LIST OF STATISTICAL CONCEPTS

Descriptive Statistics 3
Graphical methods 4
Probability theory and distributions 5
Central Limit Theorem 6
Statistical Sampling and Estimation 7
Hypothesis Testing 8
Linear Regression 9
Logistic Regression 10
Bayesian Statistics 11
Time Series Analysis 12
Multivariate analysis 13
Non-parametric statistics 14
Experimental design 15
Statistical Data Types 16
Normal distribution 17
Central Tendency 18
Statistical Model 19
Frequency Distributions 20
Type I/II errors 21
Effect Sizes 22
Data Quality 23
Data Integrity 24

Data Timeliness 25
Data Quality Analysis and Data Cleansing Methods 26
Confidence Intervals 27
Statistical Power 28
Correlation 29
Prevalence, Incidence and NNT 30
Factor Analysis 31
Regression 32
Descriptive and exploratory statistical methods 33
Data Bias 34
Data context 35
Data provenance 36
Central Limit Theorem (CLT) 37
Analysis of variance (ANOVA) 38
Data Files and Data Sets 39
Variables and Attributes 40
Saturation and Transferability 41
Statistical Significance v Operational Significance 42
Quantitative, Qualitative and Mixed Methods 43
Incidence Rates 44
Machine Learning Methods 45
Limitations of retrospective analysis and human factors 46

3 - Descriptive Statistics

Descriptive statistics is a branch of statistics that focuses on summarizing and describing the main features of a dataset.

The main tools for descriptive statistics include measures of central tendency (mean, median, mode), measures of dispersion (range, variance, standard deviation), and measures of shape (skewness and kurtosis).

Descriptive statistics provide a quick and intuitive overview of the data and can help identify patterns and relationships.

Return to Index Page

4 - Graphical methods

Graphical methods are an important part of descriptive statistics and provide visual representations of data.

Some commonly used graphical methods include bar plots, histograms, scatter plots, line plots, and box plots.
Graphical methods can help identify outliers, skewness, and other patterns in the data, and are often used to get a preliminary understanding of the data before further statistical analysis.

Together, descriptive statistics and graphical methods provide a valuable toolkit for summarizing, visualizing, and exploring data, and are critical for understanding the underlying structure and patterns in a dataset.

Return to Index Page

5 - Probability theory and distributions

Probability theory is a mathematical framework for modeling uncertain events and quantifying the likelihood of various outcomes. It provides the foundation for much of modern statistics.

A fundamental concept in probability theory is the notion of a random variable, which is a variable whose value is determined by the outcome of a random event. Probability distributions are mathematical models that describe the likelihood of different values of a random variable.

There are many different types of probability distributions, including discrete distributions (such as the binomial and Poisson distributions) and continuous distributions (such as the normal and exponential distributions).

Each distribution has its own unique characteristics and is appropriate for modeling different types of data and processes.

Return to Index Page

6 - Central Limit Theorem

The central limit theorem is an important concept in probability theory that states that the sum of a large number of independent and identically distributed random variables will converge to a normal distribution, regardless of the distribution of the individual variables.

This has important implications for hypothesis testing and estimation.

In summary, probability theory and distributions are important for understanding and quantifying uncertainty in statistical models and for choosing appropriate models for different types of data and processes.

Return to Index Page

7 - Statistical Sampling and Estimation

Sampling and estimation are two key concepts in statistics that are concerned with making inferences about a population - based on a sample of data.

Sampling refers to the process of selecting a subset of observations from a larger population for the purpose of statistical analysis.

The goal of sampling is to obtain a representative sample that accurately reflects the characteristics of the population. There are many different sampling methods, including random sampling, stratified sampling, and cluster sampling, each with their own strengths and limitations.

Estimation is the process of using the information from a sample to make inferences about a population parameter. For example, if we have a sample of heights of individuals and we want to estimate the average height of the population, we can calculate the mean of the sample heights as our estimate. Estimation can be based on point estimates, which provide a single value as an estimate of the population parameter, or interval estimates, which provide a range of values that are likely to contain the population parameter.

Return to Index Page

8 - Hypothesis Testing

Hypothesis testing is a statistical procedure used to test a claim or assumption about a population parameter based on a sample of data. The goal of hypothesis testing is to determine whether the evidence in the sample supports or refutes the claim being made about the population. The process of hypothesis testing typically involves the following steps: Formulate the null hypothesis and the alternative hypothesis: The null hypothesis is a statement of no effect or no difference, while the alternative hypothesis is the statement that we want to test.

Choose a level of significance: The level of significance is the probability of making a type I error, which is rejecting the null hypothesis when it is true. The level of significance is typically set at 0.05, meaning that we are willing to accept a 5% chance of making a type I error.

Calculate a test statistic: The test statistic is a value calculated from the sample data that provides evidence for or against the null hypothesis. Different test statistics are used for different types of data and hypotheses.

Make a decision: Based on the level of significance and the calculated test statistic, we either reject or fail to reject the null hypothesis. If we reject the null hypothesis, we conclude that there is evidence to support the alternative hypothesis.

Draw a conclusion: The conclusion is a statement about the relationship between the sample and the population based on the evidence provided by the test statistic.

Hypothesis testing is a fundamental tool in statistical inference and is used in a wide range of applications, including medical research, engineering, and social sciences. By carefully designing and conducting hypothesis tests, we can make informed decisions about complex problems and draw meaningful conclusions from data.

Return to Index Page

9 - Linear Regression

Linear regression is a statistical method used to model the relationship between a dependent variable and one or more independent variables. The goal of linear regression is to find the line of best fit that describes the relationship between the variables, allowing us to make predictions about the dependent variable based on the independent variables.

Linear regression models are based on the assumption that the relationship between the variables is linear, meaning that the change in the dependent variable is proportional to the change in the independent variables. The line of best fit is determined by minimizing the sum of the squared differences between the observed values of the dependent variable and the values predicted by the line.

Linear regression models can be simple, with only one independent variable, or multiple, with two or more independent variables. In multiple linear regression, the goal is to find the line of best fit that takes into account the effects of all the independent variables on the dependent variable.

Linear regression is a widely used statistical method with many applications in fields such as economics, finance, and engineering. It is a powerful tool for understanding the relationships between variables and for making predictions about future trends and patterns.

Return to Index Page

10 - Logistic Regression

Logistic regression is a statistical method used to model the relationship between a dependent binary variable and one or more independent variables. Unlike linear regression, which is used to model continuous dependent variables, logistic regression is used to model binary dependent variables, where the outcome can only take on one of two possible values (e.g., yes/no, success/failure).

The goal of logistic regression is to estimate the probability of the dependent variable taking on one of the two possible values, given the values of the independent variables. The estimated probabilities are then transformed into binary predictions using a threshold, typically set at 0.5.

In logistic regression, the relationship between the independent and dependent variables is modeled using a logistic function, which maps the predicted probabilities onto the range [0, 1]. The logistic function is used to model the probability of the dependent variable being equal to 1 as a function of the independent variables.

Logistic regression is widely used in fields such as medicine, social sciences, and marketing, where the goal is to model binary outcomes. It is a powerful tool for making predictions about binary outcomes and for understanding the relationships between independent variables and binary dependent variables.

Return to Index Page

11 - Bayesian Statistics

Bayesian statistics is a branch of statistics that uses Bayes' theorem to update the probability for a hypothesis as more evidence or information becomes available. It involves expressing beliefs about events as probabilities and updating those probabilities through Bayesian inference based on new data. This approach allows for the incorporation of prior knowledge and the quantification of uncertainty in the estimates. Bayesian statistics is a branch of statistics that deals with updating the beliefs or prior knowledge about the parameters of a model, based on the observed data. Here are a few examples of Bayesian statistics in action:

Medical Probability: Bayesian statistics can be used to assist medical diagnosis updating the probability of a patient having a certain disease, based on their symptoms and test results.

Image Recognition: Bayesian algorithms can be used in computer vision to recognize objects in an image based on the probabilities of different object classes, given the image features.

Spam Filtering: Bayesian methods can be used in spam filtering to determine the probability of an email being spam, based on its content and other features.

Credit Risk Assessment: Banks can use Bayesian methods to evaluate the credit risk of a loan applicant, based on their financial history, credit score, and other factors.

Recommender Systems: Bayesian algorithms can be used in recommendation systems to predict the probability of a user liking a particular item, based on their past behavior and preferences.

Return to Index Page

12 - Time Series Analysis

Time series analysis is a statistical method used to analyze time-dependent data, such as sales data, stock prices, temperature readings, etc. The goal of time series analysis is to model and understand the underlying patterns and structures in the data, in order to make predictions or decisions.

Techniques used in time series analysis include trend analysis, seasonal analysis, and forecasting models such as ARIMA (Auto Regressive Integrated Moving Average) and Exponential Smoothing. These models take into account the temporal dependencies between data points, as well as external factors such as holidays, economic indicators, etc.

Return to Index Page

13 - Multivariate analysis

Multivariate analysis is a statistical method used to study the relationships among multiple variables or features. It is used to examine the relationships between several variables simultaneously, and to determine the independent and dependent variables and their relationship to one another.

Some common methods of multivariate analysis include multiple regression, principal component analysis, discriminant analysis, and cluster analysis. The choice of method depends on the type of variables (e.g. continuous, categorical) and the research question being addressed. Multivariate analysis is useful in fields such as biology, finance, marketing, and social sciences where understanding the relationship between multiple variables is important.

Return to Index Page

14 - Non-parametric statistics

Non-parametric statistics is a branch of statistics that deals with methods that do not assume any specific distribution for the data being analyzed.

Unlike parametric statistics, which assumes that the data follows a normal or Gaussian distribution, non- parametric methods do not make any assumptions about the underlying distribution of the data. Instead, they rely on ranking, grouping, or randomization to analyze the data.

Examples of non-parametric methods include the Wilcoxon rank-sum test, the Kruskal-Wallis test, and the Mann-Whitney U test. Non-parametric methods are useful when the assumptions of normality are not met, or when the sample size is small, or when the distribution of the data is unknown.

Return to Index Page

15 - Experimental design

Experimental design in statistics refers to the planning and implementation of experiments to collect data for statistical analysis. The goal of experimental design is to obtain valid, reliable and meaningful results from the data collected. The process of experimental design involves several steps including:

1. Defining the research question and hypothesis
2. Selecting the sample and determining the sample size
3. Determining the type of experiment (randomized controlled, observational, etc.)
4. Assigning subjects to the different treatment groups (randomization)
5. Implementing the experiment and collecting the data
6. Analyzing the data and drawing conclusions

Experimental design is critical to the validity of the results because it determines the conditions under which the data will be collected. Proper experimental design can help control extraneous variables, reduce bias, and increase the precision of the results. It also helps to ensure that the results are generalizable to a larger population and that the conclusions drawn are statistically significant.

Return to Index Page

16 - Statistical Data Types

There are four commonly recognized types of statistical data, which are:

Nominal data: This type of data consists of categories or labels that do not have any inherent order or numerical value. Examples of nominal data include gender, race, and type of car.

Ordinal data: This type of data consists of categories or labels that have a natural order or ranking. However, the differences between the categories are not necessarily equal. Examples of ordinal data include income levels, education levels, and survey responses on a scale from "strongly agree" to "strongly disagree."

Interval data: This type of data consists of numerical values that have a fixed interval between them. However, there is no true zero point, and negative values are possible. Examples of interval data include temperature measured in Celsius or Fahrenheit and dates.

Ratio data: This type of data consists of numerical values that have a fixed interval between them and a true zero point, where zero means the complete absence of the variable being measured. Examples of ratio data include height, weight, and income.

Understanding the type of data being analysed is important because different statistical techniques are used depending on the type of data being analysed.

Return to Index Page

17 - Normal distribution

A normal distribution, also known as a Gaussian distribution, is a probability distribution that is symmetric and bell- shaped. It is a continuous probability distribution that describes the distribution of a large number of independent and identically distributed (i.i.d.) random variables that have a mean and a variance.

In a normal distribution, the mean, median, and mode are all equal and are located at the centre of the distribution. The distribution is characterized by two parameters: the mean and the standard deviation. The mean represents the centre of the distribution, while the standard deviation represents the spread of the distribution.

The shape of the normal distribution is determined by the mean and standard deviation. The larger the standard deviation, the wider the distribution, and the flatter the curve. The smaller the standard deviation, the narrower the distribution, and the taller the curve.

The normal distribution is important in statistical analysis because many natural phenomena follow a normal distribution, such as human height, weight, and IQ scores. Additionally, many statistical methods, such as hypothesis testing and confidence intervals, rely on assumptions of normality to make accurate inferences.

In summary the Normal or Gaussian distribution is a probability distribution that is symmetric and bell-shaped. In normal distributions the mean, median, and mode are all equal and are located at the centre of the distribution.

Return to Index Page

18 - Central Tendency

In statistics, central tendency refers to the typical or central value around which a set of data tends to cluster. It is a measure that summarizes the data by identifying the central position of the data set. There are three main measures of central tendency:
Mean: The arithmetic mean is the sum of all the values in the data set divided by the number of values. It is a common measure of central tendency that is often used when the data is normally distributed.

Median: The median is the value that separates the data into two equal halves. It is the middle value in a sorted list of data, and it is often used when the data set has extreme values or outliers that can skew the mean.

Mode: The mode is the value that occurs most frequently in the data set. It is often used when the data set is categorical or when there are multiple peaks in the distribution of the data.

Each measure of central tendency has its own strengths and weaknesses, and the choice of which measure to use depends on the nature of the data and the purpose of the analysis.

Return to Index Page

19 - Statistical Model

A statistical model is a mathematical representation of a real-world phenomenon or system. It is a set of mathematical equations or relationships that describe the underlying structure of the data and the processes that generated it. The purpose of a statistical model is to provide a framework for making inferences and predictions based on the available data.
Statistical models can be used for a wide variety of purposes, such as:
Describing relationships between variables: Statistical models can be used to identify and quantify the relationships between variables in a data set. For example, a linear regression model can be used to describe the relationship between two continuous variables.

Making predictions: Statistical models can be used to make predictions about future observations
based on the available data. For example, a time series model can be used to predict future values of a variable based on its past values.

Testing hypotheses: Statistical models can be used to test hypotheses about the relationships between variables or the effects of interventions. For example, a t-test can be used to test whether the means of two groups are significantly different.

Statistical models can be simple or complex, depending on the nature of the data and the research question. They can also be parametric or nonparametric, depending on whether they make assumptions about the distribution of the data. The choice of model depends on the nature of the data, the research question, and the available resources for analysis.

Return to Index Page

20 - Frequency Distributions

Frequency distributions are a useful tool in statistics for summarizing and visualizing data. They provide a way to organize data into categories or intervals and show how often each category or interval occurs. Here are a few interesting things about frequency distributions:

They provide a quick overview of the data: Frequency distributions allow you to see the range of values in a data set, as well as the frequency or count of each value or category. This can help you quickly identify patterns, trends, or outliers in the data.

They can be used to create histograms and other visualizations: Frequency distributions can be used to create histograms, bar charts, or other visualizations that can help you better understand the distribution of the data. Histograms, in particular, are a popular way to display frequency distributions because they provide a visual representation of the data and can highlight patterns or clusters in the distribution.

They can be used to calculate measures of central tendency and variability: Frequency distributions can be used to calculate measures of central tendency such as the mean or median, as well as measures of variability such as the standard deviation or range. These measures can provide additional information about the distribution of the data and help you make inferences or predictions based on the data.

They can be used to compare data sets: Frequency distributions can be used to compare two or more data sets by plotting them on the same histogram or bar chart. This can help you identify similarities or differences between the data sets and make comparisons or draw conclusions based on the data. Overall, frequency distributions are a powerful tool in statistics that can provide a wealth of information about a data set and help you make informed decisions or draw conclusions based on the data.

Return to Index Page

21 - Type I/II errors

Type I and Type II errors are two types of errors that can occur in statistical hypothesis testing:

Type I error: A Type I error occurs when we reject a null hypothesis that is actually true. In other words, we conclude that there is a significant effect or difference when in fact there is no effect or difference. The probability of making a Type I error is denoted by alpha (α) and is typically set at 0.05 or 0.01 in scientific research.

Type II error: A Type II error occurs when we fail to reject a null hypothesis that is actually false. In other words, we conclude that there is no significant effect or difference when in fact there is an effect or difference. The probability of making a Type II error is denoted by beta (β) and is typically set at 0.20 or 0.10 in scientific research.

The likelihood of making a Type I error and a Type II error are related, and reducing one will increase the other. For instance, if we want to reduce the likelihood of making a Type I error, we need to increase the threshold for rejecting the null hypothesis, which will also increase the likelihood of making a Type II error.

In general, the goal of hypothesis testing is to minimize the probability of making both types of errors. This can be achieved by choosing an appropriate sample size, selecting an appropriate statistical test, and setting the level of significance based on the research question and available evidence. It is important to carefully consider the potential for both types of errors when designing and interpreting statistical analyses in scientific research.

Return to Index Page

22 - Effect Sizes

Effect size is a statistical measure that quantifies the size or magnitude of the difference or relationship between two variables. It is a way to express the practical significance or meaningfulness of a result or finding in a study, beyond simply testing for statistical significance. There are different ways to calculate effect sizes depending on the research question and the type of data. Here are a few examples:

Cohen's d: Cohen's d is a common effect size measure used in studies that compare two groups on a continuous outcome variable. It is calculated by taking the difference between the means of the two groups and dividing it by the pooled standard deviation. A larger value of Cohen's d indicates a larger difference between the groups.

Correlation coefficient: The correlation coefficient is a measure of the strength and direction of the linear relationship between two continuous variables. It ranges from -1 to 1, where a value of 0 indicates no relationship, a positive value indicates a positive relationship, and a negative value indicates a negative relationship. The magnitude of the correlation coefficient indicates the strength of the relationship.

Odds ratio: The odds ratio is a measure of the strength of the association between two categorical variables. It is calculated by taking the ratio of the odds of an event occurring in one group compared to another group. A value greater than 1 indicates that the odds of the event occurring are greater in the first group, while a value less than 1 indicates that the odds are greater in the second group.

Effect sizes are important because they provide information about the practical significance of a result or finding in a study. Statistical significance only indicates whether the result is likely to occur by chance, but effect size provides information about the magnitude of the difference.

Return to Index Page

23 - Data Quality

Data quality refers to the degree to which data is accurate, complete, reliable, relevant, and consistent with its intended purpose. In other words, data quality is a measure of how well data meets the needs of its users and the requirements of its applications. There are several dimensions of data quality that are commonly used to assess the quality of data, including:

Accuracy: the degree to which data is free from errors or mistakes.

Completeness: the degree to which data includes all necessary information and is not missing any relevant data.

Consistency: the degree to which data is internally consistent and does not contain conflicting information.

Relevance: the degree to which data is appropriate and useful for its intended purpose.

Timeliness: the degree to which data is up-to-date and relevant for its intended use.

Ensuring data quality is important for several reasons.

First, it can help ensure that decisions and actions based on data are accurate and effective.

Second, it can help avoid errors, misunderstandings, and misinterpretations that can result from poor quality data.

Third, it can improve the efficiency and effectiveness of data-driven processes, such as analytics, reporting, and decision-making.

Overall, data quality is a critical aspect of data management and analysis, and ensuring high-quality process that requires attention and investment throughout the data lifecycle.

Return to Index Page

24 - Data Integrity

Data integrity refers to the accuracy, consistency, and reliability of data throughout its lifecycle. It means that data is complete, accurate, and free from errors, duplication, or corruption. In other words, data integrity ensures that data is trustworthy and can be used for its intended purposes.

Data integrity is critical in various fields, including finance, healthcare, research, and education, where inaccurate data can have significant consequences. Data integrity is maintained through various means, such as data validation, data verification, and data backup and recovery. Additionally, security measures such as access control, encryption, and authentication help to ensure that data is not tampered with or altered inappropriately.

Return to Index Page

25 - Data Timeliness

Data timeliness refers to the degree to which data is up-to-date and available when needed. It is a measure of how quickly data is collected, processed, and made available for analysis or decision-making.

Timely access to data is critical to making informed decisions. For example, in the financial industry, delays in receiving data can result in missed trading opportunities or increased risk exposure. Similarly, in healthcare, delayed access to patient data can impact the quality of care and patient outcomes.

To ensure data timeliness, organizations must have efficient and reliable data collection, processing, and delivery systems. This requires the use of technology and tools that can automate data collection and processing, as well as ensure data is available in real-time. Additionally, having well-defined data governance policies and procedures can help ensure that data is collected and maintained in a timely and accurate manner.

Return to Index Page

26 - Data Quality Analysis and Data Cleansing Methods

Data quality analysis and data cleansing are two important processes that are essential in ensuring that data is accurate, reliable, and suitable for its intended purpose.

Data quality analysis involves evaluating the accuracy, completeness, and consistency of data to ensure that it is of high quality. This process involves identifying and analyzing errors, inconsistencies, and other issues in the data, such as missing values, duplicate records, and incorrect data formats. Data quality analysis aims to identify data quality issues and determine their root cause, so that appropriate corrective measures can be taken.

Data cleansing, also known as data scrubbing, is the process of correcting or removing errors, inconsistencies, and other issues identified during data quality analysis. Data cleansing involves various techniques such as removing duplicate records, filling in missing values, and correcting formatting errors. Data cleansing aims to ensure that data is accurate, consistent, and complete, and that it is suitable for its intended use.

Some common data cleansing methods include:

Standardization - converting data into a standardized format to ensure consistency and accuracy.

Parsing - breaking data into smaller components to enable easy analysis and comparison.

Data profiling - analyzing the structure and content of data to identify inconsistencies and errors.

Data enrichment - enhancing the quality of data by adding additional information from external sources.

Record linkage - identifying duplicate records and merging them into a single record.

In summary, data quality analysis and data cleansing are crucial processes in ensuring that data is accurate, reliable, and suitable for its intended use.

These processes require a combination of technical expertise, domain knowledge, and attention to detail to ensure that the data is of high quality and can be used for its intended purpose.

Return to Index Page

27 - Confidence Intervals

Confidence intervals are a statistical measure that provides an estimate of the range of values that a population parameter, such as a mean or proportion, is likely to fall within. It is based on the sample data collected from a population and provides a range of values within which the true population parameter is likely to lie, with a certain degree of confidence.

For example, suppose we want to estimate the mean height of a population. We can collect a sample of heights from that population and calculate the mean height of the sample. However, this sample mean may not be exactly equal to the true population mean. A confidence interval provides a range of values that the true population mean is likely to fall within, with a certain level of confidence.

The level of confidence is typically expressed as a percentage, such as 95% or 99%. A 95% confidence interval means that if we were to repeat the sampling process many times, we would expect 95% of the confidence intervals to contain the true population parameter.

The width of the confidence interval is affected by the level of confidence and the sample size. A wider confidence interval indicates greater uncertainty about the true population parameter, while a narrower confidence interval indicates greater precision.

Confidence intervals are important in statistical inference because they provide a measure of the precision and uncertainty associated with estimates of population parameters. They allow researchers to draw conclusions about population parameters and make informed decisions based on the sample data.

Return to Index Page

28 - Statistical Power

Statistical power is a measure of the ability of a statistical test to detect a true effect or difference in the data. It is the probability of correctly rejecting a false null hypothesis. In other words, it is the likelihood of detecting a significant effect or difference if it truly exists in the population.

The statistical power of a test depends on several factors, including the sample size, the level of significance, the effect size, and the variability in the data. Increasing the sample size or the level of significance can increase the statistical power, while increasing the variability in the data can decrease the statistical power.

Examples of statistical power:

1. A clinical trial is designed to test the effectiveness of a new medication. The trial aims to detect a difference in the mean improvement in symptoms between the medication group and the placebo group. The statistical power of the test will depend on the sample size, the level of significance, and the variability in the data.

2. A researcher conducts a study to examine the relationship between two variables, such as age and memory performance. The statistical power of the test will depend on the sample size, the level of significance, the strength of the correlation between the variables, and the variability in the data.

An experimental study is conducted to compare the means of two groups on a continuous outcome variable. The statistical power of the test will depend on the sample size, the level of significance, the effect size, and the variability in the data.

Return to Index Page

29 - Correlation

Statistical correlation refers to the degree to which two or more variables are related to each other. It is a measure of the strength and direction of the linear relationship between two or more variables. A positive correlation means that as one variable increases, the other variable also tends to increase. A negative correlation means that as one variable increases, the other variable tends to decrease.

Statistical correlation can be measured using a correlation coefficient, which is a numerical value that ranges from -1 to +1. A correlation coefficient of +1 indicates a perfect positive correlation, while a correlation coefficient of -1 indicates a perfect negative correlation. A correlation coefficient of 0 indicates no correlation between the variables.

There are different types of correlation coefficients, including the Pearson correlation coefficient, the Spearman correlation coefficient, and the Kendall correlation coefficient.

The Pearson correlation coefficient is used to measure the strength of the linear relationship between two continuous variables, while the Spearman and Kendall correlation coefficients are used to measure the strength of the monotonic relationship between two variables.

Statistical correlation can be useful in many areas of research, including psychology, sociology, economics, and medicine. For example, researchers might use correlation to examine the relationship between two psychological constructs, such as depression and anxiety. They might also use correlation to explore the relationship between economic indicators, such as inflation and unemployment.

However, it is important to note that correlation does not imply causation, and other factors may be influencing the relationship between the variables.

Return to Index Page

30 - Prevalence, Incidence and NNT

Prevalence, incidence, and NNT (number needed to treat) are all important measures in epidemiology and clinical research.

Prevalence: Prevalence refers to the proportion of individuals in a population who have a particular disease or condition at a specific point in time or during a specified time period. It is usually expressed as a percentage or a proportion of the population. Prevalence is a useful measure for understanding the burden of a disease or condition within a population.

Incidence: Incidence refers to the number of new cases of a disease or condition that occur within a specified time period in a population that is initially free of the disease or condition. It is usually expressed as a rate, such as the number of cases per 1,000 or 100,000 people per year. Incidence is a useful measure for understanding the risk of developing a disease or condition within a population.

NNT (number needed to treat): NNT is a measure of the effectiveness of a treatment. It represents the number of people who need to be treated with a particular intervention for one person to benefit from the treatment. NNT is typically calculated based on the results of randomized controlled trials (RCTs) and is often used to compare the effectiveness of different treatments for the same condition. The lower the NNT, the more effective the treatment is considered to be.

In summary, **prevalence** measures the proportion of individuals in a population with a particular disease or condition at a specific point in time or during a specified time period, while **incidence** measures the number of new cases of the disease or condition within a population over a specified time period. **NNT** is a measure of the effectiveness of a treatment, representing the number of people who need to be treated for one person to benefit from the treatment.

Return to Index Page

31 - Factor Analysis

Factor analysis is a statistical method used to explore the relationships among a large number of variables by identifying underlying factors that the patterns of correlation within the data. It is a method used to reduce the complexity of a large data set by identifying the underlying factors that contribute to the observed correlations among the variables.

In factor analysis, a set of observed variables is analysed to identify a smaller set of unobserved variables, called factors, that capture the common variance among the observed variables. These factors are assumed to represent underlying constructs or dimensions that are not directly observable but are related to the observed variables.

Factor analysis is commonly used in fields such as psychology, sociology, and marketing to explore the relationships among a large number of variables and to identify the underlying dimensions or constructs that contribute to those relationships. For example, in psychology, factor analysis might be used to identify the underlying dimensions of a personality test or to explore the structure of a set of behavioural measures.

There are several methods of factor analysis, including principal component analysis, exploratory factor analysis, and confirmatory factor analysis. Each method has its own strengths and limitations, and the choice of method will depend on the research question and the characteristics of the data set.

Return to Index Page

32 - Regression

Statistical regression is a statistical method that is used to explore the relationship between a dependent variable and one or more independent variables. It is a technique used to model the relationship between two or more variables in order to predict or estimate the value of the dependent variable based on the values of the independent variables.

Regression analysis involves estimating the relationship between the dependent variable and the independent variable(s) by fitting a regression model to the data. The regression model is typically a mathematical equation that includes one or more independent variables and a coefficient for each independent variable that represents the strength of the relationship between that variable and the dependent variable.

The most common type of regression analysis is linear regression, which assumes a linear relationship between the dependent variable and the independent variable(s). Non-linear regression models can also be used to capture more complex relationships between the variables.

Regression analysis can be used for a variety of purposes, such as predicting the value of a dependent variable based on the values of the independent variables, identifying the important predictors of the dependent variable, and testing the significance of the relationship between the variables. It is widely used in many fields, including economics, finance, psychology, and engineering, among others.

Overall, regression analysis is a powerful statistical tool that can be used to study the relationship between variables. It is used to make predictions and test hypotheses. It is widely used in many fields, including economics, finance, healthcare, social sciences etc.

Return to Index Page

33 - Descriptive and exploratory statistical methods

Descriptive and exploratory statistical methods are two types of statistical analysis techniques used to analyse data.

Descriptive statistics involves summarizing and describing the main features of a dataset, such as measures of central tendency (mean, median, mode), measures of variability (standard deviation, range), and graphical displays (histograms, box plots, scatter plots). The primary goal of descriptive statistics is to describe the data and provide insights into the distribution and characteristics of the dataset.

On the other hand, exploratory statistics involves analysing data to discover patterns, relationships, and trends that may not be immediately apparent. Exploratory statistics techniques include clustering, principal component analysis, factor analysis, and correlation analysis. The primary goal of exploratory statistics is to gain a deeper understanding of the underlying structure of the dataset, identify potential outliers or anomalies, and generate hypotheses for further investigation.

In summary, descriptive statistics provides a summary of the data while exploratory statistics seeks to uncover hidden relationships within the data that may not be immediately apparent. Both techniques are important in the statistical analysis of data, and they are often used in conjunction with each other.

Return to Index Page

34 - Data Bias

Data bias refers to the presence of systematic errors or distortions in data that can lead to inaccurate or unfair conclusions or decisions. It can result from various factors, including the method of data collection, the selection of participants, or the interpretation of data.

Data bias can have significant implications in various fields, such as healthcare, finance, education, and criminal justice. For example, if a clinical trial only includes participants from a particular age group, gender, or ethnicity, the study's results may not be generalizable to other populations.

Similarly, if a dataset used for training an AI model has inherent biases, such as gender or racial biases, the model's predictions may reflect these biases and result in unfair decisions.

To minimize data bias, it is essential to ensure that data collection methods are unbiased, and data is representative of the population of interest. Additionally, using diverse teams to analyze data and developing algorithms that account for potential biases can help mitigate the impact of data bias.

Return to Index Page

35 - Data context

Data context refers to the circumstances, conditions, and environment surrounding data, which can influence its meaning, relevance, and interpretation. In other words, data context provides information about the broader context in which data was created, collected, and analysed, including the people, processes, and systems involved.

Data context can include various types of information, such as metadata, data sources, data quality, data lineage, and more. For example, metadata can provide information about the structure, format, and meaning of data, while data sources can provide information about the origin and provenance of data. Data quality information can provide insight into the accuracy, completeness, and consistency of data, while data lineage can provide a record of the transformation and processing of data over time.

Understanding data context is important for several reasons. First, it can help ensure the accuracy and reliability of data by providing insight into its meaning and interpretation. Second, it can help improve data integration and interoperability by providing a common understanding of data across different systems and organizations. Third, it can help improve data governance and compliance by ensuring that data is managed and used in accordance with relevant regulations and standards.

Overall, data context is an important concept in data management and analysis, particularly in contexts where data quality, interoperability, and governance are critical, such as healthcare, finance, and scientific research.

Return to Index Page

36 - Data provenance

Data provenance refers to the history of data, including its origin, processing, and transformation, as well as the people, systems, and processes involved in these activities. In other words, data provenance provides information about how data was created, modified, and used over time.

Data provenance is important for several reasons.

First, it can help ensure the accuracy, completeness, and quality of data by providing information about its origin and how it was processed.

Second, it can help improve transparency and accountability by allowing users to trace the flow of data through different systems and processes.

Third, it can help with regulatory compliance and auditing by providing a record of data usage and modifications.

There are several approaches to capturing and managing data provenance, including manual documentation, logging, and automated systems. Manual documentation involves creating and maintaining a record of data activities, while logging involves capturing data events automatically in real-time.

Automated systems, such as provenance-aware databases and workflow management systems, can capture data provenance information automatically as part of their functionality.

Overall, data provenance is an important concept in data management and analysis, particularly in contexts where data quality, transparency, and accountability are critical, such as scientific research, healthcare, and financial services.

Return to Index Page

37 - Central Limit Theorem (CLT)

The CLT is a fundamental result in probability theory and statistics that states that, under certain conditions, the distribution of the sum or mean of a large number of independent, identically distributed random variables will tend to a normal distribution, regardless of the underlying distribution of the variables themselves.

Return to Index Page

38 - Analysis of variance (ANOVA)

Analysis of variance (ANOVA) is a statistical method used to test the difference between the means of two or more groups. It is used to determine whether there is a significant difference in the population means or whether the differences observed in the sample are due to random chance.

ANOVA uses a partitioning of the total sum of squares, which measures the total variation in the dependent variable, into the variation between the groups (between-group variation) and the variation within the groups (within-group variation). The goal of ANOVA is to determine whether the between-group variation is large enough to be considered significant, compared to the within-group variation.

There are several types of ANOVA, including one-way ANOVA, which compares the means of two or more groups with one independent variable, and two-way ANOVA, which compares the means of two or more groups with two independent variables. In addition, there are also repeated measures ANOVA and mixed-design ANOVA, which are used to analyze data with a within-subjects and between-subjects design, respectively.

ANOVA is a widely used statistical method with applications in many fields, including biology, psychology, and marketing. It is a powerful tool for testing hypotheses about the difference in means between groups and for making inferences about population parameters based on sample data.

Return to Index Page

39 - Data Files and Data Sets

A data file and a data set are both related to data management and analysis, but they refer to slightly different things.

A data file is a computer file that contains data stored in a specific format. It can be in various formats such as CSV, Excel, or text file. A data file can be created by the user or generated by an application, and it can contain various types of data such as numeric, text, or date/time. Data files are used to store, manage, and exchange data between different software applications, systems, or devices.

A data set, on the other hand, refers to a collection of related data that has been organized and prepared for analysis. A data set can be created by collecting data from various sources, cleaning and transforming it into a standardized format, and then analysing it. A data set can include variables or attributes, which are characteristics or properties of the data being analysed. Data sets can be small or large and can be analysed using various statistical or machine learning techniques.

In summary, a data file is a computer file that contains data, while a data set is a collection of related data that has been prepared for analysis.

Return to Index Page

40 - Variables and Attributes

Variables and attributes are terms used in statistics and research to describe the characteristics or properties of the data being analysed.

A variable is a characteristic that can vary among individuals, groups, or objects being studied. For example, age, gender, height, weight, income, and education level are all examples of variables. Variables can be either continuous, such as age or weight, which can take on any value within a range, or categorical, such as gender or education level, which take on a limited number of values or categories.

An attribute, on the other hand, is a specific value or category that a variable can take on. For example, "male" and "female" are the attributes of the variable "gender," and "high school diploma," "bachelor's degree," and "master's degree" are the attributes of the variable "education level."

In statistical analysis, variables and attributes are used to describe and summarize the data being analysed. Different statistical methods are used for analysing variables and attributes depending on their type and the research question being investigated.

For example, descriptive statistics, such as means and standard deviations, are often used for analysing continuous variables, while frequency distributions and contingency tables are often used for analysing categorical variables.

Return to Index Page

41 - Saturation and Transferability

In qualitative research, saturation and transferability are important concepts used to assess the quality and rigor of the research findings.

Saturation refers to the point at which new data no longer provide additional insights or information on the research topic. It is the point where the researcher has collected enough data that they start to see repetition and redundancy in the data. Saturation is usually achieved when the researcher has collected enough data to represent the diversity of experiences, perspectives, and contexts related to the research question.
Transferability refers to the extent to which the findings of the study can be applied or transferred to other settings or contexts.

Transferability is important because qualitative research typically involves a small sample size and specific context, which limits the generalizability of the findings to other populations or settings. To enhance transferability, researchers need to provide detailed descriptions of the study context, sampling strategy, data collection, and analysis methods. They should also compare their findings with existing literature to identify similarities and differences in the findings.

Return to Index Page

42 - Statistical Significance v Operational Significance

Statistical significance and operational significance are two concepts that are often used in research and statistical analysis. Here is an explanation of each:

Statistical significance: Statistical significance refers to the likelihood that a result or finding is not due to chance. In other words, if a finding is statistically significant, it means that there is a low probability that the result occurred by chance. Statistical significance is typically determined by calculating a p-value, which measures the probability of obtaining a result as extreme or more extreme than the observed result, assuming that there is no true effect in the population.

Operational significance: Operational significance, also known as practical significance, refers to the magnitude or importance of a result or finding in the real world. In other words, even if a finding is statistically significant, it may not be practically significant if the effect size is small or if the finding does not have a meaningful impact on the subject of interest.

To summarize, statistical significance refers to the likelihood that a result is not due to chance, while operational significance refers to the practical importance or impact of a result in the real world. A finding can be statistically significant but not operationally significant if the effect size is small or if the finding does not have a meaningful impact.

Return to Index Page

43 - Quantitative, Qualitative and Mixed Methods

Quantitative, qualitative, and mixed methods are three different approaches to research. Here is a brief explanation of each:

Quantitative research: Quantitative research is a type of research that focuses on collecting and analysing numerical data. This data can be collected through surveys, experiments, observations, or other methods. The goal of quantitative research is to identify patterns, relationships, and trends in the data. This approach typically involves statistical analysis to help draw conclusions from the data.

Qualitative research: Qualitative research is a type of research that focuses on collecting and analysing non-numerical data. This data can be collected through interviews, focus groups, observations, or other methods. The goal of qualitative research is to gain an in-depth understanding of the experiences, perspectives, and meanings behind the data. This approach typically involves coding and categorizing the data to identify themes and patterns.

Mixed methods research: Mixed methods research is a type of research that combines elements of both quantitative and qualitative research. This approach involves collecting and analysing both numerical and non-numerical data. The goal of mixed methods research is to provide a more comprehensive understanding of the research question by drawing on the strengths of both quantitative and qualitative methods.

To summarize, quantitative research focuses on numerical data, qualitative research focuses on non-numerical data, and mixed methods research combines elements of both quantitative and qualitative research.

Return to Index Page

44 - Incidence Rates

In statistics, incidence rate is a measure of how frequently a new event or occurrence happens within a specific population during a given period of time.
Incidence rate is often used in epidemiology to measure the frequency of new cases of a particular disease or health condition within a population. It is calculated by dividing the number of new cases of the disease or condition by the total population at risk during the same time period, and then multiplying by a factor (usually 100,000) to express the rate per unit of population.

For example, if a population of 10,000 people experienced 100 new cases of a disease over the course of a year, the incidence rate would be (100 / 10,000) x 100,000 = 1,000 cases per 100,000 population per year.

Incidence rate is an important statistic because it provides a standardized way to compare the frequency of new events or occurrences across different populations or time periods. It can also be used to identify risk factors for a particular disease or condition, and to monitor the effectiveness of interventions aimed at reducing its incidence.

Return to Index Page

45 - Machine Learning Methods

Machine learning methods and metrics are critical components in the development and evaluation of machine learning models.

Machine learning involves developing algorithms and models that can learn from data and make predictions or decisions based on that data. Machine learning models are trained using historical data and are then used to make predictions or decisions on new, unseen data.

There are several types of machine learning methods, including supervised learning, unsupervised learning, and reinforcement learning. In supervised learning, the model is trained on labeled data, where the correct outcomes are known. In unsupervised learning, the model is trained on unlabelled data, where the outcomes are unknown. In reinforcement learning, the model learns by interacting with an environment and receiving feedback in the form of rewards or penalties.

To evaluate the performance of machine learning models, various metrics are used, including accuracy, precision, recall, F1 score, and ROC AUC. These metrics are used to assess how well the model is performing on the data, and they provide insights into the strengths and weaknesses of the model. Accuracy measures the percentage of correct predictions made by the model. Precision measures the proportion of true positives out of all predicted positives, while recall measures the proportion of true positives out of all actual positives. The F1 score is a harmonic mean of precision and recall, and it provides a balanced measure of the model's performance. The ROC AUC metric measures the ability of the model to distinguish between positive and negative samples and is commonly used in binary classification problems.

In summary, machine learning methods are used to develop models that can learn from data and make predictions or decisions. Metrics are used to evaluate the performance of the models, and they provide insights into the strengths and weaknesses of the model.

Return to Index Page

46 - Limitations of retrospective analysis and human factors

Retrospective analysis and human factors in modelling are two important aspects of data analysis and modelling, but they are not without limitations.

Retrospective analysis, also known as historical analysis, involves analysing past data to gain insights and make predictions about future events. One of the inherent limitations of retrospective analysis is that it relies on historical data, which may not be representative of future events. Changes in technology, consumer behaviour, or the economy can make past data obsolete, which can lead to inaccurate predictions. Another limitation of retrospective analysis is that it can be influenced by data bias. Historical data may be incomplete, inaccurate, or biased, which can lead to flawed analysis and incorrect predictions. This can occur when data is collected using biased methods or when data is incomplete, leading to incomplete analysis and incorrect conclusions.

Human factors in modelling refer to the subjective judgments and biases that can influence the development and interpretation of models. Human factors can arise from the assumptions made in the model, the selection of input variables, and the interpretation of the model outputs. For example, human factors can lead to overfitting, where the model is overly complex and fits the historical data too closely, leading to poor performance on new data.Another human factor is the interpretation of the model outputs. Models can produce complex outputs that can be difficult to interpret, and human biases can influence the interpretation of the results. This can lead to incorrect conclusions and actions based on flawed analysis.

In summary, retrospective analysis and human factors in modelling have inherent limitations that can impact the accuracy and reliability of the analysis and models. It is important to be aware of these limitations and to take steps to minimize their impact on the analysis and modelling process. This can include using diverse data sources, using unbiased data collection methods, and being aware of potential biases in the modelling process.

Return to Index Page