Auditing Course Material

4. Mean for Grouped Data

There are two scenarios in which we calculate the mean for grouped data: when the data is discrete and when it is continuous.

Let us explore the approaches for computing the arithmetic mean in both situations.

---

4. Mean for Grouped Data

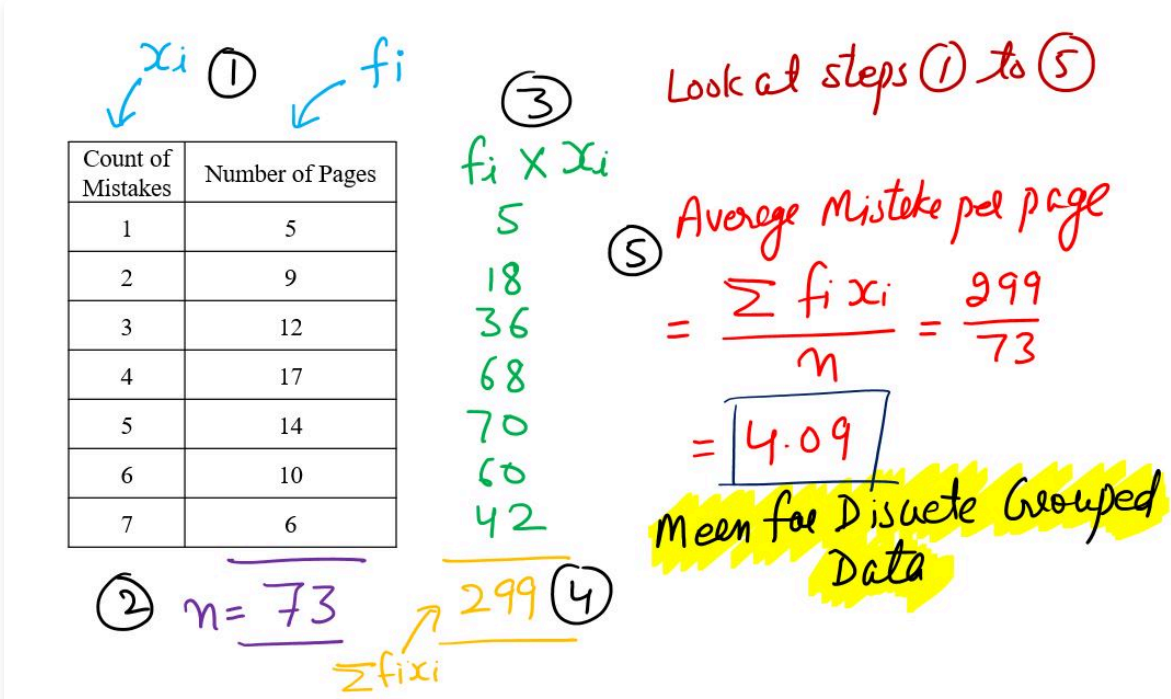
The mean for Mean for Discrete Grouped Data is computed by the following formula:

$$\bar{x} = \frac{f_1 \cdot x_1 + f_2 \cdot x_2 \ldots . f_n \cdot x_n}{f_1 + f_2 - - - - f_n} = \frac{\sum f_i . x_i}{\sum f_i}$$

where f_i is the frequency and n are total number of data points.

Illustration

Let us say we are examining a writer's 73-page body of work, aiming to determine the average number of mistakes per page.



Weighted Average Mean

The Weighted Average Mean is measure of Central Tendency of a set of quantitative observations, when not all the observations have same importance. In other words, it is important to assign weights to various data values, according to their importance. In that case, the mean is called the Weighted Mean.

If Mohan sold 5 blue balls for price of Rs 0.50 each, 15 red balls for price of Rs 0.75 each, 15 grey balls for price of Rs 0.90 each and 15 black balls for price of Rs 1.10 each, the average selling price of a ball is given by weighted mean, as calculated below:

$$\text{WEIGHTED MEAN}$$

$$\bar{x} = \frac{w_1 x_1 + w_2 x_2 + w_3 x_3 + \ldots w_n x_n}{w_1 + w_2 + w_3 - - - w_n}$$

= \(\frac{5 \times0.50+15 \times0.75+15 \times0.90+15 \times1.10}{5+15+15+15} \)

= 0.875

4. Mean for Grouped Data

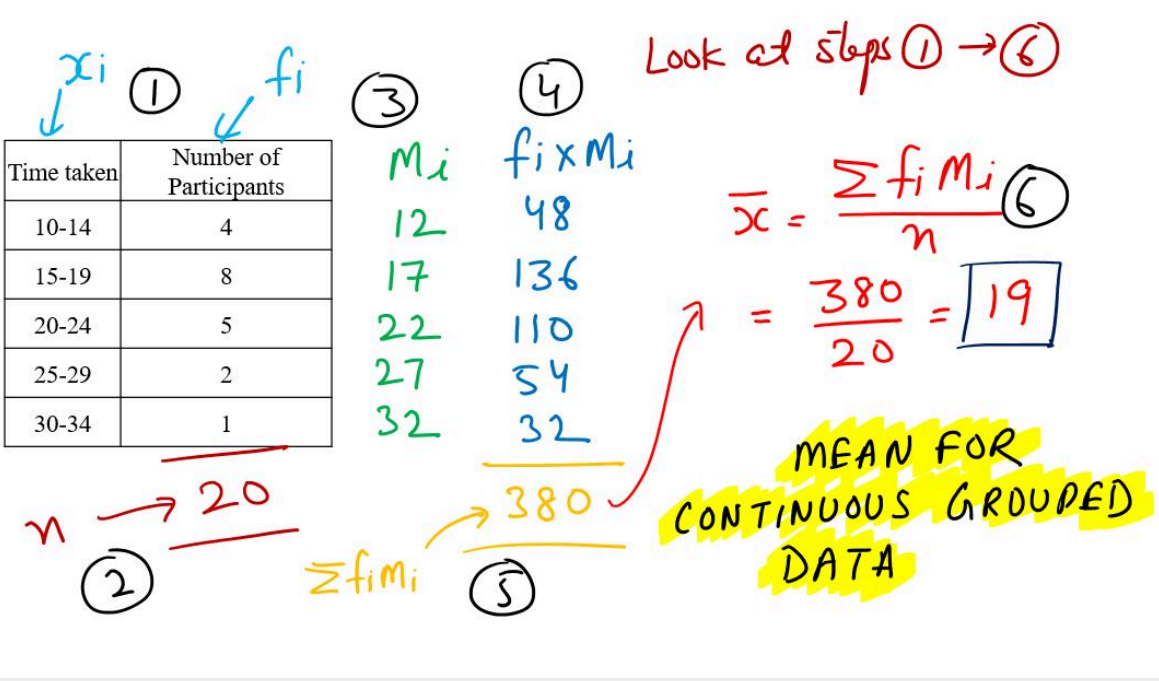
The mean for Mean for Continuous Grouped Data is computed by the following formula:

$$\bar{x} = \frac{f_1 \cdot M_1 + f_2 \cdot M_2 + \cdots f_n \cdot M_n}{f_1 + f_2 + \cdots f_n} = \frac{\sum f_i \cdot M_i}{\sum f_i}$$

where f_i is the frequency, M_i is the mid point of each class and n are total number of data points.

Illustration

An experiment records the time taken by 20 participants to complete a task, in days. Determine the average time taken to complete the task.



5. Properties of Arithmetic Mean

Here are some properties of the mean:

(i) **Balancing Property:** The sum of deviations from the mean is always zero: Σ(X - \(\bar{X} \)) = 0. This means that the positive deviations balance out the negative deviations.

(ii) **Sensitivity to Outliers**: The mean is highly affected by extreme values or outliers, as they can significantly influence its value. Outliers can skew the mean, making it less representative of central tendency.

(iii) **Affected by Changes in Values**: If each value in the dataset is increased, decreased, or multiplied by a constant, the mean also changes accordingly, increasing, decreasing, or multiplying by the same constant.

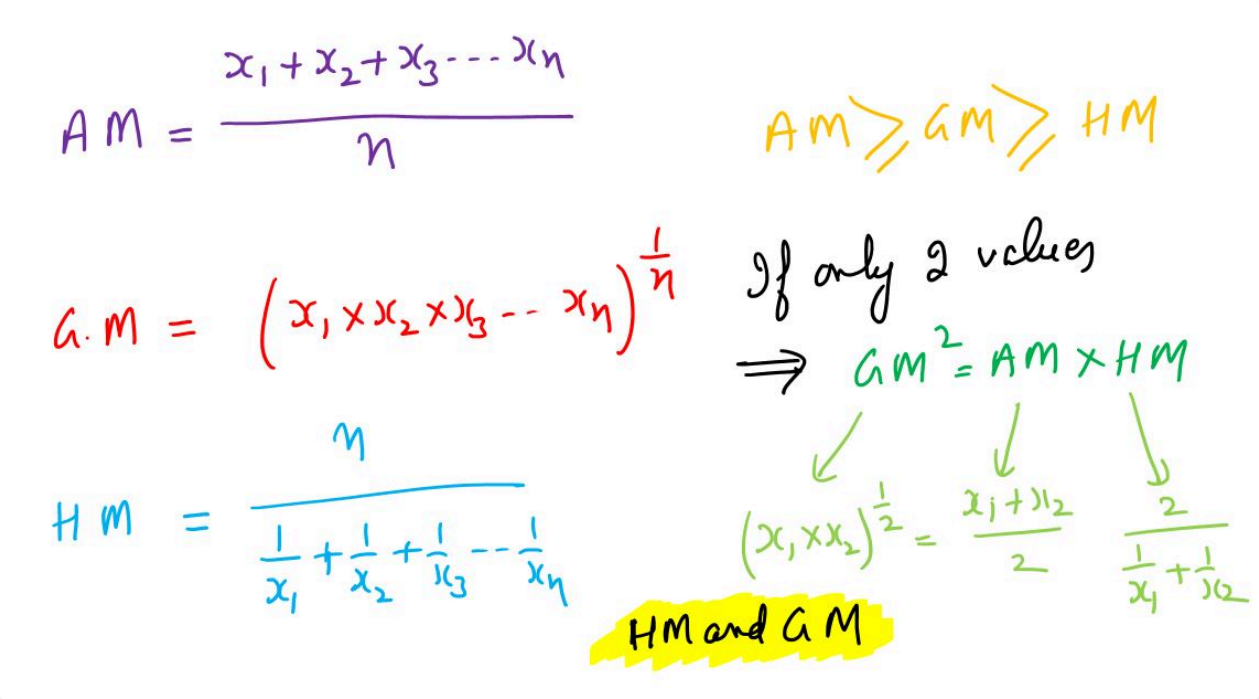
(iv) **Minimizing Deviations**: The sum of the squares of deviations of the values from the mean is the minimum possible value compared to deviations from any other value. This property defines the mean as the measure that minimizes the total squared deviations.

(v) **Affected by Skewed Distributions**: In skewed distributions, particularly where the data isn't symmetrically distributed, the mean might not accurately represent the central tendency, as it's drawn towards the tail by extreme values.

(vi) **Applicable to Interval/Ratio Data**: The mean is suitable for interval and ratio data where the numerical relationship between values exists, allowing for arithmetic operations.

---

6. Harmonic Mean and Geometric Mean



Harmonic Mean

The harmonic mean is a way to find the average of a set of numbers, but it differs from the more commonly known arithmetic mean. While the arithmetic mean sums up all values and divides by the count, the harmonic mean takes the reciprocal of each number, finds the average of those reciprocals, and then takes the reciprocal of that average. Thus if X_1, X_2, X_3,..., X_n (none of them being zero) is a series and H is its harmonic mean then

H = $\frac{n}{ \frac{1}{X_1} + \frac{1}{X_2} + \frac{1}{X_3}+..... + \frac{1}{X_n} }$

This method inherently emphasizes smaller values because when you take the reciprocal of a smaller number, the resulting value is larger. Consequently, smaller numbers have a relatively greater influence on the harmonic mean compared to larger numbers.

This characteristic makes the harmonic mean particularly useful in scenarios where rates or ratios are involved, such as average speed calculations in physics or finance, where the emphasis on smaller values is desired.

Geometric Mean

The geometric mean (GM) is another type of average used to find the central tendency of a set of numbers. Unlike the arithmetic mean that sums up values and divides by the count, the geometric mean is calculated by multiplying all the numbers together and then taking the
nth root, where n is the count of numbers.

The Geometric Mean, G of a series of values X_1, X_2, X_3,..., X_n (none of them being zero) is defined as the nth root product of n numbers.

$G = \sqrt[n]{X_1 \times X_2 \times X_3 \times..... \times X_n }$

The geometric mean is sensitive to extreme values, but less so compared to the arithmetic mean. Extreme values have less influence on the geometric mean due to the multiplicative nature of the calculation. However, if any value in the set is zero or negative, the geometric mean cannot be calculated because you can't take the root of zero or a negative number.

This makes the geometric mean suitable for situations involving rates of change, growth rates, or scenarios where values are proportional, like investment returns over multiple periods or analyzing exponential growth rates in various fields.

---

6. Harmonic Mean and Geometric Mean

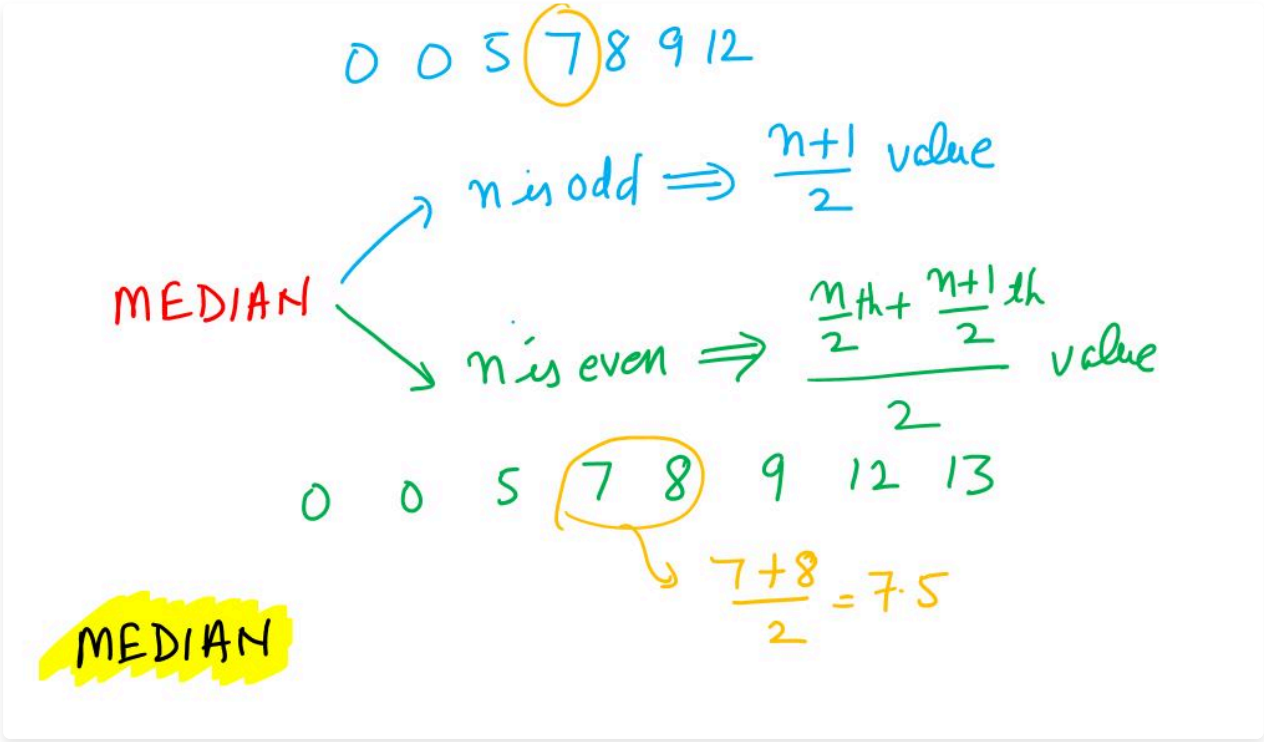
The relationship among the Arithmetic Mean (AM), Geometric Mean (GM), and Harmonic Mean (HM) is known as the inequality of means. It's a fundamental relationship that provides insights into how these different types of averages are related to each other.

If there are only two observations, the product of Arithmetic Mean (AM) and Harmonic Mean (HM) is equal to square of Geometric Mean (GM).

---

7. Median

The Median is that positional value of the variable which divides the distribution into two equal parts, one part comprises all values greater than or equal to the median value and the other comprises all values less than or equal to it. The Median is the "middle" element when the data set is arranged in order of the magnitude.



Since the median is determined by the position of different values, it remains unaffected if, say, the size of the largest value increases (unlike mean).

The median can be easily computed by sorting the data from smallest to largest and finding out the middle value. Suppose we have the following observation in a data set: 5, 7, 6, 1, 8, 10, 12, 4, and 3.

Arranging the data, in ascending order you have:
1, 3, 4, 5, 6, 7, 8, 10, 12.

The "middle score" is 6, so the median is 6. Half of the scores are larger than 6 and half of the scores are smaller.

However, if there are even numbers in the data, there will be two observations which fall in the middle. The median in this case is computed as the arithmetic mean of the two middle values.

If arranged data is 1, 3, 4, 5, 6, 7, 7, 8, 10, 12, then the Median will be
Median = $\frac{6+7}{2}$ = 6.5

Thus we conclude that:
If there are Odd number of observations, then value of $\frac{n+1}{2}$ th term will given value of Median.
If there are Even number of observations, then average of $\frac{n}{2}$ th and $\frac{n+1}{2}$ th terms will given value of Median.

---

7. Median

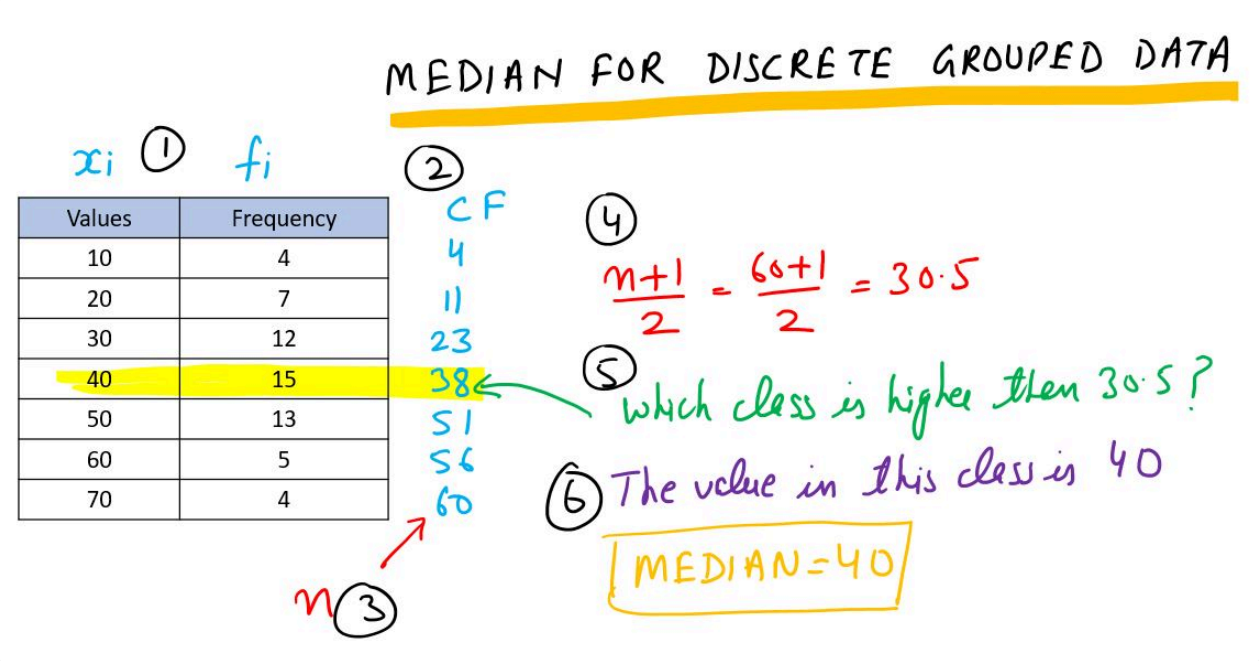
When dealing with discrete grouped data, finding the median involves locating the middle value that divides the data into two equal halves.

To find the median for grouped data, the Cumulative Frequency distribution is used.

First we find the value of $\frac{n+1}{2}$. Then we identify that class, which is higher than Cumulative Frequency of $\frac{n+1}{2}$.

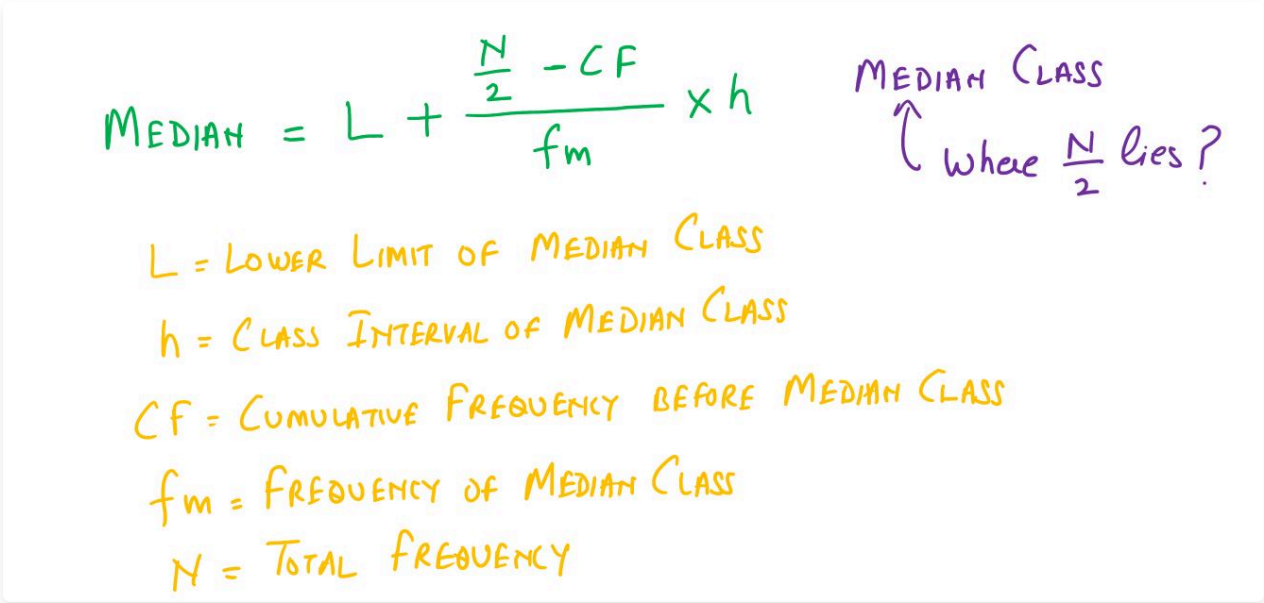
The value of this class will be Median.

Illustration



7. Median

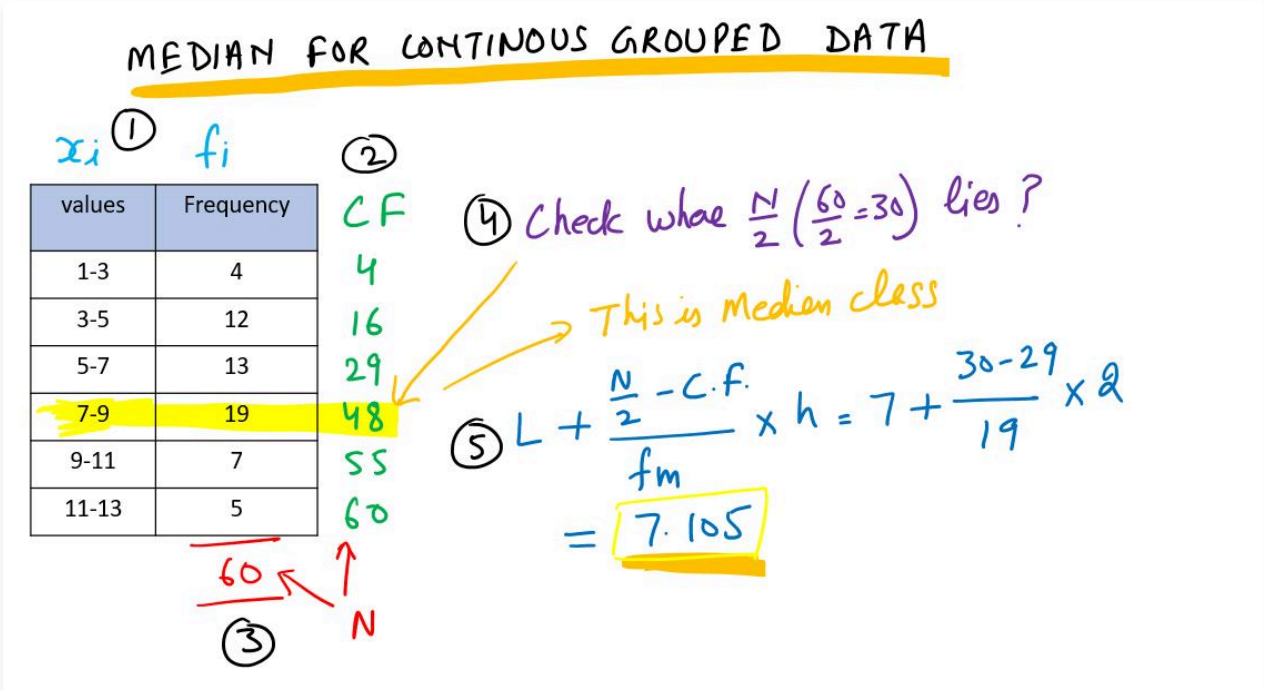
Median for continuous grouped data is given by following formula:



To find median class, we look at the class where $\frac{n}{2}$ lies.

Note: If the class intervals are 'inclusive' type, then we need to first convert it into 'exclusive' type, for calculating median. If classes are 10-19, 20-29, 30-39, we can convert them to 9.5-19.5, 19.5-29.5, 29.5-39.5 and so on.

Illustration



7. Median

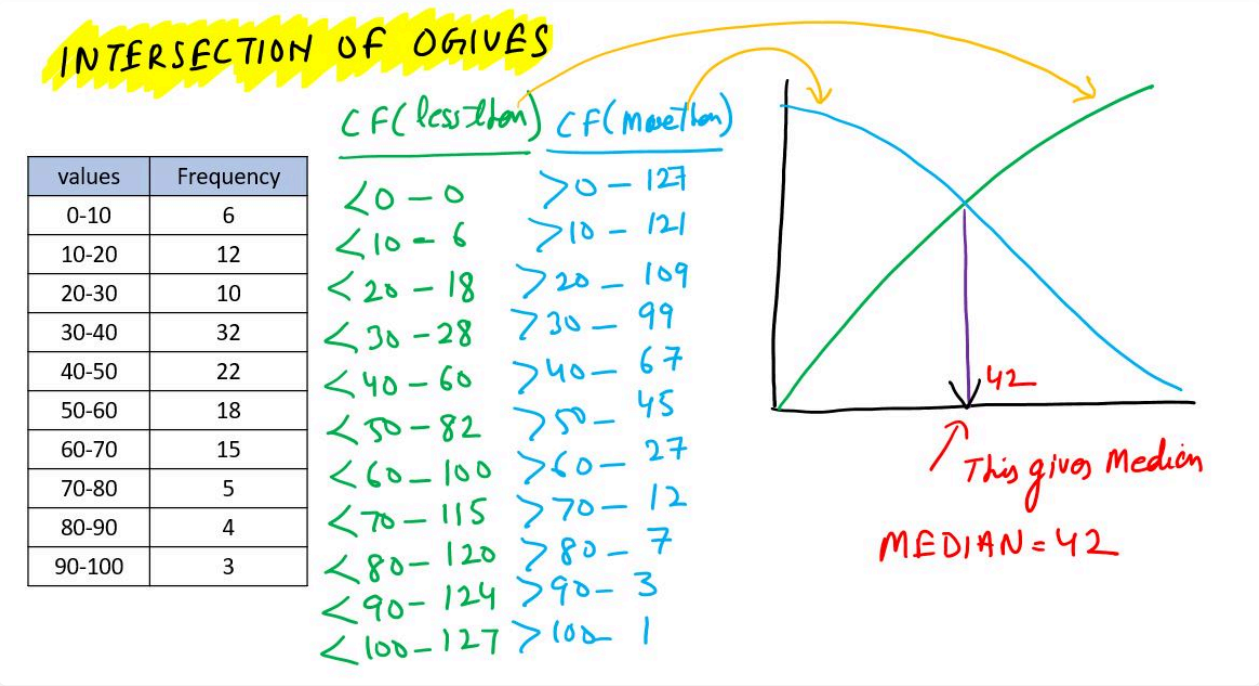
The graphical methods can determine the median using cumulative frequency graphs or ogives. The ogive graphs display cumulative frequencies against the upper or lower class boundaries.

Method 1: Intersection of Less than Ogive and More than Ogive

Plot the less than ogive (cumulative frequency curve starting from zero) and the more than ogive (cumulative frequency curve ending at the total frequency) on the same axes.

The point where these two curves intersect corresponds to the median. The value at this intersection on the horizontal axis is the median.

Illustration



values	Frequency
0-10	6
10-20	12
20-30	10
30-40	32
40-50	22
50-60	18
60-70	15
70-80	5
80-90	4
90-100	3

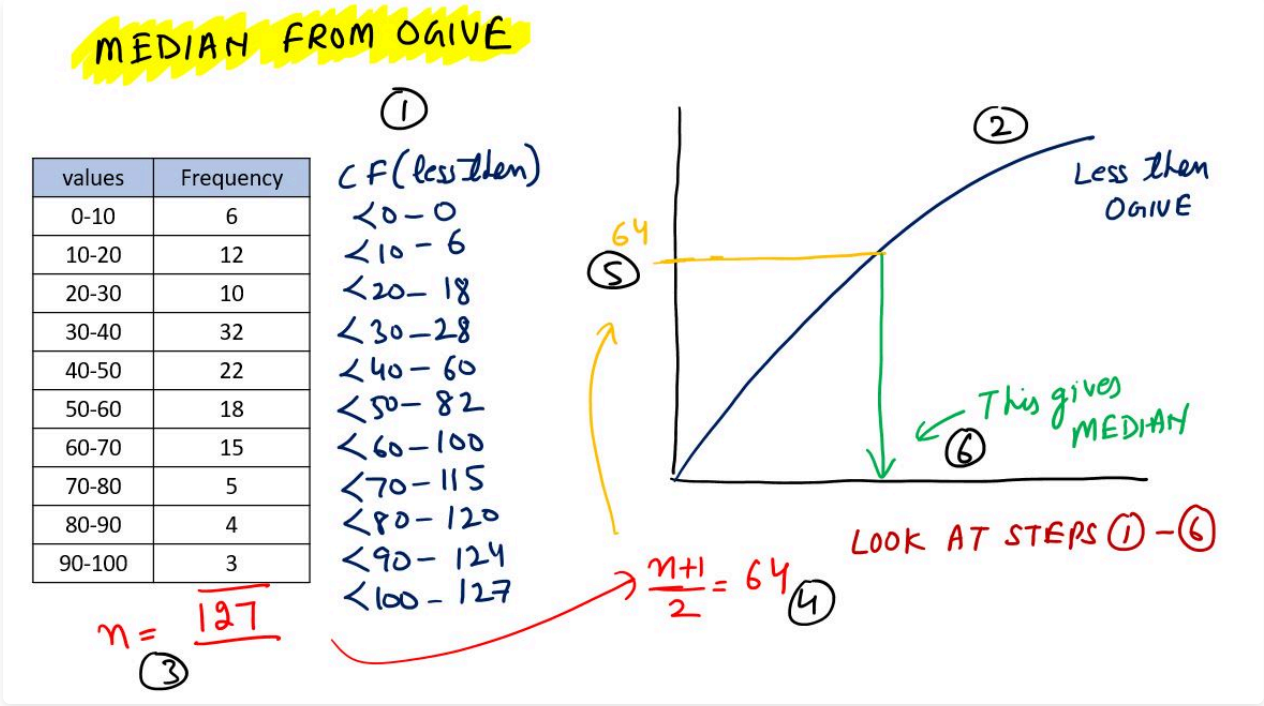
Method 2: With only one Ogive

Draw a horizontal line at the level of y = $\frac{n}{2}$ or y = $\frac{n+1}{2}$ on the y-axis, where n is the total frequency.

Where this line intersects the ogive curves, draw a vertical line downward to the x-axis.

The point of intersection on the x-axis gives the value of the median.

Illustration



---

8. Properties of the Median

Here are some properties of the median:

(i) **Robustness to Extreme Values**: The median is less affected by extreme values or outliers compared to the mean. It provides a better representation of centrality in datasets with outliers.

(ii) **Insensitive to Exact Values**: It is not influenced by the exact values of all observations, only by the order or ranking of values. This makes it suitable for ordinal data.

(iii) **Balancing Effect on Skewed Distributions**: The median is appropriate for skewed distributions as it better represents centrality in such cases. It divides the dataset into two equal halves, regardless of skewness.

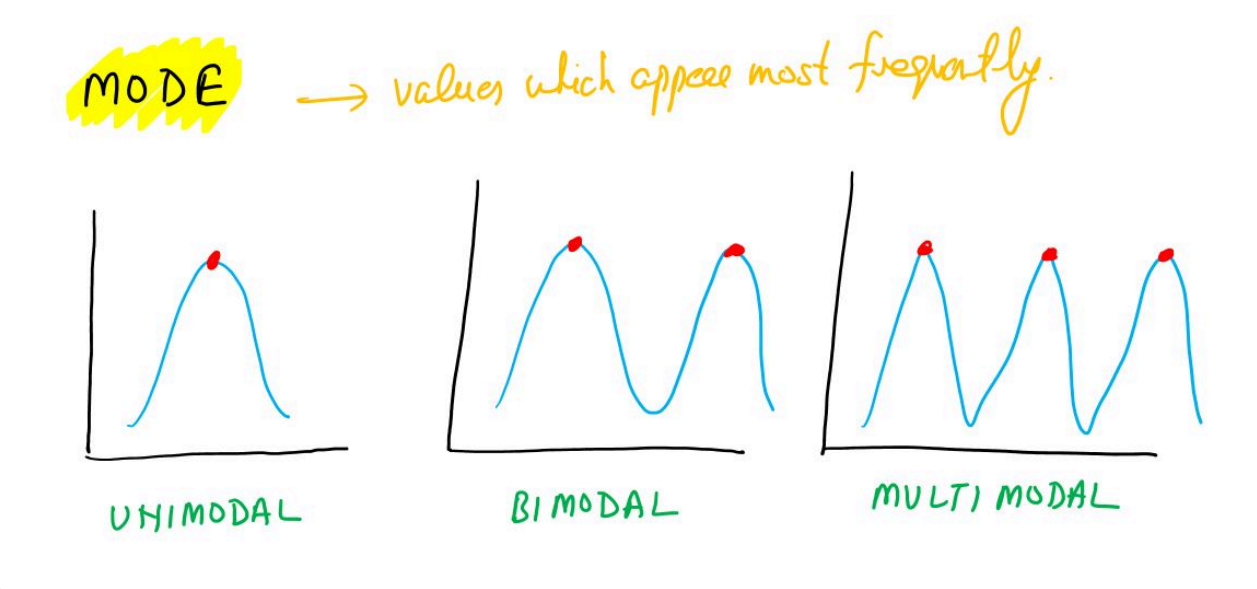
---

9. Mode

Mode is the most frequently observed data value. It is denoted by Mo. It has been derived from the French word "la Mode" which signifies the most fashionable values of a distribution, because it is repeated the highest number of times in the series.

Consider the data set 1, 2, 3, 4, 4, 5. The mode for this data is 4 because 4 occurs most frequently (twice) in the data.

In this example, as there is a unique value of mode, the data is **unimodal**. But, the mode is not necessarily unique, unlike arithmetic mean and median. You can have data with two modes (bi-modal) or more than two modes (**multi-modal**). Example of **bi-modal** data is 1, 2, 2, 3, 4, 4, 5 (because both 2 and 4 are appearing twice each).



It may be possible that there may be no mode if no value appears more frequent than any other value in the distribution. For example, in a series 1, 1, 2, 2, 3, 3, 4, 4, there is no mode.

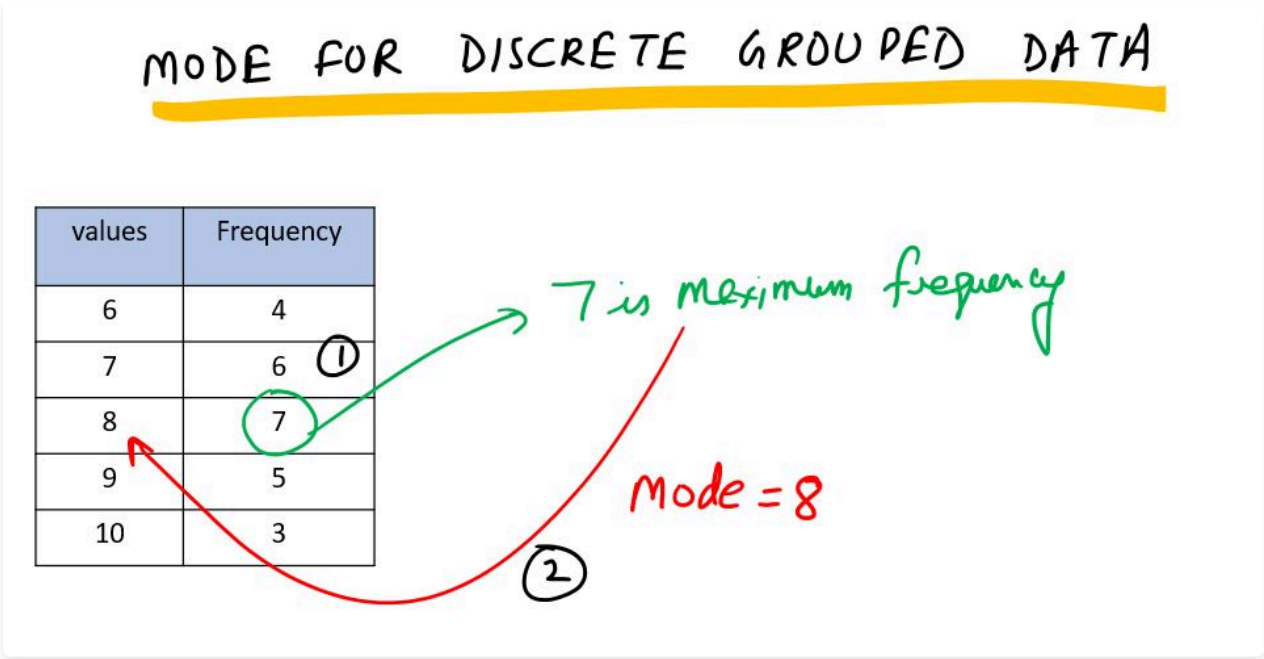
In the continuous data below, no values repeat, which means there is no mode. With continuous data, it is unlikely that two or more values will be exactly equal because there are an infinite number of values between any two values. However, you can find the mode for continuous data by locating the maximum value on a probability distribution plot. If you can identify a probability distribution that fits your data, find the peak value and use it as the mode.

---

9. Mode

The mode of discrete grouped data is the class with the highest frequency, representing the most frequently occurring category within the dataset.

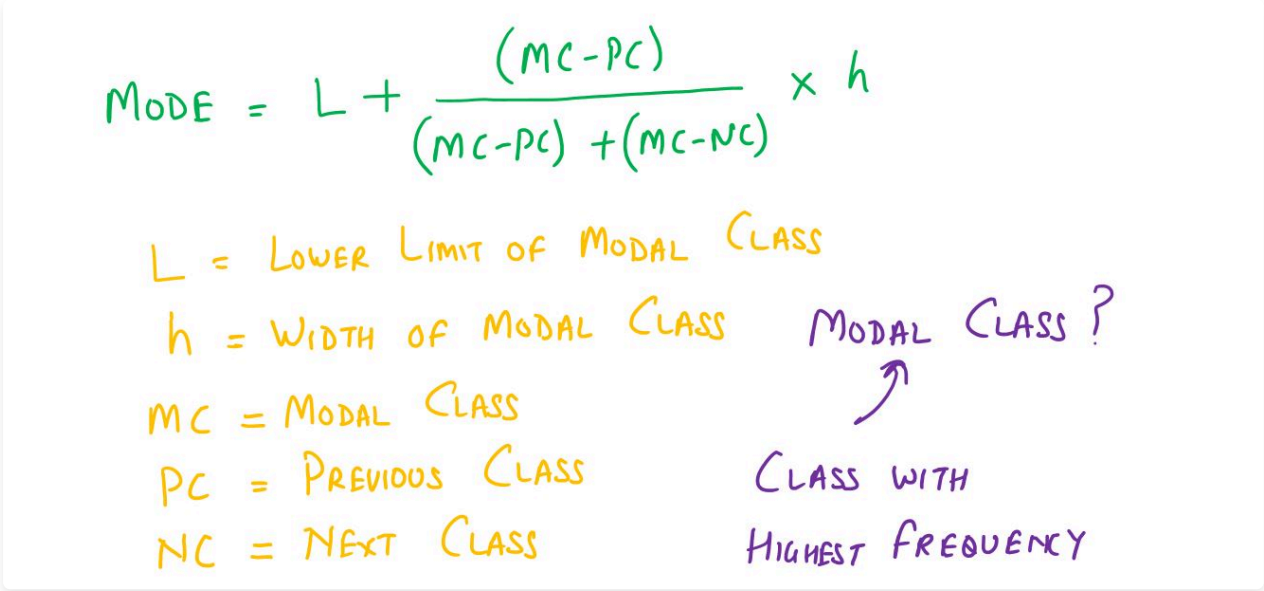
Illustration



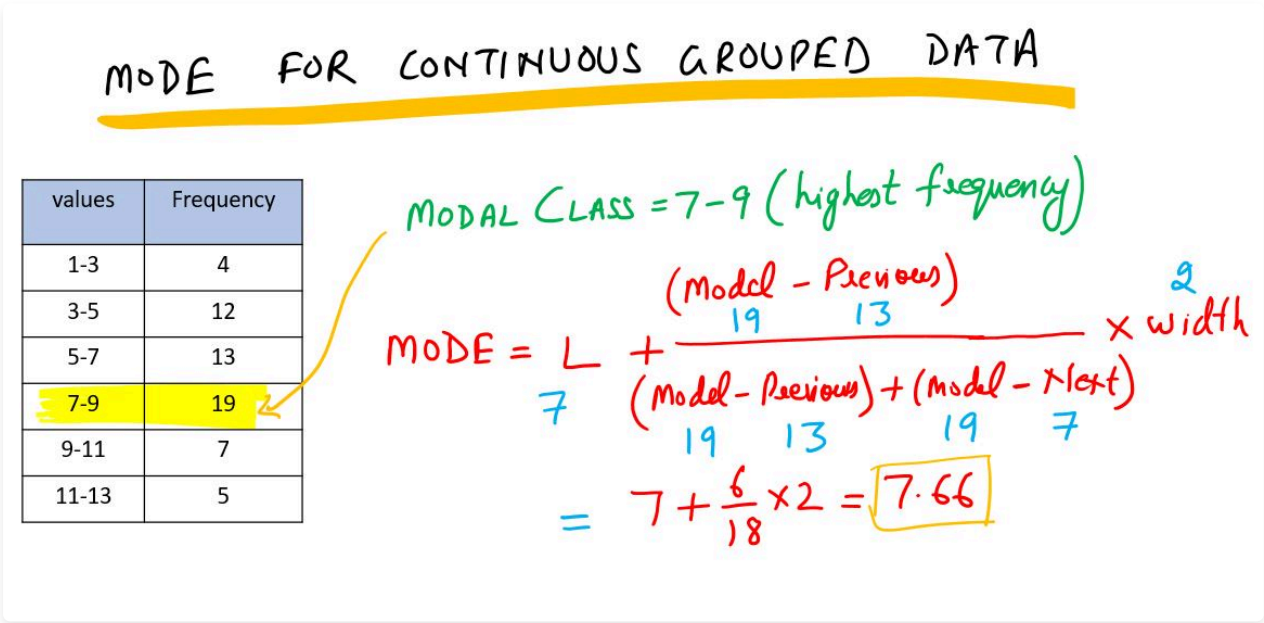
9. Mode

To find Mode for Continuous Grouped Data, first we find the Modal class. The modal class is that class, which has highest frequency.

Then the mode is calculated with following formula.



Illustration



values	Frequency
1-3	4
3-5	12
5-7	13
7-9	19
9-11	7
11-13	5

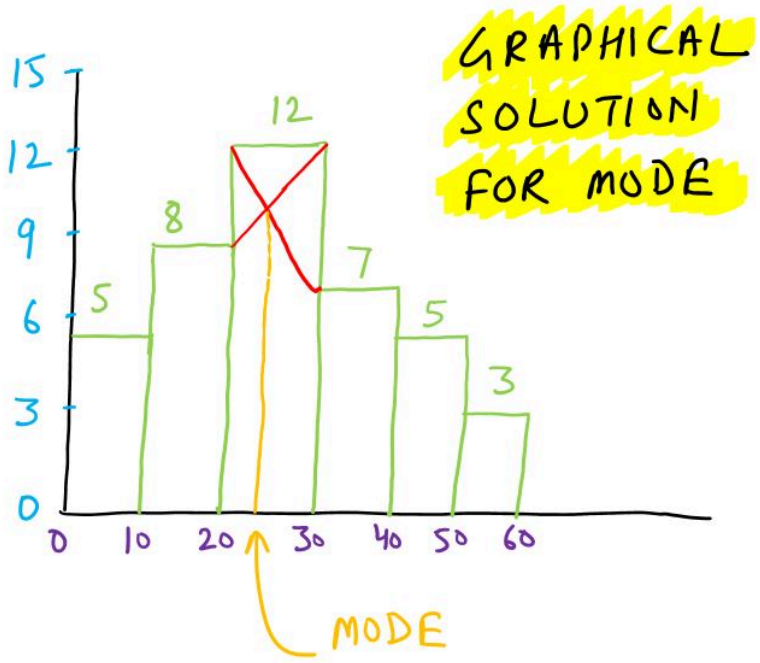
9. Mode

To find the mode, graphically, we use Histogram.

We identify the tallest bar or the highest peak in the histogram. The class corresponding to this tallest bar represents the mode for the dataset.

Illustration

Values	Frequency
0-10	5
10-20	8
20-30	12
30-40	7
40-50	5
50-60	3



---

10. Properties of Mode

Here are some properties of the mode:

(i) **Highest Frequency**: The mode is the value that appears most frequently in a dataset. It's the value with the highest frequency or occurrence.

(ii) **May Not be Unique**: A dataset can have one mode (unimodal), multiple modes (bimodal, multimodal), or no mode (no clear value repeated more than others).

(iii) **Applicability to Nominal Data**: It is suitable for nominal data where values are categorical and lack a numerical relationship.

(iv) **Resilience to Outliers**: Outliers or extreme values have minimal impact on the mode since it's solely based on frequency of occurrence.

(v) **Mode as Representative Value**: It may or may not represent the center or typical value, especially in datasets with skewed distributions or unusual frequency distributions.

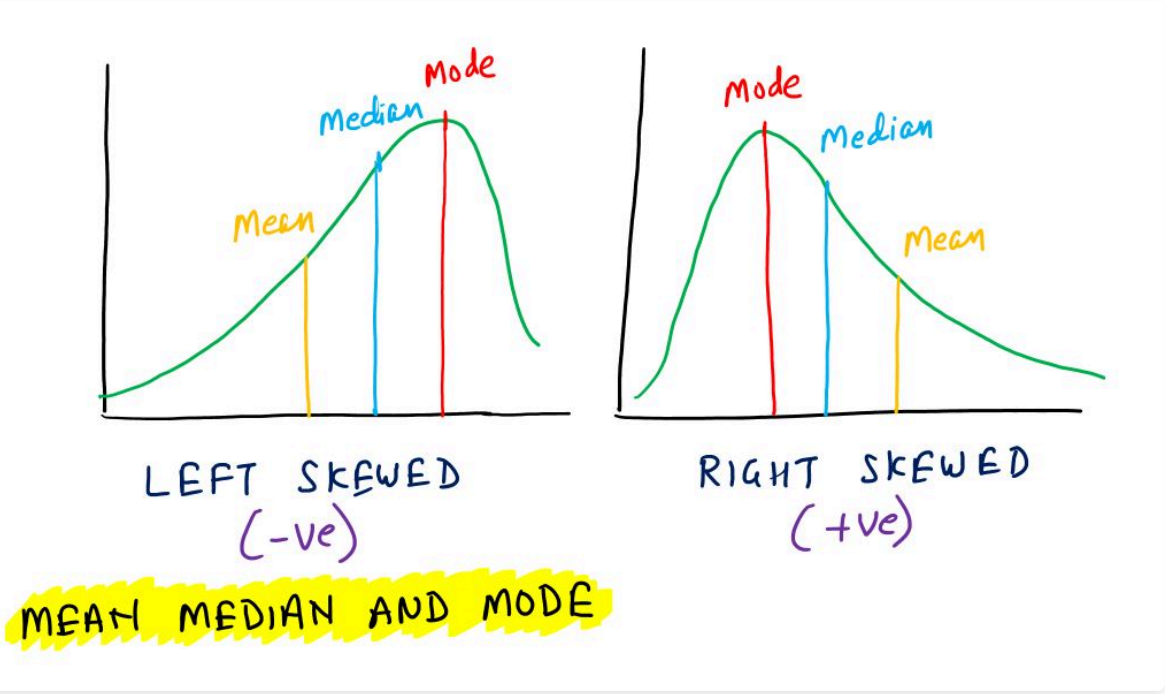
---

11. Comparing Mean, Median and Mode

For a **perfectly symmetrical distribution**, Mean, Median and Mode are always equal.

Let us understand the relationships, if the distribution is not symmetrical.

The relative magnitude of the three will be either Mean > Median > Mode or Mean < Median < Mode. To summarize, the median is always between the arithmetic mean and the mode.



For moderately asymmetrical distribution (or for asymmetrical curve), the relation **Mean – Mode = 3 (Mean - Median)**, approximately holds.

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

When you have a symmetrical distribution for continuous data, the mean, median, and mode are equal. In this case, analysts tend to use the mean because it includes all of the data in the calculations. However, if you have a skewed distribution, the median is often the best measure of central tendency.

When you have ordinal data, the median or mode is usually the best choice. For categorical data, you have to use the mode.

---

12. Partition Values

If the values of the dataset are arranged in ascending or descending order of magnitude, then we have seen above that **Median** is that value of the dataset which divides the total frequencies in two equal parts.

Similarly the given dataset can be divided into 4, 10 and 100 equal parts. The values of the dataest dividing into 4 equal parts are called **Quartile**, into 10 equal parts are called **Decile** and into 100 equal parts are called **Percentile**.

Let us understand them one by one.

---

13. Percentiles

Percentiles divide the distribution into 100 equal parts, so you can get 99 dividing positions denoted by P_1, P_2, P_3 , ..., P_{99}. The P_{50} is the median value.

To understand this, if you have scored 82 percentile in CAT exam, it means that your position is below 18 per cent of total candidates appeared in the exam. It also means that 18% candidates score marks, greater than your score.

---

13. Percentiles

To find the percentile, first we arrange the data in ascending order.

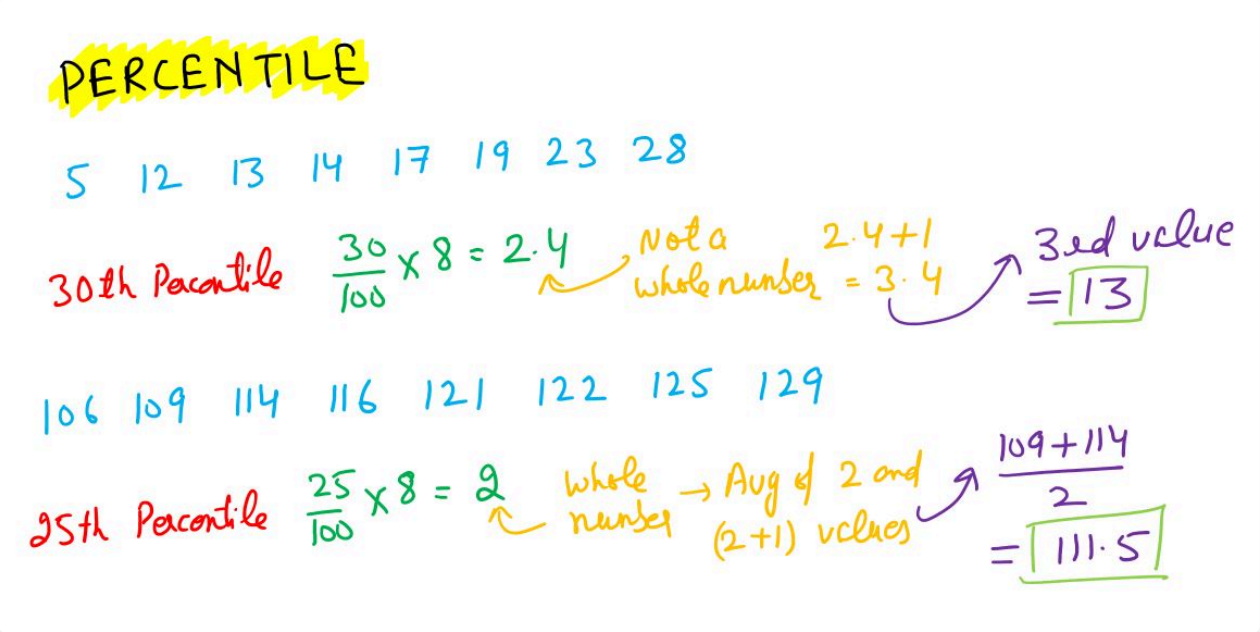
The value of p^{th} percentile is given by two steps:

1. First find the location of p^{th} percentile, i = \(\frac{p}{100} \times N \)

2. If i comes out to be whole number, then the required percentile is given by average of i^{th} and $(i+1)^{th}$ value.

If i comes out to be not a whole number, then the required percentile is given by the location of whole number part of (i+1) value.

Illustration

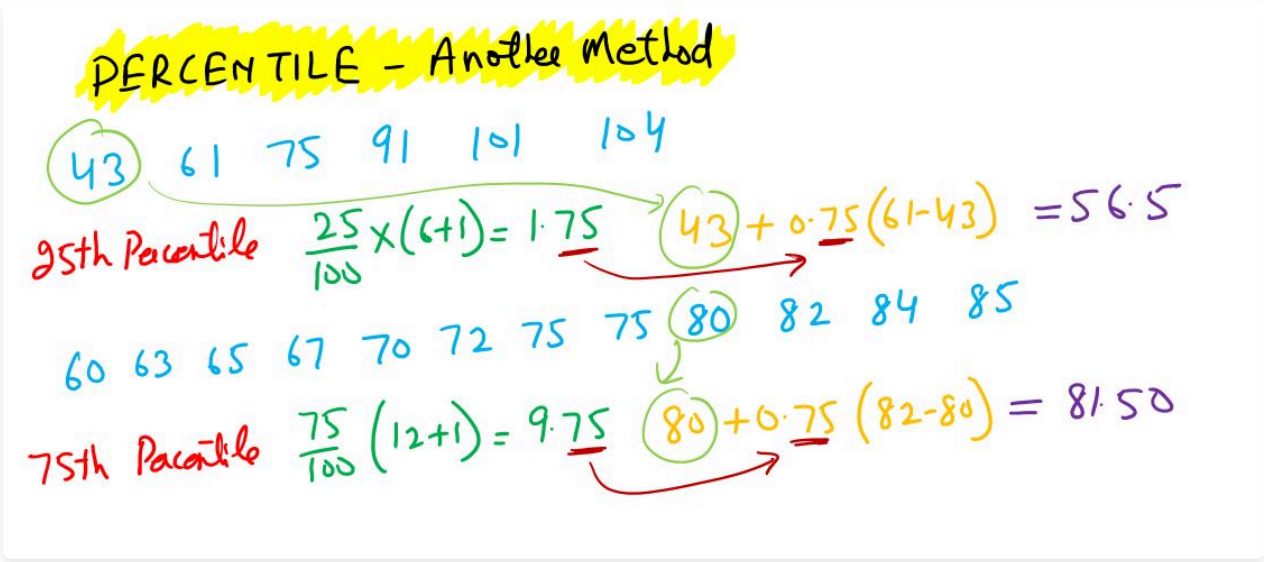


13. Percentiles

In some academic literature, another method is given for finding percentile.

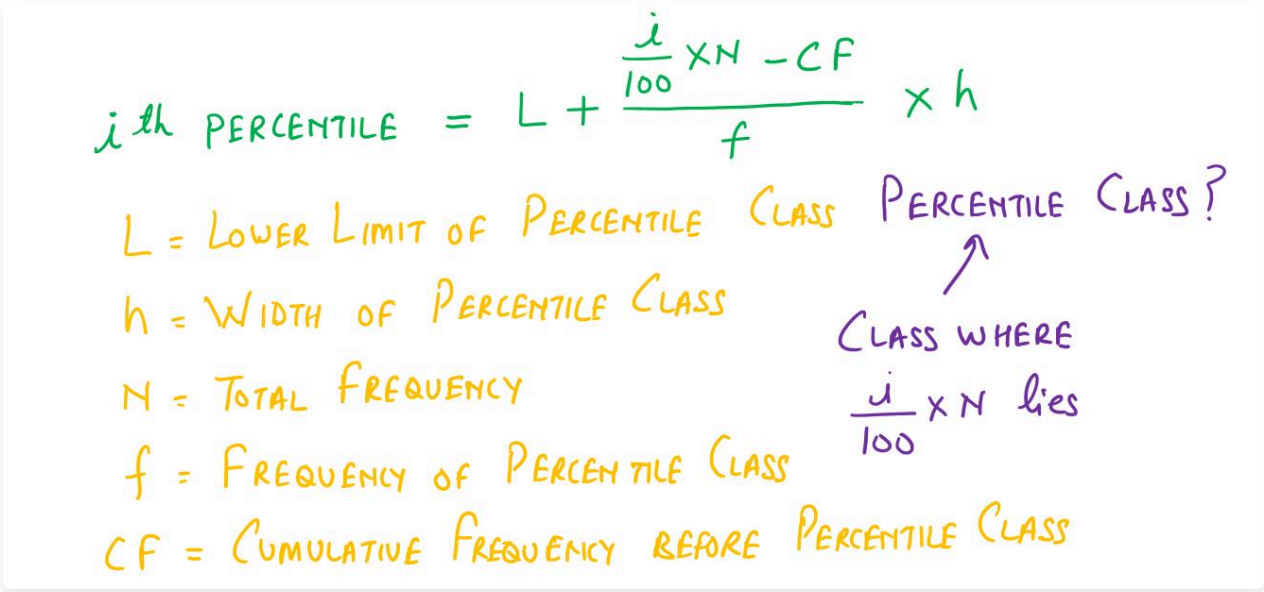
The value of p^{th} percentile is given by the location of i = $\frac{p}{100} \times (N+1)$

Illustration



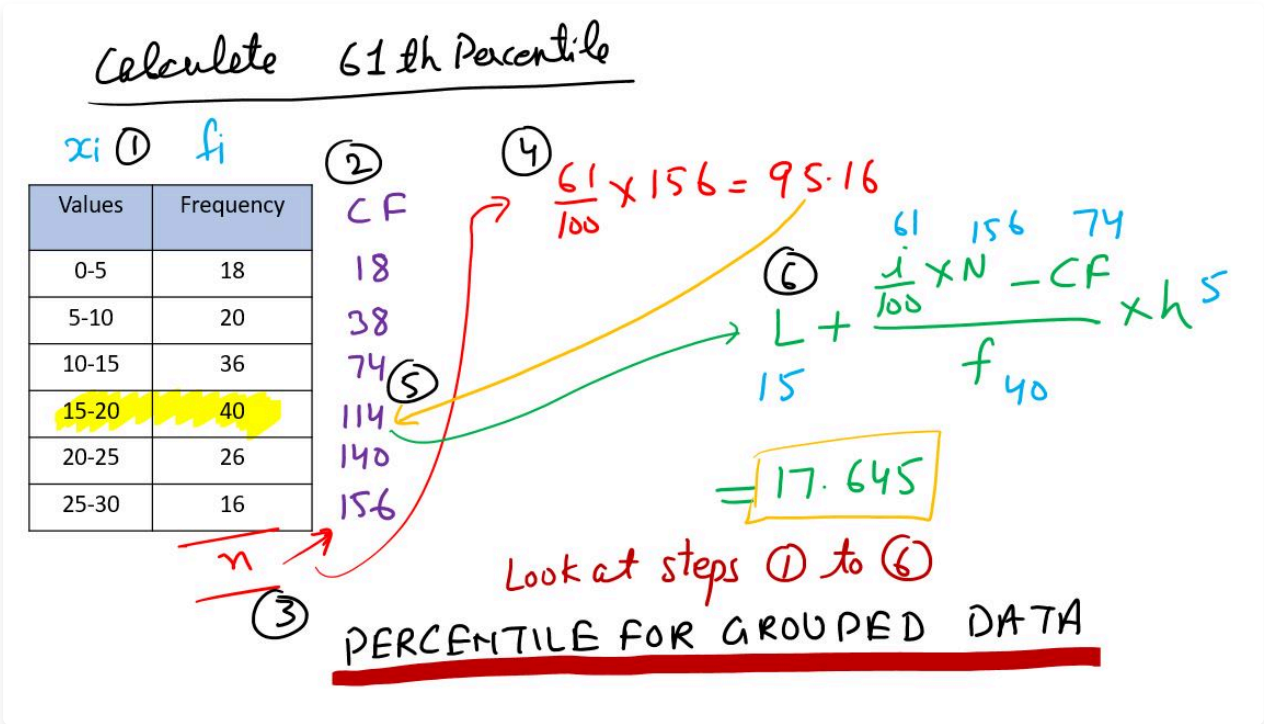
13. Percentiles

Percentile for continuous grouped data is given by following formula:



To find percentile class, we look at the class where \(\frac{p}{100} \) \(\times N \) lies.

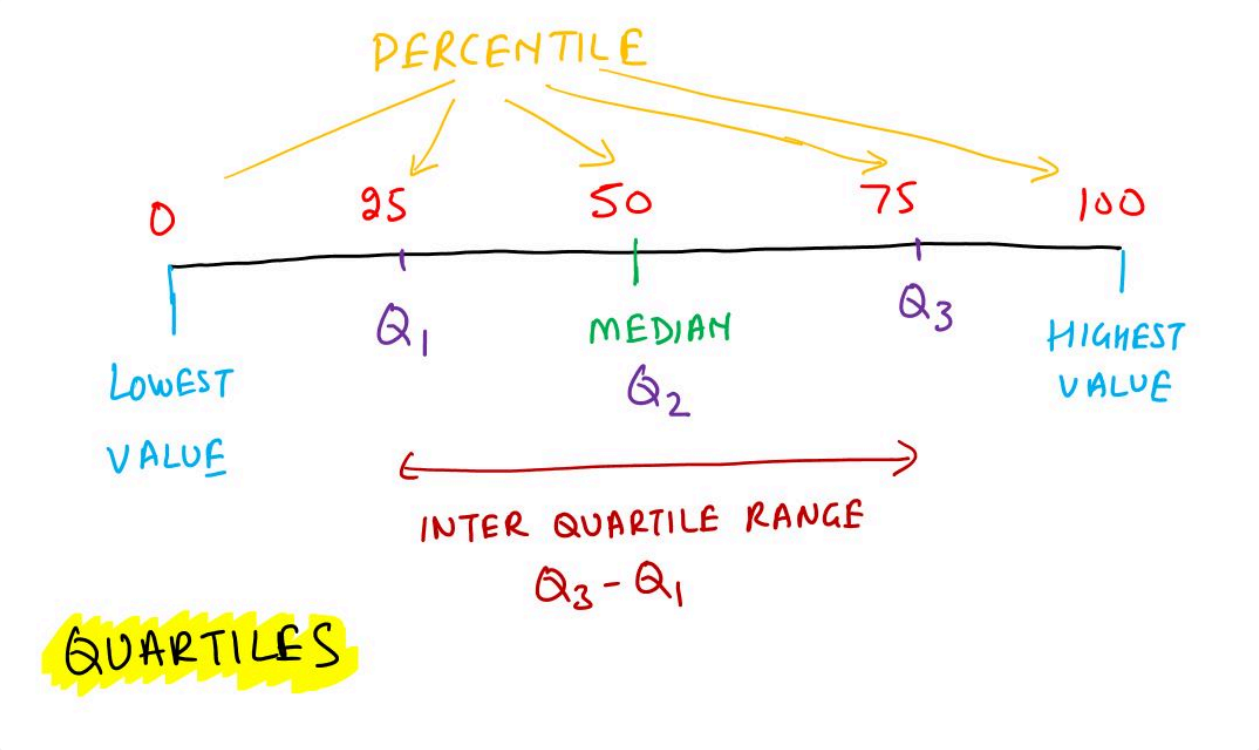
Illustration



Values	Frequency	CF
0-5	18	18
5-10	20	38
10-15	36	74
15-20	40	114
20-25	26	140
25-30	16	156
	n	

14. Quartiles

The values of the variate which divide the total frequency into **4 equal parts**, are called quartiles. There are three quartiles.



The first Quartile (denoted by Q_1) or **lower quartile** has 25% of the items of the distribution below it and 75% of the items are greater than it.

The second Quartile (denoted by Q_2) or **median** has 50% of items below it and 50% of the observations above it.

The third Quartile (denoted by Q_3) or **upper Quartile** has 75% of the items of the distribution below it and 25% of the items above it. Thus, Q_1 and Q_3 denote the two limits within which central 50% of the data lies.

---

15. Change of Origin and Scale

Change of origin and change of scale are concepts in statistics that illustrate how altering the location (origin) or the magnitude (scale) of a dataset affects various measures of central tendency or dispersion.

Change of Origin

Changing the origin refers to adding or subtracting a constant value to each data point in the dataset. For instance, if you have a dataset [3, 6, 9, 12], adding 5 to each value would change it to [8, 11, 14, 17].

Change of Scale

Changing the scale involves multiplying or dividing each data point by a constant value. For example, if you have the dataset [2, 4, 6, 8], multiplying each value by 3 would result in [6, 12, 18, 24].

Effect on Mean

The mean is impacted by both change of origin and change of scale. Adding or subtracting a constant affects the mean by the same constant value, and multiplying or dividing by a constant affects the mean proportionally.

For instance, if you add 5 to each value in a dataset, the mean also increases by 5. Similarly, if you multiply each value by 3, the mean gets multiplied by 3 as well.

Effect on Median, Mode, and Percentiles

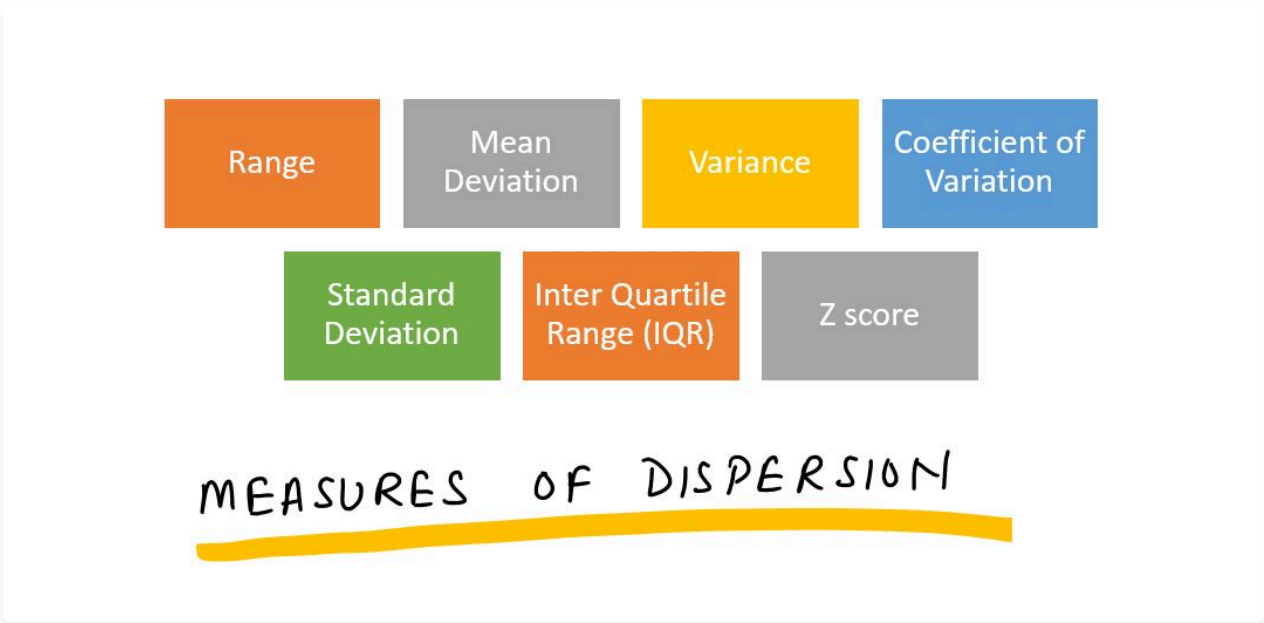
Like the mean, the median, mode, and percentiles are also affected by changes in both origin and scale.

Changing the origin shifts the median, mode, and percentiles by the same constant value. Changing the scale multiplies or divides the median, mode, and percentiles by the same constant.

---

1. Measures of Dispersion

Measures of dispersion quantify the extent to which individual data points in a dataset differ or spread out around a central tendency measure (like mean, median, or mode). They indicate the spread or variability of data points within a dataset.



These measures help to understand the distribution of data points around the central tendency. Small measures of dispersion indicate data points clustered close to the central value, while larger measures suggest greater variability or scattering of data points from the central value.

Various measures of dispersion or variation are available like Range, Mean Deviation, Variance, Coefficient of Variation, Standard Deviation, Inter Quartile Range (IQR) and Z score.

Let us discuss them one by one.

---

2. Range

The range, denoted as R, is a measure of dispersion in a dataset.

It is calculated by subtracting the smallest value S from the Largest value L in the dataset:

$$\text{RANGE} = L - S$$

$$L = \text{LARGEST VALUE}$$

$$S = \text{SMALLEST VALUE}$$

A higher value of the range indicates greater variability or dispersion in the dataset. For example, if you have a range of 50 in one dataset and 10 in another, the first dataset has more spread or variability among its values.

However, while the range provides insights into the spread of the data, it doesn't specify where the values are most concentrated within that spread. For instance, a dataset could have a wide range, but the bulk of its values might be clustered towards one end or the other.

To address this limitation and get a better understanding of the concentration of values within the range, the coefficient of range is used. The coefficient of range (CR) is calculated by dividing the range R by the sum of the extreme values (sum of the largest and smallest values, (L+S).

$$\text{COEFFICIENT OF RANGE} = \frac{L-S}{L+S}$$

The coefficient of range provides a relative measure that helps in understanding the extent of dispersion concerning the total span of the dataset (sum of the largest and smallest values). A higher coefficient of range suggests a greater dispersion relative to the total range of values in the dataset, giving an idea of how the spread relates to the overall scope of the dataset.

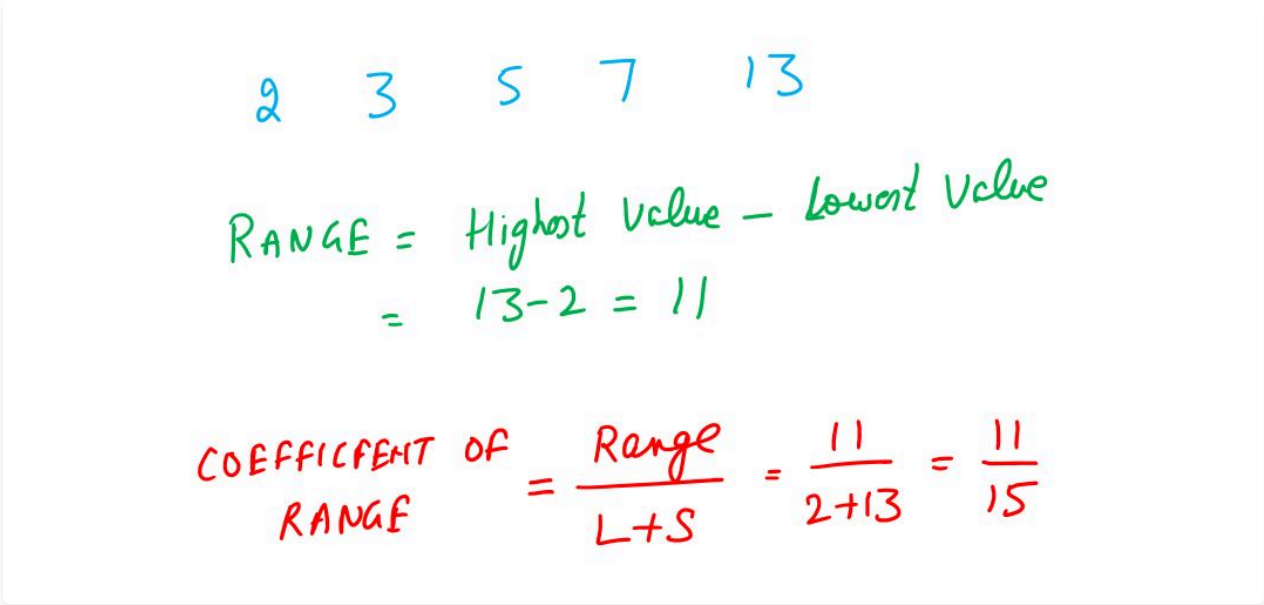
While Range is an Absolute mesaure of dispersion, Coefficient of Range provides a relative measure of dispersion.

---

2. Range

Calculate Range and Coefficient of Range for 2, 3, 5, 7 and 13.

Solution:



3. Quartile based mesaures

The presence of even one extremely high or low value in a distribution can reduce the utility of range as a measure of dispersion. Thus, you may need a measure which is not unduly affected by the outliers.

In such a situation, if the entire data is divided into 4 equal parts, each containing 25% of the values, we get the values of Quartiles and Median. The upper and lower quartiles (Q_3 and Q_1, respectively) are used to calculate Inter Quartile Range which is $Q_3 - Q_1$. Inter-Quartile Range is based upon middle 50% of the values in a distribution and is, therefore, not affected by extreme values.

$$\text{Inter Quartile Range (IQR)} = Q_3 - Q_1$$

Half of the Inter-Quartile Range is called Quartile Deviation (QD).

$$\text{Quartile Deviation} = \frac{Q_3 - Q_1}{2}$$

The Quartile Deviation (QD) is also called **Semi inter Quartile Range.** It can generally be calculated for open-ended distributions and is not unduly affected by extreme values.

The Coefficient of Quartile Deviation (also called Quartile Coefficient of Dispersion) is given by following formula:

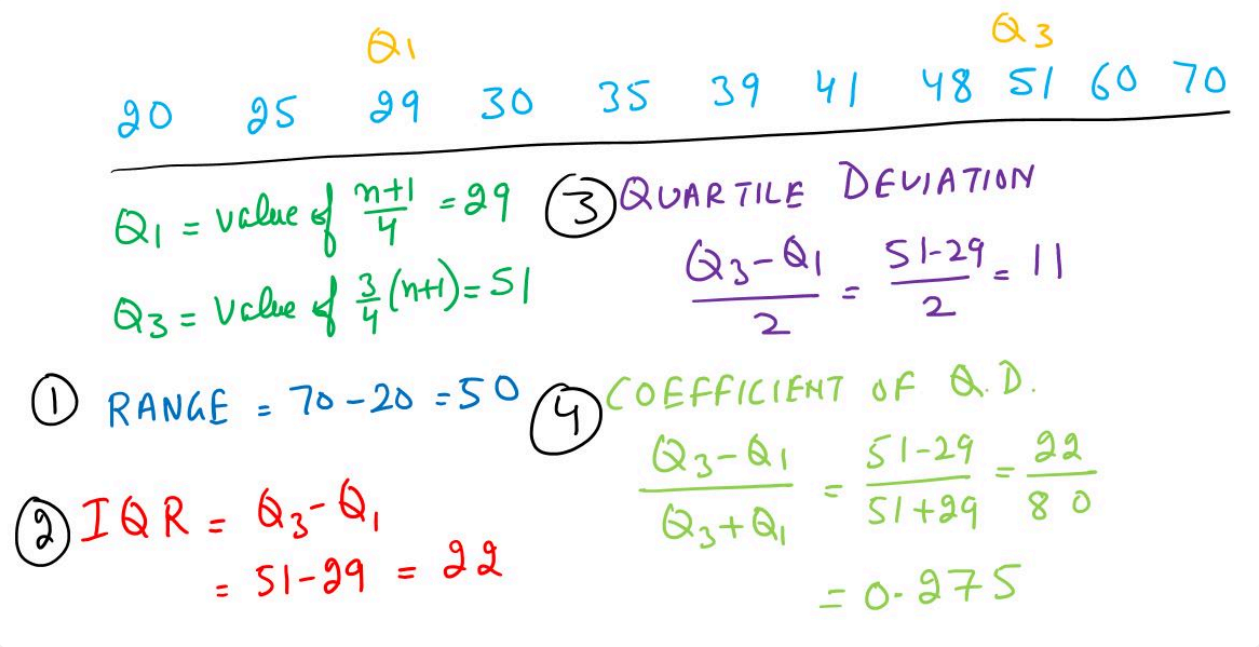
$$\text{Coefficient of Quartile Deviation} = \frac{Q_3 - Q_1}{Q_3 + Q_1}$$

---

3. Quartile based mesaures

Calculate Range, Inter Quartile Range, Quartile Deviation and Coefficient of Quartile Deviation for the following observations: 20, 25, 29, 30, 35, 39, 41, 48, 51, 60 and 70.

Solution:



4. Mean Absolute Deviation

The Mean Absolute Deviation (MAD) is the Arithmetic Mean of absolute deviations of the observations from a measure of central tendency. It is also named Average Deviation or Mean Deviation.

The steps to calculate mean deviation are given below:
(i) The A.M. of the values is calculated
(ii) Difference between each value and the A.M. is calculated. All differences are considered positive. These are denoted as |d|
(iii)The A.M. of these differences (called deviations) is the Mean Deviation.

$$\text{MEAN DEVIATION} = \frac{\sum |d|}{n}$$

Coefficient of Mean Deviation

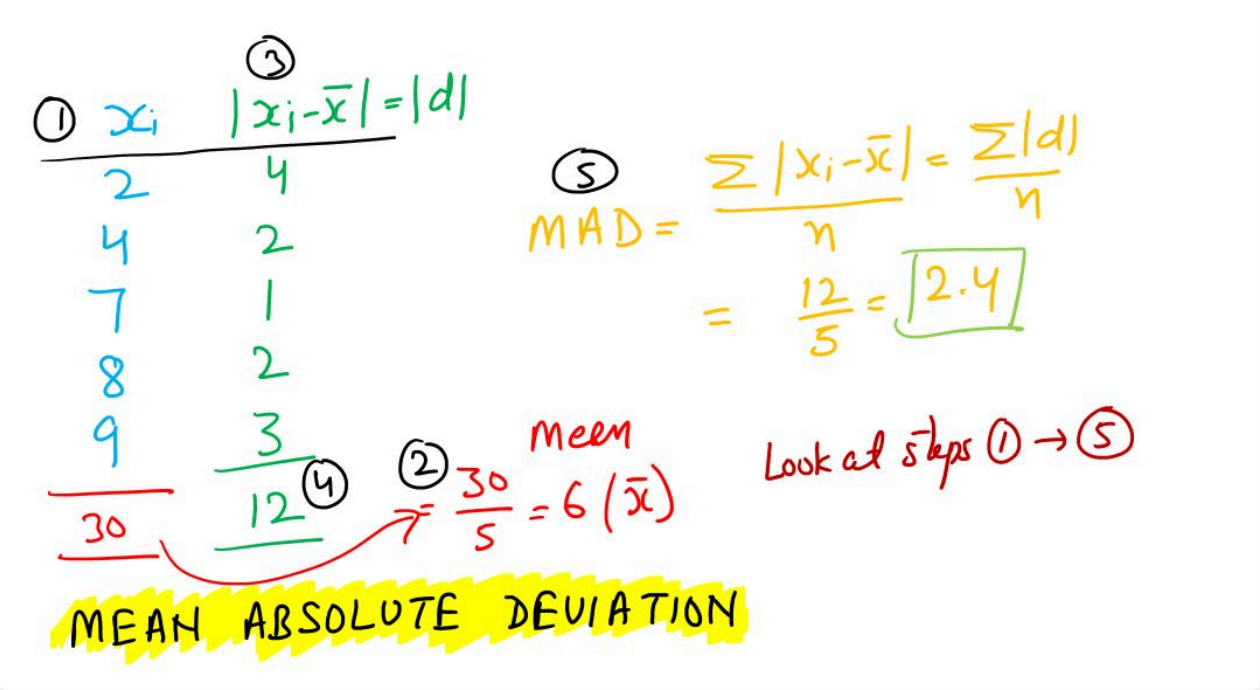
MAD is absolute mesaure of dispersion. Coefficient of Mean Deviation is the relative measure. It is given by following formula.

$$\text{COEFFICIENT OF MEAN DEVIATION} = \frac{\text{MEAN DEVIATION}}{\text{MEAN OR MEDIAN}}$$

4. Mean Absolute Deviation

Calculate the Mean Deviation of the following values: 2, 4, 7, 8 and 9.

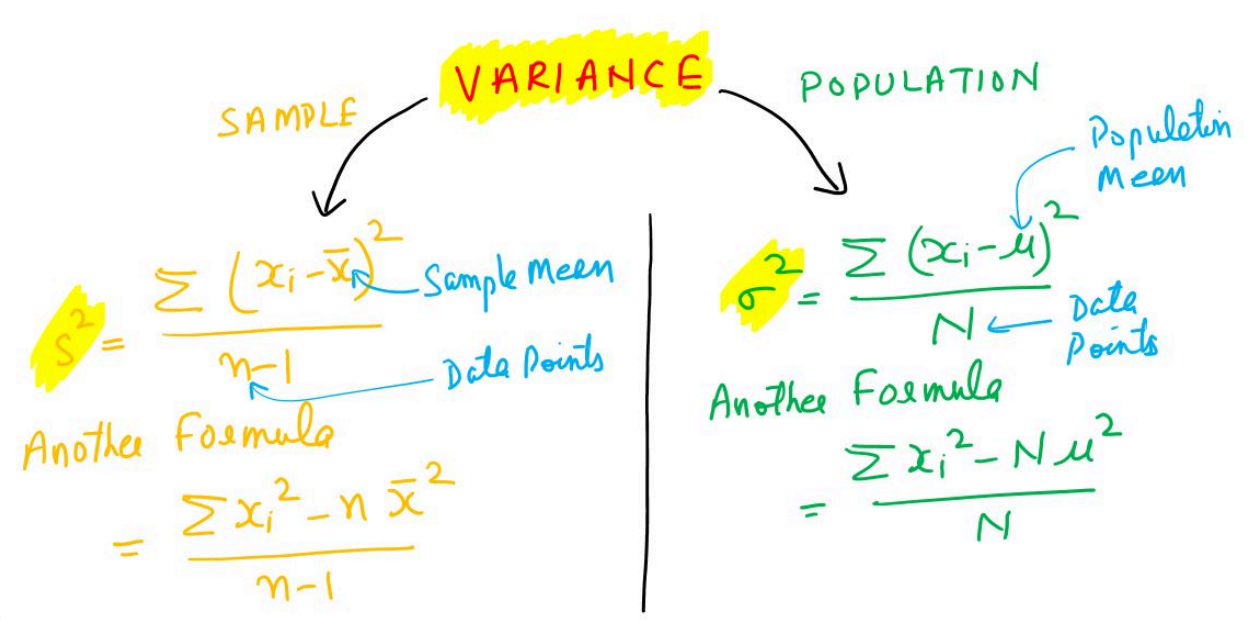
Solution:



5. Variance

Variance quantifies the dispersion or spread of a dataset around its mean. It is a measure of how much individual data points deviate from the mean value of the dataset.

Mathematically, the variance is calculated as the average of the squared differences between each data point and the mean.



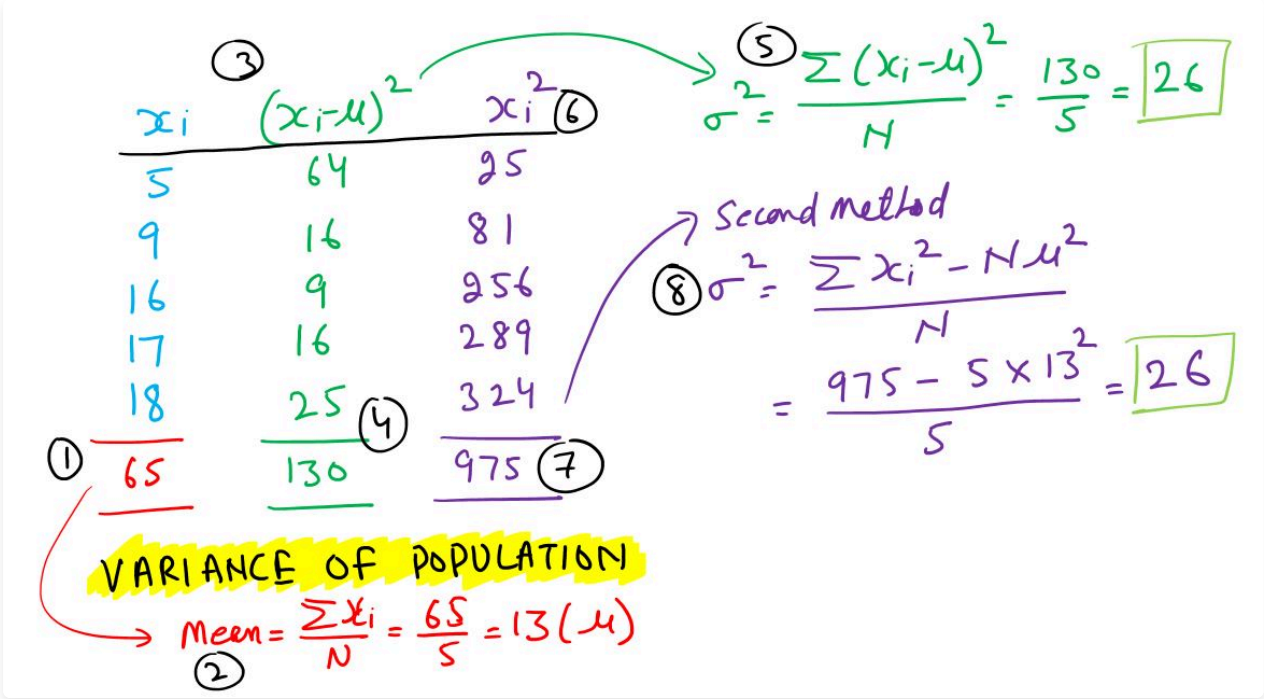
Variance indicates how spread out the values in a dataset are from the mean. A higher variance implies greater variability or dispersion of data points from the mean, while a lower variance suggests that data points are closer to the mean.

---

5. Variance

Calculate Variance of given data points of population: 5, 9, 16, 17, 18.

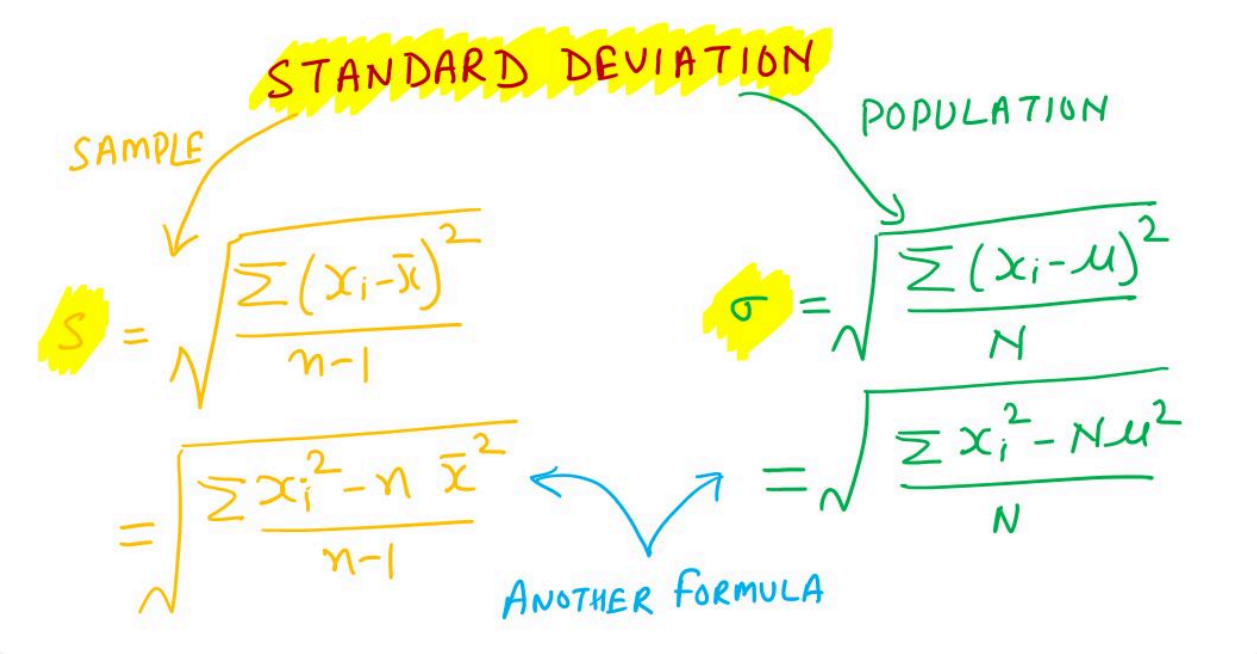
Solution:



6. Standard Deviation

Standard Deviation is the positive square root of the mean of squared deviations from mean. In other words, positive square root of the variance is the Standard deviation.

Variance gives an average of how far each data point in a set is from the mean, but in squared units. Standard deviation provides a measure of the average distance between each data point and the mean in the original units of the data set.



The standard deviation provides insight into the spread of the data points around the mean. A smaller standard deviation indicates that most data points are closer to the mean, while a larger standard deviation suggests that the data points are more spread out from the mean.

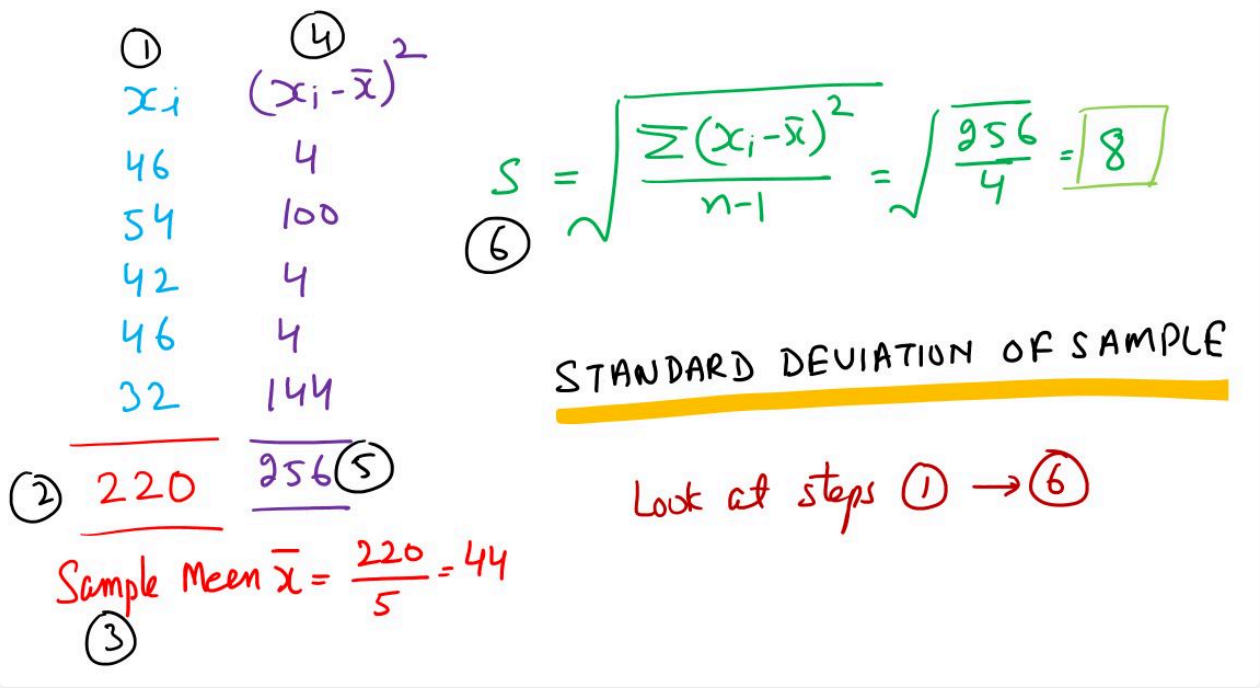
Standard Deviation, the most widely used measure of dispersion, is based on all values. Therefore, a change in even one value affects the value of standard deviation.

---

6. Standard Deviation

Calculate the standard deviation for following values of a sample: 46, 54, 42, 46, 32.

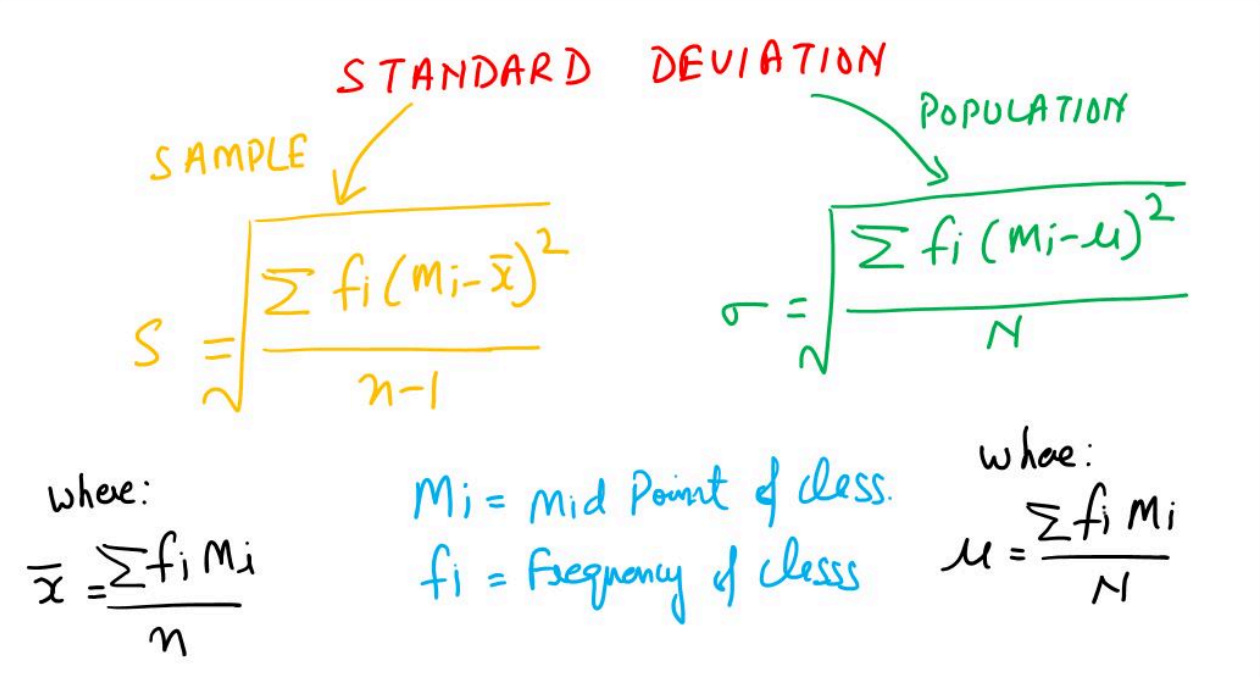
Solution:



---

7. Standard Deviation for Grouped Data

If we have to calculate standard deviation for grouped data, we use following formulas.

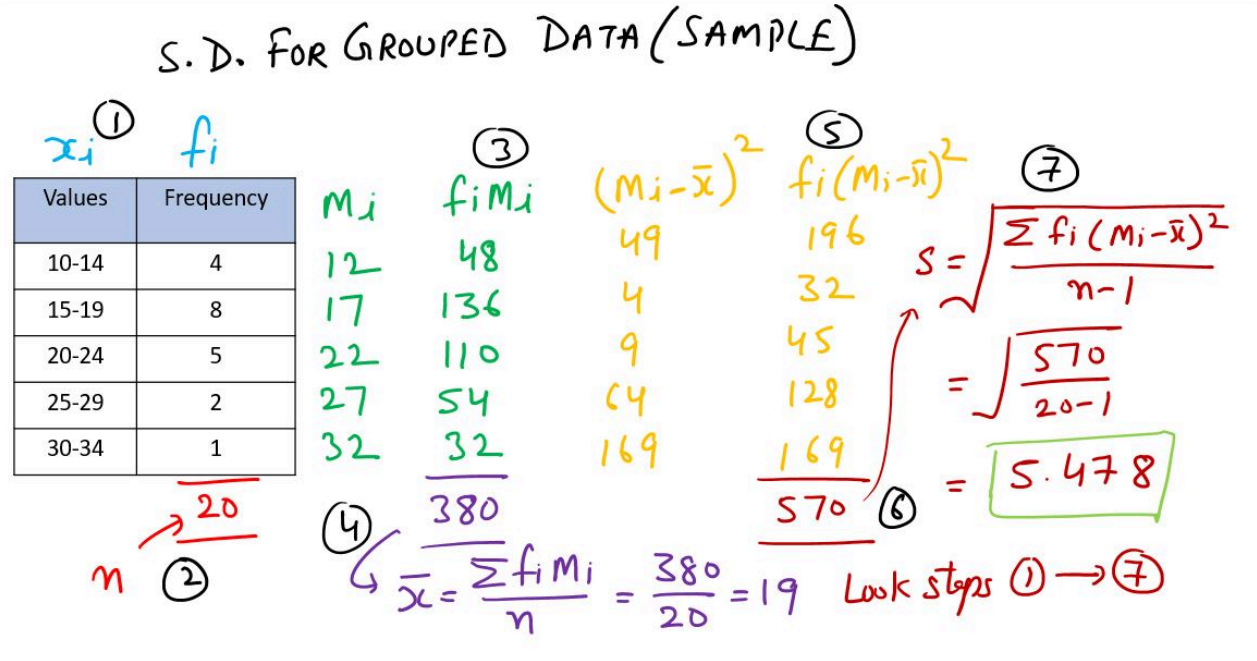


---

7. Standard Deviation for Grouped Data

Calculate standard deviation for given sample data.

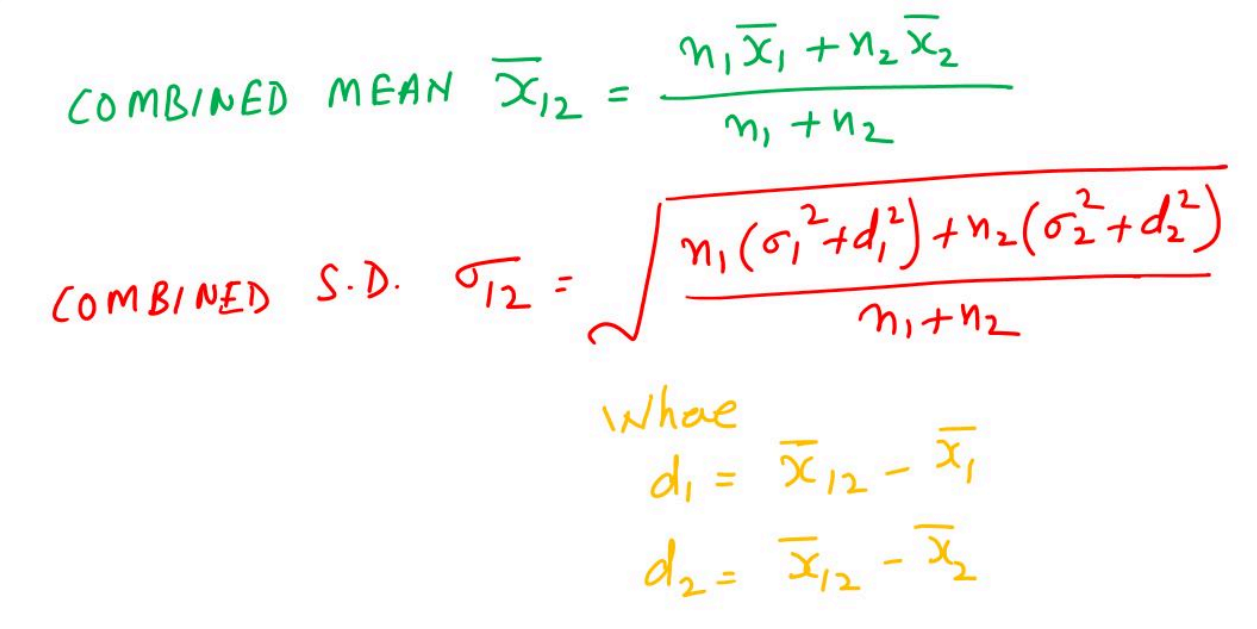
Solution:



---

8. Combined Mean and Standard Deviation

If there are two set of values having n_1 and n_2 data points, then combined mean and combined standard deviations are given by following formulas.



---

9. Coefficient of Variation

The Coefficient of Variation (CV) is a measure that describes the relative variability or dispersion in a data set, expressed as a percentage. It's used to compare the variability of datasets with different means or units of measurement.

Thus, for Standard Deviation, the relative measure is called Coefficient of Variation.

$$\text{Coefficient of Variation} = \frac{\text{Standard Deviation}}{\text{Mean}}$$

The CV is particularly useful when comparing the variability of datasets with different units or scales, allowing for a more standardized comparison. A lower CV indicates less variability relative to the mean, while a higher CV implies greater variability relative to the mean.

---

10. Relations among Deviations

In a Symmetrical Distribution, following 3 relationships exists:

1. Quartile deviation = 2/3 (Standard deviation)

2. Mean deviation = 4/5 (Standard deviation)

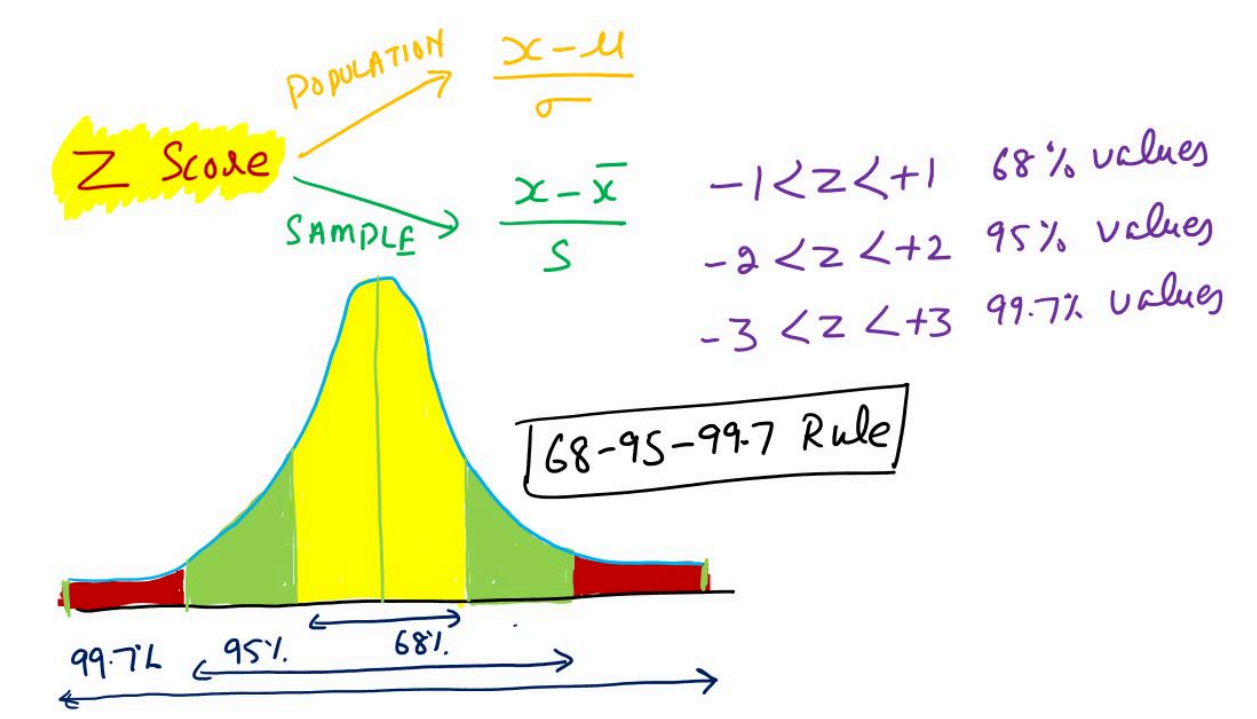
3. Quartile deviation = 5/6 (Mean deviation)

---

11. Z Score

The Z-score is a statistical measurement that describes a value's relationship to the mean in terms of standard deviations. It indicates how many standard deviations a data point is from the mean and in which direction—above or below the mean.

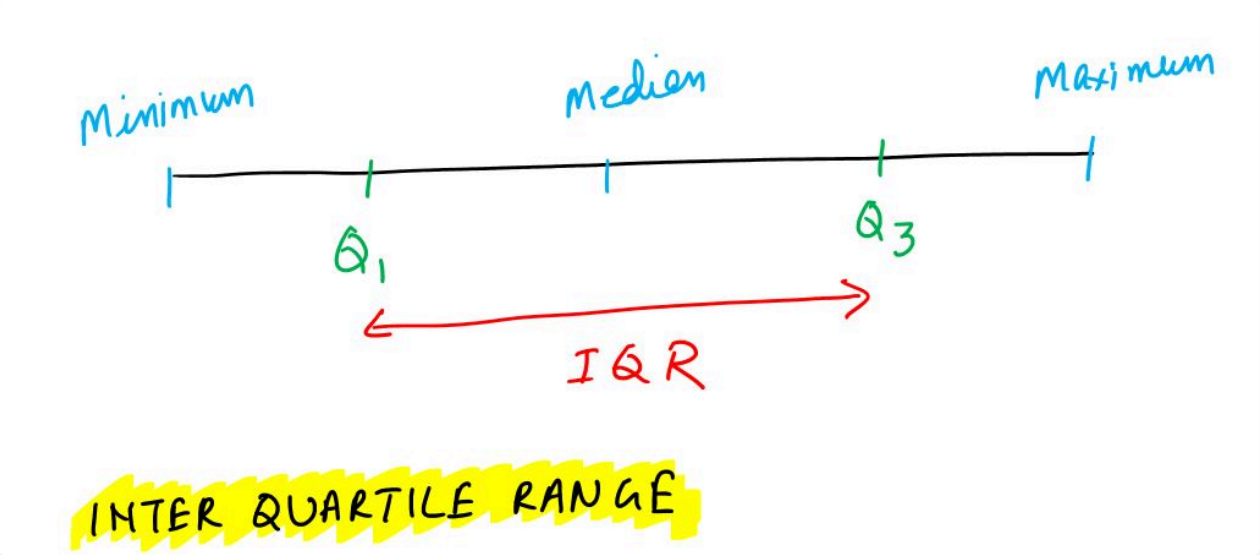
The formula for calculating the Z-score of a data point x in a normally distributed data set is given by following formula.



The Z-score helps in understanding the relative position of a particular data point within the distribution. If the Z-score is positive, the data point is above the mean, while a negative Z-score means the data point is below the mean. A Z-score of 0 means the data point is equal to the mean.

---

12. Inter Quartile Range



The Interquartile Range (IQR) is a measure of dispersion that quantifies the spread of a data set by focusing on the middle 50% of the observations. It is calculated as the difference between the third quartile (Q_3) and the first quartile (Q_1) in a data set.

$$IQR = Q_3 - Q_1$$

The IQR is considered a robust measure of dispersion because it is not influenced by extreme values or outliers in the data, unlike the range. It focuses on the spread of the central part of the data, making it more resistant to extreme values that might skew the measure. It is a resistant measure.

It is also named **Middle Fifty.**

---

13. Relative Measures of Dispersion

The absolute measures may give misleading ideas about the extent of variation specially when the averages differ significantly.

Another weakness of absolute measures is that they give the answer in the units in which original values are expressed. Consequently, if the values are expressed in kilometers, the dispersion will also be in kilometers. However, if the same values are expressed in meters, an absolute measure will give the answer in meters and the value of dispersion will appear to be 1000 times.

To overcome these problems, **relative measures** of dispersion can be used. Each absolute measure has a relative counterpart.

For Range, there is **Coefficient of Range**.
Coefficient of Range = $\frac{L-S}{L+S}$
where L = Largest value, S = Smallest value.

For Quartile Deviation, it is **Coefficient of Quartile Deviation**.
Coefficient of Quartile Deviation = $\frac{Q_3-Q_1}{Q_3+Q_1}$

For Mean Deviation, it is **Coefficient of Mean Deviation**. If Mean Deviation is calculated on the basis of the Mean, it is divided by the Mean. If Median is used to calculate Mean Deviation, it is divided by the Median.

Coefficient of Mean Deviation = $\frac{Mean \ Deviation}{Mean \ or \ Median}$

For Standard Deviation, the relative measure is called **Coefficient of Variation.**
Coefficient of Variation = $\frac{Standard \ Deviation}{Mean}$ × 100

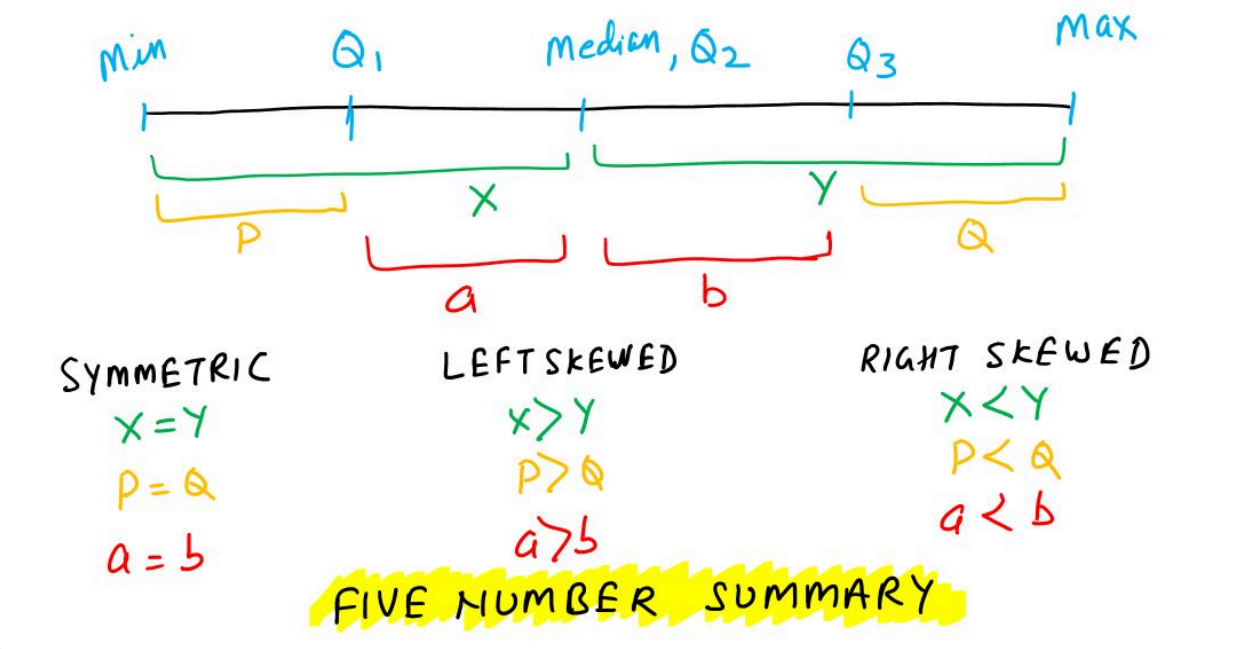
The measures of dispersion discussed so far give a numerical value of dispersion.

A graphical measure called **Lorenz Curve** is available for estimating in inequalities in distribution. You may have heard of statements like 'top 10% of the people of a country earn 50% of the national income while top 20% account for 80%'. An idea about income disparities is given by such figures. Lorenz Curve uses the information expressed in a cumulative manner to indicate the degree of inequality. For example Lorenz curve of income gives a relationship between percentage of population and its share of income in total income. It is specially useful in comparing the variability of two or more distributions by drawing two or more Lorenz curves on the same axis.

---

14. Five Number Summary

The Five Number Summary is a descriptive statistical summary of a dataset that includes 5 key values. These values help in understanding the distribution and essential characteristics of the data.



The five numbers are:

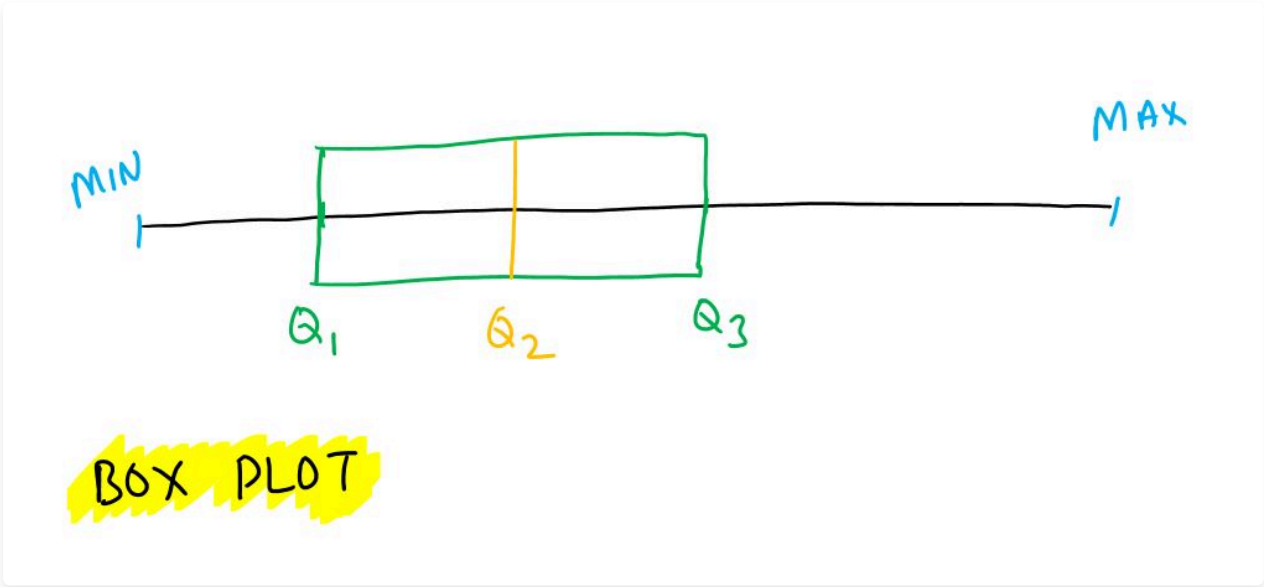
1. **Minimum**: The smallest value in the dataset.
2. **First Quartile** (Q_1): The value below which 25% of the data falls. It represents the lower 25th percentile.
3. **Median** (Q_2 or the second quartile): The middle value of the dataset when it's arranged in ascending order. It divides the data into two halves, with 50% of the data points lying below and 50% above.
4. **Third Quartile** (Q_3): The value below which 75% of the data falls. It represents the lower 75th percentile.
5. **Maximum**: The largest value in the dataset.

The Five Number Summary is commonly used in box plots, where the visual representation helps in understanding the spread, center, and distribution of the data.

These five key summary statistics provide a concise yet informative overview of the dataset's central tendency, variability, and presence of outliers.

---

14. Five Number Summary



A box plot, also known as a box-and-whisker plot, is a graphical representation based on the Five Number Summary: minimum, first quartile (Q_1), median (Q_2), third quartile (Q_3), and maximum. It displays the distribution and spread of a dataset along with potential outliers.

The box in the plot represents the interquartile range (IQR) and contains the middle 50% of the data, with the lower edge representing Q_1 and the upper edge representing Q_3. The line within the box marks the median.

The "whiskers" extend from the edges of the box to the minimum and maximum values, showing the range of the data excluding outliers. The length of the whiskers may vary based on the definition of outliers; sometimes, they represent a particular range of values or a specific multiple of the IQR.

Box plots provide a quick visual summary of the data's central tendency, variability, and skewness, making it easy to compare distributions across different groups or datasets. They're particularly useful for identifying the presence and extent of outliers and understanding the overall shape of the dataset.

---

15. Change of Original and Scale

Let us break down these statistical measures based on whether they are independent or dependent of changes in origin (shifting the data) and scale (changing the spread of the data):

Independent of change of origin and Dependent of change of scale.

- Range
- IQR
- Quartile Deviation
- Mean Deviation
- Standard Deviation
- Variance

Dependent of change of origin and Independent of change of scale

- Coefficient of Range
- Coefficient of Quartile Deviation

Dependent of change of origin and Dependent of change of scale

- Coefficient of Mean Deviation
- Coefficient of Variation

---

1. Skewness

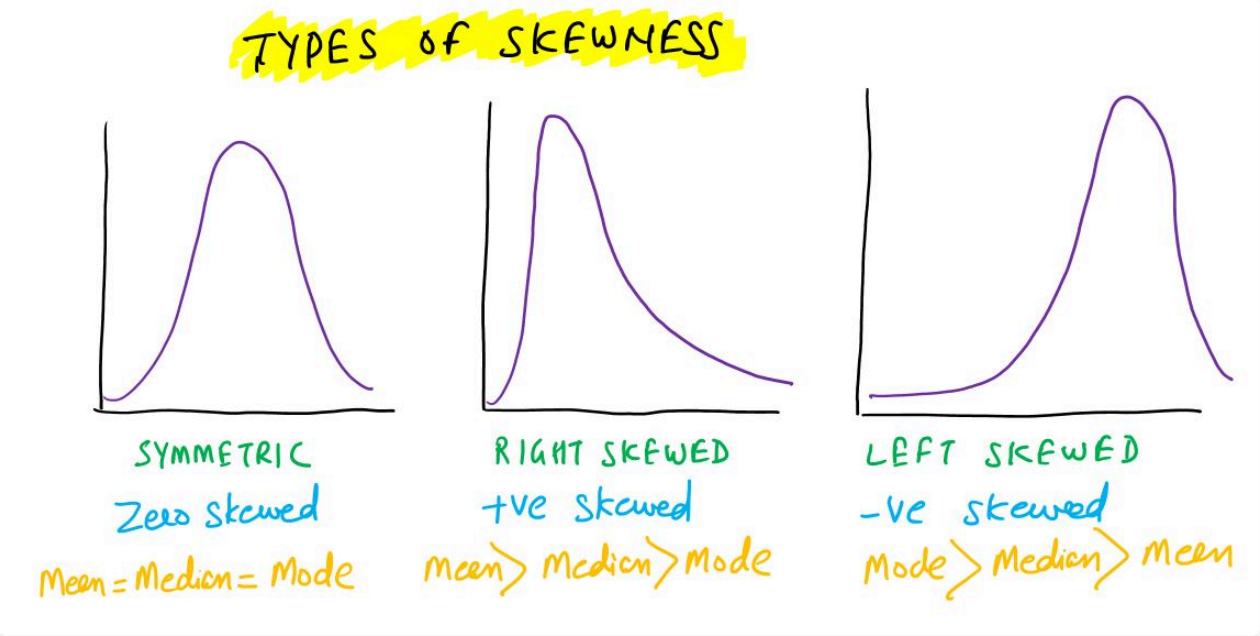
Skewness is a measure to describe the asymmetry or lack of symmetry in a probability distribution. It quantifies the degree and direction of asymmetry in the distribution of data points around the mean.

In simpler terms, it assesses the shape of the distribution.

When a distribution is symmetric, the mean, median, and mode tend to be similar. Skewness measures how far and in which direction the distribution differs from this symmetry.

---

2. Types of Skewness



Skewness provides insights into the shape and nature of the dataset, guiding analysts in understanding its characteristics. Skewness can be classified into three types:

1. Positive Skewness (Right Skewed)

This distribution has a longer tail on the right side, extending towards larger values. Most data points cluster on the left side, with a few exceptionally high values skewing the mean towards the right.

The mean is typically greater than the median and mode in a positively skewed distribution.

2. Negative Skewness (Left Skewed)

This distribution exhibits a longer tail on the left side, extending towards smaller values. The majority of data points concentrate on the right side, with a few exceptionally low values dragging the mean towards the left.

In a negatively skewed distribution, the mean is usually lower than the median and mode.

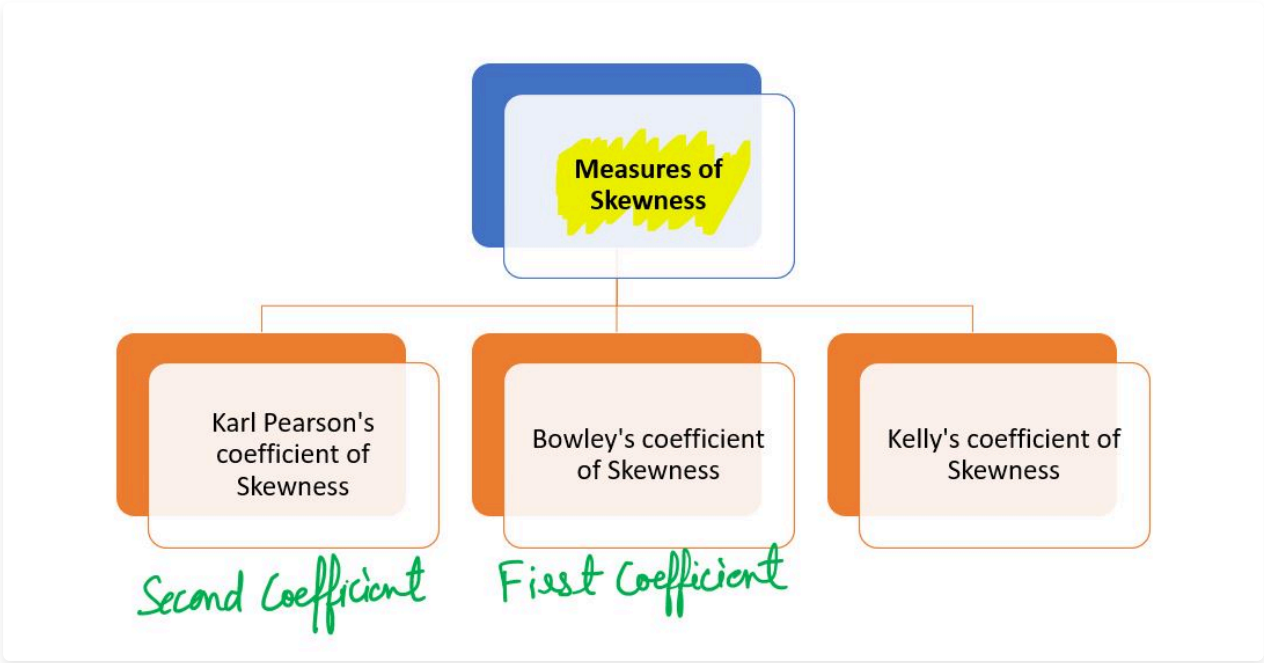
3. Zero Skewness

When skewness is zero, the distribution is perfectly symmetrical. Data points are evenly distributed around the mean, creating a symmetric distribution.

In such cases, the mean, median, and mode are all equal.

---

3. Measures of Skewness



There are three commonly used methods for measuring Skewness:

(i) Second coefficient of skewness, also known as Karl Pearson's coefficient of Skewness.

(ii) First coefficient of skewness, also known as Bowley's coefficient of Skewness.

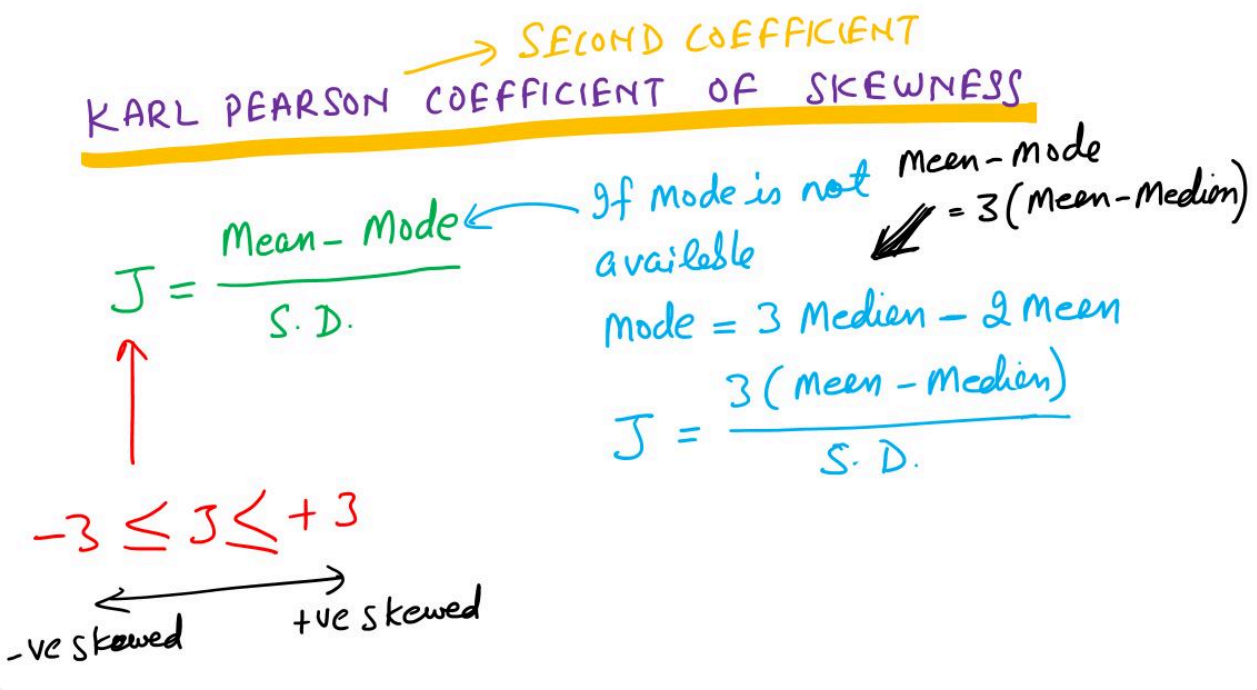
(iii) Kelly's coefficient of Skewness.

Let us discuss them one by one.

---

3. Measures of Skewness

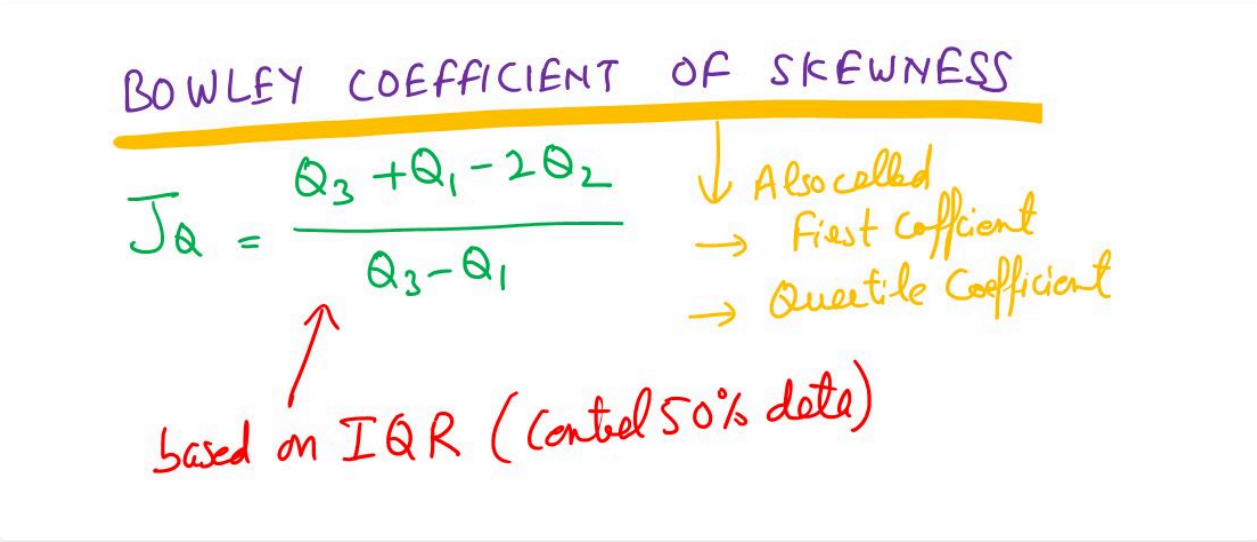
Second coefficient of skewness (also known as Karl Pearson's coefficient of Skewness, J) is calculated from mean and mode values.



---

3. Measures of Skewness

First coefficient of skewness (also known as Bowley's coefficient of Skewness, J_Q) is given by following formula.

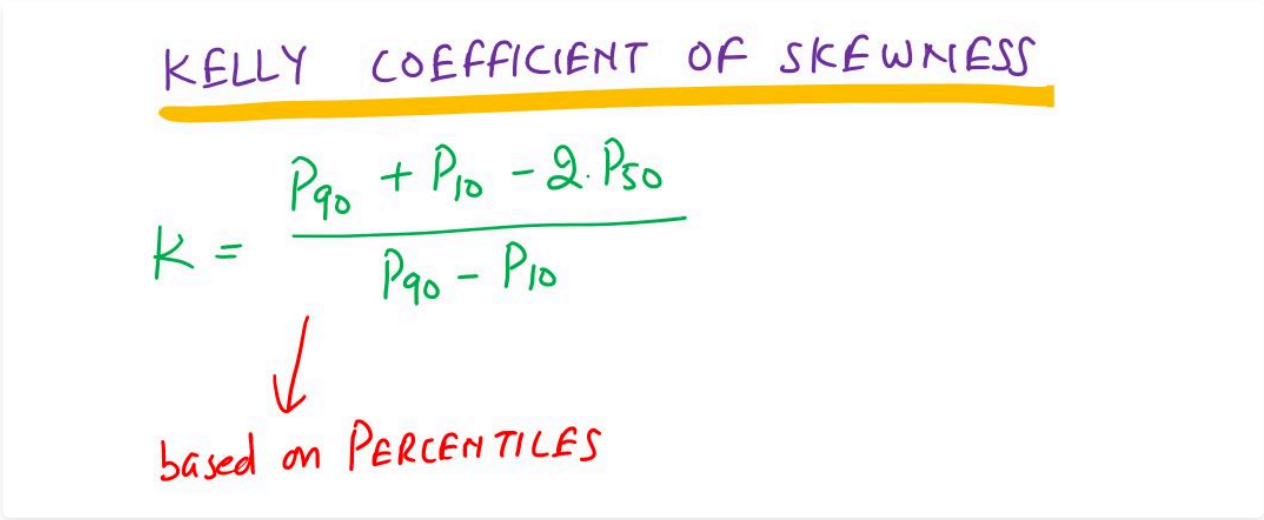


Bowley's measure of skewness is based on the middle 50% of the observations because it leaves 25% of the observations on each extreme of the distribution

---

3. Measures of Skewness

Kelly's Measure of Skewness is given by following formula.



As an improvement over Bowley's measure, Kelly has suggested a measure based on P_{10} and, P_{90} so that only 10% of the observations on each extreme are ignored.

All above 3 coefficients will be Zero, in case of perfectly symmetrical distribution.

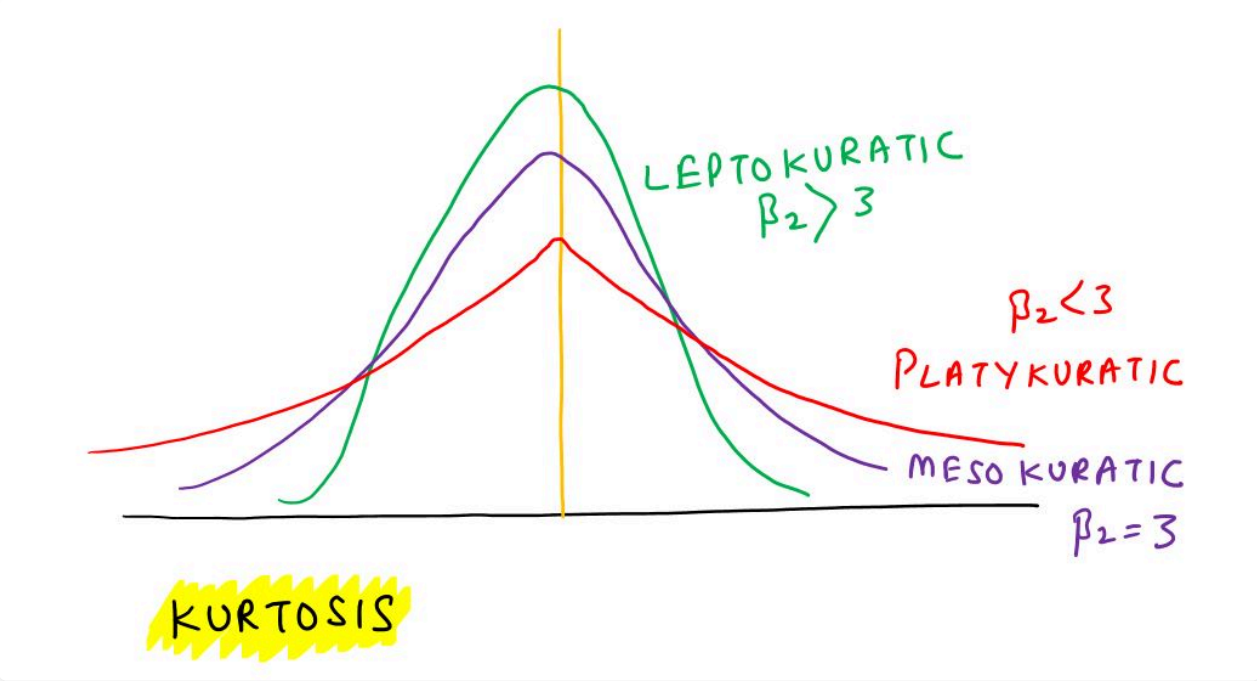
---

4. Kurtosis

In Greek language kurtosis means "bulgines". kurtosis indicates the nature of the vertex of the curve. Several statisticians defined kurtosis.

A measure of kurtosis indicates the degree to which a curve of the frequency distribution is peaked or flat-topped.

A measure of Kurtosis is given by β_2.



Karl Pearson defined following three types of curves :

1. **Normal Curve or Measokurtic Curve:** A curve which is neither flat nor peaked is called a normal curve or meso-kurtic curve. For such type of curve we have $\beta_2 = 3$

2. **Leptokurtic Curve:** A curve which is more peaked than the normal curve is called leptokurtic curve. For such type of curve, we have $\beta_2 > 3$.

3. **Platykurtic Curve**: A curve which is more flatter than the normal curve is called platykurtic curve. For such type of curve, we have $\beta_2 < 3$.

---

1. Introduction

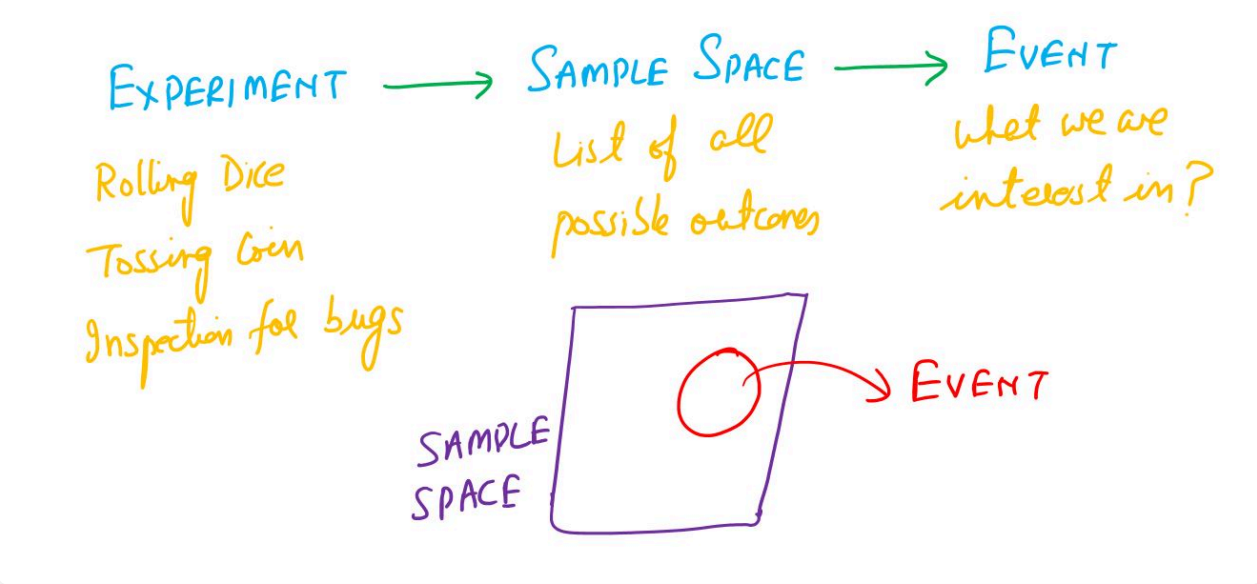
Probability represents the chance or likelihood of an event happening, ranging from impossible (0 probability) to certain (probability of 1).

It helps us quantify the uncertainty associated with various outcomes in situations governed by randomness or unpredictability.

---

2. Experiment, Sample Space and Event

Let us understand fundamental concepts of Experiment, Sample Space and Event of probability.



Experiment

An experiment is an action or process that results in well-defined outcomes.

Consider the experiment of rolling a six-sided die. When you roll the die, the action represents the experiment. The possible outcomes are the numbers that appear on the face of the die: {1, 2, 3, 4, 5, 6}.

Sample Space

The sample space is the set containing all possible outcomes of an experiment.

For the die experiment, the sample space is {1, 2, 3, 4, 5, 6}. Each individual number represents a specific outcome or sample point in the sample space.

Event

An event is any collection or subset of outcomes that we are interested in.

For instance, let's define an event A as getting an even number when rolling the die. The event A includes outcomes {2, 4, 6}.

Probability of an Event

The probability of an event is calculated by dividing the number of favorable outcomes by the total number of possible outcomes in the sample space.

In this example:

Experiment: Rolling the six-sided die.
Sample Space: The set of all possible outcomes {1, 2, 3, 4, 5, 6}.
Event: Event A, defined as getting an even number {2, 4, 6}.

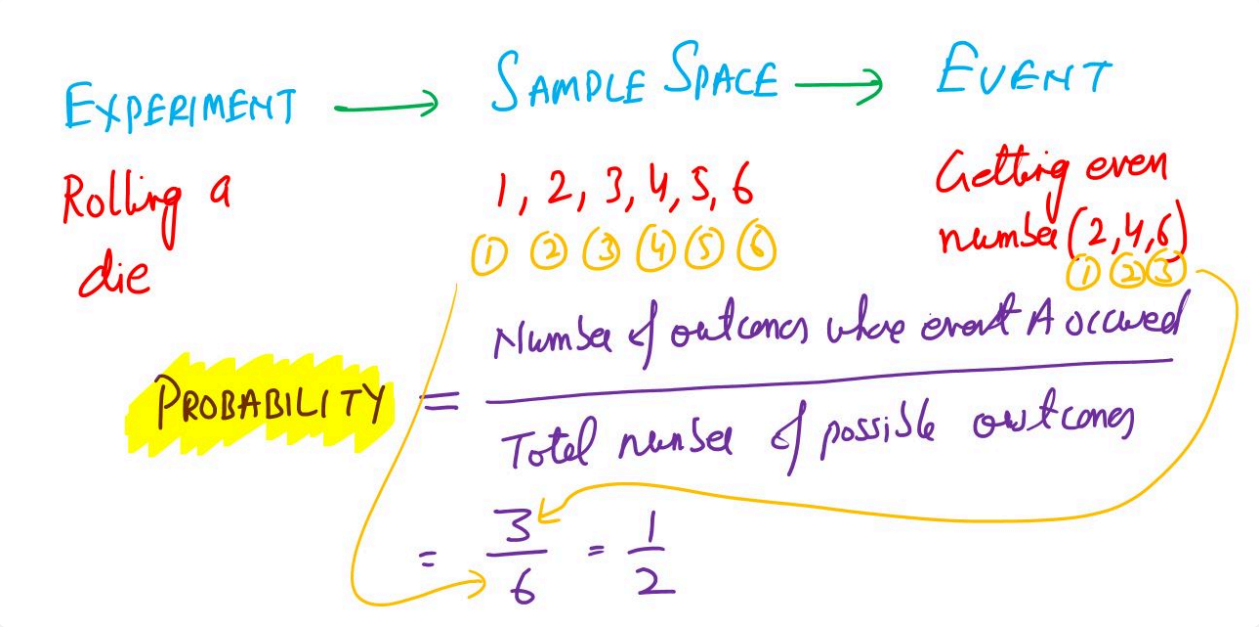
Another Example:

Experiment: Drawing a card from a standard deck.
Sample Space: The complete set of possible outcomes (52 cards).
Event: Event B, which is the subset of outcomes consisting of heart cards.

The sample space encompasses all possible outcomes resulting from the experiment. Events are subsets of the sample space, representing specific combinations of outcomes that we are interested in.

---

3. Probability of Event



The probability of an event (let's say event A) occurring is calculated by dividing the number of outcomes where event A occurs by the total number of possible outcomes within the sample space.

Probability of an event = $\frac{Number \ of \ outcomes \ where \ the \ event \ occurs}{Total \ number \ of \ possible \ outcomes}$

P (A) = Probability of event happening of event A.

A single probability means that only one event can take place. It is also known as **unconditional probability.**

---

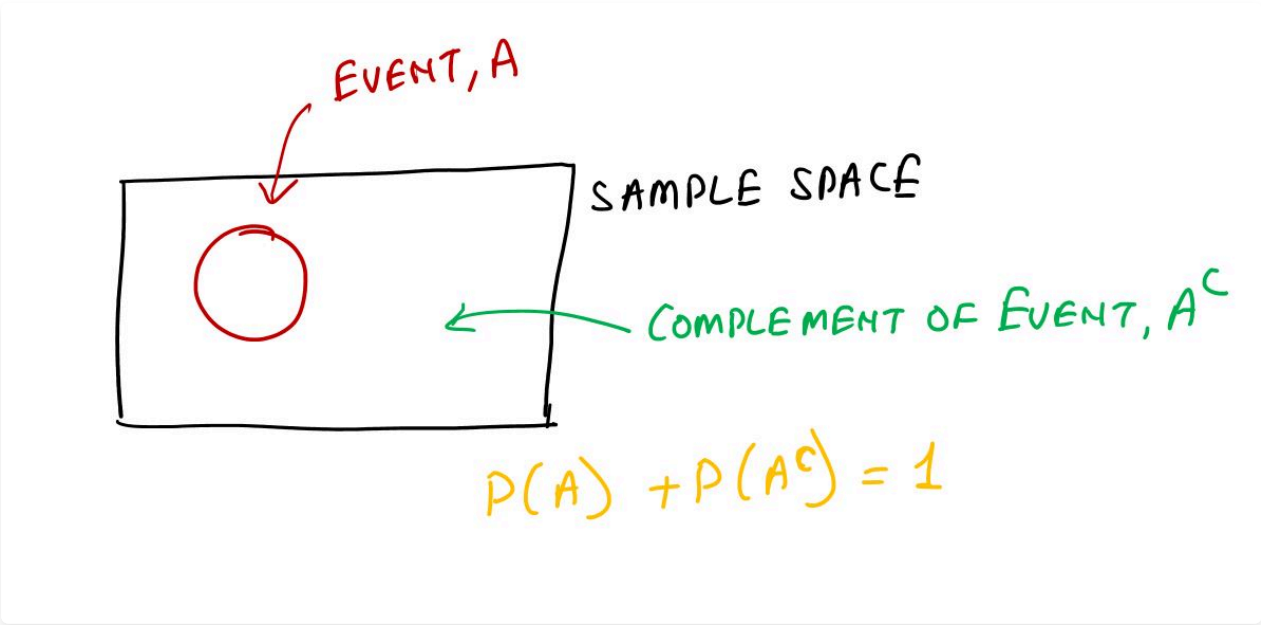
4. Properties of Events

Let us understand few inportant concepts regarding event.

---

4. Properties of Events

Given an event A, the complement of A is defined to be the event consisting of all sample points that are not in A. The complement of A is denoted by \(A^c \).



As shown in the figure, the rectangular area represents the sample space for the experiment and as such contains all possible sample points. The circle represents event A and contains only the sample points that belong to A. The shaded region of the rectangle contains all sample points not in event A and is by definition the complement of A.

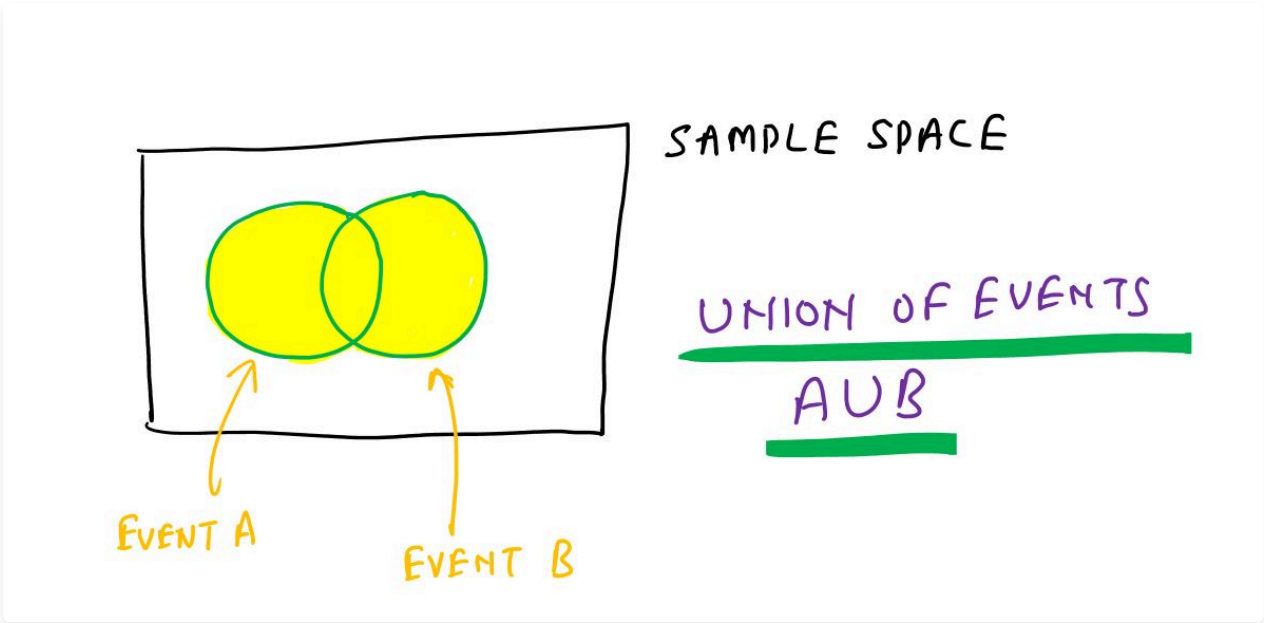
$P(A) + P(A^C) = 1$
i.e. the sum of probability of an event and its complement is always 1.

For example, if there is 0.90 probability that Mohan will pass in the exam. Using the complement, we can conclude that there is a 1-0.90 = 0.10 probability that Mohan will NOT pass the exam.

---

4. Properties of Events

The union of two events A and B (denoted as A∪B) represents the event that contains all the sample points that belong to either event A or event B, or to both events.



Let's consider two events:
Event A: Rolling an even number on a six-sided fair die.
Event B: Rolling a number greater than 3 on the same die.

The sample space of rolling a die is {1,2,3,4,5,6}.

Event A (rolling an even number) = {2,4,6}
Event B (rolling a number greater than 3) = {4,5,6}

Now, the union of A and B (A∪B) represents all the outcomes that belong to either event A or event B or to both.

$A \cup B = \{2,4,6\} \cup \{4,5,6\}$
$A \cup B = \{2,4,5,6\}$

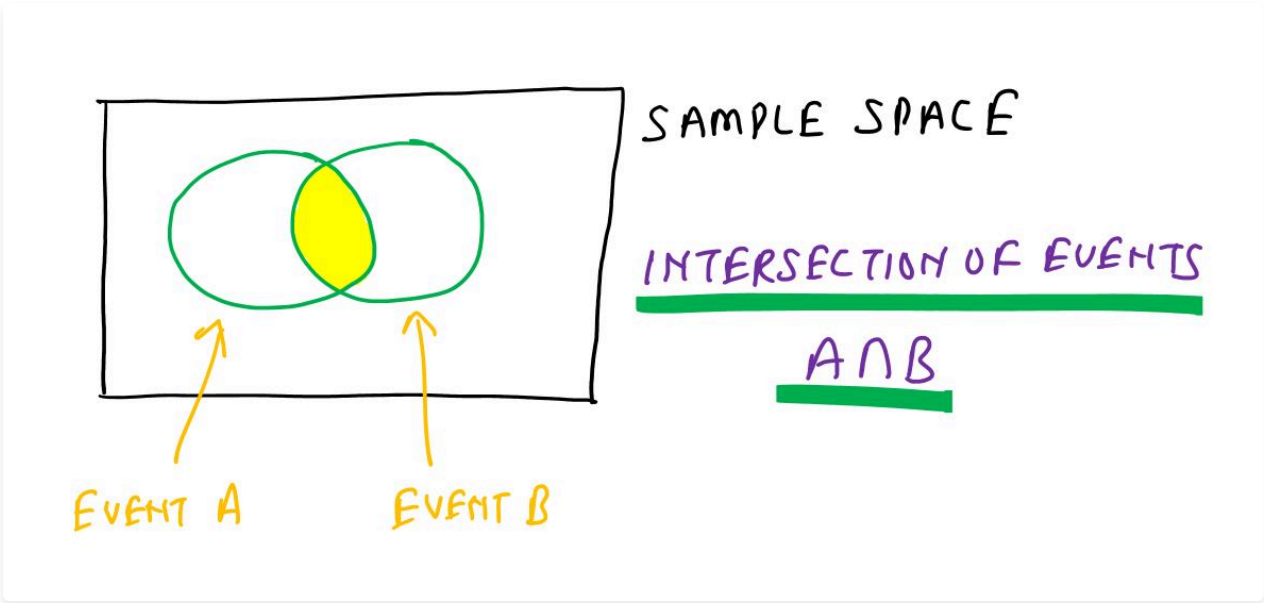
Therefore, the event A∪B includes all the outcomes that belong to event A or event B or to both events. In this example, it represents rolling an even number OR rolling a number greater than 3 OR rolling a number that satisfies both conditions, which gives us the set {2,4,5,6}.

It is represented by the shaded area in the figure.

---

4. Properties of Events

The intersection of two events A and B, denoted as A∩B, refers to the event that contains all the sample points that belong to both event A and event B simultaneously.



Let's use the same events:
Event A: Rolling an even number on a six-sided fair die.
Event B: Rolling a number greater than 3 on the same die.

The sample space of rolling a die is {1,2,3,4,5,6}.

Event A (rolling an even number) = {2,4,6}
Event B (rolling a number greater than 3) = {4,5,6}

The intersection of A and B (A∩B) represents all the outcomes that belong to both event A and event B.

$A \cap B = \{2,4,6\} \cap \{4,5,6\}$
$A \cap B = \{4,6\}$

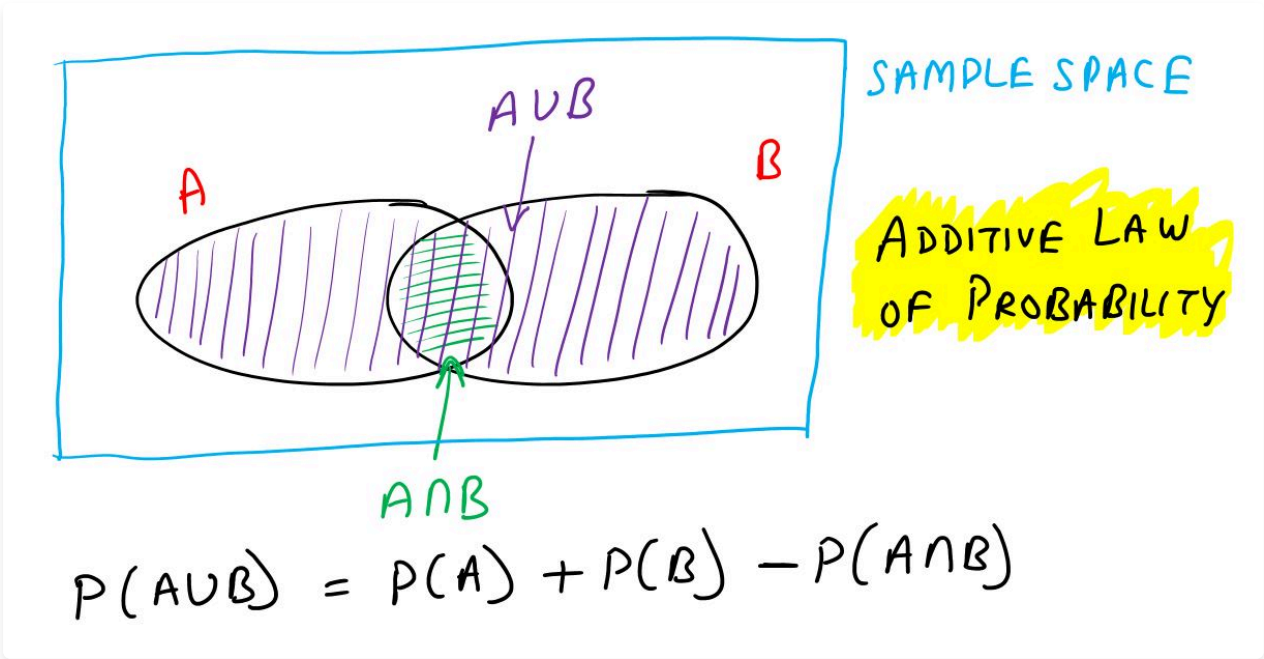
So, the event A∩B includes the outcomes that belong to both rolling an even number (Event A) and rolling a number greater than 3 (Event B), which gives us the set {4,6}. This set represents the common outcomes that satisfy both conditions.

It is represented by the figure where shaded portion gives measure of A∩B.

---

5. Additive Law of Probability

The Additive Law of Probability, also known as the Addition Rule, applies to the probability of the union of two events and deals with the probability of either event A or event B occurring or both.



For two events A and B, the additive law of probability states:

$P(A \cup B) = P(A) + P(B) - P(A \cap B)$

Illustration

Let A be the event of getting an odd number and B be the event of getting a prime number in a single throw of a die. What will be the probability that it is either an odd number or a prime number?

In a single throw of a die, the sample space would be
S = {1, 2, 3, 4, 5, 6}
The outcomes favourable to the events A and B are
A = {1, 3, 5} B = {2, 3, 5}
The outcomes favourable to the event A or B are
AUB = {1, 2, 3, 5}.
Thus, the probability of getting either an odd number or a prime number will be
P(A or B) = $\frac{4}{6} = \frac{2}{3}$

To discover an **alternate method**, we can proceed as follows:
The outcomes favorable to the event A are 1, 3 and 5.
∴ P(A) = $\frac{3}{6}$
Similarly, P(B) = $\frac{3}{6}$
The outcomes favorable to the event 'A and B' are 3 and 5.
∴ P(A and B) = $\frac{2}{6}$
Now, P(A) + P(B) - P(A and B) = $\frac{3}{6}$ + $\frac{3}{6}$ - $\frac{2}{6}$ = $\frac{4}{6}$ = $\frac{2}{3}$ = P(A or B)

Thus, we state the following law, called additive rule, which provides a technique for finding the probability of the union of two events, when they are not disjoint.

For any two events A and B of a sample space S,
P(A or B) = P(A) + P(B) – P(A and B)
or P(AUB) = P(A) + P(B) – P(A∩B)

---

5. Additive Law of Probability

A card is drawn from a well-shuffled deck of 52 cards. What is the probability that it is either a spade or a king?

Solution:
If a card is drawn at random from a well-shuffled deck of cards, the likelihood of any of the 52 cards being drawn is the same. Obviously, the sample space consists of 52 sample points.

If A and B denote the events of drawing a 'spade card' and a 'king' respectively, then the event A consists of 13 sample points, whereas the event B consists of 4 sample points. Therefore,
P(A) = $\frac{13}{52}$, P(B) = $\frac{4}{52}$
The compound event (A∩B) consist of only one sample point, viz.; king of spade. So,
P(A∩B) = $\frac{1}{52}$
Hence, the probability that the card drawn is either a spade or a king is given by

P(AUB) = P(A) + P(B) - P(A∩B) = $\frac{13}{52}$ + $\frac{4}{52}$ - $\frac{1}{52}$ = $\frac{16}{52}$ = $\frac{4}{13}$

---

5. Additive Law of Probability

In an experiment of throwing 2 fair dice, consider the events
A: The sum of numbers on the faces is 8
B: Doubles are thrown.
What is the probability of getting A or B?

Solution:
In a throw of two dice, the sample space consists of 6 × 6 = 36 sample points. The favourable outcomes to the event A (the sum of the numbers on the faces is 8) are:
A = {(2, 6), (3, 5), (4, 4), (5, 3), (6, 2)}

The favourable outcomes for the event B (Double means both dice have the same number) are
B = {(1, 1), (2, 2), (3, 3), (4, 4), (5, 5), (6, 6)}
A∩B = {(4,4)}.
Now P(A) = $\frac{5}{36}$

P(B) = $\frac{6}{36}$

P(A∩B) = $\frac{1}{36}$

Thus the probability of A or B is
P(AUB) = $\frac{5}{36}$ + $\frac{6}{36}$ - $\frac{1}{36}$ = $\frac{10}{36}$ = $\frac{5}{18}$

---

5. Additive Law of Probability

The probabilities that a student will receive an A, B, C or D grade are 0.30, 0.35, 0.20 and 0.15 respectively. What is the probability that a student will receive at least a B grade?

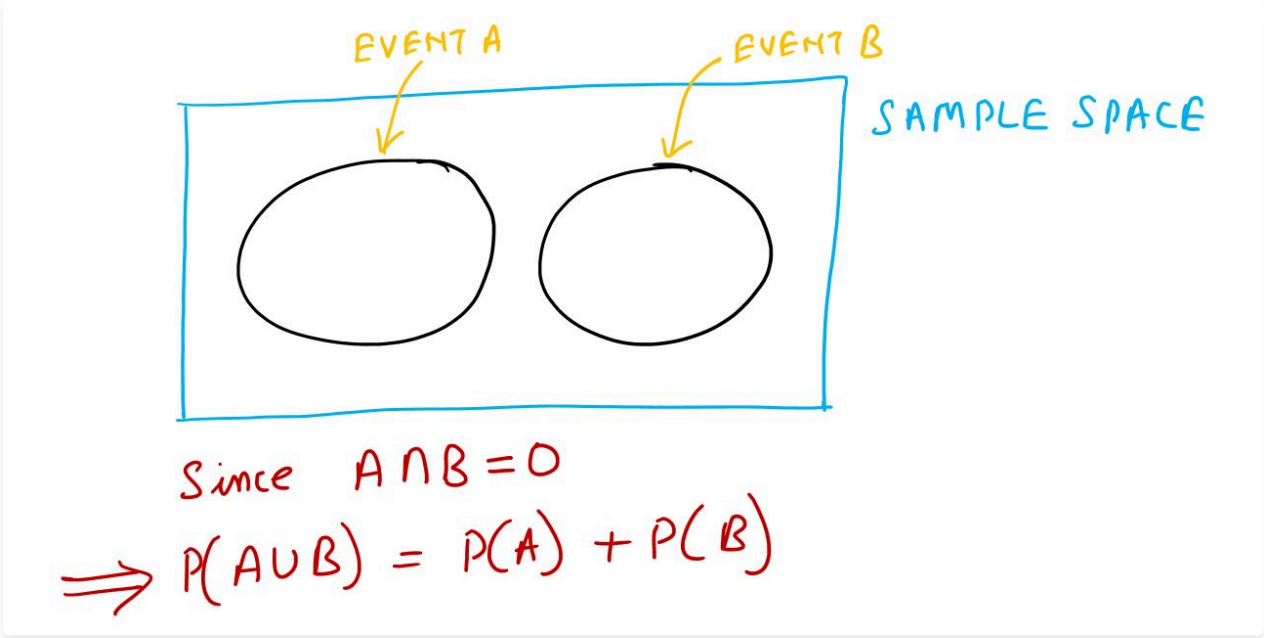
Solution:
The event at least a 'B' grade means that the student gets either a B grade or an A grade.
P(at least B grade) = P(B grade) + P(A grade)
= 0.35 + 0.30 = 0.65

---

6. Mutually Exclusive

Events are said to be mutually exclusive if one and only one can take place at a time. In other words, two events are said to be mutually exclusive if the events have no sample points in common. Thus, a requirement for A and B to be mutually exclusive is that their intersection must contain no sample points (look at Figure below). For example, if we toss a coin, only head or tail can occur but not both.

They are also called **Disjoint Events**.



Thus P (A∩B) will be equal to Zero.

So, Additive law for Mutually Exclusive events becomes:
P(AUB) = P(A) +P(B)

---

6. Mutually Exclusive

In a single throw of two dice, find the probability of a total of 9 or 11?

Solution:
Clearly, the events - a total of 9 and a total of 11 are mutually exclusive.
Now,
P (a total of 9) = P[(3,6),(4,5),(5,4),(6,3)] = $\frac{4}{36}$

P(a total of 11) = P[(5,6),(6,5)] = $\frac{2}{36}$

Thus, P(a total of 9 or 11) = $\frac{4}{36}$ + $\frac{2}{36}$ = $\frac{1}{6}$

---

7. Collectively Exhaustive

Collectively exhaustive events, also known as collectively exhaustive outcomes or exhaustive events, refer to a situation in probability theory where a set of events covers all possible outcomes of an experiment or a situation. In simpler terms, if you consider all the events together, they encompass every possible outcome that could occur.

For example, when rolling a six-sided die, the events {rolling a 1}, {rolling a 2}, {rolling a 3}, {rolling a 4}, {rolling a 5}, and {rolling a 6} form a set of collectively exhaustive events. When you roll the die, one of these events must happen—there's no other possible outcome.

In a survey asking about people's favorite colors, if the options are red, blue, green, and yellow, these options together represent collectively exhaustive events. Any respondent's favorite color would fall into one of these categories, covering all possibilities.

---

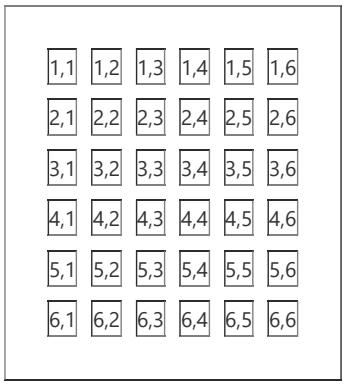
7. Collectively Exhaustive

Illustration 1

In a single throw of two dice, what is the probability that the sum is 9?

Solutions:

The number of possible outcomes is 6×6=36. We write them as given below:



Now, how do we get a total of 9. We have:
3 + 6 = 9
4 + 5 = 9
5 + 4 = 9
6 + 3 = 9
In other words, the outcomes (3,6), (4,5), (5,4) and (6,3) are favorable to the said event, i.e., the number of favorable outcomes is 4.
Hence, P (a total of 9) = $\frac{4}{36}= \frac{1}{9}$

Illustration 2

From a bag containing 10 red, 4 blue and 6 black balls, a ball is drawn at random. What is the probability of drawing (i) a red ball (ii) a blue ball (iii) not a black ball?

Solution:

There are 20 balls in all. So, the total number of possible outcomes is 20. (Random drawing of balls ensure equally likely outcomes)

(i) Number of red balls = 10
∴ P (a red ball) = $\frac{10}{20}= \frac{1}{2}$

(ii) Number of blue balls = 4
∴ P (a blue ball) = $\frac{4}{20}= \frac{1}{5}$

(iii) Number of balls which are not black = 10+ 4 = 14
∴ P (not a black ball) = $\frac{14}{20}= \frac{7}{10}$

Illustration 3

A card is drawn at random from a well shuffled deck of 52 cards. If A is the event of getting a queen and B is the event of getting a card bearing a number greater than 4 but less than 10, find P(A) and P (B).

Solution:

Well shuffled pack of cards ensures equally likely outcomes.
∴ the total number of possible outcomes is 52.

(i) There are 4 queens in a pack of cards.
∴ P(A) = $\frac{4}{52}= \frac{1}{13}$

(ii) The cards bearing a number greater than 4 but less than 10 are 5, 6, 7, 8 and 9.
Each card bearing any of the above number is of 4 suits diamond, spade, club or heart.

Thus, the number of favorable outcomes = 5 × 4 = 20
∴ = $\frac{20}{52}= \frac{5}{13}$

Illustration 4

A bag contains 3 red, 6 white and 7 blue balls. What is the probability that two balls drawn are white and blue?

Solution:
Total number of balls = 3 + 6 + 7 = 16
Now, out of 16 balls, 2 can be drawn in ${}_2^{16}C$ ways
∴ Exhaustive number of cases = ${}_2^{16}C$ = $\frac{16 \times 15 }{2}$ = 120

Out of 6 white balls, 1 ball can be drawn in ${}_1^{6}C$ ways and out of 7 blue balls, one can be drawn is ${}_1^{7}C$ ways.
Since each of the former case is associated with each of the later case,
Therefore, total number of favorable cases is ${}_1^{6}C$ × ${}_1^{7}C$ = 6 ×7 = 42.

∴ Required probability = $\frac{42}{120}$ = $\frac{7}{20}$

Note: When two or more balls are drawn from a bag containing several balls, there are two ways in which these balls can be drawn.
(i) Without replacement: The ball first drawn is not put back in the bag, when the second ball is drawn. The third ball is also drawn without putting back the balls drawn earlier and so on. Obviously, the case of drawing the balls without replacement is the same as drawing them together.
(ii) With replacement: In this case, the ball drawn is put back in the bag before drawing the next ball. Here the number of balls in the bag remains the same, every time a ball is drawn.

In these types of problems, unless stated otherwise, we consider the problem of without replacement.

Illustration 5

Find the probability of getting both red balls, when from a bag containing 5 red and 4 black balls, two balls are drawn,
(i) with replacement.
(ii) without replacement.

Solution:
(i) Total number of balls in the bag is both the draws = 5 + 4 = 9
Probability of drawing first red ball = $\frac{5}{9}$

Since the second ball is drawn after replacing the first ball, probability of drawing second red ball = $\frac{5}{9}$
Hence, probability (both red balls) = $\frac{5}{9} \times \frac{5}{9}= \frac{25}{81}$

(ii) total number of possible outcomes is equal to the number of ways of selecting 2 balls out of 9 balls = ${}_2^9C$
Number of favorable cases is equal to the number of ways of selecting 2 balls out of 5 red balls = ${}_2^5C$
Hence, P (both red balls) = $\frac{{}_2^5C}{{}_2^9C}$ = $\frac{ \frac{5 \times 4 }{1 \times 2 } }{ \frac{9 \times 8 }{1 \times 2 }}$
= $\frac{5}{18}$

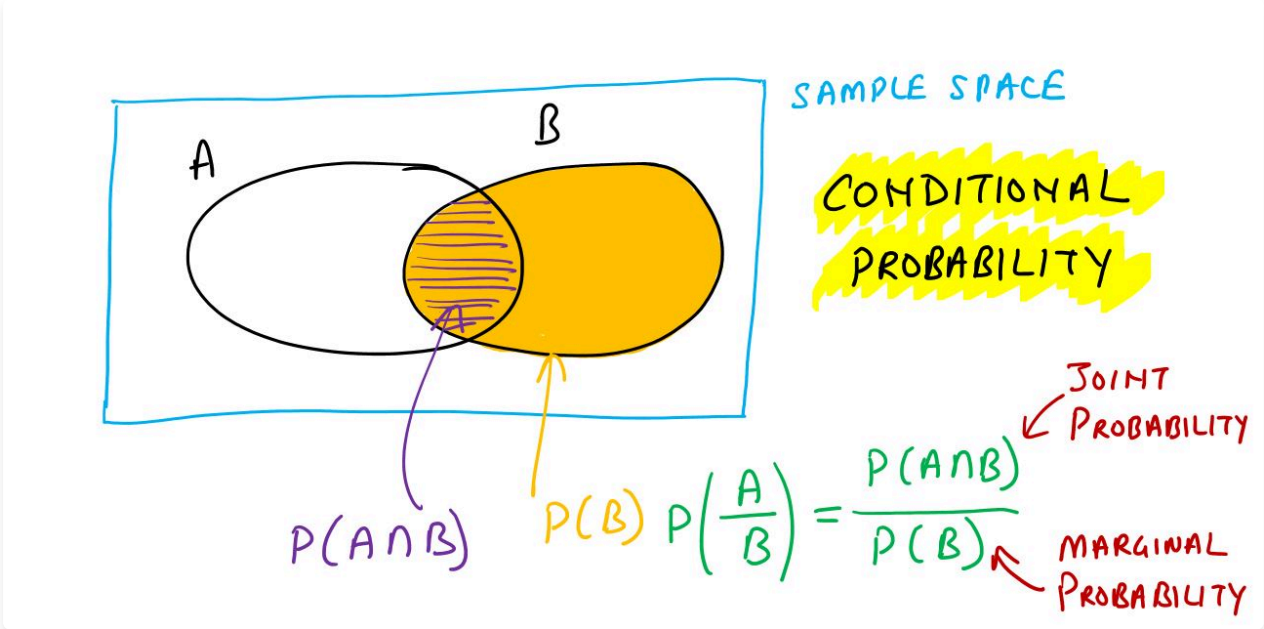
---

8. Conditional Probability

Often, the probability of an event is influenced by whether a related event already occurred. Suppose we have an event A with probability P(A). If we obtain new information and learn that a related event, denoted by B, already occurred, we will want to take advantage of this information by calculating a new probability for event A. This new probability of event A is called a conditional probability and is written P(A|B). We use the notation "|" to indicate that we are considering the probability of event A given the condition that event B has occurred. Hence, the notation P(A|B) reads "the probability of A given B."

$P(A|B) = \frac{P(A \cap B)}{P(B)}$

Here numerator $P(A \cap B)$ is called as **Joint Probability** and denominator P(B) is called as **Marginal Probability**. Thus the conditional probabilities can be computed as the ratio of a joint probability to a marginal probability. Thus, Joint probability of two independent events represents the case when both events occur together. Marginal probability of an event A refers to the probability of A in a joint probability setting.



The Venn diagram in the Figure below is helpful in obtaining an intuitive understanding of conditional probability. The circle on the right shows that event B has occurred; the portion of the circle that overlaps with event A denotes the event $(A \cap B)$. We know that once event B has occurred, the only way that we can also observe event A is for the event $(A \cap B)$ to occur. Thus, the ratio $\frac{P(A \cap B)}{P(B)}$ provides the conditional probability that we will observe event A given that event B has already occurred.

Illustration

In a company with a workforce of 1200 employees, it is observed that 960 are men and remaining are women. Among the men, 288 were promoted, while among the women, only 36 received promotions. Based on the available data, explore whether there might be evidence of gender bias in the promotion process.

Solution:
We need to calculate the conditional probability of an employee being promoted, given that the employee is a woman.

	Men	Women	TOTAL
PROMOTED	288	36	324
NOT PROMOTED	672	204	876
TOTAL	960	240	1200

PROMOTED given the MEN

$$P\left(\frac{P}{M}\right) = \frac{P(P \cap M)}{P(M)} = \frac{0.24}{0.80} = 0.30$$

PROMOTED given the WOMEN

$$P\left(\frac{P}{W}\right) = \frac{P(P \cap W)}{P(W)} = \frac{0.03}{0.20} = 0.15$$

$\frac{288}{1200} = 0.24$, $\frac{960}{1200} = 0.80$

JOINT PROBABILITY

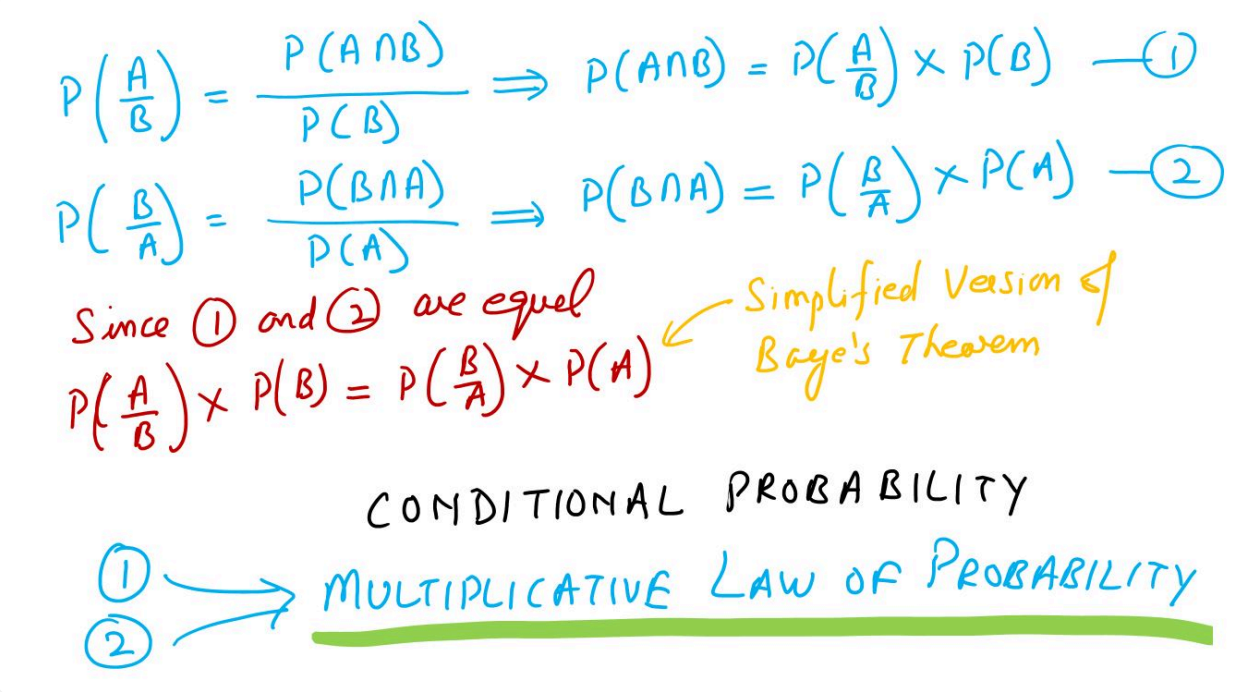
	Men	Women	TOTAL
PROMOTED	0.24	0.03	0.27
NOT PROMOTED	0.56	0.17	0.73
TOTAL	0.80	0.20	1.00

MARGINAL PROBABILITY

YES, THERE IS GENDER BIAS

8. Conditional Probability

Whereas the addition law of probability is used to compute the probability of a union of two events, the **multiplicative law** is used to compute the probability of the intersection of two events.



It states that the probability of the joint occurrence of two independent events is the product of their individual probabilities.

The multiplication law is based on the definition of conditional probability.
P(A|B) = \(\frac{P(A∩B)}{P(B)} \), We can re-write it as
$P(A \cap B) = P(A \mid B) \times P(B)$..... **X**

Similarly we can re-write P(B|A) = (P(A∩B))/(P(A)) as below
$P(A \cap B) = P(B \mid A) \times P(A)$..... **Y**

Equations **X** and **Y** shows Multiplicative Law of Probability.

---

8. Conditional Probability

Illustration 1

Assume that a certain school contains equal number of female and male students. 5% of the male population is football players. Find the probability that a randomly selected student is a male football player.

Solution:
Let M = Male, F = Football player
We wish to calculate P(M∩F). From the given data,
P(M) = $\frac{1}{2}$ (∵School contains equal number of male and female students)
P(F|M) = 0.05

But from definition of conditional probability, we have
P(F|M) = $\frac{(P(M∩F)}{P(M)}$
P(M∩F) = P(M) × P(F|M)
= $\frac{1}{2} \times 0.05 = 0.025$

Illustration 2

If A and B are two events, such that P(A) = 0.8, P(B)=0.6, P(A∩B) = 0.5, find the value of
(i) P(A∪B)
(ii) P(B|A)
(iii) P(A|B)

Solution:
(i) P(A∪B) = P(A) + P(B) - P(A∩B) = 0.8 + 0.6 - 0.5 = 0.9

(ii) P(B|A) = $\frac{P(A∩B)}{P(A)}$) = $\frac{0.5}{0.8} = \frac{5}{8}$

(iii) P(A|B) = $\frac{P(A∩B)}{P(B)}$) = $\frac{0.5}{0.6} = \frac{5}{6}$

Illustration 3

Find the chance of drawing 2 white balls in succession from a bag containing 5 red and 7 white balls, the balls drawn not being replaced.

Solution:
Let A be the event that ball drawn is white in the first draw. B be the event that ball drawn is white in the second draw.
∴ P(A∩B) = P(A) $\times$ P(B|A)

Here P(A) = $\frac{7}{12}$, P(B|A) = $\frac{6}{11}$
∴ P(A∩B) = $\frac{7}{12}$ × $\frac{6}{11}$ = $\frac{7}{22}$

9. Independent Events

In conditional probability, events are interrelated. If the occurrence of one event influences the probability of another event, these events are termed **dependent**. This means that the outcome of one event affects the likelihood of the subsequent event.

Conversely, if the events are **independent**, the occurrence of one event doesn't impact the occurrence of another event. In independent events, the outcome of one event doesn't change the probability of the other event happening.

Examples of Independent Events

1. Flipping a fair coin multiple times. The outcome of one flip (heads or tails) doesn't influence the outcome of subsequent flips. Each flip remains independent, with a consistent 50% chance of landing heads or tails.

2. The weather on one day is typically independent of the weather on the following day. For instance, if today is sunny, it doesn't affect the chances of tomorrow being sunny, rainy, or cloudy. Weather conditions are usually considered independent events from day to day.

3. Rolling a fair six-sided die multiple times. The outcome of one roll doesn't impact the outcome of the next. Each roll of the die is an independent event with an equal probability of landing on any of the six faces.

Examples of dependent events

1. If cards are drawn from a deck without replacement, the events are dependent. For instance, if you draw a card from a standard deck and don't replace it, the probability of drawing a certain card on the second draw changes because the deck has one fewer card.

2. Consider a bag containing different colored marbles. If you draw a marble and don't put it back, the probability of drawing a specific color on subsequent draws changes because the number of marbles of that color has decreased.

3. In a sequential process, one event influences the probability of another. For instance, the probability of raining today might influence the probability of carrying an umbrella tomorrow based on whether it rained or not.

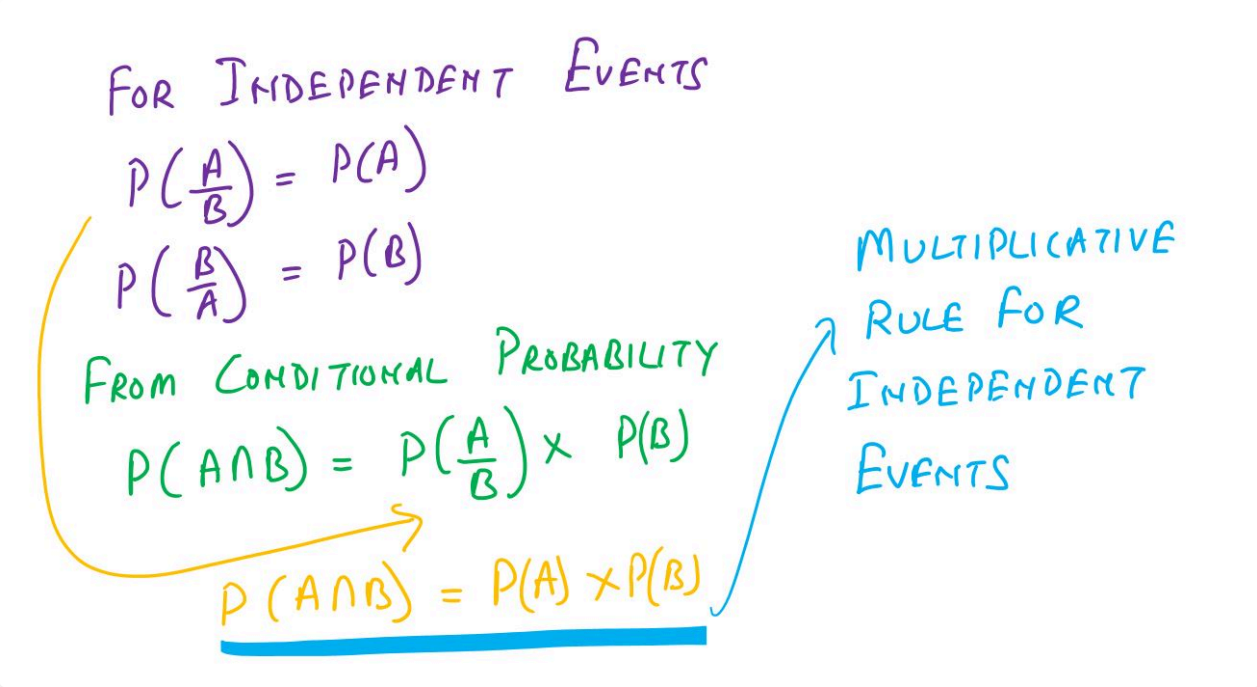
Now, applying concept of Conditional Probability on Independent Events, we get the following:

$P(A|B) = P(A)$ (Because incident of happening of B does not affect probability of A)

Similarly, $P(B|A) = P(B)$

---

9. Independent Events



We have learnt that for Independent events:
$P(B|A) = P(B)$ and $P(A|B) = P(A)$

Applying them in our formulas of conditional probability, we get:
$P(A \cap B) = P(A) \times P(B)$
This is called **Multiplicative Rule for Independent Events**.

We can extend Multiplicative rule for Independent events to more events:
$P(A \cap B \cap C) = P(A) \times P(B) \times P(C)$

Illustration

Suppose an individual applying to a college determines that he has an 80% chance of being accepted, and he knows that dormitory housing will only be provided for 60% of all of the accepted students.

The chance of the student being accepted and receiving dormitory housing is defined by:

P (Accepted and Dormitory Housing) = P (Dormitory Housing | Accepted) P(Accepted)

= (0.60) × (0.80) = 0.48.

---

9. Independent Events

Illustration 1

A die is tossed twice. Find the probability of a number greater than 4 on each throw.

Solution:
Let us denote by A, the event 'a number greater than 4' on first throw. B be the event 'a number greater than 4' in the second throw. Clearly A and B are independent events. In the first throw, there are two outcomes, namely, 5 and 6 favourable to the event A.
∴ P(A) = $\frac{2}{6} = \frac{1}{3}$
Similarly, P(B) = $\frac{1}{3}$
Hence, P (A and B) = P(A) × P(B)
= $\frac{1}{3}$ × $\frac{1}{3}$ = $\frac{1}{9}$

Illustration 2

Two balls are drawn at random with replacement from a box containing 15 red and 10 white balls. Calculate the probability that
(a) both balls are red.
(b) first ball is red and the second is white.
(c) one of them is white and the other is red.

Solution:
(a) Let A be the event that first drawn ball is red and B be the event that the second ball drawn is red. Then, as the balls drawn are with replacement,

Therefore, P(A) = $\frac{15}{25} = \frac{3}{5}$

P(B) = $\frac{3}{5}$
As A and B are independent events
Therefore P (Both red) = P(A and B)
= P(A) × P(B)
= $\frac{3}{5}$ × $\frac{3}{5}$ = $\frac{9}{25}$

(b) Let A: First ball drawn is red
B: Second ball drawn is white.
∴ P(A and B) = P(A) × P(B)
= $\frac{3}{5}$ × $\frac{2}{5}$ = $\frac{6}{25}$

(c) If WR denotes the event of getting a white ball in the first draw and a red ball in the second draw and the even RW of getting a red ball in the first draw and a white ball in the second draw. Then as 'RW' and 'WR' are mutually exclusive events, therefore
∴ P (a white and a red ball)
= P(WR or RW)
= P(WR) + P (RW)
= P(W) × P(R) + P(R) × P(W)

= $\frac{2}{5}$ × $\frac{3}{5}$ + $\frac{3}{5}$ × $\frac{2}{5}$
= $\frac{6}{25}$ + $\frac{6}{25}$ = $\frac{12}{25}$

Illustration 3

The probability that a new advertising campaign will increase sales is 0.80. The probability that the cost of developing the new advertising campaign can be kept within the original budget allocation is 0.40. Assuming that the two events are independent, what is the probability that the: *(UPSC Management Optional 2018 Paper)*
(i) cost is kept within budget or the campaign will increase sales?
(ii) cost is kept within budget and the campaign will increase sales?
(iii) cost is not kept within budget or the campaign will not increase sales?
(iv) cost is not kept within budget and the campaign will not increase sales?

Solution:

$P(S) = 0.80 \quad P(S') = 0.20 \quad P(B) = 0.40 \quad P(B') = 0.60$

(i) Budget or Scle, $P(B \cup S) = P(B) + P(S) - P(B \cap S)$

$\uparrow P(B) \times P(S)$

$= 0.80 + 0.40 - 0.80 \times 0.40$

$= \boxed{0.88}$

(ii) Budget and Scle, $P(B \cap S) = P(B) \times P(S) = 0.80 \times 0.40 = \boxed{0.32}$

(iii) Neither Budget nor Scle, $P(B' \cap S') = P(B') \times P(S') = 0.20 \times 0.60$

$= \boxed{0.12}$

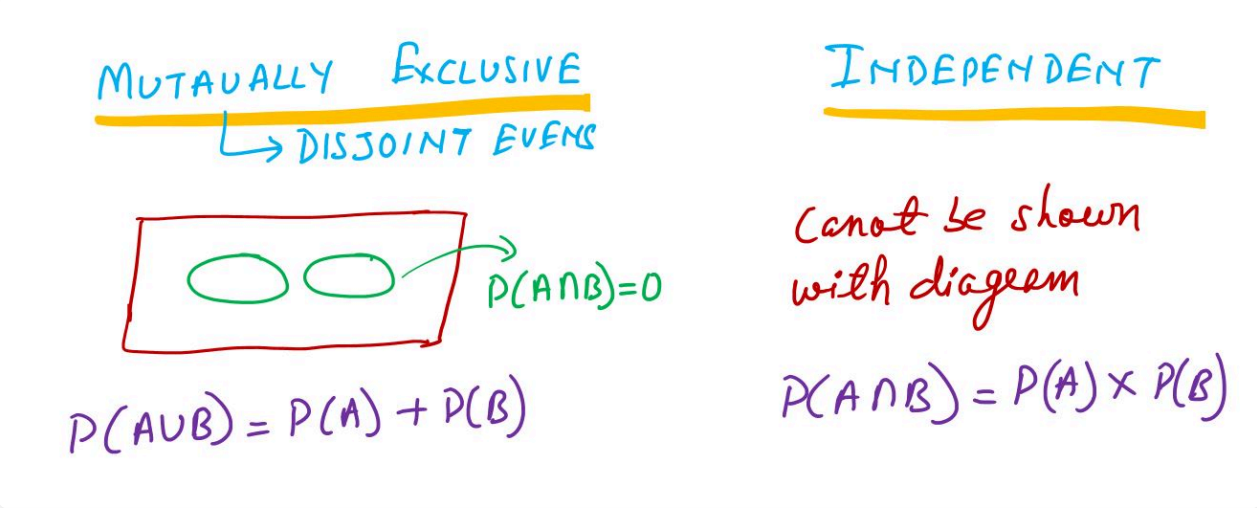
(iv) Not Budget and Not Scle, $P(B' \cup S') = P(B') + P(S') - P(B' \cap S')$

$= 0.60 + 0.20 - 0.60 \times 0.20$

$= \boxed{0.68}$

10. Mutually Exclusive and Independent events

While mutually exclusive and independent events might seem similar in terms of their relation, they represent different relationships.



Mutually Exclusive Events

These are events that cannot occur simultaneously. If one event happens, the other event cannot happen at the same time. For instance, when flipping a coin, getting a "heads" and getting a "tails" are mutually exclusive events. They cannot happen together in a single flip. In a roll of a die, getting a 1 and getting a 6 are also mutually exclusive events. The key characteristic here is that the occurrence of one event excludes the possibility of the other event occurring at the same time.

If A and B are **mutually exclusive**, then P (A and B) = 0; that is, the probability that both A and B occur is zero. They are also called **disjoint events**.

Independent Events

These are events where the occurrence or non-occurrence of one event does not affect the probability of another event occurring. For instance, when rolling a die, the outcome of one roll does not influence the outcome of the next roll. Similarly, drawing a red ball from a bag and then drawing a blue ball (without replacement) are independent events if the probability of drawing a blue ball remains unchanged after the red ball is drawn.

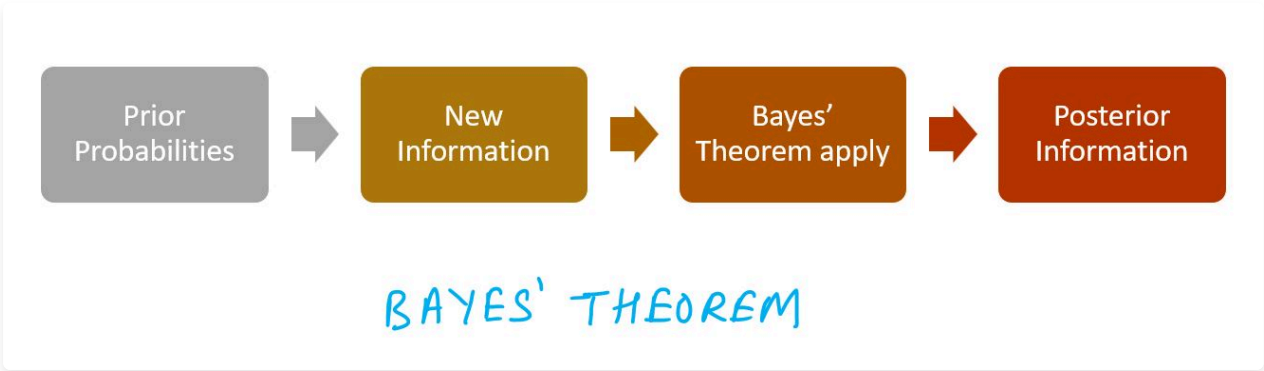
If two events A and B are **Independent**, then
P (A and B) = P(A) × P(B)

This means, the probability that both A and B occur is equal to the probability that A occurs times the probability that B occurs.

---

11. Bayes' Theorem

In the discussion of conditional probability, we indicated that revising probabilities when new information is obtained is an important phase of probability analysis. Often, we begin the analysis with initial or prior probability estimates for specific events of interest. Then, we obtain additional information about the events. Given this new information, we update the prior probability values by calculating revised probabilities, referred to as posterior probabilities. Bayes' theorem provides a means for making these probability calculations.



The steps of calculation using Bayes' Theorem are given in the figure.

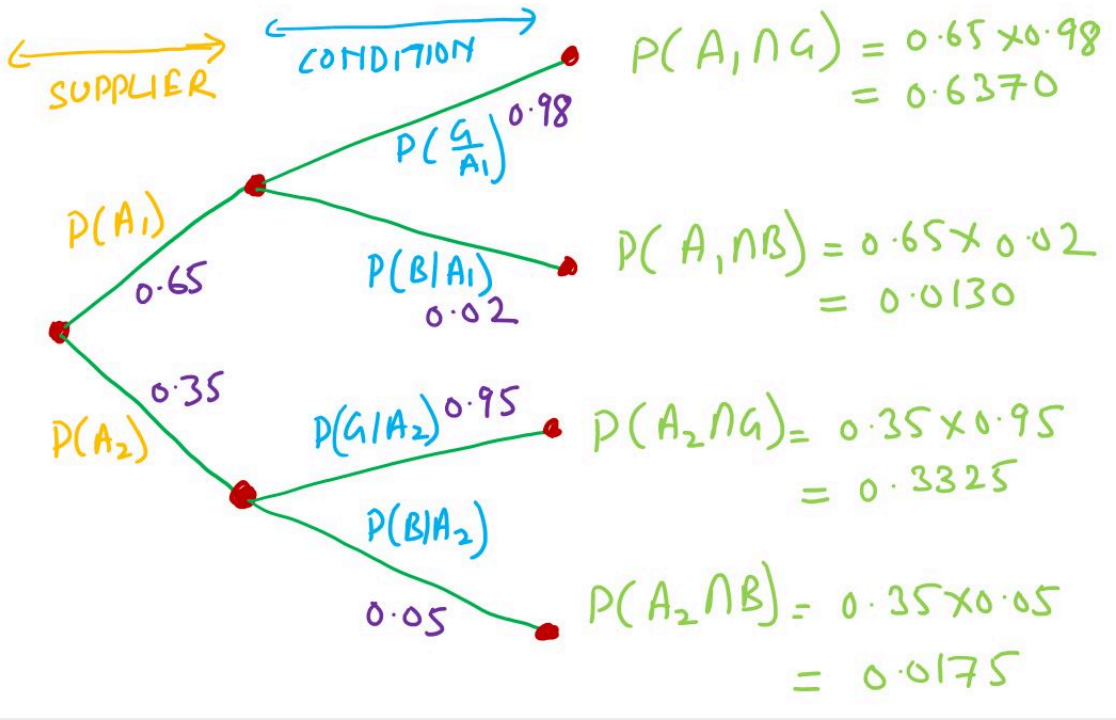
Let us understand this with an example.

A firm receives parts from Supplier 1 and Supplier 2. Let A_1 denote the event that a part is from supplier 1 and A_2 denote the event that a part is from supplier 2. Currently, 65% of the parts purchased by the company are from supplier 1 and the remaining 35% are from supplier 2.
Thus $P(A_1) = 0.65$ and $P(A_2) = 0.35$

Further let G denote the event that a part is good and B denote the event that a part is bad. We have following historic data:

Supplier	Percentage Good Parts	Percentage Bad Parts
Supplier 1	98%	2%
Supplier 2	95%	5%

The data from the table can be written as:
For Supplier 1, we have $P(G|A_1) = 0.98$ and $P(B|A_1) = 0.02$
For Supplier 2, we have $P(G|A_2) = 0.95$ and $P(B|A_2) = 0.05$



Look at the tree diagram shown in the figure. We see that four experimental outcomes are possible; two correspond to the part being good and two correspond to the part being bad.

Each of the experimental outcomes is the intersection of two events, so we can use the multiplication rule to compute the probabilities. From left to right through the tree, the probabilities for each branch at step 1 are prior probabilities and the probabilities for each branch at step 2 are conditional probabilities. To find the probabilities of each experimental outcome, we simply multiply the probabilities on the branches leading to the outcome.

In this tree diagram $P(A_1 \cap G)$ has been calculated using the definition, we learnt in, Multiplicative Law of Probability, i.e. $P(A_1 \cap G) = P(A_1) \times P(G|A_1)$

Now, Suppose we get a Bad Part and we want to find out the probability that the given bad part is from supplier 1. In other words, we want to find out $P(A_1|B)$.

As per Bayes' Theorem, $P(A_1|B)$ is given by following equation:

$$P\left(\frac{A_1}{B}\right) = \frac{P(A_1) \times P\left(\frac{B}{A_1}\right)}{P(A_1) \times P\left(\frac{B}{A_1}\right) + P(A_2) \times P\left(\frac{B}{A_2}\right)}$$

This formula can be understood from the tree diagram.

Conditional Probability that given bad part is from supplier 1

$= \frac{Probability \ of \ Bad \ part \ from \ Supplier 1}{Probability \ of \ Bad \ part \ from \ Supplier 1 + Probability \ of \ Bad \ part \ from \ Supplier 2}$

To find our answer, we put up values from the given tree:

$$P\left(\frac{A_1}{B}\right) = P\left(\frac{\text{SUPPLIER 1}}{\text{BAD PART}}\right)$$

$$= \frac{P(A_1) \times P\left(\frac{B}{A_1}\right)}{P(A_1) \times P\left(\frac{B}{A_1}\right) + P(A_2) \times P\left(\frac{B}{A_2}\right)} = \frac{0.65 \times 0.02}{0.65 \times 0.02 + 0.35 \times 0.05}$$

$$= 0.4262$$

Thus, if we find a bad part, the probability that it came from Supplier 1, is 42.62%.

Similarly, we can find Probability that this bad part came from Supplier 2, using following formula:

$$P\left(\frac{A_2}{B}\right) = \frac{P(A_2) \times P\left(\frac{B}{A_2}\right)}{P(A_1) \times P\left(\frac{B}{A_1}\right) + P(A_2) \times P\left(\frac{B}{A_2}\right)}$$

Thus, the **Bayes' theorem** can be written as:

$P(A_1|B) = \frac{P(A_1) \times P(B|A_1)}{P(A_1) \times P(B|A_1)+P(A_2) \times P(B|A_2)}$

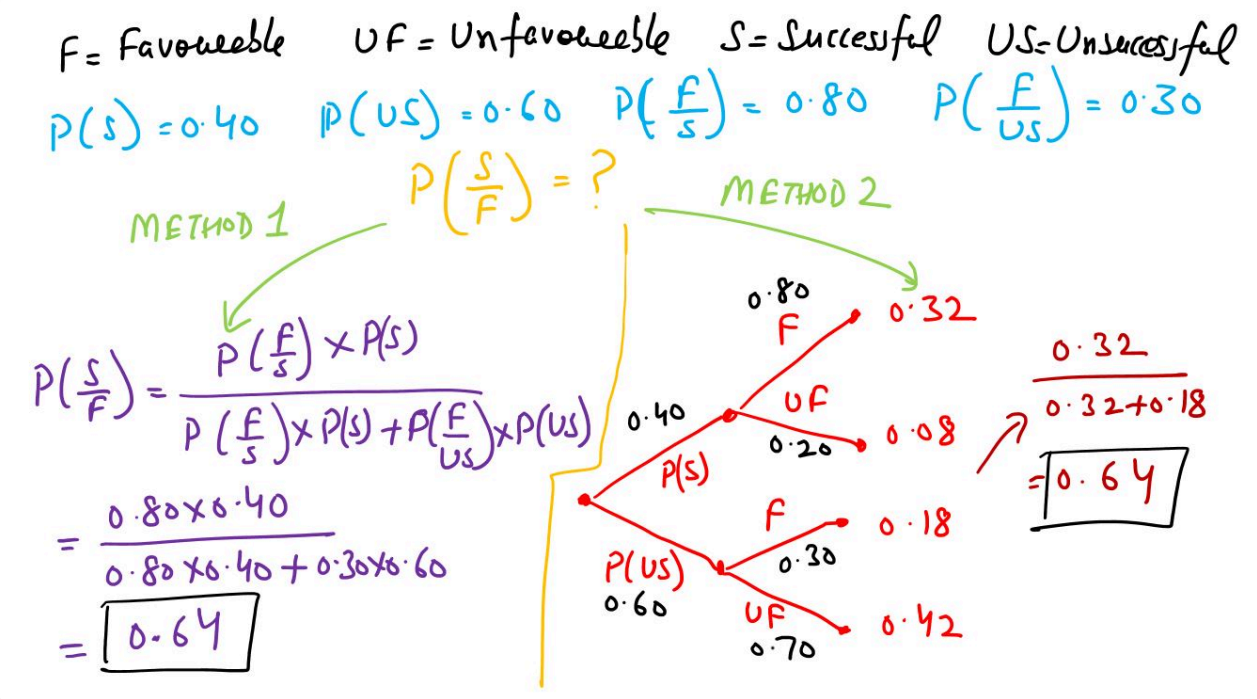
$P(A_2|B) = \frac{P(A_2) \times P(B|A_2)}{P(A_1) \times P(B|A_1)+P(A_2) \times P(B|A_2)}$

---

11. Bayes' Theorem

In the past, 40% of the movies by YRFs have been successful, and 60% have been unsuccessful. Before releasing a new movie, the marketing research department conducts an extensive study and releases a report, either favorable or unfavorable. In the past, 80% of the successful movies had received favorable market research reports, and 30% of the unsuccessful movies had received favorable reports. For a new movie under consideration, the marketing research department has issued a favorable report. What is the probability that the given movie will be successful?

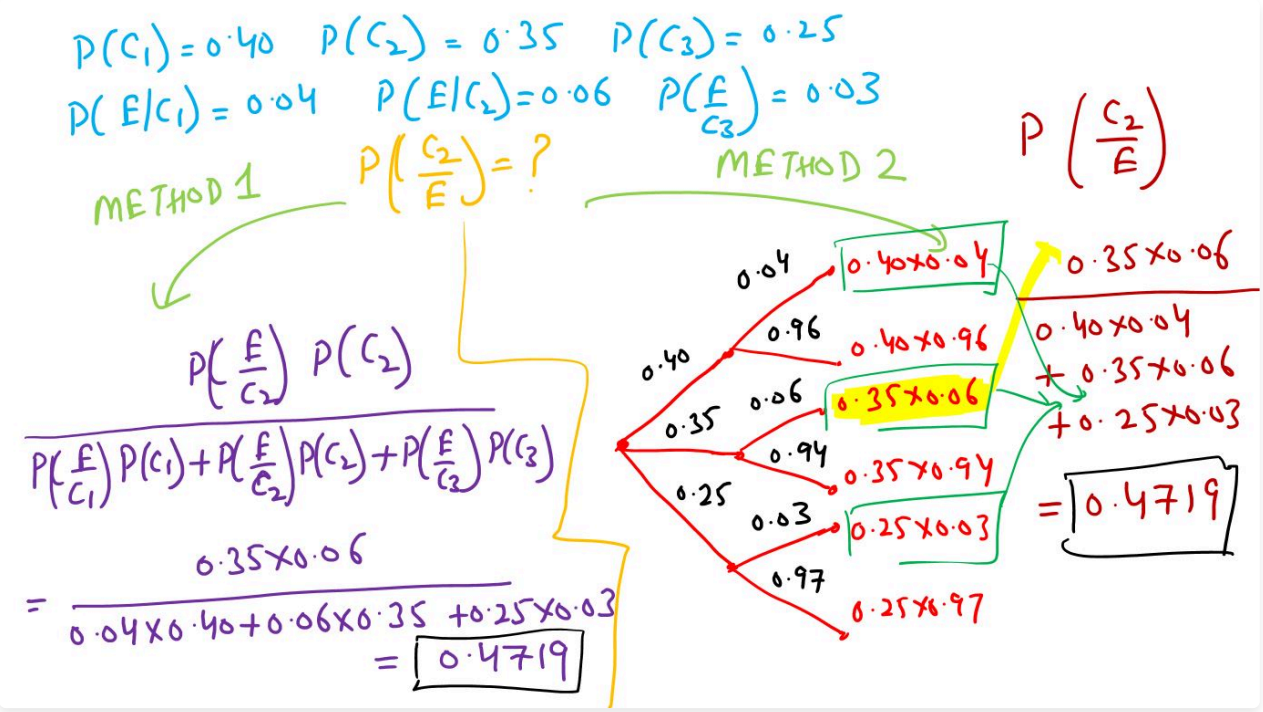
Solution:



11. Bayes' Theorem

In a Post Office, three clerks were assigned to process incoming mail. The first clerk, C1, processes 40%, the second clerk, C2, processes 35% and the third clerk, C3, processes 25% of the mail. The first clerk has an error rate of 0·04, the second clerk has an error rate of 0·06 and the third clerk has an error rate of 0·03. A mail selected at random from a day's output is found to have an error. The Postmaster wishes to know the probability that the mail was processed by the second clerk? (*UPSC Management Optional 2013 Paper*)

Solution:



11. Bayes' Theorem

A factory has two machines, A and B. Machine-A produces 60% of items and Machine-B produces 40% of the items of the total output. Further 2% of the items produced by Machine-A are defective whereas 4% produced by Machine-B are defective. If an item is drawn at random what is the probability that it is defective?

Solution:

Let A_1, be the event that the items are produced by Machine-A,

A_2 be the event that the items are produced by Machine-B.

Let B be the event of drawing a defective item.

Given $P(A_1) = \frac{60}{100}$

$P(B|A_1) = \frac{2}{100}$

$P(A_2) = \frac{40}{100}$

$P(B|A_2) = \frac{4}{100}$

We have to find the total probability of event B. Since A_1 and A_2 are mutually exclusive and exhaustive events, we have,

$P(B) = P(A_1) \times P(B|A_1) + P(A_2) \times P(B|A_2)$

$P(B) = \frac{60}{100} \times \frac{2}{100} + \frac{40}{100} \times \frac{4}{100} = 0.028$

Thus, there is 2.8% probability.

11. Bayes' Theorem

An advertising executive is studying television viewing habits of married men and women during prime time hours. Based on the past viewing records he has determined that during prime time, wives are watching television 60% of the time. It has also been determined that when the wife is watching television, 40% of the time the husband is also watching. When the wife is not watching the television, 30% of the time the husband is watching the television. Find the probability that
(i) the husband is watching the television during the prime time of television
(ii) if the husband is watching the television, the wife is not watching the television.

Solution:
Let the events be defined as follows
A_1 = Event of wife watching the television
A_2 = Wife not watching the television
B = Husband is watching the television
Given $P(A_1) = 0.60$

$P(B|A_1) = 0.40$

$P(A_2) = 1 - 0.60 = 0.40$

$P(B|A_2) = 0.30$

(i) P (Husband watching the television)
$P(B) = P(A_1) \times P(B|A_1) + P(A_2) \times P(B|A_2)$
$P(B) = (0.40) \times (0.60)+(0.30) \times (0.40)$
$P(B) = 0.24 + 0.12$
$P(B) = 0.36$
$P(B) = \frac{36}{100} = \frac{9}{25}$

Thus, the probability that the husband is watching the television during the prime time is $\frac{9}{25}$

(ii) P (if the husband is watching, the wife is not watching the television)
$P(A_2|B) = \frac{P(A_2) \times P(B|A_2)}{P(A_1) \times P(B|A_1)+P(A_2) \times P(B|A_2)}$

$= \frac{0.40 \times0.30 }{ \frac{9}{25} } = \frac{1}{3}$

Thus, the probability that if the husband is watching the television, the wife is not watching the television is $\frac{1}{3}$

---

12. Unconditional Probability

Unconditional Probability, also known as **marginal probability**, refers to the probability of an event occurring without considering any other events or conditions. It focuses solely on the likelihood of a specific outcome or event happening from a set of possible outcomes.

This type of probability is calculated independently, irrespective of any prior or related results. It doesn't take into account the occurrence or non-occurrence of any other events. It measures the chance of a single event or outcome happening on its own, without considering other events in the scenario.

Mathematically, to determine the unconditional probability of an event, you divide the number of outcomes where the event occurs by the total number of possible outcomes. This method involves counting the occurrences of the event of interest and dividing it by the total sample space.

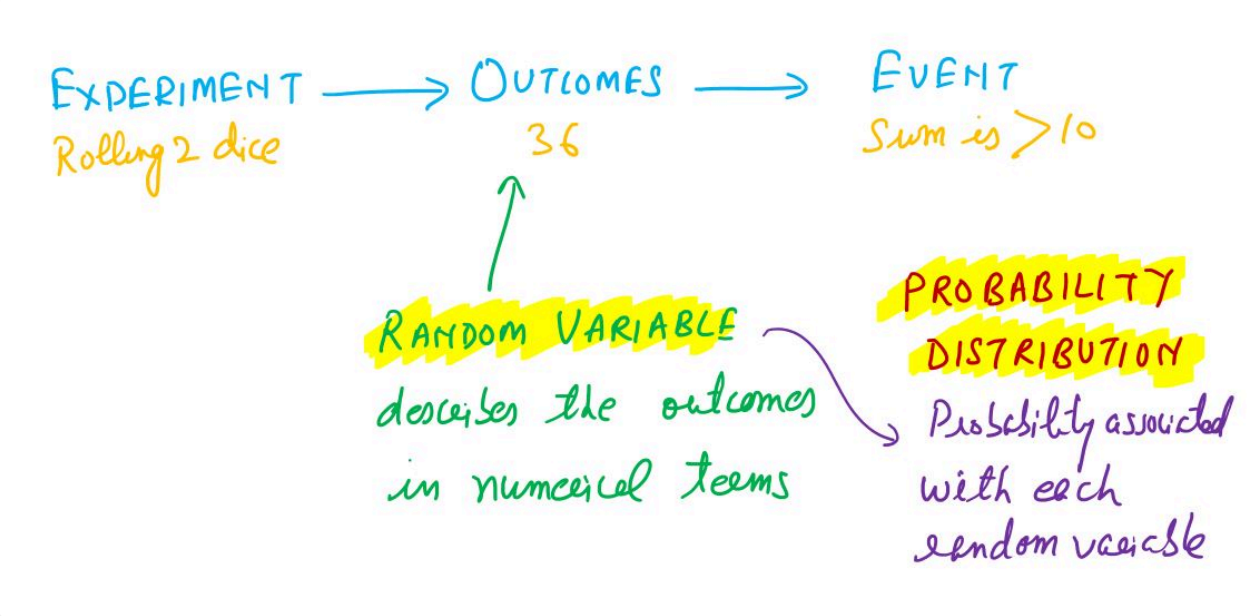
---

1. Introduction

Before we attempt to understand Probability Distribution, the concept of ‘random variable’ needs to be elaborated.

In an experiment involving the tossing of three coins, the recorded outcomes were quantified as the 'number of heads'. Denoted as 'H', this variable encompasses values of 0, 1, 2, or 3, each associated with a specific probability.

This uncertain real variable 'H' adopts varied numerical values contingent upon the experiment's outcomes, with a probability assignment allocated to every potential value. It is called as a **Random Variable** due to its inherent variability and uncertainty within the experiment.



The culmination of these values and their respective probabilities constitutes what is termed the **Probability Distribution** of 'H'.

Examples

Here are three diverse examples illustrating the concept of a random variable and its associated probability distribution:

Rolling a Six-Sided Die Twice

Consider an experiment where you roll a six-sided die twice and note the sum of the numbers rolled. Let 'X' represent this sum. The values 'X' can take range from 2 (if both rolls are 1) to 12 (if both rolls are 6). Each possible value of 'X' (2, 3, 4, ..., 12) has an associated probability determined by the combinations of outcomes (e.g., probability of getting a 7 is higher than getting a 2).

Measuring Rainfall in Inches

Imagine a scenario where you're measuring the rainfall in inches in a particular area over a day. Let 'Y' be the variable representing the amount of rainfall. 'Y' can take on any value from 0 to a certain maximum (say 5 inches). The probability distribution for 'Y' would indicate the likelihood of different amounts of rainfall occurring (e.g., probability of getting 1 inch of rain might be higher than getting 4 inches).

Testing Manufacturing Defects

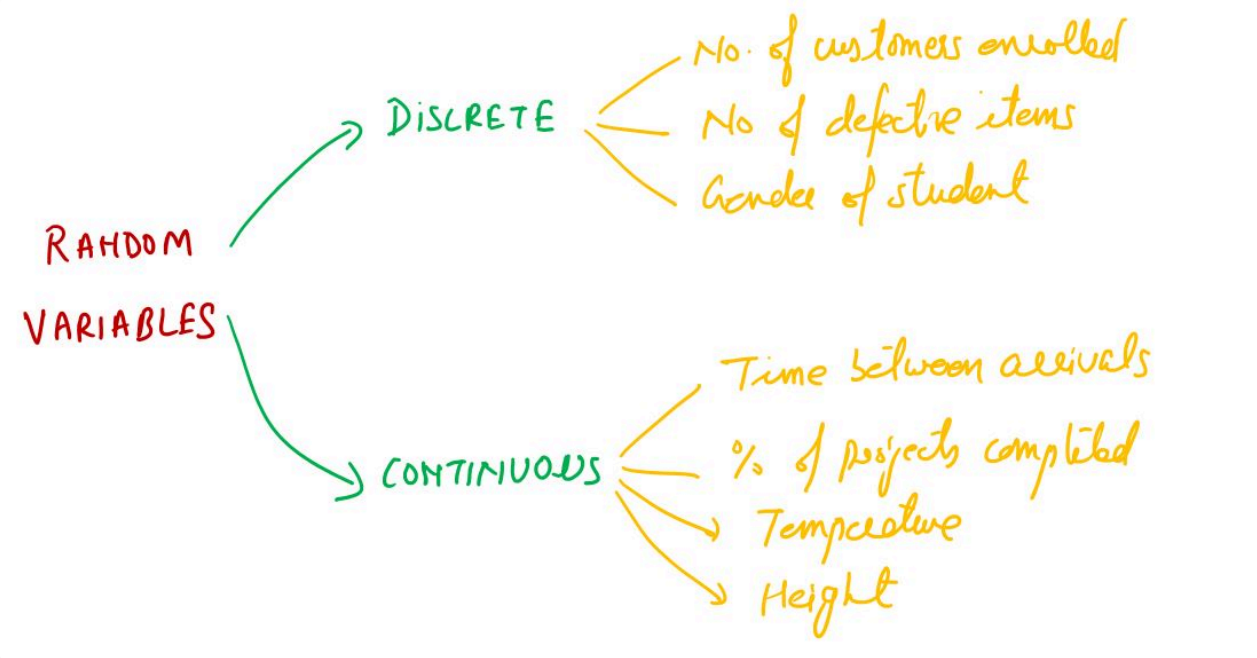
In a manufacturing process, suppose 'Z' represents the number of defects found in a batch of products. 'Z' can vary from 0 (no defects) to some upper limit (let's say 10 defects). The probability distribution for 'Z' would describe the probabilities associated with different defect counts (e.g., probability of finding no defects might be higher than finding 5 defects).

Each of these examples demonstrates a different scenario where a random variable is used to represent outcomes that can vary due to chance or randomness, and the associated probability distribution helps understand the likelihood of these outcomes occurring.

---

2. Discrete and Continuous variables

Random variables can be categorized into two main types: discrete and continuous.



Discrete Random Variables

A discrete random variable is one that can take on a countable number of distinct and separate values. These values are often integers or whole numbers and have gaps between them. For example, when rolling a die, the number showing on the die is a discrete random variable because it can only be whole numbers from 1 to 6. Another example is the number of customers entering a store in a given hour; it can be 0, 1, 2, and so on.

Continuous Random Variables

On the other hand, continuous random variables are those that can take on any value within a certain range or interval. They can take on an uncountably infinite number of possible values and can include any real number within a specified range. Examples of continuous random variables include measurements such as height, weight, temperature, time, and distances. For instance, the time it takes for a certain event to occur (like the arrival of a bus) can be measured in fractions of seconds, forming a continuous random variable.

Experiment	Random Variable (x)	Possible Values for Random Variable
Operate a toll booth	Time between arrivals of two cars in minutes	$x \geq 0$
Fill a soft drink can (max = 250 ml)	Number of ml	$0 \leq x \leq 250$
Construct a new building	Percentage of project complete after six months	$0 \leq x \leq 100$
Test a new covid drug in lab	Temperature when the desired reaction takes place (min 150°F; max 212°F)	$150 \leq x \leq 212$

Difference

The key difference lies in the nature of the values they can take. Discrete random variables have specific, separate, and countable values, while continuous random variables have an infinite number of values within a range, often represented as intervals.

For discrete random variables, we have a **Discrete Probability Distribution**, which assigns probabilities to each possible value of the variable.

Continuous random variables are associated with **Continuous Probability Distributions**, which utilize probability density functions to represent the likelihood of the variable falling within certain intervals.

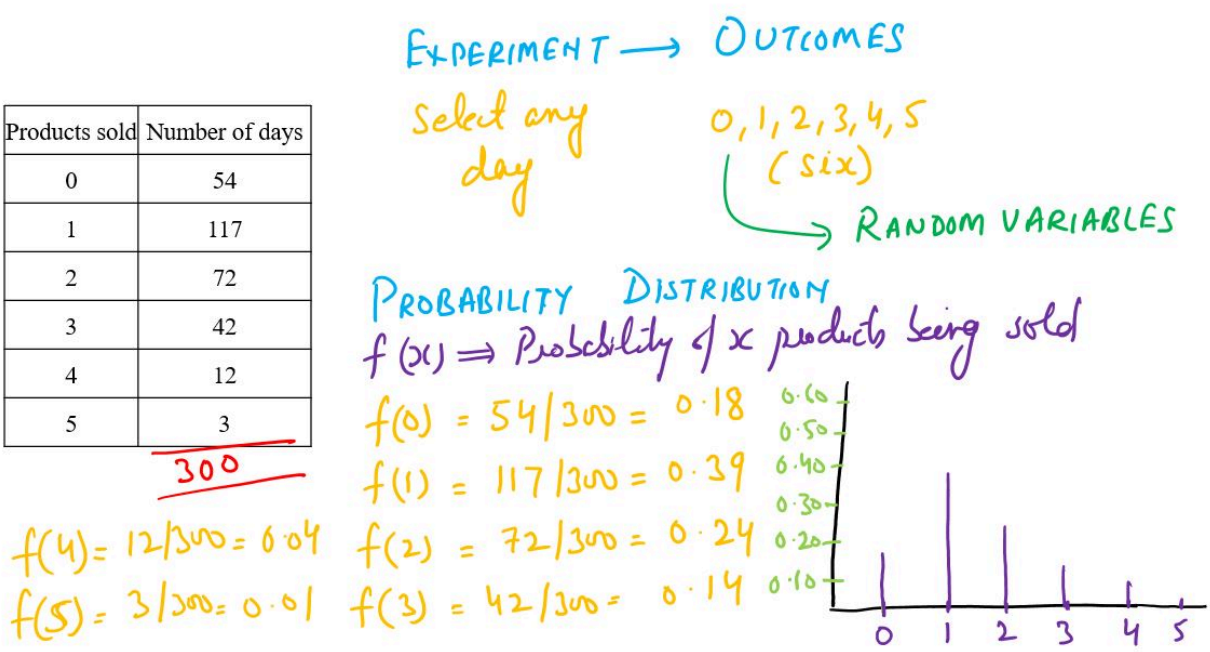
---

3. Probability Distribution Function

Probability distribution refers to the way probabilities are assigned or distributed across every possible value of a random variable, encompassing all potential outcomes of an experiment. The probability distribution function represents this allocation of probabilities to different values of the random variable.

To illustrate this concept, let's examine a dataset concerning the sales performance of a specific SAP product over 300 days by a sales team.

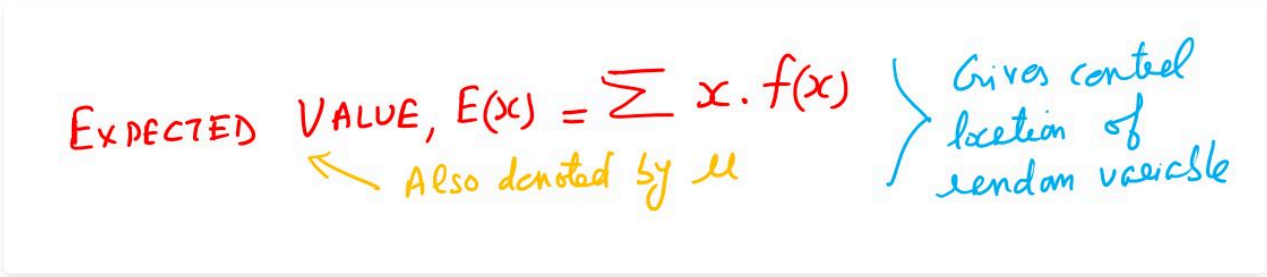
Products sold	Number of days
0	54
1	117
2	72
3	42
4	12
5	3



---

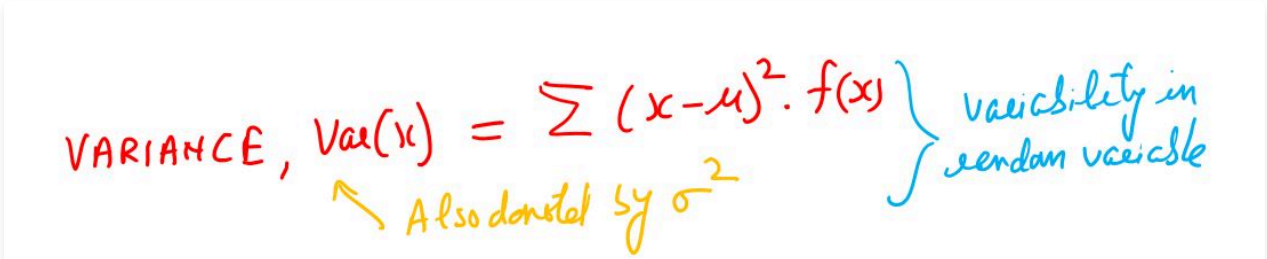
4. Mean and Variance of Random Variable

The **expected value**, or mean, of a random variable is a measure of the central location for the random variable. The formula for the expected value of a discrete random variable x follows.



Both the notations E(x) and μ are used to denote the expected value of a random variable.

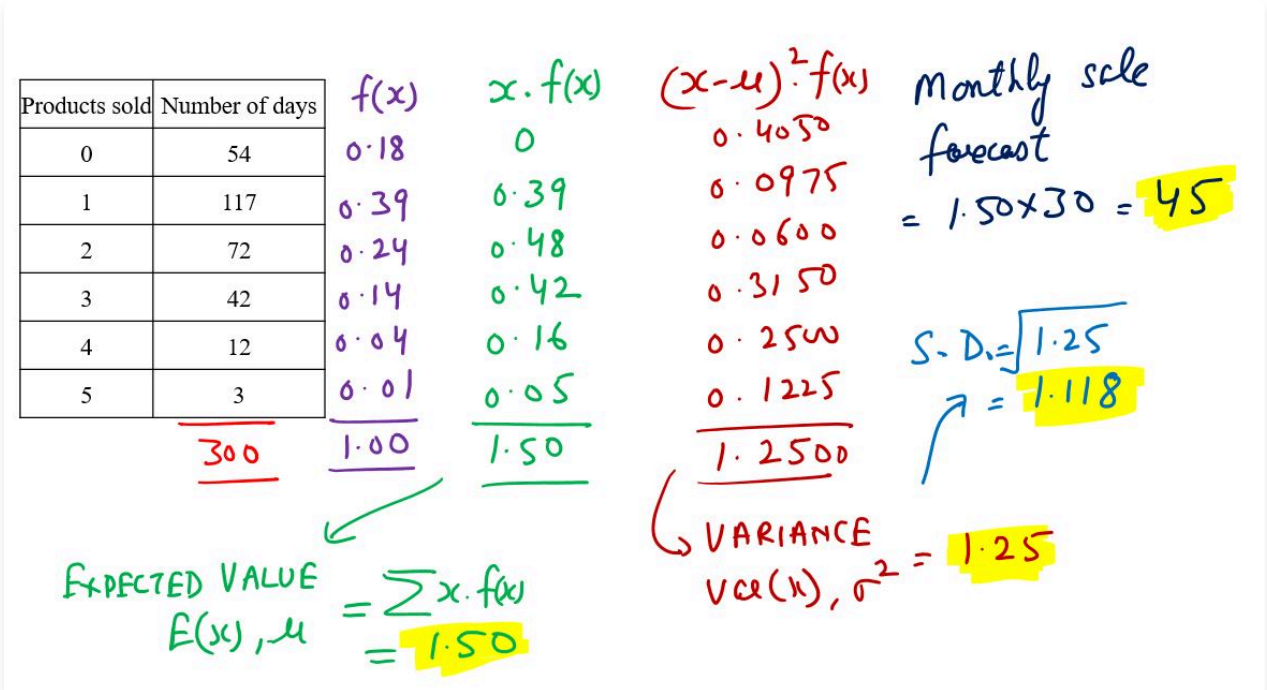
Even though the expected value provides the mean value for the random variable, we often need a measure of variability, or dispersion. We use **variance** to summarize the variability in the values of a random variable. The formula for the variance of a discrete random variable follows:



The notations Var(x) and \(\sigma^2 \) are both used to denote the variance of a random variable.

Illustration

Calculate the expected value and variance for the number of products sold per day by the sales team at SAP, considering data for 300 days? What is average monthly forecast? Also calculate standard deviation?



5. Binomial Distribution

Binomial Distribution is **Discrete probability distribution**.

Any uncertain situation or experiment that is marked by the following 4 properties is known as binomial probability distribution:

1. The experiment consists of N repeated trials.

2. Each trial can result in just two possible outcomes. We call one of these outcomes a success and the other, a failure.

3. The probability of success, denoted by p, is the same on every trial (called stationary assumption).

4. The trials are independent; that is, the outcome on one trial does not affect the outcome on other trials.

If properties 2, 3, and 4 are present, we say the trials are generated by a **Bernoulli process.** If, in addition, property 1 is present, we say we have a **Binomial experiment**.

The probability of failure is denoted by q, which is equal to 1-p.

Typical examples of Bernoulli processes are coin-tossing and success-failure situations. Similar dichotomy is preserved in testing of different pieces of a product. Each piece when tested may be defective (a failure) or non-defective (a success). We know that the production process is such that the probability of a non-defective in any trial is ‘p’ and that of a defective = q = (1 - p)

Example

Let us consider an example of testing electronic components in a production line where the components either pass or fail the test.

1. N Repeated Trials
Suppose there are 50 electronic components being tested in a production run. Each component tested represents a trial. So, in this case, N = 50 trials.

2. Two Possible Outcomes - Success and Failure
For this scenario, let's consider a "pass" as success and a "fail" as failure. Each electronic component tested can either pass the quality test (success) or fail the test (failure).

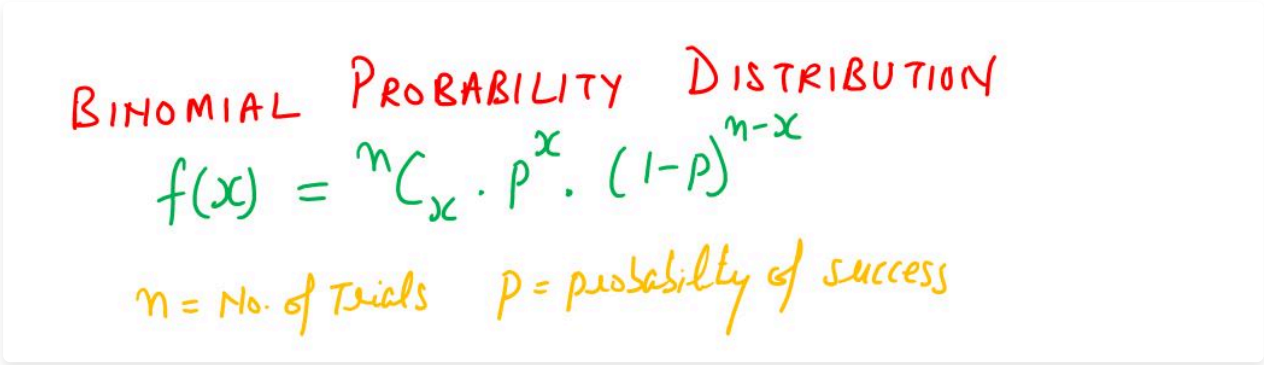
3. Constant Probability of Success (p)
Assume the quality control process ensures a consistent quality standard, so the probability of any component passing the test remains the same at 0.85 or 85%. Therefore, the probability of success (a component passing the test) denoted by p = 0.85 remains constant across all 50 trials.

4. Independence of Trials
The testing of one electronic component's success or failure does not influence the testing outcome of any other component. Each component is tested independently of the others. Whether one component passes or fails the test does not impact the testing of the subsequent components.

---

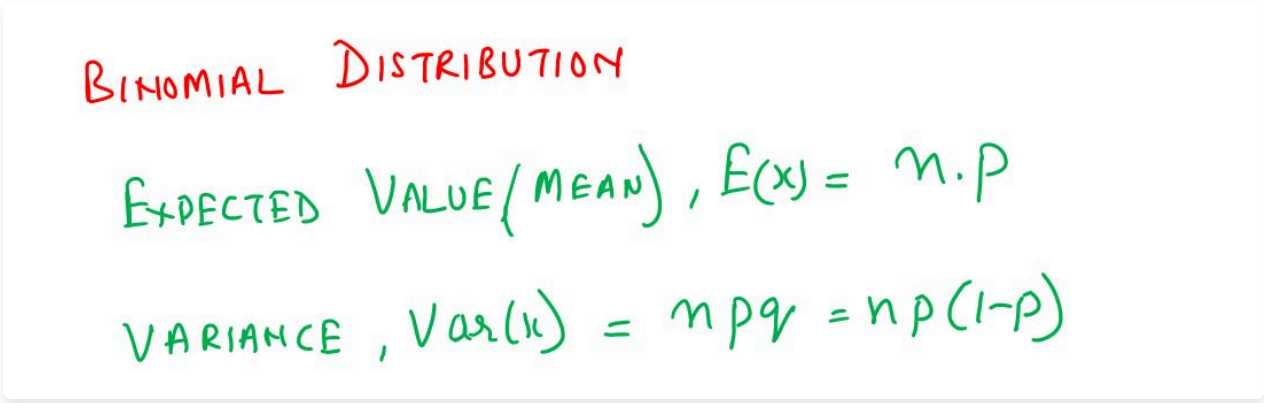
5. Binomial Distribution

The Binomial Distribution is given by following function.



Expected value or Mean of Binomial Distribution, E(x) = μ = n.p

Variance of Binomial Distribution, Var(x) = σ^2 = n.p.q (which may also be written as n.p.(1-p).



Thus in case of Binomial Distribution, Mean is 'np' and Variance is 'npq'.

It may be noted that:

- When p is small, the binomial distribution is skewed to the right.
- When p=0.5, then binomial distribution is symmetrical.
- When p is larger than 0.5, the distribution is skewed to the left.

---

5. Binomial Distribution

Illustration 1

Suppose that a real estate agent, Sumit Chadha has 5 contacts, and he believes that for each contact the probability of making a sale is 0.40.
(i) Find the probability that he makes at most 1 sale.
(ii) Find the probability that he makes between 2 and 4 sales (inclusive).

Solution:

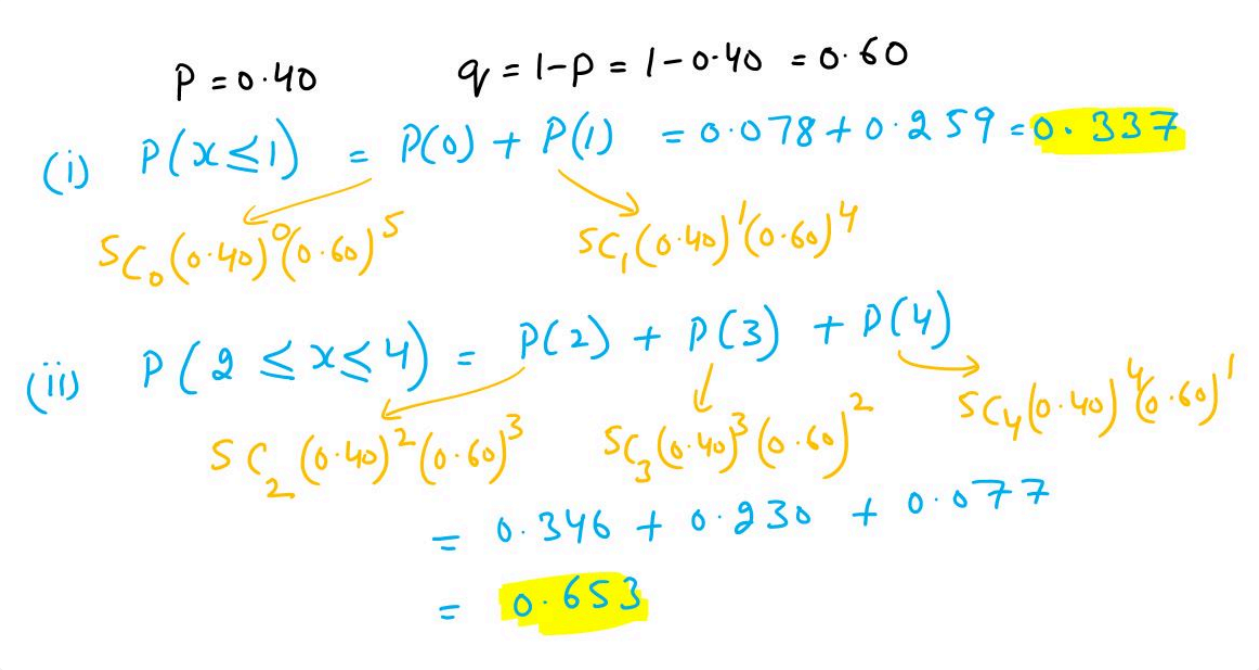
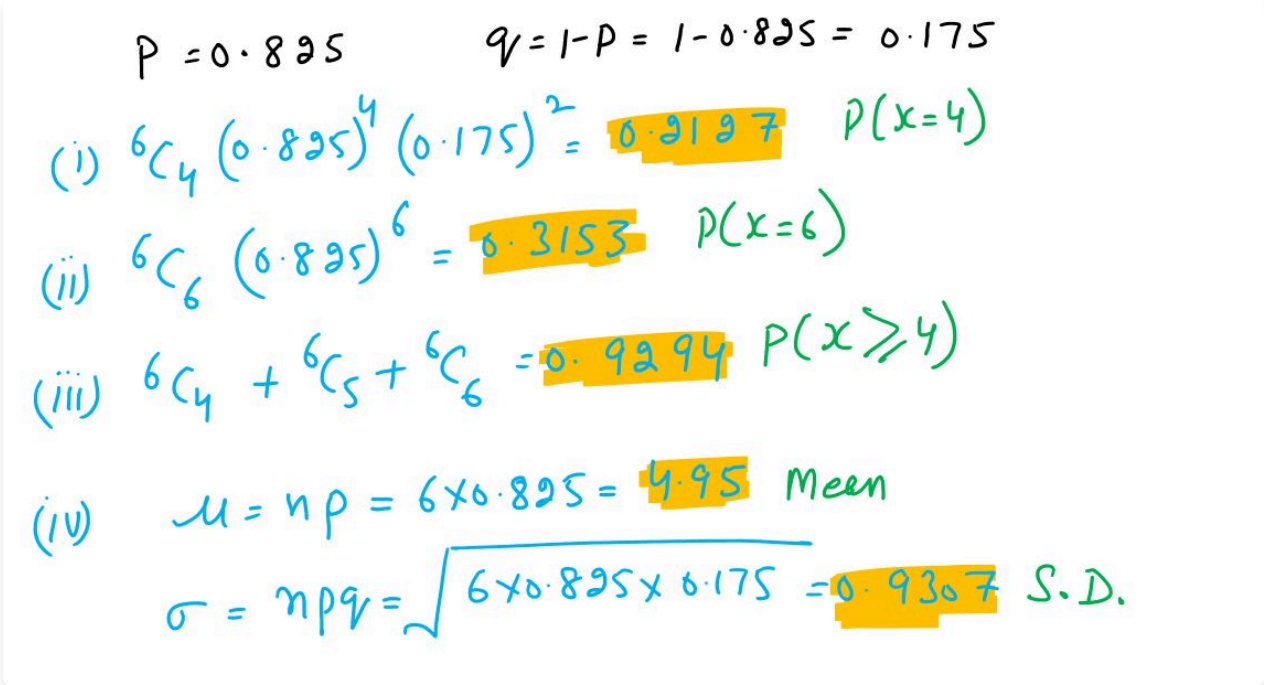


Illustration 2

Indigo airlines reports that on-time arrival rate for domestic flights is 0.825. Using the binomial distribution, what is the probability that in the next six flights:
(i) four flights will be on time?
(ii) all six flights will be on time?
(iii) at least four flights will be on time?
(iv) What are the mean and standard deviation of the number of on-time arrivals?

Solution:



---

5. Binomial Distribution

Illustration 1

Given that 51.3% of all newly born children are boys, then what is the probability that in a sample of 5 newly born children, exactly 3 are boys?

Solution:
Here we have n = 5, x = 3 and p = 0.513.

Applying the binomial formula, we get
P (3 boys) = $ {_3^5}C \times (0.513)^3 \times (1-0.513)^{5-3} $ = 0.320

Illustration 2

In a large collection of light bulbs, we assume that 98% of these bulbs will not defective. If we select 10 bulbs from this collection, then what is the probability that 8 are not defective?

Solution:
Here we have n = 10, x = 8 and p = 0.98.

Applying the binomial formula, we get
P (8 are not defective) = $ {_8^{10}}C \times (0.98)^8 \times (1-0.98)^2 $ = 0.015

Illustration 3

Of all the cars registered in Germany, 53% are German made. In a sample of 12 cars registered in Germany, what is the probability that 9 are foreign made?

Solution:
We must be careful here: the problem gives the percentage of "German made" cars, but then asks for the percentage of "foreign made" cars. If we want to go with the 9 cars foreign made we must first calculate the probability that a car is foreign made. This is 47% (calculated from 100-53).
P (9 are not foreign made) = $ {_9^{12}}C \times (0.47)^9 \times (1-0.47)^3 $ = 0.037

---

6. Poisson Distribution

The Poisson distribution serves as a model for estimating the occurrences of events within a specified time interval or area. It is associated with **discrete random variables**.

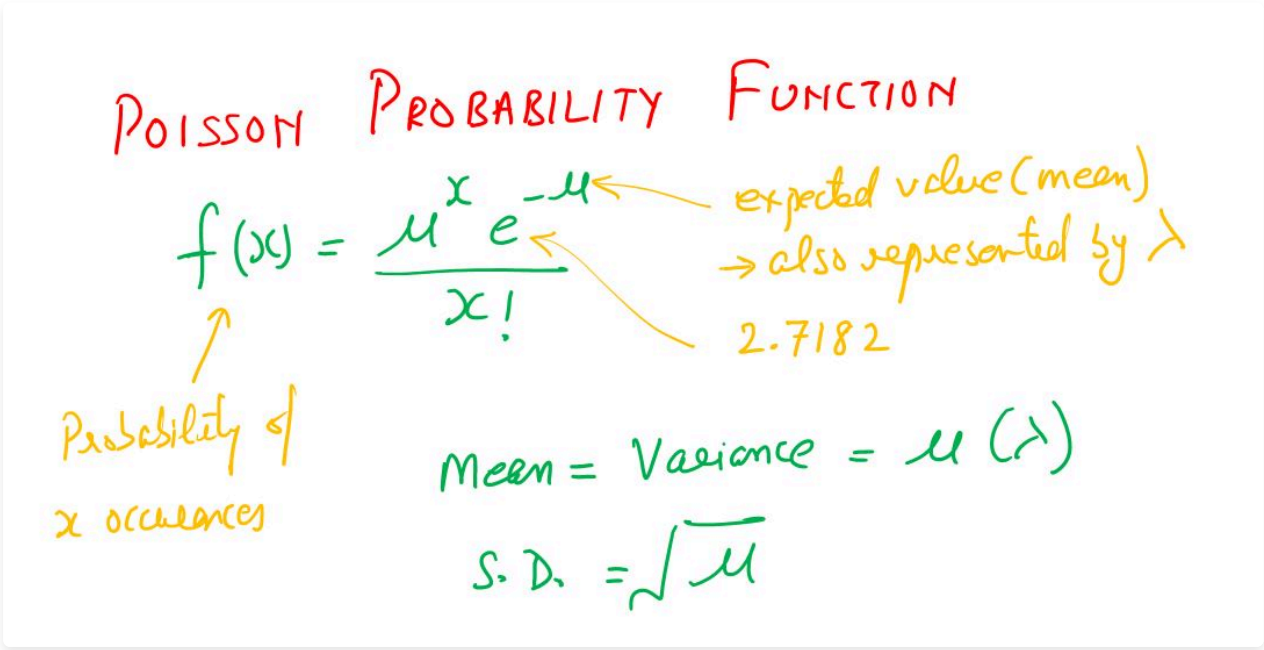
Here are few examples of Poisson distribution:
1. Number of customers arriving at an ATM during 30 minutes
2. Number of trees planted along a 1 km stretch of highway
3. Number of leaks found in a 500-meter pipeline
4. Number of times a car experiences a puncture in a year
5. Number of reported crimes in a state during a specific time period (e.g., a month, a year)

Properties of the Poisson distribution

The properties of the Poisson distribution are:
1. The probability of an occurrence is the same for any two intervals of equal length.
2. The occurrence or nonoccurrence in any interval is independent of the occurrence or nonoccurrence in any other interval.

---

6. Poisson Distribution



Equation for Poisson Distribution is given below:
$P(x) = \frac{e^{- \mu} \times \mu^x}{x!}$

Range of X = 0,1,2,3,.....∞
Expected Value or Mean, μ = λ
Variance = λ
Standard Deviation: $\sigma = \sqrt{\lambda}$

λ is the shape parameter which indicates the average number of events in the given time interval.
e: A constant equal to approximately 2.71828. (Actually, e is the base of the natural logarithm system).

The Mean and Variance of the Poisson distribution are both equal to λ.

Note: Poisson is a good approximation of the binomial distribution when n is greater than or equal to 20 (large sample size) and p is less than or equal to 0.05 (smaller values of p).

---

6. Poisson Distribution

Illustration 1

At Tata Motors, the number of work-related injuries per month in a manufacturing plant is known to follow a Poisson distribution with a mean of 2.5 work-related injuries a month. What is the probability that in a given month,
(i) no work-related injuries occur?
(ii) That at least one work-related injury occurs?

Solution:

$\lambda = \mu = 2.5$

(i) $P(x=0) = \frac{\mu^x e^{-\mu}}{x!} = \frac{2.5^0 e^{-2.5}}{0!} = 0.0821$

(ii) $P(x \geq 1) = 1 - P(0)$
$= 1 - 0.0821 = 0.9179$

Illustration 2

Customers arrive at a photocopying machine at an average rate of 2 every five minutes. Assume that these arrivals are independent, with a constant arrival rate, and that this problem follows a Poisson model, with X denoting the number of arriving customers in a 5-minute period and mean of 2. Find the probability that more than two customers arrive in a 5-minute period.

Solution:

$\lambda = \mu = 2$

$P(x > 2) = 1 - (P(0) + P(1) + P(2))$

$\frac{2^0 e^{-2}}{0!} \quad \frac{2^1 e^{-2}}{1!} \quad \frac{2^2 e^{-2}}{2!}$

$= 1 - (0.1353 + 0.2706 + 0.2706)$

$= 0.3233$

6. Poisson Distribution

Illustration 1

The average number of homes sold by the ABC Company is 2 homes per day. What is the probability that exactly 3 homes will be sold tomorrow?

Solution:
This is a Poisson experiment in which we know the following:
λ= μ = 2; since 2 homes are sold per day, on average.
x = 3; since we want to find the likelihood that 3 homes will be sold tomorrow.
e = 2.71828; since e is a constant equal to approximately 2.71828.
We plug these values into the Poisson formula as follows:
P (x; λ) = $\frac{e^{- \mu} \times \mu^x}{x!}$
P (3; 2) = $\frac{{2.71828}^{-2} \times 2^3}{3!}$
P (3; 2) = 0.18
Thus, the probability of selling 3 homes tomorrow is 0.18.

Illustration 2

Suppose the average number of calls to a helpline number in one minute is 2. What is the probability of 10 calls in 5 minutes?

Solution:

Since the average number of calls to a helpline in one minute is 2, thus the average number of calls in 5 minutes is 10. Let X represent the number of calls in 5 minutes. Then,

P(x = i) = $f_x(i) = \frac{e^{-10} \times 10^i }{i!}$, i = 0, 1, 2, K

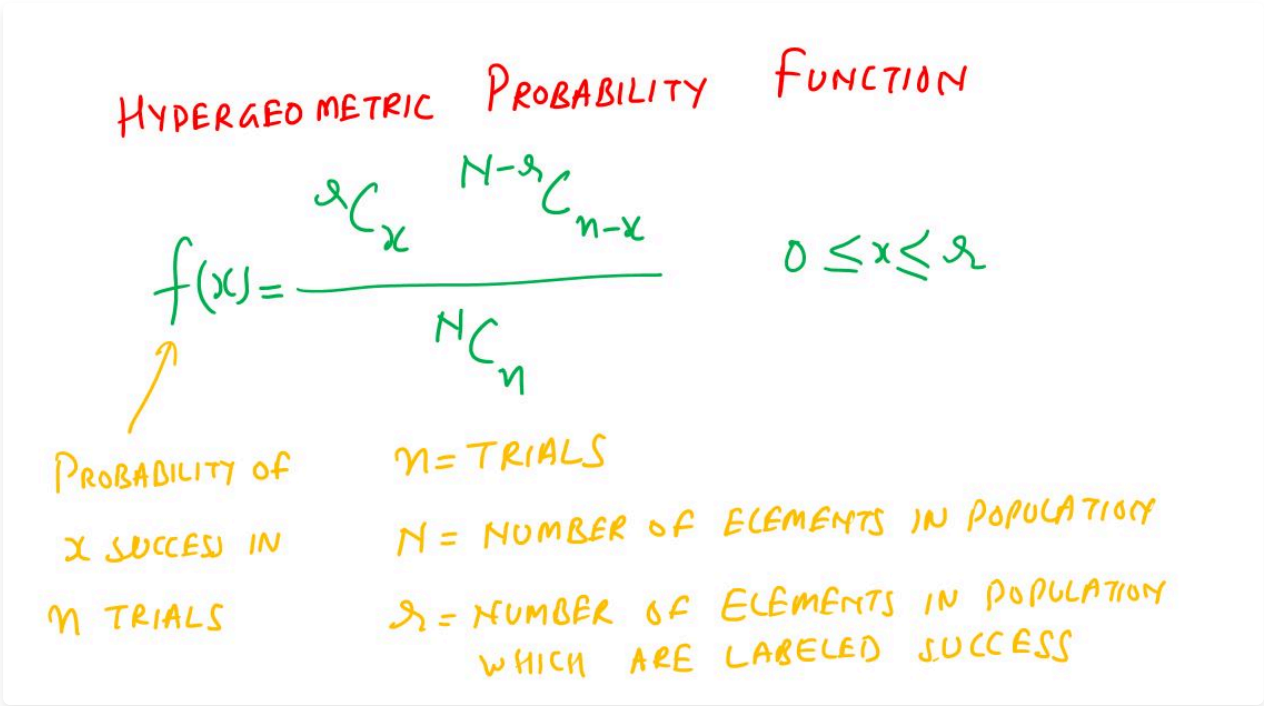
And E (X) = 10

Then, P (10 calls in 5 minutes) P(x = 10) = $f_x(10) = \frac{e^{-10} \times 10^{10} }{10!}$ = 0.1251

---

7. Hypergeometric Distribution

The hypergeometric probability function is used to calculate the probability of obtaining a certain number of successes in a fixed number of draws without replacement from a finite population.



Unlike the binomial distribution, it considers situations where the trials are not independent and where the probability of success changes from trial to trial due to the lack of replacement.

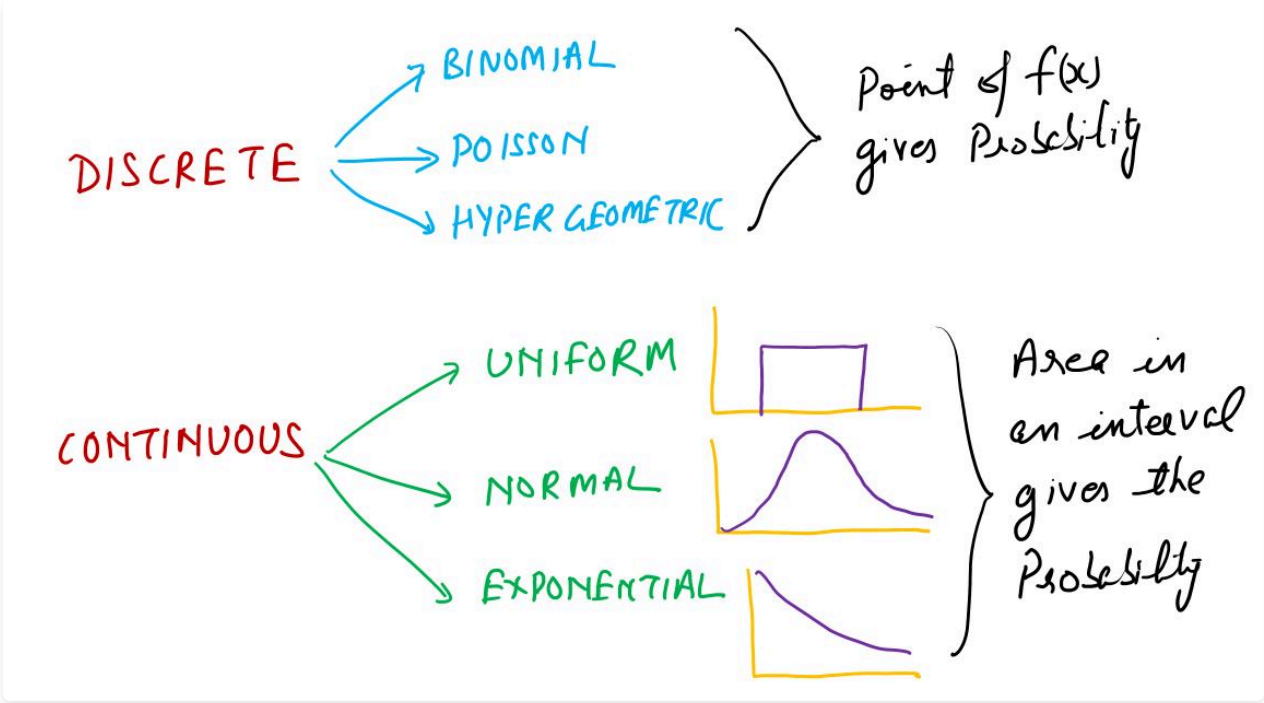
Imagine a bag containing 10 red marbles and 20 blue marbles. You want to draw 5 marbles from this bag without replacing them. The probability of success (let's define success as drawing a red marble) changes as you draw marbles, as the composition of the bag alters with each draw.

---

8. Continuous Probability Functions

So far, we've covered discrete probability distributions like Binomial, Poisson, and Hypergeometric.

Now, we'll delve into continuous probability functions, including Uniform, Normal, and Exponential distributions.



There are two significant differences when dealing with continuous random variables compared to their discrete counterparts:

1. Instead of discussing the probability of a random variable assuming a specific value, we focus on the probability of the random variable falling within a given interval.

2. The probability of a continuous random variable taking on a value within a particular interval (from x_1 to x_2) is defined as the area under the curve of its probability density function between x_1 and x_2. Since a single point has zero width, the probability of a continuous random variable landing on an exact value is zero. Additionally, the probability of a continuous random variable falling within an interval remains the same whether or not the endpoints are included.

Calculating the expected value and variance for a continuous random variable follows a similar principle to that of a discrete random variable. However, the computational process involves integral calculus.

---

9. Uniform Distribution

Uniform Distribution, also known as Rectangular Distribution, is a probability distribution where all outcomes within a certain range are equally likely to occur.

In this distribution, every possible outcome within the given range has an equal probability of occurring. For instance, if you have a uniform distribution over the range from 1 to 6, each number (1, 2, 3, 4, 5, and 6) has a probability of 1/6 of occurring in a single trial.

The probability density function of a uniform distribution is constant within the range and zero outside of it. This means that the probability of any specific value occurring is constant across the range, creating a rectangle-like shape when visualized.

The uniform distribution can be either continuous or discrete. In a continuous uniform distribution, the values within the range can take on any value within that range, while in a discrete uniform distribution, the values can only take on specific discrete values within the range.

Examples

Here are a few examples that demonstrate situations where a uniform distribution is applicable:

Rolling a Fair Die
The outcome of rolling a fair six-sided die can be considered a discrete uniform distribution, where each number (1, 2, 3, 4, 5, or 6) has an equal probability of 1/6.

Choosing a Random Time within an Hour
Suppose you randomly select any moment within an hour (from 0 to 60 minutes). Each minute within that hour has an equal chance of being chosen, demonstrating a continuous uniform distribution.

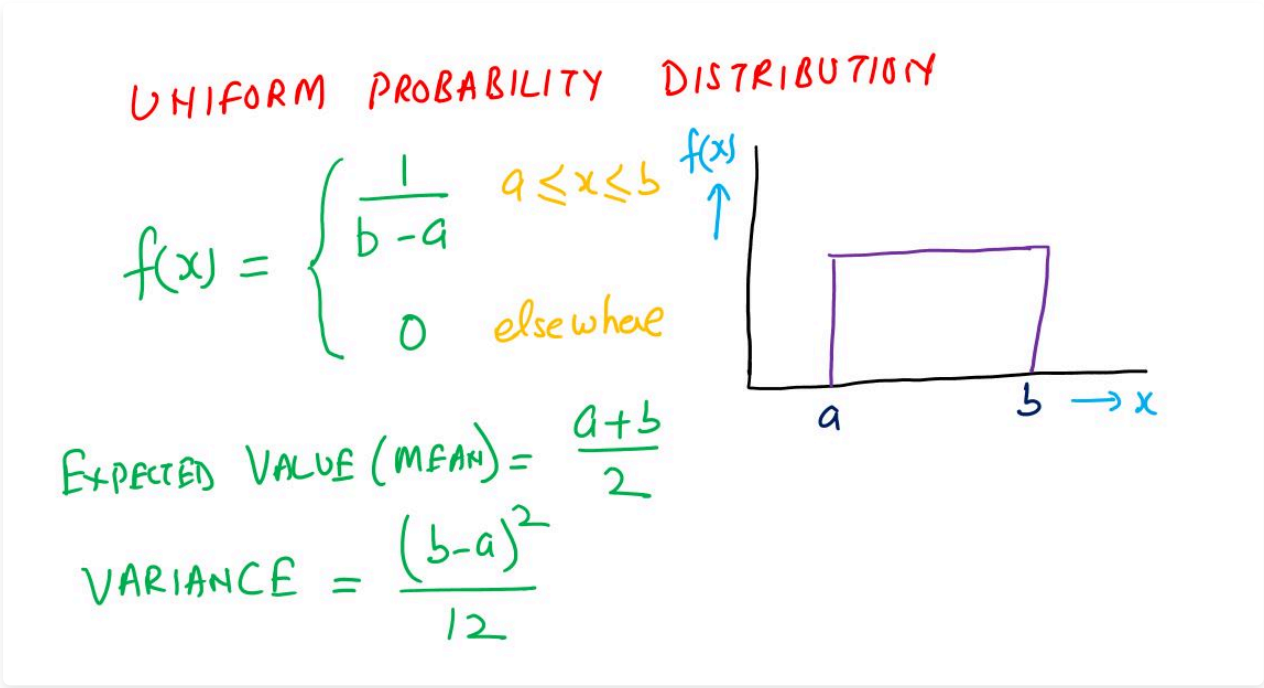
Selecting a Random Number from a Lottery Range
Consider a lottery where you can choose any number from 1 to 50. If each number has an equal chance of being drawn, this represents a discrete uniform distribution over that range.

Picking a Random Page from a Book
If you were to randomly pick a page from a 300-page book, and each page has an equal likelihood of being chosen, this situation follows a discrete uniform distribution over the pages (1 to 300).

---

9. Uniform Distribution

The uniform probability density function for a random variable x is defined by the following formula. The mean and variance are also calcualted.

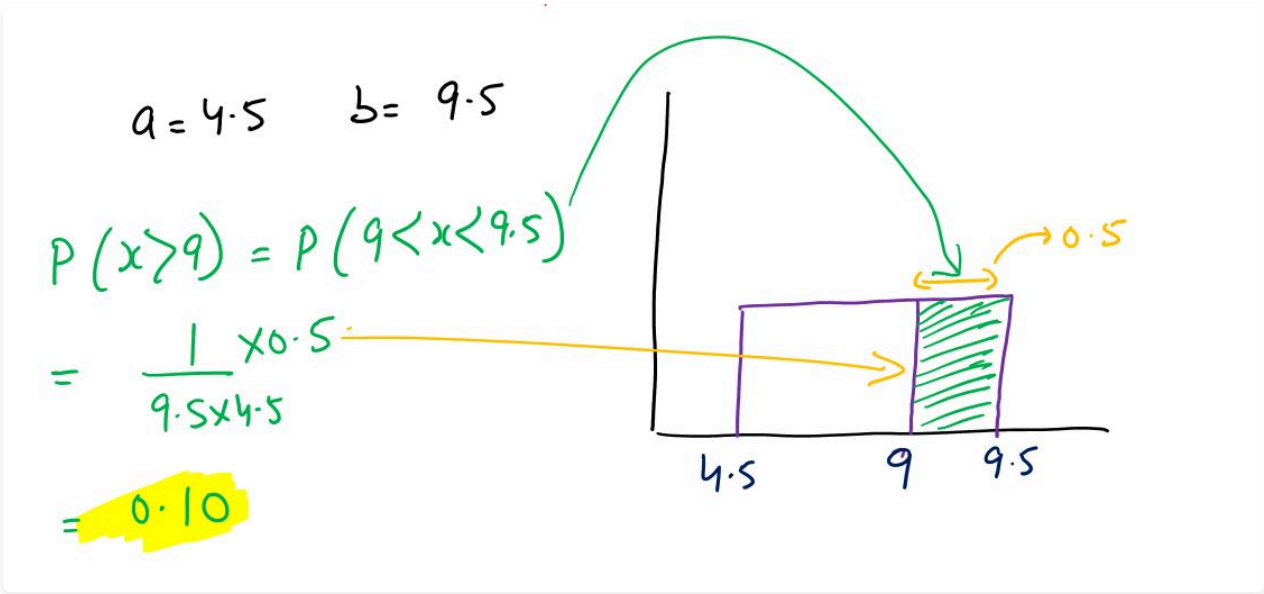


9. Uniform Distribution

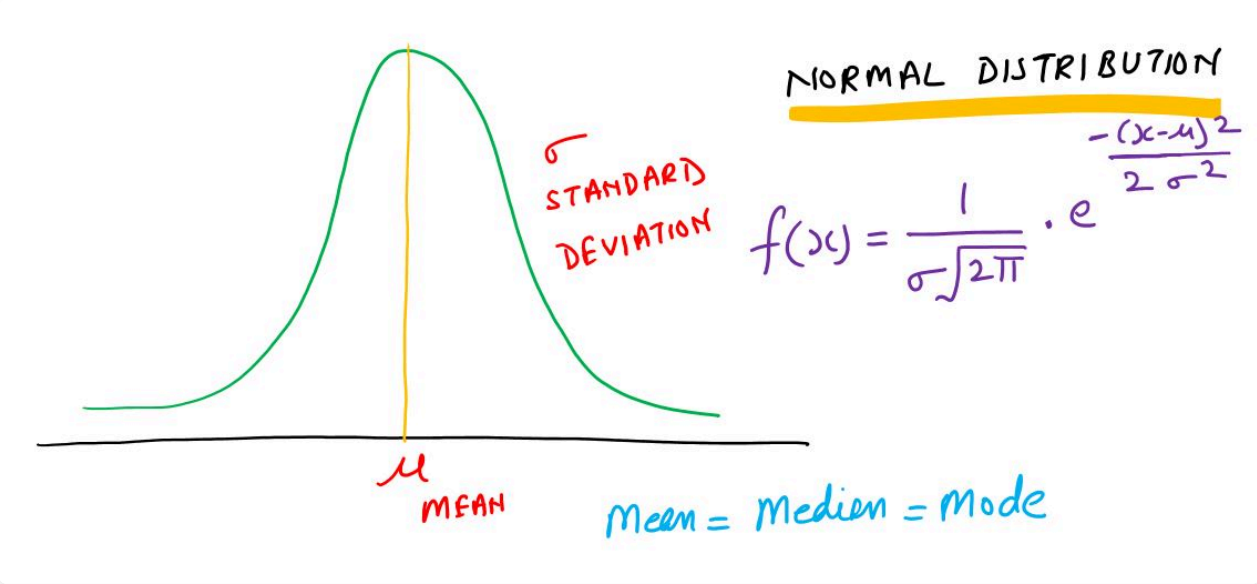
Illustration 1

The download time of Youtube videos is assumed to be uniform distribution between 4.5 and 9.5 seconds. What is the probability that a download time will take more than 9 seconds?

Solution:



10. Normal Distribution



The Normal distribution stands as one of the most prevalent probability functions in statistics. It pertains to continuous random variables, distinguishing itself from discrete distributions.

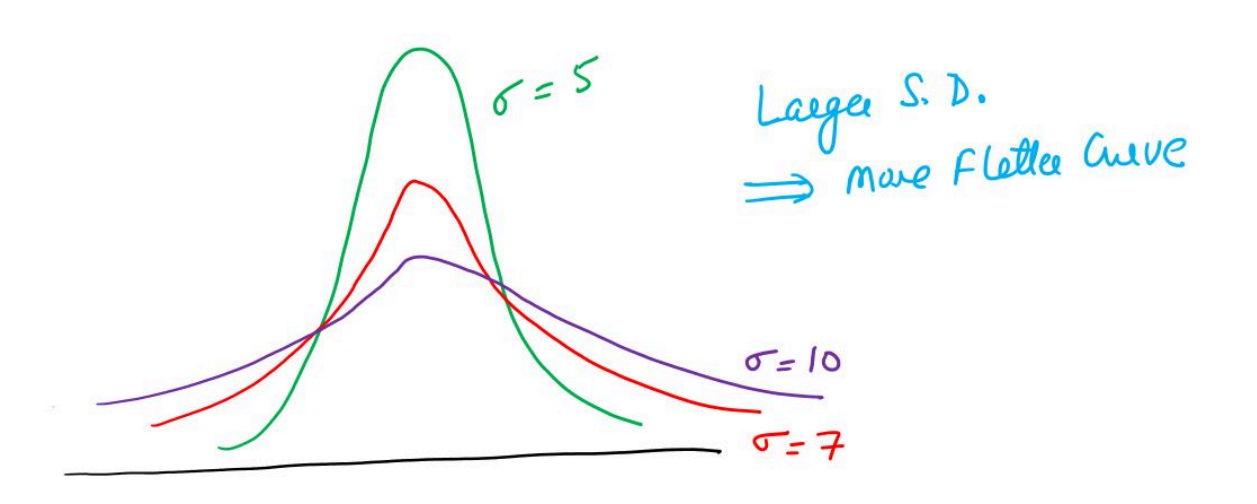
Also known as the Gaussian distribution, it assumes a bell-shaped curve, exhibiting symmetry, zero skewness, and unimodality. The graph approaches the x-axis asymptotically on both ends.

The normal probability function is given by:

\(f(x) = \frac{1}{ \sigma \sqrt{2 \pi } } \times e^ \frac{-(x- \mu)^2 }{2 \times \sigma^2 } \)

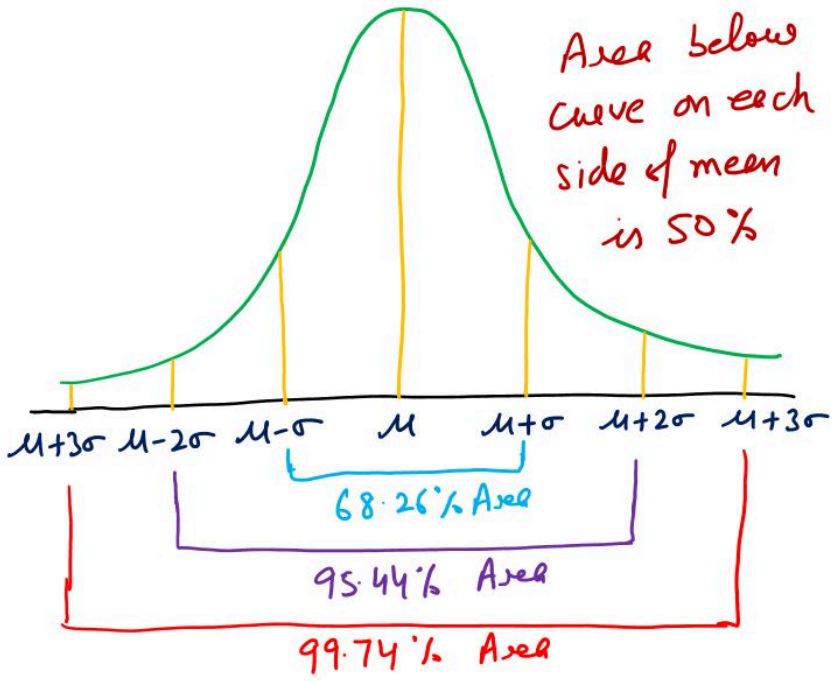
Where: \(\mu = \) mean
σ = standard deviation
π = 3.14159
e = 2.71828

This distribution is characterized by two parameters: μ (mu) representing the mean and σ (sigma) denoting the standard deviation.



A higher standard deviation results in a flatter curve, widening the spread of values across the distribution.

Probabilities for the normal random variable are given by areas under the normal curve. The total area under the curve for the normal distribution is 1. Because the distribution is symmetric, the area under the curve to the left of the mean is 0.50 and the area under the curve to the right of the mean is 0.50.



Mathematically, in a Normal Distribution:

Approximately 68.3% of all values lie within ± 1 standard deviation from mean.
Approximately 95.4% of all values lie within ± 2 standard deviation from mean.
Approximately 99.7% of all values lie within ± 3 standard deviation from mean.

Binomial Distribution approximated by Normal Distribution

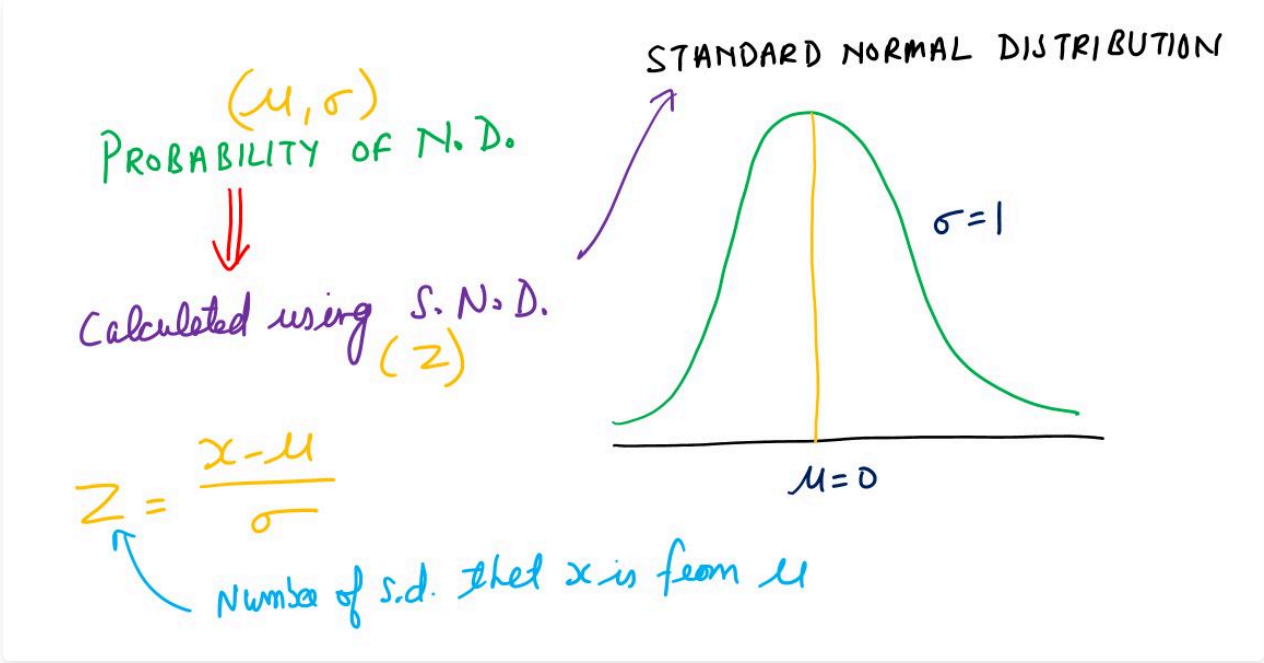
Previously, we explored how the Binomial Distribution can be approximated by the Poisson Distribution when the sample size (represented by n) exceeds or equals 20 and the probability (p) is less than or equal to 0.05.

Similarly, the Binomial Distribution can be approximated by the Normal distribution when the sample size (represented by n) becomes very large, regardless of the value of p. However, this approximation becomes more accurate when p is close to 0.50.

---

10. Normal Distribution

A random variable that has a normal distribution with a mean of zero (μ = 0) and a standard deviation of one (σ=1) is said to have a **standard normal probability distribution**. The letter z is commonly used to designate this particular normal random variable.



Thus, the formula for the standard normal probability density function, becomes:

$$f(z) = \frac{1}{ \sigma \sqrt{2 \pi } } \times e^ \frac{-(x-z)^2 }{2}$$

As with other continuous random variables, probability calculations with any normal distribution are made by computing areas under the graph of the probability density function. Thus, to find the probability that a normal random variable is within any specific interval, we must compute the area under the normal curve over that interval.

The probabilities for all normal distributions are computed by using the standard normal distribution. That is, when we have a normal distribution with any mean μ and any standard deviation σ, we answer probability questions about the distribution by first converting to the standard normal distribution. Then we can use the standard normal probability table and the appropriate z values to find the desired probabilities. The formula used to convert any normal random variable x with mean μ and standard deviation σ to the standard normal random variable z follows.

$$z = \frac{x- \mu}{ \sigma}$$

In other words, we can interpret z as the number of standard deviations that the normal random variable x is from its mean μ.
