STATISTICS

Definition:

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

Some of the real-life examples of statistics are:

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

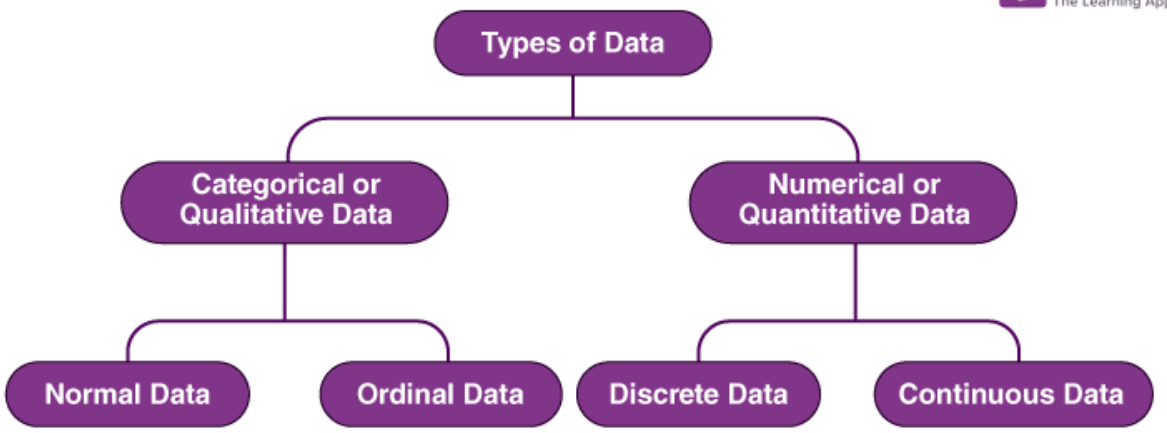
Data:

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

Types of Data:

The data is classified into majorly four categories:

- Nominal data
- Ordinal data
- Discrete data
- Continuous data



Qualitative or Categorical Data

Qualitative data, also known as the categorical data, describes the data that fits into the categories. Qualitative data are not numerical. The categorical information involves categorical variables that describe the features such as a person's gender, home town etc. Categorical measures are defined in terms of natural language specifications, but not in terms of numbers.

Sometimes categorical data can hold numerical values (quantitative value), but those values do not have a mathematical sense. Examples of the categorical data are birthdate, favourite sport, school postcode. Here, the birthdate and school postcode hold the quantitative value, but it does not give numerical meaning.

Nominal Data

Nominal data is one of the types of qualitative information which helps to label the variables without providing the numerical value. It's the simplest form of data and is commonly used for the purpose of classification or categorization. Nominal data is also called the nominal scale. It cannot be ordered and measured. But sometimes, the data

can be qualitative and quantitative. Examples of nominal data are letters, symbols, words, gender etc.

The nominal data are examined using the grouping method. In this method, the data are grouped into categories, and then the frequency or the percentage of the data can be calculated. These data are visually represented using the pie charts.

For example, consider a survey that asks respondents to select their favourite colour from a list of options (red, blue, green, yellow, etc.). The responses here represent nominal data because they categorise the respondents by preference, but don't provide any quantitative or orderable value. The colours themselves don't have a specific order and you can't perform mathematical operations on them. They are simply labels for the categories.

Ordinal Data

Ordinal data/variable is a type of data that follows a natural order. The significant feature of the nominal data is that the difference between the data values is not determined. This variable is mostly found in surveys, finance, economics, questionnaires, and so on.

For example, consider a survey that asks respondents to rate their satisfaction with a restaurant on a scale of 1 to 5, where 1 means "very dissatisfied" and 5 means "very satisfied". The responses here represent ordinal data because they categorise the respondents by satisfaction level and the categories have a clear order ($1 < 2 < 3 < 4 < 5$), but the difference between each category is not quantitatively defined (the difference in satisfaction between 1 and 2 might not be the same as between 4 and 5).

The ordinal data is commonly represented using a bar chart. These data are investigated and interpreted through many visualisation tools. The information may be expressed using tables in which each row in the table shows the distinct category.

Interval Scale Data

Interval Scale data is a type of numerical data where the difference between two values is meaningful and quantifiable, but it doesn't have a true zero point. This means that while we can interpret differences in the values of interval scale data, we cannot interpret ratios.

For example, consider temperature measured in degrees Celsius. The difference between 20°C and 30°C is the same as the difference between 30°C and 40°C - both are differences of 10 degrees. This makes it interval scale data.

However, 0°C does not mean the absence of temperature (which would be absolute zero at -273.15°C), so it's not a true zero point. Because of this, it wouldn't make sense to say that 20°C is twice as hot as 10°C, even though 20 is numerically double of 10. This lack of a true zero point is what differentiates interval scale data from ratio scale data, which does have a true zero point.

Ratio Scale Data

Ratio data is a type of numerical data that not only allows us to identify the order of the data points, but also to make meaningful statements about the ratios of measurements. Ratio data has a clear definition of zero. When the data equals zero, there is none of that variable.

An example of ratio data is the measurement of weight. If an object weighs 10 kilograms, it's twice as heavy as an object that weighs 5 kilograms. This is because weight has a true zero point. If an object weighs 0 kilograms, it truly means there is no weight. This ability to interpret both differences and ratios makes weight an example of ratio data.

Data Collection in Statistics

In statistics, data is usually collected in two ways: as a **population** or as a **sample**:

1. **Population Data**: This refers to data that relates to all members of a particular group or set. The population is the entire group that you want to draw conclusions about. For example, if you wanted to know the average height of all adult men in India, the population would be all adult men in India.
2. **Sample Data**: This refers to data collected from a subset of the population. A sample is a group of subjects selected from the population. For example, if it's not feasible to measure the height of every adult man in India, you might measure the heights of a sample of 1000 men selected randomly from across the country. The idea is that the sample represents the population and can give you a good estimate of the population parameter.

In both cases, the goal is usually to learn something about the population. When it's not practical or possible to study the entire population, then a sample is used, and statistical inference is used to draw conclusions about the population based on the sample.

Types of Statistics

Basically, there are two types of statistics.

- Descriptive Statistics
- Inferential Statistics

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

Descriptive Statistics

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

Key points:

- Descriptive statistics summarises or describes the characteristics of a data set.
- Descriptive statistics consists of three basic categories of measures: measures of central tendency, measures of variability (or spread), and frequency distribution.
- Measures of central tendency describe the centre of the data set (mean, median, mode).
- Measures of variability describe the dispersion of the data set (variance, standard deviation).
- Measures of frequency distribution describe the occurrence of data within the data set (count).

Measure of Central Tendency

Measures of central tendency are statistical measures that identify a single value as representative of an entire distribution. They are used to provide a 'typical' or 'average' value for the data set. The three most common measures of central tendency are:

Mean

The mean is a measure of central tendency that represents the average value in a dataset. It's calculated differently depending on whether we're dealing with a population or a sample.

1. **Population Mean**: The population mean, often denoted by the Greek letter μ (mu), is the average of all values in the entire population. It's calculated by summing all the values in the population and dividing by the total number of values (N). The formula for population mean is:

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$$

   where:

   - x_i represents each value in the population
   - N is the total number of values in the population

2. **Sample Mean**: The sample mean, often denoted by x̄ (x-bar), is the average of all values in a sample taken from the population. It's calculated by summing all the values in the sample and dividing by the number of values in the sample (n). The formula for sample mean is:

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

   where:

   - x_i represents each value in the sample
   - n is the number of values in the sample

For example, consider a dataset of five numbers: {1, 2, 3, 4, 5}.

- The population mean would be $(1 + 2 + 3 + 4 + 5)/5 = 3$.

- If we took a sample of three numbers from this dataset, say {1, 2, 3}, then the sample mean would be $(1 + 2 + 3)/3 = 2$.

Median

The median is a measure of central tendency that represents the middle value in a dataset when the data is sorted in ascending or descending order. If there is an odd number of observations, the median is simply the middle number. If there is an even number of observations, the median is calculated as the average of the two middle numbers.

For example, consider the following dataset: {1, 3, 3, 6, 7, 8, 9}. This dataset has 7 numbers (odd), so the median is simply the middle number, which is 6.

Now consider another dataset: {1, 2, 3, 4}. This dataset has 4 numbers (even), so the median is calculated as the average of the two middle numbers. Here, those numbers are 2 and 3. So, the median would be (2 + 3) / 2 = 2.5.

Why median?

The median is an important measure of central tendency for several reasons:

1. **Resistant to Outliers**: The median is not affected by extreme values or outliers in the dataset. This is because it only considers the middle value(s) and not the actual magnitude of the values. For example, in the dataset {1, 2, 3, 100}, the median is 2.5, which is not influenced by the outlier (100).
2. **Reflects the Center of Ordered Data**: The median gives the middle value of the dataset when it's ordered from smallest to largest. This can be more representative of "typical" values in certain datasets, especially when they're skewed.
3. **Applicable to Ordinal Data**: The median can be used with ordinal data (data that can be put into order), while the mean requires interval or ratio data.
4. **Divides Data into Two Equal Parts**: The median divides a dataset into two equal halves, meaning that 50% of the data lies below the median and 50% lies above it. This can be useful in understanding the distribution of data.

Remember, while the median has these advantages, it's just one measure of central tendency. Depending on your specific needs and the nature of your data, other measures like mean or mode might be more appropriate.

Mode

The mode is a measure of central tendency that represents the most frequently occurring value in a dataset. A dataset may have one mode (unimodal), more than one mode (bimodal or multimodal), or no mode at all.

For example, consider the following dataset: {1, 2, 2, 3, 4}. In this dataset, the number 2 appears twice, while all other numbers appear only once. Therefore, the mode of this dataset is 2 because it is the most frequently occurring value.

Now consider another dataset: {1, 1, 2, 2, 3}. In this dataset, both 1 and 2 appear twice. Therefore, this dataset has two modes (it's bimodal), which are 1 and 2.

Finally, consider a dataset where no numbers are repeated, like {1, 2, 3, 4}. This dataset has no mode because no number appears more than once.

Why mode?

The **mode** is an important measure of central tendency for several reasons:

1. **Applicable to Any Level of Measurement**: Unlike the mean and median, the mode can be used with nominal data (categories), ordinal data (rankings), interval data, and ratio data. This makes it a very versatile measure.
2. **Identifies Most Common Value**: The mode gives you the most frequently occurring value in the dataset. This can be useful in many situations where you want to know the most common occurrence, such as the most common colour of cars in a parking lot, or the most common grade in a class.
3. **Can Reveal Multiple Values**: A dataset can have more than one mode (bimodal or multimodal), which can provide additional insight into the distribution of values. For example, in a bimodal distribution, two values occur most frequently, indicating two 'peaks' or 'high points' in the data.
4. **Easy to Understand**: The concept of the "most common" or "most frequently occurring" value is easy to understand, even for people without a background in statistics.

However, like all measures of central tendency, the mode has its limitations and is not always the best measure to use. For example, it's possible for a dataset to have no mode (if no values are repeated) or multiple modes (if several values are repeated with the same frequency), which can sometimes make it less useful for summarising the data.

Measure of Dispersion

Measure of Dispersion is the numbers that are used to represent the scattering of the data. These are the numbers that show the various aspects of the data spread across various parameters. There are various measures of dispersion that are used to represent the data that includes,

- Variance
- Standard Deviation
- Mean Deviation
- Quartile Deviation
- Range, etc

Dispersion in the general sense is the state of scattering. Suppose we have to study the data for thousands of variables there we have to find various parameters that represent the crux of the given data set. These parameters are called the measure of dispersion.

Variance

Variance is defined as, “The measure of how far the set of data is dispersed from their mean value”. Variance is represented with the symbol σ^2. In other words, we can also say that the variance is the average of the squared difference from the mean.

Properties of Variance

Various properties of the Variance of the group of data are,

- As each term in the variance formula is firstly squared and then their mean is found, it is always a non-negative value, i.e. mean can be either positive or can be zero but it can never be negative.
- Variance is always measured in squared units. For example, if we have to find the variance of the height of the student in a class, and if the height of the student is given in cm then the variance is calculated in cm2.

Variance Formula

There are two formulas for Variance, that are:

- Population Variance
- Sample Variance

Formula for Population Variance

The mathematical formula to find the variance of the given data is,

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

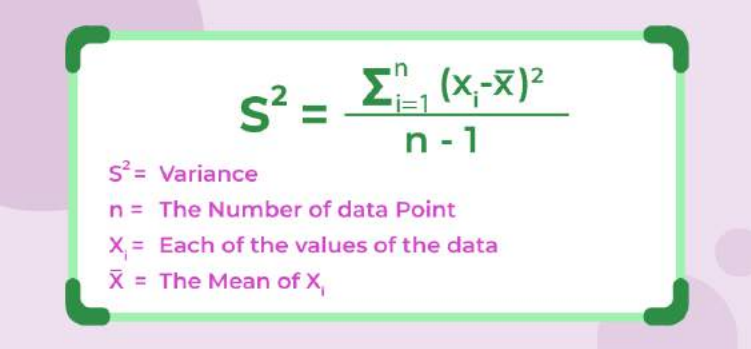
Where,

- ***σ2** is the Variance of the Population,*
- ***N** is the Number of Observation in the Population,*
- ***Xi** is the ith observation in the Population, and*
- ***μ** is the mean of the Population.*

This formula is also called the Population variance formula as it is used for finding the variation in the population data.

Formula for Sample Variance

Also, the other formula for finding the variance is the sample variance formula is discussed in the image



Standard Deviation

How far our given set of data varies along with the mean of the data is measured in standard deviation. Thus, we define standard deviation as the “spread of the statistical data from the mean or average position”. We denote the standard deviation of the data using the symbol **σ**.

We can also define the standard deviation as the square root of the variance.

Properties of Standard Deviation

Various properties of the Variance of the group of data are,

- Standard Deviation is the square root of the variance of the given data set. It is also called root mean square deviation.
- Standard Deviation is a non-negative quantity i.e. it always has positive values or zero values.
- If all the values in a data set are similar then Standard Deviation has a value close to zero. Whereas if the values in a data set are very different from each other then standard deviation has a high positive value.

Standard Deviation Formula

There are two formulas for the standard deviation listed as follows:

- Population Standard Deviation
- Sample Standard Deviation

Formula for Population Standard Deviation

The mathematical formula to find the standard deviation of the given data is,

$$\sigma = \sqrt{Population\ Variance}$$

This formula is also called the Population standard deviation formula as it is used for finding the standard deviation in the population data.

Formula for Sample Standard Deviation

Also, the other formula for finding the standard deviation is the sample space i.e. sample variance formula is discussed in the image below,

$$S = \sqrt{Sample\ Variance}$$

Relation between Standard Deviation and Variance

Variance and Standard deviation are the most common measure of the given set of data. They are used to find the deviation of the values from their mean value or the spread of all the values of the data set.

- Variance is defined as the average degree through which all the values of a given data set deviate from the mean value.
- Standard Deviation is the degree to which the values in a data set are spread out with respect to the mean value.

The relationship between Variance and Standard Deviation is discussed below.

Variance = (Standard Deviation)$_2$

OR

√(Variance) = Standard Deviation

Differences Between Standard Deviation and Variance

Standard Deviation	Variance
Standard Deviation is defined as the square root of the variance.	Variance is defined as the average of the squared differences from the mean.
The standard deviation provides a measure of the typical distance between data points and the mean.	The variance provides a measure of the average squared distance between data points and the mean.
It is represented by the Greek symbol σ.	It is represented by a square of the Greek symbol sigma i.e. σ2.

It has the same unit as the data set.	Its unit is the square of the unit of the data set.
It represents the volatility in the market or given data set.	It represents the degree to which the average return varies according to the long-term change in the market.

Examples on Variance and Standard Deviation

Example 1: Find the variance and standard deviation of all the possibilities of rolling a die.

Solution:

All possible outcomes of rolling a die are {1; 2; 3; 4; 5; 6}.

This data set has six values (n) = 6

Before finding the variance, we need to find the mean of the data set.

Mean, $\overline{x}$ = (1+2+3+4+5+6)/6 = 3.5

We can put the value of data and mean in the formula to get;

$\sigma^2 = \Sigma (xi - \bar{x})^2/n$

$\Rightarrow \sigma^2 = [(1 - 3.5)^2 + (2 - 3.5)^2 + (3 - 3.5)^2 + (4 - 3.5)^2 + (5 - 3.5)^2 + (6 - 3.5)^2]/6$

$\Rightarrow \sigma^2 = (6.25+2.25+0.25+0.25+2.25+6.25)/6$

Variance (σ2) = 2.917

Now,

Standard Deviation (σ) = √ (σ2)

⇒ Standard Deviation (σ) = √(2.917)

⇒ Standard Deviation (σ) = 1.708

Example 2: Find the variance and standard deviation of all the even numbers less than 10.

Solution:

Even Numbers less than 10 are {0, 2, 4, 6, 8}

This data set has five values (n) = 5

Before finding the variance, we need to find the mean of the data set.

Mean, $\overline{x}$ = (0+2+4+6+8)/5 = 4

We can put the value of data and mean in the formula to get;

$\sigma^2 = \Sigma (xi - \overline{x})/n$

$\Rightarrow \sigma^2 = [\,(0-4)^2 + (2-4)^2 + (4-4)^2 + (6-4)^2 + (8-4)^2\,]/5$

$\Rightarrow \sigma^2 = (16 + 4 + 0 + 4 + 16)/5 = 40/5$

Variance $(\sigma^2) = 8$

Now, Standard Deviation (σ) = √ (σ2)

⇒ *Standard Deviation (σ) = √(8)*

⇒ *Standard Deviation (σ) = 2.828*

Random Variable

Random variable is a process of mapping an output of a random process or experiment to a number.

Types of Random Variable

Random variable are of two types that are,

Discrete Random Variable

A random variable X is said to be discrete if it takes on a finite number of values. For example, tossing a coin. In this experiment, there are two possible outcomes: heads (H) and tails (T). We can define a random variable X that maps these outcomes to real numbers. For instance, we could define X such that it assigns the number 1 to the outcome ‘heads’ and the number 0 to the outcome ‘tails’.

Here's how the mapping would look:

- If the coin lands on heads, X(H) = 1
- If the coin lands on tails, X(T) = 0

So, in this case, X is a random variable because it assigns numerical values to the outcomes of a random process (tossing a coin). The value of X is not known until the coin is tossed and the outcome is observed.

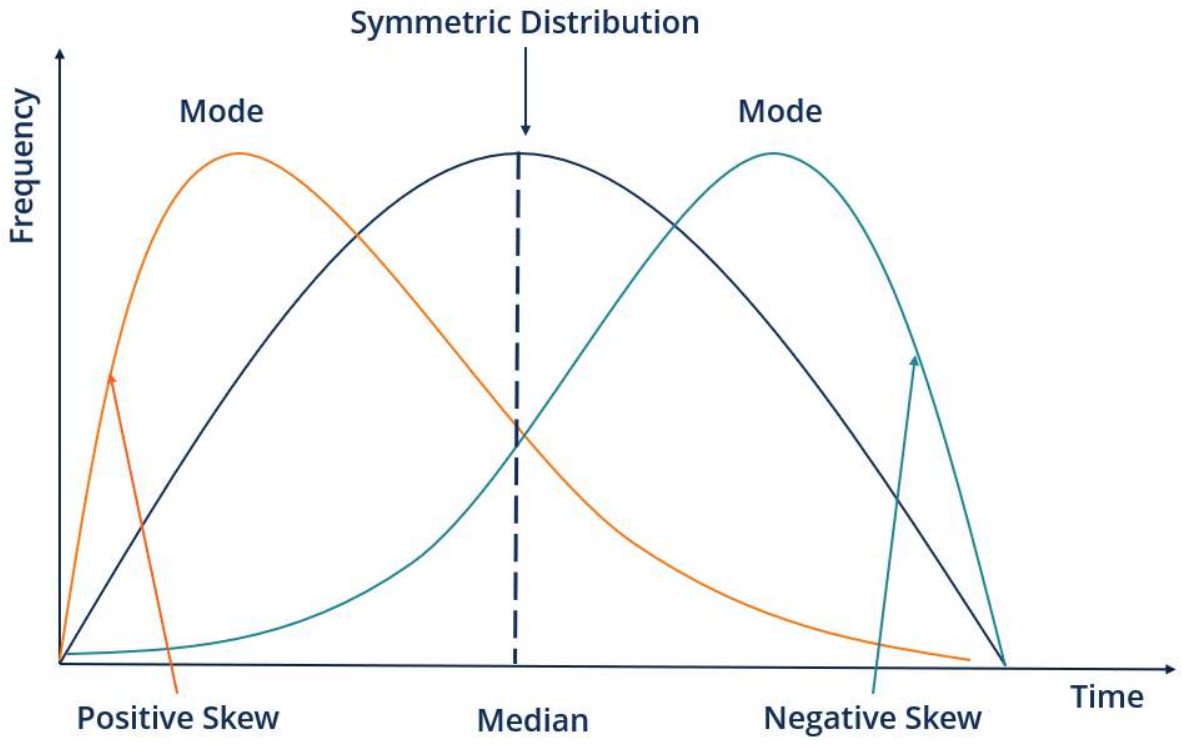
This is a simple example of a discrete random variable, as it can only take on a finite number of values (in this case, 0 and 1).

Continuous Random Variable

This type of random variable can take on any value within a continuous range. For example, the height of a person is a continuous random variable because it can take on any value within a certain range (e.g., any value between 150 cm and 200 cm).

Skewness

Skewness is a measure of asymmetry or distortion of symmetric distribution. It measures the deviation of the given distribution of a random variable from a symmetric distribution, such as normal distribution. A normal distribution is without any skewness, as it is symmetrical on both sides. Hence, a curve is regarded as skewed if it is shifted towards the right or the left.



Summary

- Skewness measures the deviation of a random variable's given distribution from the normal distribution, which is symmetrical on both sides.
- A given distribution can either be skewed to the left or the right. Skewness risk occurs when a symmetric distribution is applied to the skewed data.
- Skewness tells you where the outliers occur, although it doesn't tell you how many outliers occur.
- In normal distribution Mean = Median = Mode.

Types of Skewness

1. Positive Skewness

If the given distribution is shifted to the left and with its tail on the right side, it is a positively skewed distribution. It is also called the right-skewed distribution. A tail is referred to as the tapering of the curve differently from the data points on the other side.

As the name suggests, a positively skewed distribution assumes a skewness value of more than zero. Since the skewness of the given distribution is on the right, the mean value is greater than the median and moves towards the right, and the mode occurs at the highest frequency of the distribution (Mean > Median > Mode).

2. Negative Skewness

If the given distribution is shifted to the right and with its tail on the left side, it is a negatively skewed distribution. It is also called a left-skewed distribution. The skewness value of any distribution showing a negative skew is always less than zero. The skewness of the given distribution is on the left; hence, the mean value is less than the median and moves towards the left, and the mode occurs at the highest frequency of the distribution (Mean < Median < Mode).

Percentile And Quartiles

Percentiles

A percentile is a value below which a certain percentage of data points lie.

Example 1: X = {2,3,3,4,6,6,6,7,8,8,9,9,10,11,12}. How much percentage of data lies below point 10.

Solution:

$$\text{Percentile Rank of 10} = \frac{Number\ of\ Values\ below\ 10}{Total\ number\ of\ data\ points\ (n)} * 100$$

$$= \frac{12}{15} * 100$$

= 80

80% of data points lie below 10. Therefore, 10 is at 80th percentile.

Example 2: What value lies at 25th percentile?

Solution: Value = $\frac{Percentile}{100}$ * (n+1)

= $\frac{25}{100}$ * (15 + 1)

= 4

Value 4 lies at 25th percentile.

Note: If the value comes in decimal for example 4.5 then find the two data points closest to your percentile (the one below and the one above) and calculate its average.

Quartiles

Quartiles are a set of three values that divide a dataset into four equal parts, each containing 25% of the data points. These three values are:

- Q1 (First Quartile): The 25th percentile. This value is the middle number between the smallest number (minimum) and the median of the dataset. 25% of the data points are less than Q1.
- Q2 (Second Quartile): The 50th percentile (median). This value is the middle number of the dataset, with 50% of the data points below it and 50% above it.

- Q3 (Third Quartile): The 75th percentile. This value is the middle number between the median and the largest number (maximum) of the dataset. 75% of the data points are less than Q3.

Interpretation of Quartiles:

Quartiles, along with the median, provide valuable insights into the distribution of a dataset. They can be used to:

- Identify the spread of the data: The difference between Q1 and Q3 is known as the interquartile range (IQR). A large IQR indicates a wide spread of data, while a small IQR indicates a narrow spread.
- Identify outliers: Data points that are significantly lower than Q1 or higher than Q3 can be considered outliers.
- Compare different datasets: Quartiles allow you to compare the distributions of different datasets, even if they have different units or scales.

Examples:

- If you have a dataset of test scores, you can calculate the quartiles to see how the scores are distributed. For example, Q1 might be 70, Q2 might be 80, and Q3 might be 90. This would tell you that 25% of the students scored below 70, 50% scored between 70 and 90, and 25% scored above 90.
- If you have a dataset of the monthly rainfall in a particular city, you can calculate the quartiles to see how the rainfall is distributed throughout the year. For example, Q1 might be 2 inches, Q2 might be 5 inches, and Q3 might be 8 inches. This would tell you that 25% of the months receive less than 2 inches of rain, 50% of the months receive between 2 and 8 inches of rain, and 25% of the months receive more than 8 inches of rain.

Five Number Summary

The five-number summary is a set of five descriptive statistics that provides a concise overview of the distribution of a dataset. It consists of the following values:

1. Minimum: The smallest value in the dataset.
2. First Quartile (Q1): The 25th percentile. This is the middle number between the minimum and the median. 25% of the data points are less than or equal to Q1.
3. Median (Q2): The 50th percentile. This is the middle number of the dataset. 50% of the data points are less than or equal to the median and 50% are greater than or equal to the median.
4. Third Quartile (Q3): The 75th percentile. This is the middle number between the median and the maximum. 75% of the data points are less than or equal to Q3.
5. Maximum: The largest value in the dataset.

These five numbers provide information about the centre, spread, and range of the data:

- Centre: The median provides information about the central tendency of the data.
- Spread: The interquartile range (IQR), calculated as Q3 - Q1, provides information about how spread out the data is. A large IQR indicates a wider spread of data points, while a small IQR indicates a narrower spread.
- Range: The difference between the minimum and the maximum values provides information about the entire range of values in the dataset.

Benefits of the Five-Number Summary:

- Simple and easy to understand: The five-number summary is a quick and easy way to get a general idea of the distribution of a dataset.
- Provides information about key aspects of the data: It provides information about the centre, spread, and range of the data, which are important for understanding the overall distribution.
- Used in various statistical analyses: The five-number summary can be used as a starting point for further statistical analysis, such as calculating measures of variability or creating box plots.

Limitations of the Five-Number Summary:

- Hides details: The five-number summary only provides a limited amount of information about the data. It can mask important details about the distribution, such as the presence of outliers or skewness.
- Not suitable for small datasets: The five-number summary is not as informative for small datasets where there are few data points between the quartiles.

Overall, the five-number summary is a valuable tool for gaining a basic understanding of the distribution of a dataset. It is easy to calculate and understand, and it provides information about key aspects of the data. However, it is important to be aware of its limitations and to use it in conjunction with other statistical analyses to gain a more complete picture of the data.

How to remove outliers?

Example 1: X = {1,2,2,2,3,3,4,5,5,5,6,6,6,6,7,8,8,9,29}

Solution: First we need to find lower fence and upper fence

Lower Fence = Q1 - 1.5(IQR)

Upper Fence = Q3 + 1.5(IQR)

IQR = Q3 - Q1

Q1 = 25/100 * 20 = 5th value = 3

Q3 = 75/100 * 2. = 15th value = 7

IQR = Q3 - Q1 = 7-3 = 4

Lower Fence = 3 - 1.5(4) = -3

Higher Fence = 7 + 1.5(4) = 13

Any number less than the lower bound or greater than the upper bound is considered an outlier. Here, 29 is an outlier.

Box Plot

A box plot, also known as a whisker plot, is a graphical representation of a five-number summary of a dataset. The five-number summary consists of the following statistics:

1. **Minimum**: The smallest number in the dataset.
2. **First Quartile (Q1)**: The middle number between the minimum and the median (25th percentile).
3. **Median (Q2)**: The middle number of the dataset (50th percentile).
4. **Third Quartile (Q3)**: The middle value between the median and the maximum (75th percentile).
5. **Maximum**: The largest number in the dataset.

Here’s how these five numbers are represented in a box plot:

- The **box** in the box plot represents the interquartile range (IQR), which is the range between Q1 and Q3. The length of the box is therefore IQR = Q3 - Q1.
- The **line inside the box** represents the median of the dataset.
- The **whiskers** (lines extending from the box) represent the range of the data within 1.5 times the IQR from the box. Any data point outside this range is considered an outlier and is represented as a dot (or a different marker).
- The **ends of the whiskers** represent the minimum and maximum data values excluding outliers.

This way, a box plot provides a visual summary of the distribution of a dataset, including its central tendency (median), variability (IQR), and outliers. It's a useful tool for comparing distributions across different categories.

Correlation

Correlation refers to the statistical relationship between two variables. It measures the extent to which two variables tend to move together, either in the same direction (positive correlation) or in the opposite direction (negative correlation).

Here are some examples to illustrate the concept:

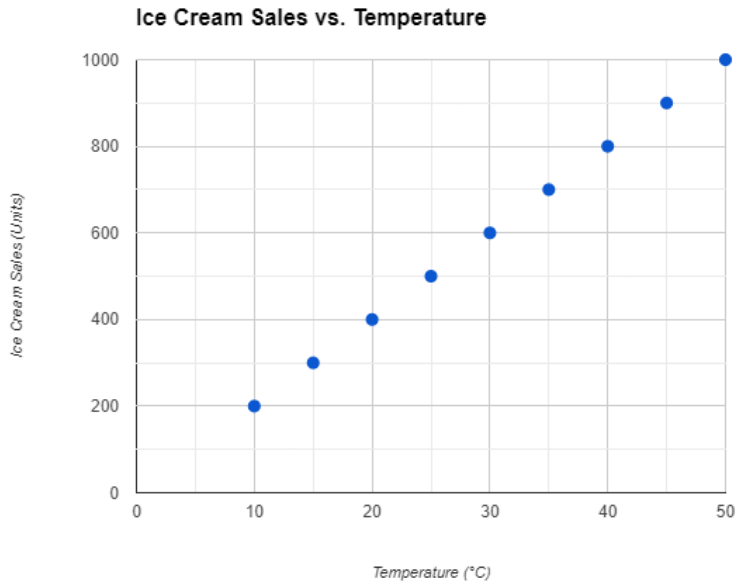
Positive Correlation:

- Ice cream sales and temperature: As the temperature rises, ice cream sales tend to go up. This is a positive correlation because both variables increase together.

Temperature and Ice Cream Sales Data Points:

Temperature (°C)	Ice Cream Sales (Units)

10	200
15	300
20	400
25	500
30	600
35	700
40	800
45	900
50	1000

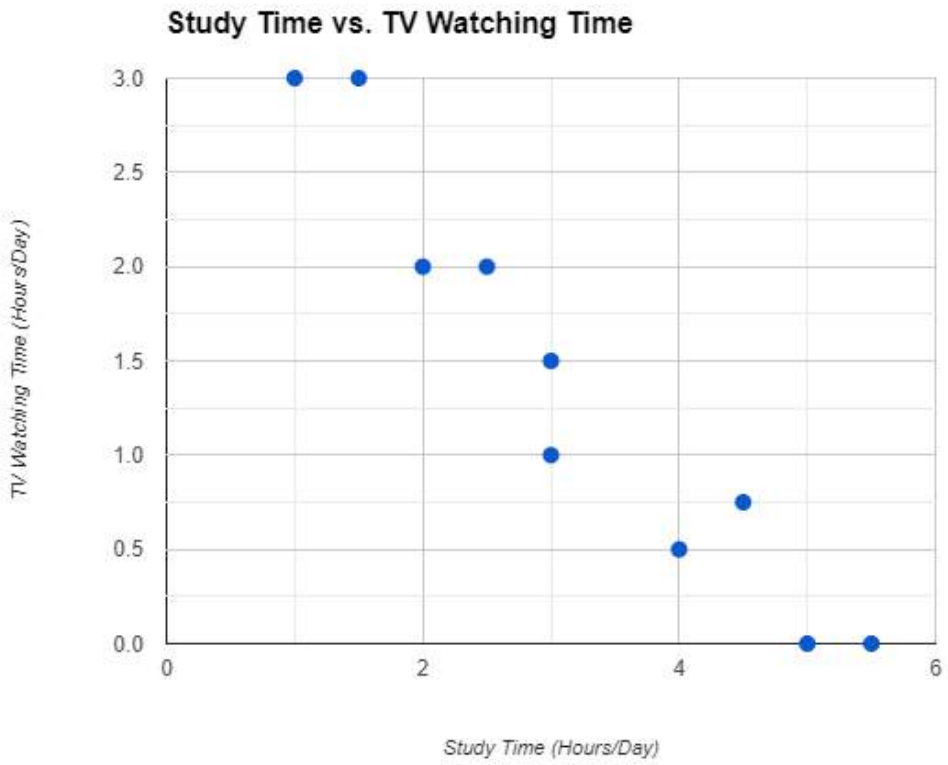


Negative Correlation:

- Study time and time spent watching TV: Students who spend more time studying tend to spend less time watching TV. This is another example of a negative correlation.

Study Time and TV Watching Data Points:

Student	Study Time (Hours/Day)	TV Watching Time (Hours/Day)
Alice	3	1
Bob	4	0.5
Charlie	2	2
Diana	1	3
Eve	5	0
Frank	4.5	0.75
Gwen	3	1.5
Harry	2.5	2
Iris	5.5	0
Jack	1.5	3



Zero Correlation:

- Shoe size and number of movies watched: There is no relationship between shoe size and the number of movies a person watches. This is a zero correlation because the two variables are not related.
- Eye colour and intelligence: There is no evidence to suggest that eye colour has any impact on intelligence. This is another example of a zero correlation.

It's important to note that correlation does not imply causation. Just because two variables are correlated does not mean that one causes the other. There could be other factors that are influencing both variables.

Here is an example:

- There is a positive correlation between the number of ice cream trucks in a neighbourhood and the crime rate. However, this does not mean that ice cream trucks cause crime. It is more likely that there is another factor, such as poverty, that is influencing both the number of ice cream trucks and the crime rate.

Correlation is a useful tool for identifying relationships between variables, but it is important to interpret it carefully and avoid drawing causal conclusions based on correlation alone.

Covariance

Covariance is a measure of the relationship between two random variables. The metric evaluates how much – to what extent – the variables change together. In other words, it is essentially a measure of the variance between two variables. However, the metric does not assess the dependency between variables.

Unlike the correlation coefficient, covariance is measured in units. The units are computed by multiplying the units of the two variables. The variance can take any positive or negative values. The values are interpreted as follows:

- **Positive covariance**: Indicates that two variables tend to move in the same direction.
- **Negative covariance**: Reveals that two variables tend to move in inverse directions.

Formula for Covariance

The covariance formula is similar to the formula for correlation and deals with the calculation of data points from the average value in a dataset. For example, the covariance between two random variables X and Y can be calculated using the following formula (for population):

$$\text{Cov}\,(X, Y) = \frac{\sum(X_i - \bar{X})(Y_j - \bar{Y})}{n}$$

For a sample covariance, the formula is slightly adjusted:

$$\text{Cov}\,(X, Y) = \frac{\sum(X_i - \bar{X})(Y_j - \bar{Y})}{n - 1}$$

Where:

- $\mathbf{X_i}$ – the values of the X-variable
- $\mathbf{Y_j}$ – the values of the Y-variable
- $\bar{\mathbf{X}}$ – the mean (average) of the X-variable
- $\bar{\mathbf{Y}}$ – the mean (average) of the Y-variable
- **n** – the number of data points

Example: X = 2,4,6 and Y = 3,5,7. Calculate sample covariance.

Solution: Covariance using formula

Here's how to calculate the covariance using the formula:

1. Calculate the mean of X and Y:

- Mean of X ($\bar{X}$) = (2 + 4 + 6) / 3 = 4
- Mean of Y ($\bar{Y}$) = (3 + 5 + 7) / 3 = 5

2. Calculate the deviations from the mean for each X and Y value:

- X deviations:

    - $X1 - \bar{X} = 2 - 4 = -2$
    - $X2 - \bar{X} = 4 - 4 = 0$
    - $X3 - \bar{X} = 6 - 4 = 2$
- Y deviations:
    - $Y1 - \bar{Y} = 3 - 5 = -2$
    - $Y2 - \bar{Y} = 5 - 5 = 0$
    - $Y3 - \bar{Y} = 7 - 5 = 2$

3. Multiply the corresponding deviations from the mean for X and Y:

- $(X1 - \bar{X}) * (Y1 - \bar{Y}) = (-2) * (-2) = 4$
- $(X2 - \bar{X}) * (Y2 - \bar{Y}) = (0) * (0) = 0$
- $(X3 - \bar{X}) * (Y3 - \bar{Y}) = (2) * (2) = 4$

4. Sum the products of deviations:

$4 + 0 + 4 = 8$

5. Divide the sum by N-1, where N is the number of data points:

N = 3, so covariance = 8 / (3-1) = 8 / 2 = 4.0

Therefore, the covariance of X and Y is 4.0.

Advantages of Covariance:

- Simple to calculate: The formula for covariance is straightforward and easy to understand and implement.
- Measures linear relationship: Covariance effectively measures the linear relationship between two variables. If two variables move together in the same direction or opposite directions, covariance can accurately capture this relationship.

Disadvantages of Covariance:

- Does not measure strength of relationship: Covariance only indicates the direction of the relationship between variables, not its strength. A high covariance value can be misleading if the units are different or the data points are clustered tightly.

Covariance vs. Correlation

Covariance and correlation both primarily assess the relationship between variables. The closest analogy to the relationship between them is the relationship between the variance and standard deviation.

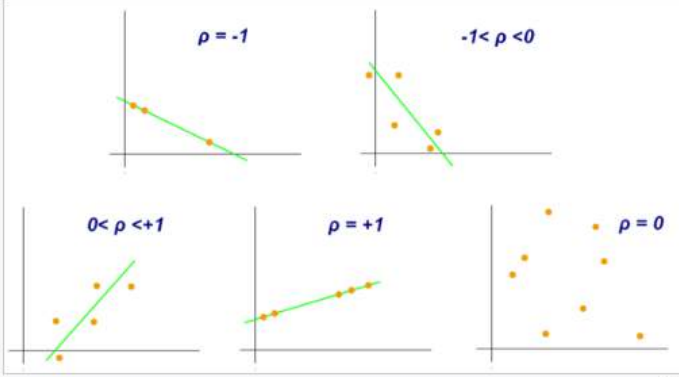
Covariance measures the total variation of two random variables from their expected values. Using covariance, we can only gauge the direction of the relationship (whether the variables tend to move in tandem or show an inverse relationship). However, it does not indicate the strength of the relationship, nor the dependency between the variables.

On the other hand, **correlation** measures the strength of the relationship between variables. Correlation is the scaled measure of covariance. It is dimensionless. In other words, the correlation coefficient is always a pure value and not measured in any units.

Pearson's coefficient

Correlation

The Pearson Correlation Coefficient, often denoted by 'r', is a statistical measure that quantifies the strength and direction of the linear relationship between two variables. It's a number between -1 and 1.



Examples of scatter diagrams with different values of correlation coefficient (ρ)

- A correlation coefficient close to +1 indicates a strong positive correlation, meaning as one variable increases, the other also increases.
- A coefficient close to -1 indicates a strong negative correlation, meaning as one variable increases, the other decreases.
- A coefficient close to 0 suggests little to no linear relationship between the variables.

The Pearson Correlation Coefficient is calculated as the covariance of the two variables divided by the product of their standard deviations. This is why it's sometimes referred to as the Pearson Product-Moment Correlation Coefficient.

For example, if we have two variables, X and Y, representing the hours studied and the marks obtained respectively, a positive Pearson Correlation Coefficient would indicate that as the hours studied increase (X), the marks obtained (Y) also tend to increase.

Remember, while the Pearson Correlation Coefficient can indicate a linear relationship between two variables, it does not imply causation. There could be other factors at play influencing the variables.

Formula for pearson's correlation coefficient is given as

$$\rho(X, Y) = \frac{\mathrm{Cov}(X, Y)}{\sigma_X \sigma_Y}$$

Spearman's Rank Correlation

Spearman's rank correlation measures the strength and direction of association between two ranked variables. It basically gives the measure of monotonicity of the relation between two variables i.e. how well the relationship between two variables could be represented using a monotonic function.

The formula for Spearman's rank coefficient is:

$$\rho = 1 - \frac{6\Sigma\, d_i^2}{n(n^2 - 1)}$$

ρ = Spearman's rank correlation coefficient

di = Difference between the two ranks of each observation

n = Number of observations

The Spearman Rank Correlation can take a value from +1 to -1 where,

- A value of +1 means a perfect association of rank
- A value of 0 means that there is no association between ranks
- A value of -1 means a perfect negative association of rank

Let's understand the concept better with the help of an example.

Consider the score of 5 students in Maths and Science that are mentioned in the table.

Students	Maths	Science
A	35	24
B	20	35
C	49	39
D	44	48
E	30	45

Step 1: Create a table for the given data.

Step 2: Rank both the data in descending order. The highest marks will get a rank of 1 and the lowest marks will get a rank of 5.

Step 3: Calculate the difference between the ranks (d) and the square value of d.

Step 4: Add all your d square values.

Students	Maths	Rank	Science	Rank	d	d square
A	35	3	24	5	2	4
B	20	5	35	4	1	1
C	49	1	39	3	2	4
D	44	2	48	1	1	1
E	30	4	45	2	2	4
						14

Step 5: Insert these values into the formula.

$$\rho = 1 - \frac{6\Sigma\, d_i^2}{n(n^2 - 1)}$$

= 1 - (6 * 14) / 5(25 - 1)

= 0.3

The Spearman's Rank Correlation for the given data is 0.3. The value is near 0, which means that there is a weak correlation between the two ranks.

Note: The formula for Spearman's Rank Correlation can also be given as

$$r_s = \frac{Cov(R(X),R(Y))}{\sigma_{R(X)} * \sigma_{R(Y)}}$$

Where,

$Cov(R(X), R(Y))$ is the covariance of the rank variables,

$\sigma_{R(X)}$ and $\sigma_{R(Y)}$ are the standard deviations of the rank variables.

Probability Distribution Function

Probability Distribution Function is a function in mathematics that gives the probability of all the possible outcomes of any event. We define probability distribution as how all the possible probabilities of any event are allocated over the distinct values for an unexpected variable.

There are two types of probability distributions:

- Discrete probability distributions
- Continuous probability distributions

You can read more about Probability Distribution Function here.

Discrete probability distributions

A discrete probability distribution is a probability distribution of a categorical or discrete variable.

Discrete probability distributions only include the probabilities of values that are possible. In other words, a discrete probability distribution doesn't include any values with a probability of zero. For example, a probability distribution of dice rolls doesn't include 2.5 since it's not a possible outcome of dice rolls.

The probability of all possible values in a discrete probability distribution add up to one. It's certain (i.e., a probability of one) that an observation will have one of the possible values.

Probability tables

A probability table represents the discrete probability distribution of a categorical variable. Probability tables can also represent a discrete variable with only a few possible values or a continuous variable that's been grouped into class intervals.

A probability table is composed of two columns:

- The values or class intervals
- Their probabilities

Example: Probability table

A robot greets people using a random greeting. The probability distribution of the greetings is described by the following probability table:

Greeting	Probability
“Greetings, human!”	.6
“Hi!”	.1
“Salutations, organic life-form.”	.2
“Howdy!”	.1

Notice that all the probabilities are greater than zero and that they sum to one.

Probability Mass Function

Probability mass function can be defined as the probability that a discrete random variable will be exactly equal to some particular value. In other words, the probability mass function assigns a particular probability to every possible value of a discrete random variable.

Probability Mass Function Formula

The probability mass function provides all possible values of a discrete random variable as well as the probabilities associated with it. Let X be the discrete random variable.

Then the formula for the probability mass function, f(x), evaluated at x, is given as follows:

$f(x) = P(X = x)$

The cumulative distribution function of a discrete random variable is given by the formula $F(x) = P(X \leq x)$.

Probability Mass Function Example

Example 1: Let's consider a simple example of tossing a fair coin once. This is a discrete random variable with two possible outcomes: heads (H) and tails (T). We can assign the value 0 to heads and 1 to tails.

Probability Mass Function (PMF): The PMF gives the probability of each possible outcome. In this case, the PMF is:

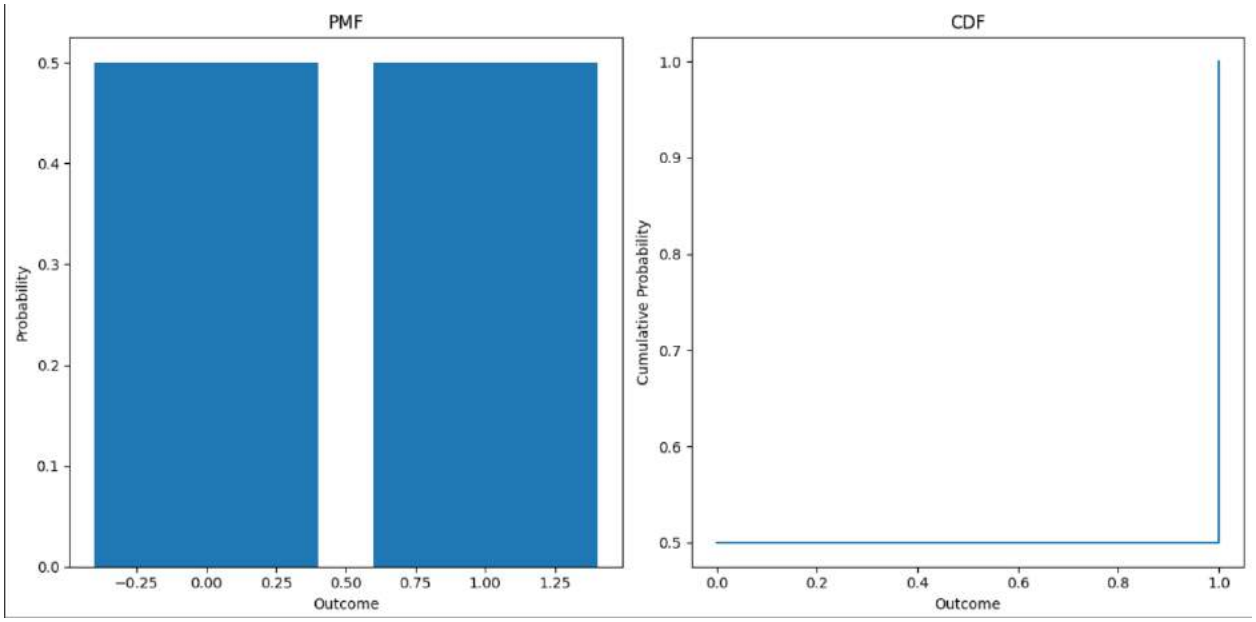
- P(X=0) = 0.5 (probability of getting heads)
- P(X=1) = 0.5 (probability of getting tails)

If we were to graph the PMF, we would have a bar at 0 with height 0.5 and a bar at 1 with height 0.5.

Cumulative Distribution Function (CDF): The CDF gives the cumulative probability up to a certain value. In this case, the CDF is:

- F(X=0) = P(X<=0) = 0.5 (probability of getting heads or less)
- F(X=1) = P(X<=1) = 1.0 (probability of getting tails or less)

If we were to graph the CDF, it would be a step function that increases from 0 to 0.5 at X=0 and from 0.5 to 1 at X=1.



So, the PMF gives the probabilities of individual outcomes, while the CDF gives the cumulative probabilities. The CDF at any point is the sum of the PMF for all outcomes up to that point.

Example 2: Let's consider another example of rolling a fair six-sided die.

PMF:

The PMF of rolling a die is the probability of each outcome (1 through 6), which is 1/6 for a fair die. We can represent it as follows:

Outcome (x) : 1 2 3 4 5 6

PMF P(X=x) : 1/6 1/6 1/6 1/6 1/6 1/6

CDF:

The CDF is the cumulative sum of the probabilities up to a certain value. For a die roll:

Outcome (x) : 1 2 3 4 5 6

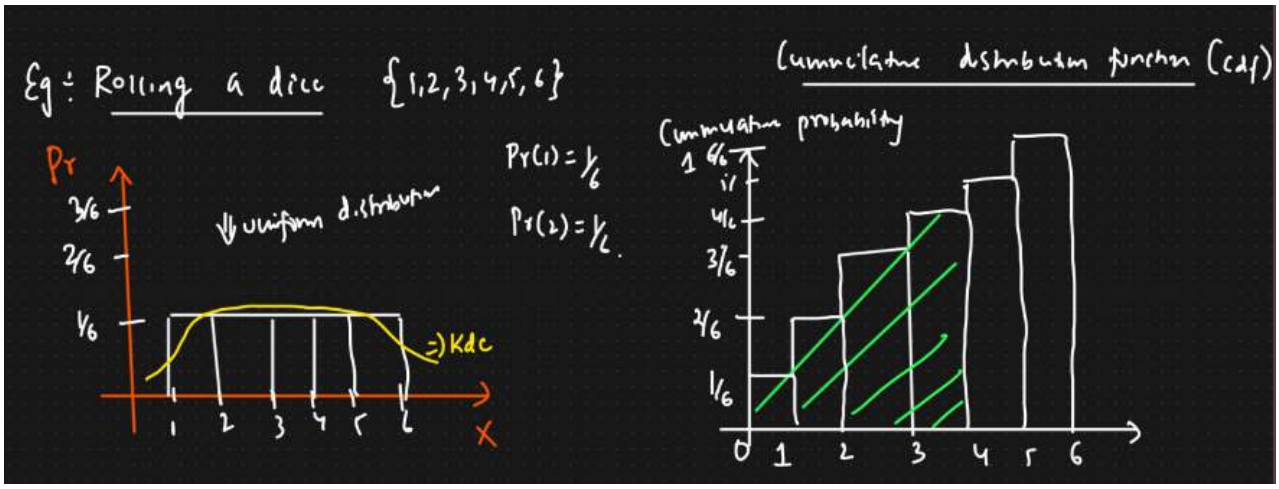
CDF F(X<=x) : 1/6 2/6 3/6 4/6 5/6 1

The Cumulative Distribution Function (CDF) at 4, denoted as F(X<=4), is the sum of the probabilities of getting 1, 2, 3, or 4. Since each of these has a probability of 1/6, the CDF at 4 is:

$$F(X \leq 4) = P(X = 1) + P(X = 2) + P(X = 3) + P(X = 4) = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{4}{6} = \frac{2}{3} \approx 0.67$$

So, there's approximately a 67% chance that a roll of the die will result in a value of 4 or less.

Graphical Representation



You can read more about PMF here.

Continuous probability distributions

A continuous probability distribution is the probability distribution of a continuous variable.

A continuous variable can have any value between its lowest and highest values. Therefore, continuous probability distributions include every number in the variable's range.

The probability that a continuous variable will have any specific value is so infinitesimally small that it's considered to have a probability of zero. However, the probability that a value will fall within a certain interval of values within its range is greater than zero.

Probability Density Function Definition

Probability density function defines the density of the probability that a continuous random variable will lie within a particular range of values. To determine this probability, we integrate the probability density function between two specified points.

Probability Density Function Formula

The probability density function of a continuous random variable is analogous to the probability mass function of a discrete random variable. Discrete random variables can be evaluated at a particular point while continuous random variables have to be evaluated between a certain interval. This is because the probability that a continuous random variable will take an exact value is 0. Given below are the various probability density function formulas.

Probability Density Function of Continuous Random Variable

Suppose we have a continuous random variable, X. Let F(x) be the cumulative distribution function of X. Then the formula for the probability density function, f(x), is given as follows:

$$f(x) = \frac{dF(x)}{dx} = F'(x)$$

If we want to find the probability that X lies between lower limit 'a' and upper limit 'b' then using the probability density function this can be given as:

$$P(a < X \leq b) = F(b) - F(a) = \int_a^b f(x)dx$$

Here, F(b) and F(a) represent the cumulative distribution function at b and a respectively.

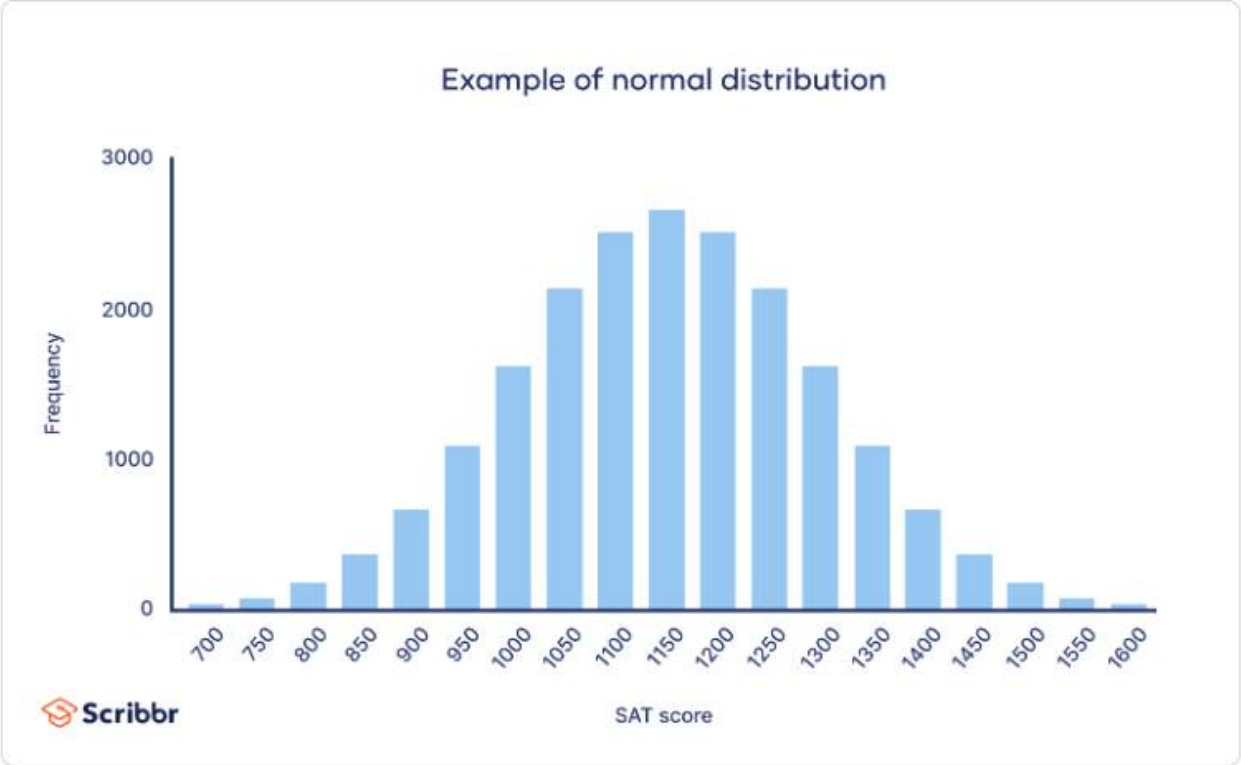
Probability Density Function Example

This YouTube video has explained PMF and PDF with the help of examples in a very easy way.

Normal Distribution:

In a normal distribution, data is symmetrically distributed with no skew. When plotted on a graph, the data follows a bell shape, with most values clustering around a central region and tapering off as they go further away from the centre.

Normal distributions are also called Gaussian distributions or bell curves because of their shape.



Why do normal distributions matter?

All kinds of variables in natural and social sciences are normally or approximately normally distributed. Height, birth weight, reading ability, job satisfaction, or SAT scores are just a few examples of such variables.

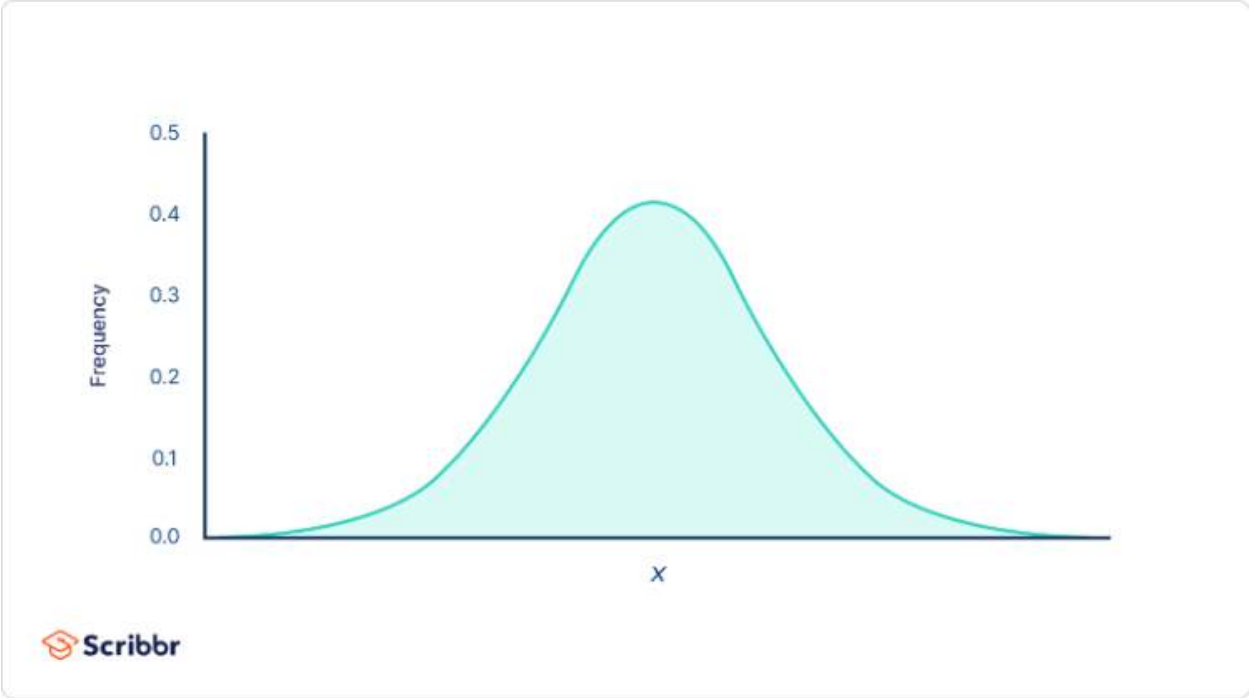
Because normally distributed variables are so common, many statistical tests are designed for normally distributed populations.

Understanding the properties of normal distributions means you can use inferential statistics to compare different groups and make estimates about populations using samples.

What are the properties of normal distributions?

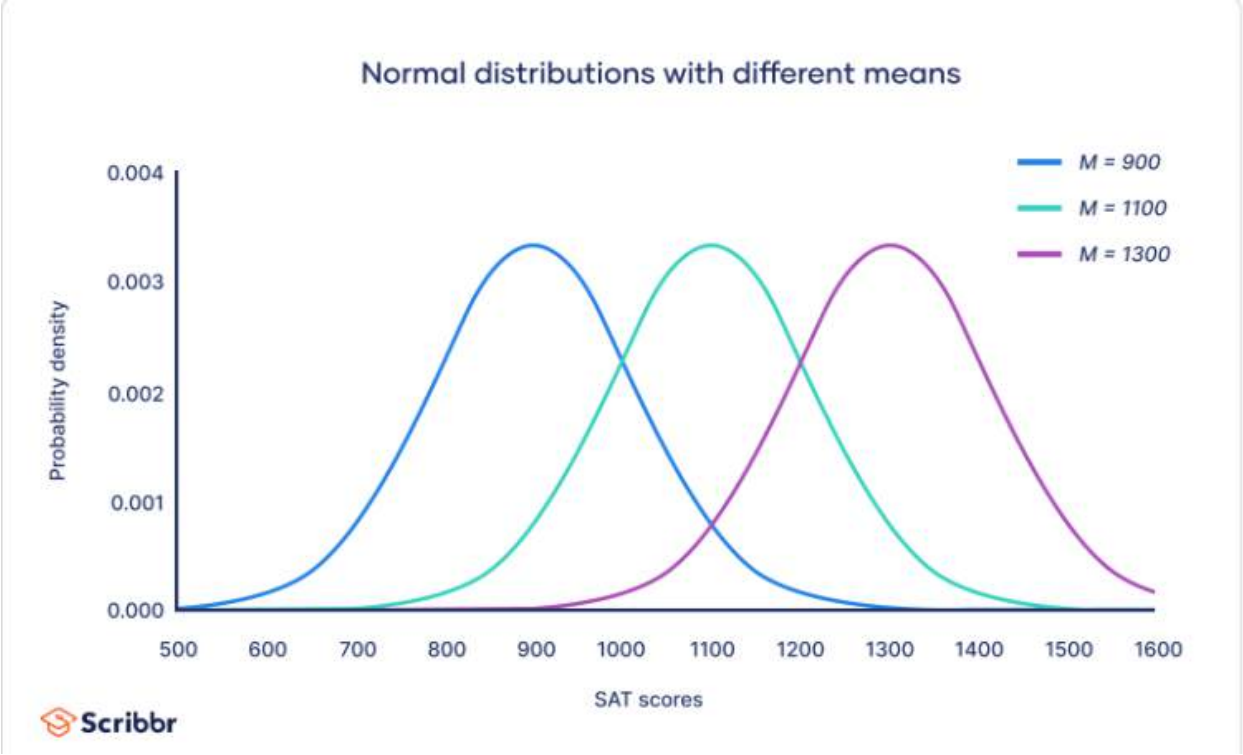
Normal distributions have key characteristics that are easy to spot in graphs:

- The mean, median and mode are exactly the same.
- The distribution is symmetric about the mean—half the values fall below the mean and half above the mean.
- The distribution can be described by two values: the mean and the standard deviation.

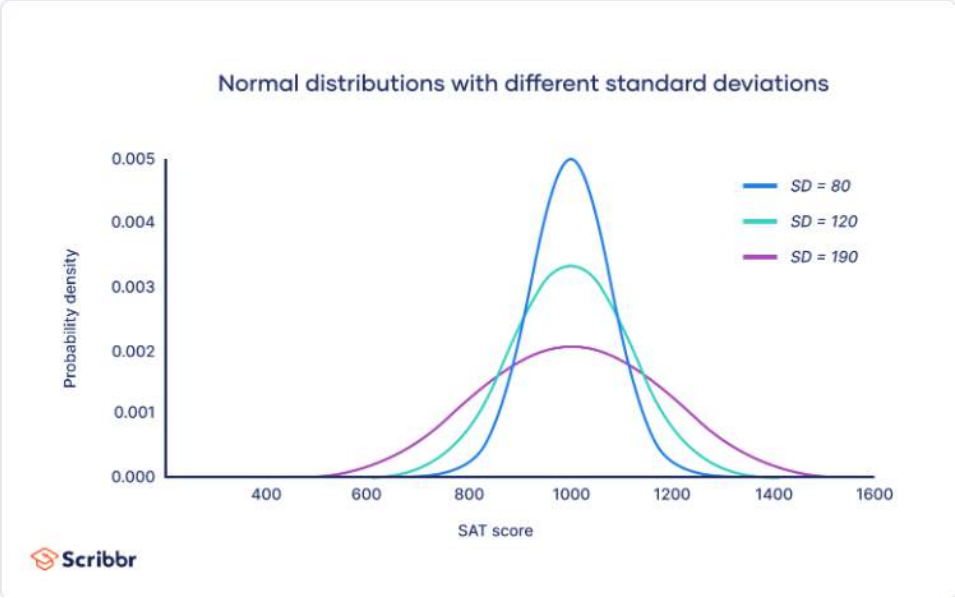


The mean is the location parameter while the standard deviation is the scale parameter.

The mean determines where the peak of the curve is centred. Increasing the mean moves the curve right, while decreasing it moves the curve left.



The standard deviation stretches or squeezes the curve. A small standard deviation results in a narrow curve, while a large standard deviation leads to a wide curve.



Empirical rule

The empirical rule, or the 68-95-99.7 rule, tells you where most of your values lie in a normal distribution:

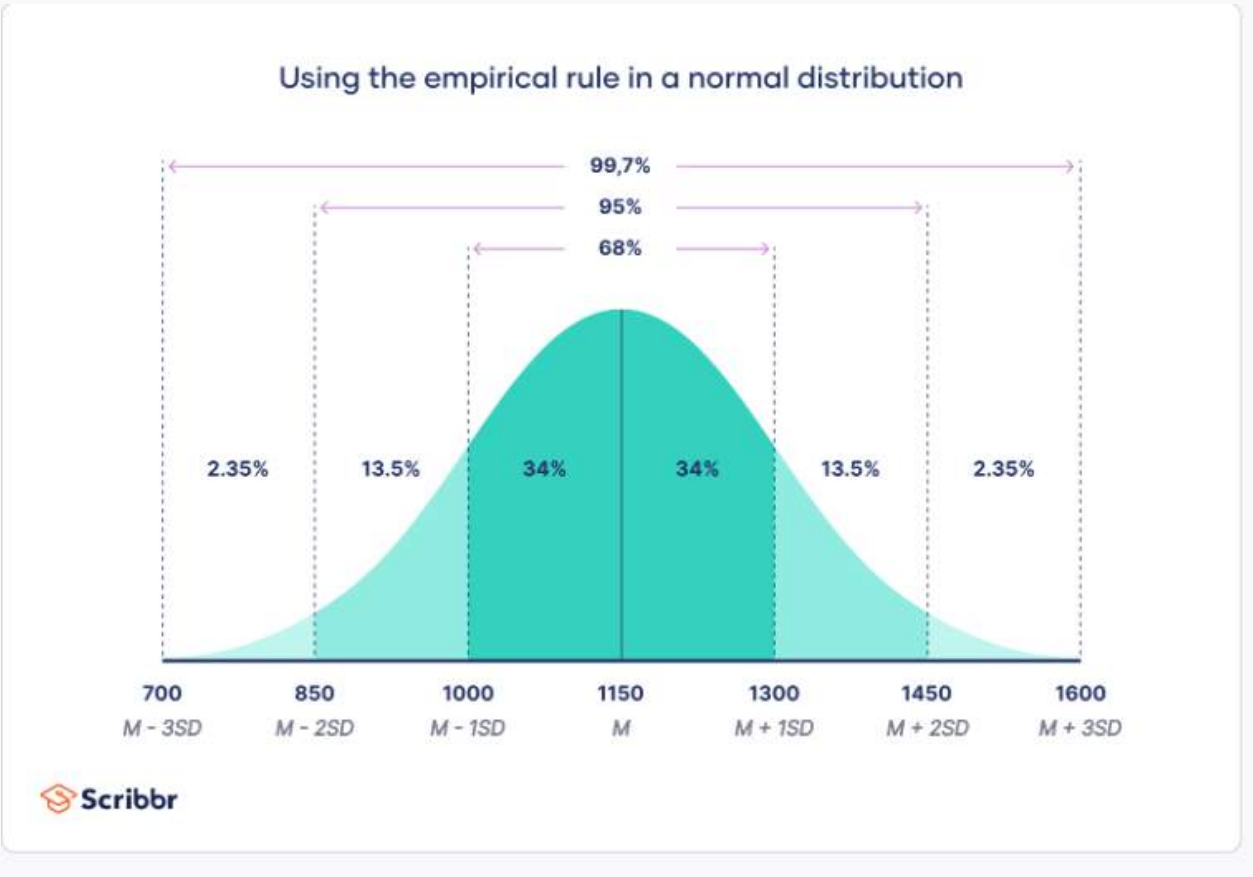
- Around 68% of values are within 1 standard deviation from the mean.
- Around 95% of values are within 2 standard deviations from the mean.
- Around 99.7% of values are within 3 standard deviations from the mean.

Example: Using the empirical rule in a normal distribution

You collect SAT scores from students in a new test preparation course. The data follows a normal distribution with a mean score (*M*) of 1150 and a standard deviation (*SD*) of 150.

Following the empirical rule:

- Around 68% of scores are between 1,000 and 1,300, 1 standard deviation above and below the mean.
- Around 95% of scores are between 850 and 1,450, 2 standard deviations above and below the mean.
- Around 99.7% of scores are between 700 and 1,600, 3 standard deviations above and below the mean.



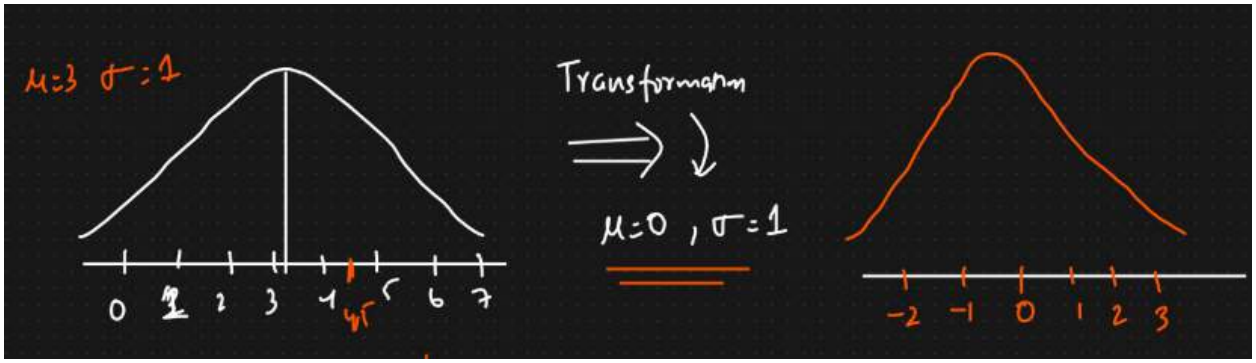
The empirical rule is a quick way to get an overview of your data and check for any outliers or extreme values that don't follow this pattern.

If data from small samples do not closely follow this pattern, then other distributions like the t-distribution may be more appropriate. Once you identify the distribution of your variable, you can apply appropriate statistical tests.

Notation	$\mathcal{N}(\mu, \sigma^2)$
Parameters	$\mu \in \mathbb{R}$ = mean (location) $\sigma^2 \in \mathbb{R}_{>0}$ = variance (squared scale)
Support	$x \in \mathbb{R}$
PDF	$\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$
CDF	$\Phi\left(\frac{x-\mu}{\sigma}\right) = \frac{1}{2}\left[1 + \operatorname{erf}\left(\frac{x-\mu}{\sigma\sqrt{2}}\right)\right]$

Standard Normal Distribution

The **standard normal distribution** is one of the forms of the normal distribution. It occurs when a normal random variable has a mean equal to zero and a standard deviation equal to one. In other words, a normal distribution with a mean 0 and standard deviation of 1 is called the standard normal distribution.



A normal distribution can be converted into a standard normal distribution by Z-Score.

$$\text{Z - Score} = \frac{(X_i - \mu)}{\sigma}$$

For example, Z-Score = (1-3)/1

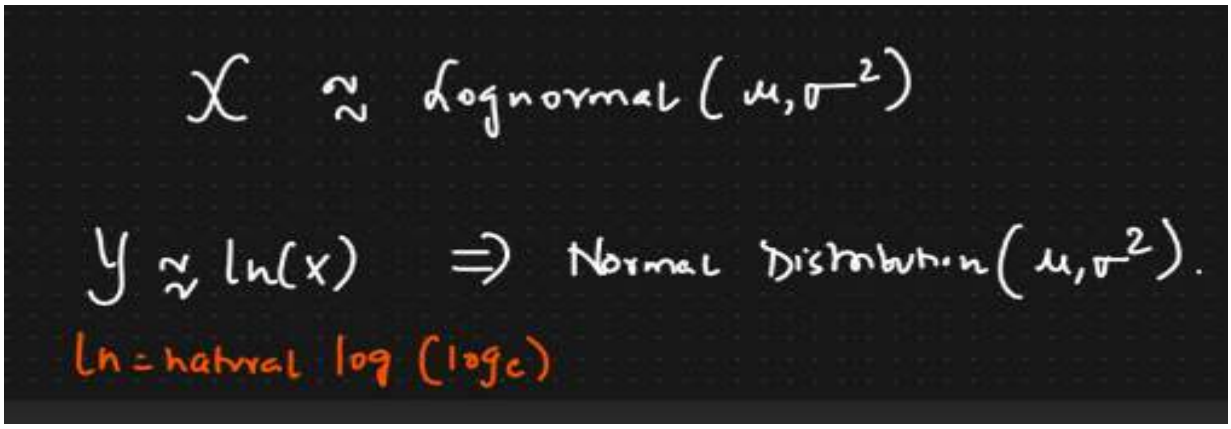
= -2

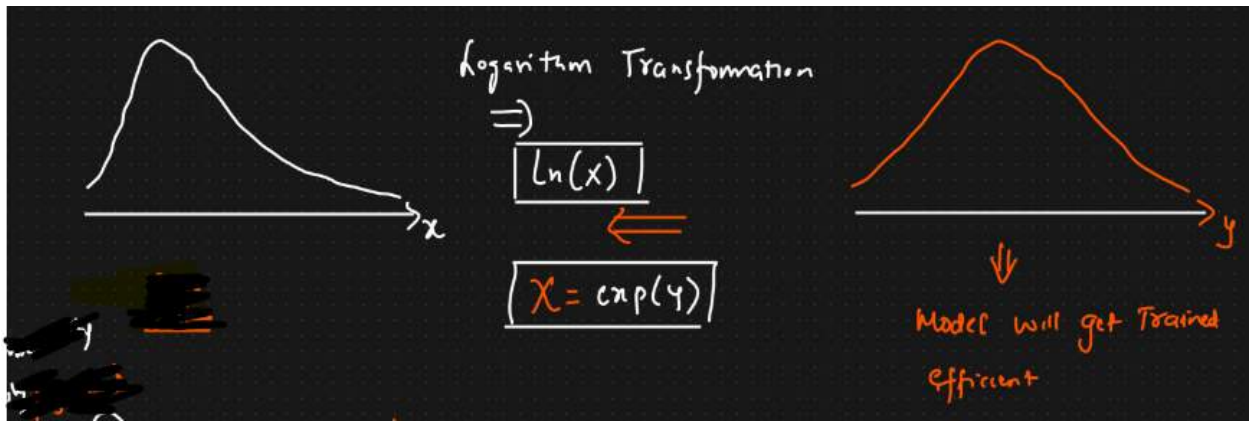
Similarly, -1,0,1,2,3

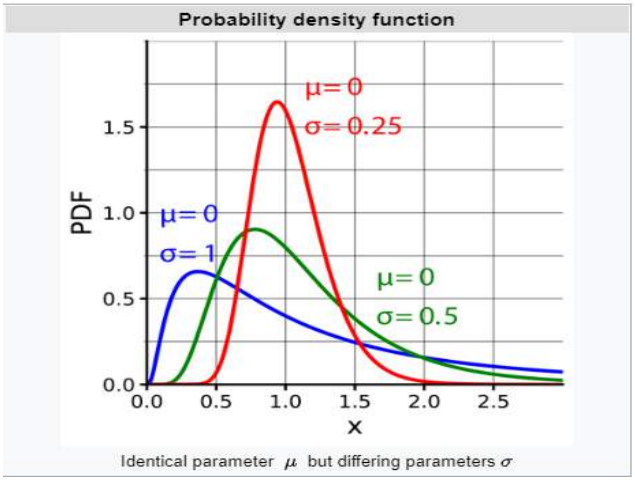
Z-Score tells us about a value, how many standard deviations it is away from the mean.

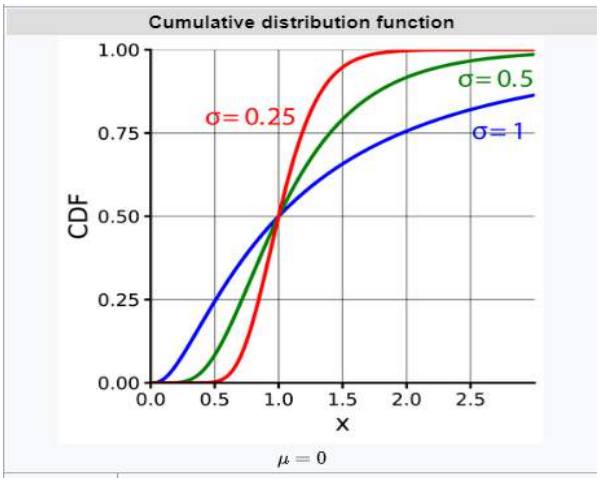
Log Normal Distribution

In probability theory, a log-normal (or lognormal) distribution is a continuous probability distribution of a random variable whose logarithm is normally distributed. Thus, if the random variable X is log-normally distributed, then $Y = \ln(X)$ has a normal distribution. Equivalently, if Y has a normal distribution, then the exponential function of Y, $X = \exp(Y)$, has a log-normal distribution.









Notation	$\operatorname{Lognormal}(\mu, \sigma^2)$
Parameters	$\mu \in (-\infty, +\infty)$ (logarithm of scale), $\sigma > 0$
Support	$x \in (0, +\infty)$
PDF	$\frac{1}{x\sigma\sqrt{2\pi}} \exp\left(-\frac{(\ln x - \mu)^2}{2\sigma^2}\right)$
CDF	$\frac{1}{2}\left[1 + \operatorname{erf}\left(\frac{\ln x - \mu}{\sigma\sqrt{2}}\right)\right] = \Phi\left(\frac{\ln(x) - \mu}{\sigma}\right)$

Examples

- Wealth distribution of the world.
- Salary distribution in a company.

Power Law Distribution

In statistics, a **power law** is a functional relationship between two quantities, where a relative change in one quantity results in a relative change in the other quantity proportional to a power of the change, independent of the initial size of those quantities:

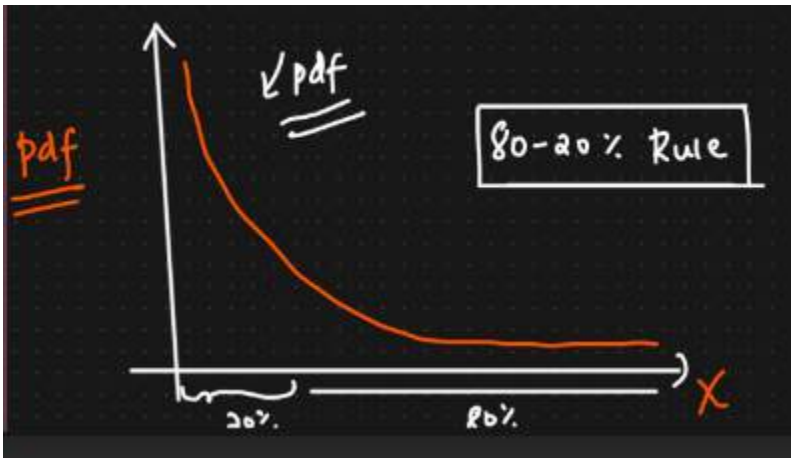
one quantity varies as a power of another. For instance, considering the area of a square in terms of the length of its side, if the length is doubled, the area is multiplied by a factor of four.

A power law is a type of relationship between two quantities. When one quantity changes, the other quantity changes in a way that's proportional to the power of the first quantity.

Here's an easy example: consider a square. If you double the length of a side of the square, the area of the square quadruples. This is because the area of a square is the side length squared (side length x side length). So, if the side length is doubled, the area increases by a factor of four (2 squared is 4).

In a power-law distribution, a small number of items are clustered at the top of a distribution, taking up most of the resources. For instance, there are very few billionaires; the bulk of the population holds very modest nest eggs. This is often referred to as the 80-20 rule, where 20% of the entities hold 80% of the value or power.

Power laws can be observed in a wide variety of phenomena, including the distribution of income, the size of cities according to population, the magnitude of earthquakes, and even word frequencies in languages.



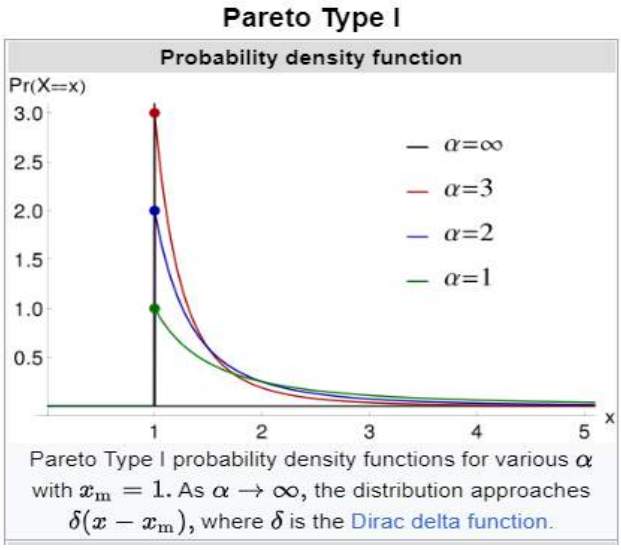
Types Of Power Law Distribution

Pareto Distribution

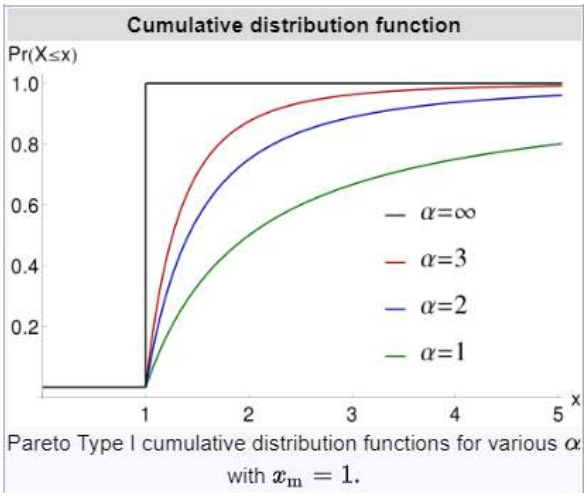
If X is a random variable with a Pareto (Type I) distribution, then the probability that X is greater than some number x, i.e., the survival function (also called tail function), is given by

$$\overline{F}(x) = \Pr(X > x) = \begin{cases} \left(\frac{x_\mathrm{m}}{x}\right)^\alpha & x \geq x_\mathrm{m}, \\ 1 & x < x_\mathrm{m}, \end{cases}$$

where x_m is the (necessarily positive) minimum possible value of X, and α is a positive parameter.



Pareto Type I probability density functions for various α with $x_\mathrm{m} = 1$. As $\alpha \to \infty$, the distribution approaches $\delta(x - x_\mathrm{m})$, where δ is the Dirac delta function.



Pareto Type I cumulative distribution functions for various α with $x_\mathrm{m} = 1$.

Parameters	$x_{\mathrm{m}} > 0$ scale (real) $\alpha > 0$ shape (real)
Support	$x \in [x_{\mathrm{m}}, \infty)$
PDF	$\frac{\alpha x_{\mathrm{m}}^{\alpha}}{x^{\alpha+1}}$
CDF	$1 - \left(\frac{x_{\mathrm{m}}}{x}\right)^{\alpha}$

Exponential Distribution

The exponential distribution is a continuous probability distribution that often concerns the amount of time until some specific event happens. It is a process in which events happen continuously and independently at a constant average rate. The exponential distribution has the key property of being memoryless. The exponential random variable can be either more small values or fewer larger variables. For example, the amount of money spent by the customer on one trip to the supermarket follows an exponential distribution.

Exponential Distribution Formula

The continuous random variable, say X is said to have an exponential distribution, if it has the following probability density function:

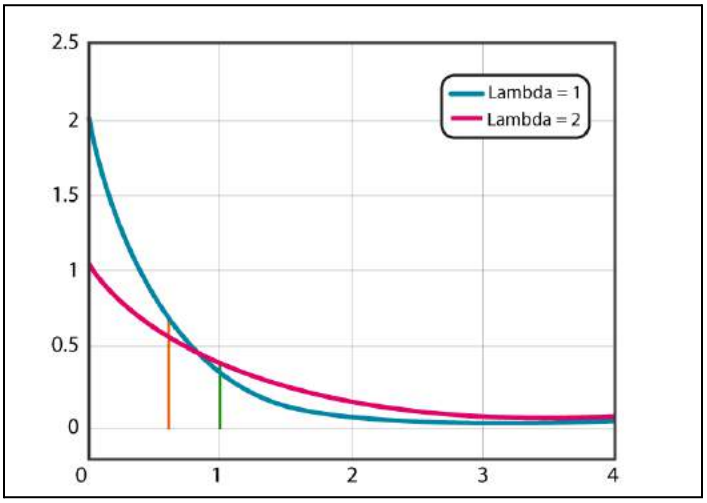
$$f_X(x|\lambda) = \begin{cases} \lambda e^{-\lambda x} & for\ x > 0 \\ 0 & for\ x \leq 0 \end{cases}$$

Where

λ is called the distribution rate.

Exponential Distribution Graph

The exponential distribution graph is a graph of the probability density function which shows the distribution of distance or time taken between events. The two terms used in the exponential distribution graph is lambda (**λ**)and x. Here, lambda represents the events per unit time and x represents the time. The following graph shows the values for λ=1 and λ=2.



Parameters	$\lambda > 0$, rate, or inverse scale
Support	$x \in [0, \infty)$
PDF	$\lambda e^{-\lambda x}$
CDF	$1 - e^{-\lambda x}$

Bernoulli Distribution

A discrete probability distribution wherein the random variable can only have 2 possible outcomes is known as a Bernoulli Distribution. If in a Bernoulli trial the random variable takes on the value of 1, it means that this is a success. The probability of success is

given by p. Similarly, if the value of the random variable is 0, it indicates failure. The probability of failure is q or 1 - p.

Bernoulli Distribution Example

Suppose there is an experiment where you flip a coin that is fair. If the outcome of the flip is heads then you will win. This means that the probability of getting heads is p = 1/2. If X is the random variable following a Bernoulli Distribution, we get P(X = 1) = p = 1/2.

Bernoulli Distribution Formula

A binomial random variable, X, is also known as an indicator variable. This is because if an event results in success then X = 1 and if the outcome is a failure then X = 0. X can be written as X ~ Bernoulli (p), where p is the parameter. The formulas for Bernoulli distribution are given by the probability mass function (pmf) and the cumulative distribution function (CDF).

Probability Mass Function for Bernoulli Distribution

We calculate the probability mass function for a Bernoulli distribution. The probability that a discrete random variable will be exactly equal to some value is given by the probability mass function. The formula for pmf, f, associated with a Bernoulli random variable over possible outcomes 'x' is given as follows:

$$\text{PMF} = f(x, p) = \begin{cases} p & \text{if } x = 1 \\ q = 1 - p & \text{if } x = 0 \end{cases}$$

We can also express this formula as,

$$f(x, p) = p^x (1 - p)^{1 - x}, x \in \{0, 1\}$$

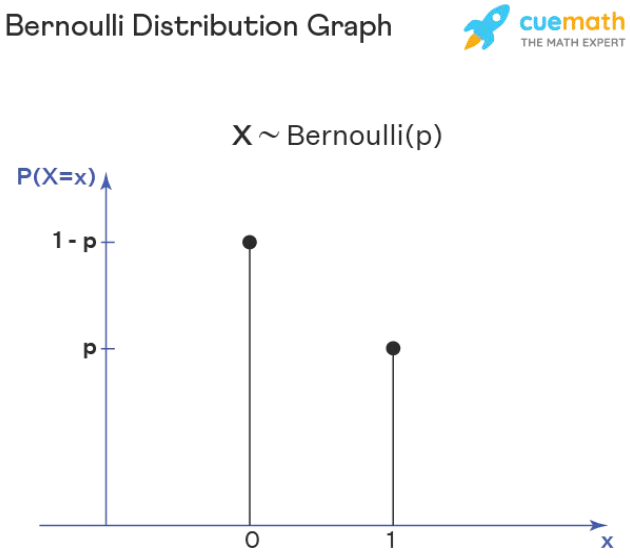
Cumulative Distribution Function for Bernoulli Distribution

The cumulative distribution function of a Bernoulli random variable X when evaluated at x is defined as the probability that X will take a value lesser than or equal to x. The formula is given as follows:

$$\text{CDF} = F(x, p) = \begin{cases} 0 & \text{if } x < 0 \\ 1 - p & \text{if } 0 \le x < 1 \\ 1 & x \ge 1 \end{cases}$$

Bernoulli Distribution Graph

The graph of a Bernoulli distribution helps to get a visual understanding of the probability density function of the Bernoulli random variable.



The graph shows that the probability of success is p when X = 1 and the probability of failure of X is (1 - p) or q if X = 0.

Parameters	$0 \leq p \leq 1$ $q = 1 - p$
Support	$k \in \{0, 1\}$
PMF	$\begin{cases} q = 1 - p & \text{if } k = 0 \\ p & \text{if } k = 1 \end{cases}$
CDF	$\begin{cases} 0 & \text{if } k < 0 \\ 1 - p & \text{if } 0 \leq k < 1 \\ 1 & \text{if } k \geq 1 \end{cases}$

Binomial Distribution

What is Binomial Distribution in Probability?

the **binomial distribution** with parameters *n* and *p* is the discrete probability distribution of the number of successes in a sequence of *n* independent experiments, each asking a yes–no question, and each with its own Boolean-valued outcome: *success* (with probability *p*) or *failure* (with probability *q* = *1-p*). A single success/failure experiment is also called a Bernoulli trial or Bernoulli experiment, and a sequence of outcomes is called a Bernoulli process; for a single trial, i.e., *n* = 1, the binomial distribution is a Bernoulli distribution.

In short, the combination of multiple Bernoulli Distributions is Binomial Distribution.

Binomial Distribution Formula

Binomial Distribution Formula which is used to calculate the probability, for random variable X = 0, 1, 2, 3,….,n is given as

$$P(x:n,p) = {}^{n}C_{x}\, p^{x}\, (1-p)^{n-x} \textbf{ Or } P(x:n,p) = {}^{n}C_{X}\, p^{x}\, (q)^{n-x}$$

where,

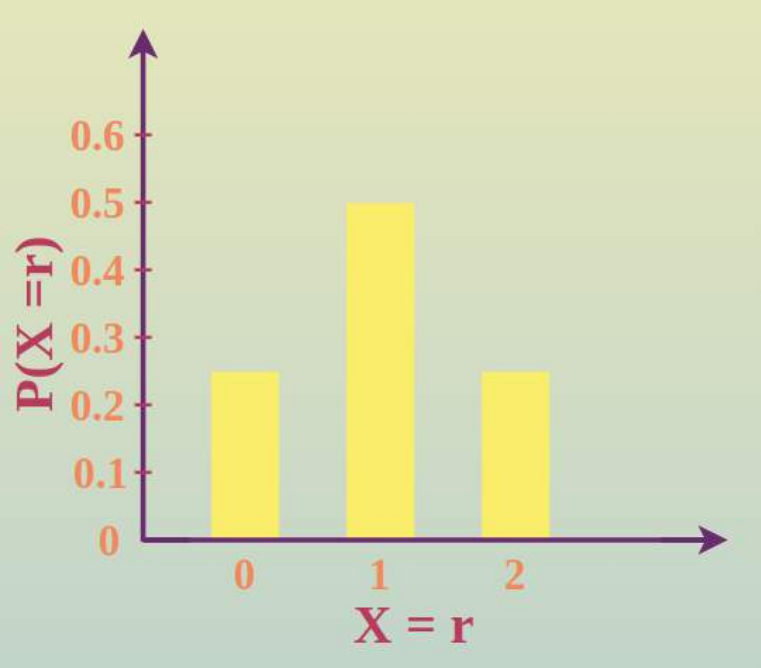
- *n = the number of experiments*
- *x = 0, 1, 2, 3, 4, …*
- *p = Probability of success in a single experiment*
- *q = Probability of failure in a single experiment (= 1 – p)*

Binomial Distribution Graph

The Binomial Distribution Graph is plotted for X and P(X). We will plot a Binomial Distribution Graph for tossing a coin twice where getting the head is a success. If we toss a coin twice, the possible outcomes are {HH, HT, TH, TT}. The binomial Distribution Table for this is given below:

X (Random Variable)	**P(X)**
X = 0 (Getting no head)	P(X = 0) = 1/4 = 0.25
X = 1 (Getting 1 head)	P(X = 1) = 2/4 = 1/2 = 0.5
X = 2 (Getting two heads)	P(X = 2) = 1/4 = 0.25

The Binomial Distribution Graph for the above table is given below:



Poisson Distribution

Poisson distribution is used to model the number of events that occur in a fixed interval of time or space, given the average rate of occurrence, assuming that the events happen independently and at a constant rate.

Poisson Distribution Formula

Let X be a discrete random variable that can assume values 0, 1, 2 . . . then, the probability function Poisson distribution of X is given as:

$$\mathbf{f(x) = P(X = x) = \frac{\lambda^x e^{-\lambda}}{x!}}$$

Where,

- ***P(X = x)*** *is the probability that an event will occur x times,*
- ***X*** *is a random variable following a Poisson distribution,*
- ***λ*** *is the average number of times an event occurs,*
- ***x*** *is the number of times an event occurs, and*
- ***e*** *is Euler's constant (≈2.718).*

Poisson Distribution with Example

To understand Poisson Distribution, let's consider an example.

Suppose there is a bakery on the corner of the street and on average 10 customers arrive at the bakery per hour. For this case, we can calculate the probabilities of different numbers of customers arriving at the bakery at any hour using the Poisson distribution. As probability mass function (PMF) or Poisson Distribution Formula is given as:

$$P(X = k) = \lambda^k e^{-\lambda}/k!$$

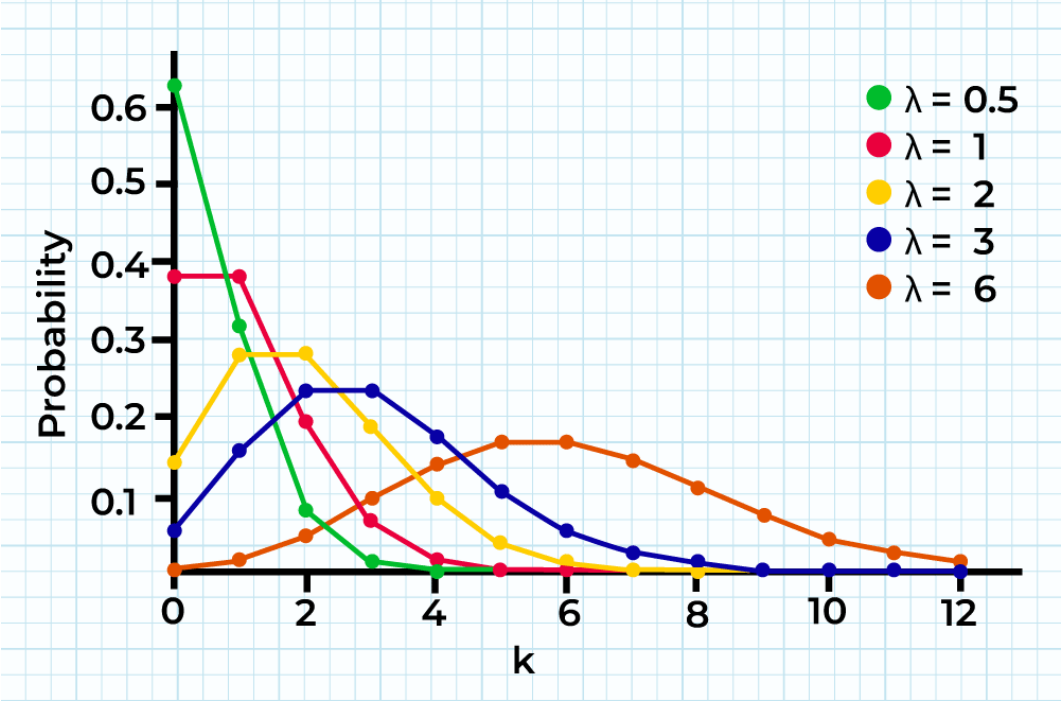
Probability of having exactly 5 customers arrive in an hour:
$P(X = 5) = 10^5 \times e^{-10}/5! \approx 0.037$

Probability of having no customers arrive in an hour:
P(X = 0) = 100 × e-10/0! ≈ 4.54 × 10-5

Probability of having at least 15 customers arrive in an hour (sum of probabilities from 15 to infinity):
$P(X \geq 15) = 1 - P(X < 15) = 1 - (P(X = 0) + P(X = 1) + \ldots + P(X = 14))$

Poisson Distribution Graph

The following illustration shows the Graph of the Poisson Distribution or Poisson Distribution Curve.



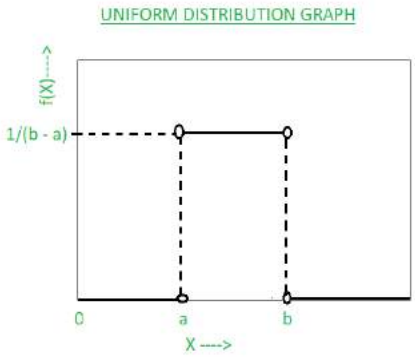
Notation	$\mathrm{Pois}(\lambda)$
Parameters	$\lambda \in (0, \infty)$ (rate)
Support	$k \in \mathbb{N}_0$ (Natural numbers starting from 0)
PMF	$\frac{\lambda^k e^{-\lambda}}{k!}$
CDF	$\frac{\Gamma(\lfloor k+1 \rfloor, \lambda)}{\lfloor k \rfloor !}$, or $e^{-\lambda} \sum_{j=0}^{\lfloor k \rfloor} \frac{\lambda^j}{j!}$, or $Q(\lfloor k+1 \rfloor, \lambda)$ (for $k \geq 0$, where $\Gamma(x, y)$ is the upper incomplete gamma function, $\lfloor k \rfloor$ is the floor function, and Q is the regularized gamma function)

Uniform Distribution

A uniform distribution is a distribution that has constant probability due to equally likely occurring events. It is also known as rectangular distribution (continuous uniform distribution). It has two parameters a and b: a = minimum and b = maximum. The distribution is written as U(a, b). Types of uniform distribution are:

1. **Continuous Uniform Distribution:** A continuous uniform probability distribution is a distribution that has an infinite number of values defined in a specified range. It has a rectangular-shaped graph so-called rectangular distribution. It works on the values which are continuous in nature. Example: Random number generator
2. **Discrete Uniform Distribution:** A discrete uniform probability distribution is a distribution that has a finite number of values defined in a specified range. Its graph contains various vertical lines for each finite value. It works on values that are discrete in nature. Example: A dice is rolled.

Graph of Uniform Distribution:



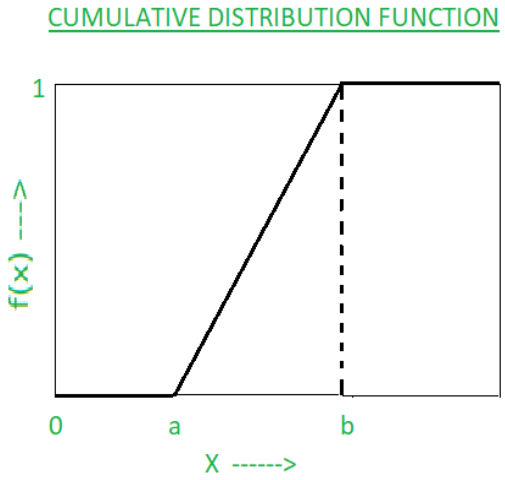
Calculating the height of the rectangle:

The maximum probability of the variable X is 1 so the total area of the rectangle must be 1.

Area of rectangle = base * height = 1

(b – a) * f(x) = 1

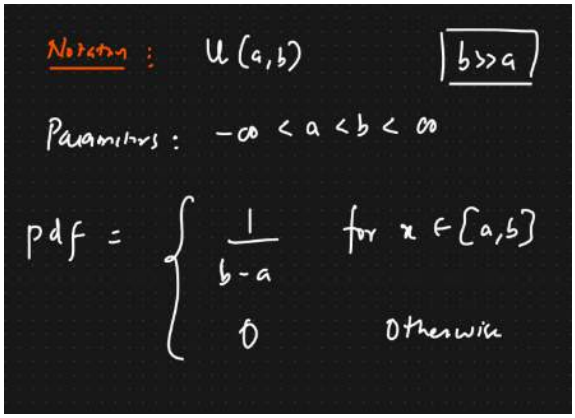
f(x) = 1/(b – a) = height of the rectangle



Note: Discrete uniform distribution: **Px = 1/n.** Where, Px = Probability of a discrete variable, n = Number of values in the range

Uniform Distribution Formula

A random variable X is said to be uniformly distributed over the interval $-\infty < a < b < \infty$.



Hypothesis Testing

Hypothesis Testing Definition

Hypothesis testing can be defined as a statistical tool that is used to identify if the results of an experiment are meaningful or not. It involves setting up a null hypothesis and an alternative hypothesis. These two hypotheses will always be mutually exclusive. This means that if the null hypothesis is true then the alternative hypothesis is false and vice versa. An example of hypothesis testing is setting up a test to check if a new medicine works on a disease in a more efficient manner.

Null Hypothesis

The null hypothesis is a concise mathematical statement that is used to indicate that there is no difference between two possibilities. In other words, there is no difference between certain characteristics of data. This hypothesis assumes that the outcomes of an experiment are based on chance alone. It is denoted as H_0. Hypothesis testing is used to conclude if the null hypothesis can be rejected or not. Suppose an experiment is

conducted to check if girls are shorter than boys at the age of 5. The null hypothesis will say that they are the same height.

Alternative Hypothesis

The alternative hypothesis is an alternative to the null hypothesis. It is used to show that the observations of an experiment are due to some real effect. It indicates that there is a statistical significance between two possible outcomes and can be denoted as H_1 or H_a. For the above-mentioned example, the alternative hypothesis would be that girls are shorter than boys at the age of 5.

Hypothesis Testing P Value

In hypothesis testing, the p value is used to indicate whether the results obtained after conducting a test are statistically significant or not. It also indicates the probability of making an error in rejecting or not rejecting the null hypothesis.This value is always a number between 0 and 1. The p value is compared to an alpha level, α or significance level. The alpha level can be defined as the acceptable risk of incorrectly rejecting the null hypothesis. The alpha level is usually chosen between 1% to 5%.

Hypothesis Testing Critical region

All sets of values that lead to rejecting the null hypothesis lie in the critical region. Furthermore, the value that separates the critical region from the non-critical region is known as the critical value.

You can read more about hypothesis testing here

Central Limit Theorem Statement

The central limit theorem relies on the concept of sampling distribution, which is the probability distribution of a statistic for a large number of samples taken from a population.

The central limit theorem says that the sampling distribution of the mean will always be normally distributed, as long as the sample size is large enough. Regardless of whether the population has a normal, Poisson, binomial or any other distribution, the sampling distribution of mean will be normal.

Read more about Central Limit Theorem here.

Standard Error

In statistics, the **standard error** is the standard deviation of the sample distribution. The sample mean of a data is generally varied from the actual population mean. It is represented as SE. It is used to measure the amount of accuracy by which the given sample represents its population.

Standard Error Meaning

The standard error is one of the mathematical tools used in statistics to estimate the variability. It is abbreviated as SE. The standard error of a statistic or an estimate of a parameter is the standard deviation of its sampling distribution. We can define it as an estimate of that standard deviation.

Standard Error Formula

The accuracy of a sample that describes a population is identified through the SE formula. The sample mean which deviates from the given population and that deviation is given as;

$$SE_{\bar{x}} = \frac{S}{\sqrt{n}}$$

Where S is the standard deviation and n is the number of observations.

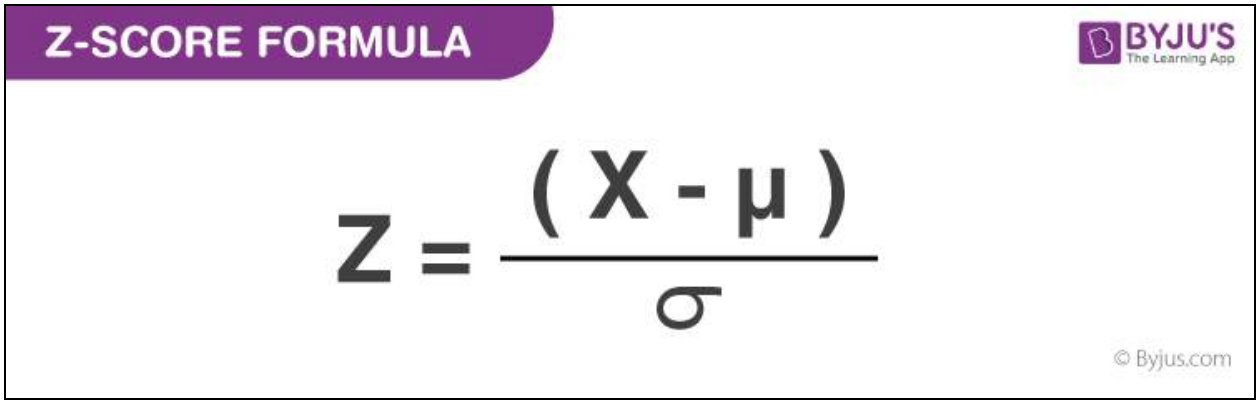
Z - Score

Z-Score, also known as the standard score, indicates how many standard deviations an entity is, from the mean.

Z-Score Formula

It is a way to compare the results from a test to a “normal” population.

If X is a random variable from a normal distribution with mean (μ) and standard deviation (σ), its Z-score may be calculated by subtracting mean from X and dividing the whole by standard deviation.



Where, x = test value

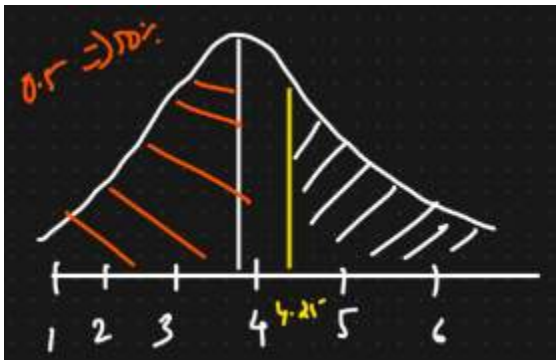
μ is mean and

σ is SD (Standard Deviation)

For the average of a sample from a population ‘n’, the mean is μ and the standard deviation is σ.

Know more about Z-Score here.

Example 1: Consider a random variable X which belongs to a normal distribution with a mean 4 and standard deviation 1, calculate the area above point 4.5.



Solution: Here x = 4.5, μ = 4 and σ = 1.

Z-score = 4.5 - 4/1 = 0.25

4.5 is 0.25 standard deviations away from the mean.

In the z table search for 0.2 and 0.05 the value we get will be 0.59871.

Using the first two digits of the z score, determine the row containing these digits of the z table. Now using the 2nd digit after the decimal, find the corresponding column. The intersection of this row and column will give a value.

STANDARD NORMAL DISTRIBUTION: Table Values Represent AREA to the LEFT of the Z score.

Z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.50000	.50399	.50798	.51197	.51595	.51994	.52392	.52790	.53188	.53586
0.1	.53983	.54380	.54776	.55172	.55567	.55962	.56356	.56749	.57142	.57535
0.2	.57926	.58317	.58706	.59095	.59483	.59871	.60257	.60642	.61026	.61409

So the area above point 4.5 will be

1 - 0.59871 = 0.4013

So, 40.13% area lies above point 4.5.

Example 2: What percentage of score lies between 3.5 to 4.5?(Take standard deviation and mean values from above example)

Solution: z-score(min) = 3.5 - 4/1 = -0.5.

The value at -0.5 in the z table is 0.3085.

STANDARD NORMAL DISTRIBUTION: Table Values Represent AREA to the LEFT of the Z score.

Z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
-0.9	.18406	.18141	.17879	.17619	.17361	.17106	.16853	.16602	.16354	.16109
-0.8	.21186	.20897	.20611	.20327	.20045	.19766	.19489	.19215	.18943	.18673
-0.7	.24196	.23885	.23576	.23270	.22965	.22663	.22363	.22065	.21770	.21476
-0.6	.27425	.27093	.26763	.26435	.26109	.25785	.25463	.25143	.24825	.24510
-0.5	.30854	.30503	.30153	.29806	.29460	.29116	.28774	.28434	.28096	.27760

(Here I have taken some cropped parts of the z table).

Similarly, z-score(max) = 4.5 - 4/1 = 0.5

The value at 0.5 in the z table is 0.6915.

Therefore the percentage of score that lies between 3.5 to 4.5 is

0.6915 - 0.3085 = 0.383 = 38.3%.

Z Test:

A z test is conducted on a population that follows a normal distribution with independent data points and has a sample size that is greater than or equal to 30. It is used to check whether the means of two populations are equal to each other when the population variance is known. The null hypothesis of a z test can be rejected if the z test statistic is statistically significant when compared with the critical value.

Let’s understand the Z Test with an example.

The average height of all students in a city is 168cm with a σ of 13.9. A doctor believes mean to be different. He measured heights of 36 individuals and found the average height to be 169.5cm.

a) State null and alternate hypothesis.
b) At a 95% confidence level, is there enough evidence to reject the null hypothesis?

Solution: Given,

μ = 168cm

σ = 13.9

$\bar{x}$ = 169.5cm

n = 36

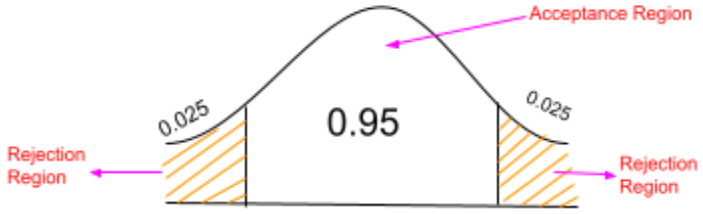
C.I = 0.95

α = 1 - C.I = 0.05

1) Null Hypothesis Ho : μ = 168cm

   Alternate Hypothesis H_1 = 169.5cm

2) Decision boundary based on confidence interval



We are checking alternate hypothesis for $\mu \neq$ 168cm. So our value can be greater than 168 or less than 168. Also that means it can come in left/right extreme ends.

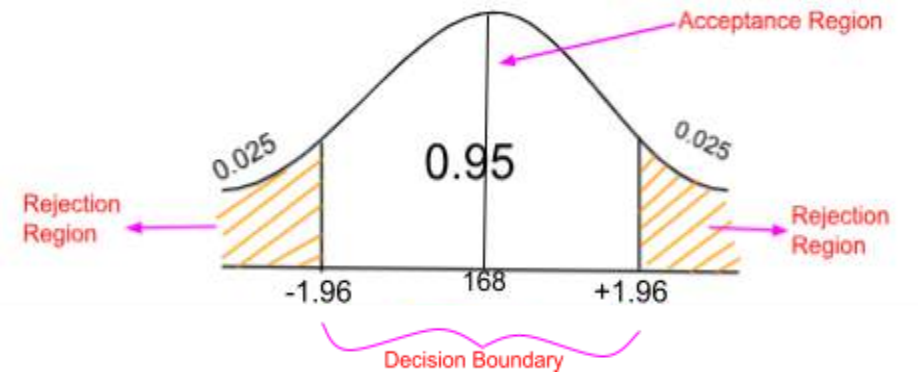
So these types of tests are called Two Tailed Tests. This Z -Test follows a Gaussian Distribution.

Now find how much far both the lines in the above figure are away from the mean.

Total area = 1

1 - 0.025 = 0.9750.

Find value 0.9750 in the Z table, we get 1.96 as the Z-Score. As both the orange areas in the above figure are symmetrical, the point to left will be -1.96 standard deviations away from the mean.



If Z test is less than -1.96 or greater than +1.96 then we reject the null hypothesis.

Whenever we take sampling our Z test formula will be

$$Z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

$$= \frac{169.5 - 168}{3.9/\sqrt{36}}$$

$$= 2.31$$

As 2.31 > 1.96, we reject the null hypothesis. Therefore the doctor is right about mean being different.

Example One-Tailed Test:

A school claimed that the students who study are more intelligent than the average school. On calculating the IQ scores of 50 students, the average turns out to be 110. The mean of the population IQ is 100 and the standard deviation is 15. State whether the claim of the principal is right or not at a 5% significance level.

- First, we define the null hypothesis and the alternate hypothesis. Our null hypothesis will be:

  $$H_0 : \mu = 100$$

  and our alternate hypothesis.

  $$H_A : \mu > 100$$

- State the level of significance. Here, our level of significance is given in this question ($\propto$ =0.05), if not given then we take $\propto$=0.05 in general.
- Now, we compute the Z-Score:

  X = 110

  Mean = 100

  Standard Deviation = 15

  Number of samples = 50

  $$\begin{aligned} \text{Z-Score} &= \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \\ &= \frac{110 - 100}{15/\sqrt{50}} \\ &= \frac{10}{2.12} \\ &= 4.71 \end{aligned}$$

- Now, we look up to the z-table. For the value of $\propto$=0.05, the z-score for the right-tailed test is 1.645.
- Here 4.71 >1.645, so we reject the null hypothesis.
- If the z-test statistics are less than the z-score, then we will not reject the null hypothesis.

T - Test:

A t-test is an inferential statistic used to determine if there is a significant difference between the means of two groups and how they are related. T-tests are used when the data sets follow a normal distribution and have unknown variances, like the data set recorded from flipping a coin 100 times.

The t-test is a test used for hypothesis testing in statistics and uses the t-statistic, the t-distribution values, and the degrees of freedom to determine statistical significance.

Degrees of freedom are the maximum number of logically independent values, which may vary in a data sample. Degrees of freedom are calculated by subtracting one from the number of items within the data sample.

Imagine you and four friends are going to a movie and you all decide to share a large popcorn. The popcorn costs Rs10. If four of your friends have already contributed their parts and you're the last one to contribute, the amount you'll contribute isn't really up to you. It's whatever is left to make the total Rs10.

In this scenario, the "degrees of freedom" is four, because four people had the freedom to give any amount they wanted. But the last person (the fifth one) doesn't have that freedom because their contribution is determined by what the others have given.

In statistics, "degrees of freedom" is similar. It's the number of values in a calculation that are free to vary without violating the constraints of the calculation.

Example of T test: In the population the average IQ is 100. A team of researchers wants to test a new medication to see if it has a positive or negative effect or no effect at all. A sample of 30 participants who have taken the medication has a mean of 140 with a standard deviation of 20. Did the medication affect intelligence?

Solution: Given,

μ = 100

s = 20

$\bar{x}$ = 140

n = 30

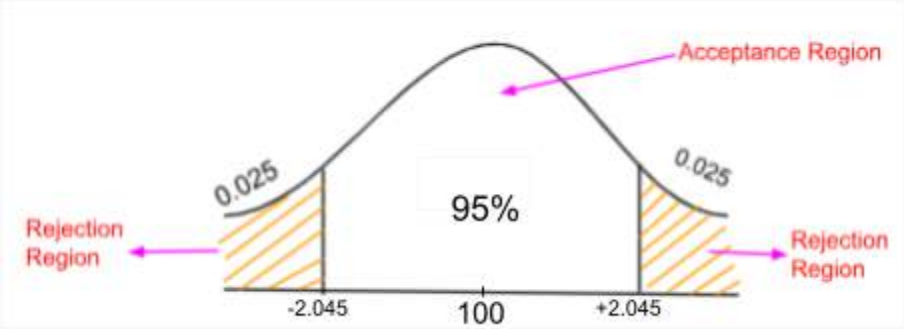
C.I = 95%

α = 1 - C.I = 0.05

1) Null Hypothesis Ho : μ = 168cm

   Alternate Hypothesis H_1 = 169.5cm

2) Degree of freedom: n-1 = 30-1 = 29
3) Decision rule:



Here we need to use T table. Find the row that corresponds to 29 degrees of freedom and the column that corresponds to the 0.05 (as it is a two tailed problem) significance level.

t Table

cum. prob	$t_{.50}$	$t_{.75}$	$t_{.80}$	$t_{.85}$	$t_{.90}$	$t_{.95}$	$t_{.975}$	$t_{.99}$	$t_{.995}$	$t_{.999}$	$t_{.9995}$
one-tail	0.50	0.25	0.20	0.15	0.10	0.05	0.025	0.01	0.005	0.001	0.0005
two-tails	1.00	0.50	0.40	0.30	0.20	0.10	0.05	0.02	0.01	0.002	0.001
df											
26	0.000	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	0.000	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	0.000	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	0.000	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	0.000	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.385	3.646

If the T test is less than -2.045 or greater than +2.045 then reject the null hypothesis.

T test statistics:

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

$$= \frac{140 - 100}{20/\sqrt{30}}$$

= 10.96

t > 2.045. Therefore, we reject the null hypothesis.

Conclusion: Medication has a positive effect on intelligence.

Point Estimate:

A point estimate definition is a calculation where a sample statistic is used to estimate or approximate an unknown population parameter. For example, the average

height of a random sample can be used to estimate the average height of a larger population. Here we can say that sample mean is point estimate of population mean.

We rarely know if our point estimate is correct because it is an estimation of the actual value.

We construct a confidence interval to help estimate what the actual value of population parameter is i.e., we specify a range. The formula to construct confidence interval is:

$$\text{Point Estimate} \pm \text{Margin of Error}$$

$$\text{Lower range of C.I.} = \text{Point Estimate} - \text{Margin of Error}$$

$$\text{Higher range of C.I.} = \text{Point Estimate} + \text{Margin of Error}$$

Example: In the verbal section of CAT Exam, a sample of 25 test takers has a mean of 520. With a standard deviation of 80. Construct a 95% C.I. about the mean.

Solution: s = 80, $\bar{x}$ = 520, n = 25, C.I = 95%, α = 1 - C.I = 0.05

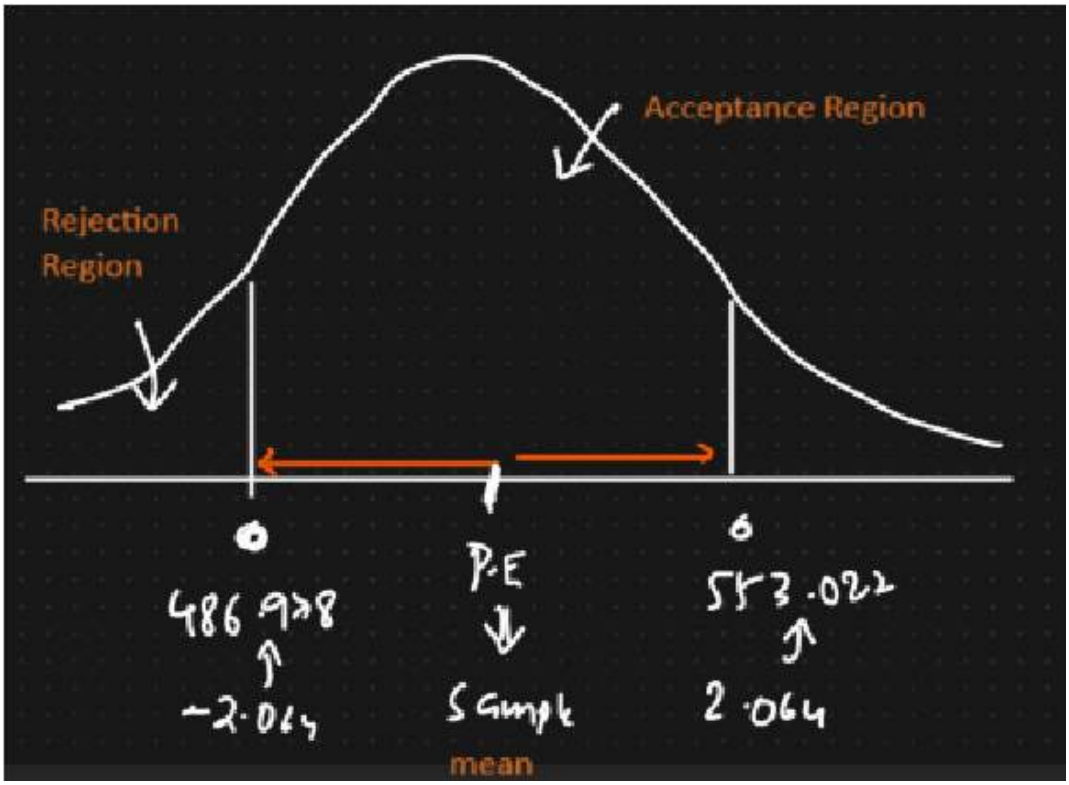
C.I. = Point Estimate $\pm$ Margin of Error

$$= \bar{x} \pm t_{\alpha/2}\left(\frac{s}{\sqrt{n}}\right)$$

$$= 520 \pm 2.064\left(\frac{80}{\sqrt{25}}\right)$$ (Here 2.064 is calculated using t table as the degree of freedom is 24 and C.I. is 0.05).

Lower C.I. = 520 - 2.064(80/5) = 486.971

Higher C.I. = 520 + 2.064(80/5) = 553.022



Chi Square Test

A **chi-squared test** (symbolically represented as χ^2) is basically a data analysis on the basis of observations of a random set of variables. Usually, it is a comparison of two statistical data sets.

Note: Chi-squared test is applicable only for categorical data, such as men and women falling under the categories of Gender, Age, Height, etc.

Example: In the 2010 census of the city, the weight of the individuals in a small city were found to be the following.

<50kg	50-75kg	>75kg
20%	30%	50%

In 2010, weights of n = 500 individuals were sampled. Below are the results.

<50kg	50-75kg	>75kg
140	160	200

Using α = 0.05, would you conclude the population differences of weights have changed in the last 10 years?

Solution: In 2010 expected,

<50kg	50-75kg	>75kg
20%	30%	50%

In 2020 observed,

<50kg	50-75kg	>75kg
140	160	200

In 2010 expected,

<50kg	50-75kg	>75kg
500 * 0.2 = 100	500 * 0.3 = 150	500 * 0.5 = 250

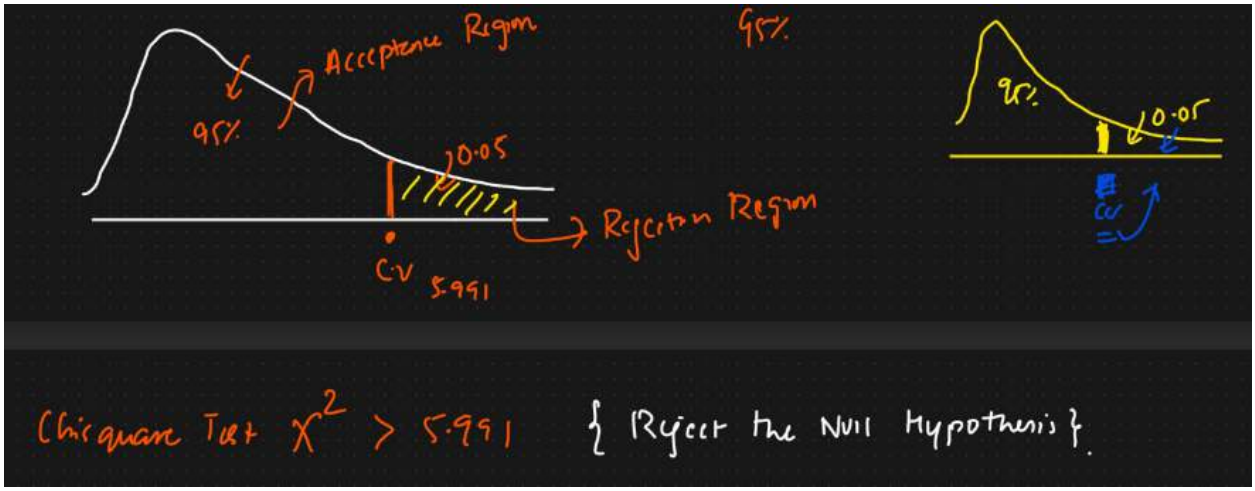
1) Null Hypothesis Ho : The data meets the expectation

   Alternate Hypothesis H_1: The data does not meet the expectation

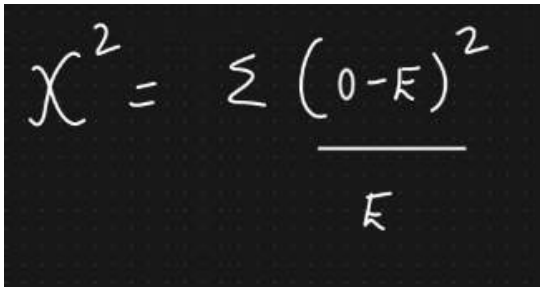
2) Degree of freedom: k-1 = 3-1 = 5

3) α = 0.05 , C.I. = 95%

4) Decision Boundary



5) Calculate Chi Square test



$$= \frac{(140-100)^2}{100} + \frac{(160-150)^2}{100} + \frac{(200-250)^2}{100}$$

= 26.67

Identify the Degrees of Freedom: The degrees of freedom for the Chi-Square test is usually denoted as ‘df’ or ‘k’. It depends on the nature of the test.

Identify the Alpha Level: The alpha level for the test is the probability of rejecting the null hypothesis when it is true.

Find the Critical Value: Locate the row in the table that corresponds to the degrees of freedom and the column that corresponds to the alpha level. The intersection of this row and column is the critical value.

Critical values of chi-square (right tail)

Significance level (α)

Degrees of freedom (*df*)	.99	.975	.95	.9	.1	.05	.025	.01
1	-------	0.001	0.004	0.016	2.706	3.841	5.024	6.635
2	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210
3	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345

$x^2 > 5.991.$

Since the chi square test result is greater than the critical value, we reject the null hypothesis.

F Distribution

The *F*-distribution with d_1 and d_2 degrees of freedom is the distribution of

$$X = \frac{S_1/d_1}{S_2/d_2}$$

where S_1 and S_2 are independent random variables with chi-square distributions with respective degrees of freedom d_1 and d_2.

F Test

F test can be defined as a test that uses the f test statistic to check whether the variances of two samples (or populations) are equal to the same value. To conduct an f test, the population should follow an f distribution and the samples must be independent events. On conducting the hypothesis test, if the results of the f test are statistically significant then the null hypothesis can be rejected otherwise it cannot be rejected.

F test is also called the Variance Ratio Test.

Example: The following data shows the number of bulbs produced daily by two workers A and B.

A	B
40	39
30	38
38	41
41	33
38	32
35	39
	40
	34

Can we consider based on the data worker B is more stable and efficient or not. α = 0.05, C.I = 95%.

Solution:

Null Hypothesis: $H_0 : \sigma_1^2 = \sigma_2^2$

Alternate Hypothesis: $H_1 : \sigma_1^2 \neq \sigma_2^2$

Calculating sample variance

For Worker A

X_1	$\bar{x}$	$X_1 - \bar{x}$
40	37	9
30	37	49
38	37	1
41	37	16
38	37	1
35	37	4
		$\sum_{i=1}^{n} (X_1 - \bar{x})^2 = 80$

For Worker B

X_1	$\bar{x}$	$X_1 - \bar{x}$
39	37	4
38	37	1
41	37	16
33	37	16
32	37	25
39	37	4

40	37	9
38	37	9
		$\sum_{i=1}^{n} (X_2 - \bar{x})^2 = 84$

Sample variance of A is

$$S_1^2 = \sum_{i=1}^{n} \frac{\left(x_i - \bar{x}\right)^2}{n-1}$$

= 80/5

= 16

Similarly, S_2^2 = 84/7 = 12

F test = S_1^2 / S_2^2

= 16/12

= 1.33

Degree of freedom df_1 = 6-1 = 5

Degree of freedom df_2 = 8-1 = 7

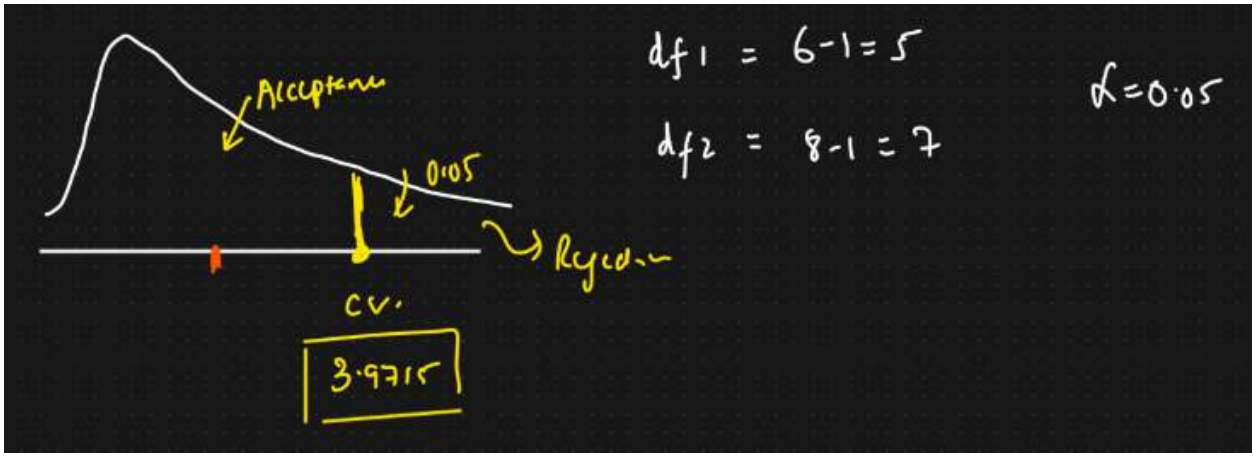
Search for df_1 in column and df_2 in rows to get the critical value for 0.05 significance level. The intersection of this row and column is the critical value.

Critical values of F for the 0.05 significance level:

	1	2	3	4	5	6	7	8	9	10
1	161.45	199.50	215.71	224.58	230.16	233.99	236.77	238.88	240.54	241.88
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.39	19.40
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64

Critical Value = 3.97

Decision Boundary



F test > 3.97

= 1.33 < 3.97 . We failed to reject the null hypothesis.