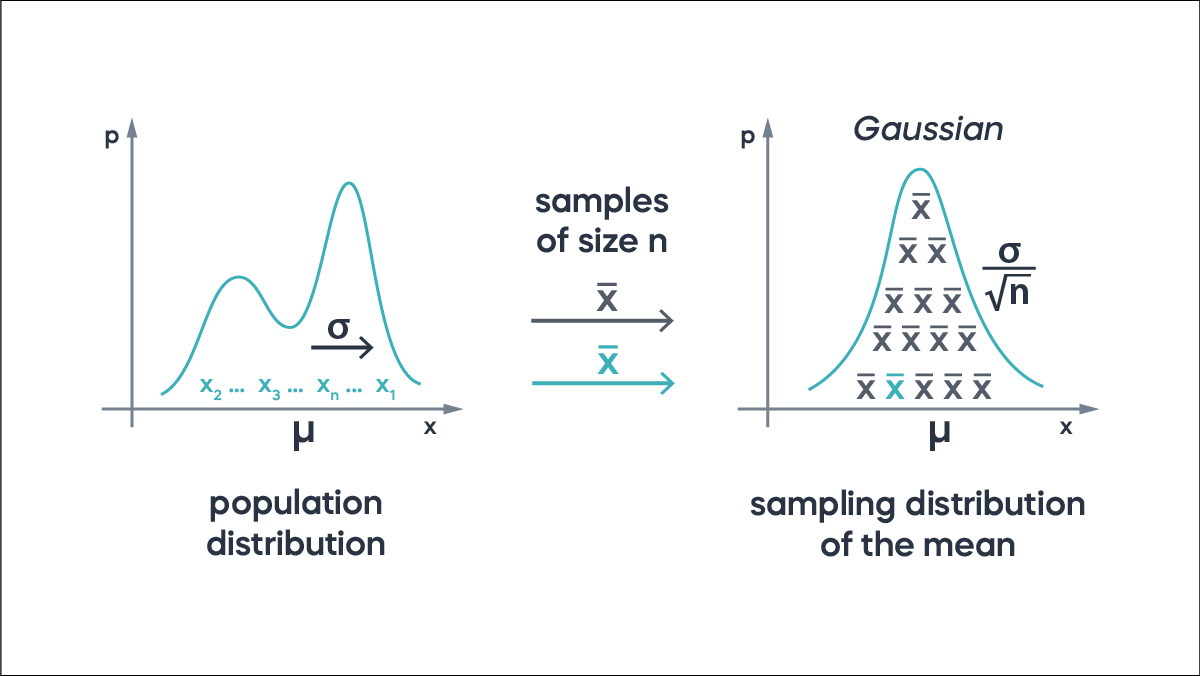
**Basic Level**

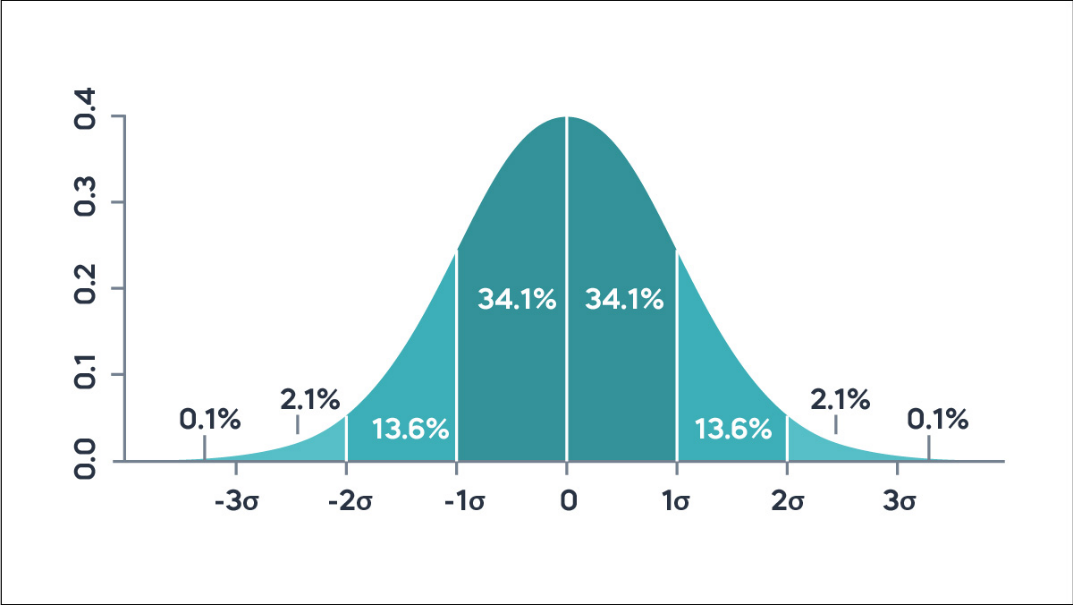
1. **What is Statistics?**
   * Statistics is the discipline that studies and develops techniques for gathering, processing, analysing, interpreting, and communicating statistical information (using information gathered from research)
2. **What is the difference between descriptive and inferential statistics?**
   * **Descriptive statistics** summarize and describe the features of a dataset (e.g., mean, median, mode, range).
   * **Inferential statistics** use a random sample of data taken from a population to describe and make inferences about the population.
3. **Explain the difference between a population and a sample.**
   * A **population** includes all elements from a set of data.
   * A **sample** consists of one or more observations drawn from the population.
4. **What are the measures of central tendency?**
   * Measures of central tendency include the **mean** (average), **median** (middle value), and **mode** (most frequent value).
5. **Define variance and standard deviation.**
   * **Variance** measures the dispersion of a set of data points around their mean value. **Standard deviation** is the square root of the variance and represents the average distance of each data point from the mean.
6. **What is the Central Limit Theorem? Why is it important?**
   * The **Central Limit Theorem** states that the distribution of the sample mean approaches a normal distribution as the sample size increases, regardless of the population's distribution.
   * It is important because it allows for the application of inferential statistics on sample data.



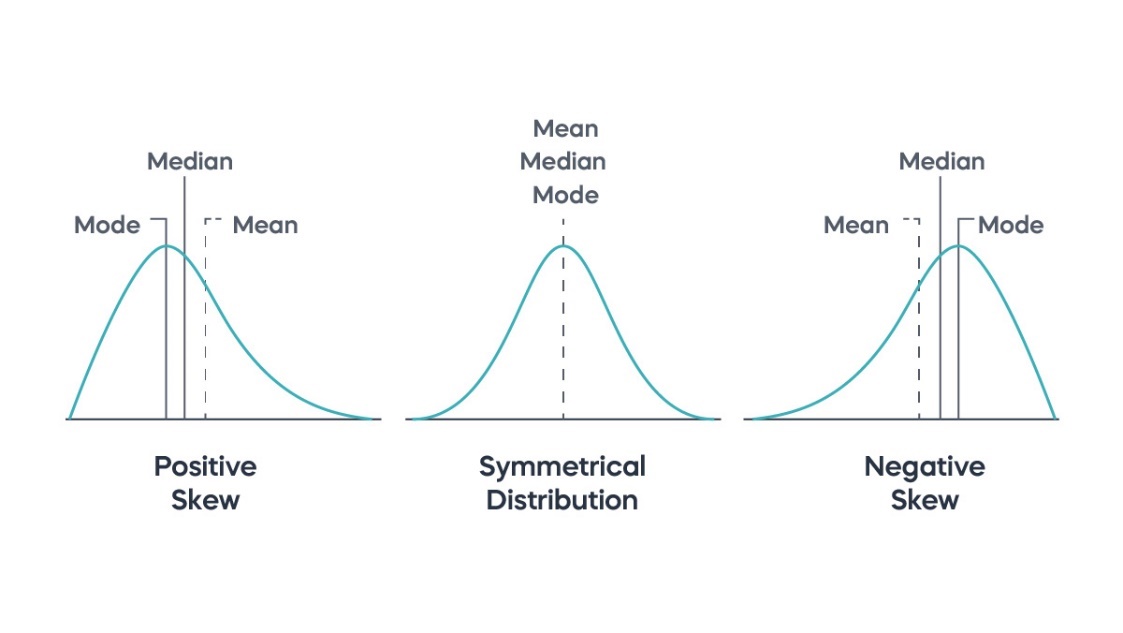
1. **What is a p-value?**
   * A **p-value** measures the strength of evidence against the null hypothesis. A low p-value (< 0.05) indicates strong evidence against the null hypothesis, leading to its rejection.
2. **What is the difference between a Type I and Type II error?**
   * A **Type I error** occurs when the null hypothesis is true but is rejected.
   * A **Type II error** occurs when the null hypothesis is false but fails to be rejected.
3. **Explain what a confidence interval?**
   * A **confidence interval** is a range of values, derived from a dataset, that is likely to contain the value of an unknown population parameter. The interval has an associated confidence level that quantifies the level of confidence that the parameter lies within the interval.
4. **What is a normal distribution?**
   * A **normal distribution** is a probability distribution that is symmetric about the mean, with data near the mean more frequent in occurrence than data far from the mean. It is often referred to as a bell curve.
5. **How is the statistical significance of an insight assessed?**
   * Statistical significance of an insight is assessed through **hypothesis testing**.
     1. First, formulate null (H0) and alternative (H1) hypotheses.
     2. Choose a significance level (α), commonly 0.05.
     3. Collect data and compute a test statistic using an appropriate statistical test (e.g., t-test, chi-square).
     4. Calculate the p-value, which indicates the probability of obtaining the observed results if H0 is true.
     5. Compare the p-value to α. If p-value ≤ α, reject H0, indicating statistical significance. If p-value > α, fail to reject H0, suggesting the result is not statistically significant.
6. **Where are long-tailed distributions used?**
   * Long-tailed distributions find applications in various fields where rare but significant events are observed more frequently than expected under a normal distribution.
   * They are commonly used in finance for modelling extreme market movements and risk management.
   * **Common Long-Tailed Distributions:**
     1. **Pareto Distribution**: Often used in economics and finance to model wealth distribution and other phenomena.
     2. **Log-Normal Distribution**: Used in various fields such as finance, biology, and reliability engineering.
7. **What is observational and experimental data in Statistics?**
   * **Observational data** in statistics involves collecting information by observing existing behaviours or characteristics without intervening.
   * It's used to study correlations and associations between variables, such as examining the relationship between diet and health outcomes in a population.
   * **Experimental data**, on the other hand, is gathered through controlled experiments where researchers manipulate variables to test hypotheses and establish cause-and-effect relationships.
   * For instance, testing the efficacy of a new medication by administering it to one group while providing a placebo to another.
8. **What is exploratory data analysis?**
   * Exploratory data analysis (EDA) is a critical approach in data science and statistics used to summarize, visualize, and interpret the main characteristics of a dataset.
   * Its primary goals are to uncover patterns, identify anomalies, test assumptions, and generate hypotheses that can guide further analysis.
   * Techniques used in EDA include summary statistics, histograms, box plots, scatter plots, correlation matrices, and more sophisticated methods like clustering or dimensionality reduction.
9. **What is the meaning of selection bias?**
   * Selection bias occurs when the way participants or data are chosen for a study leads to results that don't represent the whole group accurately.
   * For instance, imagine a study on **smartphone usage** that only surveys people in urban areas.
   * This could overlook how rural residents use smartphones differently, leading to conclusions that might not apply to everyone.
   * Selection bias can happen when certain groups are more likely to be included or excluded, skewing results and making them less reliable for understanding the entire population.
   * It's important in research to choose participants in a way that avoids favoring one group over another.
10. **What is meant by mean imputation for missing data? Why is it bad?**
    * **Mean imputation** for missing data involves replacing missing values with the mean of observed values for that variable.
    * It's not ideal because it can change the results and hide important details in the data.
    * For example, if you're looking at people's heights and a few are missing, replacing those missing heights with the average could make it seem like everyone is closer in height than they really are.
    * It's better to use methods that try to predict what the missing values might be based on the information you do have, rather than just filling in with an average.

1. **How is missing data handled in statistics?**
   * Missing data in statistics is managed through deletion or imputation techniques.
   * **Deletion methods** remove incomplete cases but reduce sample size and potentially bias results.
   * **Imputation methods** replace missing values with estimated ones, such as mean, regression-based predictions, or multiple imputations that preserve data variability.
2. **What is the meaning of an inlier?**
   * An inlier is a data point within a dataset that lies close to the majority of other data points, typically within the central tendency or main cluster of observations.
   * Unlike outliers, which are data points that significantly deviate from the majority, inliers are considered typical or representative of the overall pattern or distribution of the data.
3. **What is an outlier? How can outliers be determined in a dataset?**
   * An outlier in a dataset is a data point that significantly deviates from other observations.
   * Methods to detect outliers include visual inspection using **box plots** or **scatter plots**.
   * Statistical approaches like **Z-score** or **interquartile range (IQR**).
4. **What is DOE?**
   * DOE stands for Design of Experiments. It is a systematic approach used in statistics and engineering to determine the relationships between factors affecting a process and the outcomes of that process.
   * DOE involves planning, conducting, analyzing, and interpreting controlled tests or experiments to optimize processes or products.
5. **What is the meaning of six sigma in statistics?**
   * Six Sigma is a disciplined, data-driven methodology used to improve processes and eliminate defects or errors in various industries.
   * It originated in manufacturing but has been widely adopted across sectors including healthcare, finance, and services.
   * The term "Six Sigma" refers to a statistical concept that measures how far a process deviates from perfection.
   * It aims to achieve no more than 3.4 defects per million opportunities, corresponding to a process capability of 99.99966%.
6. **State the case where the median is a better measure when compared to the mean.**
   * The median is preferable to the mean when data is **skewed** or contains **outliers**.
   * **Skewed distributions** can pull the mean away from the center, making it less representative of typical values.
   * The median, being the middle value in sorted data, is robust to outliers and extreme values, accurately reflecting central tendency.
   * Additionally, in ordinal or categorical data where arithmetic operations lack meaning, the median offers a meaningful measure of central location without requiring the assumptions that underlie the mean calculation.
7. **What type of data does not have a log-normal distribution or a Gaussian distribution?**
   * Data that does not have a log-normal distribution or a Gaussian (normal) distribution typically falls into **categories** where values are not symmetrically distributed around a central tendency and do not follow the typical bell-shaped curve observed in these distributions.

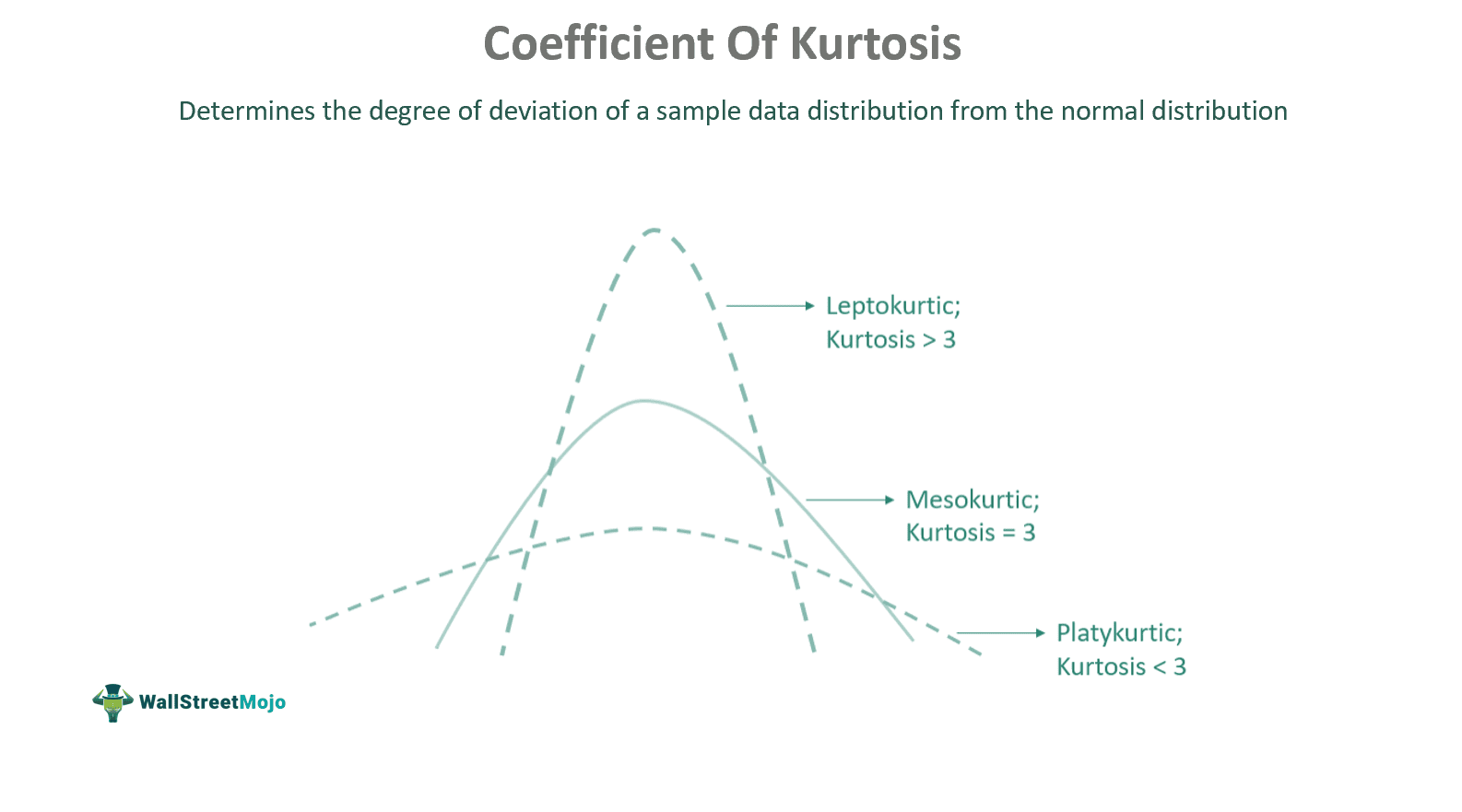
1. **What is the Pareto principle?**
   * The **Pareto Principle**, also known as the 80/20 rule, is a concept that suggests that roughly 80% of the effects come from 20% of the causes.
   * **Benefits of Applying the Pareto Principle**:
     1. Prioritization: Helps identify the most critical issues or opportunities to focus on for maximum impact.
     2. Efficiency: Encourages focusing resources on the most productive areas.
     3. Problem Solving: Guides efforts to address the most significant problems first, leading to more effective solutions.
2. **What is the meaning of the five-number summary in Statistics?**
   * The five-number summary in statistics is a descriptive statistic that provides a quick and informative overview of a dataset. It consists of five key values:
     1. **Minimum**: The smallest data point in the dataset.
     2. **First Quartile (Q1):** The value below which 25% of the data points lie. It is the median of the lower half of the dataset.
     3. **Median (Q2):** The middle value that divides the dataset into two equal parts. If the number of observations is odd, it's the middle number; if even, it's the average of the two middle numbers.
     4. **Third Quartile (Q3):** The value below which 75% of the data points lie. It is the median of the upper half of the dataset.
     5. **Maximum**: The largest data point in the dataset.
3. **What are quantitative data and qualitative data?**
   * [Quantitative data](https://intellipaat.com/blog/what-is-quantitative-methods/) is also known as numeric data.
   * Qualitative data is also known as categorical data.
4. **What is a bell-curve distribution?**
   * A bell-curve distribution, also known as a normal distribution or Gaussian distribution, is a specific type of continuous probability distribution that is symmetrical and has a characteristic bell-shaped curve. Here are its key features:
   * Key Features of a Bell-Curve Distribution
     1. **Symmetry**: The curve is perfectly symmetrical around the mean (average), meaning the left side of the curve is a mirror image of the right side.
     2. **Mean, Median, and Mode**: In a normal distribution, the mean, median, and mode are all located at the center of the distribution and are equal.
     3. **Standard Deviation:** The spread or width of the curve is determined by the standard deviation. Approximately 68% of the data falls within one standard deviation of the mean, 95% within two standard deviations, and 99.7% within three standard deviations.
     4. **Asymptotic**: The tails of the curve approach, but never actually touch, the horizontal axis. This implies that the probability of extreme values never actually reaches zero, though it becomes very small.
     5. **Unimodal**: There is only one peak (mode) in the distribution.



1. **What is skewness?**
   * Skewness is a statistical measure that describes the asymmetry of the probability distribution of a real-valued random variable about its mean.
   * It indicates whether the data points are spread out more to the left (negative skew) or to the right (positive skew) of the mean.



1. **What is kurtosis?**
   * Kurtosis is a statistical measure that describes the "tailedness" of a data distribution. It indicates whether the data have heavy or light tails compared to a normal distribution. There are three types of kurtosis:
     1. **Leptokurtic** (Positive Kurtosis): Distributions with heavy tails and more extreme outliers than a normal distribution.
     2. **Platykurtic** (Negative Kurtosis): Distributions with light tails and fewer extreme values.
     3. **Mesokurtic**: Distributions similar to the normal distribution, with moderate tails.



1. **What is correlation?**
   * A correlation measures the linear relationship between two or more variables.
   * It ranges between -1 and 1.
     1. It’s positive if the variables increase or decrease together.
     2. If it’s negative, one variable decrease while the other increases.
     3. When the value is 0, the variables aren’t related.

A graph of a number of points

Description automatically generated with medium confidence

1. **What are the characteristics of large numbers in statistics?**
   * **Law of Large Numbers**: This fundamental theorem states that as the number of observations or trials (n) increases, the sample mean (average) approaches the population mean.
   * In practical terms, this means that with a large enough sample size, random sampling will tend to reflect the true population parameters more accurately.
2. **How does linear regression work?**
   * Linear regression analyzes relationships between variables by fitting a straight line through data points. It estimates how changes in independent variables influence a dependent variable.
   * For example, linear regression may be used to study the connection between various predictors, such as age, gender, heredity, diet, and height.
3. **What Is the Confidence Interval?**
   * A confidence interval is a statistical range around a sample estimate (like a mean or proportion) that, with a specified level of confidence (e.g., 95% or 99%), likely contains the true population parameter.
   * It quantifies the uncertainty in estimating population values from sample data, typically calculated as the **sample statistic ± margin of error**.
   * A wider interval reflects greater uncertainty, while a narrower interval indicates more precise estimation.
4. **What are the types of biases that you can encounter while sampling?**
   * Sampling biases include
     1. **selection bias** (non-random sampling)
     2. **measurement bias** (flawed measurement tools),
     3. **undercoverage** bias (incomplete sampling frame).
     4. **reporting bias** (selective reporting)
     5. **time interval bias** (temporal effects)
     6. **survivorship bias** (excluding non-survivors).
5. **How can you calculate the p-value using MS Excel?**
   * The formula used in MS Excel to calculate p-value is –

=tdist(x,deg\_freedom,tails)

* + The p-value is expressed in decimals in Excel. Here are the steps to calculate it –
    1. Find the Data tab
    2. On the Analysis tab, click on the data analysis icon
    3. Select Descriptive Statistics and then click OK
    4. Select the relevant column
    5. Input the confidence level and other variables

1. **What are the necessary conditions for a Binomial Distribution?**
   * The necessary conditions for a binomial distribution are:
     1. **Fixed Number of Trials**: The experiment consists of a fixed number of identical trials or experiments(n).
     2. **Independent Trials**: Each trial must be independent of the others. The outcome of one trial should not affect the outcomes of the others.
     3. **Binary Outcomes**: Each trial must result in one of two possible outcomes, typically labeled as success or failure (denoted as 1−p).
     4. **Constant Probability of Success**: The probability of success remains constant across all trials. This is a key assumption that distinguishes the binomial distribution from other distributions.
2. **What are the different kinds of variables or levels of measurement?**
   * Variables in statistics can be classified into different levels of measurement, which indicate the nature and extent of information they provide. There are four main levels of measurement:
     1. **Nominal Level**: Variables at the nominal level are categorical and represent qualitative differences between items or subjects. They have no inherent order or ranking. Examples include:
        1. Gender (male, female)
        2. Marital status (single, married, divorced)
        3. Eye color (blue, brown, green)
     2. **Ordinal Level**: Variables at the ordinal level have a natural order or ranking, but the differences between values are not necessarily equal or quantifiable. Examples include:
        1. Educational level (high school, college, graduate school)
        2. Likert scales (e.g., strongly agree, agree, neutral, disagree, strongly disagree)
        3. Socioeconomic status (low, middle, high)
     3. **Interval Level:** Variables at the interval level have a meaningful order, and the differences between values are equal and measurable. However, there is no true zero point. Examples include:
        1. Temperature measured in Celsius or Fahrenheit (e.g., 20°C, 30°C)
        2. Dates on a calendar (e.g., January 1, February 15)
        3. IQ scores (e.g., IQ of 100, IQ of 120)
     4. **Ratio Level**: Variables at the ratio level have all the characteristics of interval variables, but they also have a true zero point, indicating the absence of the quantity being measured. This allows for ratios to be meaningful. Examples include:
        1. Height (measured in centimeters or inches)
        2. Weight (measured in kilograms or pounds)
        3. Time duration (e.g., time taken to complete a task)

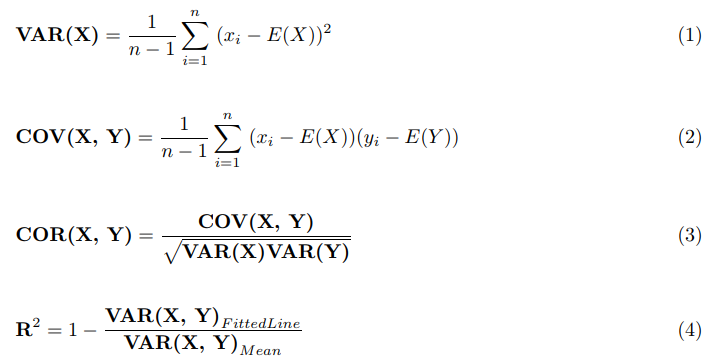
**A diagram of measurement level

Description automatically generated**

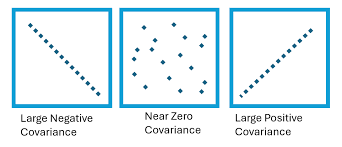
1. **Explain the Difference Between Probability Distribution and Sampling Distribution**
   * **Probability Distribution**:
     1. A probability distribution describes the likelihood of all possible outcomes of a random variable.
     2. It specifies the probabilities associated with each possible outcome or range of outcomes.
     3. Probability distributions can be discrete (like the binomial or Poisson distributions, where outcomes are countable) or continuous (like the normal or exponential distributions, where outcomes can take any value within a range).
   * **Sampling Distribution**:
     1. A sampling distribution refers to the distribution of a sample statistic (such as the sample mean or sample proportion) calculated from multiple samples of the same size taken from a population.
     2. The central idea is that as more samples are taken from a population, the distribution of the sample statistics (like means or proportions) tends to follow a predictable pattern.
     3. For example, the sampling distribution of the sample mean tends to approximate a normal distribution due to the Central Limit Theorem, regardless of the population's underlying distribution.
2. **What Are the Main Measures of Variability?**
   * The main measures of variability quantify how data points spread around the central tendency.
   * **Range** measures the difference between the maximum and minimum values.
   * **Interquartile Range** (IQR) assesses the spread of the middle 50% of data, robust to outliers.
   * **Variance** computes the average squared deviation from the mean.
   * **standard deviation** is its square root, indicating average distance from the mean. **Mean Absolute Deviation (MAD**) measures average absolute differences from the mean, less influenced by outliers.
3. **What is the meaning of KPI in statistics?**
   * KPI stands for Key Performance Indicator in statistics. It is used as a reliable metric to measure the success of a company with respect to its achieving the required business objectives.
   * There are many good examples of KPIs:
     1. Profit margin percentage.
     2. Operating profit margin
     3. Expense ratio
4. **What is a null hypothesis and an alternative hypothesis?**
   * The **null hypothesis** (H0) is a statement that there is no effect or no difference, and it serves as the starting point for statistical testing.
   * The **alternative hypothesis** (H1) is a statement that contradicts the null hypothesis, indicating the presence of an effect or difference.

**Intermediate Level**

1. **What are left-skewed and right-skewed distributions?**
   * **A left-skewed distribution** is one where the left tail is longer than that of the right tail. Here, it is important to note that the **mean < median < mode**.
   * Similarly, **a right-skewed distribution** is one where the right tail is longer than the left one. But, here **mean > median > mode**.
2. **What are the types of sampling in Statistics?**
   * There are four main types of data sampling as shown below:
     1. **Simple random**: Pure random division
     2. **Cluster**: Population divided into clusters
     3. **Stratified**: Data divided into unique subgroups
     4. **Systematical**: Picks up every ‘n’ member in the data
3. **What are the assumptions of a linear regression model?**
   * Linearity: The relationship between the predictors and the outcome is linear.
   * Independence: Observations are independent of each other.
   * Homoscedasticity: Constant variance of errors.
   * Normality: The errors of the model are normally distributed.
4. **Explain the difference between correlation and causation.**
   * **Correlation** indicates a relationship or association between two variables.
   * **Causation** indicates that one variable directly affects the other.
5. **What is ANOVA and when is it used?**
   * **ANOVA (Analysis of Variance)** is a statistical method used to test differences between two or more means.
   * It is used when comparing three or more groups for statistical significance.
6. **What is the difference between a parametric and a non-parametric test?**
   * **Parametric tests** assume underlying statistical distributions in the data.
   * **Non-parametric tests** do not assume any specific distribution and are often used when the data does not meet the assumptions of parametric tests.
7. **What is multicollinearity and how can it be detected?**
   * **Multicollinearity** occurs when predictor variables in a regression model are highly correlated.
   * It can be detected using variance inflation factor (VIF) or correlation matrices.
8. **What is covariance?**
   * Covariance is a measure that indicates the extent to which two random variables change together. It measures the relationship between two variables, showing whether they tend to increase or decrease in tandem.



* + Interpretation:
    1. Cov(X,Y)>0 : Indicates that as X increases, Y tends to increase as well (positive covariance).
    2. Cov(X,Y)<0: Indicates that as X increases, Y tends to decrease (negative covariance).
    3. Cov(X,Y)=0: Indicates no linear relationship; X and Y are independent or linearly unrelated.



1. **Explain what a chi-square test is and when you would use it.**
   * A **chi-square test** is used to determine if there is a significant association between categorical variables.
   * It compares the observed frequencies to the expected frequencies.
2. **Explain the difference between homoscedasticity and heteroscedasticity.**
   * **Homoscedasticity** means that the variance of errors is constant across all levels of an independent variable.
   * **Heteroscedasticity** means that the variance of errors varies across levels of an independent variable.
3. **What is Bessel's correction?**
   * Bessel's correction is a statistical adjustment made to correct the bias in the estimation of the population variance and covariance from a sample. It addresses the fact that the sample variance tends to underestimate the true population variance.

A white paper with black text

Description automatically generated

1. **The standard normal curve has a total area to be under one, and it is symmetric around zero. True or False?**
   * **True**, a [normal curve](https://intellipaat.com/blog/tutorial/probability-tutorial/the-normal-distribution/) will have the area under unity and the symmetry around zero in any distribution. Here, all the measures of central tendencies are equal to zero due to the symmetric nature of the standard normal curve.
2. **If a distribution is skewed to the right and has a median of 20, will the mean be greater than or less than 20?**
   * If the given distribution is a right-skewed distribution, then the mean should be greater than 20, while the mode remains to be less than 20.

1. **In an observation, there is a high correlation between the time a person sleeps and the amount of productive work he does. What can be inferred from this?**
   * A high correlation between a person's sleep duration and their productivity suggests a strong association where more sleep is likely linked to higher work output.
   * This correlation could imply a causal relationship, where sufficient sleep enhances cognitive function and energy levels, thereby boosting productivity.
   * Alternatively, shared underlying factors such as health or motivation may influence both sleep habits and work performance.
   * Contextual factors like job demands or personal routines might also contribute to this relationship.
   * Further research, including longitudinal studies and controlled experiments, would be needed to confirm causality and understand the precise mechanisms at play in this observed correlation.
2. **What is the relationship between the confidence level and the significance level in statistics?**
   * The confidence level (1−α) indicates the probability that a confidence interval contains the true population parameter.
   * The significance level (α) sets the probability of rejecting the null hypothesis when it is true.
   * **They are complements, where higher confidence corresponds to lower significance, affecting the precision of estimates and hypothesis testing.**
3. **A regression analysis between apples (y) and oranges (x) resulted in the following least-squares line: y = 100 + 2x. What is the implication if oranges are increased by 1?**
   * If the oranges are increased by one, there will be an increase of 2 apples since the equation is: y = 100 + 2x.
4. **What types of variables are used for Pearson’s correlation coefficient?**
   * Pearson's correlation coefficient, denoted as r, is used to measure the strength and direction of the linear relationship between two continuous variables. Specifically, it is appropriate for:
     1. **Interval Variables**:
        1. Variables measured on an interval scale, where the intervals between consecutive values are equal and meaningful.
        2. Examples include temperature in Celsius or Fahrenheit.
     2. **Ratio Variables**:
        1. Variables measured on a ratio scale, where there is a true zero point and ratios between values are meaningful.
        2. Examples include height, weight, and income.
5. **In a scatter diagram, what is the line that is drawn above or below the regression line called?**
   * In a scatter diagram, the line that is drawn above or below the regression line is called the "residual line" or "residual plot."
   * These lines represent the vertical distances between each data point and the corresponding predicted values on the regression line.
6. **What are the examples of symmetric distribution?**
   * Symmetric distribution means that the data on the left side of the median is the same as the one present on the right side of the median.
   * There are many examples of symmetric distribution, but the following three are the most widely used ones:
     1. Uniform distribution
     2. Binomial distribution
     3. Normal distribution
7. **Where is inferential statistics used?**
   * Inferential statistics is used for several purposes, such as research, in which we wish to draw conclusions about a population using some sample data.
   * This is performed in a variety of fields, ranging from government operations to quality control and quality assurance teams in multinational corporations.
8. **What is the relationship between mean and median in a normal distribution?**
   * In a normal distribution, the mean is equal to the median.
9. **What is the difference between the Ist quartile, the IInd quartile, and the IIIrd quartile?**
   * Quartiles are used to describe the distribution of data by splitting data into three equal portions, and the boundary or edge of these portions are called quartiles.
   * That is,
     1. The lower quartile (Q1) is the 25th [percentile](https://intellipaat.com/blog/percentile/).
     2. The middle quartile (Q2), also called the median, is the 50th percentile.
     3. The upper quartile (Q3) is the 75th percentile.
10. **What is a likelihood function?**
    * A **likelihood function** measures the probability of observing the given sample data as a function of the parameters of the statistical model.
11. **What is the difference between a Z-test and a T-test?**
    * A **Z-test** is used when the sample size is large (n > 30) and the population variance is known.
    * A **T-test** is used when the sample size is small (n ≤ 30) and the population variance is unknown.
12. **How do the standard error and the margin of error relate?**
    * The margin of error is directly related to the standard error. Specifically, for estimates such as means or proportions, the margin of error is typically calculated as **Margin of Error=Critical Value × Standard Error**, where the critical value depends on the desired confidence level (e.g., 1.96 for 95% confidence level in a normal distribution).
    * Therefore, the standard error serves as the basis for calculating the margin of error, influencing how much uncertainty is associated with an estimate. A smaller standard error leads to a smaller margin of error, indicating a more precise estimate with greater confidence in its accuracy.
13. **What is one sample t-test?**
    * This T-test is a statistical hypothesis test in which we check if the mean of the sample data is statistically or significantly different from the population’s mean.
14. **What are the applications of long-tailed distributions?**
    * **Network Traffic**: Long-tailed distributions help in modelling heavy-tailed data traffic, aiding in network capacity planning and resource allocation.
    * **Web Traffic**: Analysis of user visits, page views, or download counts often follows long-tailed distributions, guiding content optimization and marketing strategies.
    * **Income Distribution**: Long-tailed distributions in income inequality studies inform policies addressing wealth disparity and social equity.
15. **What does a degree of freedom (DF) represent in statistics?**
    * In statistics, degrees of freedom (DF) denote the number of independent pieces of information available to estimate a parameter or perform a statistical test accurately. It varies based on the context:
      1. For hypothesis testing like t-tests and ANOVA, DF adjusts for the number of groups and sample sizes.
      2. In regression analysis, DF accounts for the number of predictors used.
      3. In chi-square tests, DF reflects the number of categories minus one. DF plays a crucial role in determining the precision of estimates and the reliability of statistical tests by accounting for the variability inherent in the data, ensuring valid conclusions are drawn from statistical analyses.
16. **Symmetric distributions need to be unimodal, does it?**
    * No, symmetric distributions do not necessarily need to be unimodal, though many common symmetric distributions are indeed unimodal.
17. **How do you define empirical rule?**
    * The 68 – 95 – 99.7 rule or the Three Sigma Rule refers to the proposition that on a Normal Distribution,
    * There will be 68% within one Standard Error of the Mean of the data.
    * There will be 95% of the data within two Standard deviations of the mean.
    * There is around a 97% chance that the data will be within three standard deviations of the mean.

A diagram of a normal distribution

Description automatically generated

1. **What is a statistical interaction?**
   * A statistical interaction occurs when the effect of one independent variable (factor) on the dependent variable differs depending on the level or presence of another independent variable.
   * **Example**: In a study examining the effect of a new drug (Factor A) on blood pressure (Dependent Variable), an interaction might occur with age (Factor B). The drug’s effect on blood pressure may vary depending on whether the patient is young or old, indicating an interaction effect between the drug and age.
2. **What general conditions must be satisfied for the central limit theorem to hold?**
   * **Independence**: Observations in the sample must be independent of each other.
   * **Sample Size**: The sample size n should be sufficiently large (commonly n≥30n), although this can vary based on the population's distribution.
   * **Population** Distribution: The population should have a finite mean and variance.
   * **Random Sampling**: The sample should be selected randomly from the population to ensure representativeness.
3. **How to convert normal distribution to standard normal distribution?**
   * Any point (x) from the normal distribution can be converted into standard normal distribution (Z) using this formula –

**Z(standardized) = (x-µ) / σ**

* + Here, Z for any x value indicates how many standard deviations x is away from the mean of all values of x.

1. **What can you do with an outlier?**
   * **Identification**: Detect outliers through visual inspection, statistical methods like Z-score or interquartile range (IQR).
   * **Assessment**: Understand the nature and potential cause of outliers (e.g., data entry errors, natural variability, or significant observations).
   * **Impact Analysis**: Evaluate how outliers affect statistical analyses, such as skewing means or inflating variances.
   * **Handling**: Consider transformations (e.g., log transformation) to normalize data or use robust statistical methods less sensitive to outliers.
   * **Reporting**: Transparently report outliers to provide context and maintain data integrity.
   * **Decision**: Decide whether to exclude outliers based on their nature and impact, balancing data accuracy with analysis objectives.
2. **What is a confounding variable?**
   * **Definition**: A confounding variable is an extraneous variable that correlates with both the independent and dependent variables. It can mask the true relationship between them or falsely suggest a relationship where none exists.
   * **Impact**: If not properly controlled or accounted for, a confounding variable can lead to erroneous conclusions about the causal relationship between the independent and dependent variables.
   * **Examples**: Consider a study investigating the relationship between exercise and heart disease risk. Age could act as a confounding variable because older individuals are more likely to have higher heart disease risks regardless of their exercise habits. Without controlling for age, the study might erroneously attribute differences in heart disease risk solely to exercise levels, when in fact age could be influencing both the likelihood of exercising and the risk of heart disease independently.
3. **What does it mean if a model is heteroscedastic?**
   * A model is said to be heteroscedastic when the variation in errors comes out to be inconsistent. It often occurs in two forms – conditional and unconditional.
4. **What does autocorrelation mean?**
   * Autocorrelation measures the correlation between a variable's current value and its past values within a series. It indicates whether there is a pattern or relationship between successive observations.
5. **What is the meaning of sensitivity in statistics?**
   * Definition: Sensitivity, also known as the true positive rate or recall, quantifies the proportion of actual positives (e.g., individuals with a disease, events occurring) that are correctly identified by a test or model.
   * **Application**: Sensitivity is crucial in medical diagnostics, quality control, risk assessment, and any scenario where accurately detecting true positives is essential for decision-making and outcomes.
6. **What is a z-score and what is it used for?** 
   * A z-score is a statistical measurement that describes a value's relation to the mean of a group of values. It is measured in terms of standard deviations from the mean.
   * A z-score is used to determine how unusual a value is, and it's commonly used for hypothesis testing, outlier detection, and comparison of scores from different datasets.
7. **Describe the difference between cross-validation and bootstrapping.**
   * **Cross-validation** is a technique for evaluating the performance of a statistical model by partitioning the original sample into a training set to train the model, and a test set to evaluate it. One common method is k-fold cross-validation.
   * **Bootstrapping**, on the other hand, is a resampling technique used to estimate the distribution of a statistic (like the mean or variance) by sampling with replacement from the data. It helps assess the variability of a sample statistic and construct confidence intervals.
8. **What is the purpose of a control group in an experiment?**
   * A **control group** **serves** as a baseline to compare the effects of the treatment group. It helps to isolate the effect of the treatment and ensure that the observed effects are due to the treatment itself.

**Advanced Level**

1. **Explain Simpson's Paradox and its implications in descriptive statistics**.
   * Simpson's Paradox occurs when a trend or relationship between two variables reverses or disappears when they are examined in the context of a third variable.
   * This can happen due to confounding factors. It emphasizes the importance of considering all relevant factors when interpreting statistical relationships.
2. **Briefly explain the procedure to measure the length of all sharks in the world.**
   * Following steps can be used to determine the length of sharks:
     1. Define the confidence level (usually around 95%)
     2. Use sample sharks to measure
     3. Calculate the mean and standard deviation of the lengths
     4. Determine t-statistics values
     5. Determine the confidence interval in which the mean length lies
3. **What are some of the low and high-bias Machine Learning algorithms?**
   * There are many low and high-bias Machine Learning algorithms, and the following are some of the widely used ones:
     1. Low bias: [SVM](https://intellipaat.com/blog/tutorial/machine-learning-tutorial/svm-algorithm-in-python/), decision trees, KNN algorithm, etc.
     2. High bias: Linear and logistic regression
4. **What are some of the techniques to reduce underfitting and overfitting during model training?**
   * Underfitting refers to a situation where data has high bias and low variance, while overfitting is the situation where there are high variance and low bias.
   * For reducing underfitting:
     1. Increase model complexity.
     2. Increase the number of features.
     3. Remove noise from the data.
     4. Increase the number of trainings epochs
   * For reducing overfitting:
     1. Increase training data.
     2. Stop early while training.
     3. Lasso regularization.
     4. Use random dropouts.
5. **When creating a statistical model, how do we detect overfitting?**
   * Overfitting can be detected by cross-validation.
   * In cross-validation,
     1. We divide the available data into multiple parts and iterate on the entire dataset.
     2. In each iteration, one part is used for testing, and others are used for training.
     3. This way, the entire dataset will be used for training and testing purposes.
     4. And we can detect if the data is being overfitted.
6. **What are the scenarios where outliers are kept in the data?**
   * There are not many scenarios where outliers are kept in the data, but there are some important situations when they are kept. They are kept in the data for analysis if:
     1. Results are critical.
     2. Outliers add meaning to the data.
     3. The data is highly skewed.