Types of Statistics

Descriptive Statistics:

Descriptive statistics focuses on summarizing and describing data sets. It involves collecting, organizing, and presenting data in a meaningful way. Descriptive statistics provide simple summaries about the sample or population being studied, without making any inferences or generalizations beyond the data itself.

Real-time example: Let's say we want to understand the average height of students in a class. We measure the height of each student and calculate the mean height. This mean height is a descriptive statistic because it provides a summary of the data collected (the heights of the students) without making any broader claims about the entire population of students.

Inferential Statistics:

Inferential statistics involves using sample data to make inferences or draw conclusions about a larger population. It allows us to make predictions, test hypotheses, and generalize findings based on the data collected from a subset of the population.

Real-time example: Suppose we want to know whether a new teaching method improves student performance. We randomly select two groups of students: one group receives the new teaching method, while the other group follows the traditional method. After conducting an experiment and collecting data on their performance, we can use inferential statistics to determine if there is a significant difference in the average performance between the two groups. By analysing the sample data, we can make inferences about the larger population of students.

COMMON STATISTICAL TERMS

- **Population: In statistics, a population refers to the entire group of individuals, objects, or events that we want to study. It represents the larger set from which we collect data.**

  Real-time example: Let's say we want to determine the average income of all adults in a country. The population in this case would be all the adults living in that country.

- **Sample: A sample is a subset of the population that is selected for analysis. It represents a smaller group of individuals or observations from which we collect data.**

  Real-time example: To estimate the average income of all adults in the country mentioned earlier, we can randomly select a sample of, let's say, 1,000 adults. We would collect income data from this sample and use it to make inferences about the entire population.

- **Parameter: A parameter is a numerical summary or characteristic of a population. It represents a fixed value that describes the entire population.**

  Real-time example: In the context of the country's population, the average income of all adults would be a parameter. It provides a fixed value that represents the average income of the entire population.

- **Statistic: A statistic is a numerical summary or characteristic of a sample. It represents a computed value based on the data collected from the sample.**

  Real-time example: In our example, the average income of the 1,000 adults in the sample would be a statistic. It provides an estimate of the average income of the entire population based on the data from the selected sample.

FREQUENCY DISTRIBUTION

Frequency distribution is a way to organize and summarize data by displaying the frequency, or count, of each unique value or range of values in a dataset. It helps us understand the distribution and patterns within the data.

Real-time example: Let's say we have a dataset that represents the ages of students in a classroom. The dataset consists of the following ages: 12, 14, 13, 12, 15, 14, 13, 12, 14, 13, 15, 14, 13, 12.

To create a frequency distribution, we need to determine how many times each unique age appears in the dataset. We can then organize this information into a table, where one column represents the unique ages, and another column represents the frequency (count) of each age.

In this case, the frequency distribution table might look like this:

Age	Frequency
12	4
13	4
14	4
15	2

CENTRAL TENDENCY

Central tendency is a statistical measure that represents the typical or central value of a dataset. It provides a summary of the data by identifying a single value around which the data tends to cluster.

Real-time example: Let's consider a dataset representing the daily commute times (in minutes) of 10 individuals: 25, 30, 35, 20, 30, 35, 25, 30, 40, 30.

There are three common measures of central tendency: the mean, the median, and the mode.

- **Mean: The mean, also known as the average, is the most commonly used measure of central tendency. It is calculated by summing up all the values in the dataset and dividing the sum by the total number of observations.**

  For our example dataset, the mean commute time would be calculated as follows:

  (25 + 30 + 35 + 20 + 30 + 35 + 25 + 30 + 40 + 30) / 10 = 30

  So, the mean commute time is 30 minutes.

- **Median: The median is the middle value in a dataset when the values are arranged in ascending or descending order. If there is an even number of observations, the median is calculated as the average of the two middle values.**

  For our example dataset, the median commute time would be found by arranging the values in ascending order: 20, 25, 25, 30, 30, 30, 35, 35, 40.

  Since there are 10 observations, the median is the average of the fifth and sixth values: (30 + 30) / 2 = 30.

  So, the median commute time is also 30 minutes.

- **Mode: The mode is the value that appears most frequently in a dataset. It represents the peak or the most common value.**

  For our example dataset, the mode commute time is 30 minutes, as it appears the most number of times (3 times) compared to any other value.

These measures of central tendency provide different perspectives on the typical value of a dataset. The mean is influenced by extreme values, the median is less affected by extreme values, and the mode represents the most frequent value.

Understanding central tendency helps us summarize and describe datasets, making it easier to interpret and draw conclusions from the data.

MEASURE OF DISPERSION

Measure of dispersion, also known as variability or spread, is a statistical measure that quantifies the extent to which data points in a dataset vary or deviate from the central tendency. It provides information about how spread out or clustered the data points are.

Real-time example: Let's consider a dataset representing the heights (in centi-meters) of 10 individuals: 160, 165, 170, 155, 168, 172, 158, 175, 180, 160.

There are several common measures of dispersion, including range, variance, and standard deviation.

- **Range: The range is the simplest measure of dispersion. It is calculated by taking the difference between the maximum and minimum values in a dataset.**

  For our example dataset, the range of heights would be:

  Maximum height: 180 cm

  Minimum height: 155 cm

  Range: 180 - 155 = 25 cm

So, the range of heights is 25 cm, indicating the spread between the tallest and shortest individuals in the dataset.

- **Variance: Variance measures the average squared deviation from the mean. It quantifies how much the data points vary from the mean.**

  For our example dataset, the variance of heights would involve several steps:

  Step 1: Calculate the mean height:

  (160 + 165 + 170 + 155 + 168 + 172 + 158 + 175 + 180 + 160) / 10 = 166.3 cm

  Step 2: Calculate the squared deviation of each height from the mean:

  (160 - 166.3)^2, (165 - 166.3)^2, (170 - 166.3)^2, (155 - 166.3)^2, (168 - 166.3)^2, (172 - 166.3)^2, (158 – 166.3)^2, (175 - 166.3)^2, (180 - 166.3)^2, (160 - 166.3)^2

Step 3: Calculate the average of the squared deviations:

(sum of squared deviations) / number of observations = variance

After performing the calculations, let's say we obtain a variance of 63.21 cm^2.

- **Standard Deviation: The standard deviation is the square root of the variance. It represents the typical or average amount of deviation from the mean.**

  For our example dataset, the standard deviation would be the square root of the variance, which is approximately 7.95 cm.

The range, variance, and standard deviation are all measures of dispersion that provide insights into the spread or variability of data points. They help us understand the distribution of values and the degree to which they deviate from the central tendency.

By examining these measures, we can gain a better understanding of the dataset's variability, identify outliers, and make more informed decisions based on the data's spread.

COEFFICIENT OF VARIATIONS

The coefficient of variation (CV) is a statistical measure that expresses the relative variability of a dataset in relation to its mean. It is used to compare the variability between datasets with different means or scales.

Real-time example: Let's consider two datasets representing the heights (in centi-meters) of two groups of individuals: Group A and Group B.

Group A: 160, 165, 170, 155, 168

Group B: 185, 190, 195, 180, 200

To calculate the coefficient of variation, we need to follow these steps:

Calculate the mean of each group:

Mean of Group A: (160 + 165 + 170 + 155 + 168) / 5 = 163.6 cm

Mean of Group B: (185 + 190 + 195 + 180 + 200) / 5 = 190 cm

Calculate the standard deviation of each group:

Standard deviation of Group A: Let's assume it is 5 cm

Standard deviation of Group B: Let's assume it is 7 cm

Calculate the coefficient of variation:

Coefficient of Variation for Group A = (Standard deviation of Group A / Mean of Group A) * 100

Coefficient of Variation for Group B = (Standard deviation of Group B / Mean of Group B) * 100

Let's assume the standard deviations we obtained are accurate. The calculations would be as follows:

Coefficient of Variation for Group A = (5 cm / 163.6 cm) * 100 ≈ 3.06%

Coefficient of Variation for Group B = (7 cm / 190 cm) * 100 ≈ 3.68%

In this example, the coefficient of variation for Group A is approximately 3.06%, while for Group B, it is approximately 3.68%.

Interpreting these results, we can say that the variability of the heights within Group B (3.68%) is slightly higher compared to Group A (3.06%). The coefficient of variation allows us to assess and compare the relative variability between datasets, regardless of their scales or means.

By using the coefficient of variation, we can gain insights into the relative dispersion of datasets, making it a useful tool in comparing the variability of different groups or datasets with varying means.

FIVE-NUMBER SUMMARY

The Five-Number Summary is a set of descriptive statistics that provides a concise summary of a dataset. It consists of five values: the minimum, the first quartile (Q1), the median (Q2), the third quartile (Q3), and the maximum.

Real-time example: Let's consider a dataset representing the scores of 20 students in a math test: 56, 62, 72, 68, 81, 78, 66, 59, 77, 85, 90, 73, 65, 72, 64, 69, 76, 79, 83, 87.

To calculate the Five-Number Summary, we need to follow these steps:

- **Minimum: Find the smallest value in the dataset. In this example, the minimum score is 56.**
- **Maximum: Find the largest value in the dataset. In this example, the maximum score is 90.**
- **Median (Q2): Arrange the dataset in ascending order: 56, 59, 62, 64, 65, 66, 68, 69, 72, 72, 73, 76, 77, 78, 79, 81, 83, 85, 87, 90. The median is the middle value of the ordered dataset. If there is an even number of observations, the median is the average of the two middle values. In this example, the median is 73.**
- **First Quartile (Q1): The first quartile is the median of the lower half of the dataset. It divides the dataset into two equal parts, with 25% of the data falling below it. In this example, the lower half of the dataset is: 56, 59, 62, 64, 65, 66, 68, 69, 72, 72. The median of this lower half is 66, so Q1 is 66.**
- **Third Quartile (Q3): The third quartile is the median of the upper half of the dataset. It divides the dataset into two equal parts, with 75% of the data falling below it. In this example, the upper half of the dataset is: 76, 77, 78, 79, 81, 83, 85, 87, 90. The median of this upper half is 81, so Q3 is 81.**

The Five-Number Summary for this example is:

Minimum: 56

Q1: 66

Median (Q2): 73

Q3: 81

Maximum: 90

The Quartiles:

Quartiles are statistical measures that divide a dataset into four equal parts, each containing 25% of the data. The first quartile (Q1) is the value below which 25% of the data falls, while the third quartile (Q3) is the value below which 75% of the data falls.

In the example above, Q1 is 66 and Q3 is 81. Quartiles are particularly useful for understanding the distribution of data and identifying potential outliers.

By using the Five-Number Summary and quartiles, we can quickly grasp the key characteristics of a dataset, such as the range, central tendency, and spread, making it easier to analyse and interpret the data.

IQR-INTER QUARTILE RANGE

The Interquartile Range (IQR) is a statistical measure that represents the spread or variability of the middle 50% of a dataset. It is calculated as the difference between the third quartile (Q3) and the first quartile (Q1).

Real-time example: Let's consider the dataset of math test scores we used earlier:

56, 62, 72, 68, 81, 78, 66, 59, 77, 85, 90, 73, 65, 72, 64, 69, 76, 79, 83, 87.

To calculate the IQR, we need to determine the values of Q1 and Q3, which we already found in the previous explanation:

Q1 = 66

Q3 = 81

Now, we can calculate the IQR:

IQR = Q3 - Q1

= 81 - 66

= 15

Therefore, the IQR for this dataset is 15.

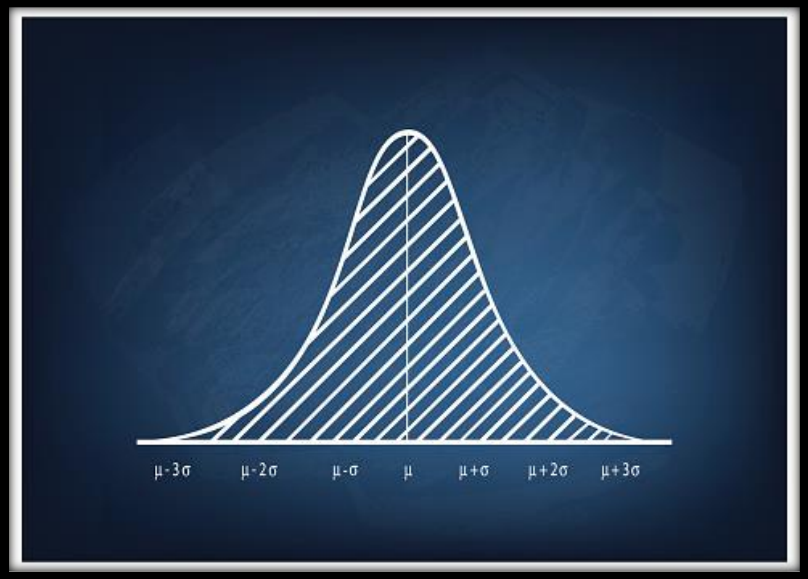
The IQR provides a measure of the dispersion or spread of the central 50% of the data. It is often used to identify and analyse potential outliers in a dataset. Values that fall significantly below Q1 - 1.5 * IQR or above Q3 + 1.5 * IQR are often considered outliers.

In our example, any value below 66 - 1.5 * 15 or above 81 + 1.5 * 15 would be considered a potential outlier. This helps in detecting extreme values that might significantly differ from the rest of the dataset.

The IQR is a robust measure of spread, meaning it is less affected by extreme values compared to the range or standard deviation. It provides a better representation of the variability within the central part of the dataset, making it a useful measure in statistical analysis and data interpretation.

NORMAL DISTRIBUTION

Normal Distribution, also known as the Gaussian distribution or bell curve, is a statistical distribution that is symmetric and characterized by its bell-shaped curve. It is a continuous probability distribution that is commonly observed in many natural and social phenomena.



Real-time example: Let's consider the heights of adult males in a population. Suppose we collect height data from a large sample of adult males and plot a histogram, where the x-axis represents height intervals and the y-axis represents the frequency or count of individuals falling within each interval.

If the distribution of heights follows a normal distribution, we would expect to see a bell-shaped curve when we plot the histogram. In a normal distribution:

The curve is symmetric: The left half of the curve is a mirror image of the right half. This means that the mean, median, and mode of the distribution are all equal and located at the centre of the curve.

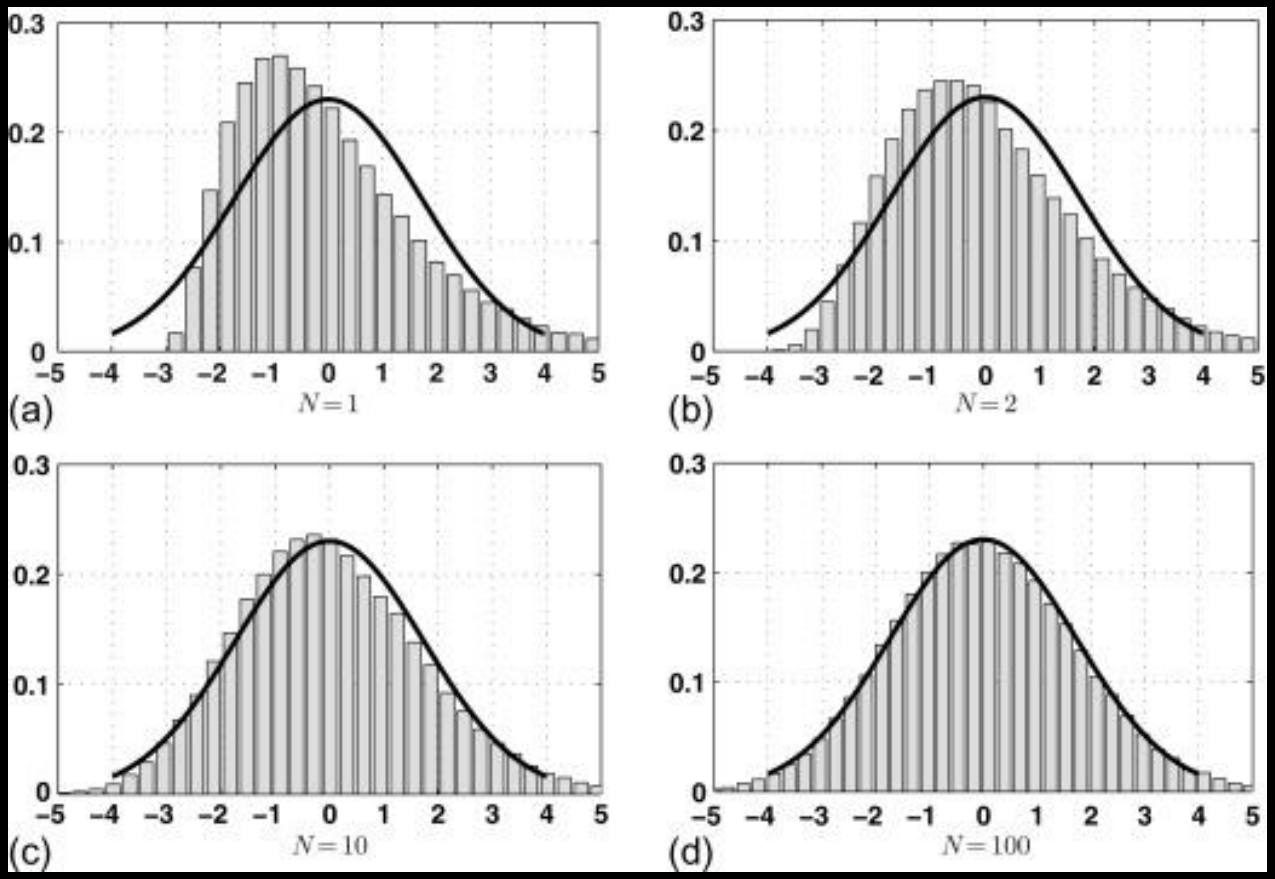
The majority of data is clustered around the mean: The highest frequency of data points occurs near the mean, and as we move away from the mean, the frequency gradually decreases.

The curve is bell-shaped: The curve is highest at the mean and tapers off as we move away from the mean in both directions. The curve is characterized by a specific standard deviation that determines the spread of the distribution.

The normal distribution is widely observed in various phenomena, such as physical measurements, exam scores, IQ scores, and many biological and social attributes. The understanding of normal distribution is important as it provides valuable insights into the characteristics of a dataset, allows for the application of statistical tests and models, and aids in making predictions and drawing conclusions.

CENTRAL LIMIT THEOREM

The Central Limit Theorem (CLT) is a statistical concept that states that when we take repeated random samples from any population, the distribution of the sample means will approach a normal distribution, regardless of the shape of the original population.



Real-time example: Imagine we have a population of 1,000 people and we want to know the average height of the population. We randomly select multiple samples of different sizes from this population. For each sample, we calculate the average height.

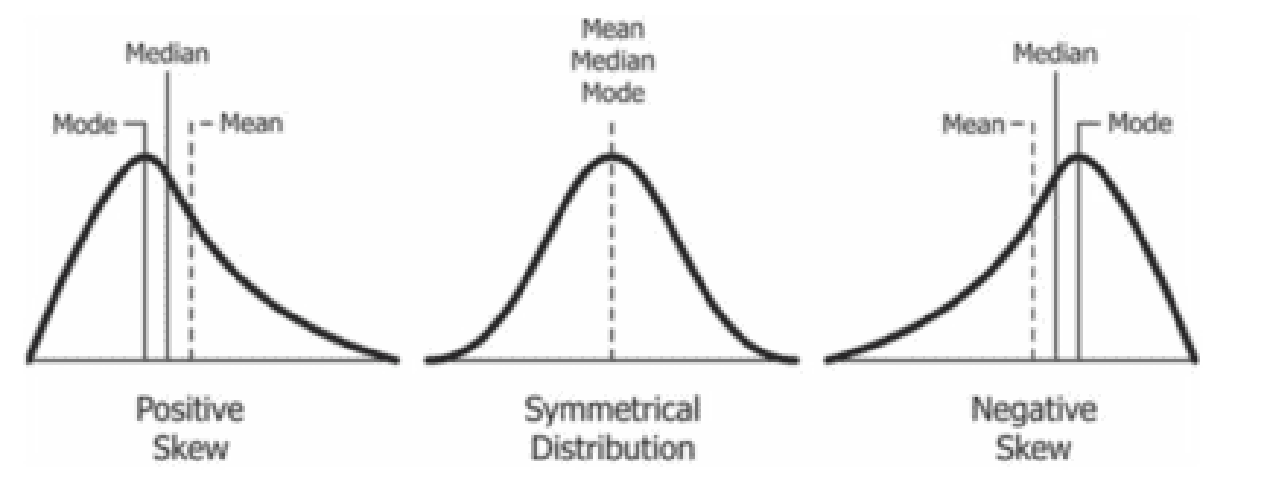
According to the Central Limit Theorem, as we increase the number of samples and the sample size, the distribution of these sample means will become approximately normal, regardless of the original population's shape. In other words, even if the population's height distribution is not normal, the distribution of the sample means will tend to be normal.

For instance, let's say we take 100 random samples of 50 people each from our population. We calculate the average height for each sample. If we plot a histogram of these sample means, we will observe that the distribution of the sample means will resemble a bell-shaped, normal distribution.

The Central Limit Theorem is valuable because it allows us to make statistical inferences about the population based on sample data. It enables us to use techniques that assume a normal distribution, such as hypothesis testing or constructing confidence intervals, even when the population itself is not normally distributed.

SKEWED DISTRIBUTION

A skewed distribution refers to an asymmetrical distribution of data points in a dataset. It occurs when the values cluster more towards one end of the distribution, resulting in a tail that extends towards the other end.



Real-time example: Let's consider a dataset representing the incomes of individuals in a particular city. Suppose we have the following income values: $20,000, $25,000, $30,000, $35,000, $40,000, $50,000, $60,000, $70,000, $100,000, $200,000.

In this dataset, we can observe that most of the incomes are clustered towards the lower end (left side) of the distribution, while a few individuals have significantly higher incomes. This results in a tail extending towards the higher end (right side) of the distribution.

If we were to plot a histogram of this dataset, we would see a skewed distribution, specifically a right-skewed or positively skewed distribution. The tail would be on the right side of the histogram, indicating the presence of a few higher income outliers.

In a skewed distribution:

- **In a right-skewed distribution: The tail extends towards the higher values, and the mean tends to be greater than the median.**

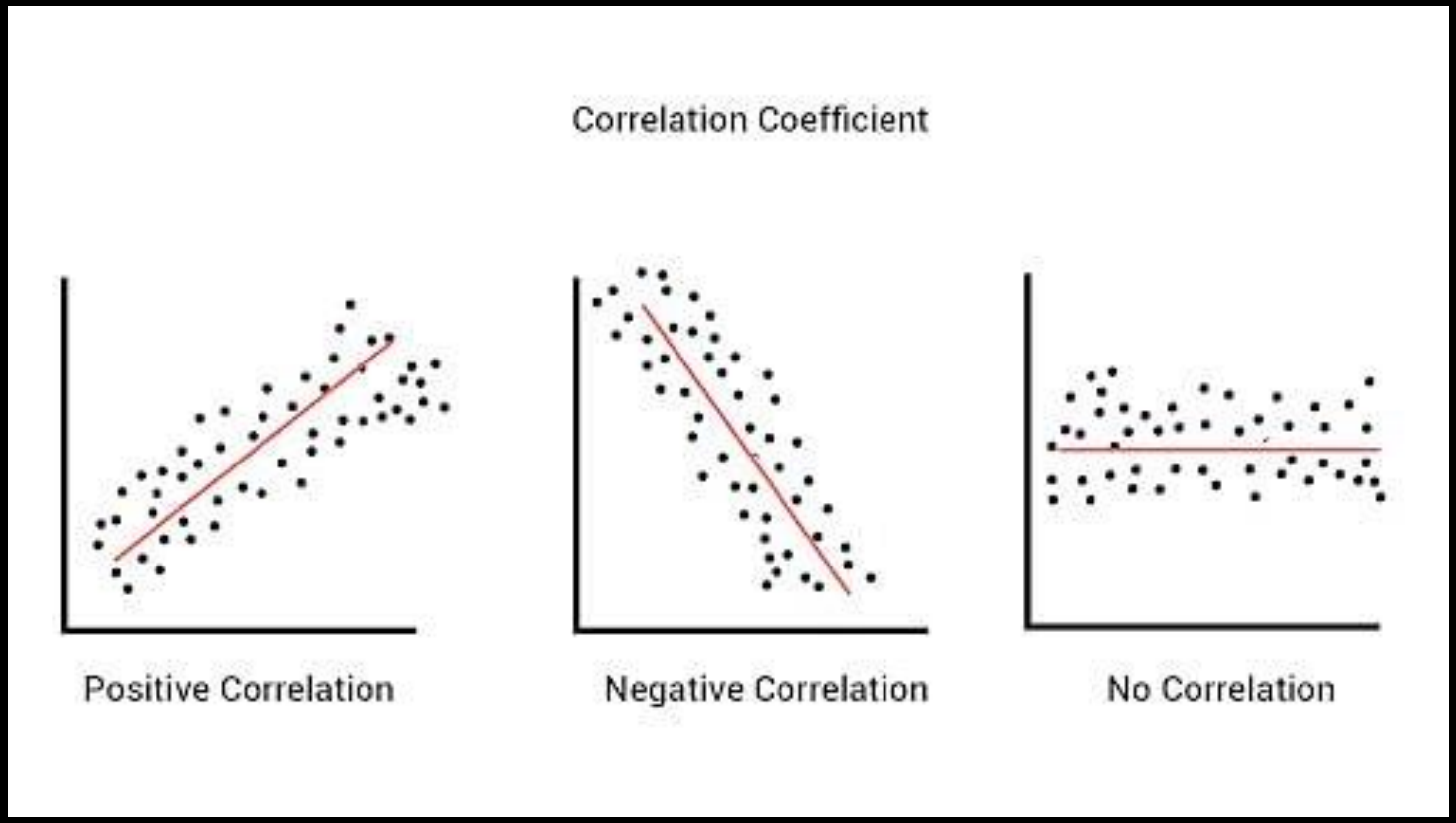
- **In a left-skewed distribution: The tail extends towards the lower values, and the mean tends to be smaller than the median.**

Understanding skewed distributions is important because they can have implications for data analysis and interpretation. Skewness provides insights into the shape and characteristics of the dataset, indicating whether it is symmetrical or exhibits a systematic bias towards one end. Skewed distributions can influence the choice of statistical methods and the interpretation of results.

CORRELATION

Correlation refers to the statistical relationship or association between two variables. It measures the extent to which changes in one variable are related to changes in another variable.

Real-time example: Let's consider a dataset that contains information about the study hours and corresponding test scores of a group of students.



- **If we observe a positive correlation between study hours and test scores, it means that as the number of study hours increases, the test scores also tend to increase. This indicates a positive relationship between the two variables.**
- **Conversely, if we observe a negative correlation between study hours and test scores, it means that as the number of study hours increases, the test scores tend to decrease. This indicates a negative relationship between the two variables.**
- **If there is no apparent relationship between study hours and test scores, it implies a lack of correlation or a weak correlation.**

Correlation is typically represented by a correlation coefficient, which is a numerical value ranging from -1 to +1. The sign of the correlation coefficient indicates the direction of the relationship (positive or negative), while the magnitude represents the strength of the relationship.

A correlation coefficient close to +1 indicates a strong positive correlation, meaning the variables move together in the same direction. A correlation coefficient close to -1 indicates a strong negative correlation, implying the variables move in opposite directions. A correlation coefficient close to 0 suggests a weak or no correlation, meaning the variables have little or no relationship.

Understanding the correlation between variables is essential in data analysis and decision-making. It helps us identify patterns, make predictions, and determine the strength and direction of relationships between variables. Correlation analysis is widely used in various fields, including social sciences, finance, and healthcare, to gain insights and make informed decisions.

HYPOTHESIS TESTING

Hypothesis testing is a statistical technique used to make inferences and draw conclusions about a population based on sample data. It involves formulating two competing hypotheses, the null hypothesis (H0) and the alternative hypothesis (Ha), and analyzing the sample data to determine which hypothesis is more supported by the evidence.

Real-time example: Let's suppose a company claims that a new marketing campaign has increased the average customer spending time on their website. As a data scientist, you want to test this claim using hypothesis testing.

Here's how the hypothesis testing process would work:

Formulating the hypotheses:

- **Null hypothesis (H0): The new marketing campaign has no effect on the average customer spending time.**
- **Alternative hypothesis (Ha): The new marketing campaign has increased the average customer spending time.**
- **Collecting and analysing the data:**

  You would collect data on customer spending time before and after the marketing campaign. Let's say you collect a sample of 100 customers' spending times before the campaign and 100 customers' spending times after the campaign.

- **Setting the significance level:**

  You would determine the significance level (often denoted as α) which represents the threshold for accepting or rejecting the null hypothesis. Commonly used values for α are 0.05 or 0.01.

- **Performing the statistical test:**

  Using appropriate statistical methods (such as a t-test or z-test), you would analyse the sample data to determine if there is enough evidence to reject the null hypothesis in favour of the alternative hypothesis. This involves calculating a test statistic and comparing it to a critical value or p-value.

- **Making a decision:**

  Based on the test statistic and the significance level, you would make a decision to either reject the null hypothesis or fail to reject it. If the evidence strongly supports the alternative hypothesis, you would reject the null hypothesis and conclude that the new marketing campaign has indeed increased the average customer spending time. If the evidence is not strong enough, you would fail to reject the null hypothesis, indicating that there is insufficient evidence to conclude that the marketing campaign had an effect.

Hypothesis testing helps us make data-driven decisions and draw conclusions about population parameters based on sample data. It provides a systematic and objective approach to evaluating claims, testing theories, and assessing the significance of relationships in various fields, including business, social sciences, healthcare, and more.