 990}{6}$$

$$= \frac{495}{6}$$

$$= 82.5$$

Since one difference was wrongly taken as 3 instead of 7,

the correct value of $\sum d^2 = 82.5 - 3^2 + 7^2 = 82.5 - 9 + 49 = 122.5$

Correct value of

$$r = 1 - \frac{6 \times 122.5}{10 \times 99} = 1 - \frac{735}{990} = 1 - 0.7424 = 0.2576$$

Illustration : 2.2.16

Eight students have obtained the following marks in Economics and Accountancy. Calculate the rank correlation coefficient.

Marks in Accountancy	25	30	38	22	50	70	30	90
Marks in Economics	50	40	60	40	30	20	40	70

Solution

X	Y	R_1	R_2	$D = R_1 - R_2$	D^2
25	50	7	3	4	16
30	40	5.5	5	0.05	0.25
38	60	4	2	2	4
22	40	8	5	3	9

50	30	3	7	-4	16
70	20	2	8	-6	36
30	40	5.5	5	0.05	0.25
90	70	1	1	0	0
					81.5

Here two correction factors are to be added to the equation, for X series, 30 is repeated twice, so the correction factor is $\frac{2^3 - 2}{12}$ is added. Similarly for Y

Series value 40 is repeated thrice, so the correction factor $\frac{3^3 - 3}{12}$ is added.

The rank correlation coefficient is

$$r = 1 - \frac{6[(\Sigma D^2 + \frac{1}{12}(m^3 - m) + \frac{1}{12}(m^3 - m) + \dots\dots]}{n(n^2 - 1)} \dots.$$

$$= 1 - \frac{6[81.5 + \frac{1}{12}(2^3 - 2) + \frac{1}{12}(3^3 - 3)]}{8(8^2 - 1)}$$

$$= 1 - \frac{6[81.5 + \frac{1}{12}(6) + \frac{1}{12}(24)]}{8(8^2 - 1)}$$

$$= 1 - \frac{6[81.5 + \frac{1}{2} + 2]}{8(64 - 1)}$$

$$= 1 - \frac{6[81.5 + 2.5]}{8 \times 63}$$

$$= 1 - \frac{6 \times 84}{8 \times 63}$$

$$= 1 - 1$$

$$r = 0$$

This indicates that there is no relationship between the marks in economics and accountancy among the eight students.

Illustration : 2.2.17

Given the following aptitude and I.Q. scores for a group of students. Find the coefficient of rank correlation

Aptitude Score	57	58	59	59	60	61	60	64
I.Q. Score	97	108	95	106	120	126	113	110

Solution

X	Y	R_1	R_2	$d = R_1 - R_2$	d^2
57	97	8	7	1	1
58	108	7	5	2	4
59	95	5.5	8	-2.5	6.25
59	106	5.5	6	-0.5	0.25
60	120	3.5	2	1.5	2.25
61	126	2	1	1	1
60	113	3.5	3	0.5	0.25
64	110	1	4	-3	9
					24

Here two correction factors are to be added to the equation, for X series, 59 is repeated twice, so the correction factor is $\frac{2^3 - 2}{12}$ is added and 60 is repeated twice, so the correction factor is $\frac{2^3 - 2}{12}$ is added.

The rank correlation coefficient is

$$r = 1 - \frac{6[(\Sigma D^2 + \frac{1}{12}(m^3 - m) + \frac{1}{12}(m^3 - m) + \dots\dots]}{n(n^2 - 1)} \dots.$$

$$r = 1 - \frac{6[24 + \frac{1}{12}(2^3 - 2) + \frac{1}{12}(2^3 - 2)]}{8(8^2 - 1)}$$

$$r = 1 - \frac{6[24 + \frac{1}{12}(6) + \frac{1}{12}(6)]}{8(8^2 - 1)}$$

$$r = 1 - \frac{6[24 + \frac{1}{2} + \frac{1}{2}]}{8(64 - \ 1)}$$

$$r = 1 - \frac{6[25\ ]}{8 \times 63}$$

$$r = 1 - \frac{150}{504}$$

$$r = 1 - \ 0.2976$$

$$r = 0.7024$$

Illustration : 2.2.18

Find the coefficient of rank correlation for the following data.

x	68	64	75	50	64	80	75	40	55	64
y	62	58	68	45	81	60	68	48	50	70

Solution

X	Y	R_1	R_2	$d = R_1 - R_2$	d^2
68	62	4	5	-1	1
64	58	6	7	-1	1
75	68	2.5	3.5	-1	1
50	45	9	10	-1	1
64	81	6	1	5	25
80	60	1	6	-5	25
75	68	2.5	3.5	-1	1
40	48	10	9	1	1
55	50	8	8	0	0
64	70	6	2	4	16
Total					**72**

Here two correction factors are to be added to the equation, for X series, 75 is repeated twice, so the correction factor is $\frac{2^3-2}{12}$ is added and 64 is repeated trice, so the correction factor is $\frac{3^3-3}{12}$ is added.

Y series, 68 is repeated twice, so the correction factor is $\frac{2^3-2}{12}$ is added.

The rank correlation coefficient is

$$r = 1 - \frac{6[(\Sigma D^2 + \frac{1}{12}(m^3 - m) + \frac{1}{12}(m^3 - m) + \ldots\ldots]}{n(n^2 - 1)} \ldots.$$

$$= 1 - \frac{6[72 + \frac{1}{12}(2^3 - 2) + \frac{1}{12}(3^3 - 3) + \frac{1}{12}(2^3 - 2)\,]}{10(10^2 - 1)}$$

$$= 1 - \frac{6[72 + \frac{1}{12}(6) + \frac{1}{12}(24) + \frac{1}{12}(6)\,]}{10(10^2 - 1)}$$

$$= 1 - \frac{6[72 + \frac{1}{2} + 2 + \frac{1}{2}]}{10(100 - 1)}$$

$$= 1 - \frac{6[75\,]}{10 \times 99}$$

$$= 1 - \frac{450}{990}$$

$$= 1 - 0.455$$

$$r = 0.545$$

Recap

- Degree of relationship between two variables is called coefficient of correlation.
- Karl Pearson's Coefficient of Correlation is the most widely used method for measuring correlation.
- Degree of Correlation is used to interpret the computed value of Karl Pearson's Correlation coefficient.
- The correlation coefficient obtained from the ranks is called Rank correlation coefficient.

Objective Questions

1. What is the purpose of measuring correlation between variables?
2. Define the Karl Person correlation coefficient.
3. Explain the interpretation of a correlation coefficient value of -0.9
4. If the computed value is positive "one", what is the nature of correlation?
5. What is the range of values for the Karl Pearson correlation coefficient?
6. What does it indicates that the correlation coefficient is -1.

Answers

1. To quantify the strength and direction of a linear relationship between two variables.
2. Karl Person correlation coefficient measures the linear relationship between two continuous variables.
3. There is a strong negative linear relationship between the variables.
4. It is perfect positive.

5. The values for the Karl Pearson correlation coefficient are from -1 to +1.

6. It is perfect negative correlation.

Assignments

1. Find Karl Pearson's coefficient of correlation, from the following series of marks secured by ten students in a class test in mathematics and statistics.

Marks in Mathematics	45	70	65	30	90	40	50	75	85	60
Marks in Statistics	35	90	70	40	95	40	60	80	80	50

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

Age of Husband (years)	23	27	28	29	30	31	33	35	36	39
Age of Wife (years)	18	22	23	24	25	26	28	20	30	32

3. Calculate the Kerl Pearson coefficient of correlation from the following data

X	77	60	30	53	14	35	90	25	56	60
Y	35	38	60	40	50	40	35	56	34	42

4. Following are the marks obtained by 10 students of two subjects Mathematics and Physics in a class test. Estimate Spearman's Rank Correlation.

Name of students	A	B	C	D	E	F	G	H
Mathematics	7	3	1	4	6	8	2	5
Physics	6	2	1	5	8	7	3	4

5. Compute the correlation coefficient between the Price and Demand

Price	50	75	60	70	95	90	88
Demand	100	140	110	115	150	134	120

6. Calculate the coefficient of correlation from the following data by the Spearman's Rank Differences method.

Price of Tea (Rs)	20	25	60	45	80	25	55	65	25	75
Price of Coffee (Rs)	52	50	55	50	60	70	72	78	80	63

7. Ten students obtained the following marks in Statistics and Accountancy. Compute Spearman's Rank Correlation Coefficient?

Statistics	115	109	112	87	98	120	98	100	98	118
Accountancy	75	73	85	70	76	82	65	73	68	80

8. Find Pearson's Co-efficient of correlation from the following data

Sales	15	18	22	28	32	46	52
Profit	52	66	78	87	96	125	141

References

1. Gupta S. and V.K. Kapoor, (2014) *Fundamentals of Applied Statistics*, S.Chand and Sons, New Delhi.

2. Monga, G.S. (2001) *Mathematics and Statistics for Economics*, Vikas Publishing, New Delhi

Suggested Readings

1. Gupta, S.P, (2014) *Statistical Methods*, Sulthan Chand and Sons, New Delhi.

2. SC Gupta and VK Kapoor, (2020) *Fundamentals of Mathematical Stastistics,* Sulthanchand and Sons, NewDelhi

3. Jan Ubøe, (2018) *Introductory Statistics for Business and Economics: Theory, Exercises and Solutions*, Springer International Publishing.

4. T. Rajaretnam, (2016) *Statistics for Social Sciences*, Sage India.

3

BLOCK

Regression

UNIT 1

Regression

Learning Outcomes

After going through this unit, the learner will be able to

- have the basic knowledge of regression analysis.
- understand and develop the statistical applications and analytical skills of data using correlate regression analysis.
- acquire the knowledge on the functioning of regression analysis

Prerequisites

Having studied measures of central tendency, measures of dispersion and correlation, we are now prepared to explore the realm of regression analysis. Imagine a scenario where a marketing analyst is meticulously studying the correlation between advertising expenditures and product sales. The correlation coefficient has indicated a positive relationship between these variables.

It is crucial to understand that correlation alone does not reveal all the details of how things are connected. Regression analysis assists in constructing a model that predicts the impact of changes in advertising on sales. This model, in turn, transforms into a valuable instrument for refining advertising strategies with the ultimate goal of achieving the highest possible impact on sales outcomes.

Correlation coefficient is a measure of degree of relationship between the variables X and Y. It is a tool of ascertain the degree of relationship between two variables. Therefore, we can not say that one variable is the cause and the other is the effect. For example, the high degree of correlation between prize and demand for a certain commodity may not suggest which is the cause and which is the effect. In regression analysis cause and effect relation is very clearly indicated.

The variable whose value is influenced is called the dependent variable, denoted by Y, and the variable which exerts the influence is the independent variable denoted by X. If we are interested in finding the relationship between the level of income and consumption, the level of income will influence the level of consumption and such income is the independent variable and the consumption is the dependent variable. The term regression was first used by Sir Francis Galton in 1877.

Keywords

Regression, Regression Lines, Regression Equation

3.1.1 Regression

Regression analysis is a mathematical measure of the average relationship between two or more variables in terms of the original units of the data.

In regression analysis there are two types of variables. The variable whose value is influenced or is to be predicted is called dependent variable and the variable which influences the values or is used for prediction is called independent variable. The independent variable is also known as regressor or predictor or explanatory variable while the dependent variable is known as regressed or explained variable.

If the given bivariate data are plotted on a graph, the points so obtained on the scatter diagram will more or less concentrate round a curve, called the curve of regression. The mathematical equation of the regression curve, usually called the regression equation, enables us to study the average change in the value of the dependent variable for any given value of the independent variable.

If the regression curve is a straight line, we say that there is linear regression between the variables under study. The equation of such a curve is the equation of a straight line, i.e., a first degree equation in the variables x and y. In case of linear regression, the values of the dependent variable increase by a constant absolute amount for a unit change in the value of the independent variable. However, if the curve of regression is not a straight line, the regression is termed as curved or non-linear regression. The non-linear regression equation will be a functional relation between x and y involving terms in x and y of degree higher than one, i.e., involving terms of the type x^2, y^2, xy, etc.

Lines of Regression

Line of regression is the line which gives the best estimate of one variable for any given value of the other variable. In case of two variables *x* and *y*, we shall have two lines of regression; regression of *y* on *x* and the other regression of *x* on y. Line of regression of y on x is the line which gives the best estimate for the value of y for any specified value of x. Similarly, regression line of x on y is the line which gives the best estimate for the value of x for any specified value of y.

Line of best fit

The term best fit is interpreted in accordance with the Principle of Least Squares which consists in minimising the sum of the squares of the errors of estimates. Error is the deviations between the given observed values of the variable and their corresponding estimated values as given by the line of best fit. We may minimise the sum of the squares of the errors parallel to y-axis or parallel to x-axis, gives the equation of the line of regression of y on x and the line of regression of x on y.

3.1.2 Regression Equations

Regression equations are algebraic expression of the regression lines. Since there are two regression lines, there are two regression equations.

The regression equation of x on y, $x = a + b\,y$, is used to describe the variation in the values of x for given changes in y and the regression equation of y on x, $y = a + b\,x$ is used to describe the variation in the values of y for given changes in

x.

Regression Equation y on x

The regression equation y on x is expressed as follows;

$$y = a + b\,x$$

In this equation, "a" and "b" are unknown constants which determines the position. These constants are called the parameters of the line. The values of "a" and "b" can be obtained by solving the following equations simultaneously.

$$\sum y = na + b\sum x$$

$$\sum xy = a\sum x + b\sum x^2$$

These equations are commonly referred to as the normal equations. By solving these normal equations and substituting the determined values of 'a' and 'b' into the equation, we obtain the regression equation for predicting 'y' based on 'x'.

Regression Equation of x on y

The regression equation x on y is expressed as follows;

$$x = a + b\,y$$

To determine the values of "a" and "b", the following two normal equations are to be solved simultaneously

$$\sum x = na + b\sum y$$

$$\sum xy = a\sum y + b\sum y^2$$

Properties of regression lines

- The regression line is constructed to minimize the sum of squared residuals (the differences between observed and predicted values), ensuring the line provides the best fit to the data.
- The regression lines pass through the mean $(\bar{x}, \bar{y})$ of X and Y variables.
- The two regression lines are perpendicular when $r = 0$.
- The regression line of y on x has the same slope as the regression line of x on y, ensuring symmetry in the relationships between variables.
- The regression line can be used to predict values of the dependent variable based on the known values of the independent variable(s).
- The regression line assumes a linear relationship between the variables. If the true relationship is nonlinear, the regression line might not accurately represent the relationship.

Illustration : 3.1.1

From the following data, obtain the two regression equations by the method of least square, and estimate the value of y when $x = 12$.

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

Solution

x	y	x^2	y^2	$x \times y$
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
Total=30	**40**	**220**	**340**	**214**

Regression equation y on x is given by $y = a + b\,x$

two normal equations are

$\sum y = na + b\sum x$

$\sum xy = a\sum x + b\sum x^2$

Substituting the values in the equation, we get;

$40 = 5a + 30b$ .....(1)

$214 = 30a + 220b$.....(2)

Solving the equations,

Eqn (1) × 6 $\qquad 240 = 30a + 180b \ldots$

Eqn (2) $\qquad 214 = 30a + 220b$

....

Subtracting $\qquad 26 = -40b$

$$b = -\frac{26}{40} = -0.65$$

Substituting the value of b in eqn. (1) we get

$$40 = 5a + 30 \times -0.65$$

$$40 = 5a - 19.5$$

$$5a = 40 + 19.5 = 59.5$$

$$a = \frac{59.5}{5} = 11.9$$

$$\therefore \quad a = 11.9, \quad b = -0.65$$

Regression equation y on x is given by $y = 11.9 - 0.65\,x$

Regression equation x on y is given by $x = a + b\,y$

two normal equations are

$\sum x = na + b\sum y$

$\sum xy = a\sum y + b\sum y^2$

Substituting the values in the equation, we get;

$30 = 5a + 40b \ldots \ldots . .(1)$

$214 = 40a + 340b \ldots \ldots \ldots (2)$

Solving the equations,

Eqn (1) × 8 $\quad 240 = 40a + 320b$

$214 = 40a + 340b$

....

Subtracting $\quad 26 = -20b$

$b = -\frac{26}{20} = -1.3$

Substituting the value of b in eqn. (1) we get

$30 = 5a + 40 \times -1.3$

$30 = 5a - 52$

$5a = 30 + 52$

$5a = \frac{82}{5} = 16.4$

$\therefore \; a = 16.4, \quad b = -1.3$

Regression equation x on y is given by $x = 16.4 - 1.3y$

In order to estimate the value of y, we shall use the line of regression of y on x ,

When $x = 12, \quad y = 11.9 - 0.65 \times 12$

$= 11.9 - 7.8$

$= 4.1$

Illustration : 3.1.2

Find two regression equations from the following data

X	20	22	25	26	27	24
Y	31	29	32	37	35	34

Solution

x	y	x^2	y^2	$x \times y$
20	31	400	961	620
22	29	484	841	638
25	32	625	1024	800
26	37	676	1369	962
27	35	729	1225	945
24	34	576	1156	816
144	198	3490	6576	4781

Regression equation y on x is given by $y = a + b\,x$

two normal equations are

$$\sum y = na + b\sum x$$

$$\sum xy = a\sum x + b\sum x^2$$

Substituting the values in the equation, we get;

$198 = 6a + 144b$ .....(1)

$4781 = 144a + 3490b$.....(2)

Solving the equations,

Eqn (1)× 24 $\qquad 4752 = 144a + 3456b$

Eqn (2) $\qquad 4781 = 144a + 3490b$

....

Subtracting $\qquad -29 = -34b$

$$b = \frac{29}{34} = 0.853$$

Substituting the value of b in eqn. (1) we get

$198 = 6a + 144 \times 0.853$

$198 = 6a + 122.832$

$6a = 198 - 122.832 = 75.168$

$a = \frac{75.168}{6} = 12.53$

$\therefore \quad a = 12.53, \quad b = 0.853$

Regression equation y on x is given by $y = 12.53 + 0.853\ x$

Regression equation x on y is given by $x = a\ +\ b\ y$

two normal equations are

$\sum x\ =\ na\ +\ b\sum y$

$\sum xy =\ a\sum y + b\sum y^2$

Substituting the values in the equation, we get;

$144 =\ 6a + 198b$

$4781 = 198a + 6576b$

Solving the equations,

Multiplying first equation by 33,

$4752 =\ 198a + 6534b$

$4781 = 198a + 6576b$

....

Subtracting $\quad -29\ =\ -42b$

$b = \frac{29}{42} = 0.69$

Substituting the value of b in the first eqn. we get,

$144 =\ 6a + 198 \times 0.69$

$144 =\ 6a + 136.62$

$6a = 144 - 136.62 = 7.38$

$a = \frac{7.38}{6} = 1.23$

$\therefore \quad a = 1.23, \quad b = 0.69$

Regression equation y on x is given by $y = 1.23 + 0.69\ x$

Illustration : 3.1.3

From the following data, obtain the two regression equations by the method of least square. Also estimate the value of x when $y = 12$.

X	10	6	10	6	8
Y	6	2	10	4	8

Solution

x	y	$x \times y$	x^2	y^2
10	6	60	100	36
6	2	12	36	4
10	10	100	100	100
6	4	24	36	16
8	8	64	64	64
Total-40	30	260	336	220

Regression equation y on x is given by $y = a + b\,x$

To determine the value of constants "a" and "b", the following two normal equations are to be solved;

$$\sum y = na + b\sum x$$

$$\sum xy = a\sum x + b\sum x^2$$

Substituting the values in the equation, we get;

$$30 = 5a + 40b \ldots\ldots\ldots..(1)$$

$$260 = 40a + 336b \ldots\ldots\ldots..(2)$$

Multiplying equation 1 by 8 we get;

$$240 = 40a + 320b$$

$$260 = 40a + 336b$$

…………………………..

Subtracting

$$-20 = -16\,b$$

$$b = \frac{20}{16} = 1.25$$

This value of b can be substituted in equation (1), we get the value of a. That is;

$30 = 5a + 40 \times 1.25$

$30 = 5a + 50$

$5a = 20$

$a = {}^{-}4$

Substituting the values of "a" and "b" in the regression equation, we get the regression line of y on x, $y = -4 + 1.25x$

Now we can calculate the regression equation of x on y, that is given by the equation;

$x = a + b\,y$, and the two normal equations are

$\sum x = n\,a + b\sum y$

$\sum xy = a\sum y + b\sum y^2$

Substituting the values in the equation, we get;

$40 = 5a + 30b \ldots\ldots\ldots\ldots(1)$

$260 = 30a + 220b \ldots\ldots\ldots\ldots\ldots(2)$

Multiplying (1) by 6, we get

$240 = 30a + 180b$

$260 = 30a + 220b$

Subtracting

$-20 = -40b$

$b = \frac{20}{40} = 0.5$

Substituting the value of "b" in equation (1), we get;

$40 = 5 \times a + 30 \times 0.5$

$= 5 \times a + 15$

$5a = 40 - 15$

$5a = 25$

$a = \frac{25}{5} = 5$

The regression line of *x* on *y* is $x = 5 + 0.5y$

When $y = 12,\ x = 5 + 0.5 \times 12$

$= 5 + 6$

$= 11$

Illustration : 3.1.4

Find two regression equations from the following data

x	3	5	6	8	9	11
y	2	3	4	6	5	8

Solution

x	y	$x \times y$	x^2	y^2
3	2	6	9	4
‘;5	3	15	25	9
6	4	24	36	16
8	6	48	64	36
9	5	45	81	25
11	8	88	121	64
Total-42	**28**	**226**	**336**	**154**

Regression equation *y* on *x* is given by $y = a + b\,x$

two normal equations are

$\sum y = na + b \sum x$

$\sum xy = a\sum x + b\sum x^2$

Substituting the values in the equation, we get;

$28 = 6a + 42b$ .....(1)

$226 = 42a + 336b$.....(2)

Solving the equations,

Eqn (1)× 7 $196 = 42a + 294b \ldots$

Eqn (2) $226 = 42a + 336b$

$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$

Subtracting $-30 = -42b$

$$b = \frac{30}{42} = 0.714$$

Substituting the value of *b* in eqn. (1) we get

$28 = 6a + 42 \times 0.714$

$28 = 6a + 29.988$

$6a = 28 - 29.988 = -1.988$

$a = -\frac{1.988}{6} = -0.331$

$\therefore \; a = -0.331, \quad b = 0.714$

Regression equation y on x is given by $y = -0.331 + 0.714\,x$

Regression equation x on y is given by $x = a + b\,y$

two normal equations are

$\sum x = na + b\sum y$

$\sum xy = a\sum y + b\sum y^2$

Substituting the values in the equation, we get;

$42 = 6a + 28b \ldots\ldots\ldots.(1)$

$226 = 28a + 154b \ldots\ldots\ldots.(2)$

Solving the equations,

Multiplying first equation by 28, and second equation by 6 we get,

$$1176 = 168a + 784b$$

$$1356 = 168a + 924b$$

$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$

Subtracting $-180 = -140b$

$$b = \frac{180}{140} = 1.286$$

Substituting the value of *b* in the first eqn. we get

$$42 = 6a + 28 \times 1.286$$

$$42 = 6a + 36.008$$

$$6a = 42 - 36.008 = 5.992$$

$$a = \frac{5.992}{6} = 0.999$$

$$\therefore \ a = 0.999, \quad b = 1.286$$

Regression equation *y* on *x* is given by $y = 0.999 + 1.286\,x$

Illustration : 3.1.5

Find out the regression equation, $x \text{ on } y$ and $y \text{ on } x$ from the following data:

x	15	20	25	30	35	40	45
y	8	14	20	26	32	38	44

Solution

x	y	$x \times y$	x^2	y^2
15	8	120	225	64
20	14	280	400	196
25	20	500	625	400
30	26	780	900	676
35	32	1120	1225	1024
40	38	1520	1600	1444
45	44	1960	2025	1936
Total-210	**182**	**6300**	**7000**	**5740**

Regression equation y on x is given by $y = a + b\,x$

Two normal equations are

$\sum y \ = \ na \ + \ b\sum x$

$\sum xy = \ a\sum x + b\sum x^2$

Substituting the values in the equation, we get;

$182 = \ 7a + 210\ b$ …..(1)

$6300 = 210a + 7000b$…..(2)

Solving the equations,

Eqn (1)× 30	$5460 = \ 210a + 6300b$ …
Eqn (2)	$6300 = 210a + 7000b$
	…………………………………………
Subtracting	$-\ 840 \ = \ -700b$
	$b = \frac{840}{700} = 1.2$

Substituting the value of *b* in eqn. (1) we get

$182 = \ 7a + 210 \times 1.2$

$182 = \ 7a + 252$

$7a = 182 - 252 = -70$

$a = -\frac{70}{7} = -10$

$\therefore \ \ a = -10, \ \ \ b = 1.2$

Regression equation y on x is given by $y = -10 + 1.2\ x$

Regression equation x on y is given by $x = a \ + \ b\ y$

two normal equations are

$\sum x \ = \ na \ + \ b\sum y$

$\sum xy = \ a\sum y + b\sum y^2$

Substituting the values in the equation, we get;

$210 = \ 7a + 182b$

$6300 = 182a + 5740b$

Solving the equations,

Multiplying first equation by 26

$$5460 = 182a + 4732b$$

$$6300 = 182a + 5740b$$

....

Subtracting $$-840 = -1008b$$

$$b = \frac{840}{1008} = 0.83$$

Substituting the value of b in the first eqn. we get

$210 = 7a + 182 \times 0.83$

$210 = 7a + 151.06$

$7a = 210 - 151.06 = 58.94$

$a = \frac{58.94}{7} = 8.42$

$\therefore \; a = 8.42, \quad b = 0.83$

Regression line of x on y is $x = 8.42 + 0.83y$

Illustration : 3.1.6

For a particular product, the sales (y) and the advertisement expenditure (x) for 10 years, provide the results

$\sum x = 15, \; \sum y = 110, \sum xy = 400, \sum x^2 = 250, \sum y^2 = 3200.$

Find the regression line of y on x and the estimated value of y for $x = 10$.

Solution

Regression equation y on x is given by $y = a + bx$

two normal equations are

$\sum y = na + b\sum x$

$\sum xy = a\sum x + b\sum x^2$

Substituting the values in the equation, we get;

$110 = 10a + 15b$ .....(1)

$400 = 15a + 250b$.....(2)

Solving the equations,

Eqn (1)÷ 5 $\quad 22 = 2a + 3b$ ....(3)

Eqn (2) ÷ 5 $\quad 80 = 3a + 50b$.........(4)

Eqn (1)× 3 $\quad 66 = 6a + 9b$ …

Eqn (2) × 2 $\quad 160 = 6a + 100b$

Subtracting $\quad -94 = -91b$

$$b = \frac{94}{91} = 1.03$$

Substituting the value of b in eqn. (3) we get

$22 = 2a + 3 \times 1.03$

$22 = 2a + 3.09$

$2a = 22 - 3.09$

$= 18.91$

$$a = \frac{18.91}{2}$$

$= 9.455$

$\therefore \quad a = 1.03, \quad b = 9.455$

Regression equation y on x is given by $y = 1.03 + 9.455\, x$

$x = 10, \qquad y = 1.03 + 9.455 \times 10$

$= 1.03 + 94.55 = 95.58$

Illustration 3.1.7

If the regression equation between the variables X and y are $3x + 2y - 26 = 0$ and $6x + y - 31 = 0$. Find $(\bar{x}, \ \bar{y})$.

Solution

The given lines are $3x + 2y - 26 = 0$ ……….(1)

$6x + y - 31 = 0$……….(2)

……………………………………

Solving the equations,

Eqn (1)×2 $6x + 4y = 52$

Eqn (2) $6x + y = 31$

………………………………

Subtracting $3y = 21$

$y = 7$

Substituting the value of y in eqn. (1) we get

$3x + 2 \times 7 - 26 = 0$

$3x = 12$

$x = 4$

Thus $x = 4, y = 7$

Since the regression lines passes through $(\bar{x},\ \bar{y})$, $\bar{x},\ = 4,\ \bar{y} = 7$

Recap

- Regression analysis is a mathematical measure of the average relationship between two or more variables in terms of the original units of the data.
- Linear regression is a regression model, that estimates the relationship between one independent variable and one dependent variable using a straight line.
- Non-linear Regression is the study more than one independent variable and one dependent variable.
- Simple Regression is the study of the relationship between two variables alone.
- Line of best fit is drawn to represent the relationship between two or more variables.

Objective Questions

1. Which mathematical measure is used to estimate the average relationship between two or more variables in terms of the original units of the data variables.
2. In simple linear regression, how many variables are involved?
3. Which method is used in regression analysis to estimate the model parameter?
4. In regression analysis, what is the primary purpose of the Method of Least Squares?

Answers

1. Regression
2. One independent variable and one dependent variable
3. Method of least square
4. Minimize the least square error

Self-Assessment Questions

1. Explain the concept of Regression with real life examples. Why should there be in general two lines of regression for each bivariate frequency distribution?

2. Explain the difference between correlation and regression analysis?

3. State the importance of line of best fit.

4. Explain the various types of Regression.

Assignments

1. Obtain the equations of the two lines of regression for the data given below :

X	1	2	3	4	5	6	7	8	9
Y	9	8	10	12	11	13	14	16	15

2. Find the regression line y on x for the data

x	1	4	2	3	5
y	3	1	2	5	4

3. Find the regression line of Y on X for the following data

$$\sum x = 15, \quad \sum y = 42, \quad \sum xy = 141, \quad \sum x^2 = 55.$$

Also find the value of y when $x = 6$

4. Given the following values of x and y

X	3	5	6	8	9	11
Y	2	3	4	6	5	8

Find the regression line y on x for the data and regression line x on y

5. Given the bivariate data

X	1	5	3	2	1	1	7	3
Y	6	1	0	0	1	2	1	5

a. Fit a regression line of y on x and hence predict y if x = 10.

b. Fit a regression line of x on y and hence predict x if y = 2.5.

6. For 10 observations on price (x) and supply (y) the following data was obtained

$$\sum x = 130, \quad \sum y = 220, \sum x^2 = 2288, \sum y^2 = 5506, \sum xy = 3467$$

Obtain the line of regression line of y on x and estimate the supply when the price is 16 units.

7. Find the regression line y on x from the following data using normal equations

X	10	20	30	40	50
Y	3	2	0	5	4

References

1. Gupta, S.P, *Statistical Methods*, Sulthan Chand and Sons, New Delhi.
2. SC Gupta and VK Kapoor, *Fundamentals of Mathematical Stastistics*

Suggested Readings

1. Jan Ubøe, *Introductory Statistics for Business and Economics*: Theory, Exercises and Solutions, Springer International Publishing.

2. T. Rajaretnam, *Statistics for Social Sciences*, Sage India.

3. Gupta S. and V.K. Kapoor, *Fundamentals of Applied Statistics*, S.Chand and Sons, New Delhi.

4. Monga, G.S. *Mathematics and Statistics for Economics*, Vikas Publishing, New Delhi

2

UNIT

Regression Coefficient

Learning Outcomes

After going through this unit, the learner will be able to:

- understand the basic knowledge of different measures of regression analysis.
- acquire the knowledge on the functioning of correlation and regression analysis.
- knowledge of regression analysis and its applicability in decision making.

Prerequisites

Regression analysis is used to explain variability in dependent variable by means of one or more of independent variables and to analyse relationships among variables to answer; the question of how much dependent variable changes with changes in each of the independent variables, and to forecast or predict the value of dependent variable based on the values of the independent variables

Regression equations are algebraic expression of the regression lines. Let us consider two variables: x and y. If y depends on x, then, the result comes in the form of simple regression. If we take the case of two variable x and y, we shall have two regression lines as the regression line of x on y and regression line of y on x. The regression line of y on x gives the most probable value of y for given value of x and the regression line of x on y given the most probable value of x for given value of y. Thus, we have two regression lines and two regression equations.

Regression equation, y on x is given by $y = a + b x$ and regression equation x on y is given by $x = a + b y$.

In the above two regression lines or regression equations, there are two regression parameters, which are a and b. Here a is unknown constant and b is also another unknown constant popularly called as regression coefficient.

Keywords

Regression Coefficients, Constants, Standard Deviation

3.2.1 Regression Coefficients

The constant *b* in the regression equation is called regression coefficient. Since there are two regression equations, there are two regression coefficients. The two regression coefficients are regression coefficient of *x* on *y* and regression coefficient of *y* on *x*.

Regression Coefficient of *x* on *y*

The regression coefficient of *x* on *y* is represented by the symbol, b_{xy} .This regression coefficient measures the change in *x*, corresponding to a unit change in *y*. The regression coefficient of *x* on *y* is given by;

$$b_{xy} = r\frac{\sigma_x}{\sigma_y}$$

where, r - Karl Pearson's Correlation Coefficient

σx - Standard deviation of *x* series

σy - Standard deviation of *y* series

Regression equation *x* on *y* is given by $x - \bar{x} = b_{xy}\ (y - \bar{y})$

Regression Coefficient of *y* on *x*

The regression coefficient of *y* on *x* is represented by byx .This regression coefficient measures the change in *y* variable corresponding to unit change in *x* variable. The value of byx is given by;

$$b_{yx} = r\frac{\sigma_y}{\sigma_x}$$

where r - Karl Pearson's correlation Coefficient

σy - Standard deviation of y series

σx - Standard deviation of x series

Regression equation y on x is given by $y - \bar{y} = b_{yx}\ (x - \bar{x})$

The regression coefficients b_{yx} and b_{xy} can be easily obtained by using the formula

$$b_{yx} = \frac{\Sigma(x - \bar{x})\ (y - \bar{y})}{\Sigma(x - \bar{x})^2}$$

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

Calculating Correlation Coefficients from Regression Coefficients

We know that $b_{xy} = r\frac{\sigma_x}{\sigma_y}$ and $b_{yx} = r\frac{\sigma_y}{\sigma_x}$

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

Cancelling the common items, we get $b_{xy} \times b_{yx} = r^2$

Thus, correlation coefficient can be calculated using the equation;

$$r = \sqrt{bxy \times byx}$$

Since the value of the correlation coefficient cannot exceed one, one of the regression coefficients must be less than one. In other words, both the regression coefficients cannot be greater than one. Similarly, both the regression coefficients will have the same sign, that is they will be either positive or negative.

3.2.2 Properties of regression coefficients

- Correlation coefficient is the geometric mean between the regression coefficients
- The coefficient of correlation will have the same sign as that of the regression coefficients, such as if the regression coefficients have a positive sign, then "r" will be positive and vice-versa.
- If one of the regression coefficients is greater than unity, the other must be less than unity.
- Arithmetic mean of the regression coefficients is greater than the correlation coefficient *r* provided $r > 0$.
- Regression coefficients are independent of the change of origin but not of scale. By origin, we mean that there will be no effect on the regression coefficients if any constant is subtracted from the value of X and Y. By scale, we mean that if the value of X and Y is either multiplied or divided by some constant, then the regression coefficients will also change.
- Both the regression coefficients will have the same sign. i.e. they will be either positive or negative. Thus, it is not possible that one regression coefficient is negative while the other is positive.
- The sign (positive or negative) of the regression coefficient indicates the direction of the relationship between variables.

- A positive regression coefficient suggests a positive correlation, while a negative coefficient implies a negative correlation.

- The magnitude of the regression coefficient reflects the strength of the relationship. A larger absolute value indicates a stronger influence of the independent variable on the dependent variable.

Illustration : 3.2.1

Find the regression coefficients of X on Y and Yon X from the following details. Also find the regression equations.

X	10	6	10	6	8
Y	6	2	10	4	8

Solution

x	y	$x-8$	$(x-8)^2$	$y-6$	$(y-6)^2$	$(x-8)(y-6)$
10	6	2	4	0	0	0
6	2	¯2	4	¯4	16	8
10	10	2	4	4	16	8
6	4	¯2	4	¯2	4	4
8	8	0	0	2	4	0
Total 40	**30**	**0**	**16**	**0**	**40**	**20**

$$n=5,\ \bar{x} = \frac{\Sigma x}{n} = \frac{40}{5} = 8\ ,\ \bar{y} = \frac{\Sigma y}{n} = \frac{30}{5} = 6$$

Regression equation y on x is given by the equation

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

We have $b_{yx} = \dfrac{\Sigma(x-\bar{x})\ (y-\bar{y})}{\Sigma(x-\bar{x})^2}$

$$= \frac{\Sigma(x-8)\ (y-6)}{\Sigma(x-\bar{x})^2}$$

$$= \frac{20}{16} = 1.25$$

Substituting in the equation

$(y-6\ ) = 1.25\ (x-8)$

$y\ =\ 1.25\,x - 1.25\ \times 8 + 6$

$y = 1.25x - 10 + 6$

$y = 1.25x - 4$

Similarly, regression equation *y* on *x* is given by the formula

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

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

$$= \frac{\Sigma(x-8)\ (y-6)}{\Sigma(y-8)^2}$$

$$= \frac{20}{40} = 0.5$$

Substituting in the equation

$(x-8\ ) = 0.5\ (y-6)$

$x\ =\ 0.5y - 0.5 \times 6 + 8$

$x\ =\ 0.5y - 3 + 8$

$x\ =\ 0.5y + 5$

Illustration : 3.2.2

From the data given below find

(a) The two regression equations

(b) The two regression coefficients..

(c) The coefficient of correlation between the marks in Economics and Statistics.

(d) The most likely marks in Statistics when marks in Economics are 30.

Marks in Economics	25	28	35	32	31	36	29	38	34	32
Marks in Statistics	43	46	49	41	36	32	31	30	33	39

Solution

x	y	$x-32$	$(x-32)^2$	$y-38$	$(y-38)^2$	$(x-32)(y-38)$
25	43	-7	49	5	25	-35
28	46	-4	16	8	64	-32
35	49	3	9	11	121	33
32	41	0	0	3	9	0
31	36	-1	1	-2	4	2
36	32	4	16	-6	36	-24
29	31	-3	9	-7	49	21
38	30	6	36	-8	64	-48
34	33	2	4	-5	25	-10
32	39	0	0	1	1	0
320	380		140		398	-93

a) n =10, $\bar{x} = \frac{\sum x}{n} = \frac{320}{10} = 32$, $\bar{y} = \frac{\sum y}{n} = \frac{380}{10} = 38$

Regression equation y on x is given by the equation

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

where $b_{yx} = \frac{\sum(x-\bar{x})\ (y-\bar{y})}{\sum(x-\bar{x})^2}$

$$= \frac{\sum(x-32)\ (y-38)}{\sum(x-\bar{x})^2}$$

$$= \frac{-93}{140}$$

Regression equation y on x is

$$(y - 38\ ) = \ -0.6643\ (x - 32)$$

$$y\ = -0.6643\ x + 0.6643\ \times 32 + 38$$

$$= -0.6643x + 21.2576 + 38$$

$$= -0.6643x + 59.2576$$

Similarly, regression equation x on y is given by the formula

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

where $b_{xy} = \dfrac{\Sigma(x - \bar{x})\ (y - \bar{y})}{\Sigma(y - \bar{y})^2}$

$$= \frac{\Sigma(x - 32)\ (y - 38)}{\Sigma(y - \bar{y})^2}$$

$$= -\frac{93}{398} = -0.2337$$

$$(x - 32)\ = -0.2337\ (y - 38)$$

$$x\ = -0.2337y + 0.2337 \times 38 + 32$$

$$x\ = \ -0.2337y + 8.8806 + 32$$

$$x\ = -0.2337y + 40.8806$$

b) Regression Coefficients are $b_{yx} = -0.6643, b_{xy} = -0.2337$

c) Correlation coefficient $r = \sqrt{\ b_{yx}.b_{xy}} = \sqrt{-0.6643 \times -0.2337}$

$$= \sqrt{0.1552} = 0.39$$

d) In order to estimate the most likely marks in Statistics (y) when marks in Economics (x) are 30, we shall use the line of regression of y on x ,

$$y = -0.6643x + 59.2576$$

When $x = 30$ $\qquad y = -0.6643 \times 30 + 59.2576$

$$= -19.929 + 59.2576$$

$$= 39.3286$$

Hence, the most likely marks in Statistics when marks in Economics are 30, is 39·3286 ~ 39.

Illustration : 3.2.3

A departmental store gives in-service training to its salesmen which is followed by a test. It is considering whether it should terminate the service of any salesman who does not do well in the test. The following data give the test scores and sales made by nine salesmen during a certain period :

Test scores	14	19	24	21	26	22	15	20	19
Sales ('000 Rs.)	31	36	48	37	50	45	33	41	39

Calculate the coefficient of correlation between the test scores and the sales. Does it indicate that the termination of services of low-test scores is justified? If the firm wants a minimum sales volume of Rs. 30,000, what is the minimum test score that will ensure continuation of service ? Also estimate the most probable sales volume of a salesman making a score of 28.

Solution

x	y	$x-20$	$(x-20)^2$	$y-40$	$(y-40)^2$	$(x-20)(y-40)$
14	31	-6	36	-9	81	54
19	36	-1	1	-4	16	4
24	48	4	16	8	64	32
21	37	1	1	-3	9	-3
26	50	6	36	10	100	60
22	45	2	4	5	25	10
15	33	-5	25	-7	49	35
20	41	0	0	1	1	0
19	39	-1	1	-1	1	1
180	360		120		346	193

a) n = 9, $\bar{x} = \frac{\sum x}{n} = \frac{180}{9} = 20$

$$\bar{y} = \frac{\sum y}{n} = \frac{360}{9} = 40$$

Regression equation y on x is given by the equation

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

$$b_{yx} = \frac{\sum(x - \bar{x})\ (y - \bar{y})}{\sum(x - \bar{x})^2}$$

$$= \frac{\sum(x - 20)\ (y - 40)}{\sum(x - \bar{x})^2}$$

$$= \frac{193}{120} = 1.6083$$

$$(y - 40\ ) = \ 1.6083\ (x - 20)$$

$$y\ =\ 1.6083\ x - \ 1.6083\ \times 20 + 40$$

$$y = \ 1.6083\ x - 32.166 + 40$$

$$y = 1.6083\ x + 7.834$$

Similarly, regression equation *y* on *x* is given by the formula

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

$$b_{xy} = \frac{\sum(x - \bar{x})\ (y - \bar{y})}{\sum(y - \bar{y})^2}$$

$$= \frac{\sum(x - 20)\ (y - 40)}{\sum(y - \bar{y})^2}$$

$$= \frac{193}{346} = 0.5578$$

$$(x - 20)\ = 0.5578\ (y - 40)$$

$$x\ = 0.5578\ y - 0.5578 \times\ 40 + 20$$

$$=\ 0.5578\ y - 22.312 + 20$$

$$x = -0.5578\ y - 2.312$$

b) Correlation coefficient $r = \sqrt{\ b_{yx}.b_{xy}} = \sqrt{1.6083 \times 0.5578}$

$$= \sqrt{0.8971} = 0.947$$

Since, the regression coefficients are positive, r is also positive.

$$\therefore r = +0\cdot 9471$$

Thus, we see that there is a very high degree of positive correlation between the test scores (x) and the sales (y). This justifies the proposal for the termination of service of those with low test scores.

Hence to ensure the continuation of service, the minimum test score (x) corresponding to a minimum sales volume (y) of Rs. 30,000 = 30 ('000 Rs.) is obtained on putting y = 30 in the regression equation

$$x = -0.5578\,y - 2.312$$

$$When\ y = 30, \quad x = 0\cdot 5578 \times 30 - 2\cdot 312$$

$$= 16\cdot 734 - 2\cdot 312$$

$$= 14\cdot 422 -\sim 14$$

Probable sales volume of a sales man making a score of 28

for x = 28

y = 1.6083(28) + 7.83

y = 52.90

The sales volume will be approximately equal to Rs. 53000

Illustration : 3.2.4

If the regression equations between the variables x and y are $4x - 5y + 33 = 0$ and $20x - 9y + 107 = 0$, find the correlation coefficient and means of the variables.

Solution

The regression equations $\quad 4x - 5y = -33$

$$20x - 9y = -107$$

Solving the equations,

Eqn (1)×5 $\quad 20x - 25y = -165 \cdots$

Eqn (2) $\quad 20x - 9y = -107$

....

Subtracting $\quad -16y = -272$

$$y = \frac{272}{16} = 17$$

Substituting the value of y in eqn. (1) we get

$$4x - 5 \times 17 = -33$$

$$4x = -33 + 85$$

$$4x = 52$$

$$x = \frac{52}{4} = 13$$

$$x = 13, \quad y = 17$$

Since the regression lines pass through $\overline{(x, \bar{y})}$ we have $\bar{x} = 13, \; \bar{y} = 17$

Rewriting the regression lines of y on x $4x - 5y + 33 = 0$ as

$$y = \frac{4}{5}x + \frac{33}{5}, \text{ we get } b_{yx} = \frac{4}{5}$$

Similarly, Rewriting the regression lines of x on y, $20x - 9y = 107$ as

$$y = \frac{9}{20}y + \frac{107}{9}, \text{ we get } b_{xy} = \frac{9}{20}$$

Thus $r = \sqrt{b_{yx} . b_{xy}} = \sqrt{\frac{4}{5} . \frac{9}{20}} = \pm 0.6$

Since b_{yx} and b_{xy} are positive, $r = 0.6$

Illustration : 3.2.5

The following data pertains to the marks in subjects A and B in a certain examination. Mean marks in $A = 39.5$, Mean marks in $B = 47.5$ standard deviation of marks in $A = 10.8$ and Standard deviation of marks in $B = 16.8$. coefficient of correlation between marks in A and marks in B is 0.42. Give the estimate of marks in B for candidate who secured 52 marks in A.

Solution

Given $\bar{x} = 39.5, \; \bar{y} = 47.5, \; \sigma x = 10.8, \; \sigma y = 16.8$ and $r = 0.42$.

Line of regression of y on x

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

$$b_{yx} = r\frac{\sigma y}{\sigma x} = \frac{0.42 \times 16.8}{10.8} = 0.65$$

Substituting these in the equation,

$$(y - 47.5\ ) = \ 0.65\ (x - 39.5)$$

$$y = \ 0.65\,x - 0.65 \times 39.5 + 47.5$$

$$y = \ 0.65\,x + 21.825$$

When $x = 52$, $y = 0.65 \times 52 + 21.825 = 55.625$

Illustration : 3.2.6

The data about the sales and advertisement expenditure of a firm is given below :

	Sales (Rs) *x*	**Advertisement Expenditure *y***
Mean	40	6
S.D.	10	1.5

Coefficient of correlation $= \ r \ = \ 0 \cdot 9$

Estimate the likely sales for a proposed advertisement expenditure of Rs. 10 crores.

(ii) What should be the advertisement expenditure if the firm proposes a sales target of 60 crores of rupees ?

Solution

Given $\bar{x} = 40$, $\bar{y} = 6$, $\sigma x = 10$, $\sigma y = 1.5$ and $r = 0.9$.

To estimate the advertisement expenditure (y) for proposed sales (x), we need the equation of line of regression of y on x

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

Where $b_{yx} = r\frac{\sigma y}{\sigma x} = \frac{0.9 \times 1.5}{10} = 0.135$

Substituting these in the equation,

$$(y - 6\ ) = \ 0.135\ (x - 40)$$

$$y = \ 0.135\,x - 0.135 \times 40 + 6$$

$$y = \ 0.135\,x - 5.4 + 6$$

$$y = \ 0.135\,x + 0.6$$

When $x = 60$, $\quad y = 0.135 \times 60 + 0.6 = 8.7$

To estimate the likely sales (x) for given advertisement expenditure (y), we need the regression equation of x on y which is given by :

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

Where $b_{xy} = r\frac{\sigma x}{\sigma y} = \frac{0.9 \times 10}{1.5} = 6$

Substituting these in the equation,

$$(x - 40\ ) = \ 6(y - 6)$$

$$x = \ 6\,y - 6 \times 6 + 40$$

$$x = \ 6y - 36 + 40$$

$$x = \ 6y + 4$$

Hence the estimated sales (x) for a proposed advertisement expenditure (y) of Rs. 10 crores are obtained on putting y = 10 in the equation $x = \ 6y + 4$ we get,

$x = 6 \times 10 + 4 = 64$ crores of Rs.

Required advertisement expenditure for a sales target of Rs. 60 crore

X $= \frac{Y - a}{b} = \frac{60 - 4}{6} = \frac{56}{6} = 9.33$ crores

Illustration : 3.2.7

For 50 students of a class the equation of marks in statistics (x) on marks in Accountancy (y) is $3y - 5x + 180 = 0$. The mean mark in Accountancy is 44 and variance of marks in statistics is $\frac{9}{16}^{th}$ of the variance of marks in Accountancy. Find mean marks in statistics and coefficient of correlation between marks in two subjects.

Solution

$(\bar{x}, \bar{y})$ be the mean mark of statistics and Accountancy.

Given $n = 50,\ 3y - 5x + 180 = 0,\ \bar{y} = 44$

Since the regression lines passes through $(\bar{x}, \bar{y})$, $3\bar{y} - 5\bar{x} + 180 = 0$.

Substituting $\bar{y} = 44$, we get, $3 \times 44 - 5\bar{x} + 180 = 0$

$$5\bar{x} = \ 3 \times 44 + 180 = 312$$

$$\bar{x} = \frac{312}{5} = 62.4$$

Mean marks in statistics = 62.4

Given $\sigma x^2 = \frac{9}{16}\sigma y^2$ $\frac{\sigma x^2}{\sigma y^2} = \frac{9}{16}$, $\frac{\sigma x}{\sigma y} = \frac{3}{4}$

From the equation $3y - 5x + 180 = 0$

$$5x = 3y + 180$$

$$x = 0.6\, y + 36$$

$$\Rightarrow b_{xy} = 0.6$$

$$r\frac{\sigma x}{\sigma y} = 0.6$$

$$r \times \frac{3}{4} = 0.6$$

$$r = 0.6 \times \frac{4}{3} = 0.8$$

∴ Coefficient of correlation = 0.8

Illustration : 3.2.8

If the regression equation between the variables x and y are $x + 2y - 5 = 0$ and $2x + 3y - 8 = 0$. Find $(\bar{x}, \ \bar{y})$, correlation coefficient r.

Solution

The given lines are $x + 2y - 5 = 0$ ....(1)

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

....

Solving the equations,

Eqn (1)×2 $2x + 4y - 10 = 0$

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

....

Subtracting $y = 2$

Substituting the value of y in eqn. (1) we get

$$x + 2 \times 2 - 5 = 0$$

$x = 1$

Thus $x = 1, y = 2$

Since the regression lines passes through ($\bar{x}$, $\bar{y}$), $\bar{x}, = 1, \bar{y} = 2$

Assume that $x + 2y - 5 = 0$ is the regression line of y on x and $2x + 3y - 8 = 0$ is the regression line of x on y.

Rewriting the equations $y = -\frac{1}{2}x + \frac{5}{2}$ and x $= -\frac{3}{2}y + 4$, we get

$b_{yx} = -\frac{1}{2}$ and $b_{xy} = -\frac{3}{2}$

$$r = \sqrt{bxy \times byx}$$

$$= \sqrt{-\frac{3}{2} \times -\frac{1}{2}}$$

$$= \sqrt{0.75}$$

$$= \pm 0.866$$

Since b_{yx} and b_{xy} are negative, $r = -0.866$

Recap

- Regression equations - are algebraic expression of the regression lines.
- Regression Coefficients - There are two regression coefficients. Regression coefficient of x on y and regression coefficient of y on x.
- It is the geometric mean between regression Coefficients.
- Coefficient of correlation and regression coefficient have the same sign.
- Sign of regression coefficient indicates direction of relationship between the variables.
- Magnitude of regression coefficient reflects the strength of relation.

Objective Questions

1. What does the constant, usually represented as 'b' in regression equation refers to?
2. Symbolically represent regression coefficient of x on y.
3. If the two coefficients of regression are negative then what is the sign of the correlation coefficient?
4. Which is the product of the regression coefficients?
5. What does the sign of regression coefficient represents?

Answers

1. Regression coefficient
2. $b_{xy} = r\dfrac{\sigma x}{\sigma y}$
3. Positive
4. Coefficient of determination
5. Direction of relationship between varibles

Self-Assessment Questions

1. Explain the term method of least square.
2. What are the properties of regression coefficient?
3. State any five differences between correlation and regression.
4. Distinguish between correlation and regression as concepts used in statistical analysis.

Assignments

1. The following data gives the age and blood pressure (BP) of 10 sports persons.

Name	A	B	C	D	E	F	G	H	I	J
Age (X)	42	36	55	58	35	65	60	50	48	51
BP (Y)	98	93	110	85	105	108	82	102	118	99

i. Find regression equation of Y on X and X on Y (Use the method of deviation from arithmetic mean)

ii. Find the correlation coefficient (r) using the regression coefficients.

iii. Estimate the blood pressure of a sports person whose age is 45.

2. The following data relate to the age of husbands and wives. Obtain the two regression equations and determine the most likely age of husband when the age of wife is 25 years.

X	25	28	30	32	35	36	38	39	42	55
Y	20	26	29	30	25	18	26	35	35	46

3. From the following data, obtain the two regression equations

Sales	91	97	108	121	67	124	51	73	111	57
Purchases	71	75	69	97	70	91	39	61	80	47

Also find correlation coefficient between sales and purchases?

4. Given the following information :

	X	Y
Mean	6	8
Standard Deviation	3	13

Coefficient of Determination = 0·64

Find :(i) $byx\ and\ bxy$ and (ii) Value of Y when X = 100

5. For a bivariate data the mean value of X is 20 and the mean value of Y is 45. The regression coefficient of Y on X is 4 and that of X on Y is 1/9. Find (i) The coefficient of correlation. (ii) The standard deviation of X if the standard deviation of Y is 12. (iii) Also write down the equations of regression lines.

6. In order to find the correlation coefficient between two variables X and Y from 12 pairs of observations, the following calculations were made :

$$\sum X = 30 ; \sum X^2 = 670 ; \sum Y = 5 ; \sum Y^2 = 285 ; \sum XY = 344$$

On subsequent verification it was discovered that the pair (X = 11, Y = 4) was copied wrongly, the correct values being (X = 10, Y = 14). After making necessary correction, find: (a) the two regression coefficients ; (b) the two regression equations ; (c) the correlation coefficient.

7. From the following data, find the most likely value of x when $y = 120$. Given $r = 0.8$

	X	Y
Mean	10	90
S.D	3	12

References

1. Gupta, S.P, *Statistical Methods*, Sulthan Chand and Sons, New Delhi.
2. SC Gupta and VK Kapoor, *Fundamentals of Mathematical Stastistics*

Suggested Readings

1. Jan Ubøe, *Introductory Statistics for Business and Economics*: Theory, Exercises and Solutions, Springer International Publishing.
2. T. Rajaretnam, *Statistics for Social Sciences*, Sage India.
3. Gupta S. and V.K. Kapoor, *Fundamentals of Applied Statistics*, S.Chand and Sons, New Delhi.
4. Monga, G.S. *Mathematics and Statistics for Economics*, Vikas Publishing, New Delhi

4

BLOCK

Index Numbers

1

UNIT

Basics of Index number

Learning Outcomes

After reading this unit, the learner will be able to:

- understand the significance of index numbers in economics
- compute the simple and weighted index number
- get an awareness on the limitations of index number.
- identify the problems in constructing index number.

Prerequisites

Index numbers are statistical tools used to measure and track changes in economic, financial, or social indicators over time. They provide a standardised way of comparing and summarising relative changes in different variables by assigning them a reference value or base period. Index numbers serve as a simplified representation of complex data, making it easier to analyse trends, compare values, and draw meaningful conclusions. They are widely used in economics, finance, and various fields to assess inflation rates, economic growth, market performance, and more. Index numbers play a crucial role in data analysis, aiding in decision-making, forecasting, and understanding trends in diverse industries and sectors.

Keywords

Index Numbers, Simple Index Numbers, Weighted Index Numbers

Discussion

4.1.1 Concept of an Index Number

An index number is a quantitative measure capturing the average change in related variables across different situations, such as varying times or locations. In economics, it is commonly used to assess fluctuations in parameters like commodity prices or production volumes. The Consumer Price Index (CPI) and Wholesale Price Index (WPI) reflect changes in living costs and a country's general price level, respectively. Constructing meaningful index numbers involves addressing challenges in handling actual price changes, often overcome by using standardised measures like price relatives. These index numbers, crucial for economic decision-making, impact policies on dearness allowances, wage adjustments, and other economic measures responding to shifts in commodity prices. Precision in constructing index numbers is vital for ensuring accurate policy adjustments align with the evolving economic landscape.

Index numbers serve a crucial purpose in economics, acting as a valuable tool for comparing the general magnitude of related variables in different situations. They provide a summarised figure that indicates the overall change in the magnitude of a group of variables, similar to how averages represent the general level of a specific variable. For example, when comparing the production of consumer-durable goods over different years, index numbers offer a comprehensive view, considering variables such as the output of refrigerators, radios, and carpets. They play a significant role in economic analysis, policy formulation, and decision-making, particularly in measuring changes in the general price level through price index numbers.

There are various types of index numbers, each designed to measure specific variables. Price index numbers, a well-known category, are widely used in economics to gauge changes in the general price level. The construction of index numbers involves selecting a base period, determining the variables to be included, and calculating relative changes over time or across different characteristics. These measures have historical significance dating back to 1774, evolving into a crucial statistical tool employed in economics, finance, and sociology. In economics, the Consumer Price Index (CPI) and Producer Price Index (PPI) are prominent examples of influencing decisions related to wages, inflation, and economic policies. The construction and analysis of these index numbers contribute to a better understanding of economic trends, aiding informed decision-making across various economic sectors. Index numbers are also called barometers of the economy as they serve as indicators of inflation or deflation.

4.1.2 Characteristics of Index Numbers

1. Index Numbers as Averages of Percentages

Index numbers are considered averages of percentages, expressing the relative change or movement in a set of related variables. For instance, the Consumer Price Index (CPI) is a widely used index representing the average change in prices of a basket of household goods and services. The CPI is calculated as a percentage change from a base year, providing an average that reflects the overall price movement in the economy.

2. Relative or Comparative Measure of Group of Items

Index numbers serve as a tool to measure the relative or comparative changes in a group of items. For example, the Dow Jones Industrial Average (DJIA) in the stock market is an index that represents the relative changes in the stock prices of a specific group of large, publicly traded companies. It offers investors and analysts insights into the stock market's overall performance.

3. Measurement of Changes Not Directly Measurable

Index numbers are instrumental in measuring changes that are not directly measurable. Consider the Human Development Index (HDI) in economics, which combines indicators like life expectancy, education, and income to assess the overall development of a country. The HDI provides a comprehensive measure of changes that may not be directly quantifiable in a single unit, facilitating international comparisons.

4. Measurement of Relative Changes Over Time or Place

Index numbers are employed to measure the relative changes over time or place, offering a standardised metric for comparison. The Producer Price Index (PPI) is an example used to measure the average change in selling prices received by domestic producers for their output over time. It helps businesses and policymakers understand inflationary trends in production costs.

In summary, index numbers play a crucial role in economics by providing averages of percentages, facilitating relative comparisons, measuring changes not directly measurable, and offering a standardised measure for assessing close changes over time or place. These characteristics make them valuable tools for analyzing economic trends and making informed decisions.

4.1.3 Uses of Index Numbers

1. Studying the Trend of a Series Over Time

Index numbers are widely used to study the trend of a series over a period of time. For instance, the Consumer Price Index (CPI) is employed to analyse the trend in the prices of a basket of goods and services consumed by households. Economists can discern inflationary or deflationary trends by comparing CPI values over several years, helping policymakers make informed decisions.

2. Economic Barometers

Index numbers act as economic barometers, reflecting the pulse of an economy. The Gross Domestic Product (GDP) growth rate serves as a crucial index indicating a country's economic health. A positive growth rate suggests economic expansion, while a negative rate indicates contraction. Governments, businesses, and investors closely monitor this index to gauge the overall economic performance.

3. Adjusting National Income for Constant Prices

Index numbers are vital in adjusting national income based on constant prices. The Consumer Price Index or the Wholesale Price Index removes the impact of inflation from nominal national income, allowing economists to find the change in real income over time. This adjustment is crucial for accurately assessing the standard of living and economic growth.

4. Comparing Levels of a Phenomenon Over Time

Index numbers are utilised to measure the level of a particular phenomenon compared to the level at a standard period. For example, the Dow Jones Industrial Average (DJIA) compares the current stock price level with a base period. This index helps investors assess the performance of the stock market over time.

5. Measuring Relative Changes

Index numbers are employed to measure relative temporal or cross-sectional changes in variables. The Producer Price Index (PPI) measures changes in selling prices received by producers over time. It provides insights into inflationary pressures in the production sector, aiding businesses in pricing strategies.

6. Assessing the Intrinsic Worth of Money

Index numbers are instrumental in determining the intrinsic worth of money versus its nominal worth. The Purchasing Power Parity (PPP) index compares the relative value of currencies based on the cost of identical baskets of goods and services in different countries. It helps assess the real value of currencies beyond nominal exchange rates.

7. Measurement for Management and Planning

Index numbers are crucial for measuring and managing government organisations or businesses. For instance, the Human Development Index (HDI) combines indicators like life expectancy, education, and income to assess a country's overall development. Governments use such indices for efficient planning and policy formulation to improve citizens' well-being.

In summary, index numbers serve various essential economic purposes, including trend analysis, economic barometer representation, income adjustment, level comparison, measurement of relative changes, assessment of money's intrinsic worth, and efficient management and planning. These applications highlight the versatility and significance of index numbers in economic analysis and decision-making.

4.1.4 Limitations of Index Numbers

1. Choosing a Suitable Base Period

Selecting an appropriate base period is a crucial challenge in constructing index numbers. For instance, if the base period chosen for the Consumer Price Index (CPI) is not representative of typical consumption patterns, it may lead to inaccurate assessments of inflation. The choice of a base period significantly influences the interpretation of the index.

2. Approximate Results and Not Exact

Index numbers provide approximate results rather than exact measurements. The Consumer Confidence Index (CCI), which gauges consumers' economic outlook, is subject to fluctuations based on perceptions and expectations. It reflects sentiments rather than precise economic conditions, highlighting the inherent approximation in index numbers.

3. Based on Only Few Items

Index numbers are often based on a limited number of items, not capturing the entirety of the economic landscape. The Dow Jones Industrial Average (DJIA), while widely followed, represents only 30 large-cap stocks, providing a narrow perspective on overall market performance.

4. Problems of Comparability and Reliability

Comparability and reliability issues may arise in index numbers. For example, comparing inflation rates between countries using different base years in their CPI calculations can lead to inaccurate cross-country comparisons. The reliability of indices depends on the consistency of methodologies.

5. Difficulty in Accounting for Product Quality Changes:

It is challenging to account for changes in product quality over time. The Consumer Electronics Price Index may struggle to reflect improvements in technology and features, potentially leading to overestimating price increases if not adjusted for quality improvements.

6. Limitations of Random Sampling

Index numbers may suffer from the limitations of random sampling used in item selection. The Unemployment Rate, often used as an economic indicator, relies on a sample survey. If the sample is not representative, the index may not accurately reflect the true unemployment situation.

7. Subject to Limitations of Averages

Being specialised averages, index numbers inherit the limitations associated with averages. For instance, the Average Hourly Earnings index may not fully represent the distribution of wage changes across different income groups, potentially masking disparities.

8. Based on Samples, Not Every Item

Index numbers are based on samples, making it impossible to account for every item in the construction of the index. The retail sales Index, derived from a sample of retail establishments, may not fully capture the diversity of consumer spending patterns.

9. Suitability of Index Numbers Depends on Circumstances

Different circumstances may require different types of index numbers. For instance, the Cost of Living Index may be more suitable for assessing the impact of inflation on households, while the Producer Price Index is more relevant for businesses tracking input costs.

10. Potential Misleading Due to Changes in Taste and Availability

Index numbers could be misleading if changes in consumer taste or the availability of certain commodities are not adequately considered. The Consumer Price Index for Food may not fully account for shifts in dietary preferences, potentially leading to inaccurate assessments of food price inflation.

In conclusion, the limitations of index numbers underscore the need for careful consideration in their construction and interpretation. The challenges range from selecting a suitable base period and dealing with approximations to addressing issues of comparability, reliability, and the impact of changing product quality and consumer preferences. Researchers and policymakers must be aware of these limitations to draw meaningful conclusions from index numbers in economic analysis.

4.1.5 Simple Index Numbers

Simple index numbers are constructed using the simple aggregative method. It is a straightforward approach to constructing a price index, particularly when dealing with multiple commodities over different time periods. This method involves the summation of total current prices in the current year $(\sum p_1)$ and the summation of total prices in the base year $(\sum p_0)$ for a specified group of commodities. The formula for the price index number (p_{01}) is then given by:

$$p_{01} = \left(\frac{\sum p_1}{\sum p_0}\right) \times 100$$

If $\sum p_1$ is the total of current prices of commodities in the current year and $\sum p_0$ is the total of prices of these commodities in the base year, then the price index number for the current year is

$$P_{01} = \frac{\sum\ p_1}{\sum\ p_0} \times 100$$

Illustration : 4.1.1

Using simple aggregate method find index number for 2007 taking 2005 as the base year from the following data:

	Prices (Rs.)	
Commodities	2005	2007
A	200	280
B	160	240
C	320	360
D	440	480

Solution:

We construct the following table:

	Prices (Rs.)	
Commodities	2005(p_0)	2007(p_1)
A	200	280
B	160	240
C	320	360
D	440	480
Total	$\sum p_0 = 1120$	$\sum p_1 = 1360$

Hence, required index number by simple aggregate method,

$$P_{01}(\text{ Price index for } 2007) = \frac{\sum p_1}{\sum p_0} \times 100$$

$$= \frac{1360}{1120} \times 100 = 121.429$$

It shows that, the prices have risen by 21.429%.

From the following data calculate index number by simple aggregate method.

Commodity	A	B	C	D
Price in 1990(Rs.)	160	258	250	139
Price in 1991(Rs.)	170	165	190	144

Commodity	Price (in Rupees)	
	1990(p_0)	1991(p_1)
A	160	170
B	258	165
C	250	190

D	139	144
Total	$\Sigma p_0 = 807$	$\Sigma p_1 = 669$

The price index number using Simple Aggregate Method is given by:

$$P_{01} = \frac{\Sigma\ p_1}{\Sigma\ p_0} \times 100$$

$$= \frac{669}{807} \times 100$$

$$= 82.90$$

Limitations of the Simple Aggregate Method

The simple aggregate method has the following limitations:

1. The relative importance of various commodities is not taken into consideration.
2. Different commodities generally have different units of measurements, e.g., wheat in Rs. per quintal, cloth per meter, petrol in Rs. Per liter and so on. It may affect the value of index number.

4.1.6 Weighted Index Numbers

The unweighted index numbers assign equal importance to all the commodities, but different commodities included are not of equal importance. In order to allow each commodity to have a reasonable influence on the index, we weight the price of each commodity by a suitable factor.

Weighted Aggregate Index Numbers

In this method, the prices in the current year and the base year are both weighted by same quantities. If w is the weight assigned to commodity then a general weighted price index will be as follows:

Weighted aggregative price index

$$P_{01} = \frac{\Sigma\ p_1 w}{\Sigma\ p_0 w} \times 100$$

There are various methods of assigning weights, thus various formulas have been developed for the construction of index numbers. Some of the important formulas are as follows:

1. Laspeyre’s index number
2. Paasche’s index number
3. Fisher’s Ideal index number
4. Marshall- Edgeworth Method
5. Kelley’s Method

1. Laspeyre’s Index Number

In this number, the base year quantities are taken as weights. Laspeyre’s price index number is given by

$$P_{01} = \frac{\Sigma\ p_1 q_0}{\Sigma\ p_0 q_0} \times 100$$

2. Paasche’s Index Number

In this number, the current year quantities are taken as weights. Paasche’s price index number is given by

$$P_{01} = \frac{\Sigma\ p_1 q_1}{\Sigma\ p_0 q_1} \times 100$$

3. Fisher’s Ideal Index Number

In this number, the geometric mean of Laspeyre’s and Paasche’s index number is taken. Fisher’s ideal price index number is given by

$$P_{01} = \sqrt{\left(\frac{\Sigma\ p_1 q_0}{\Sigma\ p_0 q_0} \times \frac{\Sigma\ p_1 q_1}{\Sigma\ p_0 q_1}\right)} \times 100$$

4. Marshall-Edgeworth Index (MEI)

The Marshall-Edgeworth Index is a price index that seeks to measure changes in the prices of a fixed basket of goods and services over time

$$P_{01} = \frac{\Sigma\ p_1 q_0 + \Sigma\ p_1 q_1}{\Sigma\ p_0 q_0 + \Sigma\ p_0 q_1} \times 100$$

5. Kelly’s Price Index

Kelly’s index number is a method of calculating index numbers that uses quantities from a given period as weights.

$$P_{01} = \frac{\Sigma\ q\ p_1}{\Sigma q\ p_0} \times 100$$

Illustration : 4.1.3

From the following data calculate price index numbers for 1980 with 1970 as base by (i) Laspeyre's method, (ii) Paasche's method, (iii) Marshall- Edgeworth method, and (iv) Fisher's ideal method.

Commodities	1970		1980	
	Price	Quantity	Price	Quantity
A	20	8	40	6
B	50	10	60	5
C	40	15	50	15
D	20	20	20	25

It is stated that Marshall- Edgeworth index number is a good approximation to Fisher's ideal index number. Verify this for the data in Part(a).

Solution

Calculations for Price Indices by different formulae

Commodities	*1970*		*1980*					
	p_0	q_0	p_1	q_1	$p_0 \cdot q_0$	$p_0 \cdot q_1$	$p_1 \times q_0$	$p_1 \times q_1$
A	20	8	40	6	160	120	320	240
B	50	10	60	5	500	250	600	300
C	40	15	50	15	600	600	750	750
D	20	20	20	25	400	500	400	500
Total					1660	1470	2070	1790

i. Laspeyre's Price Index

$$P_{01} = \frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} \times 100 = \frac{2070}{1660} \times 100 = 1.24699 \times 100 = 124.699$$

ii. Paasche's Price Index

$$P_{01} = \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1} \times 100 = \frac{1790}{1470} \times 100 = 1.2177 \times 100 = 121.77$$

iii. Marshall- Edgeworth Price Index

$$P_{01} = \frac{\Sigma p_1 q_0 + \Sigma p_1 q_1}{\Sigma p_0 q_0 + \Sigma p_0 q_1} \times 100 = \frac{2070 + 1790}{1660 + 1470} \times 100 = \frac{3860}{3130} \times 100 = 123.32$$

iv. Fisher's Price Index

$$P_{01} = \sqrt{\frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} \times \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1}} \times 100$$

$$= \sqrt{\frac{2070}{1660} \times \frac{1790}{1470}} \times 100$$

$$= \sqrt{1.24699 \times 1.2177} \times 100$$

$$= \sqrt{1.51846} \times 100$$

$$= 1.23226 \times 100$$

$$= 123.23$$

Alter:

$$P_{01} = \sqrt{P_{01}La \times P_{01}Pa}$$

$$= \sqrt{124.699 \times 121.77}$$

$$= \sqrt{15184.597}$$

$$= 123.23$$

(b) Since $P_{01}^{ME} = 123.32$ and $P_{01}^{F} = 123.23$, are approximately equal, Marshall-Edgeworth index number is a good approximation to Fisher's ideal index number.

Illustration : 4.1.4

Using weighted aggregative method find the price index number for 2005 taking 2000 as the base year from the following data:

		Prices (Rs.)	
Commodities	Weight	2000	2005
A	20	100	120
B	30	130	160
C	25	200	220

Solution:

We construct the following table:

		Prices (Rs.)			
Commodities	Weight (w)	$2000(p_0)$	$2005(p_1)$	$p_0 w$	$p_1 w$
A	20	100	120	2000	2400

B	30	130	160	3900	4800
C	25	200	220	5000	5500
Total				10900	12700

Hence, required index number by weighted aggregative method,

$$P_{01} = \frac{\sum p_1 w}{\sum p_0 w} \times 100 = \frac{12700}{10900} \times 100 = 116.5138$$

Illustration : 4.1.5

Computation of Kelly's Index Number

Materials required	*Unit*	*Quantity required*	*q*	*Price (Rs.)*			
				1963	*1973*		
				p_0	p_1	$q\ p_0$	$q\ p_1$
Cement	*100 lb*	*500 lb.*	*5*	*5.0*	*8.0*	*25*	*40*
Timber	*c.ft.*	*2,000 c.ft.*	*2000*	*9.5*	*14.2*	*19,000*	*28,400*
Steel sheets	*cwt.*	*50 cwt.*	*50*	*34.0*	*42.20*	*1,700*	*2,100*
Bricks	*per'000*	*20,000*	*20*	*12.0*	*24.0*	*240*	*480*
						$\Sigma q p_0$ =20,965	$\Sigma q p_1$ =31,020

Kelly's Price Index is given by : $P_{01} = \frac{\Sigma q\ p_1}{\Sigma q\ p_0} \times 100 = \frac{31020}{20965} \times 100 = 147.96$

4.1.7 Quantity Index Numbers

This is another index number which measures the changes occurring in the quantity of goods demanded, consumed, produced, imported or exported etc. over a given period of time.

Quantity index numbers (Q_{01}) formulae are obtained by interchanging p and q in all the formulae of price index numbers. Formulae for quantity index numbers are:

(1) Simple Aggregative Method,

$$Q_{01} = \frac{\Sigma\ q_1}{\Sigma\ q_0} \times 100$$

Where Q_1 - Quantities in the given year

Q_0 - Quantities in the base year

(2) Simple Average of Relative Method,

$$Q_{01} = \frac{\sum \frac{q_1}{q_0} \times 100}{n}$$

Where n = number of commodities

(3) Laspeyre's Quantity Index Number,

$$Q_{01} = \frac{\sum q_1 p_0}{\sum q_0 p_0} \times 100$$

(4) Paasche's Quantity Index Number,

$$Q_{01} = \frac{\sum q_1 p_1}{\sum q_0 p_1} \times 100$$

(5) Dorbish-Bowley's Quantity Index Number,

$$Q_{01} = \frac{1}{2}\left[\frac{\sum q_1 p_0}{\sum q_0 p_0} + \frac{\sum q_1 p_1}{\sum q_0 p_1}\right] \times 100$$

(6) Fisher's Quantity Index Number:

$$Q_{01} = \sqrt{\left(\frac{\sum q_1 p_0}{\sum q_0 p_0} \times \frac{\sum q_1 p_1}{\sum q_0 p_1}\right)} \times 100$$

Illustration : 4.1.6

From the following data calculate quantity index numbers for 2010 taking 2003 as base year by

(i) Laspeyre's Method

(ii) Paasche's Method

(iii) Fisher's Method

	2003		2010	
Item	Price	Quantity	Price	Quantity
A	50	10	80	12
B	10	6	12	6
C	12	6	10	8
D	25	7	30	7

Solution:

We construct the following table:

Item	2003		2010					
	p_0	q_0	p_1	q_1	p_0q_0	p_0q_1	p_1q_0	p_1q_1
A	50	10	80	12	500	600	800	960
B	10	6	12	6	60	60	72	72
C	12	6	10	8	72	96	60	80
D	25	7	30	7	175	175	210	210
Total					807	931	1142	1322

(i) Laspeyre’s quantity index number,

$$Q_{01} = \frac{\Sigma\ q_1p_0}{\Sigma\ q_0p_0} \times 100 = \frac{931}{807} \times 100 = 115.366$$

(ii) Paasche’s quantity index number,

$$Q_{01} = \frac{\Sigma\ q_1p_1}{\Sigma\ q_0p_1} \times 100 = \frac{1322}{1142} \times 100 = 115.762$$

(iii) Fisher’s quantity index number,

$$Q_{01} = \sqrt{\left(\frac{\Sigma\ q_1p_0}{\Sigma\ q_0p_0} \times \frac{\Sigma\ q_1p_1}{\Sigma\ q_0p_1}\right)} \times 100$$

$$= \sqrt{\left(\frac{931}{807} \times \frac{1322}{1142}\right)} \times 100 = 115.5635$$

Illustration : 4.1.7

From the following data calculate quantity index numbers for 2002 taking 2000 as base year by

(i) Dorbish-Bowley’s Method

(ii) Fisher’s method

	2000		2002	
Commodity	Price	Quantity	Price	Quantity
A	6	25	8	28
B	4	30	6	30
C	8	40	12	36

Solution:

We construct the following table:

Commodity	2000		2002		p_0q_0	p_0q_1	p_1q_0	p_1q_1
	p_0	q_0	p_1	q_1				
A	6	25	8	28	150	168	200	224
B	4	30	6	30	120	120	180	180
C	8	40	12	36	320	288	480	432
Total					590	576	860	836

(i) Dorbish-Bowley's quantity index number,

$$Q_{01} = \frac{1}{2}\left[\frac{\Sigma\ q_1p_0}{\Sigma\ q_0p_0} + \frac{\Sigma\ q_1p_1}{\Sigma\ q_0p_1}\right] \times 100$$

$$= \frac{1}{2}\left[\frac{576}{590} + \frac{836}{860}\right] \times 100 = 97.418$$

(ii) Fisher's quantity index number,

$$Q_{01} = \sqrt{\left(\frac{\Sigma\ q_1p_0}{\Sigma\ q_0p_0} \times \frac{\Sigma\ q_1p_1}{\Sigma\ q_0p_1}\right)} \times 100$$

$$= \sqrt{\left(\frac{576}{590} \times \frac{836}{860}\right)} \times 100 = 97.418$$

Illustration : 4.1.8

From the following data, calculate Laspeyre's quantity index number.

Commodity	**Price**		**Quantity**	
	2022	**2023**	**2022**	**2023**
A	20	22	16	14
B	16	22	12	14
C	18	25	12	10
D	20	26	8	12

Solution

Construction of quantity index number

Commodity	Price		Quantity		q_1p_0	q_0p_0
	p_0	p_1	q_0	q_1		
A	20	22	16	14	280	320
B	16	22	12	14	224	192
C	18	25	12	10	180	216
D	20	26	8	12	240	160
					Σq_1p_0 **924**	Σq_0p_0 **888**

Laspeyre's index q_{01} (L) = $\frac{\Sigma q_1 p_0}{\Sigma q_0 p_0}$ x 100

$= \frac{924}{888}$ x 100 = 104.054

Illustration 4.1.9

From the following data, calculate Paashe's quantity index number.

Commodity	Price		Quantity	
	2022	**2023**	**2022**	**2023**
A	22	26	16	14
B	18	26	12	16
C	20	25	14	10
D	29	33	8	12

Solution

Construction of quantity index number

Commodity	Price		Quantity		q_1p_1	q_0p_1
	p_0	p_1	q_0	q_1		
A	22	26	16	14	364	416
B	18	26	12	16	416	312
C	20	25	14	10	250	350
D	29	33	8	12	396	264
					Σq_1p_1 **1426**	Σq_0p_1 **1342**

Paashe's quantity index q_{01} (P) = $\frac{\Sigma q_1 p_1}{\Sigma q_0 p_1}$ x 100

$= \frac{1426}{1342}$ x 100

= 106.259

Illustration 4.1.10

From the following data, calculate Fisher's quantity index number.

Commodity	Price		Quantity	
	2022	**2023**	**2022**	**2023**
A	20	22	6	4
B	20	26	2	6
C	25	27	4	10
D	29	33	8	2

Solution

Construction of quantity index number

Commodity	Price		Quantity		q_1p_1	q_0p_1	q_0p_0	q_1p_0
	p_0	p_1	q_0	q_1				
A	20	22	6	4	88	132	120	80
B	20	26	2	6	156	52	40	120
C	25	27	4	10	270	108	100	250
D	29	33	8	2	66	264	232	58
					$\Sigma q_1 p_1$ **580**	$\Sigma q_0 p_1$ **556**	$\Sigma q_0 p_0$ **492**	$\Sigma q_1 p_0$ **508**

Fisher's quantity index q_{01} (F) = $\sqrt{\frac{\Sigma q_1 p_0}{\Sigma q_0 p_0} x \frac{\Sigma q_1 p_1}{\Sigma q_0 p_1}}$ x 100

$= \sqrt{\frac{508}{492} x \frac{580}{556}}$ x 100

= $\sqrt{1.0325x1.0431}$ x 100

= $\sqrt{1.077}$ x 100

= 103.778

Recap

- An index number is a statistical tool used to measure the relative changes in a variable or a group of related variables over time or across different locations.
- Index numbers are used to measure percentage changes in economic variables and help in economic analysis, policy formulation, and decision-making.
- Index numbers provide approximate rather than exact results, often reflecting perceptions rather than precise conditions
- Index numbers are often based on a limited set of items, which may not capture the full range of economic conditions or factors.
- The Simple Aggregative Method is a straightforward approach to constructing a price index, particularly useful when dealing with multiple commodities over different time periods
- Weighted Index Numbers assign appropriate weights to each commodity, ensuring a more accurate representation based on their relative importance.
- Weighted Index Numbers include Laspeyre's Index Number, Paasche's Index Number, Fisher's Ideal Index Number, Marshall-Edgeworth Index (MEI), and Kelly's Price Index.
- Quantity index number which measures the changes occurring in the quantity of goods demanded, consumed, produced, imported or exported, etc. over a given period of time.

Objective Questions

1. Which widely used index number reflects changes in the prices of a specified group of commodities?
2. What economic implications do deriving index numbers, such as CPI or WPI, have?

3. Which type of index measures relative changes in a single variable concerning a base?

4. What role do price index numbers play in economics?

5. What is a characteristic feature of index numbers in measuring changes in a group of items?

6. Which index represents the average change in prices of a basket of goods and services consumed by households?

7. What limitation arises from the fact that index numbers are often based on a limited number of items?

8. What is a crucial challenge in constructing index numbers, as highlighted in the limitations?

9. Which type of index number is constructed through the summation of total current prices and total prices in the base year for a specified group of commodities?

10. What is the significance of the Simple Aggregative Method in constructing simple index numbers?

11. Which characteristic is associated with the Laspeyres Index?

12. What is the primary advantage of using weighted index numbers?

Answers

1. Consumer Price Index

2. Impact on policy decisions related to commodity prices

3. Simple index

4. Influence decisions related to wages, inflation, and economic policies

5. Comparative or relative measure

6. Consumer Price Index (CPI)

7. Based on samples, not every item

8. Choosing a suitable base period

9. Simple Index

10. It is straightforward, especially for dealing with multiple commodities

11. Current-year quantities as weights

12. They provide an unbiased representation of changes

Assignments

1. Describe various approaches to profit maximisation.
2. Using an example, show the relationship between AP and MP.
3. Explain the shape of TP, AP, and MP curves.
4. Using simple aggregate method find index numbers for 2000 and 2005 taking 1995 as the base year from the following data:

Commodities	Prices (Rs.)		
	1995	2000	2005
A	60	70	90
B	44	50	60
C	108	128	150
D	40	50	70
E	30	36	45

5. Using weighted aggregative method find the price index number for 2007 taking 2004 as the base year from the following data:

Commodities		Prices (Rs.)	
	Weight	2004	2007
I	5	10	20
II	3	15	20
III	2	20	24
IV	4	18	23

6. Calculate price index number of the year 2006 with 1996 as base year from the following data using:

(i) Laspeyre's

(ii) Paasche's and

(iii) Fisher's formulae.

Commodity	1996		2006	
	Price (Rs.)	**Value (Rs.)**	**Quantity consumed**	**Value (Rs.)**
A	10	1500	160	1760
B	12	1080	100	1300
C	15	900	60	960

7. From the following data, construct the Laspeyres's, Paasche's and Fisher's indices of prices

Items	Base Year		Current Year	
	p_0	q_0	p_1	q_1
A	4	20	10	15
B	8	4	16	5
C	2	10	4	12
D	10	5	20	6

8. From the following data, calculate the price index numbers for 1992 with 1982 as base by –

a. Laspeyre's method

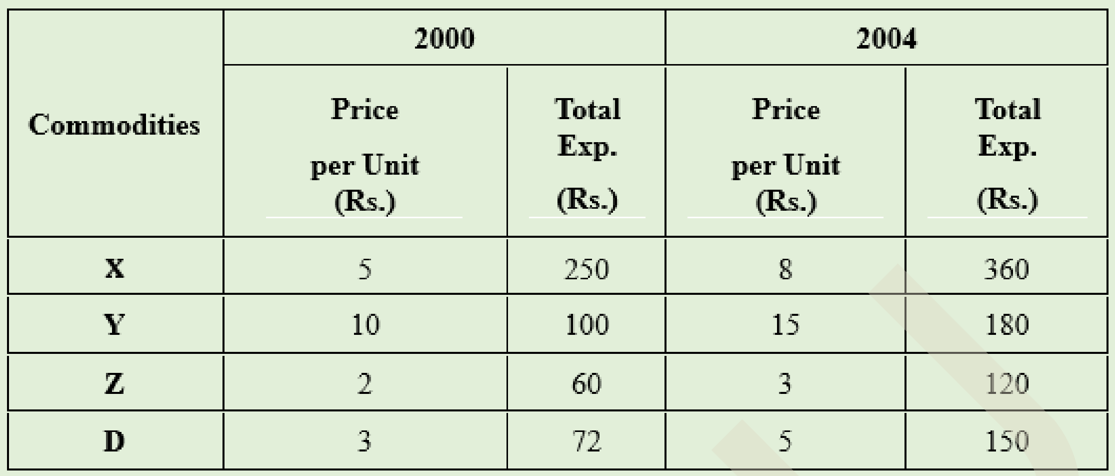
b. Paasche's method

c. Marshall – Edgeworth method

d. Fisher's ideal method

Item	1990		2000	
	Price (Rs.)	Quantity (unit)	Price (Rs.)	Quantity (unit)
Maize	70	28	140	21.0
Millet	175	35	210	17.5
Sugar	140	52.5	175	52.5
Coconut	70	70.0	70	87.5

9. Calculate Fisher's Ideal Index number from the following information:



Commodities	2000		2004	
	Price per Unit (Rs.)	Total Exp. (Rs.)	Price per Unit (Rs.)	Total Exp. (Rs.)
X	5	250	8	360
Y	10	100	15	180
Z	2	60	3	120
D	3	72	5	150

References

1. Goon, A., Gupta, M. & Dasgupta, B. (1963). *Fundamentals of Statistics*. The World Press Private Limited, Calcutta.

2. Karmel, P. H. & Polasek, M. (1970). *Applied Statistics for Economists*. Pitman Publishing, South Australia.

Suggested Readings

1. T. Rajaretnam, *Statistics for Social Sciences*, Sage India.

2. Gupta S. and V.K. Kapoor, *Fundamentals of Applied Statistics*, S.Chand and Sons, New Delhi.

3. Monga, G.S. *Mathematics and Statistics for Economics*, Vikas Publishing, New Delhi

2

UNIT

Test of Index numbers and CPI, WPI

Learning Outcomes

After reading this unit, the learner will be able to:

- calculate the various tests of index numbers
- understand the concept of Consumer Price Index (CPI)
- know the Wholesale Price Index (WPI)

Prerequisites

In the previous chapter, we explored the concept of index numbers, which are statistical tools used to track changes in variables such as prices, quantities, or other economic indicators over time. These index numbers help compare relative changes by expressing them as percentages or ratios. In this unit, we will focus on the tests of index numbers, specifically the Consumer Price Index (CPI) and Wholesale Price Index (WPI). Let us consider a practical example: your family's monthly grocery expenses. The prices of everyday items like milk, bread, or vegetables tend to fluctuate, which can have a significant impact on your budget. As a result, your consumption patterns may change. The tool that helps measure these changes in the cost of living that tool is CPI. Similarly, for a small business owner who needs to purchase raw materials in bulk, any fluctuation in the prices of these goods will affect production costs and, ultimately, the prices consumers pay. These changes can be tracked using the Wholesale Price Index (WPI). In this unit, we will delve deeper into the details of CPI and WPI and explore their practical applications.

Keywords

Unit Test, Time Reversal Test, Factor Reversal Test, Circular Test, Consumer Price Index, Wholesale Price Index

Discussion

4.2.1 Test of Index Numbers

Index numbers are essential statistical tools used to measure changes in economic variables such as prices, production, and consumption over time. However, the accuracy and reliability of an index number depend on its ability to meet specific criteria. To ensure consistency and validity, various tests are applied to index numbers. These tests, such as the time reversal test, factor reversal test, and circular test, evaluate the behavior of index numbers under different conditions.

An index number is said to be good if it satisfies these tests. The major test of index numbers are:

1. Unit Test
2. Time Reversal Test
3. Factor Reversal Test
4. Circular Test

1. Unit Test

This test is employed to ensure that the index number remains independent of the units of measurement used for prices and quantities. Most index number formulae fulfill this test, with one notable exception being the simple aggregative index number.

The principle behind this test is to ascertain that the resulting index is not influenced by the specific units in which prices and quantities are reported. The need for such independence arises to maintain the reliability and comparability of the index across different units of measurement.

While various index number formulae successfully pass this test, the simple aggregative index number fails to meet this criterion. Therefore, careful consideration and selection of an appropriate index formula are essential to ensure the robustness and validity of the index number, particularly in applications where unit independence is critical.

2. Time Reversal Test

This test is used to verify that the index number formula maintains the time consistency, i.e., it is working in both ways in time forward and backward.

Symbolically,

$$P_{01} \times P_{10} = 1$$

Laspeyre's formula, Paasche's formula does not satisfy this test. Fisher's formula, Marshall-Edgeworth's and Walsh's index number formula satisfies this test. We need to omit the factor 100 from both P_{01} and P_{10}.

3. Factor Reversal Test

This test is used to verify that the product of a price index and the quantity index is equal to the corresponding true value ratio.

Symbolically,

$$P_{01} \times Q_{01} = \frac{\sum p_1 q_1}{\sum p_0 q_0}$$

Only Fisher's index number formula satisfies this test. We need to omit the factor 100 from both ends. Since Fisher's index number formula satisfies the time reversal and factor reversal tests both. That is why Fisher's index number is called ideal index number.

4. Circular Test

This test is an extension of time reversal test for more than two periods. It is based on the shifting of the base period.

Symbolically,

For three years: $P_{01} \times P_{12} \times P_{20} = 1$

For four years: $P_{01} \times P_{12} \times P_{23} \times P_{30} = 1$

This test is satisfied by a simple aggregate index number formula and Kelly's fixed weight index number formula. Even Fisher's ideal index number formula does not satisfy this test. We need to omit the factor 100 from all indices.

Examples 1

Calculate Laspeyre's and Paasche's price index numbers for the year 2005 taking 2000 as the base year from the following data. Prove that both the formulae do not satisfy the time reversal test.

	2000		2005	
Commodity	Price	Quantity	Price	Quantity
A	4	6	5	10
B	5	8	12	
C	6	4	5	6

Solution:

We construct the following table:

Item	2000		2005		p_0q_0	p_0q_1	p_1q_0	p_1y_1
	p_0	q_0	p_1	q_1				
A	4	6	5	10	24	40	30	50
B	5	8	6	12	40	60	48	72
C	6	4	5	6	24	36	20	30
Total					88	136	98	152

Laspeyre's price index number,

$$P_{01} = \frac{\sum p_1q_0}{\sum p_0q_0} \times 100 = \frac{98}{88} \times 100 = 111.364$$

Paasche's price index number,

$$P_{01} = \frac{\sum p_1q_1}{\sum p_0q_1} \times 100 = \frac{152}{136} \times 100 = 111.765$$

Time Reversal Test: The time reversal test is satisfied if

$$P_{01} \times P_{10} = 1$$

For Laspeyre's index number,

$$P_{01} = \frac{\sum p_1q_0}{\sum p_0q_0} = \frac{98}{88} \text{ and } P_{10} = \frac{\sum p_0q_1}{\sum p_1q_1} = \frac{136}{152}$$

Now, $P_{01} \times P_{10} = \frac{98}{88} \times \frac{136}{152} \neq 1$

Thus, Laspeyre's index number does not satisfy the ime reversal test.

For Paasche's index number,

$$P_{01} = \frac{\sum p_1q_1}{\sum p_0q_1} = \frac{152}{136} \text{ and } P_{10} = \frac{\sum p_0q_0}{\sum p_1q_0} = \frac{88}{98}$$

Now, $P_{01} \times P_{10} = \frac{152}{136} \times \frac{88}{98} \neq 1$

Thus, Paasche's index number does not satisfy the time reversal test.

Example 2

Calculate Fisher's price index number for the year 2010 taking 2002 as the base year from the following data. Verify whether Fisher's formula satisfies the time reversal test.

	2002		2010	
Commodity	Price	Quantity	Price	Quantity
A	10	5	15	5
B	5	5	5	4
C	10	20	12	10
D	5	10	10	10

Solution: We construct the following table:

Item	2002			2010				
	p_0	q_0	p_1	q_1	p_0q_0	p_0q_1	p_1q_0	p_1q_1
A	10	5	15	5	50	50	75	75
B	5	5	5	4	25	20	25	20
C	10	20	12	10	200	100	240	120
D	5	10	10	10	50	50	100	100
Total					325	220	440	315

Fisher's price index number,

$$P_{01} = \sqrt{\left(\frac{\Sigma\ p_1q_0}{\Sigma\ p_0q_0} \times \frac{\Sigma\ p_1q_1}{\Sigma\ p_0q_1}\right)} \times 100$$

$$= \sqrt{\left(\frac{440}{325} \times \frac{315}{220}\right)} \times 100 = 139.229$$

Time Reversal Test: The time reversal test is satisfied if

$$P_{01} \times P_{10} = 1$$

$$P_{01} = \sqrt{\left(\frac{\Sigma\ p_1q_0}{\Sigma\ p_0q_0} \times \frac{\Sigma\ p_1q_1}{\Sigma\ p_0q_1}\right)} = \sqrt{\left(\frac{440}{325} \times \frac{315}{220}\right)}$$

$$\text{and } P_{10} = \sqrt{\left(\frac{\Sigma p_0q_1}{\Sigma p_1q_1} \times \frac{\Sigma p_0q_0}{\Sigma p_1q_0}\right)} = \sqrt{\left(\frac{220}{315} \times \frac{325}{440}\right)}$$

Now, $P_{01} \times P_{10} = \sqrt{\left(\frac{440}{325} \times \frac{315}{220}\right)} \times \sqrt{\left(\frac{220}{315} \times \frac{325}{440}\right)} = 1$

Thus, Fisher's index number satisfies the time reversal test.

4.2.2 Consumer Price Indices (CPI)

In India, Consumer Price Indices (CPI) serve as indicators tracking variations in the overall price levels of goods and services acquired by households for consumption. Widely utilized as a macroeconomic gauge for inflation, governments and central banks employ CPI as a tool for inflation targeting and to oversee price stability. Additionally, CPI functions as a deflator in national accounts and plays a pivotal role in indexing dearness allowance for employees to account for increases in prices.

The Central Statistics Office (CSO) under the Ministry of Statistics and Programme Implementation has implemented a revision of the base year for the Consumer Price Index (CPI), transitioning from 2010=100 to 2012=100. This revision comes into effect starting from the release of indices for the month of January 2015. The previous series, based on 2010=100, commenced its publication in January 2011. The updated series incorporates several methodological changes aimed at enhancing the robustness of the indices.

The Field Operations Division of the National Sample Survey Office (NSSO) and the designated State/Union Territory (UT) Directorates of Economics and Statistics gather monthly price data from 1181 villages and 1114 markets across 310 selected towns. The collection of prices is facilitated through Web Portals, with the National Informatics Centre (NIC) developing the web portal for rural prices, and the Computer Centre in the Ministry of Statistics and Programme Implementation (MoSPI) developing the web portal for urban prices.

4.2.3 Wholesale Price Index (WPI)

The Wholesale Price Index (WPI) plays a significant role in monitoring the dynamic fluctuations of prices at the wholesale level. Given the continuous changes in prices in today's dynamic environment, WPI serves as a deflator for various nominal macroeconomic indicators, including Gross Domestic Product (GDP). Its application extends to being a crucial factor in the formulation of trade, fiscal, and other economic policies by the Government, providing essential insights into inflation estimates based on WPI.

WPI is employed for the implementation of escalation clauses in the supply of raw materials, machinery, and construction work. Business firms, seeking effective strategies to adapt to price changes, frequently incorporate price adjustment (escalation) clauses in long-term sales and purchase contracts, recognizing WPI as a valuable objective indexing tool for such clauses. This recognition is widespread among business professionals, economists, statisticians, and accountants.

Recap

- The major tests of index numbers are time reversal, unit test, factor reversal test, and circular test.
- The Time Reversal Test ensures that the index number does not change if we reverse the time periods.
- Factor reversal test ensures that if the index number is computed using weights of two different periods, the results should remain the same.
- Unit reversal test ensures the unit of measurement should not affect the index number.
- Circular test ensures consistency when using indices for different base years in a circular fashion.
- CPI measures the average change in the prices of goods and services consumed by households over time.
- CPI reflects changes in the cost of living.
- WPI measures the average change in the prices of goods at the wholesale level before they reach the retail market.

Objective Questions

1. Which test assesses the impact of reversing the order of observations?
2. What is the focus of the Factor Reversal Test in time series analysis?
3. Which property does the Time Reversal Test examine in a time series?
4. Which division is responsible for gathering monthly price data for CPI in India?
5. What change has been implemented regarding the base year of the Consumer Price Index (CPI) in India?
6. In what context is WPI commonly used for the implementation of escalation clauses?
7. How does WPI contribute to the formulation of economic policies by the government?

Answers

1. Time Reversal Test
2. Multiplying time series by a constant factor
3. Statistical properties after reversing the order
4. National Sample Survey Office (NSSO)
5. Transition from 2010=100 to 2012=100
6. In the supply of raw materials, machinery, and construction work
7. By serving as a determinant for fiscal policies

Assignments

1. Calculate Fisher’s price index number for the year 2007 taking 2005 as the base year from the following data. Prove that Fisher’s index number satisfies both the time reversal and factor reversal test.

Commodity	2005		2007	
	Price	Quantity	Price	Quantity
A	5	20	25	
B	2	50	2	75
C	4	25	6	25

2. Calculate Fisher’s Ideal Index number and show that it satisfies time reversal and factor reversal test.

Commodity	1998		2010	
	Price	Quantity	Price	Quantity
I	15	55	12	65
II	16	20	13	15
III	18	13	16	17
IV	17	25	15	20

3. Calculate Fisher's Ideal Index number and show that it satisfies time reversal and factor reversal test.

Commodity	2006		2008	
	Price	Value	Price	Value
I	10	120	14	112
II	10	100	10	80
III	14	154	17	136
IV	22	330	27	270
V	7	56	10	80
VI	4	40	6	24

References

1. Goon, A., Gupta, M. & Dasgupta, B. (1963). *Fundamentals of Statistics*. The World Press Private Limited, Calcutta.

2. Karmel, P. H. & Polasek, M. (1970). *Applied Statistics for Economists*. Pitman Publishing, South Australia.

Suggested Readings

1. T. Rajaretnam, *Statistics for Social Sciences*, Sage India.

2. Gupta S. and V.K. Kapoor, *Fundamentals of Applied Statistics*, S.Chand and Sons, New Delhi.

3. Monga, G.S. *Mathematics and Statistics for Economics*, Vikas Publishing, New Delhi

5

BLOCK

Population and Sampling

UNIT 1

Methods of Sampling

Learning Outcomes

After learning the unit, the learner will be able to:

- help the students to enrich their basic understanding in the concepts of population and sample.
- increase the ability of students to familiarise the statistical concepts and terminologies like parameters, statistic etc
- help the students to know the practical concepts of the probability sampling methods and their application levels in the real life.

Prerequisites

Students should aware about the fundamentals of statistical concepts and basic practical ideologies. They have the basic potential to identify the population and sample and the basic framework of probability.

Keywords

Population, Sample, Parameter, Statistic, Probability Sampling

5.1.1 Populations and Samples

In a statistical investigation the interest usually lies in the assessment of the general magnitude and the study of variation with respect to one or more characteristics relating to individuals belonging to a group. This group of individuals under study is called population or universe. In any statistical investigation, on the basis of examining a part of the population is termed as sample, and which is drawn from the population in scientific manner. In modern decision making process in different fields of human activity, including the ordinary actions of our daily life, most of our decisions and attitudes depends very much upon the inspection or examination of only a few objects or items out of the total lot. This process of studying only the sample data and then generalizing the results to the population, i.e; drawing inferences about the population on the basis of sample study.

Here, we are going to look into the details of population and sample. First, we are discussing about the population.

Population or Universe

A population refers to a totality of all possible observations, measurements or outcomes. In statistics, population is an aggregate of objects, animate or inanimate, under study. A population is the entire group that you want to draw conclusions about. For example, the annual income of all households living in a certain locality constitutes a population, all of the students in a classroom, all possible outcomes of tossing a coin a certain number of times. If an enquiry is intended to determine the average per capita income of the people in a particular city, the population will comprise all the earning people in that city. On the other hand, if we want to study the expenditure habits of the families in that city, then the population will consist of all the households in that city. As a totality, the population varies in size from situation to situation, and is described and delimited basically by nature and extent of our interest in a given situation or context.

From the point of view of sampling, we do distinguish between various types of populations. Apart from defining the scope of sampling, the type of population to be sampled also determines the method of sampling to be used. The various populations are distinguished as finite population and infinite population.

A population containing finite number of objects or items is known as finite population.

For example, the students in a college, the population of a city or town etc.

A population having an infinite number of objects or with the number of objects as large as to appear practically infinite, is termed as infinite population. For example, the population of stars in the sky, the number of heads (or tails) to be obtained in repeated experiments of tossing two coins under the same conditions etc.

The population may be further classified as existent or hypothetical.

A population consisting of concrete objects is known as existent population. For example, the population of the books in a library, the population of the aero planes in the Indian Air Force. On the other hand, a population does not consist of concrete objects, i.e, it consists of imaginary objects then it is called hypothetical population. For example, the population of the throws of a die or a coin, thrown infinite number of times.

Populations also differ in terms of the nature of access gained through sampling. Accordingly, a distinction is made between target population and the sampled population. From this point of view, a population is defined as consisting of elementary sampling units or primary sampling units.

i. Elementary sampling units refers to events, outcomes, objects, subjects, or units one or more of whose characteristics are actually observed, measured, counted, or ranked. For example, the students of a college whose weight, height, or intelligence may be measured are the elementary units.

ii. Primary sampling units, on the other hand, refers to groups or clusters of elementary sampling units. In the above example, students studying in a different class or those doing different courses in a college form different cluster. Similarly, respondents in a survey classified according to some criterion (say, sex, occupation, or religion) also form different groups or clusters. The resultant clusters in each case are the primary units. Whereas Sampling may be done with reference to elementary or primary sampling units, the process of observation is finally done with the reference to the element sampling units. it is, therefore, necessary that the elementary units are clearly defined so that their location and identification do not cause any problem during the course of survey. Thus, a clearly defined and identified population with respect to its elementary sampling units is known as the target population.

Situations do arise when all elementary sampling units in the targeted population maynot be located, or are otherwise not accessible, for purposes of observation. In all such cases, sampling has to be restricted only to those units that are actually within the reach of investigators for observation. This calls for listing of such units. The list prepared is called a frame. Any such frame thus represents a population that is actually sampled. It is known as a sampled population,whose coverage is invariably short of the target population.

The distinction between target and sampled population offers the following three important implications:

i. It is obvious that the results based on a sample can be validly used for drawing inferences only for the sampled population, and not for the target population. The generalizations made for the sampled population maybe applicable to the target population if the difference between the two is small enough to be ignored. Where the difference is either too large, or not possible to evaluate, it is not safe to use the results based on the sampled population for making generalisations about the target population.

ii. It is highly questionable to reach any conclusion for the target population by studying a sample drawn with reference to readymade frames. Telephone directories, cooking gas and automobile registration records, registered member of cooperative stores, income tax payers' list, and the like are all ready-made frames. Taking recourse to any such frame for purpose of sampling often leads to spurious results.

iii. Any attempt so made without going into the nature of the problem at the hand, clearly defining and identifying the target population, and examining the problem of accessibilityand location of the elementary units, is most unscientific. This rudely overrules the basic tenets of sampling theory. It is, therefore, necessary that ready-made frames are carefully examined from all possible angles before being put to use.

Sample

A sample is a set of measurements taken from a process or series of experiments. It can be regarded as having been drawn from a large population of measurements covering a large number of observations.

The set of observations that form a population need not naturally be measurements, they may as well correspond to some attributes or characteristics whose presence or absence maybe under concentration.

For Example (1); 5 out of 100 items produced on a machine are examined. The sample size is 5 and the population size is 100.

For Example (2); a coin is tossed 50 times. The number of tosses at a time constitutes a sample and the observations are the heads or tails that result from it. On the basis of the sample, an inference can be made about the population. Repeated tossed of a coin comprise a population in which the observation is either head or a tail.

Sampling is simply the process of learning about the population on the basis of a sample drawn from it. Thus, in the sampling technique instead of every unit of the universe, only a part of the universe is studied and the conclusions are drawn on that basis for the entire universe. A Sample is a subset of population units. The process of sampling involves three elements:

i. Selecting the sample

ii. Collecting the information

iii. Making an inference about the population

The three elements cannot generally be considered in isolation from one another. Sample selection, data collection and estimation are all interwoven and each has an impact on the others. Sampling is not haphazard selection. It embodies definite rules for selecting the sample. But, having followed a set of rules for sample selection, we cannot consider the estimation process independent of it - estimation is guided by the manner in which sample has been selected.

Although much of the development in the theory of sampling has taken place only in recent years, the idea of sampling is a pretty old. Since times immemorial people have examined a handful of grains to ascertain the quality of the entire lot. A housewife examines only two or three grains of boiling rice to know whether the pot of rice is ready or not. A doctor examines a few drops of blood and draws conclusion about the blood content of the whole body. A business man place orders for material by examining only a small sample of the same. A teacher may put questions to one or two students and find out whether the class as a whole is following the lesson. In fact, there is hardly any field where the technique of sampling is not used either consciously or unconsciously.

It should be noted that a sample is not studied for its own sake. The basic objective of its study is to draw inference about the population. In other words, sampling is a tool which helps to know the characteristics of the universe or population by examining only a small part of it. The values obtained from the study of a sample such as the average and dispersion are known as 'statistics'. On the other hand, such values for the population are called 'parameters'.

5.1.2 Methods of Sampling

The various methods of sampling can be grouped under two broad heads:

1. Probability Sampling (Random Sampling)
2. Non-Probability Sampling (Non-Random) Sampling

Probability sampling methods are those in which every item in the universe has a known chance, or probability, of being chosen for the sample. This implies that the selection of sample items is independent of the person making the study, that is, the sampling operation is controlled so objectively that the items will be chosen strictly at random.

Non-probability sampling methods are those which do not provide every item in the universe with they known chance of being included in the sample. The selection process is, at least, partially subjective.

It may be noted that the term 'random sample' is not used to describe the data in the sample but the process employed to select the sample. Randomness is this a property of the sampling procedure instead of an individual sample. As such, randomness can and the process the sampling in a number of ways and hands random samples may be of many kinds.

5.1.3 Probability Sampling Methods

Non-random sampling is a process of sample selection without the use of randomization. In other words, Non random sample is selected on a basis other than the probability consideration such as convenience, judgement etc.

The most important difference between random and non-random sampling is that the pattern of sampling variability can be ascertained in case of random sampling, non-random sampling, there is no way of knowing the pattern of variability in the process.

5.1.4 Simple or Unrestricted Random Sampling

Simple random sampling refers to that sampling technique in which each and every unitof the population has an equal opportunity of being selected in the sample. In simple random sampling which items get selected in the sample is just a matter of chance - personal bias of the investigator does not influence the selection. It should be noted that the word random does not mean 'haphazard' or 'hit-or-miss' - it rather means that the selection process is such that chance only determines which items will be included in the sample.

As pointed out by Chou, when a sample of size n is drawn from a population with elements, the sample is a simple random sample if any of the following is true and, if any of the following is true, so are the other two:

All items of the sample are selected independently of one another and all items in the population have the same chance of being include in the sample. By independence of selection; we mean that the selection of a particular item in one draw has no influence on the probability of selection in any other draw

At each selection, all remaining items in the population have the same chance of being drawn. If sampling is made with replacement, i.e; when each unit drawn from the population is returned prior to drawing the next unit, each item has a probability of 1/N of being drawn at each selection. If sampling is without replacement, i.e; when each unit drawn from the population is not returned prior to drawing the next unit, the probabilityof selection of each item remaining in the population at the first to draw is 1/N, at the second draw is 1/(N-2) and the third draw is 1/(N-3) and so on. It should be noted that sampling with replacement has very limited and special use and statistics - we are mostlyconcerned with the sampling without replacement.

- All the possible samples of a given size n are equally likely to be selected

For example, a simple random sample would be the names of 25 employees being chosenout of a hat from a company of 250 employees. In this case, the population is all 250 employees, and the sample is random because each employee has an equal chance of being chosen.

To ensure randomness of selection one may adopt either the lottery method or consult table of random numbers

a. Lottery method

This is a very popular method of taking a random sample. Under this method, all items of the universe are numbered or named on separate slips of paper of identical size and shape. These slips are then folded and mixed up in a container or drum. A blind fold selection is then made of the number of slips required to constitute the desired sample size. The selection of items thus depends entirely on chance.

For example, if we want to take a sample of 10 persons out of a population of 100, the procedure is to write the names of the 100 persons on separate slips of paper, fold these slips,mix them thoroughly and then make a blindfold selection of 10 slips.

The example in which the names of 25 employees out of 250 are chosen out of a hat is an example of the lottery method at work. Each of the 250 employees would be assigned a number between 1 and 250, after which 25 of those numbers would be chosen at random.

The method is very popular in lottery draws where a decision about prizes is to be made. However, while adopting the lottery method, it is absolutely essential to see that the slips are of identical size, shape and colour, otherwise there is a lot of possibility of personal prejudice and bias affecting the results.

b. Table of Random Numbers

The lottery method becomes quite cumbersome as the size of population increases. An alternative method of random selection is that of using the table of random numbers.

The random numbers are generally obtained by some mechanism which, when repeated a large number of times, ensures approximately equal frequencies for the number from 0 to 9 and also proper frequency for various combination of numbers (such as 00, 01,

......,999, etc.) that could be expected in a random sequence of the digit 0 to 9.

Several standard tables of random numbers are available, among which thefollowing may be specially mentioned, as they have been tested extensively for randomness:

- Tippet's (1972), Random Number Tables consisting of 41600 random digits grouped into 10400 sets of four-digit random numbers;
- Fisher and Yates (1938) tables of random numbers with 15000 random digitsarranged into 1500 sets of ten-digit random numbers
- Kendall and Babington Smith (1939) table of random numbers consisting of 100000 random digits grouped into 25000 sets of four-digit random numbers
- Rand Corporation (1955) table of random numbers consisting of one millionrandom digits grouped into 2 lakh sets of five-digit random numbers
- C R Rao, Mitra and Mathai (1966) table of random numbers.

Tippet's table of random numbers is most popularly used in practice. We give below the first forty sets from Tippet's table as an illustration of the general appearance of random numbers.

2952	6641	3992	9792	7969	5911	3170	5624
4167	9524	1545	1396	7203	5356	1300	2693
2670	7483	3408	2762	3563	1089	6913	7991
0560	5246	1112	6107	6008	8125	4233	8776
2754	9143	1405	9025	7002	6111	8816	6446

It is important that the starting point in the table of random numbers is selected in some random fashion so that every unit has an equal chance of being selected.

One may question, and quite rightly, as to how it is ensured that these digits are random. It may be pointed out that the digits in the table were chosen haphazardly but the real guarantee of their randomness lies in practical tests. Tippet's numbers have been subjected to numerous tests and used in many investigations and their randomness has been well-established for all practical purposes. An example to illustrate how Tippet's table of random numbers maybe used is given below. Suppose, we have to select 20 items out of 6000. The procedure is to number all the 6000 items from 1 to 6000. A page from Tippet's table may then be consulted andthe first twenty numbers up to 6000 noted down. Items bearing those numbers will be included inthe sample. Making use of the portion of the table given above, the required numbers are:

2952	3992	5911	3170	5624	4167
1545	1396	5356	1300	2693	2370
3408	2762	3563	1089	0560	5246
1112	4233				

The items which bear the above numbers constitute the sample.

Universe size less than 1000. If the size of universe is less than 1000 the procedure will be different. Tippet's numbers are available only in four figures. Thus, for example, if it is desiredto take a sample of 10 items out of 400, all items from 1 to 400 should be numbered as 0001 to 0400. We may now select 10 numbers from the table which are up to 0400.

Universe size less than 100. If the size of the universe is less than 100, the table is used as follows:

Suppose, ten numbers from out of 0 to 80 are required. We start anywhere in the table and write down the numbers in pairs. The table can be read horizontally, vertically, diagonally or in any methodical way. Starting with the first and reading horizontally first (see table given above) we obtain 29, 52, 66, 41, 39, 92, 97, 92, 79, 69, 59, 11, 31, 70, 56, 24, 70, 56,24, 41, 67, and so on. Ignoring the numbers greater than 80, we obtain ten random numbers, namely 29, 52, 66, 41, 39, 79, 69, 59, 11 and 31.

Fisher and Yate's tables consist of 15000 numbers. These have been arranged in two digits in 300 blocks, each block consisting of 5 rows and 5 columns. Kendall and Smith also constructed random numbers (10000 in all) by using a randomizing machine. However, this method of random selection cannot be followed in case of articles like ghee, oil, petrol, wheatetc.

Features of Simple Random Sampling

- Since the selection of items in the sample depends entirely on chance, there is no possibility of personal bias affecting the results.
- As compared to judgement sampling a random sample represents the universe in a better way. As the size of the sample increases, it becomes increasingly representative of the population.
- The analyst can easily assess the accuracy of this estimate because sampling errors follow the principal of chance. The theory of random sampling is further developed than that of any other type of sampling which enables the analyst to provide the most reliable information at the least cost.

Limitations of Simple Random Sampling

- The use of simple random sampling necessitates a completely catalogued universe from which to draw the sample. But it is often difficult for the investigator to have up to date list of all the items of the population to be sampled. This restricts the use of this methodin economic and business data where very often we have to employee restricted random sampling designs.
- The size of the sample required to ensure statistical reliability is usually larger under random sampling than stratified sampling.
- From the point of view of field survey, it has been claimed that cases selected by random sampling tends to be too widely dispersed geographically and that the time and cost of collecting data become too large.
- Random sampling may produce the most non random looking results. For example, thirteen cards from a well shuffled pack of playing cards may consists of one suit. But theprobability of this type occurrence is very very low.

5.1.5 Restricted Random Sampling

In the case of restricted random sampling, we have classified the methods into three, i.e.; Systematic sampling, Stratified sampling, Multi-stage or Cluster sampling. We will discuss one by one.

1. Systematic Sampling

A systematic sample is formed by selecting one unit at random and then selecting additional units at evenly spaced intervals until this sample has been formed. This method is popularly used in those cases when a complete list of the population from which family is to be drawn is available. The list maybe prepared in alphabetical, geographical, numerical or some other order. The items are serially numbered. The first item is selected at random generally by following the lottery method. Subsequent items are selected by taking every k^{th} item from the list where 'k' refers to the sampling interval or sampling ratio, i.e; the ratio of population size to the size of the sample.

Let us suppose that N sampling units in the population are arranged in some systematic order and serially numbered from I to N and we want to draw a sample of size n from it such that

$$N = nk \qquad k = \frac{N}{n}$$

Where k = Sampling interval, N = Universe size and n = Sample size

Systematic sampling consists in selecting any unit at random from the first k units numbered from 1 to k and then selecting every k^{th} unit in succession subsequently. Thus, if the first unit selected at random is i^{th} unit, then the systematic sample of size n will consist of the units numbered.

$i, i + k, i + 2k, \ldots\ldots i + (n\text{-}1)k$

The random number 'i' is called random start and its value, in fact, determines the whole sample. As an example, let us suppose that we want to select 50 voters from a list of voters containing 1000 names arranged systematically. Here

$$n = 50 \text{ and } N = 1000 \qquad k = \frac{N}{n} = \frac{100}{50} = 20$$

We select any number from 1 to 20 at random and the corresponding voter in the list is selected. Suppose, the selected number is 6. Then, the systematic sample will consist of 50 voters in the list at serial numbers: 6, 26, 46, 66,…….966, 986.

While Calculating k, it is possible that we get a fractional value. In such a case, we should use approximation procedure, i.e; if the fraction is less than 0.5 it should be omitted and ifit is more than 0.5 it should be taken as 1. If it exactly 0.5, it should be omitted, if the number is even and should be taken as 1, if the number is odd. This is based on the principle that the number after approximation should preferably be even. For example, if the number of students is respectively 1020, 1150, and 1100 and we want to take a sample of 200, k will be:

(i) $k = \frac{1020}{200} = 5.1 \text{ or } 5$

(ii) $k = \frac{1150}{200} = 5.75 \text{ or } 6$

(iii) $k = \frac{1100}{200} = 5.5 \text{ or } 6$

For example, in a class, there are 96 students with roll nos. from 1 to 96. It is desired to take sample of 10 students. Use the systematic sampling method to determine the sample size.

$$k = \frac{N}{n} = \frac{96}{10} = 9.6 \text{ or } 10$$

From 1 to 96 roll nos. the first student between 1 to k, i.e, 1 to 10, will be selected at random and then we will go on taking every kth student. Suppose the first student comes out tobe 4th. The sample would then consist of the following roll nos.

4 14 24 34 44 54 64 74 84 94

Systematic sampling is relatively a simple technique and may be more efficient statistically than simple random sampling provided the lists are arranged wholly at random. However, it is rarely that is requirement is fulfilled. The nearest approach to randomness is provided by alphabetical lists such as are found in telephone directory although even these may have certain non-random characteristics.

Features of Systematic Sampling

- The systematic sampling design is simple and convenient to adopt.
- The time and work involved in sampling by this method are relatively less.
- There results obtained are also found to be generally satisfactory provided care is taken to see that there are no periodic features associated with the sampling interval.
- If populations are sufficiently large, systematic sampling can often be expected to yield result similar to those obtained by proportional stratified assembling

Limitations of Systematic Sampling

- The main limitation of the method is that it becomes less representative if we are dealing with the populations having hidden periodicities. Also, if the population is ordered in a systematic way with respect to the characteristics the investigator is interested in, then itis possible that only certain types of items will be included in the population, or at least more of certain types than others. For instance, in a study of workers' wages, the list may be such that every tenth worker on the list gets which above Rs.7500/-per month.

2. Stratified Random Sampling

Stratified random sampling is the most widely used probability sampling method. If intelligently planned, it tends to be more efficient by offering the same precision at lower cost, or higher precision at the same cost. In other words, the results based on a sample selected by using this method are more precise and reliable than those based on a simple random sample of equal size.

Stratified random sampling or simply stratified sampling is one of the random methods which, by using the available information concerning the population, attempts to design a more efficient sample than obtained by the simple random procedure.

While applying stratified random sampling method, the procedure followed is given below

- The universe to be sample is subdivided (or stratified) into groups which are mutually exclusive and include all items in the universe.
- A simple random sample is then chosen the independently from each group.

Consider the population of all the 2000 college students whose mean weight is proposed to be estimated on the basis of a sample of 50 students. For selecting a sample of this size, the population may be divided into two groups according to sex. This enables us to have two strata: boys and girls. Another criteria of dividing the population in to various strata could bethe various courses offered. Making such divisions of the population is known as stratification.

The process of stratification can be extended to undertake it at two or more stages. This depends on the problem situation and hand the degree of precision and accuracy required. For example, stratification into two strata (boys and girls) could be further extended to the nature of the course for pursued, such as arts, science and commerce, etc. The various sub groups in each strata according to the nature of course are now called sub-strata.

This sampling procedure differs from simple random sampling in which the sample items are chosen at random from the entire universe. In stratified random sampling, the sampling is designed so that a designated number of items is chosen from each stratum. In Simple random sampling, the distribution of the sample among strata is left entirely to chance.

Let's consider a situation where a research team is seeking opinions about religion amongst various age groups. Instead of collecting feedback from 326,044,985 U.S citizens, random samples of around 10000 can be selected for research. These 10000 citizens can be divided into strata according to age, i.e., groups of 18-29, 30-39, 40-49, 50-59, and 60 and above. Each stratum will have distinct members and number of members.

How to select the stratified random sample? Some of the issues involved in selecting a stratified random sample are:

1. Base of stratification:

what characteristics should be used to sub-divided the universe into different strata? As a general rule, strata are created on the basis of a variable non to be correlated with the variable of interest and for which information on each universe element is known. Strata should be constructed in a way which will minimize differences among sampling units within strata, and maximize difference among strata.

For example, if you are interested in studying the consumption pattern of the people of Thrissur, the city of Thrissur may be divided various parts (such as zones or wards) and from each part a sample may be taken at random. Before deciding on stratification, you must have knowledge of the traits of the population. Such knowledge maybe based upon expert judgement, past data, preliminary observation from pilot studies etc.

The purpose of stratification is to increase the efficiency of sampling by dividing a heterogeneous universe in such a way that (i) there is as great a homogeneity as possible within each stratum, and (ii) a marked difference is possible between the strata.

2. Number of strata:

How many strata should be constructed? The practical considerations limit the number of strata that is feasible, costs of adding more strata may soon out run benefits. As a generalization, more than six strata may be undesirable

3. Sample Size within strata:

How many of observations should be taken from each stratum? When deciding this question, we can either a proportion or a disproportional allocation. In proportional allocation, one sample each stratum in proportion to its relative weight. In disproportional sampling, this is not the case. It may be point out that proportional allocation approach is simple and if all one knows about each stratum is the number of items in that stratum, it is generally alsothe preferred procedure. In disproportional sampling, the different strata are sampled at different rates. As a general rule, when variability among observations within a stratum is high, one samples that stratum at a higher rate than for strata with less internal variation.

Proportional and Disproportional Stratified Sampling

Another important point in stratified sampling is the size of relationship of the sample with various strata and/or sub- strata, if any. As the sample selection may have a proportionate or non- proportionate relationship with population divisions into various strata and/or sub-strata, the resultant sample is a proportionate or disproportionate stratified the random sample.

1. **Proportional stratified sampling**

In a proportional stratified sampling plan, the number of items drawn from each stratum is proportional to its size. For example, if the population is divided in to five groups, their respective sizes being 10, 15, 20, 30 and 25 percent of the population and a sample of 5000 is drawn, the desired proportional sample may be obtained in the following manner:

From Stratum one	5000 (0.10)	=	500 items
From Stratum two	5000 (0.15)	=	750 items
From Stratum three	5000 (0.20)	=	1000 items
From Stratum four	5000 (0.30)	=	1500 items
From Stratum five	5000 (0.25)	=	1250 items
	Total	=	5000 items

Proportional stratification yields a sample that represents the universe with respect to the proportion in each stratum in the population. This procedure is satisfactory if there is no great difference in dispersion from stratum to stratum. But it is certainly not the most efficient procedure, especially when there is considerable variation in different strata. This indicates that in order to obtain maximum efficiency in stratification, we should assign greater representation to a stratum with a large dispersion and smaller representation to one with small variation.

Disproportional stratified sampling

In disproportional stratified sampling an equal number of cases is taken from which stratum regardless of how the stratum is represented in the universe. Thus, in the above example, an equal number of items (1000) from each stratum may be drawn. In practice, disproportional sampling is common when sampling forms a highly variable universe, wherein the variation of the measurement differs greatly from stratums to stratum.

For example, you are given the following data of the number of lectures, readers and professors in a university:

Length of Service	Lecturers	Readers	Professors	Total
Less than 5 years	2000	250	50	2300
5-10 years	3000	220	80	3300
10-15 years	1500	170	30	1700
15 years or more	880	80	40	1000
Total	7380	720	200	8300

Work out how many lecturers, readers and professors would be selected from each category if (i) we follow proportional stratified sampling method and take 10 % of the universe equivalent to the sample size, (ii) if the size of the sample is 10 % of the universe but the lecturers, readers, professors are to be in the ratio of 5:3:2 and weightage of the length of the service is to be in the ratio of 4:3:2:1

Solution, the sample size is 10 % of the universe and hence 830 persons would be selected in the sample. Since 12 strata are formed and we want to follow proportional stratified sampling method, we will take 10 % from each stratum. The number of persons selected shall be as follows:

Length of Service	Lecturers	Readers	Professors	Total
Less than 5 years	200	25	5	230
5-10 years	300	22	8	330
10-15 years	150	17	3	170
15 years or more	88	8	4	100
Total	738	72	20	830

In the second case also, the size of the sample is 830 but the lecturers, readers, professors are to be in the ratio of 5:3:2 of the sample, i.e; we take $\frac{830 \times 5}{10} = 415$ lecturers; $\frac{830 \times 3}{10} = 249$ readers, and $\frac{830 \times 2}{10} = 166$ professors.

Since the weightage to length of service is 4:3:2:1, the number selected from each category will be as given in the table below:

Length of Service	Lecturers	Readers	Professors	Total
Less than 5 years	$\frac{415 \times 4}{10} = 166.0$	$\frac{249 \times 4}{10} = 99.6$ or 100	$\frac{166 \times 4}{10} = 66.4$ or 66	332
5-10 years	$\frac{415 \times 3}{10} = 124.5$ or 124	$\frac{249 \times 3}{10} = 74.7$ or 75	$\frac{166 \times 3}{10} = 49.8$ or 50	248
10-15 years	$\frac{415 \times 2}{10} = 83.0$	$\frac{249 \times 2}{10} = 49.8$ or 50	$\frac{166 \times 2}{10} = 33.2$ or 33	166
15 years or more	$\frac{415 \times 1}{10} = 41.5$ or 42	$\frac{249 \times 1}{10} = 24.9$ or 25	$\frac{166 \times 1}{10} = 16.6$ or 17	84
Total	415	250	166	830

Features of Stratified Sampling

- **More representative:** Since the population is first divided into various strata and then a sample is drawn from each stratum, there is a little possibility of any essential group of the population being completely excluded. A more representative sample is thus secured C J Grohmann has rightly pointed out that this type of sampling balances the uncertainty of random sampling against the bias of deliberate selection

- **Greater accuracy:** Stratified sampling ensures greater accuracy. The accuracy is maximum if each stratum is so formed that it consists of uniform or homogeneous items.

- **Greater geographical concentration:** As compared with random sample, stratified samples can be more concentrated geographically, that is, the units from the different strata may be selected in such a way that all of them are localized in one geographical area. This would greatly reduce the time and expenses of interviewing.

Limitations of Stratified Sampling

- Utmost care must be exercise in dividing the population into various strata. Each strata must contain, far as possible, homogeneous items as otherwise the results may not be reliable. If proper stratification of the population is not done, the sample may have the effect of bias.

- The items from each stratum should be selected random. But this may be difficult to achieve in the absence of skilled sampling supervisors and a random selection within each stratum may not be ensured.

- Because of the likelihood that a stratified sample will be more widely distributed geographically than a simple random, sample cost per observation may be quite high.

3. **Multi Stage Sampling or Cluster Sampling**

This method is also called cluster sampling. For estimating yield of crops in India,for example, India may be considered to have a number of first stage sampling units (states and union territories), each of which has a number of second stage sampling units (districts) etc. At first, a few first stage sampling units may be selected at random from each of the first stage sampling units. From districts, certain taluks may be selected, from taluks certain villages may be selected and from villages a few plots may be selected. The yield from each of the selected plots is enquired and then the total yield is estimated.

The cluster sampling approach may be described as follows:

i. Selection of a sample of the needed size is made from one or more clusters by means of simple random sampling, and the resultant sample is called the cluster sample.

ii. As one of the two possibilities, a sample may consist of all the elementary sampling units contained in one or more randomly selected clusters. When this is done, it is known as single stage cluster sampling.

iii. Here the sample essentially consist of clusters, although the sampling units actually observed are the elementary sampling units comprising the selected cluster(s)

iv. Alternatively, selection of a sample of the desired size may be from among the elementary units contained in one or more randomly selected clusters. Sampling so effected is known as two-stage cluster sampling. In this case, the sample consists of a part of the randomly selected elementary units of the selected clusters.

v. If the sample selection passes through more than two stage of sampling, it is known as multi-stage cluster sampling

vi. The cluster sampling method involving one, two, or more stages of sampling is different from stratified random a sampling. Sampling in the case of the latter is not limited only to a specified number of randomly selected strata or groups, unlike in the case of the cluster sampling. For, a stratified sample must include elementary units from all the strata/sub- strata on proportionate or disproportionate basis, as the situation may require.

Under this method the random selection is made of primary, intermediate and final (or the ultimate) units from a given population or stratum. There are several stages in whichthe sampling process is carried out. At first, the first stage units are sampled by some suitable method, such as simple random sampling. Then, a sample of second stage units is selected from each of the selected first stage units, again by some suitable method which may be the same as ordifferent from the method employed for the first stage units. For the stages may be added as required.

Multi stage sampling introduces flexibility in the sampling method which is lacking in the other methods. It enables existing divisions and subdivisions of the population to be used as units at various stages, and permits the field work to be concentrated and yet largearea to be covered. Advantage of the method is that subdivision into second stage units (i.e., The construction of the second stage frame) need be carried out for only those first stage units which are included in the sample. It is, therefore, particularly valuable in service of under developedareas where no frame is generally sufficiently detailed and accurate for subdivision of thematerial into reasonably small sampling units.

For example, suppose in a particular survey, we wish to take a sample of 20000 students from Calicut University. We may take colleges-primary units-as the first stage, then draw departments at the second stage, and choose students at the third and last stage.

Similarly, for a company producing a popular brand of TV and intending to know the views of their brand users, sampling may be with reference to various states at the first stage, and with reference to districts at the second stage. Selecting blocks at the third stage, the sample may be selected finally with reference to those actually using the specified TV brand in the selected blocks.

When sampling at different stages is done with reference to geographical area, as in the TV case, cluster sampling is known as area sampling. It may be noticed that in selecting an area sample, sampling passes through a number of stages. The number of stages depends on geographical coverage of the population and the elementary units to be finally reached for data collection. But the selection all through the various stages remains essentially random.

Features of Multi Stage Sampling or Cluster Sampling

- Multistage sampling introduces flexibility in the sampling method which is lacking in the other methods. It enables existing divisions and subdivisions of the population to be used as units and various stages, and permits the field work to be concentrated and yet large area to be covered.

- Another feature of the method is that subdivision into second stage units, that is, the construction of the second stage frame need be carried out for only those first stage units which are included in the sample. It is, therefore, particularly valuable in surveys of underdeveloped areas where no frame is generally sufficiently detailed and accurate for sub division of the material into reasonably small sampling units.

Limitations of Multi Stage Sampling or Cluster Sampling

- However, a multi stage sample is in general less accurate than a sample containing thesame number of final stage units which have been selected by some single stage process.

Recap

- Probability sampling methods are those in which every item in the universe has a known chance, or probability, of being chosen for the sample
- Non-probability sampling methods are those which do not provide every item in the universe with they known chance of being included in the sample
- Simple random sampling refers to that sampling technique in which each and every unit of the population has an equal opportunity of being selected in the sample
- Under lottery method, all items of the universe are numbered or named on separate slips of paper of identical size and shape
- While adopting the lottery method, it is absolutely essential to see that the slips are of identical size, shape and colour, otherwise there is a lot of possibility of personal prejudice and bias affecting the results
- A systematic sample is formed by selecting one unit at random and then selecting additional units at evenly spaced intervals until this sample has been formed
- Stratified random sampling or simply stratified sampling is one of the random methods which, by using the available information concerning the population, attempts to design amore efficient sample than obtained by the simple random procedure.
- In a proportional stratified sampling plan, the number of items drawn from each stratum is proportional to its size
- Proportional stratification yields a sample that represents the universe with respect to the proportion in each stratum in the population
- In disproportional stratified sampling an equal number of cases is taken from which stratum regardless of how the stratum is represented in the universe
- Multi stage sampling is also called cluster sampling
- Selection of a sample of the needed size is made from one or more clusters by means of simple random sampling, and the resultant sample is called the cluster sample

- As one of the two possibilities, a sample may consist of all the elementary sampling units contained in one or more randomly selected clusters. When this is done, it is known as single stage cluster sampling

- The selection of a sample of the desired size may be from among the elementary units contained in one or more randomly selected clusters. Sampling so effected is known as two-stage cluster sampling

- If the sample selection passes through more than two stage of sampling, it is known as multi-stage cluster sampling

Objective Questions

1. What is a stratified random sample?
2. What is a random sample? Describe two methods of simple random sampling?
3. Describe any two methods of selecting a sample.
4. What are the probability sampling methods?
5. Explain the restricted random sampling methods.
6. Explain the difference between systematic sampling and stratified sampling.
7. How to select the stratified random sample?
8. Difference between proportional stratified sampling and disproportional stratifiedsampling.
9. Define Cluster sampling.
10. What are the main features of Simple Random Sampling?

Answers

1. Stratified random sampling or simply stratified sampling is one of the random methods which, by using the available information concerning the population, attempts to design amore efficient sample than obtained by the simple random procedure

2. Random sampling is a part of the sampling technique in which each sample has an equal probability of being chosen. A sample chosen randomly is meant to be an unbiased representation of the total population. 1. Simple or Unrestricted Random Sampling and 2.Restricted Random Sampling

3. Non -Random Sampling Methods and Random Sampling Methods

4. Simple or Unrestricted Random Sampling, Systematic Sampling, Stratified Sampling, Multi-Stage or Cluster Sampling

5. Systematic Sampling, Stratified Sampling, Multi-Stage or Cluster Sampling

6. A systematic sample is formed by selecting one unit at random and then selecting additional units at evenly spaced intervals until this sample has been formed and stratified random sampling or simply stratified sampling is one of the random methods which, by using the available information concerning the population, attempts to design a more efficient sample than obtained by the simple random procedure.

7. Base of stratification, Number of strata, Sample Size within strata

8. In a proportional stratified sampling plan, the number of items drawn from each stratum is proportional to its size and in disproportional stratified sampling an equal number of cases is taken from which stratum regardless of how the stratum is represented in the universe

9. Under this method the random selection is made of primary, intermediate and final (or the ultimate) units from a given population or stratum. There are several stages in which the sampling process is carried out

10. Since the selection of items in the sample depends entirely on chance, there is no possibility of personal bias affecting the results. As compared to judgement sampling a random sample represents the universe in a better way. As the size of the sample increases, it becomes increasingly representative of the population. The analyst can easily assess the accuracy of this estimate because sampling errors follow the principal of chance. The theory of random sampling is further developed than that of any other type of sampling which enables the analyst to provide the most reliable information at the leastcost.

References

1. Gupta, S. P.(2001). "*Statistical methods, sultan chand & sons*." New Delhi
2. Monga, Gopal Sohanlal (2001). *Mathematics and statistics for economics*. VikasPublishing House
3. Gupta, S. C., and V. K. Kapoor (2020). *Fundamentals of mathematical statistics*. SultanChand & Sons
4. Shenoy G V, Madan Pant (1994). *Statistical Methods in Business and Social Sciences*, Macmillan India Limited
5. Seymour Lipschutz and John Schiller, *Introduction to probability and Statistics*, Schaum's Outlines, McGraw Hill Education (India) Private Limited

Suggested Readings

1. Hooda R P (2008), *Statistics for Business and Economics*, Macmillian India Ltd
2. Taro Yamane, *Statistics; An introductory Analysis*, Harper International Edition
3. Agarwal, Basant Lal (2006). *Basic statistics*. New Age International

2 UNIT

Non-Probability Sampling Methods

Learning Outcomes

After reading this unit, the learner will be able to:

- apply non-probability sampling techniques in the real world research

Prerequisites

Basic knowledge in Sampling theory

Keywords

Non-Probability Sampling Methods, Convenience Sampling, Quota Sampling

5.2.1 Non-Random or Non-Probability Sampling Method

Non-probability sampling involves choosing a sample of individual units relying on personal judgment instead of employing randomisation methods. In this method, the researcher's subjective judgment plays a crucial role in determining the selection, and the resulting sample may lack representation of the entire population. Purposive or judgmental sampling is a frequently employed technique within non-probability sampling.

In economics, consider studying the impact of a specific government policy on businesses. A researcher, relying on personal judgment, may purposively select businesses known to have been directly affected by the policy. This may entail selecting businesses within a specific industry, geographic area, or size range guided by the researcher's expertise and understanding. While this approach allows for a targeted investigation of businesses most relevant to the research question, it lacks the randomisation characteristic of probability sampling.

Following are the features of Non- Probability Sampling

1. No Random Selection: Non-probability sampling does not choose participants randomly. Instead, the researcher picks them based on convenience or specific criteria. Not everyone in the population has an equal chance of being selected.

2. Subjectivity in Selection: The researcher's personal judgment plays a big role in choosing the sample. This can lead to biased results. For example, choosing participants who are easily available or seem to know a lot about the topic.

3. Focused on Specific Groups: Non-probability sampling often targets a specific subgroup of the population that fits certain criteria or is of interest to the researcher. For example, only selecting senior managers for a study on leadership.

4. Cost-Effective and Time-Saving: Non-probability sampling is faster and cheaper than probability sampling methods. For example, giving surveys to people nearby instead of randomly selecting from the whole population.

5. Useful for Exploratory Research: Non-probability sampling is great for pilot studies or when you want to explore ideas, create hypotheses, or study hard-to-reach groups. For example, studying marginalized communities using snowball sampling.

6. Limited Generalisability: Because the sample is not representative of the whole population, the results cannot be applied to everyone. The findings are specific to the sample and might not be true for others.

7. Higher Risk of Bias: Non-probability sampling has a higher risk of bias because the sample is chosen non-randomly. For example, using convenience sampling might over represent one group and exclude others.

8. Reliant on Researcher Expertise: The quality of the sample depends on the researcher's knowledge and ability to choose participants well. For example, purposive sampling requires the researcher to identify the most relevant participants.

9. Flexible Approach: Non-probability sampling allows researchers to adjust their methods based on the study's context and available resources. For example, changing the selection process as more participants become available through snowball sampling.

10. Suitable for Small Sample Sizes: Non-probability sampling is often used when studying a small population or when the population size is unknown. For example, conducting in-depth interviews with a small group of experts in a niche field.

Non-Random sampling is of four types:

i. Convenience Sampling

ii. Quota Sequential Sampling

iii. Judgment Sampling

iv. Snowball Sampling

1. Convenience Sampling

Convenience sampling involves selecting sample units based on their easy availability and accessibility, rather than through a systematic or random approach. It is frequently employed due to its practicality. In convenience sampling, the selected sample is not necessarily representative of the entire population, as certain segments may be overrepresented or underrepresented.

In an economic context, consider a study examining consumer preferences for a new product in a shopping mall. A researcher conducting convenience sampling might approach individuals who happen to be present at the mall during a specific time, making them readily available for participation in the study. Shoppers present at a particular time may not accurately represent the broader population of potential consumers, leading to biased conclusions about overall preferences.

Let us discuss the features of Convenience Sampling

1. Easy to Use: This method is quick and requires minimal effort and resources.
2. Not Random: Participants are chosen because they are easy to access, not because they are randomly selected.
3. Low Cost: This method is great for preliminary studies or when budgets are limited.
4. May Not Represent the Population: Because the sample is not random, the results may not accurately represent the population.
5. Useful for Pilot Studies: This method is often used for pilot studies or exploratory research.

2. Quota Sampling

Quota sampling involves setting specific quotas based on predetermined parameters of the population, and each investigator or interviewer is assigned a certain number of units to investigate or interview. This method is often used for its practicality and cost-effectiveness, especially in studies requiring personal interviews. However, quota sampling has potential drawbacks, such as the risk of biased estimates and the influence of the investigator's judgment on unit selection.

Consider a survey to understand consumer spending patterns in a specific city has to be conducted. Quotas may be set based on demographics such as age, income, and occupation to ensure a representative sample. Interviewers are then assigned quotas for each category, and they have the flexibility to select individuals who meet these criteria. While this approach allows for efficient data collection and cost savings, it introduces the possibility of bias.

Following are the features of Quota Sampling

1. Representing Subgroups: This method ensures each subgroup in the population is represented.
2. Non-Random Selection: Participants within each quota are selected based on convenience or purposive methods.
3. Flexible: This method allows for customization of quotas based on research objectives.
4. Potential for Bias: Non-random selection can affect representativeness.
5. Useful for Market Research: This method is often used in market research and public opinion polls.

3. Judgment Sampling

Judgment sampling involves selecting a sample for a specific purpose, and the choice of sampling units depends entirely on the investigator's judgment. This method is typically employed when random procedures for sample selection are not feasible or practical. For instance, in situations where the researcher aims to study a specific and relatively rare condition, such as patients suffering from a rare disease like blood cancer, purposive sampling may be employed.

Consider a study focusing on the financial management practices of successful entrepreneurs. If the researcher aims to interview individuals who have demonstrated exceptional financial success, a purposive sampling approach might involve selecting business leaders or entrepreneurs based on their reputation, financial achievements, or specific criteria related to financial management skills. In such cases, the investigator's expertise and judgment are crucial in identifying and selecting participants who align with the study's objectives.

Features of Judgmental or Purposive Sampling

1. Selecting the Right People: Participants are chosen based on specific criteria related to the study.
2. Targeted Approach: This method ensures that only knowledgeable or experienced individuals are included.
3. Subjective: The researcher's judgment is used to select participants, which can introduce bias.
4. Small Sample Sizes: This method is often used for in-depth studies with a limited number of participants.
5. Useful for Qualitative Research: This method is commonly used in qualitative research, expert interviews, and case studies.

4. Snowball Sampling

Snowball sampling is a non-probability sampling technique where existing study participants recruit future participants from among their acquaintances or network. This method is often employed when it's challenging to identify and access members of a specific population directly. The process involves starting with a small group of individuals who meet certain criteria and then expanding the sample by asking those individuals to refer others who fit the criteria. This chain referral process continues, resembling the way a snowball grows larger as it rolls downhill.

In an economic research context, snowball sampling might be utilized in studying informal or underground economic activities that are difficult to identify through traditional means. For instance, researchers interested in understanding the dynamics of informal economies, such as unregistered small businesses or off-the-books employment, may face challenges in accessing relevant participants directly. By initiating the study with a small group of individuals engaged in such activities, the researchers can then leverage these participants to refer others within their networks who may be willing to participate.

Features of Snowball Sampling

1. Using Networks: Existing participants recruit future participants from their acquaintances or networks.
2. Reaching Hidden Populations: This method is effective for studying hard-to-reach or sensitive groups.
3. Chain Reaction: The sample grows as participants refer others, potentially leading to exponential growth.
4. Risk of Bias: This method may lead to homogeneity as referrals often come from similar backgrounds or networks.

Recap

- ♦ Non-Probability Sampling Methods: Non-probability sampling methods, including convenience, quota, judgment, and snowball sampling. They are characterized by the absence of randomization, relying on subjective judgment or practical considerations in selecting samples.
- ♦ Non-probability sampling is cost-effective, time-saving, and flexible but has limited generalisability.
- ♦ Convenience Sampling: Convenience sampling involves selecting sample units based on easy availability and accessibility, lacking the rigor of systematic or random methods.

- Convenience sampling is easy, low-cost, and suitable for pilot studies but may not reflect the population.

- Quota Sampling: Quota sampling sets specific quotas based on predetermined parameters, and investigators are assigned to investigate or interview a certain number of units, making it a practical yet potentially biased method.

- Quota sampling ensures subgroup representation but introduces bias due to non-random participant selection.

- Judgment Sampling: Judgment sampling relies on the researcher's subjective judgment to select sample units, often applied when random procedures are impractical, allowing for targeted investigations based on expertise.

- Judgment sampling is effective for small sample sizes and qualitative research but heavily depends on researcher bias.

- Snowball Sampling: Snowball sampling is a technique where existing participants recruit future participants from their network, suitable for studying hard-to-reach populations or phenomena with limited identification.

- Snowball sampling grows through participant networks, suitable for studying sensitive groups but risks sampling bias.

Objective Questions

1. What is the most important feature of convenience sampling?
2. What is the purpose of quota sampling?
3. What is the application of snowball sampling?

Answers

1. Easy to Use
2. To ensure that the sample is representative
3. Used to study hard-to-reach or hidden populations

Assignments

1. Discuss in detail the various methods in which the non-probability sampling can be conducted in the real world.

References

1. Khalid, M. M., Javaid, S. & Shoeb, Q. (2012). *Statistical Methods in Economics*. Global Research Publications, New Delhi.
2. Levin, R. I. & Rubin, D. S. (2003). *Statistics for Management*, Seventh Edition. Prentice Hall of India Pvt. Ltd., New Delhi.
3. Spiegel, M. R., Schiller, J. J. & Srinivasan, R. A. (2010). *Probability and Statistics*, Third Edition, Schaum's Outlines. McGraw Hill Education (India) Private Limited, New Delhi.
4. Goon, A., Gupta, M. & Dasgupta, B. (1963). *Fundamentals of Statistics*. The World Press Private Limited, Calcutta.

Suggested Readings

1. Hooda R P (2008), *Statistics for Business and Economics*, Macmillian India Ltd
2. Taro Yamane, *Statistics; An introductory Analysis*, Harper International Edition
3. Agarwal, Basant Lal (2006). *Basic statistics*. New Age International

6

BLOCK

Probability

1

UNIT

Basic Concepts of Probability

Learning Outcomes

After reading this unit, the learner will be able to:

- understand the concept of probability
- compare and contrast the classical, empirical, and axiomatic approaches to probability
- calculate and interpret conditional probabilities in various scenarios

Prerequisites

A basic understanding of set theory, permutations and combinations helps in calculating probabilities, especially in the classical approach, which assumes all outcomes are equally likely. For the empirical approach, familiarity with data collection and interpreting frequencies is essential. The axiomatic approach, rooted in logical reasoning, requires an understanding of mathematical definitions and rules. Conditional probability demands knowledge of ratios, independence, and the concept of updating probabilities based on new information.

Keywords

Probability, Classical Approach, Empirical Approach, Axiomatic Approach, Conditional Probability

6.1.1 The Concept of Probability

Probability is a fundamental concept in statistics. It provides a quantitative measure of uncertainty associated with events in a random experiment. We assign a probability value between 0 and 1 to express this uncertainty. A probability of 1 indicates certainty that the event will occur. A probability of 0 signifies certainty that the event will not occur. This numerical assignment allows us to quantify and communicate our expectations regarding the likelihood of various outcomes.

The probability scale ranges from 0 to 1. It represents the entire spectrum of uncertainty. A probability of 1, or 100%, indicates complete certainty that the event will occur. Conversely, a probability of 0 implies absolute certainty that the event will not occur. For example, if the probability is 1/4, we interpret it as a 25% chance of occurrence and a 75% chance of non-occurrence. This numerical representation facilitates a clear understanding of confidence or uncertainty associated with a particular event.

Understanding probability and expressing it numerically has practical implications in various fields, including risk assessment, decision-making, etc. Whether assessing the likelihood of success in a business venture, predicting outcomes in games of chance, or making informed decisions based on uncertain information, the concept of probability and its numerical representation serve as a valuable tool for quantifying and managing uncertainty in diverse scenarios.

There are different approaches through which we can estimate the probability of an event.

Classical Approach

The Classical Approach to probability assumes equally likely outcomes. If an event can occur in h different ways out of a total number of n possible ways, all of which are equally likely, then the probability of the event is h/n.

For example, consider a fair six-sided die. If we want to find the probability of rolling a 4, there is only one way to roll a 4, with six possible outcomes. So, according to the Classical Approach, the probability of rolling a 4 is 1/6.

6.1.2 Frequency Approach

The Frequency Approach, or the Empirical Approach, relies on observed frequencies in real-world experiments.

If after n repetitions of an experiment, where n is very large, an event is observed to occur in h of these, then the probability of the event is h/n. This is also called the empirical probability of the event.

Continuing with the die example, we might roll the die 600 times (n) in a Frequency

Approach and observe that a 4 comes up 100 times. The estimated probability of rolling a 4 is then 100/600 =1/6, coincidentally the same as the Classical probability. The larger the number of repetitions (n), the closer the observed frequency-based probability tends to get to the theoretical probability from the Classical Approach.

6.1.3 The Axiomatic Approach

The axiomatic approach of probability contains three axioms

1. Axiom of Non-negativity

The first axiom of probability, known as the axiom of non-negativity, states that for any event A in the sample space S, the probability assigned to A, denoted as P(A), must be a non-negative real number. In mathematical terms, this means that $P(A) \geq 0$ for all events A. This axiom ensures that probabilities are always non-negative, reflecting the intuitive notion that the likelihood of any event occurring cannot be negative. Whether it is the probability of rolling a specific number on a fair die or the chance of rain on a given day, the probability value must be non-negative.

2. Axiom of Normalization

The second axiom, often referred to as the axiom of normalisation or the sum rule, asserts that the probability of the entire sample space S equals 1. Symbolically, this is expressed as $P(S) = 1$. In other words, the sum of probabilities for all possible outcomes or events in the sample space is unity. This axiom ensures that the total probability mass in the sample space is accounted for, aligning with the idea that, under any circumstance, something in the sample space must occur.

3. Axiom of Additivity

The third axiom, the axiom of additivity, addresses the probabilities of combinations of events. It states that for any sequence of mutually exclusive events $\{A_1, A_2, ..., An\}$, the probability of the union of these events is equal to the sum of their probabilities. Mathematically, this is expressed as $P(A_1 \cup A_2 \cup ... \cup An) = P(A_1) + P(A_2) + ... + P(An)$. The additivity axiom handles scenarios where events are not mutually exclusive by adjusting for their potential overlap. It ensures a coherent and consistent way to calculate probabilities for complex situations involving multiple events.

Some Important Theorems on Probability

From the above axioms, we can now prove various theorems on probability that are important in further work

Theorem 1

If $A_1 \subset A_2$, then $P(A_1) \leqq P(A_2)$ and $P(A_2 - A_1) = P(A_2) - P(A_1)$.

If $A_1 \subseteq A_2$, then $P(A_1) < P(A_2)$, which intuitively means that the probability of all set is less than or equal to the probability of a larger set. Additionally,

$P(A_2 - A_1) = P(A_2) - P(A_1)$, emphasizing the relationship between the probabilities of the sets A_1 and A_2. This theorem is foundational in comparing the likelihoods of different events.

Theorem 2: Probability Bounds

For every event A,

$$0 \leqq P(A) \leqq 1,$$

i.e., a probability is between 0 and 1.

For any event A, the probability $P(A)$ lies between 0 and 1. This theorem reflects the inherent constraints on probabilities. The probability cannot be negative $(P(A) \geq 0)$, and it cannot exceed certainty $(P(A) \leq 1)$. Probability serves as a measure of the likelihood of occurrence, always bounded within this range.

Theorem 3: Null Event Probability

$$P(\emptyset) = 0$$

i.e., the impossible event has a probability of zero.

The probability of the null event $\emptyset$ (impossible event) is zero $(P(\emptyset) = 0)$. This makes intuitive sense as an event that cannot happen should have a probability of zero.

Theorem 4: Complement Probability

If A' is the complement of A, then

$$P(A') = 1 - P(A)$$

For the complement of an event A, denoted as A', the probability $P(A')$ is equal to $1 - P(A)$. This theorem establishes a relationship between the probability of an event and its complement, emphasising that the total probability in the sample space is accounted for.

Theorem 5: Additivity of Mutually Exclusive Events

If $A = A_1 \cup A_2 \cup \dots \cup A_n$, where $A_1, A_2, \dots, A_n$ are mutually exclusive events, then

$$P(A) = P(A_1) + P(A_2) + \cdots + P(A_n)$$

In particular, if $A = S$, the sample space, then

$$P(A_1) + P(A_2) + \cdots + P(A_n) = 1$$

For any event A that can be expressed as the union of mutually exclusive events $A_1, A_2, \dots, A_n, P(A) = P(A_1) + P(A_2) + \dots + P(A_n)$. This theorem provides a formula for calculating the probability of the union of mutually exclusive events, allowing for the decomposition of complex events into simpler components.

Theorem 6: Inclusion-Exclusion Principle for Two Events

If A and B are any two events, then

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

More generally, if A_1, A_2, A_3 are any three events, then

$$P(A_1 \cup A_2 \cup A_3) = P(A_1) + P(A_2) + P(A_3)$$

$$-P(A_1 \cap A_2) - P(A_2 \cap A_3) - P(A_3 \cap A_1) + P(A_1 \cap A_2 \cap A_3)$$

Generalizations to n events can also be made.

For any two events A and B, $P(A \cup B) = P(A) + P(B) - P(A \cap B)$. This theorem extends the inclusion-exclusion principle to the union of two events, accounting for the overlap between them.

Theorem 7: Probability of an Event in terms of Intersection

For any events A and B,

$$P(A) = P(A \cap B) + P(A \cap B')$$

For any events A and B, $P(A) = P(A \cap B) + P(A \cap B')$. This theorem expresses the probability of event A in terms of the probabilities of its intersections with B and B', where B' is the complement of B.

Theorem 8: Total Probability for Exhaustive Events

If an event A must result in the occurrence of one of the mutually exclusive events $A_1, A_2, \ldots, A_n$, then

$$P(A) = P(A \cap A_1) + P(A \cap A_2) + \cdots + P(A \cap A_n)$$

If an event A must result in the occurrence of one of the mutually exclusive events $A_1, A_2, \ldots, A_n$, then $P(A) = P(A \cap A_1) + P(A \cap A_2) + \ldots + P(A \cap A_n)$.

This theorem establishes a way to calculate the probability of event A when it necessitates the occurrence of one of a set of mutually exclusive events.

6.1.4 Conditional Probability

In the context of conditional probability, let A and B be two events where $P(A) > 0$ and denote by $P(B|A)$ the probability of event B given that event A has occurred. Conditional probability deals with the likelihood of an event occurring, considering that another event has already taken place.

In this scenario, when event A is known to have occurred, it essentially becomes the new sample space, replacing the original sample space *S*. This adjustment is made to reflect the updated information that *A* has taken place. The conditional probability $P(B|A)$ is the probability of *B* occurring within this new context where *A* is considered certain.

The formula for conditional probability is given by:

$$P(B|A) = \frac{P(A \cap B)}{P(A)}$$

Here, $P(A \cap B)$ represents the probability of both events A and B occurring together. The denominator $P(A)$ accounts for the fact that the new sample space is A. The conditional probability $P(B|A)$ provides a measure of the likelihood of event B occurring given the information that event A has already taken place.

Illustration : 6.1.1

A bag contain 3 red and 4 white balls. Two draws are made without replacement. What is the probability that both the balls are red?

Solution:

P(drawing a red ball) = $\frac{3}{7}$

$i.e. P(A) = \frac{3}{7}$

P(drawing a red ball in the second draw given that first ball drawn is red) = $\frac{2}{6}$

$$P\left(\frac{B}{A}\right) = \frac{2}{6}$$

$$\therefore P(AB) = P(A) \times P(\frac{B}{A})$$

$$= \frac{3}{7} \times \frac{2}{6} = \frac{1}{7}$$

The probability that both balls are red is: $\frac{1}{7}$

Illustration : 6.1.2

Find the probability of drawing a queen and a king from a pack of cards in two consecutive draws, the cards drawn not being replaced.

Solution

P(drawing a queen card) = $\frac{4}{52}$

$i.e. P(A) = \frac{4}{52}$

P(drawing a King after a queen has been drawn) = $\frac{4}{51}$

$$P\left(\frac{B}{A}\right) = \frac{4}{51}$$

$$\therefore P(AB) = P(A) \times P(\frac{B}{A})$$

$$= \frac{4}{52} \times \frac{4}{51} = \frac{4}{663}$$

The probability of drawing a queen and then a king in two consecutive draws (without replacement) is: $\frac{4}{663}$

Illustration : 6.1.3

A bag contain 5 white and 3 black balls. Two balls are drawn at random one after the other without replacement. What is the probability that both the balls are black?

Solution

P(drawing a black ball in the first draw) = $\frac{3}{8}$

$$i.e. P(A) = \frac{3}{8}$$

P(drawing a second black ball given that the first ball drawn is black) = $\frac{2}{7}$

$$P\left(\frac{B}{A}\right) = \frac{2}{7}$$

$$\therefore P(AB) = P(A) \times P(\frac{B}{A})$$

$$= \frac{3}{8} \times \frac{2}{7} = \frac{3}{28}$$

The probability that both balls drawn are black is: $\frac{3}{28}$

Recap

- Probability is a value between 0 and 1
- In the Classical Approach to probability, outcomes are assumed to be equally likely, and the probability of an event is calculated as the ratio of favorable outcomes to the total possible outcomes
- The Empirical Approach to probability involves estimating the likelihood of an event based on observed frequencies from real-world experiments or data
- The Axiomatic Approach formulates probability theory using a set of axioms or fundamental principles
- Conditional Probability is the likelihood of one event occurring given that another event has already occurred, expressed as the probability of A given B, denoted as P(A|B)

Objective Questions

1. Let A and B be two events such that the occurrence of A implies occurrence of B, but not vice-versa, then the correct relation between P(A) and P(B) is?
2. What would be the probability of an event 'G' if H denotes its complement, according to the axioms of probability?
3. If the probability of an event is 1/3, how would you interpret it?
4. What does a probability of 0 signify?
5. What does the probability of the complement of an event A' establish?
6. What does the formula $P(B\,|\,A) = P(A \cap B) / P(A)$ represent in the context of conditional probability?

Answers

1. P (G) = 1 – P (H)
2. 33% chance of occurrence and 67% chance of non-occurrence
3. Certainty that the event will not occur
4. P(A') = 1 - P(A)
5. Probability of both events A and B occurring together
6. Probability of event A

Assignments

1. Write short note on the axioms of probability with suitable examples.

References

1. Spiegel, M. R., Schiller, J. J. & Srinivasan R. A. (2010). *Probability and Statistics*, Third Edition (Schaum's Outlines). McGraw Hill Education (India) Private Limited, New Delhi.
2. Stephens, L. J. (2009). *Schaum's Outline of Beginning Statistics*, Second Edition. McGraw Hill Education (India) Private Limited, New Delhi.
3. Spiegel, M. R., Schiller, J. J. & Srinivasan R. A. (2020). *Schaum's Easy Outline of Probability and Statistics (SCHAUM'S Easy Outlines)*. McGraw Hill Education (India) Private Limited, New Delhi.

Suggested Readings

1. Gupta, S. C., & Kapoor, V. K. (2020). *Fundamentals of Mathematical Statistics*. Sultan Chand & Sons.

2. Goon, A. M., Gupta, M. K., & Dasgupta, B. (2017). *An Outline of Statistical Theory: Volume I*. World Press.

3. Agarwal, B. L. (2014). *Basic Statistics*. New Age International Publishers.

4. Pandey, H. D. (2018). *Statistics for Economics and Business*. Himalaya Publishing House.

5. Sharma, J. K. (2020). *Business Statistics*. Vikas Publishing House.

2

UNIT

Probability Distributions and Distribution Functions

Learning Outcomes

After completing this unit, learners will be able to:

- understand probability distributions
- analyse distribution functions
- explore binomial and normal distributions

Prerequisites

To study probability distributions, distribution functions, and specific distributions like binomial and normal, one must understand the basics of probability theory, including concepts like random variables, outcomes, and events. Familiarity with descriptive statistics is crucial for interpreting distribution parameters. Knowledge of combinatorics (e.g., combinations) aids in understanding the binomial distribution, while a basic knowledge of calculus, particularly integration, is essential for continuous distributions like the normal distribution. Additionally, a foundation in basic algebra and graphing functions is helpful for visualising and analysing distribution curves and their behaviour.

Keywords

Probability Distributions, Distribution Functions, Binomial and Normal Distributions

6.2.1 Random Variables

Consider an experiment of throwing a coin twice. The outcomes {HH, TT, HT, TH} constitute the sample space. Let X be the number of heads in each throw. Then X can take the values 0, 1, and 2.

Event	HH	HT	TH	TT
Value	2	1	1	0

X is a variable defined over the sample space of a random experiment called random variable.

There are two types of random variables, discrete random variable and continuous random variables.

6.2.1.1 Discrete and continuous Random Variables

Discrete Probability Distributions (Probability Mass function)

For a discrete random variable X, the set of possible values it can assume is denoted by x_1, x_2, x_3, ... arranged in some order. The probabilities associated with each of these values are specified by the Probability Mass Function (PMF), denoted by $f(x_k)$, where x_k represents a particular value that X can take. The PMF describes the likelihood of X assuming each of its possible values.

The probability mass function is formally expressed as:

$$P(X = x_k) = f(x_k) \text{ for } k = 1, 2, \ldots \qquad (1)$$

Here, $P(X = x_k)$ represents the probability that the discrete random variable X takes on the specific value x_k, and $f(x_k)$ is the associated probability mass function for that value.

To streamline notation, it is convenient to introduce the probability function or probability distribution, denoted by $P(X = x)$, where x is a generic variable representing any of the possible values that X can assume. This probability function is defined by:

$$P(X = x) = f(x) \qquad (2)$$

In summary, the probability mass function $f(x)$ specifies the probabilities associated with individual values of a discrete random variable, and the probability distribution $P(X = x)$ generalises this to express the probability of X being any specific value x.

For $x = x_k$, this reduces to (1) while for other values of x, $f(x) = 0$.

In general, $f(x)$ is a probability function if

$f(x) \geq 0$

$\Sigma f(x) = 1$

Continuous Probability Function

A function $f(x)$ that satisfies the above requirements is called a probability function or probability distribution for a continuous random variable, but it is more often called a probability density function or simply density function. Any function $f(x)$ satisfying Properties 1 and 2 above will automatically be a density function.

A random variable X is said to be absolutely continuous, or simply continuous, if the Probability density function is

$f_i(x_i) \geq 0$

$\int_{-\infty}^{\infty} f_i(x_i)\, dx_i = 1$

The cumulative distribution function (CDF), denoted by $F(x)$, is a fundamental concept in probability theory and statistics. It provides a comprehensive representation of the probability distribution of a random variable X. The CDF is defined for any real number x and is expressed as $F(x) = P(X \leq x)$, where P denotes the probability.

The properties of the distribution function $F(x)$ are crucial in understanding its behavior and significance in probability theory:

1. **Non decreasing Function**: The distribution function $F(x)$ is non decreasing, meaning that if $x_1 \leq x_2$, then $F(x_1) \leq F(x_2)$. This property ensures that as we move to larger values of x, the cumulative probability never decreases. It aligns with the intuitive notion that the probability of observing a value less than or equal to x should not decrease as x increases.

2. **Limits at Infinity**: The limits of $F(x)$ as x approaches negative infinity and positive infinity have specific values. As x approaches negative infinity, $F(x)$ approaches 0, indicating that the cumulative probability of values less than or equal to negative infinity is zero. As x approaches positive infinity, $F(x)$ approaches 1, signifying that the cumulative probability of values less than or equal to positive infinity is one.

3. **Right-Continuity**: The distribution function $F(x)$ is continuous from the right. This means that the limit of $F(x + h)$ as h approaches 0 from the right is equal to $F(x)$ for all x. In other words, the probability of observing a value less than or equal to x does not change abruptly as x changes slightly. This property ensures the smoothness of the cumulative distribution function.

These properties collectively contribute to the interpretation and application of the cumulative distribution function. The nondecreasing nature ensures coherence with the probability structure, while the limits at infinity establish the overall coverage of the probability space. The right-continuity ensures that small changes in x result in small changes in the cumulative probability, emphasising the gradual accumulation of probabilities as x increases.

The distribution function $F(x)$ is defined as

$F(x) = P(X \leq x) = \sum_0^x f(x)$ (for discrete random variable)

$F(x) = \int_{-\infty}^{x} f(x) \ dx$ (for continuous random variable)

X is a continuous random variable, then the probability that X takes on any one particular value is zero, whereas the interval probability that X lies between two different values, say, a and b, is given by

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

Illustration : 6.2.1

Given the following probability distribution

x	0	1	2	3	4	5	6	7
$f(x)$	0	C	2c	2c	3c	c^2	$2c^2$	$7c^2 + c$

Find 1) c 2) $P[X \geq 5]$ c)

Solution

1. $c + 2c + 2c + 3c + c^2 + 2c^2 + 7c^2 + c = 1$

$10c^2 + 9c - 1 = 0$

$10c^2 + 10c - c - 1 = 0$

$10c(c+1) - (c+1) = 0$

$(10c - 1)(c+1) = 0$

$c = -1, \ c = \frac{1}{10}$

Since $P(x) \geq 0, \ c = \frac{1}{10}$

2. $P[X \geq 5] = P[X = 5,6,7]$

$= c^2 + 2c^2 + 7c^2 + c$

$= 10 \ c^2 + c$

$$=\frac{10}{100}+\frac{1}{10}=\frac{2}{10}=\frac{1}{5}$$

x	0	1	2	3	4	5	6	7
$f(x)$	0	c	2c	2c	3c	c^2	$2c^2$	$7c^2+c$
$F(x)$	0	c	3c	5c	8c	$8c+c^2$	$8c+c^2+2c^2$	$8c+c^2+2c^2+7c^2+c$
	0	$\frac{1}{10}$	$\frac{3}{10}$	$\frac{5}{10}$	$\frac{8}{10}$	$\frac{81}{100}$	$\frac{83}{100}$	1

Illustration : 6.2.2

Find the value of k if $f(x)=k(2-x)\ \ 0<x<2$ is a Probability density function.

Solution

Since $f(x)$ is a continuous probability density function $\int_0^2 f(x)\ dx=1$

$$\int_0^2 k(2-x)\ dx=1$$

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

$$k\left(4-\frac{4}{2}\right)=1$$

$$k(4-2)=1$$

$$k=\frac{1}{2}$$

Illustration : 6.2.3

Show that $f(x)=\frac{x+1}{2}\quad |x|<1$

$=0$ elsewhere represents the Probability density function of a random variable X.

Solution

If $f(x)$ is a Probability density function then $\int f(x)\ dx=1$

Consider $\displaystyle\int_{-1}^{1}\frac{x+1}{2}=\frac{1}{2}\int_{-1}^{1}x+1\quad dx$

$$= \frac{1}{2}\left(\frac{x^2}{2} + x\right)_{-1}^{1}$$

$$= \frac{1}{2}\left(\frac{1}{2} + 1 - (\frac{1}{2} - 1)\right)_{-1}^{1}$$

$$= \frac{1}{2} \times 2$$

$$= 1$$

Random variables can take on different forms based on the nature of the outcomes they represent. A discrete random variable is one that can assume a finite or countably infinite set of distinct values. For instance, the number of students in a classroom or the count of defective items in a production batch are examples of discrete random variables. On the other hand, a nondiscrete random variable takes on a noncountably infinite number of values and often corresponds to continuous phenomena, such as the height of individuals or the temperature in a given location.

It is essential to note that multiple random variables can be defined on the same sample space, each capturing different aspects or characteristics of the underlying experiment. For instance, in a coin-tossing experiment, besides the random variable representing the number of heads, one could define another random variable representing the square of the number of heads or the difference between the number of heads and tails. These additional variables provide diverse perspectives for analysis.

6.2.3 Binomial Distribution

Let us assume a salesperson is attempting to close deals with potential clients. Each interaction with a client can be viewed as a trial, and the outcome of each trial is either a successful deal (success) or an unsuccessful attempt (failure). Let us denote the probability of successfully closing a deal in any single trial as p, and the probability of failure as $q = 1 - p$.

Now, suppose the salesperson conducts a series of n trials, trying to close deals with different clients independently. The objective is to understand the probability distribution of the number of successful deals (x) among these n trials. This is where the probability function for a binomial distribution comes into play.

The probability mass function (PMF) for a binomial distribution is given by:

$$P(X = x) = \binom{n}{x} p^x q^{n-x}$$

Here, $\binom{n}{x}$ represents the number of ways to choose x successes from n trials, p^x is the probability of having x successes, and q^{n-x} is the probability of having $n - x$ failures.

In our economic example, let us say the salesperson has a 20% success rate $(p = 0.2)$ in closing deals. If the salesperson conducts 10 independent trials, the probability of closing exactly 2 deals $(x = 2)$ can be calculated using the binomial distribution PMF. This probability calculation provides insights into the likelihood of achieving a specific number of successful deals in a given number of trials, which is valuable information for sales forecasting and performance evaluation.

Suppose that we have an experiment such as tossing a coin or die repeatedly or choosing a marble from an urn repeatedly. Each toss or selection is called a trial. In any single trial there will be a probability associated with a particular event such as head on the coin, 4 on the die, or selection of a red marble. In some cases, this probability will not change from one trial to the next (as in tossing a coin or die). Such trials are then said to be independent and are often called Bernoulli trials after James Bernoulli who investigated them at the end of the seventeenth century.

Let p be the probability that an event will happen in any single Bernoulli trial (called the probability of success). Then $q = 1 - p$ is the probability that the event will fail to happen in any single trial (called the probability of failure). The probability that the event will happen exactly x times in n trials (i.e., successes and $n - x$ failures will occur) is given by the probability function.

$$f(x) = P(X = x) = \binom{n}{x} p^x q^{n-x} = \frac{n!}{x!\,(n-x)!} p^x q^{n-x} \qquad (1)$$

where the random variable X denotes the number of successes in n trials and $x = 0,1, \ldots, n$.

The discrete probability function $P(X = x)$ for the number of successes in n trials, where $x = 0,1, \ldots, n$, is commonly referred to as the binomial distribution. This distribution is so named because, for each value of x, it corresponds to the coefficients of the binomial expansion of $(q + p)^n$, where q and p are the probabilities of failure and success, respectively. The binomial expansion is a mathematical expression obtained by raising the binomial $(q + p)$ to the power of n.

The binomial distribution formula is given by:

$$P(X = x) = \binom{n}{x} p^x q^{n-x}$$

This formula includes the probability of obtaining exactly x successes in n independent and identical trials, with each trial having two possible outcomes: success (with probability p) or failure (with probability $q = 1 - p$).

The binomial distribution is further illustrated by the binomial expansion formula, which expands $(q + p)^n$ into a sum of terms, each representing the probability of a specific number of successes. The coefficients $\binom{n}{x}$ in the expansion correspond to the number of ways to choose x successes from n trials.

In the special case where $n = 1$, the binomial distribution reduces to the Bernoulli distribution, which represents a single Bernoulli trial. The Bernoulli distribution is characterized by the probability of success (p) and the probability of failure $(q = 1 - p)$ in a single trial, making it a fundamental building block for the broader binomial distribution.

Overall, the binomial distribution is a powerful tool in probability theory and statistics, widely used in various fields, including economics, to model and analyse phenomena involving repeated trials with binary outcomes.

The discrete probability function is often called the binomial distribution since for $x = 0, 1, 2, \dots, n$, it corresponds to successive terms in the binomial expansion

$$(q+p)^n = q^n + \binom{n}{1} q^{n-1}p + \binom{n}{2} q^{n-2}p^2 + \cdots + p^n = \Sigma_{x=0}^{n} \binom{n}{x} p^x q^{n-x} \qquad (2)$$

The special case of a binomial distribution with $n = 1$ is also called the Bernoulli distribution.

Some Properties of The Binomial Distribution

1. **Mean (μ) :** The mean of a binomial distribution is given by $\mu = np$, where n is the number of trials and p is the probability of success in a single trial. This provides the average number of successes expected in n trials.

2. **Variance (σ^2) :** The variance of a binomial distribution is calculated using $\sigma^2 = npq$, where $q = 1 - p$ is the probability of failure.

3. **Standard Deviation (σ) :** The standard deviation is the square root of the variance and is given by $\sigma = \sqrt{npq}$.

4. **Moment Generating Function $(M(t))$:** The moment generating function is a function used to derive moments of a distribution. For a binomial distribution, $M(t) = (q + pe^t)^n$, where t is a parameter.

5. **Characteristic Function $(\phi(\omega))$:** The characteristic function is another way to describe a distribution. For a binomial distribution, $\phi(\omega) = \left(q + pe^{i\omega}\right)^n$, where i is the imaginary unit and ω is a parameter.

Illustration : 6.2.4

Seventy-five percent of employed women say their income is essential to support their family. Let X be the number in a sample of 200 employed women who will say their income is essential to support their family. What is the mean and standard deviation of X .

Solution

X is a binomial random variable with $\mathrm{n} = 200$ and $\mathrm{p} = .75$. The mean is $\mu = \mathrm{np} = 200 \times .75 = 150$, and the standard deviation is $\sigma = \sqrt{\mathrm{npq}} = \sqrt{37.5} = 6.12$.

Illustration.6.2.5

A binomial distribution has a mean equal to 8 and a standard deviation equal to 2. Find the values for n and p.

Solution

The following equations must hold: $8 = np$ and $4 = npq$. Substituting 8 for np in the second equation gives $4 = 8q$, which gives $q = .5$. Since $p + q = 1, p = 1 - .5 = .5$. Substituting .5 for p in the first equation gives $\mathrm{n}(.5) = 8$, and it follows that $\mathrm{n} = 16$.

Illustration : 6.2.6

If on the average rainfall on 10 days in every 30 days, obtain the probability that rain will fall on at least 3 days of a given week.

Solution

The probability density function foe a binomial distribution is

$$P(X = x) = \binom{n}{x} p^x q^{n-x}$$

Given $p = \frac{10}{30} = \frac{1}{3}, \quad n = 7, \quad q = 1 - \frac{1}{3} = \frac{2}{3}$

$$P[X \geq 3] = P[X < 3]$$

$$= 1 - P[X = 0,1,2]$$

$$= 1 - P[X = 0] + P[X = 1] + P[X = 2]$$

$$P(X = 0) = \binom{7}{0}\left(\frac{1}{3}\right)^0\left(\frac{2}{3}\right)^{7-0} = \left(\frac{2}{3}\right)^7$$

$$P(X = 1) = \binom{7}{1}\left(\frac{1}{3}\right)^1\left(\frac{2}{3}\right)^{7-1} = \binom{7}{1}\left(\frac{2}{3}\right)^6\left(\frac{1}{3}\right)$$

$$P(X = 1) = \binom{7}{2}\left(\frac{1}{3}\right)^2\left(\frac{2}{3}\right)^{7-2} = \binom{7}{2}\left(\frac{2}{3}\right)^5\left(\frac{1}{3}\right)^2$$

$$P[X \geq 3] = 1 - 0.5706 = 0.4293$$

Illustration : 6.2.7

Ten coins are thrown simultaneously. Find the probability of getting at least seven heads?

Solution

p = Probability of getting a head $= \frac{1}{2}$

q = Probability of not getting a head = $\frac{1}{2}$

The probability of getting x heads in a random throw of 10 coins is

$$P(x) = \binom{10}{x} p^x q^{10-x} = \binom{10}{x}\left(\frac{1}{2}\right)^{10}, \quad x = 0,1,2 \ldots .10$$

probability of getting at least seven heads = $P[X \geq 7]$

$$= P[X = 7] + P[X = 8] + P[X = 9] + P[X = 10]$$

$$= \binom{10}{7}\left(\frac{1}{2}\right)^{10} + \binom{10}{8}\left(\frac{1}{2}\right)^{10} + \binom{10}{9}\left(\frac{1}{2}\right)^{10} + \binom{10}{10}\left(\frac{1}{2}\right)^{10}$$

$$= \left(\frac{1}{2}\right)^{10}\left[\binom{10}{7} + \binom{10}{8} + \binom{10}{9} + \binom{10}{10}\right]$$

$$= \left(\frac{1}{2}\right)^{10}[120 + 45 + 10 + 1] = \frac{176}{1024}$$

Illustration : 6.2.8

A die is tossed 3 times. A success is getting 1 or 6 on a toss. Find the mean and variance of the number of success.

Solution

Given $n = 3, \quad p = p(getting\ 1\ or\ 6) = \frac{1}{6} = \frac{1}{6} = \frac{1}{3}$

Mean = $np = 3 \times \frac{1}{3} = 1$

Variance = $npq = 3 \times \frac{1}{3} \times \frac{2}{3} = \frac{2}{3}$

6.2.4 Normal Distribution

One of the most important examples of a continuous probability distribution is the normal distribution, sometimes called the Gaussian distribution. The density function for this distribution is given by

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2} \qquad -\infty < x < \infty$$

where μ and σ are the mean and standard deviation, respectively. The corresponding distribution function is given by

$$F(x) = P(X \leq x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{-(v-\mu)^2/2\sigma^2} dv$$

If X has the distribution function, we say that the random variable X is normally distributed with mean μ and variance σ^2.

If we let Z be the standardized variable corresponding to X, i.e., if we let

$$z = \frac{x-\mu}{\sigma} \approx (0,1)$$

then the mean or expected value of Z is 0 and the variance is 1. In such cases the density function for Z can be formally placing $\mu = 0$ and $\sigma = 1$, yielding

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}$$

This is often referred to as the standard normal density function.

A graph of the density function, sometimes called the standard normal curve. In this graph we have indicated the areas within 1, 2, and 3 standard deviations of the mean (i.e., between $z = -1$ and $+1$, $z = -2$ and $+2$, $z = -3$ and +3) as equal, respectively, to 68.27%, 95.45% and 99.73% of the total area, which is one. This means that

$$P(-1 \leq Z \leq 1) = 0.6827$$
$$P(-2 \leq Z \leq 2) = 0.9545$$
$$P(-3 \leq Z \leq 3) = 0.9973$$

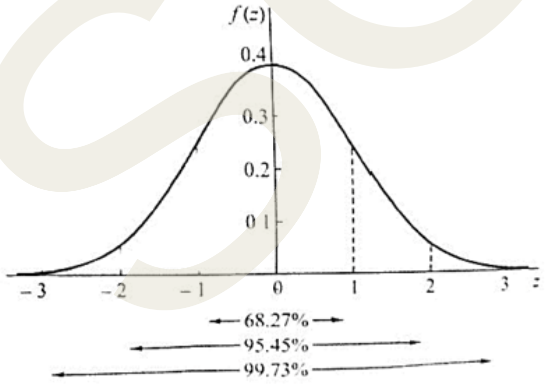
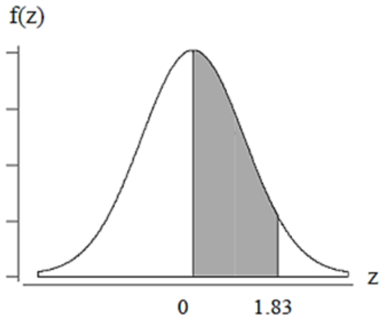
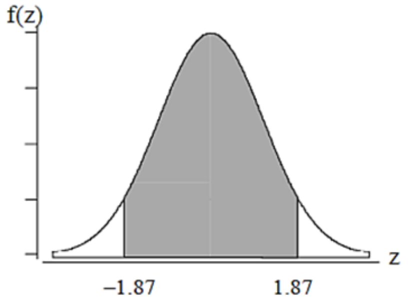


Illustration : 6.2.9

Express the areas shown in the following two standard normal curves as a probability statement and find the area of each one.





Solution

The area under the curve on the left is represented as $\mathrm{P}(0 < \mathrm{Z} < 1.83)$ and from the standard normal distribution table is equal to .9664. The area under the curve on the right is represented as $\mathrm{P}(-1.87 < \mathrm{Z} < 1.87)$ and from the standard normal distribution table is $2 \times .4693 = .9386$.

Illustration : 6.2.10

The distribution of complaints per week per 100,000 passengers for all airlines in a country is normally distributed with $\mu = 4.5$ and $\sigma = 0.8$. Find the standardized values for the following observed values of the number of complaints per week per 100,000 passengers: (a) 6.3; (b) 2.5 ; (c) 4.5 ; (d) 8.0 .

Solution

(a) The standardized value for 6.3 is found by $z = \dfrac{x-\mu}{\sigma} = \dfrac{6.3-4.5}{.8} = 2.25$

(b) The standardized value for 2.5 is found by $z = \dfrac{x-\mu}{\sigma} = \dfrac{2.5-4.5}{.8} = -2.50$

(c) The standardized value for 4.5 is found by $z = \dfrac{x-\mu}{\sigma} = \dfrac{4.5-4.5}{.8} = 0.00$

.

(d) The standardized value for 8.0 is found by $z = \dfrac{x-\mu}{\sigma} = \dfrac{8.0-4.5}{.8} = 4.38$

Illustration : 6.2.11

The net worth of senior citizens is normally distributed with mean equal to $\$225{,}000$ and standard deviation equal to $\$35{,}000$. What percent of senior citizens have a net worth less than $\$300{,}000$?

Solution

Ans. Let $\mathbf{X}$ represent the net worth of senior citizens in thousands of dollars. The percent of senior citizens with a net worth less than $\$300{,}000$ is found by multiplying $P(X < 300)$ times 100. The probability $P(X < 300)$ is shown in figure below. The event $X < 300$ is equivalent to the event $Z < \frac{300-225}{35} = 2.14$. The probability that $Z < 2.14$ is represented as the shaded area in Fig. 4.4. The probability that Z is less than

2.14 is found by adding $P(0 < Z < 2.14)$ to .5, which equals $.5 + .4838 = .9838$. We can conclude that 98.38% of the senior citizens have net worths less than $\$300,000$.

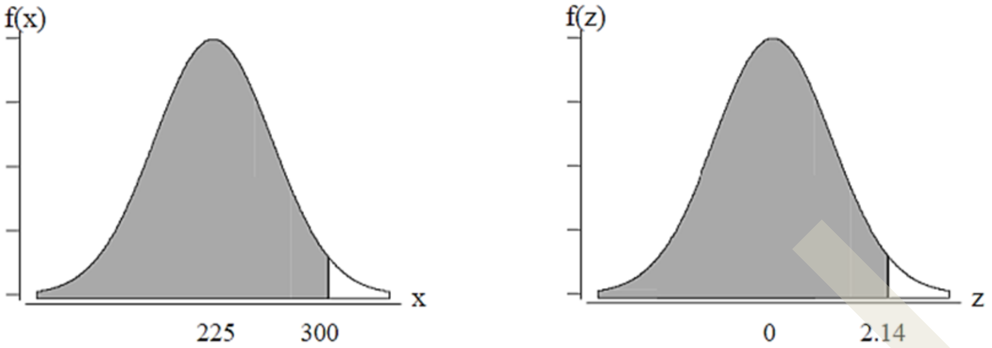


Illustration : 6.2.12

The average test marks in a particular class is 79 and standard deviation is 5. If the marks are normally distributed how many students in a class of 200 did not receive marks between 75 and 82?

Solution

Given $\mu = 79,\ \sigma = 5,\ n = 200$

$$z = \frac{x-\mu}{\sigma} = \frac{75-79}{5} = -0.8$$

$$z = \frac{x-\mu}{\sigma} = \frac{82-79}{5} = 0.6$$

$P[75 < X < 82] = P[-0.8 < Z < 0.6]$

$= 0.2881 + 0.2257 = 0.5138$

The probability of receiving marks outside this range is the complement:

$P(X < 75 \text{ } or \text{ } X > 82) = 1 - P(75 \leq X \leq 82)$

$P(X < 75 \text{ } or \text{ } X > 82) = 1 - 0.5138 = 0.4862$

In a class of 200 students, the expected number of students who scored outside the range is:

Number of students = 200 × 0.4862 = 97.24

Rounding to the nearest whole number:

Number of students≈97

Approximately 97 students did not receive marks between 75 and 82.

Recap

- Random variables are mathematical functions that assign numerical values to the outcomes of a random experiment
- Discrete variables are random variables that take on distinct, separate values with gaps between them, often associated with counting and finite outcomes in probability distributions
- Continuous variables are random variables that can take any value within a range, often associated with measurements and infinite possible outcomes
- Binomial distribution is a discrete probability distribution that models the number of successes in a fixed number of independent and identical Bernoulli trials
- Normal distribution, also known as the Gaussian distribution or bell curve, is a continuous probability distribution characterized by a symmetric, bell-shaped curve, where the majority of observations cluster around the mean

Objective Questions

1. What is the probability that a continuous random variable X takes on any one particular value?
2. What is the term used for a nondiscrete random variable that takes on a noncountably infinite number of values?
3. What is a random variable in probability theory?
4. What is the characteristic of a discrete random variable?
5. A pair of dice is tossed 180 times. What is the probability that the sum on the faces is equal to 7 on 20% or more of the tosses?
6. An Airport claims that 85% of their flights are on time. If the claim is correct, what is the probability that in a sample of 20 flights at the Dallas-Fort Worth Airport that 15 or more of the sample flights are on time?
7. A psychological study involving the troops in the Bosnia peacekeeping force was conducted. If 12 percent of the 21,496 troops are females, what is the probability that in a sample of 50 randomly selected individuals that five or fewer are female?

Answers

1. P(B)≥P(A)
2. Continuous random variable
3. A function that assigns numerical values to outcomes in the sample space.
4. It assumes a finite or countably infinite set of distinct values.
5. 0.1170
6. 0.9327
7. 0.4353

Assignments

1. Thirty percent of the trees in a forest are infested with a parasite. Fifty trees are randomly selected from this forest and X is defined to equal the number of trees in the 50 sampled that are infested with the parasite. The infestation is uniformly spread throughout the forest. Identify the values for n, p, and q. Suppose we define Y to be the number of trees in the 50 sampled that are not infested with the parasite. Then Y is a binomial random variable.

   a. What are the values of n, p, and q for Y ?

   b. The event X = 20 is equivalent to the event that Y = a. Find the value for a.

2. The hospital cost for individuals involved in accidents who do not wear seat belts is normally distributed with mean Rs 7500 and standard deviation Rs 1200.

   (a) Find the cost for an individual whose standardized value is 2.5.

   (b) Find the cost for an individual whose bill is 3 standard deviations below the average.

3. The average TV-viewing time per week for children ages 2 to 11 is 22.5 hours and the standard deviation is 5.5 hours. Assuming the viewing times are normally distributed, find the following.

a. What percent of the children have viewing times less than 10 hours per week?

b. What percent of the children have viewing times between 15 and 25 hours per week?

c. What percent of the children have viewing times greater than 40 hours per week?

4. The amount that airlines spend on food per passenger is normally distributed with mean Rs 8.00 and standard deviation Rs 2.00.

   (a) What percent spend less than Rs 5.00 per passenger?

   (b) What percent spend between Rs 6.00 and Rs 10.00 ?

   (c) What percent spend more than Rs 12.50 ?

5. Find the value of a in each of the following probability statements involving the standard normal variable Z.

   (a) $P(0 < Z < a)\ ) = .4616$

   (b) $P(Z < a) = .8980$

   (c) $P(-a < Z < a) = .8612$

   (d) $P(Z < a) = .1894$

   (e) $P(Z > a) = .1894$

   (f) $P(Z = a) = .5000$

References

1. Spiegel, M. R., Schiller, J. J. & Srinivasan R. A. (2010). *Probability and Statistics*, Third Edition (Schaum's Outlines). McGraw Hill Education (India) Private Limited, New Delhi.

2. Stephens, L. J. (2009). *Schaum's Outline of Beginning Statistics*, Second Edition. McGraw Hill Education (India) Private Limited, New Delhi.

3. Spiegel, M. R., Schiller, J. J. & Srinivasan R. A. (2020). *Schaum's Easy Outline of Probability and Statistics (SCHAUM'S Easy Outlines)*. McGraw Hill Education (India) Private Limited, New Delhi.

Suggested Readings

1. Gupta, S. C., & Kapoor, V. K. (2020). *Fundamentals of Mathematical Statistics*. Sultan Chand & Sons.

2. Goon, A. M., Gupta, M. K., & Dasgupta, B. (2017). *An Outline of Statistical Theory: Volume I*. World Press.

3. Agarwal, B. L. (2014). *Basic Statistics*. New Age International Publishers.

4. Pandey, H. D. (2018). *Statistics for Economics and Business*. Himalaya Publishing House.

5. Sharma, J. K. (2020). *Business Statistics*. Vikas Publishing House.

സർവ്വകലാശാലാഗീതം

വിദ്യയാൽ സ്വതന്ത്രരാകണം
വിശ്വപൗരരായി മാറണം
ഗ്രഹപ്രസാദമായ് വിളങ്ങണം
ഗുരുപ്രകാശമേ നയിക്കണേ

കൂരിരുട്ടിൽ നിന്നു ഞങ്ങളെ
സൂര്യവീഥിയിൽ തെളിക്കണം
സ്നേഹദീപ്തിയായ് വിളങ്ങണം
നീതിവൈജയന്തി പാറണം

ശാസ്ത്രവ്യാപ്തിയെന്നുമേകണം
ജാതിഭേദമാകെ മാറണം
ബോധരശ്മിയിൽ തിളങ്ങുവാൻ
ജ്ഞാനകേന്ദ്രമേ ജ്വലിക്കണേ

കുരീപ്പുഴ ശ്രീകുമാർ

SREENARAYANAGURU OPEN UNIVERSITY

Regional Centres

Kozhikode

Govt. Arts and Science College
Meenchantha, Kozhikode,
Kerala, Pin: 673002
Ph: 04952920228
email: rckdirector@sgou.ac.in

Thalassery

Govt. Brennen College
Dharmadam, Thalassery,
Kannur, Pin: 670106
Ph: 04902990494
email: rctdirector@sgou.ac.in

Tripunithura

Govt. College
Tripunithura, Ernakulam,
Kerala, Pin: 682301
Ph: 04842927436
email: rcedirector@sgou.ac.in

Pattambi

Sree Neelakanta Govt. Sanskrit College
Pattambi, Palakkad,
Kerala, Pin: 679303
Ph: 04662912009
email: rcpdirector@sgou.ac.in

Statistics for Economics

COURSE CODE: B21EC04DC

Sreenarayanaguru Open University
Kollam, Kerala Pin- 691601, email: info@sgou.ac.in, www.sgou.ac.in Ph: +91 474 2966841

ISBN 978-81-982853-5-5
9 788198 285355