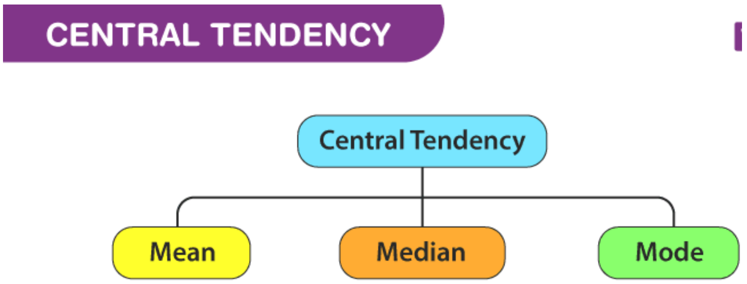
Unit II:

Measures of Central Tendency

The central tendency of the dataset can be found out using the three important measures namely mean, median and mode.



Introduction to Central Tendency: Central tendency is a statistical concept used to describe the center or typical value in a dataset. It provides a way to summarize and understand the central or most representative value within a set of data points. The three most common measures of central tendency are Mean, Median, and Mode, each serving a different purpose in data analysis.

1. Mean:

- Definition: The mean, also known as the average, is the sum of all values in a dataset divided by the number of data points. It represents the "typical" value when all values are added together and then divided equally among the data points.
- Formula: For a dataset with "n" data points, the mean (μ or $\bar{x}$) is calculated as follows: $\bar{x}=\sum x/n$

μ (Population Mean) = Σ (Sum of all values) / n

$\bar{x}$ (Sample Mean) = Σ (Sum of all values) / n

Mean for Ungrouped Data

- $\bar{x}=\sum x/n$

Example:

What is the mean of 2, 4, 6, 8 and 10?

Solution:

First, add all the numbers.

2 + 4 + 6 + 8 + 10 = 30

Now divide by 5 (total number of observations).

Mean = 30/5 = 6

Unit II:

Mean for Grouped Data

For grouped data, we can find the mean using either of the following formulas.

1. Direct method:

$$Mean, \bar{x} = \frac{\sum_{i=1}^{n} f_i x_i}{\sum_{i=1}^{n} f_i}$$

2. Assumed mean method:

$$Mean, (\bar{x}) = a + \frac{\sum f_i d_i}{\sum f_i}$$

3. Step-deviation method:

$$Mean, (\bar{x}) = a + h\frac{\sum f_i u_i}{\sum f_i}$$

Q. Find the mean for the following distribution.

x_i	11	14	17	20
f_i	3	6	8	7

Solution:

For the given data, we can find the mean using the direct method.

x_i	f_i	f_ix_i
11	3	33
14	6	84
17	8	136
20	7	140
	$\sum f_i = 24$	$\sum f_i x_i = 393$

Mean = $\sum f_ix_i/\sum f_i$ = 393/24 = 16.4

Q. Find the mean for the following distribution.

SOLUTION:

Class Interval	10-25	25-40	40-55	55-70	70-85	85-100
Number of Students	2	3	7	6	6	6

Class Interval	Number of students (f_i)	Class Mark (x_i)	$d_i = x_i$ – 47.5	$f_i d_i$
10-25	2	17.5	-30	-60
25-40	3	32.5	-15	-45
40-55	7	47.5	0	0
55-70	6	62.5	15	90
70-85	6	77.5	30	180
85-100	6	92.5	45	270
Total	Σf_i = 30			$\Sigma f_i d_i$ =435.

We can write $\bar{x} = a + \frac{\sum f_i d_i}{\sum f_i}$

Now, substitute the values of a, Σf_i , and $\Sigma f_i di$ in the above formula to get the mean,

Therefore, x̄ = 47.5 + (435/30)

x̄ = 47.5 + 14.5

x̄ = 62.

Step Deviation Method

Consider the same example as given above. In the step deviation method, we will add one more column to the table to find the mean, which is $u_i = (x_i – a)/h$

Where "a" is the assumed mean and "h" is the class size, which is equal to 15 (i.e) width of the class interval.

Class Interval	Number of students (f_i)	Class Mark (x_i)	$d_i = x_i – 47.5$ $d_i = x_i – a$	$u_i =(x_i\ a)/h$ (h=15)	f_iu_i
10-25	2	17.5	-30	-2	-4
25-40	3	32.5	-15	-1	-3
40-55	7	47.5	0	0	0
55-70	6	62.5	15	1	6
70-85	6	77.5	30	2	12
85-100	6	92.5	45	3	18
Total	$\Sigma f_i = 30$				$\Sigma f_iu_i =29$

$$\bar{x} = a + h\frac{\sum f_i u_i}{\sum f_i}$$

Now, substitute the values of a, h,Σf_i , and Σf_iui in the above formula to get the mean,

x̄ = 47.5 + 15(29/30)

x̄ = 47.5 + 15(0.967)

x̄= 47.5+ 14.5

x̄ = 62

Mean of Negative Numbers:-

We have seen examples of finding the mean of positive numbers till now. But what if the numbers in the observation list include negative numbers. Let us understand with an instance,

Example:

Find the mean of 9, 6, -3, 2, -7, 1.

Solution:

Add all the numbers first:

Total: 9+6+(-3)+2+(-7)+1 = 9+6-3+2-7+1 = 8

Now divide the total from 6, to get the mean.

Mean = 8/6 = 1.33

Types of Mean

There are majorly three different types of mean value that you will be studying in statistics.

Unit II:

1. Arithmetic Mean
2. Geometric Mean
3. Harmonic Mean

Arithmetic Mean:

When you add up all the values and divide by the number of values it is called Arithmetic Mean. To calculate, just add up all the given numbers then divide by how many numbers are given.

Example: What is the mean of 3, 5, 9, 5, 7, 2?

Now add up all the given numbers:

3 + 5 + 9 + 5 + 7 + 2 = 31

Now divide by how many numbers are provided in the sequence:

316= 5.16

5.16 is the answer.

Geometric Mean:

The geometric mean of two numbers x and y is xy. If you have three numbers x, y, and z, their geometric mean is 3xyz.

$$GeometricMean = \sqrt[n]{x_1 x_2 x_3 \ldots\ldots x_n}$$

Example: Find the geometric mean of 4 and 3 ?

$$GeometricMean = \sqrt{4 \times 3} = 2\sqrt{3} = 3.46$$

Harmonic Mean:

The harmonic mean is used to average ratios. For two numbers x and y, the harmonic mean is 2xy(x+y). For, three numbers x, y, and z, the harmonic mean is 3xyz(xy+xz+yz)

$$HarmonicMean(H) = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \frac{1}{x_2} + \frac{1}{x_3} + \cdots \ldots \frac{1}{x_n}}$$

Root Mean Square (Quadratic):

The root mean square is used in many engineering and statistical applications, especially when there are data points that can be negative.

$$X_{rms} = \sqrt{\frac{x_1^2 + x_2^2 + x_3^2 \ldots . x_n^2}{n}}$$

Contraharmonic Mean:

The contraharmonic mean of x and y is (x2 + y2)/(x + y). For n values,

$$\frac{(x_1^2 + x_2^2 + \cdots . + x_n^2)}{(x_1 + x_2 + \cdots . . x_n)}$$

2. Median:

- Definition: The median is the middle value in a dataset when all values are arranged in ascending or descending order. If there's an even number of data points, the median is the average of the two middle values. The median is less affected by extreme values (outliers) than the mean.
- Formula: To find the median:

Median of Ungrouped Data :

- If the total number of observations (n) is odd, then the median is (n+1)/2 th observation.
- If the total number of observations (n) is even, then the median will be average of n/2th and the (n/2)+1 th observation.

1. Sort the data in ascending or descending order.
2. If there is an odd number of data points, the median is the middle value.
3. If there is an even number of data points, the median is the average of the two middle values.

Q. For example, 6, 4, 7, 3 and 2 is the given data set.

SOLUTION :To find the median of the given dataset, arrange it in ascending order.

Therefore, the dataset is 2, 3, 4, 6 and 7.

In this case, the number of observations is odd. (i.e) n= 5

Hence, median = (n+1)/2 th observation.

Median = (5+1)/2 = 6/2 = 3rd observation.

Therefore, the median of the given dataset is 4.

Median of Grouped Data :

To find the median class, we have to find the cumulative frequencies of all the classes and n/2. After that, locate the class whose cumulative frequency is greater than (nearest to) n/2. The class is called the median class.

After finding the median class, use the below formula to find the median value.

$$Median = l + (\frac{\frac{n}{2} - cf}{f}) \times h$$

Where
l is the lower limit of the median class
n is the number of observations
f is the frequency of median class
h is the class size
cf is the cumulative frequency of class preceding the median class.
Example:

The following data represents the survey regarding the heights (in cm) of 51 girls of Class x. Find the median height.

Height (in cm)	Number of Girls
Less than 140	4
Less than 145	11
Less than 150	29
Less than 155	40
Less than 160	46
Less than 165	51

Solution:

To find the median height, first, we need to find the class intervals and their corresponding frequencies.
The given distribution is in the form of being less than type,145, 150 …and 165 gives the upper limit. Thus, the class should be below 140, 140-145, 145-150, 150-155, 155-160 and 160-165.
From the given distribution, it is observed that,
4 girls are below 140. Therefore, the frequency of class intervals below 140 is 4.

Unit II:

11 girls are there with heights less than 145, and 4 girls with height less than 140

Hence, the frequency distribution for the class interval 140-145 = 11-4 = 7

Likewise, the frequency of 145 -150= 29 – 11 = 18

Frequency of 150-155 = 40-29 = 11

Frequency of 155 – 160 = 46-40 = 6

Frequency of 160-165 = 51-46 = 5

Therefore, the frequency distribution table along with the cumulative frequencies are given below:

Class Intervals	Frequency	Cumulative Frequency
Below 140	4	4
140 – 145	7	11
145 – 150	18	29
150 – 155	11	40
155 – 160	6	46
160 – 165	5	51

Here, n= 51.

Therefore, n/2 = 51/2 = 25.5

Thus, the observations lie between the class interval 145-150, which is called the median class.

Therefore,

Lower class limit = 145

Class size, h = 5

Frequency of the median class, f = 18

Cumulative frequency of the class preceding the median class, cf = 11.

We know that the formula to find the median of the grouped data is:

$$Median = l + (\frac{\frac{n}{2} - cf}{f}) \times h$$

Now, substituting the values in the formula, we get

$$Median = 145 + (\frac{25.5 - 11}{18}) \times 5$$

Median = 145 + (72.5/18)

Median = 145 + 4.03

Median = 149.03.

Therefore, the median height for the given data is 149. 03 cm.

3. Mode:

- Definition: The mode is the value that appears most frequently in a dataset. It represents the most common value or values in the dataset. A dataset can have one mode (unimodal), more than one mode (multimodal), or no mode if all values occur with the same frequency.
- Formula: There is no specific formula for finding the mode. It is determined by observing the data and identifying the value(s) that occur most frequently.

In summary:

- Mean represents the average and is calculated by adding all values and dividing by the number of data points.
- Median represents the middle value when the data is ordered and is less influenced by extreme values.
- Mode represents the most frequently occurring value(s) in the dataset.

For example, the wickets taken by a bowler in 10 cricket matches are 2, 6, 4, 5, 0, 2, 1, 3, 2, 3. Find the mode of the given data.

Solution :- 2 is the number of wickets taken by the bowler in the maximum number of cricket matches i.e., 3. Therefore, the mode of the given data is 2.

Mode of Grouped Data:-

In the case of grouped data,

$$Mode = l + \left(\frac{f_1 - f_0}{2f_1 - f_0 - f_2}\right) \times h$$

Where,

f_1 is the frequency of the modal class

f_0 is the frequency of the class preceding the modal class

f_2 is the frequency of the class succeeding the modal class

h is the size of the class intervals

l is the lower limit of the modal class

Question: A survey has been conducted by a group of students on 20 households in a locality as shown in the following frequency distribution table. Find the mode for the given data.

Size of Family	1-3	3-5	5-7	7-9	9-11
No. of Families	7	8	2	2	1

Solution: From the given table, it is observed that the maximum class frequency is 8, and the corresponding class interval is 3-5.

Therefore, the modal class for the given data is 3-5.
The lower limit of modal class, l = 3
Class size, h = 2
Frequency of modal class, f_1 = 8
Frequency of class proceeding to modal class, f_0 = 7
Frequency of class succeeding to modal class, f_2 = 2
We know that the formula to find the mode of the grouped data is:

$$Mode = l + (\frac{f_1 - f_0}{2f_1 - f_0 - f_2}) \times h$$

Now, substituting the values in the mode formula, we get,

$$Mode = 3 + \left(\frac{8 - 7}{2(8) - 7 - 2}\right) \times 2$$

Mode = 3 + (2/7)
Mode = (21+2)/7
Mode = 23/7
Mode = 3.286.
Therefore, the mode of the given grouped data is 3.286.

Mean Deviation Definition

The mean deviation is defined as a statistical measure that is used to calculate the average deviation from the mean value of the given data set. The mean deviation of the data values can be easily calculated using the below procedure.
Step 1: Find the mean value for the given data values
Step 2: Now, subtract the mean value from each of the data values given (Note: Ignore the minus symbol)

Step 3: Now, find the mean of those values obtained in step 2.

Mean Deviation Formula

The formula to calculate the mean deviation for the given data set is given below.

Mean Deviation = [Σ |X – μ|]/N

Here,

Σ represents the addition of values

X represents each value in the data set

μ represents the mean of the data set

N represents the number of data values

Mean Deviation for Frequency Distribution

To present the data in the more compressed form we group it and mention the frequency distribution of each such group. These groups are known as class intervals.

Grouping of data is possible in two ways:

1. Discrete Frequency Distribution
2. Continuous Frequency Distribution

In the upcoming discussion, we will be discussing mean absolute deviation in a discrete frequency distribution.

Let us first know what is actually meant by the discrete distribution of frequency.

Mean Deviation for Discrete Distribution Frequency

As the name itself suggests, by discrete we mean distinct or non-continuous. In such a distribution the frequency (number of observations) given in the set of data is discrete in nature.

If the data set consists of values x_1,x_2, x_3………x_n each occurring with a frequency of f_1, f_2… f_n respectively then such a representation of data is known as the discrete distribution of frequency.

To calculate the mean deviation for grouped data and particularly for discrete distribution data the following steps are followed:

Step I: The measure of central tendency about which mean deviation is to be found out is calculated. Let this measure be a.

If this measure is mean then it is calculated as,

Unit II:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i f_i}{\sum_{i=1}^{n} f_i}$$

$$\Rightarrow \bar{x} = \frac{1}{N}\sum_{i=1}^{n} x_i f_i$$

where

$$N = \sum_{i=1}^{n} \quad f_i$$

If the measure is median then the given set of data is arranged in ascending order and then the cumulative frequency is calculated then the observations whose cumulative frequency is equal to or just greater than N/2 is taken as the median for the given discrete distribution of frequency and it is seen that this value lies in the middle of the frequency distribution.

Step II: Calculate the absolute deviation of each observation from the measure of central tendency calculated in step (I)

StepIII: The mean absolute deviation around the measure of central tendency is then calculated by using the formula

$$M.A.D(a) = \frac{\sum_{i=1}^{n} f_i \, |x_i - a|}{N}$$

If the central tendency is mean then,

$$M.A.D(\bar{x}) = \frac{\sum_{i=1}^{n} f_i \, |x_i - \bar{x}|}{N}$$

In case of median

$$M.A.D(M) = \frac{\sum_{i=1}^{n} f_i \, |x_i - M|}{N}$$

Example :

Determine the mean deviation for the data values 5, 3,7, 8, 4, 9.

Solution:

Given data values are 5, 3, 7, 8, 4, 9.

We know that the procedure to calculate the mean deviation.

First, find the mean for the given data:

Mean, μ = (5+3+7+8+4+9)/6

μ = 36/6

μ = 6

Therefore, the mean value is 6.

Now, subtract each mean from the data value, and ignore the minus symbol if any (Ignore”-”)

5 – 6 = 1

3 – 6 = 3

7 – 6 = 1

8 – 6 = 2

4 – 6 = 2

9 – 6 = 3

Now, the obtained data set is 1, 3, 1, 2, 2, 3.

Finally, find the mean value for the obtained data set

Therefore, the mean deviation is

= (1+3 + 1+ 2+ 2+3) /6

= 12/6

= 2

Hence, the mean deviation for 5, 3,7, 8, 4, 9 is 2.

Example :

In a foreign language class, there are 4 languages, and the frequencies of students learning the language and the frequency of lectures per week are given as:

Language	Sanskrit	Spanish	French	English
No. of students(x_i)	6	5	9	12
Frequency of lectures(f_i)	5	7	4	9

Calculate the mean deviation about the mean for the given data.

Solution: The following table gives us a tabular representation of data and the calculations

x_i	f_i	$x_i f_i$	$\lvert x_i - \bar{x} \rvert$	$f_i \lvert x_i - \bar{x} \rvert$
6	5	30	2.36	11.8
5	7	35	3.36	23.52
9	4	36	0.64	2.56
12	9	108	3.64	32.76
	$\sum f_i = 25$	$\bar{x} = \frac{1}{N}\sum_{i=1}^{n} x_i f_i = 8.36$		$\sum_{i=1}^{n} f_i \lvert x_i - \bar{x} \rvert$ $= 70.64$

Unit II:

Standard Deviation:

Standard Deviation Formula

The standard deviation formula is given as:

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(X_i-\mu)^2}$$

Here,
σ = Population standard deviation
N = Number of observations in population
Xi = ith observation in the population
μ = Population mean
Similarly, the sample standard deviation formula is:

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2}$$

Here,
s = Sample standard deviation
n = Number of observations in sample
xi = ith observation in the sample

$\bar{x}$= Sample mean

Variance Formula:

The variance formula is given by:

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(X_i-\mu)^2$$

The sample variance formula is given by:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2$$

Q. Consider the number of gold coins 5 pirates have; 4, 2, 5, 8, 6.

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

$$= \frac{x_1 + x_2 + x_3 + x_4 \ldots\ldots + x_n}{n}$$

= (4 + 2 + 5 + 6 + 8) / 5

= 5

$$x_n - \bar{x} \; for\ every\ value\ of\ the\ sample:$$

$$x_1 - \bar{x} = 4 - 5 = -1$$

$$x_2 - \bar{x} = 2 - 5 = -3$$

$$x_3 - \bar{x} = 5 - 5 = 0$$

$$x_4 - \bar{x} = 8 - 5 = 3$$

$$x_5 - \bar{x} = 6 - 5 = 1$$

$$\sum (x_n - \bar{x})^2$$

$$= (x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_5 - \bar{x})^2$$

$$= (-1)^2 + (-3)^2 + 0^2 + 3^2 + 1^2$$

= 20

Standard deviation:

$$S.D = \sqrt{\frac{\sum (x_n - \bar{x})^2}{n-1}}$$

$$= \sqrt{\frac{20}{4}}$$

= √5 = 2.236

Standard deviation of Grouped Data

In case of grouped data or grouped frequency distribution, the standard deviation can be found by considering the frequency of data values.

Question: Calculate the mean, variance and standard deviation for the following data:

Class Interval	0-10	10-20	20-30	30-40	40-50	50-60
Frequency	27	10	7	5	4	2

Solution:

Class Interval	Frequency (f)	Mid Value (x_i)	fx_i	fx_i^2

0 – 10	27	5	135	675
10 – 20	10	15	150	2250
20 – 30	7	25	175	4375
30 – 40	5	35	175	6125
40 – 50	4	45	180	8100
50 – 60	2	55	110	6050
	∑f = 55		$\sum fx_i$ = 925	$\sum fx_i^2$ = 27575

N = ∑f = 55

Mean = $(\sum fx_i)/N$ = 925/55 = 16.818

Variance = $1/(N – 1) [\sum fx_i^2 – 1/N(\sum fx_i)^2]$

= $1/(55 – 1) [27575 – (1/55) (925)^2]$

= (1/54) [27575 – 15556.8182]

= 222.559

Standard deviation = √variance = √222.559 = 14.918

SKEWNESS:-

The skewness in statistics is a measure of asymmetry or the deviation of a given random variable’s distribution from a symmetric distribution (like normal Distribution).

In Normal Distribution, we know that: Median = Mode = Mean

Skewness in statistics can be divided into two categories. They are:

- Positive Skewness
- Negative Skewness

Positive Skewness

The extreme data values are higher in a positive skew distribution, which increases the mean value of the data set. To put it another way, a positive skew distribution has the tail on the right side.

It means that, Mean > Median > Mode in positive skewness

Negative Skewness

The extreme data values are smaller in negative skewness, which lowers the dataset's mean value. A negative skew distribution is one with the tail on the left side.
Hence, in negative Skewness, Mean < Median < Mode.

Skewness Formula in Statistics:-

The skewness formula is called so because the graph plotted is displayed in a skewed manner. Skewness is a measure used in statistics that helps reveal the asymmetry of a probability distribution. It can either be positive or negative, irrespective of the signs. To calculate the skewness, we have to first find the mean and variance of the given data.

The skewness formula is given by:

$$g = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^3}{(n-1)s^3}$$

Where,

$\bar{x}$ is the sample mean

x_i is the ith sample
n is the total number of observations
s is the standard deviation
g= sample skewness

Question. Find the skewness in the following data.

Height (inches)	Class Marks	Frequency
59.5 – 62.5	61	5
62.5 – 65.5	64	18
65.5 – 68.5	67	42
68.5 – 71.5	70	27
71.5 – 74.5	73	8

To know how skewed these data are as compared to other data sets, we have to compute the skewness.
Sample size and sample mean should be found out.

N = 5 + 18 + 42 + 27 + 8 = 100

$$\bar{x} = \frac{(61 \times 5) + (64 \times 18) + (67 \times 42) + (70 \times 27) + (73 \times 8)}{100}$$

$$\bar{x} = \frac{6745}{100} = 67.45$$

Now with the mean, we can compute the skewness.

Class Mark, *x*	Frequency, *f*	*xf*	$(x - \bar{x})$	$(x - \bar{x})^2 \times f$	$(x - \bar{x})^3 \times f$
61	5	305	-6.45	208.01	-1341.68
64	18	1152	-3.45	214.25	-739.15
67	42	2814	-0.45	8.51	-3.83
70	27	1890	2.55	175.57	447.70
73	8	584	5.55	246.42	1367.63
		6745	n/a	852.75	-269.33
		67.45	n/a	8.5275	-2.6933

Now, the skewness is

$$g = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^3}{(n-1)s^3}$$

s=√[(8.5275/(100-1))=0.2935]

g=√[(-2.693/[99 * (0.295)^3] = -1.038

For interpreting we have the following rules as per Bulmer in the year 1979:

- If the skewness comes to less than -1 or greater than +1, the data distribution is highly skewed
- If the skewness comes to between -1 and -1/2 or between 1/2 and +1, the data distribution is moderately skewed.

- If the skewness is between -1/2 and 1/2, the distribution is approximately symmetric.

Karl Pearson’s coefficient of skewness:-

coefficient of skewness is :

Using Mode: $\dfrac{\overline{x} - \text{Mode}}{s}$

Using Median: $\dfrac{3(\overline{x} - \text{Median})}{s}$

where, $\bar{x}$ is the mean and s is the standard deviation.

Calculate Coefficient of Skewness:-

Q. Suppose the mean of a data set is 60.5, the mode is 75, the median is 70 and the standard deviation is 10. The steps to calculate the coefficient of skewness are as follows:

Using Mode

- Step 1: Subtract the mode from the mean. 60.5 - 75 = -14.5
- Step 2: Divide this value by the standard deviation to get the coefficient of skewness. Thus, sk_1 = -14.5 / 10 = -1.45.

Using Median

- Step 1: Subtract the median from the mean. 60.5 - 70 = -9.5
- Step 2: Multiply this value by 3. This gives -28.5.
- Step 3: Divide the value from step 2 by the standard deviation to obtain the coefficient of skewness. Thus, sk_2 = -28.5 / 10 = -2.85

Important Notes on Coefficient of Skewness

- The coefficient of skewness is used to measure the extent and direction of skewness of a sample or distribution.
- The coefficient of skewness can be positive, negative, or zero.
- There are two formulas, given by Karl Pearson, that can be used to calculate the coefficient of skewness.

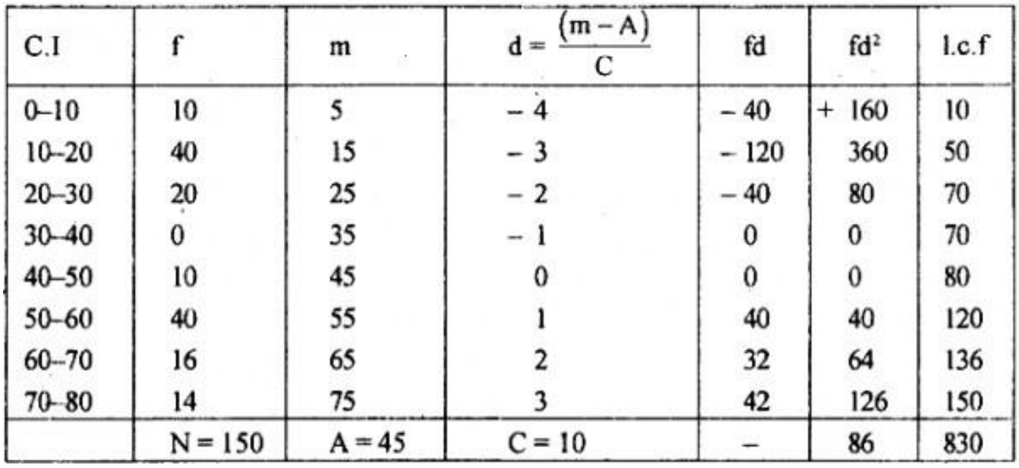
Q. Calculate Karl – Pearson’s coefficient of skewness from the following data.

Year	0-10	10-20	20-30	30-40	40-50	50-60	60-70	70-80
No. of persons	10	40	20	0	10	40	16	14

SOLUTION : The frequency distribution is bi-modal, so using Karl-Pearson’s coefficient of skewness for ill-defined data:

= 3(X -m)/σ

Let C.I and f be the Age and no. of persons



C.I	f	m	$d = \frac{(m-A)}{C}$	fd	fd²	l.c.f
0–10	10	5	– 4	– 40	+ 160	10
10–20	40	15	– 3	– 120	360	50
20–30	20	25	– 2	– 40	80	70
30–40	0	35	– 1	0	0	70
40–50	10	45	0	0	0	80
50–60	40	55	1	40	40	120
60–70	16	65	2	32	64	136
70–80	14	75	3	42	126	150
	N = 150	A = 45	C = 10	–	86	830

$$\bar{X} = A + \left(\frac{\Sigma fd}{N} \times c\right)$$

$$= 45 - \left(\frac{86}{150} \times 10\right)$$

$$= 45 - \left(\frac{86}{150} \times 10\right)$$

$$\bar{X} = 39.27$$

$$\text{Median class} = \frac{N^{th}}{2}\text{ item}$$

$$= \frac{150}{2} = 75^{th}\text{ item}$$

$$= (40 - 50)$$

$$M = L + \left(\frac{\frac{N}{2} - c.f}{f} \times i \right)$$

$$= 40 + \left(\frac{75 - 70}{2} \times 10 \right)$$

$$= 40 + 5$$

$$\therefore M = 45$$

$$\sigma = \sqrt{\frac{\Sigma fd^2}{N} - \left(\frac{\Sigma fd}{N}\right)^2} \times c$$

$$\sigma = 22.8136$$

$$= \sqrt{\frac{830}{150} - \left(\frac{-86}{150}\right)^2} \times 10$$

$$K.P.C.S.K = 3\left(\frac{\bar{X} - M}{\sigma}\right)$$

$$= \frac{3(39.27 - 45)}{22.8136}$$

The distribution is negatively skewed.

Kurtosis:

Kurtosis is a statistical measure that describes the distribution of data in a dataset. It provides information about the tails and the shape of the distribution. There are three main types of kurtosis: mesokurtic, leptokurtic, and platykurtic.

Mesokurtic:

A mesokurtic distribution has kurtosis equal to zero.

The distribution has tails that are neither too heavy (leptokurtic) nor too light (platykurtic).

It has a similar shape to the normal distribution.

Leptokurtic:

A leptokurtic distribution has positive kurtosis.

The tails of the distribution are heavier than those of a normal distribution.

This means that there are more extreme values in the dataset.

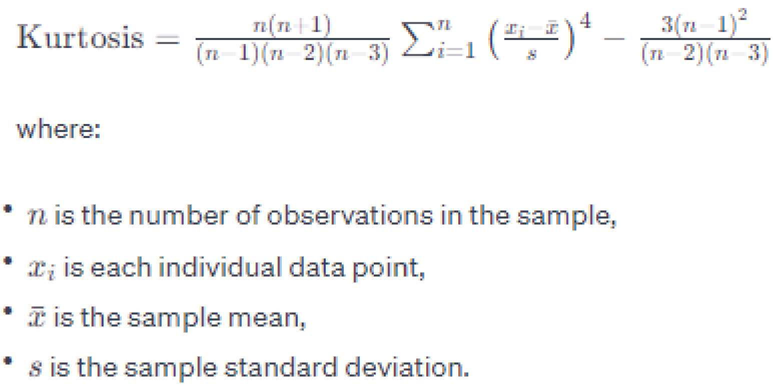
Platykurtic:

A platykurtic distribution has negative kurtosis.

The tails of the distribution are lighter than those of a normal distribution.

This implies that there are fewer extreme values in the dataset.

Kurtosis



$$\text{Kurtosis} = \frac{n(n+1)}{(n-1)(n-2)(n-3)} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s}\right)^4 - \frac{3(n-1)^2}{(n-2)(n-3)}$$

where:

- n is the number of observations in the sample,
- x_i is each individual data point,
- $\bar{x}$ is the sample mean,
- s is the sample standard deviation.

Key points about kurtosis:

Positive kurtosis indicates heavier tails, while negative kurtosis indicates lighter tails compared to a normal distribution.

Kurtosis is a dimensionless quantity, meaning it is not in the units of the original data.

In a normal distribution, the kurtosis is 3 (excess kurtosis is zero), and any deviation from this value indicates a departure from normality.

It's important to note that kurtosis alone does not provide a complete picture of the shape of the distribution. It is often used in conjunction with other measures, such as skewness and histograms, to fully understand the characteristics of a dataset.

EXAMPLE:

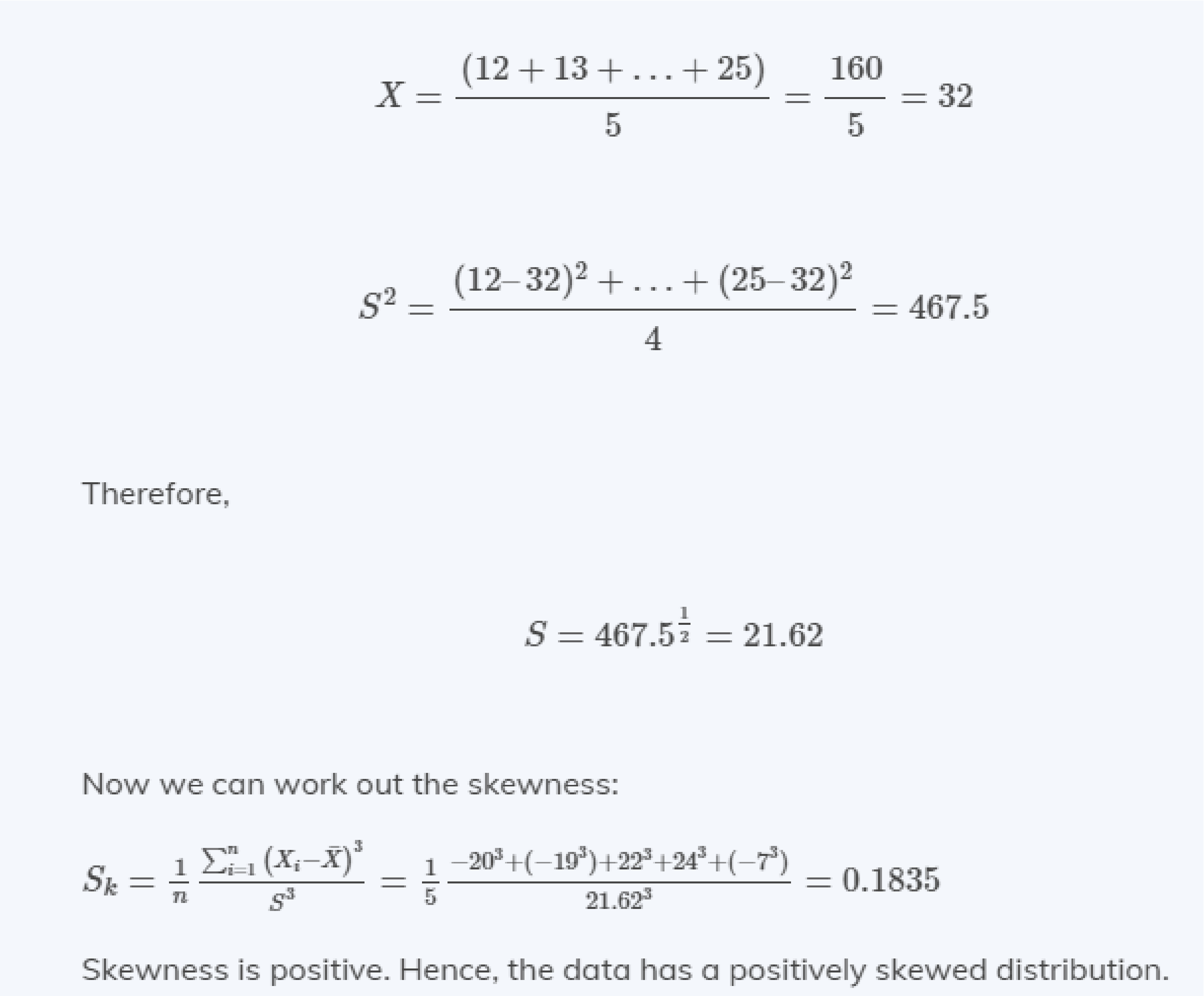
Suppose we have the following observations:

{12 13 54 56 25}

Determine the skewness of the data.

Solution

First, we must determine the sample mean and the sample standard deviation:



$$X = \frac{(12 + 13 + \ldots + 25)}{5} = \frac{160}{5} = 32$$

$$S^2 = \frac{(12\text{–}32)^2 + \ldots + (25\text{–}32)^2}{4} = 467.5$$

Therefore,

$$S = 467.5^{\frac{1}{2}} = 21.62$$

Now we can work out the skewness:

$$S_k = \frac{1}{n}\frac{\sum_{i=1}^{n}(X_i - \bar{X})^3}{S^3} = \frac{1}{5}\frac{-20^3 + (-19^3) + 22^3 + 24^3 + (-7^3)}{21.62^3} = 0.1835$$

Skewness is positive. Hence, the data has a positively skewed distribution.

Question

Using the data from the example above (12 13 54 56 25), determine the type of kurtosis present.

A. Mesokurtic distribution

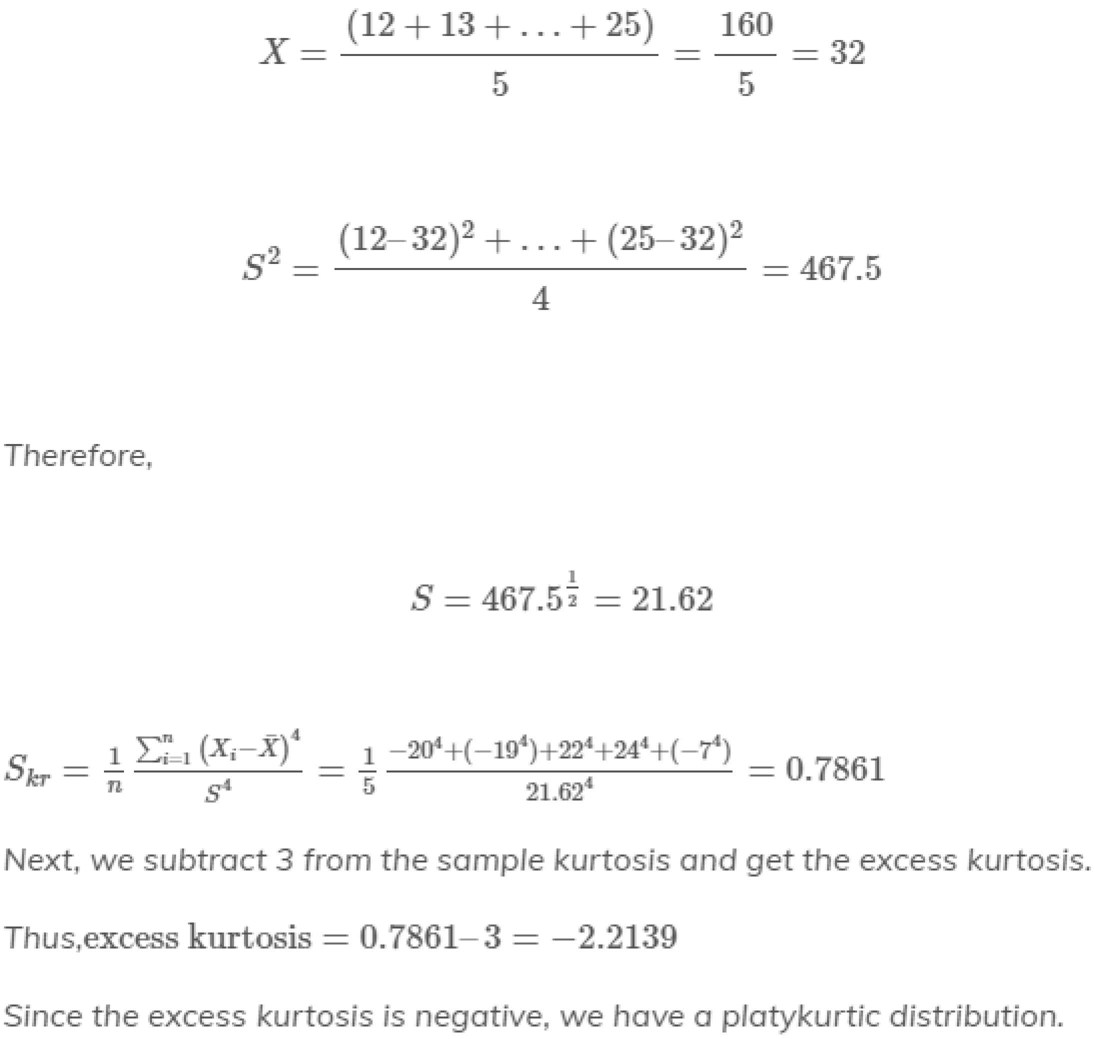
B. Platykurtic distribution

C. Leptokurtic distribution

Solution

The correct answer is B.

Unit II:



$$X = \frac{(12 + 13 + \ldots + 25)}{5} = \frac{160}{5} = 32$$

$$S^2 = \frac{(12 – 32)^2 + \ldots + (25 – 32)^2}{4} = 467.5$$

Therefore,

$$S = 467.5^{\frac{1}{2}} = 21.62$$

$$S_{kr} = \frac{1}{n}\frac{\sum_{i=1}^{n}\left(X_i – \bar{X}\right)^4}{S^4} = \frac{1}{5}\frac{-20^4+(-19^4)+22^4+24^4+(-7^4)}{21.62^4} = 0.7861$$

Next, we subtract 3 from the sample kurtosis and get the excess kurtosis.

Thus, $\text{excess kurtosis} = 0.7861 – 3 = -2.2139$

Since the excess kurtosis is negative, we have a platykurtic distribution.