**Chapter- 2**

**Measures of Central Tendency**

In the study of a population with respect to one in which we are interested we may get a large number of observations. It is not possible to grasp any idea about the characteristic when we look at all the observations. So it is better to get one number for one group. That number must be a good representative one for all the observations to give a clear picture of that characteristic. Such representative number can be a central value for all these observations. This central value is called a measure of central tendency or an average or a measure of locations.

**Types of Averages:**

There are five averages. Among them *mean, median and mode* are called ***simple averages*** and the other two averages *geometric mean and harmonic mean* are called ***special averages***.

# **Characteristics for a good or an ideal average:**

The following properties should possess for an ideal average.

* 1. It should be rigidly defined.
	2. It should be easy to understand and compute.
	3. It should be based on all items in the data.
	4. Its definition shall be in the form of a mathematical formula.
	5. It should be capable of further algebraic treatment.
	6. It should have sampling stability.
	7. It should be capable of being used in further statistical computations or processing.

**Arithmetic mean**

The arithmetic mean (or, simply average or mean) of a set of numbers is obtained by dividing the sum of numbers of the set by the number of numbers. If the variable x assumes n values x1, x2 …xn then the mean, is given by



# **Example*:* Calculate the mean for 2, 4, 6, 8, and 10.**

**Solution:** Mean = = = 6

 **(i) Direct method :** If the observations  x1,x2,x3........xn have frequencies  f1,f2,f3........fn  respectively, then the mean is given by :



This method of finding the mean is called the direct method.

# **Example:**

***Given the following frequency distribution, calculate the arithmetic mean***

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| ***Marks (x)*** | ***50*** | ***55*** | ***60*** | ***65*** | ***70*** | ***75*** |
| ***No of Students (f)*** | ***2*** | ***5*** | ***4*** | ***4*** | ***5*** | ***5*** |

**Solution:**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Marks (x)** | 50 | 55 | 60 | 65 | 70 | 75 | Total |
| **No of Students (f)** | 2 | 5 | 4 | 4 | 5 | 5 | 25 |
| **fx** | 100 | 275 | 240 | 260 | 350 | 375 | 1600 |



 = = 64

**(ii) Short cut method:** In some problems, where the number of variables is large or the values of xiorfiare larger, then the calculations become tedious. To overcome this difficulty, we use short cut or deviation method in which an approximate mean, called assumed mean is taken. This assumed mean is taken preferably near the middle, say A, and the deviation di =xi − Ais calculated for each variable Then the mean is given by the formula:



**Mean for a grouped frequency distribution**

# **Example: *Given the following frequency distribution, calculate the arithmetic mean***

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Marks (x) | 50 | 55 | 60 | 65 | 70 | 75 |
| No of Students (f) | 2 | 5 | 4 | 4 | 5 | 5 |

**Solution:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| x | f | fx | d=x-A | fd |
| 50 | 2 | 100 | -10 | -20 |
| 55 | 5 |  275 | -5 | -25 |
| **60** | 4 | 240 | 0 | 00 |
| 65 | 4 | 260 |  | 20 |
| 70 | 5 | 350 |  | 50 |
| 75 | 5 | 375 |  | 75 |
|  | 25 | 1600 |  | 100 |

**By Direct method:**



 = = 64

**By Short-cut method:**

*x* *A* *fd*

 *N*

 = 60 + = 60 + 4 = 64

**Mean for a grouped frequency distribution**

Find the class mark or mid-value x, of each class, as



# **Example:**

***Following is the distribution of persons according to different income groups. Calculate arithmetic mean.***

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Income SR (100) | 0-10 | 10-20 | 20-30 | 30-40 | 40-50 | 50-60 | 60-70 |
| Number of persons | 6 | 8 | 10 | 12 | 7 | 4 | 3 |

**Solution:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Income C.I | Number of Persons (f) | Mid X | d = x Ac | fd |
| 0-10 | 6 | 5 | -3 | -18 |
| 10-20 | 8 | 15 | -2 | -16 |
| 20-30 | 10 | 25 | -1 | -10 |
| 30-40 | 12 | A =35 | 0 | 0 |
| 40-50 | 7 | 45 | 1 | 7 |
| 50-60 | 4 | 55 | 2 | 8 |
| 60-70 | 3 | 65 | 3 | 9 |
| Total | 50 |  |  | -20 |



= 35 + × 10

 = 35 – 4 = 31

# **Merits and demerits of Arithmetic mean:**

# **Merits:**

1. It is rigidly defined.
2. It is easy to understand and easy to calculate.
3. If the number of items is sufficiently large, it is more accurate and more reliable.
4. It is a calculated value and is not based on its position in the series.
5. It is possible to calculate even if some of the details of the data are lacking.
6. Of all averages, it is affected least by fluctuations of sampling.
7. It provides a good basis for comparison.

# **Demerits:**

1. It cannot be obtained by inspection nor located through a frequency graph.
2. It cannot be in the study of qualitative phenomena not capable of numerical measurement i.e. Intelligence, beauty, honesty etc.,
3. It can ignore any single item only at the risk of losing its accuracy.
4. It is affected very much by extreme values.
5. It cannot be calculated for open-end classes.
6. It may lead to fallacious conclusions, if the details of the data from which it is computed are not given.

# **Harmonic mean (H.M.):**

Harmonic mean of a set of observations is defined as the reciprocal of the arithmetic average of the reciprocal of the given values. If x1,x2…..xn are n observations,

For a frequency distribution

**Example:**

***From the given data calculate H. M. 5, 10, 17, 24, and 30.***

**Solution:**

|  |  |
| --- | --- |
| X | 1*x* |
| 5 | 0.2000 |
| 10 | 0.1000 |
| 17 | 0.0588 |
| 24 | 0.0417 |
| 30 | 0.0333 |
| Total | 0.4338 |

Hence,

= = 11.52

**Example:**

***The marks secured by some students of a class are given below. Calculate the harmonic mean.***

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Marks** | 20 | 21 | 22 | 23 | 24 | 25 |
| **Number of Students** | 4 | 2 | 7 | 1 | 3 | 1 |

**Solution:**

|  |  |  |  |
| --- | --- | --- | --- |
| Marks*X* | No of students f | 1*x* | (1/*x*) |
| 20 | 4 | 0.0500 | 0.2000 |
| 21 | 2 | 0.0476 | 0.0952 |
| 22 | 7 | 0.0454 | 0.3178 |
| 23 | 1 | 0.0435 | 0.0435 |
| 24 | 3 | 0.0417 | 0.1251 |
| 25 | 1 | 0.0400 | 0.0400 |
|  | 18 |  | 0.8216 |

Hence,

 **=**  = 21.91

**Geometric Mean (G.M.):**

The geometric mean of a series containing n observations is the nth root of the product of the values. If x1, x2…, xn are observations then

G.M. =

 = (*x1.x2.x3……xn*)(1/n)

Log G.M. = (log*x1 + logx2 + logx3 +……+ logxn*)

Log G.M. **=**

G.M. = Antilog

**Example:**

***Calculate the geometric mean (G.M.) of the following series of monthly income of a batch of families 180, 250, 490, 1400, 1050.***

**Solution:**

|  |  |
| --- | --- |
| **x** | **Log x** |
| 180 | 2.2553 |
| 250 | 2.3979 |
| 490 | 2.6902 |
| 1400 | 3.1461 |
| 1050 | 3.0212 |
|  | 13.5107 |

G.M. = Antilog = Antilog = Antilog 2.70 = 503.6

**Example:**

**Calculate the average income per head from the data given below .Use geometric mean.**

|  |  |  |
| --- | --- | --- |
| Class of people | Number of families | Monthly income per head (SR) |
|  Landlords | 2 | 5000 |
| Cultivators | 100 | 400 |
| Landless – labours | 50 | 200 |
| Money – lenders | 4 | 3750 |
| Office Assistants | 6 | 3000 |
| Shop keepers | 8 | 750 |
| Carpenters | 6 | 600 |
| Weavers | 10 | 300 |

**Solution:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Class of people** | **Annual income ( SR) X** | **Number of families (f)** | **Log x** | **f logx** |
| Landlords | 5000 | 2 | 3.6990 | 7.398 |
| Cultivators | 400 | 100 | 2.6021 | 260.210 |
| Landless – labours | 200 | 50 | 2.3010 | 115.050 |
| Money – lenders | 3750 | 4 | 3.5740 | 14.296 |
| Office Assistants | 3000 | 6 | 3.4771 | 20.863 |
| Shop keepers | 750 | 8 | 2.8751 | 23.2008 |
| Carpenters | 600 | 6 | 2.7782 | 16.669 |
| Weavers | 300 | 10 | 2.4771 | 24.771 |
|  |  | 186 |  | 482.257 |

G.M. = Antilog

 = Antilog

 = Antilog (2.5928)

= SR 391.50

**Combined Mean:**

If the arithmetic averages and the number of items in two or more related groups are known, the combined or the composite mean of the entire group can be obtained by

Combined Mean, X =

**Example:**

Find the combined mean for the data given below:

n1 = 20; mean (x1) = 4; n2 = 30 and mean (x2) = 3

**Solution:**

Combined Mean, X = = = = = 3.4

**Positional Averages (Median and Mode):**

These averages are based on the position of the given observation in a series, arranged in an ascending or descending order. The magnitude or the size of the values does matter as was in the case of arithmetic mean. It is because of the basic difference that the median and mode are called the positional measures of an average.

**Median:**

The median is the middle value of a distribution i.e., median of a distribution is the value of the variable which divides it into two equal parts. It is the value of the variable such that the number of observations above it is equal to the number of observations below it.

**Ungrouped or Raw data:**

Arrange the given values in the increasing or decreasing order. If the numbers of values are odd, median is the middle value. If the numbers of values are even, median is the mean of middle two values.

By formula,

Median, Md = th item

# **When odd numbers of values are given:-**

# **Example:**

***Find median for the following data***

***25, 18, 27, 10, 8, 30, 42, 20, 53***

# **Solution:**

Arranging the data in the increasing order 8, 10, 18, 20, 25, 27, 30, 42, 53

Here, numbers of observations are odd (N= 9)

Hence, Median, Md = th item = th item= th item

The middle value is the 5th item i.e., 25 is the median value.

**When even numbers of values are given:-**

# **Example:**

***Find median for the following data***

***5, 8, 12, 30, 18, 10, 2, 22***

# **Solution:**

Arranging the data in the increasing order 2, 5, 8, 10, 12, 18, 22, 30

Here median is the mean of the middle two items (ie) mean of (10, 12) i. e., = 11

**Example:**

***The following table represents the marks obtained by a batch of 10 students in certain class tests in statistics and Accountancy.***

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Serial No | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| Marks (Statistics) | 53 | 55 | 52 | 32 | 30 | 60 | 47 | 46 | 35 | 28 |
| Marks (Accountancy) | 57 | 45 | 24 | 31 | 25 | 84 | 43 | 80 | 32 | 72 |

**Solution:**

For such question, median is the most suitable measure of central tendency. The marks in the two subjects are first arranged in increasing order as follows:

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Serial No | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| Marks in Statistics | 28 | 30 | 32 | 35 | 46 | 47 | 52 | 53 | 55 | 60 |
| Marks in Accountancy | 24 | 25 | 31 | 32 | 43 | 45 | 57 | 72 | 80 | 84 |

Median value for Statistics = (Mean of 5th and 6th items) = = 46.5

Median value for Accountancy = (Mean of 5th and 6th items) = = 44

Therefore, the level of knowledge in Statistics is higher than that in Accountancy.

**Grouped Data:**

In a grouped distribution, values are associated with frequencies. Grouping can be in the form of a discrete frequency distribution or a continuous frequency distribution. Whatever may be the type of distribution, cumulative frequencies have to be calculated to know the total number of items.

**Discrete Series:**

*Step1:* Find cumulative frequencies.

*Step 2:* Find

*Step 3:* See in the cumulative frequencies the value just greater than

*Step4:* Then the corresponding value of x will be median.

**Example:**

***The following data are pertaining to the number of members in a family. Find median size of the family.***

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Number of members **x** | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
| Frequency F | 1 | 3 | 5 | 6 | 10 | 13 | 9 | 5 | 3 | 2 | 2 | 1 |

**Solution:**

|  |  |  |
| --- | --- | --- |
| **X** | **f** | **cf** |
| 1 | 1 | 1 |
| 2 | 3 | 4 |
| 3 | 5 | 9 |
| 4 | 6 | 15 |
| 5 | 10 | 25 |
| 6 | 13 | 38 |
| 7 | 9 | 47 |
| 8 | 5 | 52 |
| 9 | 3 | 55 |
| 10 | 2 | 57 |
| 11 | 2 | 59 |
| 12 | 1 | 60 |
| N= | 60 |  |

Median = Size of item = Size of item = 30.5th item.

The cumulative frequency just greater than 30.5 is 38 and the value of x corresponding to 38 is 6. Hence the median size is 6 members per family.

***Note:***

*It is an appropriate method because a fractional value given by mean does not indicate the average number of members in a family.*

**Continuous Series:**

The steps given below are followed for the calculation of median in continuous series.

Step1: Find cumulative frequencies.

Step 2: Find

Step3: See in the cumulative frequency the value first greater than Then the corresponding class interval is called the Median Class. Then apply the formula for Median,

Md = *l+ ×h*

Where,

*l* = lower limit of the median class

 Σfi = n = number of Observations

*f* = frequency of the median class

*h* = size of the median class (assuming class size to be equal)

*cf* = cumulative frequency of the class preceding the median class.

 N = Total frequency.

**Note:**

*If the class intervals are given in inclusive type convert them into exclusive type and call it as true class interval and consider lower limit in this.*

**Example:**

***The following table gives the frequency distribution of 325 workers of a factory, according to their average monthly income in a certain year.***

|  |  |
| --- | --- |
| Income group (in Rs) | Number of workers |
| Below 100 | 1 |
| 100-150 | 20 |
| 150-200 | 42 |
| 200-250 | 55 |
| 250-300 | 62 |
| 300-350 | 45 |
| 350-400 | 30 |
| 400-450 | 25 |
| 450-500 | 15 |
| 500-550 | 18 |
| 550-600 | 10 |
| 600 and above | 2 |
| N= | 325 |

***Calculate median income.***

**Solution:**

|  |  |  |
| --- | --- | --- |
| **Income group (Class-interval)** | **Number of workers (Frequency)** | **Cumulative frequency c.f** |
| Below 100 | 1 | 1 |
| 100-150 | 20 | 21 |
| 150-200 | 42 | 63 |
| 200-250 | 55 | 118 |
| 250-300 | 62 | 180 |
| 300-350 | 45 | 225 |
| 350-400 | 30 | 255 |
| 400-450 | 25 | 280 |
| 450-500 | 15 | 295 |
| 500-550 | 18 | 313 |
| 550-600 | 10 | 323 |
| 600 and above | 2 | 325 |
|  | 325 |  |

Here, = = 162.5

So, l = 250; n/2 = 162.5; cf = 118; f = 62 and h = 50

Median, Md = *l+ ×h*

 = 250+ ×50

 = 250 + 35.89 = 285.89

**Example:**

***Calculate median from the following data:***

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Class Interval** | 1. 4
 | 5- 9 | 10- 14 | 15- 19 | 20- 24 | 25- 29 | 30- 34 | 35- 39 |
| **Frequency** | 5 | 8 | 10 | 12 | 7 | 6 | 3 | 2 |

**Solution:**

Here, class intervals are in inclusive type so first we should convert it into exclusive type as done below:

|  |  |  |
| --- | --- | --- |
| **Class Interval** | **Frequency** | **Cumulative Frequency (cf)**  |
| 0.5- 4.5 | 5 | 5 |
| 4.5- 9.5 | 8 | 13 |
| 9.5- 14.5 | 10 | 23 |
| 14.5- 19.5 | 12 | 35 |
| 19.5- 24.5 | 7 | 42 |
| 24.5- 29.5 | 6 | 48 |
| 29.5- 34.5 | 3 | 51 |
| 34.5- 39.5 | 2 | 53 |
|  | N= 53 |  |

Here, = = 26.5

So, l = 14.5; n/2 = 26.5; cf = 23; f = 12 and h = 5

Median, Md = *l+ ×h*

 = 14.5+ ×5

 = 14.5 + 1.46

 = 15.96

**Example:**

***Following are the daily wages of workers in a textile. Find the median.***

|  |  |
| --- | --- |
| Wages (in SR.) | Number of workers |
| less than 100 | 5 |
| less than 200 | 12 |
| less than 300 | 20 |
| less than 400 | 32 |
| less than 500 | 40 |
| less than 600 | 45 |
| less than 700 | 52 |
| less than 800 | 60 |
| less than 900 | 68 |
| less than 1000 | 75 |

**Solution:**

We are given upper limit and less than cumulative frequencies. First find the class-intervals and the frequencies. Since the values are increasing by 100, hence the width of the class interval equal to 100.

|  |  |  |
| --- | --- | --- |
| **Class Interval** | **f** | **c.f** |
| 0-100 |  5 | 5 |
| 100-200 | 7 | 12 |
| 200-300 | 8 | 20 |
| 300- 400 | 12 | 32 |
| 400-500 | 8 | 40 |
| 500-600 | 5 | 45 |
| 600-700 | 7 | 52 |
| 700-800 | 8 | 60 |
| 800-900 | 8 | 68 |
| 900-1000 | 7 | 75 |
|  | N= 75 |  |

Here, = = 37.5

So, l = 400; n/2 = 37.5; cf = 32; f = 8 and h = 100

Median, Md = *l+ ×h*

 = 400+ ×100

 = 400 + 68.75

 = 468.75

**Example: *Find median for the data given below.***

|  |  |
| --- | --- |
| **Marks** | **Number of students** |
| Greater than 10 | 70 |
| Greater than 20 | 62 |
| Greater than 30 | 50 |
| Greater than 40 | 38 |
| Greater than 50 | 30 |
| Greater than 60 | 24 |
| Greater than 70 | 17 |
| Greater than 80 | 9 |
| Greater than 90 | 4 |

**Solution:**

Here we are given lower limit and more than cumulative frequencies.

|  |  |  |  |
| --- | --- | --- | --- |
| Class interval | f | More than c.f | Less than c.f |
| 10-20 | 8 | 70 | 8 |
| 20-30 | 12 | 62 | 20 |
| 30-40 | 12 | 50 | 32 |
| 40-50 | 8 | 38 | 40 |
| 50-60 | 6 | 30 | 46 |
| 60-70 | 7 | 24 | 53 |
| 70-80 | 8 | 17 | 61 |
| 80-90 | 5 | 9 | 66 |
| 90-100 | 4 | 4 | 70 |
|  | 70 |  |  |

Here, = = 35

So, l = 40; n/2 = 35; cf = 32; f = 38 and h = 10

Median, Md = *l+ ×h*

 = 40+ ×10 = 40 + 3.75 = 43.75

**Example:**

***Compute median for the following data.***

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Mid-Value | 5 | 15 | 25 | 35 | 45 | 55 | 65 | 75 |
| Frequency | 7 | 10 | 15 | 17 | 8 | 4 | 6 | 7 |

**Solution:**

Here values in multiples of 10, so width of the class interval is 10.

|  |  |  |  |
| --- | --- | --- | --- |
| Mid x | C.I | f | c.f |
| 5 | 0-10 | 7 | 7 |
| 15 | 10-20 | 10 | 17 |
| 25 | 20-30 | 15 | 32 |
| 35 | 30-40 | 17 | 49 |
| 45 | 40-50 | 8 | 57 |
| 55 | 50-60 | 4 | 61 |
| 65 | 60-70 | 6 | 67 |
| 75 | 70-80 | 7 | 74 |
|  |  | N= 74 |  |

Here, = = 37

So, l = 30; n/2 = 37; cf = 32; f = 17 and h = 10

Median, Md = *l+ ×h*

 = 30+ ×10

 = 30 + 2.94

 = 32.94

**Quartiles:**

The quartiles divide the distribution in four parts. There are three quartiles. The second quartile divides the distribution into two halves and therefore is the same as the median. The first (lower) quartile (Q1) marks off the first one-fourth, the third (upper) quartile (Q3) marks off the three-fourth.

**Raw or ungrouped data:**

First arrange the given data in the increasing order and use the formula for Q1 and Q3 then quartile deviation, Q.D. is given by

Q.D. =

Where, Q1 = )th item and Q3 = )th item

**Example:**

Compute quartiles for the data given below 25,18, 30, 8, 15, 5, 10, 35, 40, 45

**Solution:**

5, 8, 10, 15, 18, 25, 30, 35, 40, 45

 Q1 = )th item

= )th item

= (2.75) th item.

= 2nd item + ) (3rd item- 2nd item)

= 8 + ) (10 - 8)

= 8 + )×2

= 9.5

 Q3 = )th item

= 3(2.75) th item.

= 8.25th item

= 8th item + )(9th item- 8th item)

= 35 +) (40 - 35)

= 35 + 1.25

= 36.25

**Discrete Series:**

*Step1:* Find cumulative frequencies

*Step2:* Find )

*Step3:* See in the cumulative frequencies, the value just greater than )then the corresponding value of *x* is Q1

*Step4:* Find 3)

*Step5:* See in the cumulative frequencies, the value just greater than 3) then the corresponding value of x is Q3.

**Example:**

***Compute quartiles for the data given below.***

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| X | 5 | 8 | 12 | 15 | 19 | 24 | 30 |
| f | 4 | 3 | 2 | 4 | 5 | 2 | 4 |

**Solution:**

|  |  |  |  |
| --- | --- | --- | --- |
| **x** | **f** | **c.f** |  |
|  |
| 5 | 4 | 4 |
| 8 | 3 | 7 |
| 12 | 2 | 9 |
| 15 | 4 | 13 |
| 19 | 5 | 18 |
| 24 | 2 | 20 |
| 30 | 4 | 24 |
| Total | 24 |  |

 Q1 = )th item

= )th item

= )th item

= 6.25th = 8

 Q3 = )th item

=  th item

= 18.75th item = 24

**Continuous Series:**

Step1: Find cumulative frequencies;

Step2: Find)

Step3: See in the cumulative frequencies, the value just greater), then the corresponding class interval is called first quartile class.

Step4: Find 3), See in the cumulative frequencies the value just greater than 3) then the corresponding class interval is called 3rd quartile class. Then apply the respective formulae

Q1 = *l1+ ×h1*

Q3 = *l3+ ×h3*

Where *l*1 = lower limit of the first quartile class

*f*1 = frequency of the first quartile class

*h*1 = width of the first quartile class

*cf*1 = c.f. preceding the first quartile class

*l*3 = 1ower limit of the 3rd quartile class

*f*3 = frequency of the 3rd quartile class

h3 = width of the 3rd quartile class

*cf*3 = c.f. preceding the 3rd quartile class

**Example:**

***The following series relates to the marks secured by students in an examination.***

|  |  |
| --- | --- |
| **Marks** | **No. of students** |
| 0-10 | 11 |
| 10-20 | 18 |
| 20-30 | 25 |
| 30-40 | 28 |
| 40-50 | 30 |
| 50-60 | 33 |
| 60-70 | 22 |
| 70-80 | 15 |
| 80-90 | 12 |
| 90-100 | 10 |

***Find the quartiles.***

**Solution:**

|  |  |  |
| --- | --- | --- |
| **C.I.** | **f** | **cf** |
| 0-10 | 11 | 11 |
| 10-20 | 18 | 29 |
| 20-30 | 25 | 54 |
| 30-40 | 28 | 82 |
| 40-50 | 30 | 112 |
| 50-60 | 33 | 145 |
| 60-70 | 22 | 167 |
| 70-80 | 15 | 182 |
| 80-90 | 12 | 194 |
| 90-100 | 10 | 204 |
|  | 204 |  |

Here, ) = ) = 51 and 3) = 3× 51 = 153

Q1 = *l1+ ×h1*

 = 20+ ×10 = 28.8

Q3 = *l3+ ×h3*

= 60+ ×10 = 63.63

**Deciles:**

These are the values, which divide the total number of observation into 10 equal parts. These are 9 deciles D1, D2…D9. These are all called first decile, second decile…etc.

**Deciles for Raw data or ungrouped data**

**Example: *Compute D5 for the data given below 5, 24, 36, 12, 20, 8***.

**Solution:**

Arranging the given values in the increasing order 5, 8, 12, 20, 24, 36

D5 = 5 observation

 = 5observation

 = observation

 = 3rd item +( [ 4th item – 3rd item]

 = 12 + ( [ 20 – 12]

 = 16.

**Deciles for Grouped data:**

Same as quartile.

# **Percentiles:**

The percentile values divide the distribution into 100 parts each containing 1 percent of the cases. The percentile (Pk) is that value of the variable up to which lie exactly k% of the total number of observations.

# **Relationship:**

P25 = Q1; P50 = D5 = Q2 = Median and P75 = Q3

**Percentile for Raw Data or Ungrouped Data:**

**Example: *Calculate P15 for the data given below: 5, 24, 36 , 12 , 20 , 8.***

**Solution:**

Arranging the given values in the increasing order. 5, 8, 12, 20, 24, 36

P15 = 15 item

= 15 item

= (1.05)th item

= 1st item + 0.5(2nd item – 1st item)

= 5 + 0.5(8 - 5)

= 5.15

**Percentile for Grouped Data:**

# **Example: *Find P53 for the following frequency distribution.***

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Class Interval | 0-5 | 5-10 | 10-15 | 15-20 | 20-25 | 25-30 | 30-35 | 35-40 |
| Frequency | 5 | 8 | 12 | 16 | 20 | 10 | 4 | 3 |

# **Solution:**

|  |  |  |
| --- | --- | --- |
| **Class Interval** | **Frequency** | **cf** |
| 0-5 | 5 | 5 |
| 5-10 | 8 | 13 |
| 10-15 | 12 | 25 |
| 15-20 | 16 | 41 |
| 20-25 | 20 | 61 |
| 25-30 | 10 | 71 |
| 30-35 | 4 | 75 |
| 35-40 | 3 | 78 |
| Total | 78 |  |

P53 = *l+ ×h* = *20+ ×5 =* 20.085

**Mode:**

The mode or modal value of a distribution is that value of the variable for which the frequency is the maximum. It refers to that value in a distribution which occurs most frequently. It shows the center of concentration of the frequency in around a given value. Therefore, where the purpose is to know the point of the highest concentration it is preferred. It is, thus, a positional measure.

Its importance is very great in marketing studies where a manager is interested in knowing about the size, which has the highest concentration of items. For example, in placing an order for shoes or ready-made garments the modal size helps because these sizes and other sizes around in common demand.

# **Computation of the mode:**

# **Ungrouped or Raw Data:**

For ungrouped data or a series of individual observations, mode is often found by mere inspection.

# **Example:**

2, 7, 10, 15, 10, 17, 8, 10, 2

Mode = M0 = 10

In some cases the mode may be absent while in some cases there may be more than one mode

# **Example:** (1) 12, 10, 15, 24, 30 (no mode)

 (2) 7, 10, 15, 12, 7, 14, 24, 10, 7, 20, 10

the modes are 7 and 10

# **Grouped Data:**

For Discrete distribution, see the highest frequency and corresponding value of X is mode.

**Continuous distribution:**

See the highest frequency then the corresponding value of class interval is called the modal class. Then apply the following formula:

Mode, Mo = *l+ ×h*

Where, *l* = lower limit of the modal class

*f* = frequency of the modal class

f1 = frequency of the class preceding the modal class

f2 = frequency of the class following the modal class.

*h* = size of the modal class

***Remarks:***

1. *If (2f1-f0-f2) comes out to be zero, then mode is obtained by the following formula taking absolute differences within vertical lines;*
2. *Mode, Mo = l+ ×h*
3. *If mode lies in the first class interval, then f is taken as zero.*
4. *The computation of mode poses no problem in distributions with open-end classes, unless the modal value lies in the open-end class.*

# **Example: *Calculate mode for the following:***

|  |  |
| --- | --- |
| **C- I** | **f** |
| 0-50 | 5 |
| 50-100 | 14 |
| 100-150 | 40 |
| 150-200 | 91 |
| 200-250 | 150 |
| 250-300 | 87 |
| 300-350 | 60 |
| 350-400 | 38 |
| 400 and above | 15 |

# **Solution:**

The highest frequency is 150 and corresponding class interval is 200 – 250, which is the modal class.

Here, *l* = 200; *f* = 150; *f1*= 91; *f2* = 87 and *h* = 50

Mode, Mo = *l+ ×h*

 *=* 200+ ×50

*=* 200+ 24.18

*= 224.18*

**Determination of Modal class:**

For a frequency distribution modal class corresponds to the maximum frequency. But in any one (or more) of the following cases-

1. If the maximum frequency is repeated
2. If the maximum frequency occurs in the beginning or at the end of the distribution
3. If there are irregularities in the distribution, the modal class is determined by the method of grouping.

**Steps for Calculation:**

We prepare a grouping table with 6 columns

1. In column I, we write down the given frequencies;
2. Column II is obtained by combining the frequencies two by two;
3. Leave the 1st frequency and combine the remaining frequencies two by two and write in column III;
4. Column IV is obtained by combining the frequencies three by three;
5. Leave the 1st frequency and combine the remaining frequencies three by three and write in column V;
6. Leave the 1st and 2nd frequencies and combine the remaining frequencies three by three and write in column VI.

Mark the highest frequency in each column. Then form an analysis table to find the modal class. After finding the modal class, use the formula to calculate the modal value.

# **Example:**

***Calculate mode for the following frequency distribution.***

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Class interval | 1. 5
 | 5- 10 | 10- 15 | 15- 20 | 20- 25 | 25- 30 | 30- 35 | 35- 40 |
| Frequency | 9 | 12 | 15 | 16 | 17 | 15 | 10 | 13 |

# **Grouping Table**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| C I | f | 2 | 3 | 4 | 5 | 6 |
| 0- 55-1010-1515-2020-2525-3030-3535-40 | 9121516**17**151013 | 2131**32**23 | 27**33**25 | 36**48** | **43**42 | **48**38 |

**Analysis Table**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Columns** | **0-5** | **5-10** | **10-15** | **15-20** | **20-25** | **25-30** | **30-35** | **35-40** |
| 1 |  | 1 | 1 | 1 | 1 | 1 |  |  |
| 2 | 1 |
| 3 | 1 |  |
| 4 | 1 | 1 | 1 |
| 5 | 1 |  |
| 6 | 1 | 1 | 1 |
| Total |  | 1 | 2 | 4 | 5 | 2 |  |  |

The maximum occurred corresponding to 20-25, and hence it is the modal class.

Here, *l* = 20; *f* = 16; *f1*= 15; *f2* = 17 and *h* = 5

Mode, Mo = *l+ ×h*

 *=* 20+ ×5

*=* 20+ ×5

*So, Mode, Mo = l+ ×h*

 = 20+ ×5

 *=* 20*+ =* 20+ 1.67 *=* 21.67

# **Empirical Relationship between Averages**

In a symmetrical distribution the three simple averages mean = median = mode. For a moderately asymmetrical distribution, the relationship between them are brought by Prof. Karl Pearson as

*Mode = 3 Median - 2 Mean*

# **Example:**

***If the mean and median of a moderately asymmetrical series are 26.8 and 27.9 respectively, what would be its most probable mode?***

# **Solution:**

Using the empirical formula Mode = 3 median 2 mean

= 3 27.9 2 26.8

= 30.1

**Example:**

***In a moderately asymmetrical distribution the values of mode and mean are 32.1 and 35.4 respectively. Find the median value.***

# **Solution:**

Using empirical Formula

Median =

 = = 34.3

**Measures of Dispersion – Skewness and Kurtosis**

# **Introduction:**

The measures of central tendency serve to locate the center of the distribution, but they do not reveal how the items are spread out on either side of the center. This characteristic of a frequency distribution is commonly referred to as dispersion. In a series all the items are not equal. There is difference or variation among the values. The degree of variation is evaluated by various measures of dispersion. Small dispersion indicates high uniformity of the items, while large dispersion indicates less uniformity. For example consider the following marks of two students.

|  |  |
| --- | --- |
| Student I | Student II |
| 68 | 85 |
| 75 | 90 |
| 65 | 80 |
| 67 | 25 |
| 70 | 65 |

Both have got a total of 345 and an average of 69 each. The fact is that the second student has failed in one paper. When the averages alone are considered, the two students are equal. But first student has less variation than second student. Less variation is a desirable characteristic.

# **Characteristics of a good measure of dispersion:**

An ideal measure of dispersion is expected to possess the following properties

* + 1. It should be rigidly defined
		2. It should be based on all the items.
		3. It should not be unduly affected by extreme items.
		4. It should lend itself for algebraic manipulation.
		5. It should be simple to understand and easy to calculate

# **Absolute and Relative Measures:**

There are two kinds of measures of dispersion, namely (1).Absolute measure of dispersion and (2).Relative measure of dispersion.

Absolute measure of dispersion indicates the amount of variation in a set of values in terms of units of observations. For example, when rainfalls on different days are available in mm, any absolute measure of dispersion gives the variation in rainfall in mm. On the other hand relative measures of dispersion are free from the units of measurements of the observations. They are pure numbers. They are used to compare the variation in two or more sets, which are having different units of measurements of observations.

The various absolute and relative measures of dispersion are listed below.

|  |  |
| --- | --- |
| Absolute measure | Relative measure |
| Range | Co-efficient of Range |
| Quartile deviation | Co-efficient of Quartile deviation |
| Mean deviation | Co-efficient of Mean deviation |
| Standard deviation | Co-efficient of variation |

# **Range and coefficient of Range:**

**Range:**

This is the simplest possible measure of dispersion and is defined as the difference between the largest and smallest values of the variable.

In symbols, Range = L – S.

Where L = Largest value. S = Smallest value.

In individual observations and discrete series, L and S are easily identified. In continuous series, the following two methods are followed,

**Method 1:**

L = Upper boundary of the highest class

S = Lower boundary of the lowest class

**Method 2:**

L = Mid value of the highest class.

S = Mid value of the lowest class.

**Co-efficient of Range:**

Co-efficient of Range =

**Example:**

Find the value of range and its co-efficient for the following data.

7, 9, 6, 8, 11, 10, 4

**Solution:**

L=11, S = 4.

 Range = L – S

 = 11- 4 = 7

Co-efficient of Range = = = = 0.4667

**Example: *Calculate range and its co efficient from the following distribution.***

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Size:** | 60- 63 | 63- 66 | 66- 69 | 69- 72 | 72- 75 |
| **Number:** | 5 | 18 | 42 | 27 | 8 |

**Solution:**

L = Upper boundary of the highest class = 75

 S = Lower boundary of the lowest class = 60

Range = L – S = 75 – 60 = 15

 Co-efficient of Range = = = = 0.1111

**Quartile Deviation and Co efficient of Quartile Deviation:**

**Quartile Deviation (Q.D.):**

**Definition:** Quartile Deviation is half of the difference between the first and third quartiles. Hence, it is called Semi Inter Quartile Range.

In symbol, Q.D. =

Among the quartiles Q1, Q2 and Q3, the range Q3- Q1 is called inter quartile range and, semi inter quartile range.

**Co-efficient of Quartile Deviation:**

Co-efficient of Quartile Deviation =

**Example:** Find the Quartile Deviation for the following data:

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

# **Solution:**

Arrange the given values in ascending order.

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

Position of Q1 is = = 2.75th item

Q1 = 2nd value + 0.75 (3rd value – 2nd value)

= 391 + 0.75 (407 – 391)

= 391 + 0.75 16

= 391 + 12

= 403

Position of Q3 is = 3×2.75 = 8.25th item

Q3 = 8th value + 0.25 (9th value – 8th value)

= 777 + 0.25 (1490 – 777)

= 777 + 0.25 (713)

= 777 + 178.25 = 955.25

Q.D. = = = = 276.125

**Example:**

***Weekly wages of labors are given below. Calculate Q.D. and Coefficient of Q.D.***

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Weekly Wage (Rs.) | :100 | 200 | 400 | 500 | 600 |
| No. of Weeks | : 5 | 8 | 21 | 12 | 6 |

**Solution:**

|  |  |  |
| --- | --- | --- |
| Weekly Wage (Rs.) | No. of Weeks | Cum. No. of Weeks |
| 100 | 5 | 5 |
| 200 | 8 | 13 |
| 400 | 21 | 34 |
| 500 | 12 | 46 |
| 600 | 6 | 52 |
| Total | N=52 |  |

Position of Q1 is = = 13.25th item

 Q1 = 13th value + 0.25 (14th Value – 13th value)

= 13th value + 0.25 (400 – 200)

= 200 + 0.25 (400 – 200)

= 200 + 0.25 (200)

= 200 + 50 = 250

Position of Q3 is = 3×13.25 = 39.25th item

 Q3 = 39th value + 0.75 (40th value – 39th value)

= 500 + 0.75 (500 – 500)

= 500 + 0.75 0

= 500

Q.D. = = = = 125

Co-efficient of Quartile Deviation = = = = 0.33

**Example:**

***For the date given below, give the quartile deviation and coefficient of quartile deviation.***

|  |  |  |  |
| --- | --- | --- | --- |
| X : 351 – 500 501 – 650 | 651 – 800 | 801–950 | 951–1100 |
| f : 48 189 | 88 | 4 | 28 |

**Solution:**

|  |  |  |  |
| --- | --- | --- | --- |
| x | f | True class Intervals | Cumulative frequency |
| 351- 500 | 48 | 350.5- 500.5 | 48 |
| 501- 650 | 189 | 500.5- 650.5 | 237 |
| 651- 800 | 88 | 650.5- 800.5 | 325 |
| 801- 950 | 47 | 800.5- 950.5 | 372 |
| 951- 1100 | 28 | 950.5- 1100.5 | 400 |
| Total | N = 400 |  |  |

Since, N/4 = 100

Therefore, Q1 Class is 500.5- 650.5

Hence, *l1= 500.5; n/4 = 100; cf1 =48; f1= 189; h1= 150*

Q1 = *l1+ ×h1*

 = 500.5+ ×150 = 541.77

Now, for Q3

= 3*×*100= 300

Hence, Q3 Class is 650.5- 800.5

*l3= 650.5; = 300; cf3= 237; f3= 88; h3= 150*

Q3 = *l3+ ×h3*

= 650.5+ ×150 = 757.89

Q.D. = = = = 108.06

Co-efficient of Quartile Deviation = = = = 0.1663

**Mean Deviation and Coefficient of Mean Deviation:**

**Mean Deviation:** The range and quartile deviation are not based on all observations. They are positional measures of dispersion. They do not show any scatter of the observations from an average. The mean deviation is measure of dispersion based on all items in a distribution.

**Definition:**

Mean deviation is the arithmetic mean of the deviations of a series computed from any measure of central tendency; i.e., the mean, median or mode, all the deviations are taken as positive i.e., signs are ignored.

We usually compute mean deviation about any one of the three averages mean, median or mode. Sometimes mode may be ill defined and as such mean deviation is computed from mean and median. Median is preferred as a choice between mean and median. But in general practice and due to wide applications of mean, the mean deviation is generally computed from mean. M.D can be used to denote mean deviation.

**Coefficient of mean deviation:**

Mean deviation calculated by any measure of central tendency is an absolute measure. For the purpose of comparing variation among different series, a relative mean deviation is required. The relative mean deviation is obtained by dividing the mean deviation by the average used for calculating mean deviation.

Coefficient of Mean Deviation =

*If the result is desired in percentage*,

The coefficient of mean deviation = ×100

**Computation of mean deviation – Individual Series:**

* + - 1. Calculate the average mean, median or mode of the series.
			2. Take the deviations of items from average ignoring signs and denote these deviations by |D|.
			3. Compute the total of these deviations, i.e., |D|
			4. Divide this total obtained by the number of items.

Symbolically,

M.D. =

**Example:**

***Calculate mean deviation from mean and median for the following data:***

***100, 150, 200, 250, 360, 490, 500, 600, and 671.***

***Also calculate co- efficient of M.D.***

**Solution:**

Mean = = = = 369

Now arrange the data in ascending order

100, 150, 200, 250, 360, 490, 500, 600, 671

Median, Md = Value of ()th item = ()th item = 5th item = 360

|  |  |  |
| --- | --- | --- |
| **X** | **D ⃒X Mean⃒** | **⃒D⃒ ⃒x Md⃒** |
| 100 | 269 | 260 |
| 150 | 219 | 210 |
| 200 | 169 | 160 |
| 250 | 119 | 110 |
| 360 | 9 | 0 |
| 490 | 121 | 130 |
| 500 | 131 | 140 |
| 600 | 231 | 240 |
| 671 | 302 | 311 |
| 3321 | 1570 | 1561 |

M.D. from mean = = = 174.44

Co-efficient of M.D. = = = 0.47

M.D. from median = = = 173.44

Co-efficient of M.D. = = = 0.48

**Mean Deviation- Discrete Series:**

 M.D. =

**Example:**

***Compute Mean deviation from mean and median from the following data:***

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Height in cms | 158 | 159 | 160 | 161 | 162 | 163 | 164 | 165 | 166 |
| No. of persons | 15 | 20 | 32 | 35 | 33 | 22 | 20 | 10 | 8 |

***Also compute coefficient of mean deviation.***

**Solution:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Height X | No. of persons f | d= x- A A =162 | fd | |D| =|X- mean| | f|D| |
| 158 | 15 | - 4 | - 60 | 3.51 | 52.65 |
| 159 | 20 | - 3 | - 60 | 2.51 | 50.20 |
| 160 | 32 | - 2 | - 64 | 1.51 | 48.32 |
| 161 | 35 | - 1 | - 35 | 0.51 | 17.85 |
| **162** | 33 | 0 | 0 | 0.49 | 16.17 |
| 163 | 22 | 1 | 22 | 1.49 | 32.78 |
| 164 | 20 | 2 | 40 | 2.49 | 49.80 |
| 165 | 10 | 3 | 30 | 3.49 | 34.90 |
| 166 | 8 | 4 | 32 | 4.49 | 35.92 |
|  | 195 |  | - 95 |  | 338.59 |

Mean = = = 162 – 0.49 = 161.51

M.D. = = = 1.74

 Coefficient of M.D. = = = 0.0108

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Height x | No. of persons f | c.f. | D =⃒X Median⃒ | f ⃒D⃒ |
| 158 | 15 | 15 | 3 | 45 |
| 159 | 20 | 35 | 2 | 40 |
| 160 | 32 | 67 | 1 | 32 |
| 161 | 35 | 102 | 0 | 0 |
| 162 | 33 | 135 | 1 | 33 |
| 163 | 22 | 157 | 2 | 44 |
| 164 | 20 | 177 | 3 | 60 |
| 165 | 10 | 187 | 4 | 40 |
| 166 | 8 | 195 | 5 | 40 |
|  | 195 |  |  | 334 |

 Median = Size of ()th item = ()th item = (96)th item = 161

M.D. = = = 1.71

Coefficient of M.D. = = = 0.0106

**Mean Deviation-Continuous Series:**

The method of calculating mean deviation in a continuous series same as the discrete series. In continuous series we have to find out the mid points of the various classes and take deviation of these points from the average selected. Thus

M.D. =

**Example:**

***Find out the mean deviation from mean and median from the following series.***

|  |  |
| --- | --- |
| **Age in years** | **No of persons** |
| 0-10 | 20 |
| 10-20 | 25 |
| 20-30 | 32 |
| 30-40 | 40 |
| 40-50 | 42 |
| 50-60 | 35 |
| 60-70 | 10 |
| 70-80 | 8 |

***Also compute co-efficient of mean deviation.***

**Solution:**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| X | m | f | d = m Ac(A=35,C=10) | fd | D  ⃒m  x⃒ | f ⃒D⃒ |
| 0-10 | 5 | 20 | -3 | -60 | 31.5 | 630.0 |
| 10-20 | 15 | 25 | -2 | -50 | 21.5 | 537.5 |
| 20-30 | 25 | 32 | -1 | -32 | 11.5 | 368.0 |
| 30-40 | **35** | 40 | 0 | 0 | 1.5 | 60.0 |
| 40-50 | 45 | 42 | 1 | 42 | 8.5 | 357.0 |
| 50-60 | 55 | 35 | 2 | 70 | 18.5 | 647.5 |
| 60-70 | 65 | 10 | 3 | 30 | 28.5 | 285.0 |
| 70-80 | 75 | 8 | 4 | 32 | 38.5 | 308.0 |
|  |  | 212 |  | 32 |  | 3193.0 |

Mean =

 = = 36.5

M.D. = = = 15.06

**Calculation of median and M.D. from median:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| X | m | f | c.f | |D| = |m-Md| | f |D| |
| 0-10 | 5 | 20 | 20 | 32.25 | 645.00 |
| 10-20 | 15 | 25 | 45 | 22.25 | 556.25 |
| 20-30 | 25 | 32 | 77 | 12.25 | 392.00 |
| 30-40 | 35 | 40 | 117 | 2.25 | 90.00 |
| 40-50 | 45 | 42 | 159 | 7.75 | 325.50 |
| 50-60 | 55 | 35 | 194 | 17.75 | 621.25 |
| 60-70 | 65 | 10 | 204 | 27.75 | 277.50 |
| 70-80 | 75 | 8 | 212 | 37.75 | 302.00 |
| N= 212 | Total | 3209.50 |

 = = 106

*l= 30; cf= 77; f= 40; h= 10*

Median, Md = *l+ ×h=* 30+ ×10= 37.25

M.D. = = = 15.14

Coefficient of M.D. = = = 0.41

# **Merits and Demerits of M.D:**

# **Merits:**

1. It is simple to understand and easy to compute.
2. It is rigidly defined.
3. It is based on all items of the series.
4. It is not much affected by the fluctuations of sampling.
5. It is less affected by the extreme items.
6. It is flexible, because it can be calculated from any average.
7. It is better measure of comparison.

# **Demerits:**

1. It is not a very accurate measure of dispersion.
2. It is not suitable for further mathematical calculation.
3. It is rarely used. It is not as popular as standard deviation.
4. Algebraic positive and negative signs are ignored. It is mathematically unsound and illogical.

# **Standard Deviation and Coefficient of variation:**

# **Standard Deviation:**

Karl Pearson introduced the concept of standard deviation in 1893. It is the most important measure of dispersion and is widely used in many statistical formulae. Standard deviation is also called Root-Mean Square Deviation. The reason is that it is the square–root of the mean of the squared deviation from the arithmetic mean. It provides accurate result. Square of standard deviation is called Variance.

# **Definition:**

It is defined as the positive square-root of the arithmetic mean of the Square of the deviations of the given observation from their arithmetic mean. The standard deviation is denoted by the Greek letter  (sigma).

# **Calculation of Standard deviation-Individual Series:**

There are two methods of calculating Standard deviation in an individual series.

* + - 1. Deviations taken from Actual mean; and
			2. Deviation taken from Assumed mean
1. **Deviation taken from Actual mean:**

This method is adopted when the mean is a whole number.

Steps:

1. Find out the actual mean of the series ( *x* )
2. Find out the deviation of each value from the mean (*x = X- X*)
3. Square the deviations and take the total of squared deviations x2
4. Divide the total ( x2 ) by the number of observation, ()
5. The square root of () is standard deviation.

Thus, Standard Deviation (SD or ) = =

1. **Deviations taken from assumed mean:**

This method is adopted when the arithmetic mean is fractional value. Taking deviations from fractional value would be a very difficult and tedious task. To save time and labour, we apply short– cut method; deviations are taken from an assumed mean. The formula is:

 S.D. or 

Where d-stands for the deviation from assumed mean = (X-A)

# **Steps:**

* 1. Assume any one of the item in the series as an average (A)
	2. Find out the deviations from the assumed mean; i.e., X-A denoted by d and also the total of the deviations d
	3. Square the deviations; i.e., d2 and add up the squares of deviations, i.e, d2
	4. Then substitute the values in the following formula:

S.D. or 

**Example:**

Calculate the standard deviation from the following data. 14, 22, 9, 15, 20, 17, 12, 11

**Solution:**

*Deviations from actual mean.*

|  |  |  |
| --- | --- | --- |
| Values (X) |  |  |
| 14 | -1 | 1 |
| 22 | 7 | 49 |
| 9 | -6 | 36 |
| 15 | 0 | 0 |
| 20 | 5 | 25 |
| 17 | 2 | 4 |
| 12 | -3 | 9 |
| 11 | -4 | 16 |
| 120 |  | 140 |

Mean, (*X bar*) = = = 15

Thus, Standard Deviation (SD or ) = = = = = 4.18

**Example:**

The table below gives the marks obtained by 10 students in statistics. Calculate standard deviation.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Student Nos : 1 2 3 | 4 | 5 6 | 7 | 8 | 9 | 10 |
| Marks : 43 48 65 | 57 | 31 60 | 37 | 48 | 78 | 59 |

**Solution:** (Deviations from assumed mean)

|  |  |  |  |
| --- | --- | --- | --- |
| Nos. | Marks (x) | d=X-A (A=57) | d2 |
| 1 | 43 | -14 | 196 |
| 2 | 48 | -9 | 81 |
| 3 | 65 | 8 | 64 |
| 4 | 57 | 0 | 0 |
| 5 | 31 | -26 | 676 |
| 6 | 60 | 3 | 9 |
| 7 | 37 | -20 | 400 |
| 8 | 48 | -9 | 81 |
| 9 | 78 | 21 | 441 |
| 10 | 59 | 2 | 4 |
| n = 10 |  | d=-44 | d2 =1952 |

 S.D. or   = = = 13.26

# **Calculation of standard deviation:**

# **Discrete Series:**

There are three methods for calculating standard deviation in discrete series:

1. Actual mean methods: If the actual mean in fractions, the calculation takes lot of time and labour; and as such this method is rarely used in practice.
2. Assumed mean method: Here deviations are taken not from an actual mean but from an assumed mean. Also this method is used, if the given variable values are not in equal intervals.
3. Step-deviation method: If the variable values are in equal intervals, then we adopt this method.

**Example:**

***Calculate Standard deviation from the following data.***

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| X : | 20 | 22 | 25 | 31 | 35 | 40 | 42 | 45 |
| f : | 5 | 12 | 15 | 20 | 25 | 14 | 10 | 6 |

**Solution:**

Deviations from assumed mean

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| x | f | d = x –A (A = 31) | d2 | fd | fd2 |
| 20 | 5 | -11 | 121 | -55 | 605 |
| 22 | 12 | -9 | 81 | -108 | 972 |
| 25 | 15 | -6 | 36 | -90 | 540 |
| 31 | 20 | 0 | 0 | 0 | 0 |
| 35 | 25 | 4 | 16 | 100 | 400 |
| 40 | 14 | 9 | 81 | 126 | 1134 |
| 42 | 10 | 11 | 121 | 110 | 1210 |
| 45 | 6 | 14 | 196 | 84 | 1176 |
|  | N=107 |  |  | fd=167 | fd2= 6037 |

S.D. or   = = = 7.35

**Calculation of Standard Deviation –Continuous Series:**

In the continuous series the method of calculating standard deviation is almost the same as in a discrete series. But in a continuous series, mid-values of the class intervals are to be found out. The step- deviation method is widely used.

**Coefficient of Variation:**

The Standard deviation is an absolute measure of dispersion. It is expressed in terms of units in which the original figures are collected and stated. The standard deviation of heights of students cannot be compared with the standard deviation of weights of students, as both are expressed in different units, i.e heights in centimeter and weights in kilograms. Therefore the standard deviation must be converted into a relative measure of dispersion for the purpose of comparison. The relative measure is known as the coefficient of variation.

The coefficient of variation is obtained by dividing the standard deviation by the mean and multiplies it by 100. Symbolically,

Coefficient of Variation (CV) = =

If we want to compare the variability of two or more series, we can use C.V. The series or groups of data for which the C.V. is greater indicate that the group is more variable, less stable, less uniform, less consistent or less homogeneous. If the C.V. is less, it indicates that the group is less variable, more stable, more uniform, more consistent or more homogeneous.

**Example:**

***In two factories A and B located in the same industrial area, the average weekly wages (in SR) and the standard deviations are as follows:***

|  |  |  |  |
| --- | --- | --- | --- |
| **Factory** | **Average (x)** | **Standard Deviation (σ)** | **No. of workers** |
| **A B** | **34.5****28.5** | **5****4.5** | **476****524** |

1. ***Which factory A or B pays out a larger amount as weekly wages?***
2. ***Which factory A or B has greater variability in individual wages?***

**Solution:**

Given N1= 476; X1= 34.5 and σ1= 5

N2 = 524, X2 = 28.5, 2 = 4.5

1. Total wages paid by factory A

= 34.5 476

= SR16.422

Total wages paid by factory B

= 28.5 524

= SR.14,934.

Therefore factory A pays out larger amount as weekly wages.

1. C.V. of distribution of weekly wages of factory A and B are

CV (A) = = = 14.49

CV (B) = = = 15.79

Factory B has greater variability in individual wages, since C.V. of factory B is greater than C.V of factory A.

# **Example:**

***Prices of a particular commodity in five years in two cities are given below:***

|  |  |
| --- | --- |
| ***Price in city A*** | ***Price in city B*** |
| ***20*** | ***10*** |
| ***22*** | ***20*** |
| ***19*** | ***18*** |
| ***23*** | ***12*** |
| ***16*** | ***15*** |

***Which city has more stable prices?***

# **Solution:** *Actual mean method*

|  |  |
| --- | --- |
| City A | City B |
| Prices (X) | Deviations from X=20 dx | dx2 | Prices (Y) | Deviations from Y =15 dy | dy2 |
| 20 | 0 | 0 | 10 | -5 | 25 |
| 22 | 2 | 4 | 20 | 5 | 25 |
| 19 | -1 | 1 | 18 | 3 | 9 |
| 23 | 3 | 9 | 12 | -3 | 9 |
| 16 | -4 | 16 | 15 | 0 | 0 |
| x=100 | dx=0 | dx2=30 | y=75 | dy=0 | dy2 =68 |

**City A:**

Mean = = = 20;

SD ( = = = = 2.45

CV (City A) = = = = 12.25%

**City B:**

Mean = = = 15;

SD ( = = = = 3.69

CV (City B) = = = = 24.6%

Therefore, City A had more stable prices than City B, because the coefficient of variation is less in City A.

**Skewness**

**Meaning:**

Skewness means ‘lack of symmetry’. We study skewness to have an idea about the shape of the curve which we can draw with the help of the given data. If in a distribution mean = median = mode, then that distribution is known as symmetrical distribution. If in a distribution mean median mode, then it is not a symmetrical distribution and it is called a skewed distribution and such a distribution could either be positively skewed or negatively skewed.

**Symmetrical distribution:**

It is clear from the diagram that in a symmetrical distribution the values of mean, median and mode coincide. The spread of the frequencies is the same on both sides of the center point of the curve



**Positively skewed distribution:**

It is clear from the above diagram, in a positively skewed distribution, the value of the mean is maximum and that of the mode is least, the median lies in between the two. In the positively skewed distribution the frequencies are spread out over a greater range of values on the right hand side than they are on the left hand side.

**Negatively skewed distribution:**

It is clear from the above diagram, in a negatively skewed distribution, the value of the mode is maximum and that of the mean is least. The median lies in between the two. In the negatively skewed distribution the frequencies are spread out over a greater range of values on the left hand side than they are on the right hand side.

**Measures of skewness:**

The important measures of skewness are

1. Karl – Pearason’ s coefficient of skewness
2. Bowley’ s coefficient of skewness

**Karl – Pearson’s Coefficient of skewness:**

 According to Karl – Pearson, the absolute measure of skewness = mean – mode. This measure is not suitable for making valid comparison of the skewness in two or more distributions because the unit of measurement may be different in different series. To avoid this difficulty use relative measure of skewness called Karl – Pearson’ s coefficient of skewness given by

Karl – Pearson’s Coefficient Skewness =

In case of mode is ill – defined, the coefficient can be determined by the formula:

Karl – Pearson’s Coefficient Skewness =

# **Bowley’ s Coefficient of skewness:**

In Karl – Pearson’ s method of measuring skewness the whole of the series is needed. Prof. Bowley has suggested a formula based on relative position of quartiles. In a symmetrical distribution, the quartiles are equidistant from the value of the median; ie.,

Median – Q1 = Q3 – Median. But in a skewed distribution, the quartiles will not be equidistant from the median. Hence Bowley has suggested the following formula:

Bowley’ s Coefficient of skewness (sk) =

**Kurtosis**

The expression ‘ Kurtosis’ is used to describe the peakedness of a curve. The three measures – central tendency, dispersion and skewness describe the characteristics of frequency distributions. But these studies will not give us a clear picture of the characteristics of a distribution.

As far as the measurement of shape is concerned, we have two characteristics – skewness which refers to asymmetry of a series and kurtosis which measures the peakedness of a normal curve. All the frequency curves expose different degrees of flatness or peakedness. This characteristic of frequency curve is termed as kurtosis. Measure of kurtosis denote the shape of top of a frequency curve. Measure of kurtosis tell us the extent to which a distribution is more peaked or more flat topped than the normal curve, which is symmetrical and bell-shaped, is designated as Mesokurtic. If a curve is relatively more narrow and peaked at the top, it is designated as Leptokurtic. If the frequency curve is more flat than normal curve, it is designated as platykurtic.



# **Measure of Kurtosis:**

The measure of kurtosis of a frequency distribution based moments is denoted by 2 and is given by



If 2 =3, the distribution is said to be normal and the curve is mesokurtic.

If 2 >3, the distribution is said to be more peaked and the curve is leptokurtic.

If 2< 3, the distribution is said to be flat topped and the curve is platykurtic.

**Correlation**

# **Introduction:**

The term correlation is used by a common man without knowing that he is making use of the term correlation. For example when parents advice their children to work hard so that they may get good marks, they are correlating good marks with hard work.

The study related to the characteristics of only variable such as height, weight, ages, marks, wages, etc., is known as univariate analysis. The statistical Analysis related to the study of the relationship between two variables is known as Bi-Variate Analysis. Sometimes the variables may be inter-related. In health sciences we study the relationship between blood pressure and age, consumption level of some nutrient and weight gain, total income and medical expenditure, etc. The nature and strength of relationship may be examined by correlation and Regression analysis.

Thus Correlation refers to the relationship of two variables or more. (e-g) relation between height of father and son, yield and rainfall, wage and price index, share and debentures etc.

Correlation is statistical Analysis which measures and analyses the degree or extent to which the two variables fluctuate with reference to each other. The word relationship is important. It indicates that there is some connection between the variables. It measures the closeness of the relationship. Correlation does not indicate cause and effect relationship. Price and supply, income and expenditure are correlated.

# **Uses of correlation:**

1. It is used in physical and social sciences.
2. It is useful for economists to study the relationship between variables like price, quantity etc. Businessmen estimates costs, sales, price etc. using correlation.
3. It is helpful in measuring the degree of relationship between the variables like income and expenditure, price and supply, supply and demand etc.
4. Sampling error can be calculated.
5. It is the basis for the concept of regression.

# **Types of Correlation:**

Correlation is classified into various types. The most important ones are

1. Positive and negative.
2. Linear and non-linear.
3. Partial and total.
4. Simple and Multiple.

# **Positive and Negative Correlation:**

It depends upon the direction of change of the variables. If the two variables tend to move together in the same direction (ie) an increase in the value of one variable is accompanied by an increase in the value of the other, (or) a decrease in the value of one variable is accompanied by a decrease in the value of other, then the correlation is called positive or direct correlation. Price and supply, height and weight, yield and rainfall, are some examples of positive correlation.

If the two variables tend to move together in opposite directions so that increase (or) decrease in the value of one variable is accompanied by a decrease or increase in the value of the other variable, then the correlation is called negative (or) inverse correlation. Price and demand, yield of crop and price, are examples of negative correlation.

# **Linear and Non-linear correlation:**

If the ratio of change between the two variables is a constant then there will be linear correlation between them.

Consider the following.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| X | 2 | 4 | 6 | 8 | 10 | 12 |
| Y | 3 | 6 | 9 | 12 | 15 | 18 |

Here the ratio of change between the two variables is the same. If we plot these points on a graph we get a straight line.

If the amount of change in one variable does not bear a constant ratio of the amount of change in the other. Then the relation is called Curvi-linear (or) non-linear correlation. The graph will be a curve.

# **Simple and Multiple correlation:**

When we study only two variables, the relationship is simple correlation. For example, quantity of money and price level, demand and price. But in a multiple correlation we study more than two variables simultaneously. The relationship of price, demand and supply of a commodity are an example for multiple correlation.

# **Partial and total correlation:**

The study of two variables excluding some other variable is called **Partial correlation**. For example, we study price and demand eliminating supply side. In total correlation all facts are taken into account.

# **Computation of correlation:**

When there exists some relationship between two variables, we have to measure the degree of relationship. This measure is called the measure of correlation (or) correlation coefficient and it is denoted by ‘r’.

**Co-variation:**

The covariation between the variables x and y is defined as-

Cov (XY) =

 =

Where, X bar is the mean of X and Y bar is the mean of Y. x and y are deviations from its mean.

**Karl Pearson’s Coefficient of Correlation:**

 It is most widely used method in practice and it is known as Pearsonian Coefficient of Correlation. It is denoted by ‘r’. The formula for calculating ‘r’ is-

1. r = ; Where =
2. r =
3. r =

The third formula is easy to calculate, and it is not necessary to calculate the standard deviations of x and y series respectively.

**Properties of Correlation Coefficient:**

*Property 1:* Correlation coefficient lies between –1 and +1.

*Property 2:* ‘r’ is independent of change of origin and scale.

*Property 3:* It is a pure number independent of units of measurement.

*Property 4:* Independent variables are uncorrelated but the converse is not true.

*Property 5:* Correlation coefficient is the geometric mean of two regression coefficients.

*Property 6:* The correlation coefficient of x and y is symmetric. rxy = ryx.

# **Limitations:**

* 1. Correlation coefficient assumes linear relationship regardless of the assumption is correct or not.
	2. Extreme items of variables are being unduly operated on correlation coefficient.
	3. Existence of correlation does not necessarily indicate cause- effect relation.

# **Interpretation:**

The following rules helps in interpreting the value of ‘ r’ .

1. When r = 1, there is perfect +ve relationship between the variables.
2. When r = -1, there is perfect –ve relationship between the variables.
3. When r = 0, there is no relationship between the variables.
4. If the correlation is +1 or –1, it signifies that there is a high degree of correlation. (+ve or –ve) between the two variables.

If r is near to zero (ie) 0.1,-0.1, (or) 0.2 there is less correlation.

# **Example:**

Find Karl Pearson’s coefficient of correlation from the following data between height of father (x) and son (y).

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| X | 64 | 65 | 66 | 67 | 68 | 69 | 70 |
| Y | 66 | 67 | 65 | 68 | 70 | 68 | 72 |

Comment on the result.

**Solution:**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| X | Y | *x* = *X* *X**x* = x – 67 | *x2* |  *y* = *Y* *Y**y* = Y - 68 | *y2* | *xy* |
| 64 | 66 | -3 | 9 | -2 | 4 | 6 |
| 65 | 67 | -2 | 4 | -1 | 1 | 2 |
| 66 | 65 | -1 | 1 | -3 | 9 | 3 |
| 67 | 68 | 0 | 0 | 0 | 0 | 0 |
| 68 | 70 | 1 | 1 | 2 | 4 | 2 |
| 69 | 68 | 2 | 4 | 0 | 0 | 0 |
| 70 | 72 | 3 | 9 | 4 |  16 |  12 |
| 469 | 476 | 0 | 28 | 0 |  34 |  25 |

Mean of X = = = 67;

Mean of Y = = = 68.

Hence, Karl Pearson’s Coefficient of Correlation, r = = = = = 0.81. Since r = + 0.81, the variables are highly positively correlated i. e., tall fathers have tall sons.

**Example:**

***Calculate coefficient of correlation from the following data.***

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| X | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| Y | 9 | 8 | 10 | 12 | 11 | 13 | 14 | 16 | 15 |

**Example:**

***Calculate Pearson’s Coefficient of correlation.***

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| X | 45 | 55 | 56 | 58 | 60 | 65 | 68 | 70 | 75 | 80 | 85 |
| Y | 56 | 50 | 48 | 60 | 62 | 64 | 65 | 70 | 74 | 82 | 90 |

**Rank Correlation**

It is studied when no assumption about the parameters of the population is made. This method is based on ranks. It is useful to study the qualitative measure of attributes like honesty, colour, beauty, intelligence, character, morality etc. The individuals in the group can be arranged in order and there on, obtaining for each individual a number showing his/her rank in the group. This method was developed by Edward Spearman in 1904. It is defined as-

ρ =

Where, ρ (rho) = rank correlation coefficient;

 = sum of squares of differences between the pairs of ranks; and

 N = number of pairs of observations.

The value of ρ lies between –1 and +1. If ρ = +1, there is complete agreement in order of ranks and the direction of ranks is also same. If ρ = -1, then there is complete disagreement in order of ranks and they are in opposite directions.

Computation for tied observations: There may be two or more items having equal values. In such case the same rank is to be given. The ranking is said to be tied. In such circumstances an average rank is to be given to each individual item. For example if the value so is repeated twice at the 5th rank, the common rank to be assigned to each item is = = 5.5 which is the average of 5 and 6 given as 5.5, appeared twice.

If the ranks are tied, it is required to apply a correction factor which is (m3- m). A slight formula is used when there is more than one item having the same value. The formula is-

ρ =

Where m is the number of items whose ranks are common and should be repeated as many times as there are tied observations.

**Example:**

In a marketing survey the price of tea and coffee in a town based on quality was found as shown below. Could you find any relation between and tea and coffee price.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Price of tea | 88 | 90 | 95 | 70 | 60 | 75 | 50 |
| Price of coffee | 120 | 134 | 150 | 115 | 110 | 140 | 100 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Price of tea** | **Rank** | **Price of coffee** | **Rank** | **D** | **D2** |
| 88 | 3 | 120 | 4 | 1 | 1 |
| 90 | 2 | 134 | 3 | 1 | 1 |
| 95 | 1 | 150 | 1 | 0 | 0 |
| 70 | 5 | 115 | 5 | 0 | 0 |
| 60 | 6 | 110 | 6 | 0 | 0 |
| 75 | 4 | 140 | 2 | 2 | 4 |
| 50 | 7 | 100 | 7 | 0 | 0 |
|  |  |  |  |  | ∑D2 = 6 |

ρ = = = = 1- 0.1071 = 0.8929

The relation between price of tea and coffee is positive at 0.89. Based on quality the association between price of tea and price of coffee is highly positive.

**Example:**

***In an evaluation of answer script the following marks are awarded by the examiners.***

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ***1st*** | ***88*** | ***95*** | ***70*** | ***960*** | ***50*** | ***80*** | ***75*** | ***85*** |
| ***2nd*** | ***84*** | ***90*** | ***88*** | ***55*** | ***48*** | ***85*** | ***82*** | ***72*** |

***Do you agree the evaluation by the two examiners is fair?***

**Solution:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **x** | **R1** | **y** | **R2** | **D** | **D2** |
| 88 | 2 | 84 | 4 | 2 | 4 |
| 95 | 1 | 90 | 1 | 0 | 0 |
| 70 | 6 | 88 | 2 | 4 | 16 |
| 60 | 7 | 55 | 7 | 0 | 0 |
| 50 | 8 | 48 | 8 | 0 | 0 |
| 75 | 5 | 82 | 5 | 0 | 0 |
| 80 | 4 | 85 | 3 | 1 | 1 |
| 85 | 3 | 75 | 6 | 3 | 9 |
|  |  |  |  |  | 30 |

ρ = = = = 1- 0.357 = 0.643

ρ = 0.643 shows fair in awarding marks in the sense that uniformity has arisen in evaluating the answer scripts between the two examiners.

**Example:**

***Rank Correlation for tied observations. Following are the marks obtained by 10 students in a class in two tests.***

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ***Students*** | ***A*** | ***B*** | ***C*** | ***D*** | ***E*** | ***F*** | ***G*** | ***H*** | ***I*** | ***J*** |
| ***Test 1*** | ***70*** | ***68*** | ***67*** | ***55*** | ***60*** | ***60*** | ***75*** | ***63*** | ***60*** | ***72*** |
| ***Test 2*** | ***65*** | ***65*** | ***80*** | ***60*** | ***68*** | ***58*** | ***75*** | ***63*** | ***60*** | ***70*** |

***Calculate the rank correlation coefficient between the marks of two tests.***

**Solution:**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Student | Test 1 | R1 | Test 2 | R2 | D | D2 |
| A | 70 | 3 | 65 | 5.5 | -2.5 | 6.25 |
| B | 68 | 4 | 65 | 5.5 | -1.5 | 2.25 |
| C | 67 | 5 | 80 | 1.0 | 4.0 | 16.00 |
| D | 55 | 10 | 60 | 8.5 | 1.5 | 2.25 |
| E | 60 | 8 | 68 | 4.0 | 4.0 | 16.00 |
| F | 60 | 8 | 58 | 10.0 | -2.0 | 4.00 |
| G | 75 | 1 | 75 | 2.0 | -1.0 | 1.00 |
| H | 63 | 6 | 62 | 7.0 | -1.0 | 1.00 |
| I | 60 | 8 | 60 | 8.5 | 0.5 | 0.25 |
| J | 72 | 2 | 70 | 3.0 | -1.0 | 1.00 |
|  |  |  |  |  |  | ∑D2 = 50.00 |

60 is repeated 3 times in test 1. 60, 65 is repeated twice in test 2. m = 3; m = 2; m = 2.

ρ =

ρ =

ρ = = = 0.68

**Interpretation:** There is uniformity in the performance of students in the two tests.