Prerequisites to Machine Learning:

- **Statistics**
- Probability
- Linear Algebra
- Calculus
- Programming Language

Statistics is the study of the collection, analysis, interpretation, presentation, and organization of data. In other words, it is a mathematical discipline to collect, summarize data. Also, we can say that statistics is a branch of applied mathematics. However, there are two important and basic ideas involved in statistics; they are uncertainty and variation. The uncertainty and variation in different fields can be determined only through statistical analysis. These uncertainties are basically determined by the probability that plays an important role in statistics.

What is Statistics?

Statistics is simply defined as the study and manipulation of data. As we have already discussed in the introduction that statistics deals with the analysis and computation of numerical data. Let us see more definitions of statistics given by different authors here.

According to **Merriam-Webster dictionary**, statistics is defined as “classified facts representing the conditions of a people in a state – especially the facts that can be stated in numbers or any other tabular or classified arrangement”.

According to statistician **Sir Arthur Lyon Bowley**, statistics is defined as “Numerical statements of facts in any department of inquiry placed in relation to each other”.

Statistics Examples

Some of the real-life examples of statistics are:

- To find the mean of the marks obtained by each student in the class whose strength is 50. The average value here is the statistics of the marks obtained.
- Suppose you need to find how many members are employed in a city. Since the city is populated with 15 lakh people, hence we will take a survey here for 1000 people (sample). Based on that, we will create the data, which is the statistic.

Basics of Statistics

The basics of statistics include the measure of central tendency and the measure of dispersion. The central tendencies are mean, median and mode and dispersions comprise variance and standard deviation.

Mean is the average of the observations. Median is the central value when observations are arranged in order. The mode determines the most frequent observations in a data set.

Variation is the measure of spread out of the collection of data. Standard deviation is the measure of the dispersion of data from the mean. The square of standard deviation is equal to the variance.

Mathematical Statistics

Mathematical statistics is the application of Mathematics to Statistics, which was initially conceived as the science of the state — the collection and analysis of facts about a country: its economy, and, military, population, and so forth.

Mathematical techniques used for different analytics include mathematical analysis, linear algebra, stochastic analysis, differential equation and measure-theoretic probability theory.

Types of Statistics

Basically, there are two types of statistics.

- Descriptive Statistics
- Inferential Statistics

In the case of descriptive statistics, the data or collection of data is described in summary. But in the case of inferential stats, it is used to explain the descriptive one. Both these types have been used on large scale.

Descriptive Statistics

The data is summarised and explained in descriptive statistics. The summarization is done from a population sample utilising several factors such as mean and standard deviation. Descriptive statistics is a way of organising, representing, and explaining a set of data using charts, graphs, and summary measures. Histograms, pie charts, bars, and scatter plots are common ways to summarise data and present it in tables or graphs. Descriptive statistics are just that: descriptive. They don't need to be normalised beyond the data they collect.

Inferential Statistics

We attempt to interpret the meaning of descriptive statistics using inferential statistics. We utilise inferential statistics to convey the meaning of the collected data after it has been collected, evaluated, and summarised. The probability principle is used in inferential statistics to determine if patterns found in a study sample may be extrapolated to the wider population from which the sample was drawn. Inferential statistics are used to test hypotheses and study correlations between variables, and they can also be used to predict population sizes. Inferential statistics are used to derive conclusions and inferences from samples, i.e. to create accurate generalisations.

Statistics Formulas

The formulas that are commonly used in statistical analysis are given in the table below.

Sample Mean, x¯	∑xn
Population Mean, μ	∑xN
Sample Standard Deviation, (s)	∑(x−x¯)2n−1
Population Standard Deviation, σ	σ=(x−μ)2N
Sample Variance, s2	s2=∑(xi−x¯)2n−1
Population Variance, σ2	σ2=∑(xi−μ)2N
Range, (R)	Largest data value – smallest data value

Summary Statistics

In Statistics, summary statistics are a part of descriptive statistics (Which is one of the types of statistics), which gives the list of information about sample data. We know that statistics deals with the presentation of data visually and quantitatively. Thus, summary statistics deals with summarizing the statistical information. Summary statistics generally deal with condensing the data in a simpler form, so that the observer can understand the information at a glance. Generally, statisticians try to describe the observations by finding:

- The measure of central tendency or mean of the locations, such as arithmetic mean.
- The measure of distribution shapes like skewness or kurtosis.
- The measure of dispersion such as the standard mean absolute deviation.
- The measure of statistical dependence such as correlation coefficient.

Scope of Statistics

Statistics is used in many sectors such as psychology, geology, sociology, weather forecasting, probability and much more. The goal of statistics is to gain understanding from the data, it focuses on applications, and hence, it is distinctively considered as a mathematical science.

Methods in Statistics

The methods involve collecting, summarizing, analysing, and interpreting variable numerical data. Here some of the methods are provided below.

- Data collection
- Data summarization
- Statistical analysis

What is Data in Statistics?

Data is a collection of facts, such as numbers, words, measurements, observations etc.

Types of Data

1. **Qualitative data**- it is descriptive data.

   Example- She can run fast, He is thin.

2. **Quantitative data**- it is numerical information.

   Example- An Octopus is an Eight legged creature.

Types of quantitative data

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

Representation of Data

There are different ways to represent data such as through graphs, charts or tables. The general representation of statistical data are:

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

Measures of Central Tendency

In Mathematics, statistics are used to describe the central tendencies of the grouped and ungrouped data. The three measures of central tendency are:

- Mean
- Median
- Mode

All three measures of central tendency are used to find the central value of the set of data.

Measures of Dispersion

In statistics, the dispersion measures help interpret data variability, i. e. to understand how homogenous or heterogeneous the data is. In simple words, it indicates how squeezed or scattered the variable is. However, there are two types of dispersion measures, absolute and relative. They are tabulated as below:

Absolute measures of dispersion	**Relative measures of dispersion**
1. Range 2. Variance 3. Standard deviation 4. Quartiles and Quartile deviation 5. Mean and Mean deviation	1. Co-efficient of Range 2. Co-efficient of Variation 3. Co-efficient of Standard Deviation 4. Co-efficient of Quartile Deviation 5. Co-efficient of Mean Deviation

Skewness in Statistics

Skewness, in statistics, is a measure of the asymmetry in a probability distribution. It measures the deviation of the curve of the normal distribution for a given set of data.

The value of skewed distribution could be positive or negative or zero. Usually, the bell curve of normal distribution has zero skewness.

ANOVA Statistics

ANOVA Stands for Analysis of Variance. It is a collection of statistical models, used to measure the mean difference for the given set of data.

Degrees of freedom

In statistical analysis, the degree of freedom is used for the values that are free to change. The independent data or information that can be moved while estimating a parameter is the degree of freedom of information.

Applications of Statistics

Statistics have huge applications across various fields in Mathematics as well as in real life. Some of the applications of statistics are given below:

- Applied statistics, theoretical statistics and mathematical statistics
- Machine learning and data mining
- Statistics in society
- Statistical computing
- Statistics applied to the mathematics of the arts

In statistics, the mean is one of the measures of central tendency, apart from the mode and median. Mean is nothing but the average of the given set of values. It denotes the equal distribution of values for a given data set. The mean, median and mode are the three commonly used measures of central tendency. To calculate the mean, we need to add the total values given in a datasheet and divide the sum by the total number of values. The Median is the middle value of a given data when all the values are arranged in ascending order. Whereas mode is the number in the list, which is repeated a maximum number of times.

Definition of Mean in Statistics

Mean is the average of the given numbers and is calculated by dividing the sum of given numbers by the total number of numbers.

Mean = (Sum of all the observations/Total number of observations)

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

In the case of a discrete probability distribution of a random variable X, the mean is equal to the sum over every possible value weighted by the probability of that value; that is, it is computed by taking the product of each possible value x of X and its probability P(x) and then adding all these products together.

Mean Symbol (X Bar)

The symbol of mean is usually given by the symbol '$\bar{x}$'. The bar above the letter x, represents the mean of x number of values.

$\bar{X}$ = (Sum of values ÷ Number of values)

$\bar{X} = (x_1 + x_2 + x_3 + \ldots + x_n)/n$

Mean Formula

The basic formula to calculate the mean is calculated based on the given data set. Each term in the data set is considered while evaluating the mean. The general formula for mean is given by the ratio of the sum of all the terms and the total number of terms. Hence, we can say;

Mean = Sum of the Given Data/Total number of Data

To calculate the arithmetic mean of a set of data we must first add up (sum) all of the data values (x) and then divide the result by the number of values (n). Since ∑ is the symbol used to indicate that values are to be summed (see Sigma Notation) we obtain the following formula for the mean (x̄):

x̄=∑x/n

How to Find Mean?

As we know, data can be grouped data or ungrouped data so to find the mean of given data we need to check whether the given data is ungrouped. The formulas to find the mean for ungrouped data and grouped data are different. In this section, you will learn the method of finding the mean for both of these instances.

Mean for Ungrouped Data

The example given below will help you in understanding **how to find the mean** of ungrouped data.

Example:

In a class there are 20 students and they have secured a percentage of 88, 82, 88, 85, 84, 80, 81, 82, 83, 85, 84, 74, 75, 76, 89, 90, 89, 80, 82, and 83.

Find the mean percentage obtained by the class.

Solution:

Mean = Total of percentage obtained by 20 students in class/Total number of students

= [88 + 82 + 88 + 85 + 84 + 80 + 81 + 82 + 83 + 85 + 84 + 74 + 75 + 76 + 89 + 90 + 89 + 80 + 82 + 83]/20

= 1660/20

= 83

Hence, the mean percentage of each student in the class is 83%.

Mean for Grouped Data

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

Direct method:

Mean,x—=∑i=1nfixi∑i=1nfi

Assumed mean method:

Mean,(x—)=a+∑fidi∑fi

Step-deviation method:

Mean,(x—)=a+h∑fiui∑fi

Go through the example given below to understand how to calculate the mean for grouped data.

Example:

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

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.

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=x1x2x3…..xnn

Example: Find the geometric mean of 4 and 3 ?

GeometricMean=4×3=23=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)=n1x1+1x2+1x2+1x3+……1xn

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.

Xrms=x12+x22+x32….xn2n

Contraharmonic Mean

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

(x12+x22+….+xn2)(x1+x2+…..xn)

Real-life Applications of Mean

In the real world, when there is huge data available, we use statistics to deal with it. Suppose, in a data table, the price values of 10 clothing materials are mentioned. If we have to find the mean of the prices, then add the prices of each clothing material and divide the total sum by 10. It will result in an average value. Another example is that if we have to find the average age of students of a class, we have to add the age of individual students present in the class and then divide the sum by the total number of students present in the class.

Practice Problems

Q.1: Find the mean of 5,10,15,20,25.

Q.2: Find the mean of the given data set: 10,20,30,40,50,60,70,80,90.

Q.3: Find the mean of the first 10 even numbers.

Q.4: Find the mean of the first 10 odd numbers.

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

Example: The median of 2,3,4 is 3.

In Maths, the median is also a type of average, which is used to find the centre value. Therefore, it is also called measure of central tendency.

Apart from the median, the other two central tendencies are mean and mode. Mean is the ratio of the sum of all observations and total number of observations. Mode is the value in th e given data-set, repeated most of the time.

In geometry, a median is also defined as the centre point of a polygon. For example, the median of a triangle is the line segment joining the vertex of a triangle to the centre of the opposite sides. Therefore, a median bisects the sides of a triangle.

Median in Statistics

The median of a set of data is the middlemost number or centre value in the set. The median is also the number that is halfway into the set.

To find the median, the data should be arranged first in order of least to greatest or greatest to the least value. A median is a number that is separated by the higher half of a data sample, a population or a probability distribution from the lower half. The median is different for different types of distribution.

For example, the median of 3, 3, 5, 9, 11 is 5. If there is an even number of observations, then there is no single middle value; the median is then usually defined to be the mean of the two middle values: so the median of 3, 5, 7, 9 is (5+7)/2 = 6.

Median Formula

The formula to calculate the median of the finite number of data set is given here. The median formula is different for even and odd numbers of observations. Therefore, it is necessary to recognise first if we have odd number of values or even number of values in a given data set.

The formula to calculate the median of the data set is given as follows.

Odd Number of Observations

If the total number of observations given is odd, then the formula to calculate the median is:

Median=(n+12)th term

where n is the number of observations

Even Number of Observations

If the total number of observation is even, then the median formula is:

Median=(n2)th term+(n2+1)th term2

where n is the number of observations

How to Calculate the Median?

To find the median, place all the numbers in ascending order and find the middle.

Example 1:

Find the Median of 14, 63 and 55

solution:

Put them in ascending order: 14, 55, 63

The middle number is 55, so the median is 55.

Example 2:

Find the median of the following:

4, 17, 77, 25, 22, 23, 92, 82, 40, 24, 14, 12, 67, 23, 29

Solution:

When we put those numbers in the order, we have:

4, 12, 14, 17, 22, 23, 23, 24, 25, 29, 40, 67, 77, 82, 92,

There are fifteen numbers. Our middle is the eighth number:

The median value of this set of numbers is 24.

Example 3:

Rahul's family drove through 7 states on summer vacation. The prices of Gasoline differ from state to state. Calculate the median of gasoline cost.

1.79, 1.61, 2.09, **1.84,** 1.96, 2.11, 1.75

Solution:

By organizing the data from smallest to greatest, we get:

1.61, 1.75, 1.79, 1.84 , 1.96, 2.09, 2.11

Hence, the median of gasoline cost is 1.84. There are three states with greater gasoline costs and 3 with smaller prices.

In statistics, the **mode** is the value that is repeatedly occurring in a given set. We can also say that the value or number in a data set, which has a high frequency or appears more frequently, is called mode or **modal value**. It is one of the three measures of central tendency, apart from mean and median. For example, the mode of the set {3, 7, 8, 8, 9}, is 8. Therefore, for a finite number of observations, we can easily find the mode. A set of values may have one mode or more than one mode or no mode at all.

In this article, you will understand the meaning of mode in statistics, formula for mode for grouped data and how to find the mode for the given data, i.e. for ungrouped and grouped data along with solved examples in detail.

Mode Definition in Statistics

A mode is defined as the value that has a higher frequency in a given set of values. It is the value that appears the most number of times.

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.

Statistics deals with the presentation, collection and analysis of data and information for a particular purpose. We use tables, graphs, pie charts, bar graphs, pictorial representation, etc. After the proper organization of the data, it must be further analyzed to infer helpful information.

For this purpose, frequently in statistics, we tend to represent a set of data by a representative value that roughly defines the entire data collection. This representative value is known as the measure of central tendency. By the name itself, it suggests that it is a value around which the data is centred. These measures of central tendency allow us to create a statistical summary of the vast, organized data. One such measure of central tendency is the mode of data.

Bimodal, Trimodal & Multimodal (More than one mode)

- When there are two modes in a data set, then the set is called **bimodal**

For example, The mode of Set A = {2,2,2,3,4,4,5,5,5} is 2 and 5, because both 2 and 5 is repeated three times in the given set.

- When there are three modes in a data set, then the set is called **trimodal**

For example, the mode of set A = {2,2,2,3,4,4,5,5,5,7,8,8,8} is 2, 5 and 8

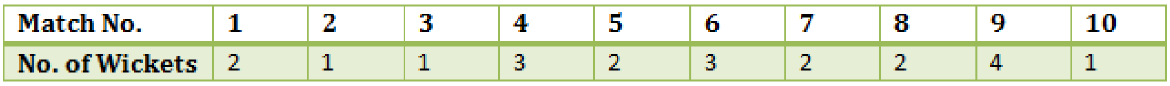
- When there are four or more modes in a data set, then the set is called **multimodal**

Mode Formula in Statistics (Ungrouped Data)

The value occurring most frequently in a set of observations is its mode. In other words, the mode of data is the observation having the highest frequency in a set of data. There is a possibility that more than one observation has the same frequency, i.e. a data set could have more than one mode. In such a case, the set of data is said to be multimodal.

Let us look into an example to get a better insight.

Example: The following table represents the number of wickets taken by a bowler in 10 matches. Find the mode of the given set of data.



Match No.	1	2	3	4	5	6	7	8	9	10
No. of Wickets	2	1	1	3	2	3	2	2	4	1

It can be seen that 2 wickets were taken by the bowler frequently in different matches. Hence, the mode of the given data is 2.

Mode Formula For Grouped Data

In the case of grouped frequency distribution, calculation of mode just by looking into the frequency is not possible. To determine the mode of data in such cases we calculate the modal class. Mode lies inside the modal class. The mode of data is given by the formula:

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

Where,

l = lower limit of the modal class

h = size of the class interval

f1 = frequency of the modal class

f0 = frequency of the class preceding the modal class

f2 = frequency of the class succeeding the modal class

Let us take an example to understand this clearly.

How to Find the Mode

Let us learn here how to find the mode of a given data with the help of examples.

Example 1: Find the mode of the given data set: 3, 3, 6, 9, 15, 15, 15, 27, 27, 37, 48.

Solution: In the following list of numbers,

3, 3, 6, 9, 15, 15, 15, 27, 27, 37, 48

15 is the mode since it is appearing more number of times in the set compared to other numbers.

Example 2: Find the mode of 4, 4, 4, 9, 15, 15, 15, 27, 37, 48 data set.

Solution: Given: 4, 4, 4, 9, 15, 15, 15, 27, 37, 48 is the data set.

As we know, a data set or set of values can have more than one mode if more than one value occurs with equal frequency and number of time compared to the other values in the set.

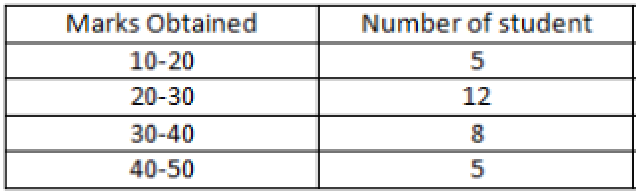
Hence, here both the number 4 and 15 are modes of the set.

Example 3: Find the mode of 3, 6, 9, 16, 27, 37, 48.

Solution: If no value or number in a data set appears more than once, then the set has no mode.

Hence, for set 3, 6, 9, 16, 27, 37, 48, there is no mode available.

Example 4: In a class of 30 students marks obtained by students in mathematics out of 50 is tabulated as below. Calculate the mode of data given.



Marks Obtained	Number of student
10-20	5
20-30	12
30-40	8
40-50	5

Solution:

The maximum class frequency is 12 and the class interval corresponding to this frequency is 20 – 30. Thus, the modal class is 20 – 30.

Lower limit of the modal class (l) = 20

Size of the class interval (h) = 10

Frequency of the modal class (f1) = 12

Frequency of the class preceding the modal class (f0) = 5

Frequency of the class succeeding the modal class (f2)= 8

Substituting these values in the formula we get;

$$Mode = l + \left(\frac{f_1 - f_0}{2f_1 - f_0 - f_2}\right) \times h = 20 + \left(\frac{12 - 5}{2 \times 12 - 5 - 8}\right) \times 10 = 26.364$$

Mean Median Mode Comparison

Mean	**Median**	**Mode**
Mean is the average value that is equal to the ration of sum of values in a data set and total number of values. Mean = Sum of observations/Number of observations	**Median** is the central value of given set of values when arranged in an order.	**Mode** is the most repetitive value of a given set of values.
For example, if we have set of values = 2,2,3,4,5, then;		
Mean = (2+2+3+4+5)/5 = 3.2	**Median = 3**	**Mode = 2**

Mode Median Mean Formula

There exists an empirical relationship between mode, median and mean and this can be expressed using the formula:

Mode=3Median–2Mean

Practice Problems

1. Find the mode of the following marks (out of 10) obtained by 20 students:

4, 6, 5, 9, 3, 2, 7, 7, 6, 5, 4, 9, 10, 10, 3, 4, 7, 6, 9, 9

2. Find the mode for the following data set.

41, 39, 48, 52, 46, 62, 54, 40, 96, 52, 98, 40, 42, 52, 60

3. Find the mode of the given distribution.

Class Interval	10 – 25	25 – 40	40 – 55	55 – 70	70 – 85	85 – 100
Frequency	12	9	17	16	20	16

In Statistics, you might have studied the methods of finding a representative value for the given data, i.e. the **measure of central tendency**. To recall, mean, median, and mode are three measures of central tendency. As we know, measuring central tendency gives us an idea of where data points are centred. However, to interpret the data thoroughly, we should also see how the data are scattered. And how much they have bunched about a measure of central tendency.

Range in Statistics

As the measures of central tendency are not enough to give complete information about a given data. Variability is another determinant that is required to be studied under statistics. Similar to measures of central tendency, we need to have a single number to describe variability. This single number has expressed a measure of dispersion. In this article, you will learn about one of the measures of dispersion called range.

Range Meaning

In statistics, the difference between the highest and lowest observations in a given data is called its Range.

Range Formula

The formula to calculate **the range of a data** set is given below:

Range = Maximum value – Minimum value

However, we can define the formulas to find the range of grouped and ungrouped data.

Range of Ungrouped Data

The formula to find the range of ungrouped data or discrete distribution of data is given as:

Range = Highest value of the data set – Lowest value of the data set

Range of Grouped Data

In the case of continuous frequency distribution or grouped data, the range is defined as the difference between the upper limit of the maximum interval of the grouped data and the lower limit of the minimum interval. It is the simplest measure of dispersion. It gives a comprehensive view of the total spread of the observations. Thus, the formula to calculate the range of a grouped data is given below:

Range = Upper-class boundary of the highest interval – Lower class boundary of the lowest interval

How to Find the Range

To understand the method of calculating range for different types of data in statistics, go through the solved examples given below. These examples help in understanding how to find the range for discrete and continuous data.

Solved Examples

Example 1: Find the range of the data: 21, 6, 17, 18, 12, 8, 4, 13

Solution:

Given,

21, 6, 17, 18, 12, 8, 4, 13

Highest value = 21

Lowest value = 4

Range = Highest value – Lowest value

= 21 – 4

= 17

Example 2: Age (in years) of 6 boys and 6 girls are recorded as below:

Girls	6	7	9	8	10	10
Boys	7	9	12	14	13	17

(a) Find the range for each group.

(b) Find the range if the two groups are combined together.

Solution:

(a) The range for group of girls = 10 – 6 = 4

The range for group of boys = 17 – 7 = 10

(b) If the ages of the group of boys and girls are combined, then the range will be:

17 – 4 = 13

Example 3: Calculate the range for the given frequency distribution.

Class Interval	10 – 20	20 – 30	30 – 40	40 – 50	50 – 60	60 – 70	70 – 80
Frequency	2	3	14	8	3	8	2

Solution:

We know that the range of grouped data is given by the formula:

Range = Upper-class boundary of the highest interval – Lower class boundary of the lowest interval

Here, the Upper-class boundary of the highest interval = 80

Lower class boundary of the lowest interval = 10

Therefore, range = 80 – 10 = 70

Example 4: Find the range of the following data.

CI	16 – 20	21 – 25	26 – 30	31 – 35	36 – 40	41 – 45	46 – 50	51 – 55
f	5	6	12	14	26	12	16	9

Solution:

Given data is not continuous frequency distribution.

Now, we have to convert the given data into continuous frequency distribution by subtracting 0.5 from the lower limit and adding 0.5 to the upper limit of each class interval.

CI	15.5 – 20.5	20.5 – 25.5	25.5 – 30.5	30.5 – 35.5	35.5 – 40.5	40.5 – 45.5	45.5 – 50.5	50.5 – 55.5
f	5	6	12	14	26	12	16	9

Here,

Upper-class boundary of the highest interval = 55.5

Lower class boundary of the lowest interval = 15.5

Therefore, range = 55.5 – 15.5 = 40

Practice Problems

1.

1. The marks in a subject for 12 students are as follows:

31, 37, 35, 38, 42, 23, 17, 18, 35, 25, 35, 29

For the given data, find the range.

2. Given below are heights of 15 students of a class measured in cm:

128, 144, 146, 143, 136, 142, 138, 129, 140, 152, 144, 140, 150, 142, 154

Find the range of the given data.

1. Calculate the range of the data given below:

Class	30 – 40	40 – 50	50 – 60	60 – 70	70 – 80	80 – 90	90 – 100
Frequency	2	3	8	15	12	7	3

Variance is the expected value of the squared variation of a random variable from its mean value, in probability and statistics. Informally, variance estimates how far a set of numbers (random) are spread out from their mean value.

The value of variance is equal to the square of standard deviation, which is another central tool.

Variance is symbolically represented by $\boldsymbol{\sigma^2}$, $\mathbf{s^2}$, or **Var(X).**

The formula for variance is given by:

$Var(X)=E[(X-\mu)2]$

Definition

Variance is a measure of how data points differ from the mean. According to Layman, a variance is a measure of how far a set of data (numbers) are spread out from their mean (average) value.

Variance means to find the expected difference of deviation from actual value. Therefore, variance depends on the standard deviation of the given data set.

The more the value of variance, the data is more scattered from its mean and if the value of variance is low or minimum, then it is less scattered from mean. Therefore, it is called a measure of spread of data from mean.

For the purpose of solving questions, the formula for variance is given by:

$Var(X)=E[(X-\mu)2]$

Put into words; this means that variance is the expectation of the squared deviation of a random set of data from its mean value. Here,

X = Random variable

"μ" is equal to E(X) so the above equation may also be expressed as,

$\mathrm{Var}(X) = E[(X - E(X))^2]$

$\mathrm{Var}(X) = E[\, X^2 - 2X\, E(X) + (E(X))^2]$

$\mathrm{Var}(X) = E(X^2) - 2\, E(X)\, E(X) + (E(X))^2$

$\mathrm{Var}(X) = E(X^2) - (E(X))^2$

Sometimes the covariance of the random variable itself is treated as the variance of that variable. Symbolically,

Var(X) = Cov(X, X)

Formula

As we know already, the variance is the square of standard deviation, i.e.,

Variance = (Standard deviation)^2= σ^2

The corresponding formulas are hence,

Population standard deviation $\sigma=\sum(X-\mu)2N$

Sample standard deviation $s=\sum(x-x—)2n-1$

Where X (or x) = Value of Observations

μ = Population mean of all Values

n = Number of observations in the sample set

x¯=Sample mean

N = Total number of values in the population

Properties

The variance, var(X) of a random variable X has the following properties.

1. Var(X + C) = Var(X), where C is a constant.

2. Var(CX) = C^2.Var(X), where C is a constant.
3. Var(aX + b) = a^2.Var(X), where a and b are constants.
4. If X_1, X_2,……., X_n are n independent random variables, then

$Var(X_1 + X_2 + \ldots\ldots + X_n) = Var(X_1) + Var(X_2) + \ldots\ldots\ldots + Var(X_n)$.

Now let’s have a look at the relationship between Variance and Standard Deviation.

Variance and Standard Deviation

Standard deviation is the positive square root of the variance. The symbols σ and s are used correspondingly to represent population and sample standard deviations.

Standard Deviation is a measure of how spread out the data is. Its formula is simple; it is the square root of the variance for that data set. It’s represented by the Greek symbol sigma (σ).

How to Calculate Variance

Variance can be calculated easily by following the steps given below:

- Find the mean of the given data set. Calculate the average of a given set of values
- Now subtract the mean from each value and square them
- Find the average of these squared values, that will result in variance

Say if $x_1, x_2, x_3, x_4, \ldots, x_n$ are the given values.

Therefore, the mean of all these values is:

$\bar{x} = (x_1+x_2+x_3+\ldots+x_n)/n$

Now subtract the mean value from each value of the given data set and square them.

$(x_1-\bar{x})^2, (x_2-\bar{x})^2, (x_3-\bar{x})^2, \ldots\ldots, (x_n-\bar{x})^2$

Find the average of the above values to get the variance.

$Var(X) = [(x_1-\bar{x})^2 + (x_2-\bar{x})^2 + (x_3-\bar{x})^2 + \ldots\ldots + (x_n-\bar{x})^2]/n$

Hence, the variance is calculated.

Example of Variance

Let’s say the heights (in mm) are 610, 450, 160, 420, 310.

Mean and Variance is interrelated. The first step is finding the mean which is done as follows,

Mean = (610+450+160+420+310)/ 5 = 390

So the mean average is 390 mm.

To calculate the Variance, compute the difference of each from the mean, square it and find then find the average once again.

So for this particular case the variance is :

$= (220^2 + 60^2 + (-230)^2 + 30^2 + (-80)^2)/5$

= (48400 + 3600 + 52900 + 900 + 6400)/5

Final answer : Variance = 22440

Problem & Solution

Example: Find the variance of the numbers 3, 8, 6, 10, 12, 9, 11, 10, 12, 7.

Solution:

Given,

3, 8, 6, 10, 12, 9, 11, 10, 12, 7

Step 1: Compute the mean of the 10 values given.

Mean = (3+8+6+10+12+9+11+10+12+7) / 10 = 88 / 10 = 8.8

Step 2: Make a table with three columns, one for the X values, the second for the deviations and the third for squared deviations. As the data is not given as sample data so we use the formula for population variance. Thus, the mean is denoted by μ.

Value X	$X - \mu$	$(X - \mu)^2$
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:

$\sigma2=\sum(X-\mu)2N$

$= 73.6 / 10$

$= 7.36$

Points to Remember

1. In statistics, the variance is used to understand how different numbers correlate to each other within a data set, instead of using more comprehensive mathematical methods such as organising numbers of the data set into quartiles.

2. Variance considers all the deviations from the mean are the same despite their direction. However, the squared deviations cannot sum to zero and provide the presence of no variability at all in the given data set.

3. One of the disadvantages of finding variance is that it gives combined weight to extreme values, i.e. the numbers that are far from the mean. When squaring these numbers, there is a chance that they may skew the given data set.

4. Another disadvantage of variance is that sometimes it may conclude complex calculations.

Note: If the data values are identical in a set, then their variance will be zero (0).

Variance and Standard deviation are the two important topics in Statistics. It is the measure of the dispersion of statistical data. Dispersion is the extent to which values in a distribution differ from the average of the distribution. To quantify the extent of the variation, there are certain measures namely:

(i) Range

(ii) Quartile Deviation

(iii) Mean Deviation

(iv) Standard Deviation

The degree of dispersion is calculated by the procedure of measuring the variation of data points. In this article, you will learn what is variance and standard deviation, formulas, and the procedure to find the values with examples.

What are the Variance and Standard Deviation?

In statistics, Variance and standard deviation are related with each other since the square root of variance is considered the standard deviation for the given data set. Below are the definitions of variance and standard deviation.

What is variance?

Variance is the measure of how notably a collection of data is spread out. If all the data values are identical, then it indicates the variance is zero. All non-zero variances are considered to be positive. A little variance represents that the data points are close to the mean, and to each other, whereas if the data points are highly spread out from the mean and from one another indicates the high variance. In short, the variance is defined as the average of the squared distance from each point to the mean.

What is Standard deviation?

Standard Deviation is a measure which shows how much variation (such as spread, dispersion, spread,) from the mean exists. The standard deviation indicates a "typical" deviation from the mean. It is a popular measure of variability because it returns to the original units of measure of the data set. Like the variance, if the data points are close to the mean, there is a small variation whereas the data points are highly spread out from the mean, then it has a high variance. Standard deviation calculates the extent to which the values differ from the average. Standard Deviation, the most widely used measure of dispersion, is based on all values. Therefore a change in even one value affects the value of standard deviation. It is independent of origin but not of scale. It is also useful in certain advanced statistical problems.

Variance and Standard Deviation Formula

The formulas for the variance and the standard deviation is given below:

Standard Deviation Formula

The population standard deviation formula is given as:

σ=1N∑i=1N(Xi−μ)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=1n−1∑i=1n(xi−x—)2

Here,

s = Sample standard deviation

n = Number of observations in sample

xi = ith observation in the sample

x—

= Sample mean

Variance Formula:

The population variance formula is given by:

σ2=1N∑i=1N(Xi−μ)2

The sample variance formula is given by:

s2=1n−1∑i=1n(xi−x—)2

How is Standard Deviation calculated?

The formula for standard deviation makes use of three variables. The first variable is the value of each point within a data set, with a sum-number indicating each additional variable (x, x1, x2, x3, etc). The mean is applied to the values of the variable M and the number of data that is assigned to the variable n. Variance is the average of the values of squared differences from the arithmetic mean.

To calculate the mean value, the values of the data elements have to be added together and the total is divided by the number of data entities that were involved.

Standard deviation, denoted by the symbol σ, describes the square root of the mean of the squares of all the values of a series derived from the arithmetic mean which is also called the root-mean-square deviation. 0 is the smallest value of standard deviation since it cannot be negative. When the elements in a series are more isolated from the mean, then the standard deviation is also large.

The statistical tool of standard deviation is the measures of dispersion that computes the erraticism of the dispersion among the data. For instance, mean, median and mode are the measures of central tendency. Therefore, these are considered to be the central first order averages. The measures of dispersion that are mentioned directly over are averages of deviations that result from the average values, therefore these are called second-order averages.

Standard Deviation Example

Let's calculate the standard deviation for the number of gold coins on a ship run by pirates.

There are a total of 100 pirates on the ship. Statistically, it means that the population is 100. We use the standard deviation equation for the entire population if we know a number of gold coins every pirate has.

Statistically, let's consider a sample of 5 and here you can use the standard deviation equation for this sample population.

This means we have a sample size of 5 and in this case, we use the standard deviation equation for the sample of a population.

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

Mean:

x¯=∑xn

=x1+x2+x3+x4…..+xnn

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

= 5

xn−x¯ for every value of the sample:

x1−x¯=4–5=−1

x2−x¯=2–5=−3

x3−x¯=5–5=0

x4−x¯=8–5=3

x5−x¯=6–5=1

∑(xn−x¯)2

=(x1−x¯)2+(x2−x¯)2+…+(x5−x¯)2

=(−1)2+(−3)2+02+32+12

= 20

Standard deviation:

S.D=∑(xn−x¯)2n−1

=204

= √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. This can be understood with the help of an example.

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 = (∑fxi)/N = 925/55 = 16.818

Variance = 1/(N – 1) [∑fxi2 – 1/N(∑fxi)2]

= 1/(55 – 1) [27575 – (1/55) (925)2]

= (1/54) [27575 – 15556.8182]

= 222.559

Standard deviation = √variance = √222.559 = 14.918

Practice Problems on Standard Deviation

1. Calculate the standard deviation of the following values:

5, 10, 25, 30, 50.

2. Find the mean and standard deviation for the following data.

x	60	61	62	63	64	65	66	67	68
f	2	1	12	29	25	12	10	4	5

3. The diameters of circles (in mm) drawn in a design are given below:

Diameters	33 – 36	37 – 40	41 – 44	45 – 48	49 – 52
No.of circles	15	17	21	22	25

Calculate the standard deviation and mean diameter of the circles.

[Hint: First make the data continuous by making the classes as 32.5-36.5, 36.5-40.5, 40.5-44.5, 44.5 – 48.5, 48.5 – 52.5 and then proceed.]

Outliers are data points that are far from other data points. In other words, they're unusual values in a dataset. Outliers are problematic for many statistical analyses because they can cause tests to either miss significant findings or distort real results.

Unfortunately, there are no strict statistical rules for definitively identifying outliers. Finding outliers depends on subject-area knowledge and an understanding of the data collection process. While there is no solid mathematical definition, there are guidelines and statistical tests you can use to find outlier candidates.

The Tukey method for finding outliers uses the interquartile range to filter out very large or very small numbers. It's practically the same as the procedure above, but you might see the formulas written slightly differently and the terminology is a little different as well. For example, the Tukey method uses the concept of "fences".

The formulas are:

Low outliers = Q1 – 1.5(Q3 – Q1) = Q1 – 1.5(IQR)

High outliers = Q3 + 1.5(Q3 – Q1) = Q3 + 1.5(IQR)

Where:

Q1 = first quartile

Q3 = third quartile

IQR = Interquartile range

These equations give you two values, or "fences". You can think of them as a fence that cordons off the outliers from all of the values that are contained in the bulk of the data.

Sample question: Use Tukey's method to find outliers for the following set of data: 1,2,5,6,7,9,12,15,18,19,38.

Step 1: Find the Interquartile range:

1. Find the median: 1,2,5,6,7,9,12,15,18,19,38.

2. Place parentheses around the numbers above and below the median — it makes Q1 and Q3 easier to find.

(1,2,5,6,7),9,(12,15,18,19,38)

3. Find Q1 and Q3. Q1 can be thought of as a median in the lower half of the data. Q3 can be thought of as a median for the upper half of data.

(1,2,5,6,7), 9, (12,15,18,19,38). Q1=5 and Q3=18.

4. Subtract Q1 from Q3. 18-5=13.

Step 2: Calculate 1.5 * IQR:

1.5 * IQR = 1.5 * 13 = 19.5

Step 3: Subtract from Q1 to get your lower fence:

5 – 19.5 = -14.5

Step 4: Add to Q3 to get your upper fence:

18 + 19.5 = 37.5.

Step 5:Add your fences to your data to identify outliers:

(-14.5) 1,2,5,6,7,9,12,15,18,19,(37.5),38.

Anything outside of the fences is an outlier. For this data set, 38 is the only outlier.

That's how to find outliers with the Tukey method!

How to Find Outliers Using the Interquartile Range(IQR)

An outlier is defined as being any point of data that lies over 1.5 IQRs below the first quartile (Q1) or above the third quartile (Q3)in a data set.

High = (Q3) + 1.5 IQR

Low = (Q1) – 1.5 IQR

Example Question: Find the outliers for the following data set: 3, 10, 14, 22, 19, 29, 70, 49, 36, 32.

Step 1: **Find the IQR, Q1(25th percentile) and Q3(75th percentile).** Follow the steps: Interquartile Range in Statistics: How to find it.

IQR = 22

Q1 = 14

Q3 = 36

Step 2: Multiply the IQR you found in Step 1 by 1.5:

IQR * 1.5 = 22 * 1.5 = 33.

Step 3: Add the amount you found in Step 2 to Q3 from Step 1:

33 + 36 = 69.

This is your upper limit. Set this number aside for a moment.

Step 3: Subtract the amount you found in Step 2 from Q1 from Step 1:

14 – 33 = -19.

This is your lower limit. Set this number aside for a moment.

Step 5: Put the numbers from your data set in order:

3, 10, 14, 19, 22, 29, 32, 36, 49, 70

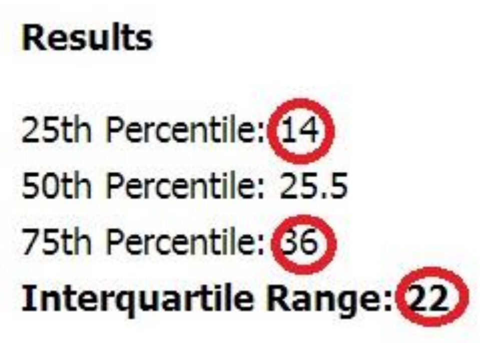
Step 6: Insert your low and high values into your data set, in order:

-19, 3, 10, 14, 19, 22, 29, 32, 36, 49, 69, 70

Step 6: Highlight any number below or above the numbers you inserted in Step 6:

-19, 3, 10, 14, 19, 22, 29, 32, 36, 49, 69, 70

That's it!



Results

25th Percentile: 14
50th Percentile: 25.5
75th Percentile: 36
Interquartile Range: 22

In statistics, a histogram is a graphical representation of the distribution of data. The histogram is represented by a set of rectangles, adjacent to each other, where each bar represent a kind of data. Statistics is a stream of mathematics that is applied in various fields. When numerals are repeated in statistical data, this repetition is known as Frequency and which can be written in the form of a table, called a frequency distribution. A Frequency distribution can be shown graphically by using different types of graphs and a Histogram is one among them. In this article, let us discuss in detail about what is a histogram, how to create the histogram for the given data, different types of the histogram, and the difference between the histogram and bar graph in detail.

What is Histogram?

A histogram is a graphical representation of a grouped frequency distribution with continuous classes. It is an area diagram and can be defined as a set of rectangles with bases along with the intervals between class boundaries and with areas proportional to frequencies in the corresponding classes. In such representations, all the rectangles are adjacent since the base covers the intervals between class boundaries. The heights of rectangles are proportional to corresponding frequencies of similar classes and for different classes, the heights will be proportional to corresponding frequency densities.

In other words, a histogram is a diagram involving rectangles whose area is proportional to the frequency of a variable and width is equal to the class interval.

How to Plot Histogram?

You need to follow the below steps to construct a histogram.

1. Begin by marking the class intervals on the X-axis and frequencies on the Y-axis.

2. The scales for both the axes have to be the same.

3. Class intervals need to be exclusive.

4. Draw rectangles with bases as class intervals and corresponding frequencies as heights.

5. A rectangle is built on each class interval since the class limits are marked on the horizontal axis, and the frequencies are indicated on the vertical axis.

6. The height of each rectangle is proportional to the corresponding class frequency if the intervals are equal.

7. The area of every individual rectangle is proportional to the corresponding class frequency if the intervals are unequal.

Although histograms seem similar to graphs, there is a slight difference between them. The histogram does not involve any gaps between the two successive bars.

When to Use Histogram?

The histogram graph is used under certain conditions. They are:

- The data should be numerical.

- A histogram is used to check the shape of the data distribution.

- Used to check whether the process changes from one period to another.

- Used to determine whether the output is different when it involves two or more processes.

- Used to analyse whether the given process meets the customer requirements.

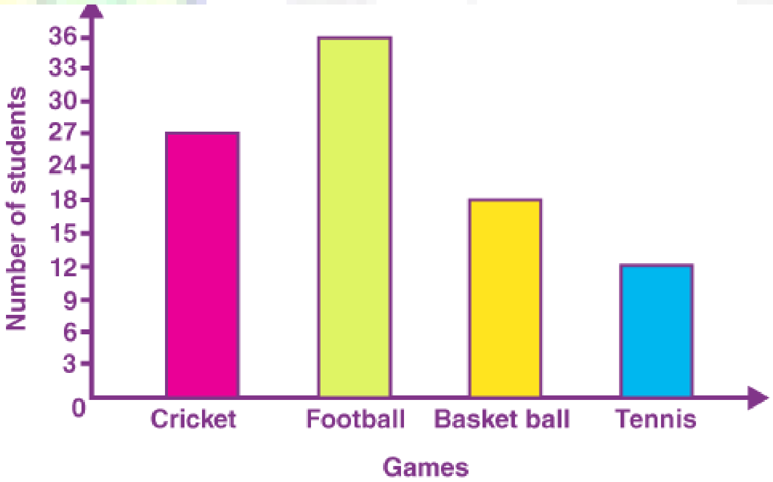
Difference Between Bar Graph and Histogram

A histogram is one of the most commonly used graphs to show the frequency distribution. As we know that the frequency distribution defines how often each different value occurs in the data set. The histogram looks more similar to the bar graph, but there is a difference between them. The list of differences between the bar graph and the histogram is given below:

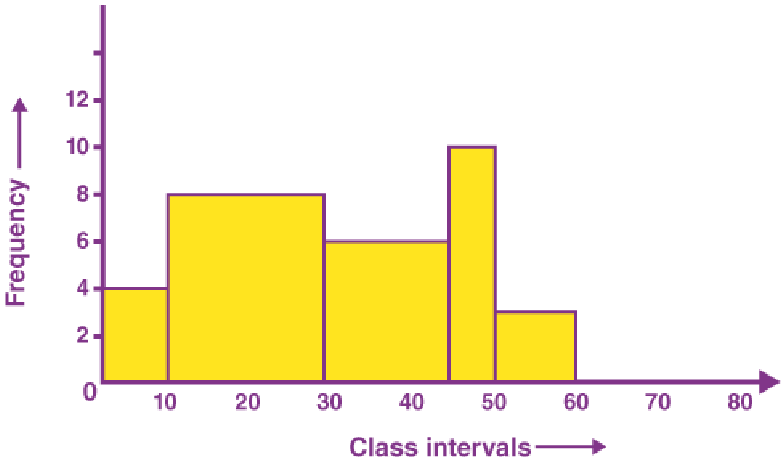
Histogram	**Bar Graph**
It is a two-dimensional figure	It is a one-dimensional figure
The frequency is shown by the area of each rectangle	The height shows the frequency and the width has no significance.
It shows rectangles touching each other	It consists of rectangles separated from each other with equal spaces.

The above differences can be observed from the below figures:

Bar Graph (Gaps between bars)



Histogram (No gaps between bars)

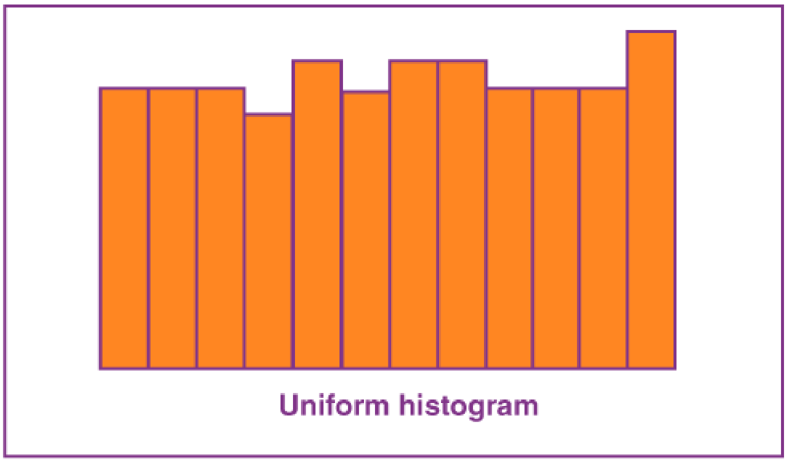


Types of Histogram

The histogram can be classified into different types based on the frequency distribution of the data. There are different types of distributions, such as normal distribution, skewed distribution, bimodal distribution, multimodal distribution, comb distribution, edge peak distribution, dog food distribution, heart cut distribution, and so on. The histogram can be used to represent these different types of distributions. The different types of a histogram are:

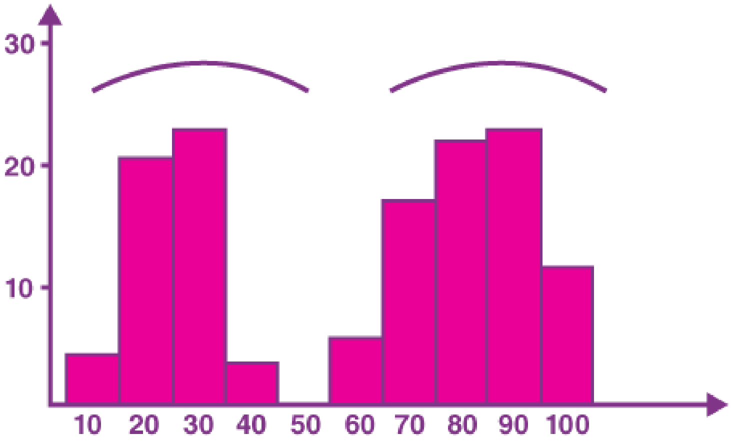
- Uniform histogram
- Symmetric histogram
- Bimodal histogram
- Probability histogram

Uniform Histogram



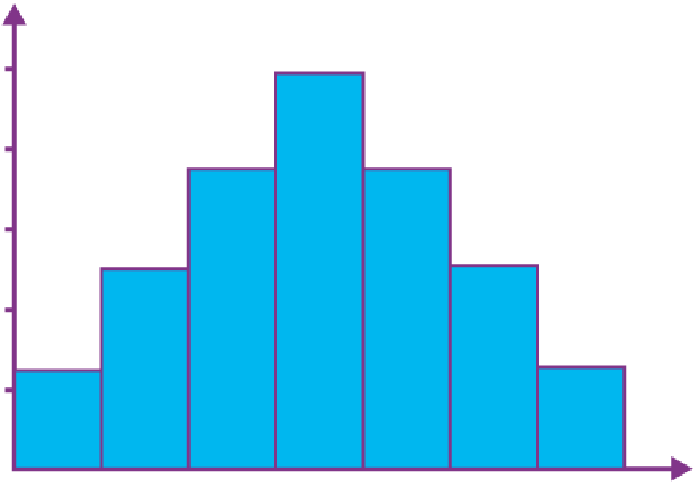
A uniform distribution reveals that the number of classes is too small, and each class has the same number of elements. It may involve distribution that has several peaks.

Bimodal Histogram



If a histogram has two peaks, it is said to be bimodal. Bimodality occurs when the data set has observations on two different kinds of individuals or combined groups if the centers of the two separate histograms are far enough to the variability in both the data sets.

Symmetric Histogram



Symmetric histogram

A symmetric histogram is also called a bell-shaped histogram. When you draw the vertical line down the center of the histogram, and the two sides are identical in size and shape, the histogram is said to be symmetric. The diagram is perfectly symmetric if the right half portion of the image is similar to the left half. The histograms that are not symmetric are known as skewed.

Probability Histogram

A Probability Histogram shows a pictorial representation of a discrete probability distribution. It consists of a rectangle centered on every value of x, and the area of each rectangle is proportional to the probability of the corresponding value. The probability histogram diagram is begun by selecting the classes. The probabilities of each outcome are the heights of the bars of the histogram.

Applications of Histogram

The applications of histograms can be seen when we learn about different distributions.

Normal Distribution

The usual pattern that is in the shape of a bell curve is termed normal distribution. In a normal distribution, the data points are most likely to appear on a side of the average as on the other. It is to be noted that other distributions appear the same as the normal distribution. The calculations in statistics are utilised to prove a distribution that is normal. It is required to make a note that the term "normal" explains the specific distribution for a process. For instance, in various processes, they possess a limit that is natural on a side and will create distributions that are skewed. This is normal which means for the processes, in the case where the distribution isn't considered normal.

Skewed Distribution

The distribution that is skewed is asymmetrical as a limit which is natural resists end results on one side. The peak of the distribution is the off-center in the direction of the limit and a tail that extends far from it. For instance, a distribution consisting of analyses of a product that is unadulterated would be skewed as the product cannot cross more than 100 per cent purity. Other instances of natural limits are holes that cannot be lesser than the diameter of the drill or the call-receiving times that cannot be lesser than zero. The above distributions are termed right-skewed or left-skewed based on the direction of the tail.

Multimodal Distribution

The alternate name for the multimodal distribution is the plateau distribution. Various processes with normal distribution are put together. Since there are many peaks adjacent together, the tip of the distribution is in the shape of a plateau.

Edge peak Distribution

This distribution resembles the normal distribution except that it possesses a bigger peak at one tail. Generally, it is due to the wrong construction of the histogram, with data combined together into a collection named "greater than".

Comb Distribution

In this distribution, there exist bars that are tall and short alternatively. It mostly results from the data that is rounded off and/or an incorrectly drawn histogram. For instance, the temperature that is rounded off to the nearest 0.2o would display a shape that is in the form of a comb provided the width of the bar for the histogram were 0.1o.

Truncated or Heart-Cut Distribution

The above distribution resembles a normal distribution with the tails being cut off. The producer might be manufacturing a normal distribution of product and then depending on the inspection to segregate what lies within the limits of specification and what is out. The resulting parcel to the end-user from within the specifications is heart cut.

Dog Food Distribution

This distribution is missing something. It results close by the average. If an end-user gets this distribution, someone else is receiving a heart cut distribution and the end-user who is left gets dog food, the odds and ends which are left behind after the meal of the master. Even if the end-user receives within the limits of specifications, the item is categorised into 2 clusters namely – one close to the upper specification and another close to the lesser specification limit. This difference causes problems in the end-users process.

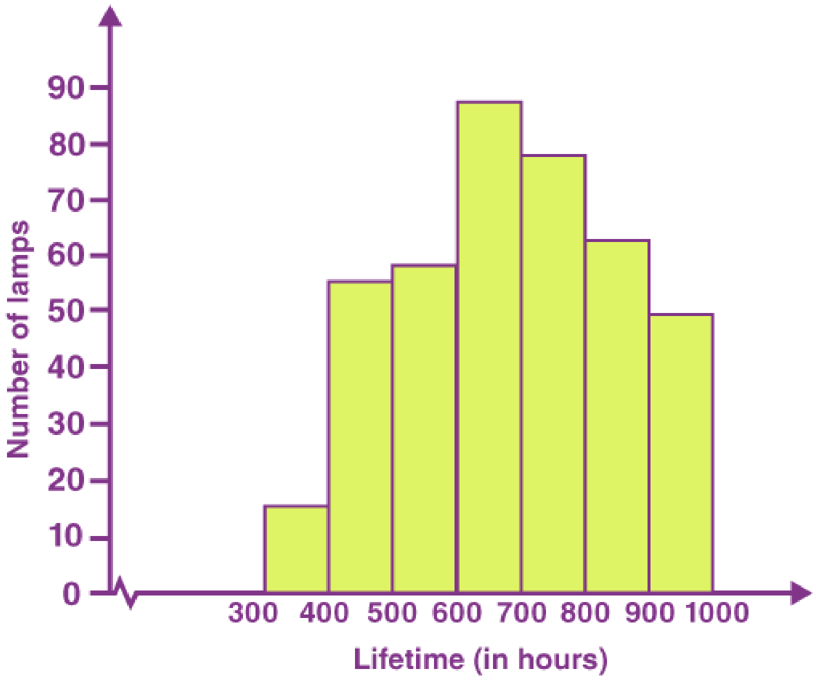
Histogram Solved Example

Question: The following table gives the lifetime of 400 neon lamps. Draw the histogram for the below data.

Lifetime (in hours)	Number of lamps
300 – 400	14
400 – 500	56
500 – 600	60
600 – 700	86
700 – 800	74
800 – 900	62
900 – 1000	48

Solution:

The histogram for the given data is:


90
80
70
60
50
40
30
20
10
0
Number of lamps
300
400
500
600
700
800
900
1000
Lifetime (in hours)