UNIT-1

Statistics

statistics is the branch of science that deals with the study of collection, compilation, analysis, interpretation and presentation of data.

Biostatistics

Biostatistics is the branch of science that deals with study of data derived from biological or medical science.

stages in investigation of data

1) Collection of data → The process of obtaining measurements or counts of measurements.

2) Organisation of data → Includes editing, classifying, and tabulating the data collected.

3) Presentation of data → It can be done in the forms of tables, graphs or diagrams.

4) Analysis of data → To take out the useful information from the data for decision making. Information includes mean, median, mode, range, variance, etc.

5) Interpretation of data → concerned with drawing conclusions from the data collected and analyzed and giving meaning to the results.

Applications of Biostatistics

→ It simplifies the information of data, figures.

→ It facilitate comparisons

→ It help in testing hypothesis.

→ It help in prediction and decision making.

Frequency Distribution

Frequency → It is the number of a particular value that occurs in a data.

For example, heart beats 70 times per minute 80, 70 is the frequency.

Frequency distribution → the graphical or tabular representation that displays the number of observations within a given interval (frequency).

A frequency distribution is a means to organize a large amount of data.

Example, The marks of a class of 20 students are 11, 27, 18, 14, 28, 18, 2, 22, 11, 24, 22, 11, 8, 20, 25, 28, 30, 12, 11, 8.

Prepare a frequency table for the same.

Answer → The range of marks of the students is 2-28. Let us take class intervals 0-5, 5-10, 10-15, 15-20, 20-25, 25-30.

Marks	Frequency
0-5	1
5-10	2
10-15	6
15-20	2
20-25	4
25-30	5

Types of frequency distribution

1) **Discrete or ungrouped frequency distributions**

The data in the table is represented as descrete numbers or quantity. The data is ungrouped and not given in class intervals.

For example, if the following data shows the number of children in 20 families:

1, 1, 2, 3, 4, 3, 2, 1, 1, 4, 5, 2, 4, 2, 2, 1, 3, 3, 2, 5

The data may be put in the form of a discrete frequency distribution as follows:

no. of children	Frequency
1	5
2	6
3	4
4	3
5	2

This method is convenient only where the values are largely repeating and the difference between the greatest and the smallest observations is not very large.

2) Continuous or grouped frequency distribution

If the number of observations in data is large and the difference between the greatest and the smallest observations is large, then we condense the data into classes or groups. There are two methods of classifying the data according to the class intervals:

1) Exclusive Method

2) Inclusive Method

1) Exclusive Method → when the upper limit of a class interval coincide with the lower limit of next class interval.

Example, let the marks obtained by 30 students of a class in a test be: 39, 25, 5, 33, 19, 21, 12, 48, 13, 21, 9, 1, 10, 8, 12, 17, 19, 17, 41, 40, 12, 46, 37, 17, 27, 30, 6, 2, 23, 19.

we can arrange these marks as: the class intervals include the items having the range from the lower limit to the value just below the upper limit.

Marks (class intervals)	No. of students (frequency)
0 – 10	6
10 – 20	11
20 – 30	5
30 – 40	4
40 – 50	4

2) Inclusive Method → The series with class intervals, in which all the items having the range from the lower limit upto the upper limit are included. The upper limit of one class does not coincide with the lower limit of next class.

Example,

Marks	No. of students
0 – 10	7
11 – 20	10
21 – 30	6
31 – 40	4
41 – 50	3

Solution Examples

Example → Given below are the ages of 25 students of a class. Prepare a descrete frequency distribution.

15, 16, 16, 14, 17, 17, 16, 15, 15, 16, 16, 15, 17, 16, 16, 14, 16, 15, 14, 15, 16, 16, 15, 14, 15.

Solution → Discrete frequency distribution :

Age	Frequency
14	4
15	8
16	10
17	3

Example → The water tax bills of 30 houses in a locality are given. Construct a grouped frequency distribution with class size of 10.

30, 32, 45, 54, 74, 78, 108, 112, 66, 76, 88, 40, 14, 20, 15, 35, 44, 66, 75, 84, 95, 96, 102, 110, 88, 74, 112, 14, 34, 44.

Solution→ Here, the maximum and minimum values are 112 and 14 respectively.

Range = 112 - 14 = 98

class size given is 10.

Bill	Frequency
14-24	4
24 - 34	2
34 - 44	3
44 - 54	3
54 - 64	1
64 - 74	2
74 - 84	5
84 - 94	3
94 - 104	3
104 - 114	4

3. **Cumulative Frequency Distribution (c.f.)**

If the frequency of the first class is added to that of the second class and this sum is added to that of the third and so on, then the frequencies obtained are known as cumulative frequencies (c.f.).

For example, if we draw the table given in the example of discrete frequency distribution.

no. of children	frequency	c.f.
1	5	5
2	6	11
3	4	15
4	3	18
5	2	20
	20	

In a grouped frequency distribution, the cumulative frequency of a class is the total of all frequencies upto and including that particular class.

Example → The distances covered by 24 cars are given below:

125, 128, 140, 108, 96, 149, 136, 112, 84, 123, 130, 120, 103, 89, 65, 103, 145, 97, 102, 87, 68, 78, 98, 126

Represent them as a cumulative frequency Table using 60 as the lower limit of the first group and all groups having class size of 15.

Solution

Maximum value = 149; Minimum value = 65

Range = 149 − 65 = 84

$$\text{no. of classes} = \frac{84}{15} = 5.6 \sim 6$$

Class interval	Frequency	c.f.
60 − 75	2	2
75 − 90	4	6
90 − 105	6	12
105 − 120	2	14
120 − 135	6	20
135 − 150	4	24
	24	

Measures of Central Tendency

Central tendency is the single value that is typically the representative of the collected data. The most commonly used measures:

i) Arithmetic Mean (A.M.) or simply Mean

ii) Median

iii) Mode

1) Arithmatic Mean

a) Individual Observations or ungrouped data

Arithmatic mean of a set of observations is equal to their sum divided by the total number of observations (n). Arithmatic mean is denoted by $\bar{x}$.

Example – Heights of 5 persons are 144, 152, 151, 158, and 155 in centimeters.

$n = 5$

$$\text{Mean height} = \frac{144 + 152 + 151 + 158 + 155}{5} = \frac{760}{5}$$

$$= \boxed{152\text{ cm}}$$

Example-2 → If the mean of 6, 4, 7, P and 10 is 8, find the value of P.

since $\bar{x} = 8$

$$8 = \frac{6+4+7+P+10}{5} =$$

$$40 = 6+4+7+P+10$$

$$40 = 27+P \Rightarrow \boxed{P = 13}$$

b) Mean of discrete frequency distribution

If a variable x takes values $x_1, x_2, \ldots\ldots, x_n$ with corresponding frequencies $f_1, f_2, f_3 \cdots f_n$ respectively, then the arithmatic mean of these values is

$$\bar{x} = \frac{f_1x_1 + f_2x_2 + \cdots\cdots + f_nx_n}{f_1 + f_2 + \cdots\cdots + f_n}$$

or, $\bar{x} = \dfrac{\sum_{i=1}^{n} f_ix_i}{N}$ where, $N = \sum_{i=1}^{n} f_i = f_1 + f_2 \cdots f_n$

Solution Examples

① Find the mean of the following distribution,

x :	4	6	9	10	15
f :	5	10	10	7	8

Solution

$\boxed{x}$	$\boxed{f}$	$\boxed{fx}$
4	5	20
6	10	60
9	10	90
10	7	70
15	8	120
	N = 40	360

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

$$= \frac{360}{40}$$

$$\boxed{\bar{x} = 9}$$

②

x :	2	4	6	10	p+5
f :	3	2	3	1	2

[and $\bar{x} = 6$ find P = ?]

$\boxed{x}$	$\boxed{f}$	$\boxed{fx}$
2	3	6
4	2	8
6	3	18
10	1	10
p+5	2	2p+10
	N = 11	2p+52

$$\bar{x} = \frac{2p+52}{11} = 6$$

$$66 = 2p+52$$

$$2p = 14$$

$$\boxed{P = 7}$$

③ Find the missing frequencies in the following frequency distribution if the mean is 1.46.

x:	0	1	2	3	4	5	Total
f:	46	?	?	25	10	5	200

solution → Let the missing frequencies be f_1 & f_2.

$\boxed{x}$	$\boxed{f}$	$\boxed{fx}$
0	46	0
1	f_1	f_1
2	f_2	$2f_2$
3	25	75
4	10	40
5	5	25

$[\Sigma fx = 140 + f_1 + 2f_2]$

We know,

$N = 86 + f_1 + f_2 = 200$

$f_1 + f_2 = 114$ ——— (i)

And,

Mean given = 1.46

$$1.46 = \frac{\Sigma fx}{N}$$

$$1.46 = \frac{140 + 2f_2 + f_1}{200}$$

$292 = 140 + f_1 + 2f_2$

$f_1 + 2f_2 = 152$ ——— (ii)

On solving equation (i) and (ii), we get

$f_1 = 76$; $f_2 = 38$

c) Mean of continuous frequency distribution

When class intervals are given, then we don't know the actual values of x.

The values of $x_1, x_2, x_3, \ldots, x_n$ are taken as the mid-points of class intervals.

The formula to obtain mid-value of a class interval $= \frac{1}{2}$ (lower limit + upper limit)

Example

class intervals:	0-20	30-60	40-80
mid-value:	$\frac{1}{2}(0+20)$ $= 10$	$\frac{1}{2}(30+60)$ $= 45$	$\frac{1}{2}(40+80)$ $= 60$

$\boxed{\text{Solution Examples}}$

① Find the mean of following frequency distribution

Interval:	0-10	10-20	20-30	30-40	40-50
f :	7	10	15	8	10

Solution

Class-interval	Mid-values(x)	Frequency (f)	fx
0-10	5	7	35
10-20	15	10	150
20-30	25	15	375
30-40	35	8	280
40-50	45	10	450
		$N = 50$	$\Sigma fx = 1290$

$$\text{Mean} = \frac{\Sigma fx}{\Sigma f} = \frac{1290}{50}$$

$$\boxed{\bar{X} = 25.8}$$

② If the mean of the following distributions is 54, find the value of p.

class:	0-20	20-40	40-80	60-80	80-100
f :	7	p	10	9	13

Solution

class	Mid-values(x)	frequency (f)	fx
0-20	10	7	70
20-40	30	p	30p
40-60	50	10	500
60-80	70	9	630
80-100	90	13	1170
		$N = 39+p$	$\Sigma fx = 2370 + 30p$

$$\text{Mean} = \frac{\Sigma fx}{\Sigma f}$$

$$54 = \frac{2370 + 30p}{39 + p}$$

$$2106 + 54p = 2370 + 30p$$

$$24p = 264$$

$$p = \frac{264}{24}$$

$$\boxed{p = 11}$$

Median

Median is the middle value of the given list of data when arranged in an order. The arrangement of data or observations can be made either in ascending order or descending order.

① Median of individual observations

For a data $x_1, x_2, \dots\dots, x_n$, the median is calculated by using following algorithm.

→ Arrange the observations in ascending or descending order.

→ Determine the total number of observations, say it is n.

→ If n is odd, then

$$\text{Median} = \left(\frac{n+1}{2}\right)^{th} \text{observation in the data}$$

And if n is even, then

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

Solution Examples

① Find the median of the following data.

25, 34, 31, 23, 22, 26, 35, 28, 20, 32

Solution

Arranging the data in ascending order,

20, 22, 23, 25, 26, 28, 31, 32, 34, 35

Here, number of observations, $n = 10$ (even)

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

$$= \frac{5^{th} \text{observation} + 6^{th} \text{observation}}{2}$$

$$= \frac{26+28}{2}$$

$$\boxed{\text{Median} = 27}$$

③ Find the median of the following values:

37, 31, 42, 43, 46, 25, 39, 45, 32

Solution

Arranging the data in ascending order

25, 31, 32, 37, 39, 42, 43, 45, 46

Here, number of observations, $n = 9$ (odd)

$$\text{Median} = \left(\frac{9+1}{2}\right)^{th} \text{observation}$$

$$= 5^{th} \text{observation}$$

$$\boxed{\text{Median} = 39}$$

② Median of discrete frequency distribution

In case of a discrete frequency distribution, we calculate the median by using following algorithm.

→ Find the cumulative frequencies (c.f.).

→ Find $N/2$, where N is sum of frequencies.

→ see the c.f. just greater than $N/2$ and determine the corresponding value of x.

→ The value obtained of x is the median.

Example → obtain the median for the following frequency distribution;

x:	1	2	3	4	5	6	7	8	9
f:	8	10	11	16	20	25	15	9	6

Solution

x	f	cf
1	8	8
2	10	18
3	11	29
4	16	45
5	20	65
6	25	90
7	15	105
8	9	114
9	6	120
	$N = 120$	

$$\frac{N}{2} = \frac{120}{2} = 60$$

c.f. just greater than 60 is 65 and the value of x corresponding to 65 is 5.

so, $\boxed{\text{median} = 5}$

(3) Median of continuous frequency distribution

To calculate this, we the following algorithm.

1) Obtain the frequency distribution.

2) Find cumulative frequencies and obtain N.

3) Find N/2.

4) See the c.f. just greater than N/2 and determine the corresponding class interval. This is known as median class.

5) Now, to calculate median, we the formula;

$$\text{Median} = l + \left\{\frac{\frac{N}{2} - F}{f}\right\} \times h$$

where,

l = lower limit of median class

f = frequency of the median class

h = size of the median class

F = c.f. of the class preceding the median class.

Solved examples

(1) Calculate median from following distribution:

class:	5-10	10-15	15-20	20-25	25-30	30-35	35-40
f:	5	6	15	10	5	4	2

Solution

class	frequency	c.f.
5-10	5	5
10-15	6	11
15-20	15	26
20-25	10	36
25-30	5	41
30-35	4	45
35-40	2	47
	N = 47	

$N/2 = \frac{47}{2} = 23.5$

The c.f. just greater than N/2 is 26 and corresponding class is 15-20.

so, 15-20 is the median class.

$l = 15$; $f = 15$; $F = 11$ and $h = 5$

$$\text{Median} = l + \left[\frac{\frac{N}{2} - F}{f}\right] \times h$$

$$= 15 + \frac{23.5 - 11}{15} \times 5 = 15 + \frac{12.5}{3} = 19.16$$

② Calculate the median from following data:

x:	0-10	10-30	30-60	60-80	80-90
f:	5	15	30	8	2

Solution → Here, the class intervals are of unequal width. If the class intervals are of unequal width, the frequencies need not be adjusted to make the class intervals equal.

x	f	c.f.
0-10	5	5
10-30	15	20
30-60	30	50
60-80	8	58
80-90	2	60
	N=60	

$N/2 = \frac{60}{2} = 30$

c.f. just greater than $\frac{N}{2}$ is 50 and corresponding class is 30-60.

$l = 30$; $f = 30$; $F = 20$

$h = 30$

$$\text{Median} = l + \left(\frac{N/2 - F}{f}\right) \times h$$

$$= 30 + \frac{30-20}{30} \times 30 = 40$$

$\boxed{\text{Mode}}$

Mode is the value that appears most often in a set of data values.

For example, if a set of numbers contains 1, 1, 3, 5, 6, 6, 7, 7, 7, 8, the mode would be 7 as it appears the most times.

Example – Find mode from following data:

110, 120, 130, 120, 110, 140, 130, 120, 140, 120

Solution

Value	frequency
110	2
120	4
130	2
140	2

Since, the value 120 occurs maximum number of times, i.e. 4. Hence, the mode is 120.

Mode of discrete series

① Compute the modal value for the following frequency distribution;

x:	95	105	115	$\boxed{125}$	135	145	155	165
f:	4	2	18	$\boxed{22}$	21	19	10	3

since the frequency of 125 is largest. so; 125 is the modal value.

Mode of Continuous frequency distribution

→ Obtain the frequency distribution.

→ Determine the class of maximum frequency. This class is called modal class.

→ Formula,

$$\text{Mode} = l + \frac{f - f_1}{2f - f_1 - f_2} \times h$$

where, l = lower limit of modal class

f = frequency of modal class

h = width (or size) of modal class

f_1 = frequency of preceding class of modal class

f_2 = frequency of class following modal class

Solved examples

① Compute the mode for the following distributio

class:	0-4	4-8	8-12	$\boxed{12-16}$	16-20	20-24	24-28
f:	5	7	9	$\boxed{17}$	12	10	6

Solution → Here, the modal class is 12-16 because it has highest frequency.

$l = 12$; $h = 4$; $f = 17$; $f_1 = 9$; $f_2 = 12$

$$\text{Mode} = l + \frac{f - f_1}{2f - f_1 - f_2} \times h$$

$$= 12 + \frac{17 - 9}{34 - 9 - 12} \times 4$$

$$= 12 + \frac{8}{13} \times 4 = 12 + \frac{32}{13} = 14.46$$

② Find mode.

Class :	3-6	6-9	9-12	12-15	15-18	18-21	21-24
f :	2	5	10	23	21	12	3

solution → The modal class is 12-15.

$l = 12$; $h = 3$; $f = 23$; $f_1 = 10$; $f_2 = 21$

$$\text{Mode} = l + \frac{f - f_1}{2f - f_1 - f_2} \times h$$

$$= 12 + \frac{23 - 10}{(2 \times 23) - 10 - 21} \times 3$$

$$= 12 + \frac{13}{46 - 10 - 21} \times 3$$

$$= 12 + \frac{13}{15} \times 3$$

$$= 12 + \frac{13}{5}$$

$$\text{Mode} = 14.6$$

Measures of Dispersion

Dispersion is the measure of variations in the values of a data set. or the scatteredness of the observations in a distribution around the central value. Example,

Let us consider the runs scored by two batsmen B_1 and B_2 in their last ten matches as:

Match:	1	2	3	4	5	6	7	8	9	10
B_1 :	30	91	0	64	42	80	30	5	117	71
B_2 :	53	46	48	50	53	53	58	60	57	52

The mean and median of scores of both the batsmen are 53 and 53, i.e., same. But, scores of batsman B_1 are more scattered or have more variation than that of batsman B_2.

This variation is called dispersion.

commonly used measures of dispersion are:

1) Range.
2) Mean deviation
3) standard deviation

Range

Range is the difference between two extreme observations of the distribution. For example, if A and B are the greatest and smallest observations respectively in a distribution, then its range is A-B.

Range = Maximum value- Minimum value

And coefficient of range is the ratio of difference 'A' and 'B' and sum of 'A' and 'B'.

$$\text{coefficient of range} = \frac{A-B}{A+B}$$

Ex- Calculate the range and coefficient of range for the Hb % of 10 patients,

8.4, 9.4, 10.1, 10.9, 9.3, 11.2, 9.6, 12.1, 11.7, 13.2

Solution

Highest value = 13.2 ; Lowest value = 8.4

Range = 13.2 - 8.4 = 4.8

$$\text{coefficient of range} = \frac{13.2-8.4}{13.2+8.4}$$

$$= \frac{4.8}{21.6} = 0.22$$

Mean Deviation

① Mean deviation (M.D) for individual observations

Calculations will follow the following algorithm

1) Calculate the central value (mean or median) about which mean deviation is to be calculated.

2) Calculate the deviation of observations from the central value.

3) Obtain the total of these deviations.

Formula

$$M.D. = \frac{1}{n} \times \Sigma |x - A|$$

where, n = no. of observations (x)

A = central value (mean or median)

Ex- Calculate the mean deviation about median from the following data.

340, 150, 210, 240, 300, 310, 320

Solution→ Arranging them in ascending order,

150, 210, 240, 300, 310, 320, 340

clearly, the middle observation is 300. so, median = 300.

Now, determination of deviation of observations from the central value (300) which is $|x-300|$.

$\boxed{x}$	$\lvert x-300\rvert$
340	40
150	150
210	90
240	60
300	0
310	10
320	20
$n=7$	370

$$M.D. = \frac{1}{n} \times \Sigma |x-300|$$

$$= \frac{1}{7} \times 370$$

$$M.D. = \frac{370}{7}$$

$$M.D = 52.8$$

Ex:- Find M.D. from the mean of following data.

6, 7, 10, 12, 13, 4, 8, 20

Solution.→ Let $\bar{x}$ be the mean of the given data.

$$\bar{x} = \frac{6+7+10+12+13+4+8+20}{8}$$

$$\bar{x} = 10$$

Now, determination of deviation of observations from the mean value (i.e., 10).

$\boxed{x}$	$\lvert x-10\rvert$
6	4
7	3
10	0
12	2
13	3
4	6
8	2
20	10
$n=8$	30

$$M.D. = \frac{1}{n} \times \Sigma |x-10|$$

$$= \frac{1}{8} \times 30$$

$$= \frac{30}{8}$$

$$M.D. = 3.75$$

② Mean deviation of a discrete frequency

Apply following algorithm.

1) Calculate the central value (mean or median) about which mean deviation is to be calculated

2) Take deviations from the central value and ignore signs. $|x-A|$

3) Multiply these deviations by respective frequencies and obtain $\Sigma f\,|x-A|$. A is central value.

4) Divide the total obtained in step-III by the number of observations, i.e., $N=\Sigma f$ to obtain mean deviation,

Formula

$$M.D. = \frac{1}{N}\Sigma f\,|x-A|$$

Solved examples

① Calculate the deviation about mean for the following data:

x:	3	9	17	23	27
f:	8	10	12	9	5

Solution

$\boxed{x}$	$\boxed{f}$	$\boxed{fx}$	$\lvert x-15\rvert$	$f\lvert x-15\rvert$
3	8	24	12	96
9	10	90	6	60
17	12	204	2	24
23	9	207	8	72
27	5	135	12	60
	N = 44	660		312

$$\text{Mean} = \frac{1}{N}(\Sigma fx) = \frac{660}{44} = 15$$

$$M.D. = \frac{1}{N} \times \Sigma f\,|x-15|$$

$$= \frac{1}{44} \times 312 = 7.09$$

② Calculate the M.D. from median for the following distribution:

x:	10	15	20	25	30	35	40	45
f:	7	3	8	5	6	8	4	9

Solution → First, we will calculate median.

$\boxed{x}$	$\boxed{f}$	$\boxed{cf}$	$\boxed{\lvert x-30\rvert}$	$\boxed{f\lvert x-30\rvert}$
10	7	7	20	140
15	3	10	15	45
20	8	18	10	80
25	5	23	5	25
30	6	29	0	0
35	8	37	5	40
40	4	41	10	40
45	9	50	15	135
	N=50			505

The c.f. just greater than N/2 is 29 and the corresponding value of x is 30. So,

median = 30

$$M.D. = \frac{1}{N} \times \Sigma f\,|x-30|$$

$$= \frac{1}{50} \times 505 = 10.1$$

③ Mean deviation for continuous frequency:

Here, the calculation of mean deviation is same as in descrete frequency distribution. The only difference is that, here we have to obtain the mid-points of various classes and take the deviations of these mid-points from the given central value (mean or median).

Solved examples

(i) Find the M.D. about median of following data

Class:	0-6	6-12	12-18	18-24	24-30
f:	8	10	12	9	5

Solution → First, to calculate median;

Class	x	f	c.f.	$\lvert x-14\rvert$	$f\,\lvert x-14\rvert$
0-6	3	8	8	11	88
6-12	9	10	18	5	50
12-18	15	12	30	1	12
18-24	21	9	39	7	63
24-30	27	5	44	13	65
		N=44			278

$N = 44$; $\frac{N}{2} = 22$ [x here represents mid-values of classes]

The c.f. just greater than $\frac{N}{2}$ is 30. Thus, 12-18 is the median class.

$$\text{Median} = l + \frac{N/2 - F}{f} \times h$$

$$= 12 + \frac{22-18}{12} \times 6 = 12 + \frac{4\times6}{12} = 14.$$

$$\text{M.D.} = \frac{1}{N} \times \Sigma f\,|x-14|$$

$$= \frac{1}{44} \times 278 = 6.318$$

(ii) Find the mean deviation from the mean of following data:

class:	10-20	20-30	30-40	40-50	50-60	60-70	70-80
frequencies:	2	3	8	14	8	3	2

Solution.

$[\bar{x} = 45]$

Classes	x	f	fx	$\lvert x-\bar{x}\rvert$	$f\lvert x-\bar{x}\rvert$
10-20	15	2	30	30	60
20-30	25	3	75	20	60
30-40	35	8	280	10	80
40-50	45	14	630	0	0
50-60	55	8	440	10	80
60-70	65	3	195	20	60
70-80	75	2	150	30	60
		N = 40	1800		400

$$\text{Mean} = \frac{1}{N} \Sigma fx = \frac{1800}{40} = 45$$

And,

$\Sigma f\,|x-\bar{x}| = 400$ and $N = \Sigma f = 40$

$$\text{M.D.} = \frac{1}{N} \Sigma f\,|x-\bar{x}|$$

$$= \frac{1}{40} \times 400$$

$$\text{M.D.} = 10$$

Variance and Standard Deviation

Variance → A measurement of how far each number in a data set is from the mean and thus from every other number in the set. It is denoted by Var(X) or σ^2.

standard deviation → A measure of the dispersion of a dataset relative to its mean and is calculated as the square root of the variance. It is denoted by σ.

$$\text{standard deviation} = +\sqrt{\text{Var}(X)}$$

① Variance of individual observations

calculation by following algorithm.

1) Compute the mean, $\bar{x}$ of the given observations.

2) Take the deviations of the observations from the mean, i.e., find $x-\bar{x}$.

3) Square the deviations and obtain their sum as $\Sigma(x-\bar{x})^2$.

$$\text{Formula of variance} = \frac{1}{n}\left\{\Sigma(x-\bar{x})^2\right\}$$

where n = no. of observations.

solved examples

(i) Compute the variance and standard deviation of the following observations of marks of 5 students.

8, 12, 13, 15, 22

solution

$$\text{Mean, } \bar{x} = \frac{8+12+13+15+22}{5} = 14$$

x	$x-\bar{x}$	$(x-\bar{x})^2$
8	−6	36
12	−2	4
13	−1	1
15	1	1
22	8	64
		$\Sigma(x-\bar{x})^2 = 106$

$n = 5$ and $\Sigma(x-\bar{x})^2 = 106$

$$\text{Var}(X) = \frac{1}{n}\Sigma(x-\bar{x})^2 = \frac{106}{5} = 21.2$$

$$\text{standard deviation} = \sqrt{\text{Var}(X)} = \sqrt{21.2} = 4.604$$

② Variance of discrete frequency distribution

Calculate by following algorithm.

1) Obtain the frequency distribution.

2) Find mean ($\bar{x}$).

3) Compute deviations $(x-\bar{x})$ from the mean.

4) Find the squares of deviations $-(x-\bar{x})^2$.

5) Multiply the squared deviations by their respective frequencies. And obtain their total as $\Sigma f(x-\bar{x})^2$.

Formula

$$\text{variance} = \frac{1}{N}\left\{\Sigma f(x-\bar{x})^2\right\}$$

where, N = sum of frequencies.

Solved examples

(i) Find the variance and standard Deviation(S.D) of the following distribution;

x:	2	4	6	8	10	12	14	16
f:	4	4	5	15	8	5	4	5

Solution.

x	f	fx	$x-\bar{x}$	$(x-\bar{x})^2$	$f(x-\bar{x})^2$
2	4	8	−7	49	196
4	4	16	−5	25	100
6	5	30	−3	9	45
8	15	120	−1	1	15
10	8	80	1	1	8
12	5	60	3	9	45
14	4	56	5	25	100
16	5	80	7	49	245
	N=50	450			754

N=50 and $\Sigma fx = 450$

$$\text{Mean} = \frac{\Sigma fx}{N} = \frac{450}{50} = 9$$

And,

$$\Sigma f(x-\bar{x})^2 = 754$$

$$\text{variance}, \sigma^2 = \frac{1}{N}\Sigma f(x-\bar{x})^2$$

$$= \frac{754}{50} = 15.08$$

$$\text{S.D.} = \sqrt{\sigma^2} = \sqrt{15.08} = 3.88$$

③ Variance of a group or continuous distribution

Calculate by following algorithm.

1) Find the mid-points of various classes.

2) Take the deviations of these mid-points from an assumed mean. $(x-\bar{x})$.

3) Divide these deviations by the class size, h which is denoted as, $u = \frac{x-\bar{x}}{h}$.

4) Multiply u of each class with their corresponding frequency and obtain Σfu.

5) Square the values of u and multiply them with corresponding frequencies to obtain Σfu^2.

6) Substitute the values of Σfu, Σfu^2 and $N = \Sigma f$ in the following formula.

$$var(X) \text{ or } \sigma^2 = h^2\left[\frac{1}{N}\Sigma fu^2 - \left(\frac{1}{N}\Sigma fu\right)^2\right]$$

Solved example

Calculate standard deviation for following distribution;

Class :	20-30	30-40	40-50	50-60	60-70	70-80	80-90
f :	3	6	13	15	14	5	4

Solution → Let us assume the mean to be 55.

class	x	f	$u = \frac{x-55}{10}$	fu	u^2	fu^2
20-30	25	3	-3	-9	9	27
30-40	35	6	-2	12	4	24
40-50	45	13	-1	-13	1	13
50-60	55	15	0	0	0	0
60-70	65	14	1	14	1	14
70-80	75	5	2	10	4	20
80-90	85	4	3	12	9	36
		N=60		$\Sigma fu = 2$		$\Sigma fu^2 = 134$

$$var(X) = h^2\left[\frac{1}{N}\Sigma fu^2 - \left(\frac{1}{N}\Sigma fu\right)^2\right]$$

$$= 100\left[\frac{134}{60} - \left(\frac{2}{60}\right)^2\right] = 222.9$$

$$S.D. = \sqrt{Var(X)} = \sqrt{222.9} = 14.94$$

Relation between Mean, Median and Mode

The empirical relation between Mean, Median and Mode is :

$$\text{Mode} = 3\,\text{Median} - 2\,\text{Mean}$$

Example - Calculate the median when Mean and Mode of distribution are 38.6 and 32.6 respectively.

Solution.

$$\text{Median} = \text{Mode} + \frac{2}{3}(\text{Mean} - \text{Mode})$$

$$= 32.6 + \frac{2}{3}(38.6 - 32.6)$$

$$= 32.6 + \frac{2}{3} \times 6$$

$$= 32.6 + 4$$

$$\text{Median} = 36.6$$

Correlation

If these are two variables and changes in the value of one variable will affect the value of the other variable, then both the variables are correlated. Correlation is a statistical tool used to measure the relationship between two sets of variables.

Correlation types

1) Positive correlation → when two variables vary in the same direction.

For example: sale and purchase both increases

Sales:	50	60	70	80	90	100
Purchase:	11	15	19	23	27	31

2) Negative correlation → when both the variables vary in opposite direction.

For example, consider the correlation between production and price of crop.

Production (kg):	100	200	300	400	500	600	700	800
price (per kg) :	10	9	8	7	6	5	4	3

Degree of correlation

1) <u>Perfect correlation</u> → If two variables change in the same direction and in the same proportion, the correlation is <u>perfect positive correlation</u>. The coefficient of correlation in this case is +1.

If the variables change in the opposite direction and in same proportion, the correlation is called <u>perfect negative correlation</u>. In this case, the coefficient of correlation is −1.

2) <u>Absence of correlation (No correlation)</u>

If two variables show no relation between them or change in one variable does not lead to change in other variable, it is the absence of correlation. The coefficient of correlation is zero.

<u>coefficient of correlation</u> → The extent of measurement as to how much one number can be influenced by changes in another number is known as correlation coefficient. It is denoted by 'r'. This is also known as Karl Pearson's correlation coefficient.

Coefficient of correlation lies between −1 and +1.

when :

(i) $r = -1$, shows perfect negative correlation between variables.

(ii) $r = 0$, no correlation

(iii) $r = +1$, shows a perfect positive correlation

<u>Calculation of Karl Pearson's correlation coefficient</u>

$$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \Sigma y^2}} \quad \text{or} \quad r = \frac{\Sigma(X-\bar{X})(Y-\bar{Y})}{\sqrt{\Sigma(X-\bar{X})^2(Y-\bar{Y})^2}}$$

where,

$x = X - \bar{X}$ (deviation from A.M. of X-series

$y = Y - \bar{Y}$ (deviation from A.M. of y-series)

N = no. of observations.

<u>Example</u>- Calculate Karl Pearson coefficient of correlation

X:	42	52	55	60	66	68	65	60	58	34
Y:	11	13	18	22	26	40	31	27	24	18

solution

X	Y	$x = X - \bar{X}$	x^2	$y = Y - \bar{Y}$	y^2	xy
42	11	−14	199	−12	144	168
52	13	−4	16	−10	100	40
55	18	−1	1	−5	25	5
60	22	4	16	−1	1	−4
66	26	10	100	3	9	30
68	40	12	144	17	289	204
65	31	9	81	8	64	72
60	27	4	16	4	16	16
58	24	2	4	1	1	2
34	18	−22	484	−5	25	110
560	230		$\Sigma x^2 = 1058$		$\Sigma y^2 = 674$	$\Sigma xy = 643$

$$\bar{X} = \frac{\Sigma X}{N} = \frac{560}{10} = 56$$

$$\bar{Y} = \frac{\Sigma Y}{N} = \frac{230}{10} = 23$$

$$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \Sigma y^2}} = \frac{643}{\sqrt{1058 \times 643}}$$

$$= \frac{643}{\sqrt{680294}}$$

$$r = \frac{643}{824.8} = 0.779$$

variance-covariance method

$$r = \frac{\operatorname{Cov}(X,Y)}{\sqrt{\operatorname{Var}(X)}\sqrt{\operatorname{var}(Y)}}$$

Example – If covariance between X and Y variables is 12.5 and the variance of X and Y are respectively 16.4 and 13.8. Find r.

Solution

$\operatorname{cov}(X,Y) = 12.5$; $\operatorname{var}(X) = 16.4$; $\operatorname{var}(Y) = 13.8$

$$r = \frac{12.5}{\sqrt{16.4}\sqrt{13.8}} = \frac{12.5}{\sqrt{226.32}}$$

$$r = \frac{12.5}{15.04}$$

$$r = 0.83$$

Multiple Correlation

When the value of a variable is influenced by another variable, the relationship between them is a simple correlation.

But, when a single variable is influenced by two or more (multiple) other variables, the relationship between them is the multiple correlation.

The single variable which is getting influenced by other multiple variables is called dependent variable. And those which influence the single variable are called independent variables. It means independent variables cause changes in the dependent variables.

'R' represents the correlation between the dependent variable and all independent variables collectively.

'r' represents the correlation between a single dependent variable and a single independent variable.

Suppose X_1 is a dependent variable which is influenced by other variables $X_2, X_3 \ldots\ldots$ considered together. The multiple correlation is a measure of the relationship between X_1 and $X_2, X_3 \ldots\ldots$

Notations

→ $R_{1.23}$ denotes the multiple correlation of the dependent variable, X_1 with two independent variables X_2 and X_3.

→ $R_{2.13}$ is the multiple correlation of the dependent variable X_2 with two independent variables X_1 and X_3.

→ $R_{1.234}$ is the multiple correlation of the dependent variable X_1 with three independent variables X_2, X_3 and X_4.

Also,

→ r_{12} is the correlation between dependent variable X_1 and independent variable X_2.

→ r_{13} is the correlation between dependent variable X_1 and independent variable X_3.

Properties of Multiple correlation

1) The values of coefficient of multiple correlation lies between 0 and 1. It is never negative. Hence, $0 \le R \le 1$.

2) The position of the subscripts to the right of the dot does not make any difference. Thus,

$R_{1.34} = R_{1.43}$ and $R_{2.134} = R_{2.341} = R_{2.431}$

3) If $r_{12} = 0$ and $r_{13} = 0$, then $R_{1.23} = 0$ and vice versa.

4) $R_{1.23} \ge r_{12}$ and r_{13} ; $R_{2.13} \ge r_{12}$ and r_{23}

5) Coefficient of a multiple determination is calculated by the square of the coefficient of a multiple correlation.

$$R^2 = \frac{\text{Explained variation}}{\text{Total variation}}$$

= coefficient of multiple correlation

Methods of Calculation

The method for calculating multiple correlation coefficient is based on the number of independent variables.

1) When idependent variables are two:→ In this case, we have three multiple correlation coefficients: These are:

(i) $R_{1.23}$ (ii) $R_{2.13}$ (iii) $R_{3.12}$

Formulae

(i) Multiple correlation between dependent variable X_1 and the group of independent variables X_2 and X_3 is calculated as follows:

$$R_{1.23} = \sqrt{\frac{r_{12}^2 + r_{13}^2 - r_{12} r_{13} r_{23}}{1 - r_{23}^2}}$$

or,

$$R_{1.23} = \sqrt{r_{12}^2 + r_{13.2}^2 (1 - r_{12}^2)}$$

(ii) Multiple correlation between dependent variable X_2 and independent variables X_1 & X_3.

$$R_{2.13} = \sqrt{\frac{r_{12}^2 + r_{23}^2 - 2r_{12}r_{13}r_{23}}{1 - r_{13}^2}}$$

or,

$$R_{2.13} = \sqrt{r_{12}^2 + r_{23.1}^2 (1 - r_{12}^2)}$$

(iii) Multiple correlation between dependent variable X_3 and independent variables X_1 and X_2.

$$R_{3.12} = \sqrt{\frac{r_{13}^2 + r_{23}^2 - 2r_{12}r_{13}r_{23}}{1 - r_{12}^2}}$$

or,

$$R_{3.12} = \sqrt{r_{13}^2 + r_{23.1}^2 (1 - r_{13}^2)}$$

2) When independent variables are three → In this case, we have four multiple correlation coefficients as:

(i) $R_{1.234}$ (ii) $R_{2.134}$ (iii) $R_{3.124}$ (iv) $R_{4.123}$

Formulae

(i) $R_{1.234} = \sqrt{1 - (1 - r_{14}^2)(1 - r_{13.4}^2)(1 - r_{12.34}^2)}$

(ii) $R_{2.134} = \sqrt{1 - (1 - r_{24}^2)(1 - r_{23.4}^2)(1 - r_{12.34}^2)}$

(iii) $R_{3.124} = \sqrt{1 - (1 - r_{34}^2)(1 - r_{23.4}^2)(1 - r_{13.24}^2)}$

(iv) $R_{4.123} = \sqrt{1 - (1 - r_{34}^2)(1 - r_{24.3}^2)(1 - r_{14.23}^2)}$

Solved Examples

① Calculate the multiple correlation coefficients if dependent variable $r_{12} = 0.96$ and independent variables are $r_{13} = 0.49$ and $r_{23} = 0.46$.

<u>solution</u>

$R_{1.23}$ = Multiple correlation between X_1 on one hand and X_2 and X_3 on the other

$$R_{1,23} = \sqrt{\frac{r_{12}^2 + r_{13}^2 - 2r_{12}r_{13}r_{23}}{1 - r_{23}^2}}$$

$$= \sqrt{\frac{(0.96)^2 + (0.49)^2 - 2(0.96)(0.49)(0.46)}{1-(0.46)^2}}$$

$$= \sqrt{\frac{0.9216 + 0.2401 - 0.432768}{0.7884}}$$

$$= \sqrt{\frac{0.728932}{0.7884}} = 0.96$$