



ARULMIGU PALANIANDAVR ARTS COLLEGE FOR WOMEN

(Autonomous)

(Re-Accredited with 'B⁺⁺' Grade by NAAC 3rd Cycle)

Run by Arulmigu Dhandayuthapani Swamy Thirukoil, H.R & C.E Dept. Government of Tamil Nadu
A Government Aided College - Affiliated to Mother Teresa Women's University, Kodaikanal
CHINNAKALAYAMPUTHUR(PO), PALANI - 624615



DEPARTMENT OF COMPUTER SCIENCE

AND

DEPARTMENT OF COMPUTER APPLICATION

LEARNING RESOURCE

STATISTICS

What is Statistics?

Statistics is a branch of mathematics that deals with the collection, review, and analysis of data. It is known for drawing the conclusions of data with the use of quantified models. Statistical analysis is a process of collecting and evaluating data and summarizing it into mathematical form.

Statistics can be defined as the study of the collection, analysis, interpretation, presentation, and organization of data. In simple words, it is a mathematical tool that is used to collect and summarize data.

Uncertainty and fluctuation in different fields and parameters can be determined only through statistical analysis. These uncertainties are determined by the probability that plays a very important role in statistics.

Basic terminology of Statistics:

Population

It is actually a collection of set of individuals or objects or events whose properties are to be analyzed.

Sample

It is the subset of a population.

Scope of Statistics

Numerous fields, including psychology, geology, sociology, forecasting the weather, probability, and many more, use statistics. Statistics is considered a mathematical science since its concentration is on applications and its purpose is to get insight from the data.

Methods in Statistics

The techniques include gathering, condensing, examining, and understanding fluctuating numerical data. Here, a few of the techniques are listed below.

- Data gathering
- Summarising data
- Statistic evaluation

What is Data in Statistics?

Data is a collection of observations, it can be in the form of numbers, words, measurements, or statements.

Types of Data

1. **Qualitative Data:** This data is descriptive. For example – She is beautiful, He is tall, etc.
2. **Quantitative Data:** This is numerical information. For example- A horse has four legs.

Types of Quantitative Data

1. **Discrete Data:** It has a particular fixed value and can be counted.
2. **Continuous Data:** It is not fixed but has a range of data and can be measured.

Statistical Data

When complete census data cannot be obtained, **statisticians** gather sample data through the creation of complex experiment designs and survey samples. Statistics, in and of itself, offers tools for prediction and forecasting via **statistical models**.

The scientific discipline of probability theory includes sampling theory. In mathematical statistics, probability is used to investigate the sampling distributions of sample statistics and, more broadly, the properties of statistical procedures like organizing and grouping data through graphs, pie charts, etc as discussed in the articles below:

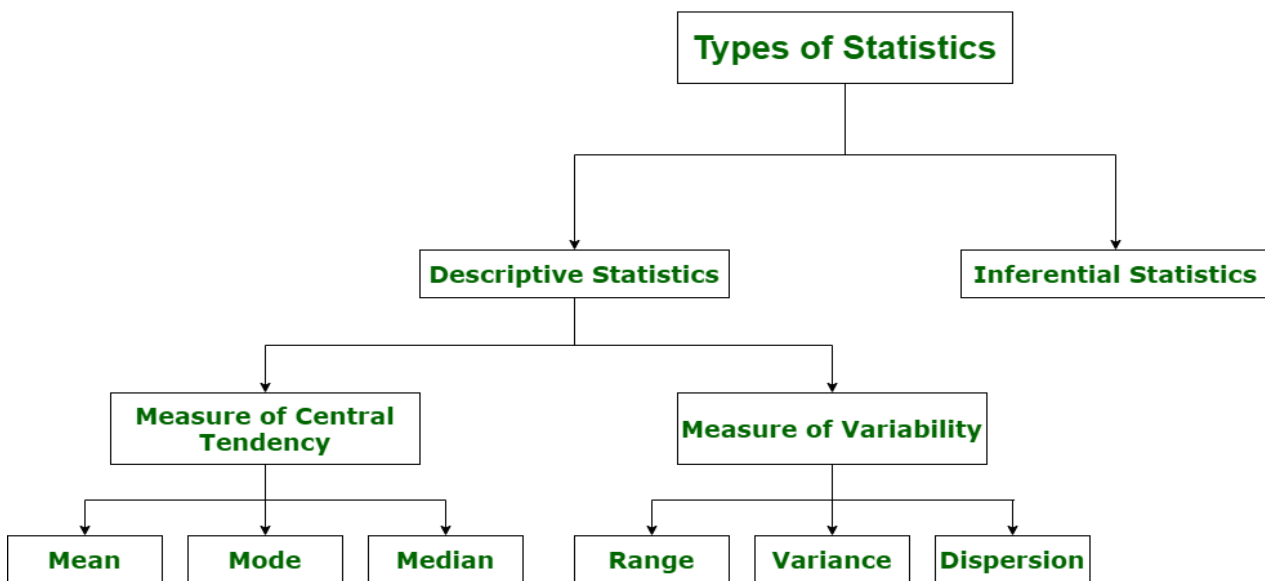
1. Data Handling
2. Organizing Data
3. Grouping Data
4. Pie Chart
5. Chance and Probability
6. Introduction to Graphs
7. Linear Graphs
8. Presentation of data
9. Graphical Representation of Data
10. Bar graphs and Histograms

Representation of data

Data may be represented in various ways, including tables, charts, and graphs. In general, statistical data are represented as follows:

- Bar Graph
- Pie Chart
- Line Graph
- Pictograph
- Histogram
- Frequency Distribution

Types of Statistics



Descriptive Statistics :

Descriptive statistics uses data that provides a description of the population either through numerical calculation or graph or table. It provides a graphical summary of data. It is simply used for summarizing objects, etc. There are two categories in this as following below.

(a). Measure of central tendency –

Measure of central tendency is also known as summary statistics that is used to represents the center point or a particular value of a data set or sample set.

In statistics, there are three common measures of central tendency as shown below:

What is Mean?

Mean is the sum of all the values in the data set divided by the number of values in the data set. It is also called the Arithmetic Average. Mean is denoted as \bar{x} and is read as x bar.

Mean Formula

The formula to calculate the mean is,

Mean (\bar{x}) = Sum of Values / Number of Values

If $x_1, x_2, x_3, \dots, x_n$ are the values of a data set then the mean is calculated as: $\bar{x} = (x_1 + x_2 + x_3 + \dots + x_n) / n$

Example:

Find the mean of data sets 10, 30, 40, 20, and 50

Solution:




Mean of the data 10, 30, 40, 20, 50 is

Mean = (sum of all values) / (number of values)

Mean = (10 + 30 + 40 + 20 + 50) / 5 = 30

Mean of Grouped Data

Mean for the grouped data can be calculated by using various methods. The most common methods used are discussed in the table below,

-  Direct Method
-  Assumed Mean Method
-  Step Deviation Method

Direct method:

Mean

$$\bar{x} = \frac{\sum f_i x_i}{\sum f_i}$$

where, $\sum f_i$ is the sum of all frequencies

Example:

Find the mean of the given distribution, which contains the students' exam grades.

Grade	37	89	58	98	70	15	66	35
No. of students	3	9	5	1	19	20	36	7

Ans:

Let's make a table to figure out the total:

$$\text{Mean} = (\sum f_i x_i) / \sum f_i$$

$$= 5551/100$$

$$= 55.51$$

Thus, the average of the provided distribution is 55.51.

Marks (xi)	No. of students (fi)	fixi
37	3	111
89	9	801
58	5	290
98	1	98
70	19	1330
15	20	300
66	36	2376
35	7	245
Sum	100	5551

Assumed Mean Method

$$\text{Mean } \bar{x} = a + \frac{\sum f_i x_i}{\sum f_i}$$

where, a is Assumed mean

d_i is equal to $x_i - a$

$\sum f_i$ the sum of all frequencies

Example:

The results of a test taken by 150 students are shown in the table below.

Class	0-10	10-20	20-30	30-40	40-50
Frequency	29	34	23	26	38

Solution:

To calculate x_i , d_i , and $f_i d_i$,

we'll make a table.

Assumed mean = $a = 25$

Class (CI)	Frequency (fi)	Class mark (xi)	$d_i = x_i - a$	$f_i d_i$
0-10	29	5	$5 - 25 = -20$	-580
10-20	34	15	$15 - 25 = -10$	-340
20-30	23	$25 = a$	$25 - 25 = 0$	0
30-40	26	35	$35 - 25 = 10$	260
40-50	38	45	$45 - 25 = 20$	760
Total	$\Sigma f_i = 150$			$\Sigma f_i d_i = -100$

$$\text{Mean of the data} = a + \left(\frac{\Sigma f_i d_i}{\Sigma f_i} \right)$$

$$= 25 + \left(\frac{-100}{150} \right)$$

$$= 25 - \left(\frac{1}{15} \right)$$

$$= \frac{(375-1)}{15}$$

$$= \frac{374}{15}$$

$$= 23.375$$

Hence, the mean marks of the students = 23.375.

Step Deviation Method

Mean

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

where,

a is Assumed mean

$$u_i = (x_i - a)/h$$

h is Class size

$\sum f_i$ the sum of all frequencies

Example:

Calculate the mean of the following data using the Step Deviation method.

Age(in year)	20-24	24-28	28-32	32-36	36-40	40-44	44-48
Frequency	3	6	8	5	5	2	1

Range of the data is 20 to 48, for assumption of mean, let's take average of the range values,

$$\text{Assumed mean} = (20 + 48) / 2 = 68/2 = 34$$

Let's A = 34 be the assumed mean of the data,

Now, using assumed mean value, let's create the table for step deviation as follows:

Age (in years)	Frequency(f_i)	Class Mark(x_i)	Deviation($d_i = x_i - A$)	Step Deviation ($u_i = d_i/h$)	$f_i \times u_i$
20 – 24	3	22	-12	-3	-9
24 – 28	6	26	-8	-2	-12
28 – 32	8	30	-4	-1	-8
32 – 36	5	34	0	0	0
36 – 40	5	38	4	1	5
40 – 44	2	42	8	2	4

44 – 48	1	46	12	3	3
	$\sum f_i = 20$				$\sum f_i u_i = -17$

Thus,

$$\text{Mean} = 34 + 4 \times (-17)/20$$

$$= 34 + 4 \times (-0.85)$$

$$= 34 - 3.4$$

$$= 30.6$$

Thus, the mean age of data using step deviation method is 30.6

Properties of Arithmetic Mean:

- 1) It is rigidly defined.
- 2) It is based on all the observations.
- 3) It is easy to comprehend.
- 4) It is simple to calculate.
- 5) The presence of extreme observations has the least impact on it.
- 6) The sum of deviations of the items from the arithmetic mean is always zero.
- 7) The Sum of the squared deviations of the items from A.M. is minimum, which is less than the sum of the squared deviations of the items from any other values.
- 8) If each item in the series is replaced by the mean, then the sum of these substitutions will be equal to the sum of the individual items.) It is amenable to mathematical treatment or properties.

Draw backs of arithmetic mean:

- 1) It is very much affected by sampling fluctuation.
- 2) Arithmetic mean cannot be advocated to open end classification.

Merits:

1. It is straightforward to calculate and comprehend. It is for this reason that it is the most widely used central tendency measure.
2. Every item has an impact because it is included in the calculation.
3. The result remains the same since the mathematical formula is rigid.




4. When repeated samples are gathered from the same population, fluctuations are minimal for this measure of central tendency.
5. Unlike other measures like as mode and median, it can be subjected to algebraic treatment.
6. A.M. has an advantage in that it is a calculated quantity that is not depending on the order of terms in a series.
7. Due to its strict definition, it is mostly used to compare issues.

Demerits or Limitations:

1. It cannot be located graphically.
2. A single component can have a significant impact on the outcome. If there are three terms, for example, 4, 7, and 10, X is 7. The new X is $4+7+10+95/4 = 116/4 = 29$ when we add a new term 95. When compared to the size of the X - in the first three terms, this is a significant change.
3. Only if the frequency is regularly distributed will it be useful. If the skewness is greater, the results will be ineffectual.
4. In the case of open end class intervals, we must assume the intervals' boundaries, and a small fluctuation in X is possible. This is not the case with median and mode, as the open end intervals are not used in their calculations.
5. Because data cannot be stated numerically, qualitative forms such as Cleverness and Riches cannot provide X .
6. Unlike the mode and median, X cannot be found by inspection.
7. It can sometimes come to absurd or impossible conclusions, for example, if three courses have 60, 50, and 12 pupils, the average number of students is $60+50+12/3 = 50.67$, which is impossible because students cannot be in fractions.

Types of Mean

There are majorly 3 distinct types of mean value that you will find in statistics.

-  Arithmetic mean
-  Geometric Mean
-  Harmonic Mean

Arithmetic Mean

The Arithmetic Mean is the average of the numbers/data or can be understood as the calculated central value of a set of numbers. To determine Arithmetic Mean:

- Add all the numbers/data given.
- Divide the total obtained in the above steps by the total numbers/data.

Here N= Total number of observations.

Geometric Mean Formula

$$\bar{X}_{geom} = \sqrt[n]{\prod_{i=1}^n x_i} = \sqrt[n]{x_1 \cdot x_2 \cdot \dots \cdot x_n}$$

Example:

Find the geometric mean of 1,3,5,7,9

Solution :

The GM is given as $(x_1 \times x_2 \times x_3 \dots \times x_n)^{1/n}$

$$= (1 \times 3 \times 5 \times 7 \times 9)^{1/5}$$

$$= (945)^{1/5}$$

$$= 3.936$$

Therefore, GM = 3.936

Geometric Mean

Grouped data

G.M can be calculated using: $GM = \text{Antilog} (\sum f \log x_i) / n$,

where $n = f_1 + f_2 + \dots + f_n$.

Example:

Calculate Geometric mean from the following grouped data

X	1	2	3	4	5
Frequency	5	10	6	3	2

Solution:

Geometric mean :

x	f	$f \log(x)$
1	5	0
2	10	3.0103
3	6	2.8627
4	3	1.8062
5	2	1.3979
---	---	---
--	$n=26$	$\sum f \log(x)=9.0771$

$$\text{GM of } X = \text{Antilog} \left(\frac{\sum f \log(x)}{n} \right)$$

$$= \text{Antilog} \left(\frac{9.0771}{26} \right)$$

$$= \text{Antilog}(0.3491)$$

$$= 2.2342$$

Examples:

Find the geometric mean of the following grouped data for the frequency distribution of weights.

Weights of ear heads (g)	No of ear heads (f)
60-80	22
80-100	38
100-120	45

120-140	35
140-160	20
Total	160

Solution:

Weights of ear heads (g)	No of ear heads (f)	Mid x	Log x	f log x
60-80	22	70	1.845	40.59
80-100	38	90	1.954	74.25
100-120	45	110	2.041	91.85
120-140	35	130	2.114	73.99
140-160	20	150	2.176	43.52
Total	160			324.2

From the given data, $n = 160$

We know that the G.M for the grouped data is

$$GM = \text{Antilog} (324.2 / 160)$$

$$GM = \text{Antilog} (2.02625)$$

$$GM = 106.23$$

Therefore, the G.M = 106.23

Some real-life uses of geometric mean:

1. Aspect Ratios:

The geometric mean has been used in film and video also to find the appropriate aspect ratios i.e. the proportion of the width to the height of a screen or image. It is used to find an appropriate balancing between the two aspect ratios as well as for distorting or cropping both ratios equally.

2. Computer Science:

Computers use mind-boggling amounts of large data which generally requires the summarization for many applications using various statistical measurements.

3. Medicine:

The Geometric Mean has many applications in the medical industry also. It is known as the “gold standard” for some measurements, including for the calculation of gastric emptying time.

4. Proportional Growth:

It is very useful in finding the growth rate. The geometric mean is used for calculating the proportional growth as well as demand growth.

Merits and Demerits of Geometric Mean

The merits and demerits of the G.M. can be outlined in the light of the characteristics of an ideal average as follows:

Merits

1. It is rigidly defined.
2. It is based on all the observations of the series.
3. It is suitable for measuring the relative changes.

4. It gives more weights to the small values and fewer weights to the large values.
5. It is used in averaging the ratios, percentages and in determining the rate gradual increase and decrease.
6. It is capable of further algebraic treatment. As such, if the G.M. and the number of items of two or more series are given, we can readily find out the combined G.M. of all the series by the following formula: $G_{1.2} = \text{AL} \{ N_1 \log G_1 + N_2 \log G_2 / N_1 + N_2 \}$

Demerits

1. It is not easy to understand by a man of ordinary prudence as it involves logarithmic operations. As such it is not popular like that of arithmetic average.
2. It is difficult to calculate as it involves finding out of the root of the products of certain values either directly, or through logarithmic operations.
3. It cannot be calculated, if the number of negative values is odd.
4. It cannot be calculated, if any value of a series is zero.
5. At times it gives a value which may not be found in the series, and may even be assured or impracticable.

Harmonic Mean

Definition

The Harmonic Mean (HM) is defined as the reciprocal of the average of the reciprocals of the data values. It is based on all the observations, and it is rigidly defined. Harmonic mean gives less weightage to the large values and large weightage to the small values to balance the values correctly.

In general, the harmonic mean is used when there is a necessity to give greater weight to the smaller items. It is applied in the case of times and average rates.

Harmonic Mean Formula

Since the harmonic mean is the reciprocal of the average of reciprocals, the formula to define the harmonic mean “HM” is given as follows:

If $x_1, x_2, x_3, \dots, x_n$ are the individual items up to n terms, then,

$$\text{Harmonic Mean, HM} = n / [(1/x_1) + (1/x_2) + (1/x_3) + \dots + (1/x_n)]$$

How to Find a Harmonic Mean?

If a, b, c, d, ... are the given data values, then the steps to find the harmonic mean are as follows:

Step 1: Calculate the reciprocal of each value (1/a, 1/b, 1/c, 1/d, ...)

Step 2: Find the average of reciprocals obtained from step 1.

Step 3: Finally, take the reciprocal of the average obtained in step 2.

Harmonic Mean Formula

Harmonic Mean Formula

Harmonic mean formula is a sort of average calculator that is estimated by dividing the number of utilities or values by the total of the reciprocals of each value of the data series.

The harmonic mean serves to find multiplicative or divisor relations among fractions without worrying about common denominators. Harmonic means are often practised in averaging things.

If $x_1, x_2, x_3, \dots, x_n$ are the individual elements then

$$H.M = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \frac{1}{x_3} + \dots + \frac{1}{x_n}} = \frac{n}{\sum_{i=1}^n \frac{1}{x_i}}$$

For a frequency distribution, the harmonic mean formula is

$$H.M = \frac{N}{\sum_{i=1}^n f \frac{1}{x_i}}$$

Here N=summation of f.

Harmonic Mean vs Geometric Mean

Harmonic Mean	Geometric Mean
We can calculate the harmonic mean for a series by dividing the total number of terms by the sum of reciprocal terms.	We can calculate the geometric mean for a series by multiplying all the terms and taking the nth root.
The value of Harmonic mean is less than all the other means, i.e. GM and AM.	The value of Geometric mean is always greater than HM, but less than AM.
We can call HM as the arithmetic mean of the data set with reciprocal transformation.	We can call GM as the arithmetic mean of the data set with log transformation.
Example: For a sequence: 1, 2, 4, 7 n = 4 HM= 2.113	Example: For a sequence: 1, 2, 4, 7 n = 4 GM=2.735

Harmonic Mean vs Arithmetic Mean

Harmonic Mean	Arithmetic Mean
We can find the HM of the given series by finding the reciprocal of the AM of the reciprocal terms in the data set.	We can find the arithmetic mean by dividing the sum of all the elements of the data set by the number of elements.
Harmonic mean gives the lowest value among all the three means.	Arithmetic mean gives the highest value among all the three means.
We cannot use harmonic mean on data set that consist of negative or zero values.	We can calculate the arithmetic mean of all kinds of data sets containing positive, negative, or zero value.
Example: For a sequence: 3, 2, 1, 6 n = 4 HM= 2	Example: For a sequence: 3, 2, 1, 6 n = 4 AM=3

Harmonic Mean Examples

Example 1: What is the harmonic mean for the given data 5, 10, 15, 20.

Solution: Given there are a total of 4 numbers.

Therefore the Harmonic Mean according to the formula will be:

If $x_1, x_2, x_3, \dots, x_n$ are the individual elements then

$$H.M = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \frac{1}{x_3} + \dots + \frac{1}{x_n}} = \frac{n}{\sum_{i=1}^n \frac{1}{x_i}}$$

$$\text{Harmonic mean} = \frac{4}{\frac{1}{5} + \frac{1}{10} + \frac{1}{15} + \frac{1}{20}}$$

$$H.M = \frac{4}{0.2+0.1+0.66+0.05}$$

$$H.M = \frac{4}{1.01}$$

$$H.M = 3.960$$

Hence the harmonic mean for the given data is 3.960.

Merits and Demerits of Harmonic Mean

The benefits of the harmonic mean are as follows:

1. It is very clearly defined.
2. It is calculated using all of a series' observations; it cannot be calculated without considering any of the series' items.
3. It can be further algebraically treated.
4. It produces greater results when the goals to be reached are the same for all methods used.
5. It gives the smallest component in a series the most weight.
6. It can be calculated for any negative value in a series.
7. It transforms a skewed distribution into a normal distribution.
8. It produces a curve that is straighter than the arithmetic and geometric mean.

The harmonic mean has the following drawbacks:

1. It is difficult to comprehend for a prudent person.
2. Its calculation is time-consuming because it requires obtaining the reciprocals of the numbers.
3. When the means used for diverse ends are the same, it does not produce better or more accurate results.
4. Its algebraic treatment is substantially more constrained than the arithmetic mean's.
5. The values of the extreme items have a significant impact.
6. It is impossible to determine whether any of the things are zero.

Harmonic Mean Uses

The following are the most common applications of harmonic means:

- In finance, the harmonic mean is used to calculate average multiples such as the price-earnings ratio.
- It is also utilised by market technicians to discover patterns such as Fibonacci Sequences.

What is Median?

A Median is a middle value for sorted data. The sorting of the data can be done either in ascending order or descending order. A median divides the data into two equal halves.

Median Formula

The formula for the median is,

If the number of values (n value) in the data set is odd then the formula to calculate the median is,

$$\text{Median} = [(n + 1)/2]\text{th term}$$

If the number of values (n value) in the data set is even then the formula to calculate the median is:

$$\text{Median} = [(n/2)\text{th term} + \{(n/2) + 1\}\text{th term}] / 2$$

Example: Find the median of given data set 30, 40, 10, 20, and 50

Solution:

Median of the data 30, 40, 10, 20, 50 is,

Step 1: Order the given data in ascending order as:

10, 20, 30, 40, 50

Step 2: Check n (number of terms of data set) is even or odd and find the median of the data with respective 'n' value.

Step 3: Here, n = 5 (odd)

$$\text{Median} = [(n + 1)/2]\text{th term}$$

$$\begin{aligned}\text{Median} &= [(5 + 1)/2]\text{th term} \\ &= 30\end{aligned}$$

Median of Grouped Data

The median of the grouped data median is calculated using the formula,

$$\text{Median} = l + [(n/2 - cf) / f] \times h$$

where

l is lower limit of median class

n is number of observations

f is frequency of median class

h is class size

cf is cumulative frequency of class preceding the median class.

Example:

Calculate the median for the following data.

Class	10 – 20	20 – 30	30 – 40	40 – 50	50 – 60
Frequency	5	10	12	8	5

Solution:

Create the following table for the given data.

Class	Frequency	Cumulative Frequency
10 – 20	5	5
20 – 30	10	15
30 – 40	12	27
40 – 50	8	35
50 – 60	5	40

As $n = 40$ and $n/2 = 20$,

Thus, 30 – 40 is the median class.

$l = 30$, $c_f = 15$, $f = 12$, and $h = 10$

Putting the values in the formula $\text{Median} = l + \frac{N/2 - c_f}{f} \times h$

$$\text{Median} = 30 + (20 - 15)/12 \times 10$$

$$\Rightarrow \text{Median} = 30 + (5/12) \times 10$$

$$\Rightarrow \text{Median} = 30 + 4.17$$

⇒ Median = 34.17

So, the median value for this data set is 34.17

Merits and Demerits of Median

Median as an average of position has a number of merits and demerits. These are outlined as under keeping in view the characteristics of an ideal average.

Merits

1. It is simple to understand.
2. It does not require all the observations of the data for its determination.
3. It is not affected by the extreme values of a series.
4. It can be determined graphically which is shown a little later along with the quartiles etc.
5. It can be determined easily in open and series without estimating the lowest or highest class limits.
6. It can be determined straightway in case of a series of unequal class intervals without converting the series into equal class intervals as it is done in case of a Mode.
7. It is considered suitable for computation of the mean deviation as the sum of the deviation as the sum of the deviations taken from it is the minimum.
8. It is capable of being expressed in qualitative form as it is not computed but located.
9. It gives a value, which very much exists in the series, and is a round figure in most of the cases:

Demerits

1. It is not rigidly, and as such, its value cannot be computed but located.
2. It is not based on all the observations of the series.
3. It is not capable of further algebraic treatment like mean, geometric mean and harmonic mean.
4. It needs the arrangement of a series in ascending or descending order and more particularly in a frequency distribution it needs the arrangement of the series in ascending order.
5. It is very much affected by fluctuations in sampling.
6. It gives erroneous result, if the number of items is very small.

7. If the number of items of an individual series is of even nature, its value is determined by the process of arithmetic average.
8. At times, it produces a value which is never found in the series.
9. At times it gives fractional and impracticable results.

Definition of Mode





In statistics, the mode is the value that appears most frequently in a data set. It is a measure of central tendency and can be calculated for both numerical and categorical data.

Example:

In the given set of data: 2, 4, 5, 5, 6, 7, the mode of the data set is 5 since it has appeared in the set twice.

Types of Mode in Statistics

A dataset may contain numerous modes if two or more values occur equally often. The dataset is referred to as multimodal in these circumstances. A dataset without recurring values, however, lacks a mode. Depending upon the number of modal solutions, mode is classified into the following categories:

-  Unimodal
-  Bimodal
-  Trimodal
-  Multimodal

Unimodal

When there is only one and only one mode of any dataset, then that dataset is called a unimodal dataset.

For example, In set $X = \{1, 2, 2, 3, 6, 7, 7, 7, 8, 9\}$, only 7 appears thrice. Thus, there is only one mode of the dataset X.

Bimodal

When there exist two modes in the given data set, then it is called a bimodal.

For example, In Set $A = \{1,1,1,3,4,4,6,6,6\}$ the mode is 1 and 6 because both 1 and 6 have the highest frequency in the given set.

Trimodal

When there exist three modes in the given data set, then it is called trimodal.

For example, In Set A = {2, 2, 2, 3, 4, 4, 6, 6, 6, 7, 9, 9, 9} the mode is 2, 6, and 9 because 2, 6, and 9 have the highest frequency in the given set.

Multimodal

When there are four or more modes in the given data set, then it is called multimodal.

For example,

In Set A = {1, 1, 1, 3, 4, 4, 6, 6, 6, 7, 9, 9, 9, 11, 11, 11} 1, 6, 9, and 11 are the mode because 1, 6, 9, and 11 have the highest frequency in the given set.

Mode for Ungrouped Data

To calculate the mode of any given ungrouped data set, use the following steps: Step 1: Sort the data in ascending or descending order, whichever is more convenient.

Step 2: Determine the value that occurs most frequently in the data set. This value is the mode.

Step 3: If there are two or more values that occur with the same highest frequency, then the data set has multiple modes.

Example:

Find the mode in the given set of data: 4, 6, 8, 16, 22, 24, 41, 24, 42, 24, 15, 13, 61, 24, 29.

Solution:

Arrange the given set of data in ascending order,

4, 7, 8, 13, 15, 16, 22, 24, 24, 24, 24, 29, 41, 42, 61.

The mode of the data set is 24 as it appeared in the given most.

Mode of Grouped Data

The mode of the grouped data is calculated using the formula,

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

where,

f₁ is the frequency of the modal class

f₀ is the frequency of the class preceding the modal class

f₂ is the frequency of the class succeeding the modal class

h is the size of class intervals

l is the lower limit of modal class

Example:

Find the mode of the dataset which is given as follows.

Class Interval	10-20	20-30	30-40	40-50	50-60
Frequency	5	8	12	16	10

Solution:

As the class interval with the highest frequency is 40-50, which has a frequency of 16.

Thus, 40-50 is the modal class.

Thus, $l = 40$, $h = 10$, $f_1 = 16$, $f_0 = 12$, $f_2 = 10$

Plugging in the values in formula,

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

we get

$$\text{Mode} = 40 + \frac{(16 - 12)}{(2 \times 16 - 12 - 10)} \times 10$$

$$\Rightarrow \text{Mode} = 40 + \frac{4}{10} \times 10$$

$$\Rightarrow \text{Mode} = 40 + 4$$

$$\Rightarrow \text{Mode} = 44$$

Therefore, the mode for this set of data is 44.

Merits and Demerits of Mode

Mode as an average of position has a number of merits and demerits. These are outlined here as under:

Merits

1. It gives the most representative value of a series.
2. It is not affected by the extreme values of a series. For example, let a series be as under.

Here, the value of the Mode is 15. If, an extreme value, say 60 is added to the series, or 10 is deleted from the series, the value of the mode will remain the same 15.

3. It can be determined straight way from an open end series without estimating the two extreme class limits.
4. It is capable of studying qualitative data as its determination depends on the frequencies rather than the values of the items.
5. It can be determined graphically either through a histogram, or through a Frequency Polygon.
6. It is considered a reliable average for studying skewness of a distribution.
7. It is understood by a layman as it refers to a value that occurs for maximum times. Thus, when we talk of modal size of shoes. a layman easily understands that it refers a size of shoes which demanded by the maximum number of customers.
8. It is very much useful in the field of business, and commerce as it helps a businessman in taking a decision on the varieties of the goods he should procure in large quantities to enhance his sales.

Demerits

1. It is not rigidly defined, and so in certain cases it may come out with different results.
2. It is not based on all the observations of a series but on the concentration of frequencies of the items. If any non-modal value is left out of the series, or is added thereto, the value of the mode is not altered.
3. It is not capable of further algebraic treatment like A.M., G.M. or H.M.
4. In case of a continuous and bimodal series, its determination becomes difficult, and lingering as it involves passing through a number of trials and use of interpolation formulae in those cases.
5. It cannot be easily determined graphically, if two, or more values of a series have the same highest frequency. Thus, for the following series, Mode can not be easily located through histogram.

Marks :	1-3	3-5	5-7	7-9	9-11
F :	5	15	20	20	4

6. It cannot be determined from a series with unequal class intervals unless they are equalized on the assumption that the frequencies are evenly distributed, and such assumption may not also hold good.
7. In certain cases, it contradicts its very meaning and nature when certain value with lesser frequency is determined as the modal values.
8. There are different methods, and different formula which field different results of Mode, and so it is rightly remarked as the most ill defined average.

Relation between Mean Median Mode

For any group of data, the relation between the three central tendencies mean, median, and mode is

Mean Median Mode Formula

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

Difference Between Mean and Median

The mathematical average is known as the mean of the data set, whereas the positional average is considered the Median.

The difference between Mean and Median is understood by the following example.

In a school, there are 8 teachers whose salaries are 20000 rupees, a principal with a salary of 35000, find their mean salary and median salary.

$$\begin{aligned} \text{Mean} &= \frac{(20000+20000+20000+20000+20000+20000+20000+20000+35000)}{9} \\ &= \frac{195000}{9} \\ &= 21666.67 \end{aligned}$$

Therefore, the mean salary is ₹21,666.67.

For median, in ascending order: 20000, 20000, 20000, 20000, 20000, 20000, 20000, 20000, 35000.

$n = 9$,

$$\text{Thus, } (9 + 1)/2 = 5$$

Thus, the median is the 5th observation.

$$\text{Median} = 20000$$

Therefore, the median is ₹20,000.

After comparing the mean and median for the above data set. It is evident that the mean salary is Rs 21666.67, and the median is Rs 20,000

Example :

Find the mean, median, mode, and range for the given data

190, 153, 168, 179, 194, 153, 165, 187, 190, 170, 165, 189, 185, 153, 147, 161, 127, 180

Solution:

For Mean:

190, 153, 168, 179, 194, 153, 165, 187, 190, 170, 165, 189, 185, 153, 147, 161, 127, 180

Number of observations = 18

Mean = (Sum of observations) / (Number of observations)

$$\begin{aligned} &= (190+153+168+179+194+153+165+187+190+170+165+189+185+153+147 \\ &+161+127+180) / 18 \\ &= 2871/18 \\ &= 159.5 \end{aligned}$$

Therefore, the mean is 159.5

For Median:

The ascending order of given observations is,

127, 147, 153, 153, 153, 161, 165, 165, 168, 170, 179, 180, 185, 187, 189, 190, 190, 194

Here, n = 18

Median = $1/2 [(n/2) + (n/2 + 1)]$ th observation

$$= 1/2 [9 + 10] \text{th observation}$$

$$= 1/2 (168 + 170)$$

$$= 338/2$$

$$= 169$$

Thus, the median is 169

For Mode:

The number with the highest frequency = 153

Thus, mode = 53

For Range:

$$\begin{aligned}\text{Range} &= \text{Highest value} - \text{Lowest value} \\ &= 194 - 127 \\ &= 67\end{aligned}$$

Example:

Find the Median of the data 25, 12, 5, 24, 15, 22, 23, 25

Solution:

25, 12, 5, 24, 15, 22, 23, 25

Step 1: Order the given data in ascending order as:

5, 12, 15, 22, 23, 24, 25, 25

Step 2: Check n (number of terms of data set) is even or odd and find the median of the data with respective 'n' value.

Step 3: Here, $n = 8$ (even) then,

$$\text{Median} = [(n/2)\text{th term} + \{(n/2) + 1\}\text{th term}] / 2$$

$$\begin{aligned}\text{Median} &= [(8/2)\text{th term} + \{(8/2) + 1\}\text{th term}] / 2 \\ &= (22+23) / 2 \\ &= 22.5\end{aligned}$$

Example:

Find the mode of given data 15, 42, 65, 65, 95.

Solution:

Given data set 15, 42, 65, 65, 95

The number with highest frequency = 65

Mode = 65

What are the differences between Mean, Median, and Mode?

These three terms are related to each other. There's a relationship between mean, median and mode and is called an empirical relationship between them. Below are some of the most integral differences between the mean, median and mode.

Sl. No.	Mean	Median	Mode
1.	The average taken for a set of numbers is called a mean.	The middle value in the data set is called the Median.	The number that occurs the most in a given list of numbers is called a mode.
2.	Add all of the numbers together and divide the sum by the total number of values.	Place all the given numbers in an ascending order	It shows the frequency of occurrence.
3.	The result is the mean or average score.	The next step is to find the middle number on the list. It is called the median.	We can have more than one mode or no mode at all.
4.	<p>Example: To find the average of the four numbers 2, 4, 6, and 8, we need to add the number first.</p> <ul style="list-style-type: none"> $2 + 4 + 6 + 8 = 20$ Divide the sum by the total number of numbers, i. e 4. $20/4 = 5$ is the average or mean 	<p>Example: 4, 2, 8, 10, 19.</p> <ul style="list-style-type: none"> Arrange the numbers in ascending order. i .e., 2, 4, 8, 10, 19. As the total numbers are 5, so the middle number 8 is the median here. 	<p>Example: 3, 3, 5, 6, 7, 7, 8, 1, 1, 1, 4, 5, 6.</p> <ul style="list-style-type: none"> Find the frequency of each number. For number 3, it's 2. For 5, it's 2. For 6, it's 2. For 7, it's 2. For 8, it's one. For 1, it's 3. For 4, it's 1. The number with the highest frequency is the mode. Hence, the mode of the given sequence of numbers is 1.

Measures of Dispersion

In statistics, the dispersion measures help in understanding how homogeneous or heterogeneous the data is. In simpler words, it shows how constrained or dispersed the variable is. Absolute and relative dispersion metrics are the two different categories that exist. They are as follows:

- Range
- Variance
- Standard Deviation
- Quartiles and Quartile Deviation
- Mean and Mean Deviation

Range

It is the difference between the highest value and the lowest value. It is a way to understand how the numbers are spread in a data set.

Range Formula

The formula to find the Range is:

$$\text{Range} = \text{Highest value} - \text{Lowest Value}$$

Merits and Demerits of Range

Merits or Uses:

1. It is easiest to calculate and simplest to understand even for a beginner.
2. It is one of those measures which are rigidity defined.
3. It gives us the total picture of the problem even with a single glance.

It is used to check the quality of a product for quality control. Range plays an important role in preparing R- charts, thus quality is maintained.

5. The idea about the price of Gold and Shares is also made taking care of the range in which prices have moved for the past some periods.
6. Meteorological Dep't. Also makes forecasts about the weather by keeping range of temp, in view.

Demerits or Limitations or Drawbacks:

1. Range is not based on all the terms. Only extreme items reflect its size. Hence range cannot be completely representative of the data as all other middle values are ignored.
2. Due to above reason range is not a reliable measure of dispersion.
3. Range does not change even the least even if all other, in between, terms and variables are changed.
4. Range is too much affected by fluctuation of sampling. Range changes from sample to sample. As the size of sample increases range increases and vice versa.
5. It does not tell us anything about the variability of other data.
6. For open-end intervals, range is indeterminate because lower and upper limits of first and last interval are not given.

Example: Find the range of the given data set 12, 19, 6, 2, 15, 4

Solution:

Given set is {12, 19, 6, 2, 15, 4}

Here,

Lowest Value = 2

Highest Value = 19

Range = 19 - 2
= 17

What is Interquartile Range?

Range only takes the two extreme values (largest and smallest) into consideration; therefore, it is a crude measure of dispersion. This effect of extreme values on range can be avoided by using the measure of interquartile range. The difference between the values of two quartiles is known as **Interquartile Range**.

Interquartile Range Formula

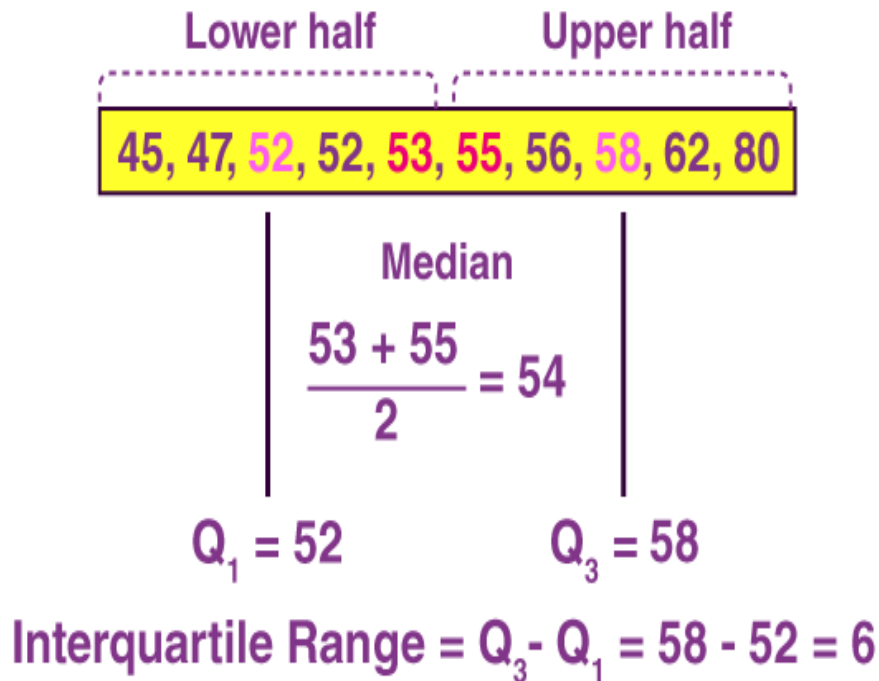
The difference between the upper and lower quartile is known as the interquartile range. The formula for the interquartile range is given below

Interquartile range = Upper Quartile – Lower Quartile

= $Q_3 - Q_1$

where Q_1 is the first quartile and Q_3 is the third quartile of the series.

The below figure shows the occurrence of median and interquartile range for the data set.



Semi Interquartile Range

The semi-interquartile range is defined as the measures of dispersion. Semi interquartile range also is defined as half of the interquartile range. It is computed as one half the difference between the 75th percentile (Q_3) and the 25th percentile (Q_1). The semi-interquartile range is one-half of the difference between the first and third quartiles. The Formula for Semi Interquartile Range is

$$\text{Semi Interquartile Range} = (Q_3 - Q_1) / 2$$

Median and Interquartile Range

The median is the middle value of the distribution of the given data. The interquartile range (IQR) is the range of values that resides in the middle of the scores. When a distribution is skewed, and the median is used instead of the mean to show a central tendency, the appropriate measure of variability is the Interquartile range.

Q_1 – Lower Quartile Part

Q_2 – Median

Q_3 – Upper Quartile Part

It is a measure of dispersion based on the lower and upper quartile. Quartile deviation is obtained from interquartile range on dividing by 2, hence also known as semi interquartile range.

How to Calculate the Interquartile Range?

The procedure to calculate the interquartile range is given as follows:

- ✚ Arrange the given set of numbers into increasing or decreasing order.
- ✚ Then count the given values. If it is odd, then the center value is median otherwise obtain the mean value for two center values. This is known as Q_2 value. If there are even number of values, the median will be the average of the middle two values.
- ✚ Median equally cuts the given values into two equal parts. They are described as Q_1 and Q_3 parts.
- ✚ The median of data values below the median represents Q_1 .
- ✚ The median of data values above the median value represents Q_3 .
- ✚ Finally, we can subtract the median values of Q_1 and Q_3 .
- ✚ The resulting value is the interquartile range.

Example:

Determine the interquartile range value for the first ten prime numbers.

Solution:

Given: The first ten prime numbers are:

2, 3, 5, 7, 11, 13, 17, 19, 23, 29

This is already in increasing order.

Here the number of values = 10

10 is an even number. Therefore, the median is mean of 11 and 13

That is $Q_2 = (11 + 13)/2 = 24/2 = 12$.

Now we have to get two parts i.e. lower half to find Q_1 and the upper half to find Q_3 .

Q_1 part : 2, 3, 5,7,11

Here the number of values = 5

5 is an odd number. Therefore, the center value is 5, that is $Q_1 = 5$

Q_3 part : 13, 17, 19, 23, 29

Here the number of values = 5

5 is an odd number. Therefore, the center value is 19, that is $Q_3 = 19$

The subtraction of Q_1 and Q_3 value is $19 - 5 = 14$

Therefore, **14 is the interquartile range value.**

Example:

Calculate Interquartile Range of the following data:

150, 110, 200, 300, 180, 320

Solution:

S.No.	Items arranged in ascending order
1	110
2	150
3	180
4	200
5	300
6	320
N = 6	

$$Q1 = \text{Size of 1.75th item} = \text{Size of 1st item} + 0.75(\text{Size of 2nd item} - \text{Size of 1st item})$$

$$Q1 = 110 + 0.75(150 - 110)$$

$$Q1 = 110 + 30$$

$$\mathbf{Q1 = 140}$$

$$Q3 = \text{Size of 5.25th item} = \text{Size of 5th item} + 0.25(\text{Size of 6th item} - \text{Size of 5th item})$$

$$Q3 = 300 + 0.25(320 - 300)$$

$$Q3 = 300 + 5$$

$$\mathbf{Q3 = 305}$$

$$\text{Interquartile Range} = Q3 - Q1 = 305 - 140$$

$$\mathbf{\text{Interquartile Range} = 165}$$

Quartile Deviation Formula

Suppose Q_1 is the lower quartile, Q_2 is the median, and Q_3 is the upper quartile for the given data set, then its quartile deviation can be calculated using the following formula.

$$QD = (Q_3 - Q_1)/2$$

In the next section, you will learn how to calculate these quartiles for both ungrouped and grouped data separately.

Quartile Deviation for Ungrouped Data

For an ungrouped data, quartiles can be obtained using the following formulas,

$$Q_1 = [(n+1)/4]\text{th item}$$

$$Q_2 = [(n+1)/2]\text{th item}$$

$$Q_3 = [3(n+1)/4]\text{th item}$$

Where n represents the total number of observations in the given data set.

Also, Q_2 is the median of the given data set, Q_1 is the median of the lower half of the data set and Q_3 is the median of the upper half of the data set.

Before, estimating the quartiles, we have to arrange the given data values in ascending order. If the value of n is even, we can follow the similar procedure of finding the median.

Quartile Deviation for Grouped Data

For a grouped data, we can find the quartiles using the formula,

$$Q_r = l_1 + \frac{r\left(\frac{N}{4}\right) - c}{f} (l_2 - l_1)$$

Here,

Q_r = the rth quartile

l_1 = the lower limit of the quartile class

l_2 = the upper limit of the quartile class

f = the frequency of the quartile class

c = the cumulative frequency of the class preceding the quartile class

N = Number of observations in the given data set

Merits of Quartile Deviation:

1. It can be easily calculated and simply understood.
2. It does not involve much mathematical difficulties.
3. As it takes middle 50% terms hence it is a measure better than Range and Percentile Range.

4. It is not affected by extreme terms as 25% of upper and 25% of lower terms are left out.
5. Quartile Deviation also provides a short cut method to calculate Standard Deviation using the formula $6 \text{ Q.D.} = 5 \text{ M.D.} = 4 \text{ S.D.}$
6. In case we are to deal with the center half of a series this is the best measure to use.

Demerits or Limitation Quartile Deviation:

As Q_1 and Q_3 are both positional measures hence are not capable of further algebraic treatment.

2. Calculation are much more, but the result obtained is not of much importance.
3. It is too much affected by fluctuations of samples.
4. 50% terms play no role; first and last 25% items ignored may not give reliable result.
5. If the values are irregular, then result is affected badly.
6. We can't call it a measure of dispersion as it does not show the scatterness around any average.
7. The value of Quartile may be same for two or more series or Q.D. is not affected by the distribution of terms between Q_1 and Q_3 or outside these positions.

Example :

Find the quartiles and quartile deviation of the following data:

17, 2, 7, 27, 15, 5, 14, 8, 10, 24, 48, 10, 8, 7, 18, 28

Solution:

Given data:

17, 2, 7, 27, 15, 5, 14, 8, 10, 24, 48, 10, 8, 7, 18, 28

Ascending order of the given data is:

2, 5, 7, 7, 8, 8, 10, 10, 14, 15, 17, 18, 24, 27, 28, 48

Number of data values = $n = 16$

Q_2 = Median of the given data set

n is even, median = $(1/2) [(n/2)\text{th observation and } (n/2 + 1)\text{th observation}]$

= $(1/2)[8\text{th observation} + 9\text{th observation}]$

= $(10 + 14)/2$

= $24/2$

= 12

$Q_2 = 12$

Now, lower half of the data is:

2, 5, 7, 7, 8, 8, 10, 10 (even number of observations)

Q_1 = Median of lower half of the data

$$= (1/2)[4\text{th observation} + 5\text{th observation}]$$

$$= (7 + 8)/2$$

$$= 15/2$$

$$= 7.5$$

Also, the upper half of the data is:

14, 15, 17, 18, 24, 27, 28, 48 (even number of observations)

Q_3 = Median of upper half of the data

$$= (1/2)[4\text{th observation} + 5\text{th observation}]$$

$$= (18 + 24)/2$$

$$= 42/2$$

$$= 21$$

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

$$= (21 - 7.5)/2$$

$$= 13.5/2$$

$$= 6.75$$

Therefore, the quartile deviation for the given data set is 6.75.

Example :

Calculate the quartile deviation for the following distribution.

Class	0-10	10-20	20-30	30-40	40-50	50-60	60-70	70-80	80-90	90-100
Frequency	5	3	4	3	3	4	7	9	7	8

Solution:

Let us calculate the cumulative frequency for the given distribution of data.

Class	Frequency	Cumulative Frequency
0 – 10	5	5

10 – 20	3	$5 + 3 = 8$
20 – 30	4	$8 + 4 = 12$
30 – 40	3	$12 + 3 = 15$
40 – 50	3	$15 + 3 = 18$
50 – 60	4	$18 + 4 = 22$
60 – 70	7	$22 + 7 = 29$
70 – 80	9	$29 + 9 = 38$
80 – 90	7	$38 + 7 = 45$
90 – 100	8	$45 + 8 = 53$

Here, $N = 53$

We know that,

$$Q_r = l_1 + \frac{r\left(\frac{N}{4}\right) - c}{f} (l_2 - l_1)$$

Finding Q_1 :

$$r = 1$$

$$N/4 = 53/4 = 13.25$$

Thus, Q_1 lies in the interval 30 – 40.

In this case, quartile class = 30 – 40

l_1 = the lower limit of the quartile class = 30

l_2 = the upper limit of the quartile class = 40

f = the frequency of the quartile class = 3

c = the cumulative frequency of the class preceding the quartile class = 12

Now, by substituting these values in the formula we get:

$$Q_1 = 30 + [(13.25 - 12)/3] \times (40 - 30)$$

$$= 30 + (1.25/3) \times 10$$

$$= 30 + (12.5/3)$$

$$= 30 + 4.167$$

$$= 34.167$$

Finding Q₃:

$$r = 3$$

$$3N/4 = 3 \times 13.25 = 39.75$$

Thus, Q₃ lies in the interval 80 – 90.

In this case, quartile class = 80 – 90

l₁ = the lower limit of the quartile class = 80

l₂ = the upper limit of the quartile class = 90

f = the frequency of the quartile class = 7

c = the cumulative frequency of the class preceding the quartile class = 38

Now, by substituting these values in the formula we get:

$$Q_3 = 80 + [(39.75 - 38)/7] \times (90 - 80)$$

$$= 80 + (1.75/7) \times 10$$

$$= 80 + (17.5/7)$$

$$= 80 + 2.5$$

$$= 82.5$$

Finally, the quartile deviation = $(Q_3 - Q_1)/2$

$$QD = (82.5 - 34.167)/2$$

$$= 48.333/2$$

$$= 24.1665$$

Hence, the quartile deviation of the given distribution is 24.167 (approximately).

What is Variance?

Variance is another measure of dispersion and is based on standard deviation. The term variance was first used by R.A. Fisher in 1913 and means the square of the standard deviation of the given distribution. Symbolically, Variance is denoted by σ^2 .

$$\text{Variance} = \sigma^2$$

Standard Deviation and Variance are two measures of dispersion and are closely related to each other. The only difference between them is that Standard Deviation is the square root of Variance; however, Variance is the average squared deviation from the mean.

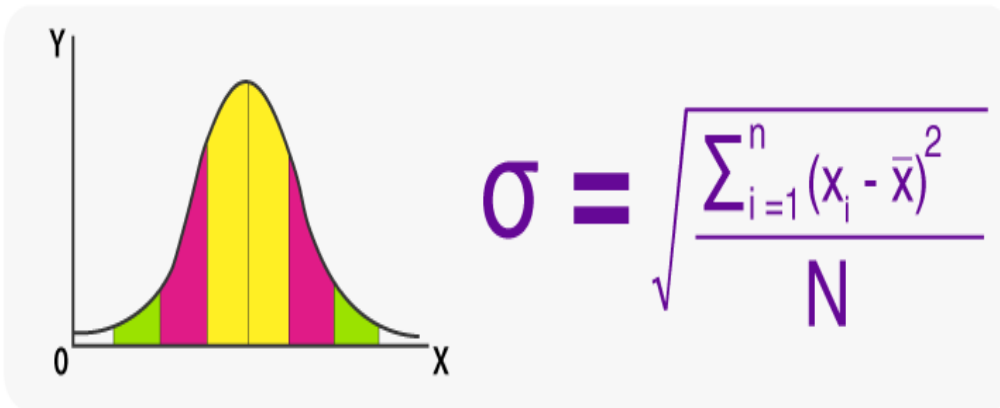
Standard deviation

Standard deviation formula is used to find the values of a particular data that is dispersed. In simple words, the standard deviation is defined as the deviation of the values or data from an average mean. Lower standard deviation concludes that the values are very close to their average. Whereas higher values mean the values are far from the mean value. It should be noted that the standard deviation value can never be negative.

Standard Deviation is of two types:

1. Population Standard Deviation
2. Sample Standard Deviation

Standard Deviation Formula



Formula to Calculate Standard Deviation

Actual Mean Method:

Example:

Determine the variance of the following distribution using the Actual Mean Method.

No. of Workers	12	15	17	10	14	11	13	12
-----------------------	----	----	----	----	----	----	----	----

Solution:

No. of Workers (X)	$x = X - \bar{X}$	x^2
12	$12 - 13 = -1$	1
15	$15 - 13 = 2$	4
17	$17 - 13 = 4$	16
10	$10 - 13 = -3$	9
14	$14 - 13 = 1$	1
11	$11 - 13 = -2$	4
13	$13 - 13 = 0$	0
12	$12 - 13 = -1$	1
$\Sigma X = 104$		$\Sigma x^2 = 36$

Arithmetic Mean =13

Standard Deviation =2.12

Variance = $2.12^2 = 4.5$

Variance = 4.5

Assumed Mean Method:

Example:

Calculate the variance of the data given below using Assumed Mean Method.

Size (X)	10	20	30	40	50	60
Frequency (f)	6	2	4	1	5	2

Solution:

Size (X)	Frequency (f)	d = X - A A = 30	fd	d ²	fd ²
10	6	-20	-120	400	2,400
20	2	-10	-20	100	200
30 (A)	4	0	0	0	0
40	1	10	10	100	100
50	5	20	100	400	2,000
60	2	30	60	900	1,800
	N = $\sum f = 20$		$\sum fd = 30$		$\sum fd^2 = 6,500$

Standard Deviation = 17.9

Variance = $17.9^2 = 322.75$

Variance = 322.75

Step-Deviation Method:

Example:

Calculate the Variance of the data given below using the Step-Deviation Method.

Marks (X)	0-10	10-20	20-30	30-40	40-50
No. of Students (f)	5	4	3	6	2

Solution:

Marks (X)	No. of Students (f)	Mid-Point (m)	$d = m - A$ $A = 25$	$d' = \frac{m-A}{C}$ $C = 10$	fd'	d'^2	fd'^2
0-10	5	5	-20	-2	-10	4	20
10-20	4	15	-10	-1	-4	1	4
20-30	3	25 (A)	0	0	0	0	0
30-40	6	35	10	1	6	1	6
40-50	2	45	20	2	4	4	8
	$N = \sum f = 20$				$\sum fd' = -4$		$\sum fd'^2 = 38$

Standard Deviation = 1.36

Variance = $1.36^2 = 1.86$

Variance = 1.86

Example:

Find the standard deviation of the numbers given (3, 8, 6, 10, 12, 9, 11, 10, 12, 7).

Solution:

Step 1: First compute the mean of the 10 values given.

$$X^- = (3+8+6+10+12+9+11+10+12+7)/10$$

$$= 88/10$$

$$= 8.8$$

Step 2: Make a table as following with three columns, one for the X values, the second for the deviations and the third for squared deviations.

Value (X)	$X-X^-$	$(X-X^-)^2$
3	-5.8	33.64
8	-0.8	0.64

6	-2.8	7.84
10	1.2	1.44
12	3.2	10.24
9	0.2	0.04
11	2.2	4.84
10	1.2	1.44
12	3.2	10.24
7	-1.8	3.24
Total	0	73.6

Step 3:

As the data is not given as sample data, thus we use the formula for population variance.

$$= 73.610$$

$$= 7.36$$

Thus standard Deviation = $\sqrt{7.36}$

Therefore, standard deviation = 2.71

What is Mean Deviation?

Deviation in statistics is a measure that refers to the difference between the observed and expected values of a variable. In layman's terms, a deviation is a distance from the centre point. Mean,

median, and mode are all data set centre points. Similarly, the mean deviation is used to calculate the distance between the values in a data collection and the centre point.

Mean deviation is a statistical measure that computes the average deviation from the average value of a given data collection. The mean deviation can be calculated using various data series, such as – continuous data series, discrete data series and individual data series.

Mean Deviation Formula

The mean deviation is the mean of the absolute deviations of the observations or values from a suitable average. This suitable average may be the mean, median or mode. We also know it as the mean absolute deviation.

The basic formula to calculate mean deviation for a given data set is as follows:

$$\text{Mean Deviation} = \frac{\Sigma |X - \bar{X}|}{N}$$

where,

X = denotes each value in the data set

\bar{X} = denotes the mean value of the data set

N = total number of data values

$| \quad |$ = represents absolute value, i.e. it ignores the sign

How to calculate Mean Deviation?

Step 1 – Calculate the mean, median or mode value of the given data set.

Step 2 – Then we must find the absolute difference between each value in the data set with the mean, ignoring the signs.

Step 3 – We then sum up all the deviations.

Step 4 – Finally, we find the mean or average of those values found in Step 3. The result obtained is the mean deviation.

Let us look at a simple example to understand the working of the above steps. Suppose we have a dataset {2, 4, 8, 10} and we want to calculate the mean deviation about the mean.

Step 1 – We find the mean of the dataset i.e. $(2+4+8+10)/4 = 6$.

Step 2 – We then subtract each value in the dataset with the mean, get their absolute values i.e. $|2-6| = 4$, $|4-6| = 2$, $|8-6| = 2$, $|10-6| = 4$

Step 3 – And add them i.e. $4+2+2+4 = 12$.

Step 4 – Finally, we divide this sum by the total number of values in the dataset (4) that will give us the mean deviation. The answer is $12/4 = 3$.

Mean Deviation Types

Individual Data Series – When the given data set is on an individual basis.

Age	5	10	15	20	25	30
-----	---	----	----	----	----	----

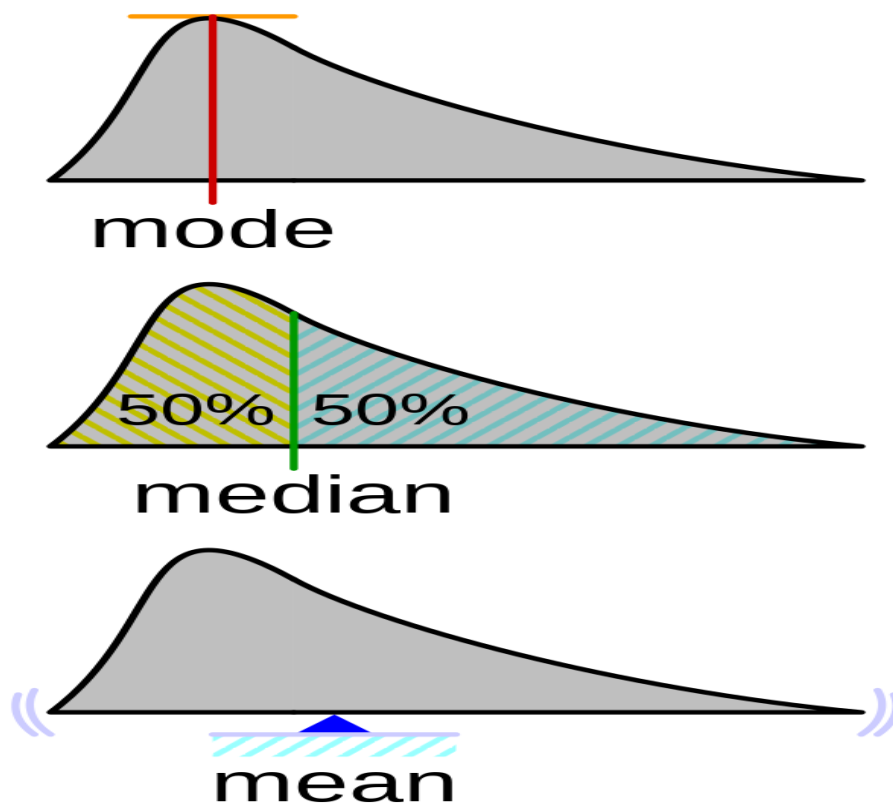
Discrete Data Series – When the data set is given along with their frequencies.

Shoe Size	7	8	9	10
Frequency	5	10	7	13

Continuous Data Series – When the data set is based on ranges along with their frequencies.

Age	5-10	10-15	15-20	20-25
Frequency	19	23	31	26

We shall now learn more about some important formulas, for example, the mean deviation formula for ungrouped data, ungrouped data as well as for an individual series or a continuous series, etc.



Mean Deviation Formula for Ungrouped Data

Ungrouped data is data that has not been sorted or categorized into groups and is still in its raw form. Generally, ungrouped data includes individual data series. The formula to calculate mean deviation for ungrouped data is as follows:

$$\text{Mean Deviation} = \frac{\sum_1^n |x_i - \bar{x}|}{n}$$

where,

x_i = i^{th} observation

\bar{x} = central point of the data (mean, median or mode)

n = number of observations

Mean Deviation Formula for Grouped Data

Grouped data is data that has been sorted and classified into groups. Continuous and discrete frequency distributions are used to group data.

- **For Discrete Frequency Distribution** – The formula to calculate mean deviation is as follows:

$$\text{Mean Deviation} = \frac{\sum_1^n f_i |x_i - \bar{x}|}{\sum_1^n f_i}$$

where,

x_i = specified individual observation

f_i = frequency of the occurrence of that observation

- **For Continuous Frequency Distribution** – The formula to calculate mean deviation is as follows:

$$\text{Mean Deviation} = \frac{\sum_1^n f_i |x_i - \bar{x}|}{\sum_1^n f_i}$$

where,

x_i = mid-value of the class intervals

f_i = frequency of repetition of x_i

Mean Deviation about Mean

The mean is calculated by taking the sum of all observations and dividing it by the total number of observations. Formulas for mean deviation about the mean are given below:

- **For Individual Data –**

$$\text{Mean Deviation} = \frac{\sum_1^n |x_i - \mu|}{n}$$

$$\text{where the mean is } \mu = \frac{x_1 + x_2 + \dots + x_n}{n}$$

- **For Continuous/Discrete Data –**

$$\text{Mean Deviation} = \frac{\sum_1^n f_i |x_i - \mu|}{\sum_1^n f_i}$$

$$\text{where the mean of grouped data is } \mu = \frac{\sum_1^n f_i x_i}{\sum_1^n f_i}$$

Mean Deviation about Median

The median is the value that separates the lower and upper halves of the data. The median is the number in the middle of a sorted, ascending or descending list of numbers. Formulas for mean deviation about the median are given below:

- **For Individual Data –**

$$\text{Mean Deviation} = \frac{\sum_1^n |x_i - M|}{n}$$

where median (M) –

- if n is odd – $M = \left(\frac{n+1}{2}\right)^{\text{th}}$ observation
- if n is even – $M = \frac{\frac{n}{2}^{\text{th}} \text{ obs} + \left(\frac{n}{2} + 1\right)^{\text{th}} \text{ obs}}{2}$

- **For Discrete Data –**

$$\text{Mean Deviation} = \frac{\sum_1^n f_i |x_i - M|}{\sum_1^n f_i}$$

The median of discrete data is calculated as above

- **For Continuous Data –**

$$\text{Mean Deviation} = \frac{\sum_1^n f_i |x_i - M|}{\sum_1^n f_i}$$

$$\text{where the median of continuous data (M)} = l + \frac{\frac{\sum_1^n f_i}{2} - c.f}{f} \times h$$

where,

c.f = cumulative frequency preceding the median class

l = lower value of the median class

f = frequency of the median class

h = length of the median class

Mean Deviation about Mode

The value that occurs the most frequently in a given data collection is defined as the mode. Formulas to calculate mean deviation about the mode is given below:

- **For Individual Data –**

$$\text{Mean Deviation} = \frac{\sum_1^n |x_i - \text{mode}|}{n}$$

where mode = most frequently occurring value in the data set

- **For Discrete Data –**

$$\text{Mean Deviation} = \frac{\sum_1^n f_i |x_i - \text{mode}|}{\sum_1^n f_i}$$

The mode of discrete data is calculated as above.

- **For Continuous Data –**

$$\text{Mean Deviation} = \frac{\sum_1^n f_i |x_i - \text{mode}|}{\sum_1^n f_i}$$

$$\text{where the mode of continuous data} = l + \left(\frac{f - f_1}{2f - f_1 - f_2} \right) \times h$$

where,

l = lower value of the modal class

h = size of the modal class

f = frequency of the modal class

f_1 = frequency of the class preceding the modal class

f_2 = frequency of the class succeeding the modal class

Difference between Mean Deviation and Standard Deviation

Mean Deviation	Standard Deviation
We use central points (mean, median, mode) to calculate the mean deviation.	To calculate the standard deviation we only use the mean.
To calculate the mean deviation, we take the absolute value of the deviations.	We use the square of the deviations to calculate the standard deviation.
It is less frequently used.	It is one of the most commonly used measures of variability and frequently used.
When there are a greater number of outliers in the data, mean absolute deviation is employed.	When there are fewer outliers in the data, the standard deviation is employed.

Merits of Mean Deviation

1. Mean deviation is easy to understand and calculate.
2. It gets least affected by extreme values (outliers).

3. Mean Deviation is calculated by considering all the items in the data set.
4. When compared to other statistical measures, it exhibits the fewest sample volatility.
5. It is a useful comparison metric because it is based on deviations from the mean.

Demerits of Mean Deviation

1. It is not strictly defined because it can be calculated in relation to the mean, median, and mode.
2. Because we take the absolute value, we ignore both negative and positive indications. This can result in inaccuracies in the outcome.
3. It is not a well-defined statistic because the mean deviation from different averages (mean, median, and mode) will differ.
4. It cannot be further algebraically treated.
5. The fluctuations in sampling have a significant impact on it.

Formula for the Co-efficient of Mean Deviation

- Co-efficient of Mean Deviation from Mean = $\frac{M.D}{\bar{X}}$
- Co-efficient of Mean Deviation from Median = $\frac{M.D}{M}$
- The Co-efficient of Mean Deviation from Mode = $\frac{M.D}{Mode}$

Example:

Calculate the mean deviation from the median and the co-efficient of mean deviation from the following data:

Marks of the students: 86, 25, 87, 65, 58, 45, 12, 71, 35.

Solution:

Arrange the data in ascending order: 12, 25, 35, 45, 58, 65, 71, 86, 87.

Median = Value of the $(N+1)/2$ th term

= Value of the $(9+1)/2$ th term

= 58

Calculation of mean deviation:

X	 X-M
12	46
25	33
35	23
45	13
58	0
65	7
71	13
86	28
87	29
N = 9	$\Sigma X-M =460$

$$\text{M.D.} = \Sigma(|X-M|)/N$$

$$= 460/9$$

$$= 51.11$$

$$\text{Co-efficient of Mean Deviation from Median} = \text{M.D./M}$$

$$= 51.11/58$$

$$= 0.881$$

Example:

Calculate the mean deviation from mean for the following data.

x	12	9	6	18	10
f	7	3	8	1	2

Answer.

x	f	x.f	 x - μ 	f. x - μ
12	7	84	2.619	18.33
9	3	27	0.381	1.143
6	8	48	3.381	27.048
18	1	18	8.619	8.619
10	2	20	0.619	1.238
Total	21	197		56.378

We first find the Mean of the given dataset,

$$\text{Mean } (\mu) = \frac{\sum_1^5 f_i x_i}{\sum_1^5 f_i} = \frac{197}{21} = 9.381$$

Finally, we substitute values in the mean deviation about mean formula,

$$\text{Mean Deviation} = \frac{\sum_1^5 f_i |x_i - \mu|}{\sum_1^5 f_i} = \frac{56.378}{21} = 2.684$$

Hence, the mean deviation about the mean is found to be 2.684

Example:

Calculate the mean deviation for the following data.

Class Interval	0 – 2	2 – 4	4 – 6	6 – 8
Frequency	4	2	5	3

Answer.

Class Interval	Mid-point (x)	Frequency (f)	f.x	 x – μ = x – 4 	f. x – μ
0 – 2	1	4	4	3	12
2 – 4	3	2	6	1	2
4 – 6	5	5	25	1	5
6 – 8	7	3	21	3	9
Total		14	56		28

$$\text{Mean } (\mu) = \frac{\sum_1^n f_i x_i}{\sum_1^n f_i} = \frac{56}{14} = 4$$

Finally, we substitute values in the mean deviation formula,

$$\text{Mean Deviation} = \frac{\sum_1^n f_i |x_i - \mu|}{\sum_1^n f_i} = \frac{28}{14} = 2$$