Notes on Descriptive and Inferential Statistics by Suraj

Descriptive Statistics

It is the branch of statistics that deals with the collection, analysis and interpretation of data. It involves the use of various techniques to summarize and describe the important features of a dataset.

The goal of descriptive statistics is to provide a clear and concise summary of the data that can be easily understood and interpreted. It can be used to gain insights into a dataset and to identify pattern and trends. It can also be used to identify outliers, assess the normality of the data, and identify potential sources of bias or errors in the data.

Inferential Statistics

Inferential statistics is a branch of statistics that involves using statistical methods to draw conclusions or make inferences about a population based on a sample of data. The main goal of inferential statistics is to make generalizations about a population based on the analysis of a smaller group of individuals or data points.

Inferential statistics is used to test hypotheses and make predictions about the behavior of a population based on the behavior of a sample from that population. It involves calculating probabilities and using statistical tests to determine whether the results of a sample are likely to be representative of the population as a whole.

Examples of inferential statistics include hypothesis testing, confidence intervals, and regression analysis.

Table of Contents

5 Normal distribution and Z-scores.

6 Confidence intervals for means and proportions.

7 Hypothesis testing for means and proportions.

8 Contigency tables and Chi-Square test of independence.

9 Analysis of variance(ANOVA)

10 Non parametric tests(e.g - Mann-Whitney U Test, Wilcoxon Signed-rank test)

11 Time series analysis(e.g. - moving averages, exponential smoothing, seasonal decomposition, ARIMA models, AR models)

12 Index numbers and inflation.

13 Sampling distributions.

14 Type I and Type II errors.

15 Power of a statistical test.

16 Paired t-test.

17 Forecasting methods(e.g- exponential smoothing, regression analysis)

18 Quality control and process improvements(e.g - control charts, six sigma)

19 Data visualisation (e.g - heat maps, tree maps, network diagrams)

20 Data cleaning and preprocessing.

21 Missing data analysis and imputation.

22 Outlier detect and treatment.

23 Principal component analysis.

24 Factor analysis.

25 Cluster analysis.

26 Discriminant analysis.

27 Multidimensional scaling(MDS)

28 Correspondence analysis.

29 Survival analysis(e.g - Kaplan-Meier estimator, Cox proportional hazards)

30 Bayesian statistics.

31 Machine learning for descriptive statistics(e.g- Linear regression, Multiple regression, Decision Trees and Random Forests).

32 Data visualisation using advanced techniques (e.g - interactive visualizations, network graphs, 3D plots)

33 T-Distribution.

1. Measures of central tendency

Measures of central tendency are statistical measures that describe where the center of a distribution of data is located. The three most commonly used measures of central tendency are the mean, median, and mode.

Mean: The mean is the arithmetic average of a dataset, and is calculated by adding up all the values in the dataset and then dividing by the number of values. For example, if we have the following dataset: 4, 6, 8, 10, 12, the mean would be (4 + 6 + 8 + 10 + 12) / 5 = 8.

Median: The median is the middle value in a dataset when the values are arranged in numerical order. If the dataset has an odd number of values, the median is the middle value. For example, if we have the following dataset: 4, 6, 8, 10, 12, the median would be 8. If the dataset has an even number of values, the median is the average of the two middle values. For example, if we have the following dataset: 4, 6, 8, 10, the median would be (6 + 8) / 2 = 7.

Mode: The mode is the value in a dataset that occurs most frequently. For example, if we have the following dataset: 4, 6, 8, 10, 12, 8 is the mode because it appears twice, which is more than any other value.

2. Measures of Variability

Measures of variability, also known as measures of dispersion, are statistical measures that describe how spread out or dispersed a set of data is. They are useful in understanding the distribution of a dataset, and can help in identifying outliers and other anomalies. The most commonly used measures of variability are range, variance, and standard deviation.

3. Histogram, Box plot, Stem and leaf display

Histogram A histogram is a graphical representation of the frequency distribution of a dataset. It is a way of showing the distribution of data by grouping the data into intervals or bins, and representing the frequency or count of each bin by a bar. The x-axis represents the intervals, while the y-axis represents the frequency or count.

They are commonly used to visualise the distribution of data in various fields.

Box Plot: A box plot is a graphical representation of the distribution of data through five statistics: minimum, first quartile (Q1), median, third quartile (Q3), and maximum. The box itself represents the interquartile range (IQR), which is the distance between Q1 and Q3. The whiskers represent the range of the data, excluding outliers, and the dots or circles represent the outliers.

Interpretation of Box Plot: The box plot allows us to quickly see the range of the data and identify any outliers. The length of the box shows the range of the middle 50% of the data. If the box is short, the data is tightly clustered, while if it is long, the data is more spread out. The median (middle line of the box) indicates the center of the data. The whiskers show the spread of the data, and any points beyond the whiskers are considered outliers.

Stem and Leaf display: Stem and Leaf display is a graphical representation of a dataset, where each data point is split into two parts: the stem and the leaf. The stem is the first digit (or digits) of the data point, and the leaf is the last digit. The stems are then arranged in a vertical column, and the corresponding leaves are listed next to each stem.

Interpretation of Stem and Leaf display: A stem and leaf display can provide quick insights into the distribution of the data. The stems give an idea of the general range of the data, while the leaves show the frequency of each data point. By looking at the stem and leaf display, one can see if the data is skewed, if there are any outliers, and if there are any patterns in the data. The display can be especially useful for smaller datasets, where patterns may be harder to discern in a histogram or box plot.

4. Measures of association

Measures of association are statistical techniques used to determine the strength and direction of the relationship between two or more variables. They are used to explore the degree to which changes in one variable are associated with changes in another variable. The commonly used measures of association are correlation and regression.

Correlation -

Correlation is a measure of association between two continuous variables. It indicates the strength and direction of the linear relationship between the two variables. Correlation can range from -1 to 1, with -1 indicating a strong negative correlation, 0 indicating no correlation, and 1 indicating a strong positive correlation.

Regression :

Regression is a statistical technique used to model the relationship between a dependent variable and one or more independent variables. It helps to predict the value of the dependent variable based on the value of one or more independent variables.

5. Normal distribution and z scores

Normal Distribution, also known as the Gaussian distribution or bell curve, is a probablity distribution that is widely used in statistics to model real-world phenomena such as height, weight, and test scores. The distribution has a symmetric, bell-shaped curve and is characterized by two parameters: the mean and the standard deviation.

Z-score, also known as standard score, is a measure of how many standard deviations an observation or data point is away from the mean of the population. Z-scores are often used in statistical analysis to compare individual observations or data points across different datasets.

6. Confidence interval for means and proportions

Confidence intervals are a statistical tool used to estimate the range of values within which a population parameter is likely to fall with a certain level of confidence. In particular, confidence intervals for means and proportions are used to estimate the range of values within which the true mean or proportion of a population is likely to fall based on a sample.

Confidence intervals for means:
Suppose we want to estimate the average weight of all apples produced by a certain farm, but it is not feasible to measure the weight of all apples. We can randomly select a sample of apples from the farm and compute the sample mean weight. However, the sample mean weight is likely to differ from the true population mean weight. We can construct a confidence interval to estimate the range of values within which the true population mean weight is likely to fall.

Confidence intervals for proportions: Suppose we want to estimate the proportion of customers who are satisfied with a certain product, but it is not feasible to survey all customers. We can randomly select a sample of customers and compute the sample proportion of satisfied customers. However, the sample proportion is likely to differ from the true population proportion. We can construct a confidence interval to estimate the range of values within which the true population proportion is likely to fall.

7. hypothesis testing for means and proportions

Hypothesis testing is a statistical method used to make decisions based on data, where we compare a null hypothesis and an alternative hypothesis. The null hypothesis represents the default assumption, while the alternative hypothesis represents the claim that we want to test.

When it comes to hypothesis testing for means and proportions, we typically use a z-test or a t-test for means and a chi-squared test for proportions. The steps for hypothesis testing include:

State the null and alt. hypothesis.
Determine the appropriate test statistic and level of significance
calc. test statistics.
Determine the p-value.
Make decision and interpret the results

8. Contigency tests and chi-square test of independence

Contingency tables and Chi-Square are statistical tools used to analyze the relationship between two categorical variables.

A contingency table is a table that displays the frequency distribution of two or more categorical variables. Each cell in the table represents the frequency count of a specific combination of values from the two variables. Contingency tables are also known as cross-tabulations.

The chi-square test is a statistical test used to determine if there is a significant association between two categorical variables in a contingency table. It compares the observed frequencies in the contingency table with the expected frequencies under the null hypothesis of no association.

9. Analysis of variance (ANOVA)

ANOVA (Analysis of Variance) is a statistical technique used to compare means of two or more groups of data. It helps to determine whether there are any statistically significant differences between the means of the groups.

Use cases

In medicine, it can be used to compare the effectiveness of different treatments on a particular disease.

In education, it can be used to compare the performance of diff. schools.

In manufacturing, it can be used to compare the quality of products manuf. by different production lines.

10. Non parametric tests(e.g - Mann-Whitney U Test, Wilcoxon signed-rank test)

Non-parametric tests are statistical tests that do not make any assumptions about the underlying distribution of the data. These tests are used when the data does not follow a normal distribution or when the sample size is small. Two commonly used non-parametric tests are the Mann-Whitney U test and the Wilcoxon signed-rank test.

Mann-Whitney U Test:

The Mann-Whitney U test is a non-parametric test used to compare two independent samples. It is also known as the Wilcoxon rank-sum test. The test is used to determine whether the two samples come from populations with the same distribution. The null hypothesis for the test is that the two populations have the same distribution.

Use cases - In medical, it is used to compare the effectiveness of two treatments.

Wilcoxon signed-rank test:

The Wilcoxon signed-rank test is a non-parametric test used to compare two related samples. It is used to determine whether the differences between the two samples are significant. The null hypothesis for the test is that there is no difference between the two populations.

11. Time series analysis

Time series analysis is a statistical method used to analyze data that is measured over time. It helps to identify patterns and trends in the data and make predictions based on past

observations. There are various methods used in time series analysis such as moving averages, exponential smoothing, seasonal decomposition, ARIMA models, and AR models.

Moving averages: Moving averages is a method used to smooth out variations in a time series data by calculating the average of a specified number of consecutive observations.

Exponential smoothing: Exponential smoothing is a method used to forecast time series data that assumes the future value of a variable is a weighted average of past observations, with the weights decreasing exponentially as the observations get older.

Seasonal decomposition: Seasonal decomposition is a method used to identify and separate the seasonal component, trend component, and random component of a time series data.

ARIMA models: Autoregressive Integrated Moving Average (ARIMA) models are used to analyze time series data that is not stationary. It involves differencing the data to make it stationary and fitting the model to the differenced data.

AR models: Autoregressive (AR) models are used to analyze time series data that is stationary. It involves using past values of the variable to predict future values.

Use cases:

Predicting future sales of a company based on past sales data.
Forecasting stock prices based on past prices and other economic factors.
Analyzing website traffic to identify trends and make predictions about future traffic.
Forecasting demand for products based on historical sales data.
Predicting the weather based on past weather patterns.

12. Index numbers inflation

Index numbers are statistical measures that allow us to track changes in the value of a variable over time.

Inflation is the rate at which the general level of prices for goods and services is rising, and subsequently, purchasing power is falling.

Let's say we have data on the prices of different commodities over the years. We can calculate the price index for each year, which will give us an idea of the changes in prices over time. The formula for calculating the price index is:

Price index = (Price of commodity in current year / Price of commodity in base year) x 100

13. Sampling distribution

In statistics, a sampling distribution is the probability distribution of a statistic based on a random sample. In other words, it is the distribution of all possible values of a statistic that could be obtained from samples of a given size taken from a population.

Sampling distributions are important in statistical inference, as they allow us to make inferences about a population based on a sample. By understanding the properties of sampling distributions, we can estimate population parameters with a certain degree of confidence and test hypotheses about the population.

14. Type I and 2 errors

Type I and Type II errors are concepts used in hypothesis testing to describe the errors that can occur when making a decision based on the results of a statistical test.

Type I error occurs when the null hypothesis is rejected even though it is true. This means that the test incorrectly identifies a significant difference when in fact there is none. Type I error is also known as a false positive.

Type II error occurs when the null hypothesis is not rejected even though it is false. This means that the test fails to identify a significant difference when in fact there is one. Type II error is also known as a false negative.

15. Power of a statistical test

In statistics, the power of a statistical test is the probability of rejecting a null hypothesis when it is actually false (i.e., correctly detecting a real effect). In other words, it is the probability of correctly identifying a significant difference or relationship in a study.

The power of a statistical test is influenced by several factors, including the sample size, effect size, level of significance, and variability of the data.

16. Paired t-test

Paired t-test is a statistical test used to determine whether there is a significant difference between two related samples. It is used when the samples being compared are not independent of each other, i.e., they are paired. In other words, the paired t-test is used when the same individuals are measured or tested twice, and the differences between the two measurements or tests are being evaluated.

In a paired t-test, the null hypothesis is that there is no significant difference between the two related samples, and the alternative hypothesis is that there is a significant difference.

17. Forecasting methods

Forecasting methods are techniques used to predict future values of a time series based on past observations. There are several forecasting methods, including exponential smoothing, regression analysis, and ARIMA modeling, among others.

Exponential smoothing is a popular forecasting method used for time series analysis. It involves using a weighted average of past observations to predict future values. The weights assigned to each observation decrease exponentially as the observations get older. Exponential smoothing is useful when there is a trend or a seasonal pattern in the data.

Regression analysis is another commonly used forecasting method that involves fitting a regression model to the time series data. This method can be used to identify relationships between the

dependent variable (the time series data) and one or more independent variables (e.g., time, economic indicators) that may affect the time series. Once the regression model is fit, it can be used to predict future values of the time series.

It can be used to forecast future sales, inventory levels, or demand for a particular product, among other things.

18. quality control proc improvement

Quality control and process improvement are essential components of any production or manufacturing process. They help organizations maintain consistent quality and improve processes to reduce waste, increase efficiency, and minimize defects. Two popular methods for achieving these goals are control charts and Six Sigma.

Control charts are statistical tools used to monitor and control a process. They help identify when a process is out of control and provide information for making adjustments. A control chart consists of a series of data points plotted in time order, with upper and lower control limits calculated based on the process data. The chart helps identify trends, cycles, and patterns in the data and distinguish between normal and abnormal variations in the process.

Six Sigma is a data-driven approach to process improvement that seeks to minimize defects and variations in a process. It uses a systematic approach to identify, measure, and eliminate defects in the process. Six Sigma involves defining the problem, measuring the process, analyzing the data, improving the process, and controlling the process to maintain the improvements.

19. Data visualization

Data visualization is the process of representing data in graphical or pictorial format. It is an effective way of summarizing and communicating large amounts of data in a way that is easy to understand and interpret.

There are many types of data visualizations available in Python, including heat maps, tree maps,

and network diagrams. Each type of visualization has its own strengths and weaknesses, and can be used in different ways to explore and communicate different types of data.

One popular Python library for data visualization is Matplotlib, which provides a wide range of tools for creating high-quality charts, graphs, and other visualizations.

Tree maps, on the other hand, are a type of data visualization that displays hierarchical data using nested rectangles. Each rectangle represents a node in the tree, and its size and color can be used to encode additional information about the data.

Network diagrams are another type of data visualization that can be used to represent complex networks, such as social networks or computer networks. Nodes in the network are represented by circles or squares, and edges between nodes are represented by lines or arrows.

20. Data cleaning and pre-processing

Data cleaning and pre-processing refer to the process of preparing raw data for analysis by identifying and correcting or removing inaccurate, incomplete, or irrelevant data. The goal is to transform the data into a usable and consistent format for analysis.

Data cleaning and pre-processing are essential steps in any data analysis project. It helps to ensure the accuracy and consistency of the data, which in turn leads to more accurate insights and better decision-making. Some real-world applications include:

Healthcare: In healthcare, data cleaning and pre-processing can be used to identify and remove errors in patient data, ensuring that doctors and nurses have access to accurate information for diagnosis and treatment.

Retail: In the retail industry, data cleaning and pre-processing can be used to identify and remove duplicate records, ensuring that customer data is accurate and up-to-date.

Finance: In finance, data cleaning and pre-processing can be used to identify and remove outliers,

ensuring that financial data is accurate and reliable for investment decisions.

21. Missing data analysis and imputation

Missing data analysis is the process of identifying and handling missing values in a dataset. It is a critical step in data pre-processing, as missing values can cause bias, errors, and inaccuracies in statistical analysis and machine learning models.

22. Outlier detection and treatment

Outlier detection is the process of identifying and handling data points that are significantly different from the rest of the data. These data points can skew the results of analysis and modeling, and it's important to either remove them or treat them in a way that doesn't affect the analysis.

23. Principal component analysis

PCA (Principal Component Analysis) is a dimensionality reduction technique used to reduce the number of variables in a dataset while preserving most of the important information. It involves transforming the original variables into a new set of variables, called principal components, which are uncorrelated and capture the maximum amount of variation in the data.

24. Factor analysis

Factor analysis is a statistical method used to identify underlying latent variables, or factors, that explain the correlation structure among a set of observed variables.

25. Cluster analysis

Cluster analysis is a method used to group similar objects or observations into clusters based on their similarities or distances. The objective is to create homogeneous groups that differ from each other, and thus, the observations within each cluster should be as similar as possible.

26. Discriminant analysis

Discriminant analysis is a statistical technique used to determine the relationship between a categorical dependent variable and a set of continuous or categorical independent variables. It helps in identifying the characteristics of the independent variables that differentiate the groups in the dependent variable.

27. multidimensional scaling

Multidimensional scaling (MDS) is a technique used to visualize the similarity or dissimilarity between objects in a dataset. It aims to project high-dimensional data onto a lower-dimensional space while preserving the pairwise distances between the objects.

One common application of MDS is to analyze customer preferences based on product ratings or survey responses. In this case, the objects are the customers, and the dimensions represent the different attributes of the products or survey questions. MDS can help identify groups of customers with similar preferences and uncover underlying patterns in the data.

Most important use case

Visualizing high-dimensional data in a lower-dimensional space for easier interpretation and analysis

28. Correspondence analysis

Correspondence analysis (CA) is a statistical technique used to explore the relationships between categorical variables. It is similar to principal component analysis (PCA), but instead of working with quantitative variables, it works with categorical variables.

In correspondence analysis, a contingency table is created to display the frequency of co-occurrences between the categorical variables. The technique then produces a graphical representation of the relationships between the variables in the form of a biplot.

A biplot is a scatterplot that shows both the observations and the variables, where the distance between the points reflects the strength of the relationship between the variables. The biplot

also shows the contribution of each variable to the overall variability in the data.

Correspondence analysis can be used in a variety of real-world applications, such as market research to analyze the relationships between different products and customer segments, or in linguistics to analyze the relationships between words and documents.

29. Survival analysis

Survival analysis is a statistical method used to analyze the time it takes for an event of interest to occur

Survival analysis is particularly useful when studying events that are rare, such as death or disease occurrence, or when the event of interest is not observed in all subjects.

In survival analysis, we typically use a survival function to model the time until the event of interest occurs. The survival function represents the probability that an individual survives past a given time. We also use hazard functions, which represent the probability of an event occurring at a given time, to model the risk of the event of interest.

30. Bayesian statistics

Bayesian statistics is a branch of statistics that deals with updating probabilities based on new evidence or data. It involves formulating probability statements about unknown parameters or quantities of interest, given the observed data and any prior knowledge or beliefs about those parameters.

In Bayesian statistics, probabilities are interpreted as degrees of belief or uncertainty rather than frequencies of events. Bayes' theorem is used to update the probability of a hypothesis or parameter based on the observed data and any prior knowledge or belief. The resulting probability distribution is called the posterior distribution, which summarizes the updated information about the parameter of interest.

31. ML for descriptive statistics

Machine learning is a branch of artificial intelligence that deals with the development of algorithms that can learn from data and make predictions or decisions based on that data. One important aspect of machine learning is the use of statistical models, which can be used to describe and analyze data. Descriptive statistics is a branch of statistics that deals with summarizing and describing data, and machine learning techniques can be used to build descriptive statistical models.

Linear regression

Linear regression is a popular machine learning algorithm used in descriptive statistics. It is a method for modeling 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 explains the relationship between the variables. The equation for a simple linear regression model with one independent variable is:

$y = b0 + b1x + bias$

where y is the dependent variable, x is the independent variable, b0 and b1 are the intercept and slope coefficients, and bias is the error term. The goal of linear regression is to estimate the values of the coefficients that minimize the sum of squared errors between the predicted values and the actual values.

Multiple Linear Regression

Multiple regression is an extension of linear regression that allows for modeling the relationship between a dependent variable and multiple independent variables. The equation for a multiple linear regression model with n independent variables is:

$y = b0 + b1x1 + b2x2 + ... + bnxn + bias$

where y is the dependent variable, x1, x2, ..., xn are the independent variables, b0, b1, b2, ..., bn are

the intercept and slope coefficients, and bias is the error term. The goal of multiple regression is to estimate the values of the coefficients that minimize the sum of squared errors between the predicted values and the actual values.

Decision Trees and Random Forests

Decision Trees and Random Forests are two popular machine learning algorithms used for both classification and regression tasks. They are based on the concept of decision trees, which can be thought of as a flowchart-like structure where each node represents a feature and each branch represents a possible value or outcome for that feature. At the end of each branch is a leaf node, which represents the final classification or regression result.

Decision Trees

Decision Trees use a top-down, recursive approach to divide a dataset into smaller and smaller subsets. The algorithm starts by selecting the best feature to split the dataset based on a particular criterion, such as entropy or Gini impurity. The dataset is then split into two or more subsets based on the values of that feature. This process is repeated recursively for each subset until a stopping criterion is met, such as reaching a maximum depth or a minimum number of samples per leaf node.

The final result is a tree-like structure where each internal node represents a feature and each leaf node represents a class label or regression value. To make a prediction for a new sample, the algorithm follows the path from the root to a leaf node based on the values of the features in the sample.

Mathematically, the decision tree algorithm can be represented as follows:

Given a dataset D, the goal is to partition it into subsets based on the values of the features x1, x2, ..., xn such that the resulting subsets are as homogeneous as possible with respect to the target variable y.

Start with the entire dataset D.

Select the best feature xk to split the dataset based on a criterion such as entropy or Gini impurity.

Partition the dataset into subsets D1, D2, ..., Dm based on the values of xk.

For each subset Di, repeat the above steps recursively until a stopping criterion is met.

Assign the class label or regression value of the majority of samples in each leaf node.

Random Forests

Random Forests are an extension of Decision Trees that use an ensemble of multiple trees to improve performance and reduce overfitting. The algorithm works by building a set of decision trees on bootstrapped samples of the dataset, where each tree is trained on a random subset of features. The final prediction is made by averaging the predictions of all the trees.

Mathematically, the Random Forest algorithm can be represented as follows:

Given a dataset D, the goal is to build an ensemble of decision trees T1, T2, ..., Tn such that each tree Ti has high accuracy and low correlation with the other trees.

For each tree Ti in the ensemble:

a. Sample a bootstrap subset Di from the dataset D.

b. Select a random subset of features Fk from the total set of features.

c. Build a decision tree Ti on the subset Di using the features Fk.

Make a prediction for a new sample by averaging the predictions of all the trees in the ensemble.

The mathematics behind decision trees and random forests involve concepts such as entropy, information gain, and Gini impurity for determining the splitting criteria at each node. Random forests combine multiple decision trees to reduce overfitting and improve accuracy through a process called bagging.

32. Data visualisation using advanced techniques

Data visualization is the graphical representation of data and information. It is a powerful tool for understanding and communicating complex data sets. Advanced techniques in data visualization can help to convey insights and patterns in data that may not be apparent through traditional methods.

Interactive visualizations allow users to interact with the data, enabling them to explore and analyze it in real-time. Network graphs provide a way to visualize relationships between nodes or entities in a network, while 3D plots allow for a more immersive experience of the data.

33. T-distribution

The t-distribution, also known as the Student's t-distribution, is a probability distribution used in hypothesis testing when the sample size is small (less than 30) or the population standard deviation is unknown. It is similar to the standard normal distribution, but with slightly heavier tails, allowing for more variability in the sample data.

The t-distribution is commonly used in the following scenarios:

Testing hypotheses about a population mean or difference between two population means when the population standard deviation is unknown.

Constructing confidence intervals for a population mean when the population standard deviation is unknown.

Comparing the means of two small samples.

Regression analysis when the residuals are not normally distributed.

34. Degrees of freedom

Degrees of freedom (df) is a concept in inferential statistics that refers to the number of values in a calculation that are free to vary. In general, degrees of freedom is the number of

observations in a sample that are independent and not fixed by any constraints.

In the context of hypothesis testing, degrees of freedom is used to determine the critical value of a test statistic and to calculate the p-value of a test. The number of degrees of freedom is typically calculated as the sample size minus one. The degrees of freedom determine the distribution of the test statistic, which in turn determines the probability of obtaining the observed test statistic by chance alone.

For example, suppose we have a sample of $n = 10$ values and we want to calculate the variance of the sample. The variance is defined as the sum of the squared deviations from the mean divided by the degrees of freedom. Since we are estimating the variance based on the sample, we need to subtract one from the sample size to get the degrees of freedom. Therefore, the degrees of freedom for this calculation would be $df = n - 1 = 10 - 1 = 9$.

Degrees of freedom is an important concept in inferential statistics that is used in a variety of statistical tests and calculations. It is particularly important in hypothesis testing, where it is used to determine the critical value and p-value of a test statistic. In general, the more degrees of freedom a calculation has, the more reliable the estimate is likely to be.

35. Paired sample tests

Paired sample tests are a type of statistical test used to compare the means or differences between two related or matched samples. The samples are considered paired because they come from the same individuals or subjects, and the pairing is based on some characteristic or factor that is common to both samples.

For example, suppose we want to compare the effectiveness of two different weight loss programs on the same group of individuals. We could use a paired sample test to compare the weight loss results of each individual after completing each program, with the pairing based on each individual's weight loss from the first program.

Paired sample tests are useful in a variety of applications, such as in clinical trials to compare

the effectiveness of different treatments on the same patients, or in before-and-after studies to evaluate the impact of a program or intervention on a particular outcome.

36. Null hypothesis

Null hypothesis is a statement that assumes that there is no significant difference between two groups or variables being studied. It is denoted by H0 and is usually used as a starting point for statistical analysis to test the significance of the research hypothesis.

For example, suppose we want to test whether there is a significant difference in the mean weight of apples from two different gardens. The null hypothesis, in this case, would be that there is no significant difference in the mean weight of apples from the two gardens.

If the p-value is greater than significance level(0.05), We accept the null hypothesis that is there is no significant difference between two groups of apples.

37. Alternate hypothesis

The alternative hypothesis (Ha) is a statement that asserts that there is a difference or relationship between two variables of interest. It is the opposite of the null hypothesis (H0), which assumes that there is no difference or relationship between the variables.

For example, suppose we are interested in whether there is a difference in mean heights between men and women in a population. The null hypothesis would be that the mean height of men and women is the same, while the alternative hypothesis would be that the mean height of men and women is different.

If p-value is greater than significance level(0.05) we reject the alternate hypothesis.

38. One tailed test

A one-tailed test is a hypothesis test in which the alternative hypothesis is formulated in one direction only. In other words, the test is designed to determine whether a parameter is greater than or less than a specific value. One-tailed tests are often used when a researcher has a specific prediction or directional hypothesis about the relationship between two variables.

For example, suppose a researcher wants to determine if a new teaching method is more effective than the traditional method. They might set up a one-tailed test with the null hypothesis that there is no difference in the mean test scores between the two methods and the alternative hypothesis that the new method results in higher mean test scores. In this case, the researcher is only interested in testing the alternative hypothesis in one direction, and a one-tailed test is appropriate.

39. Two tailed tests

A two-tailed test is a statistical test in which the null hypothesis is tested against an alternative hypothesis that can differ from the null hypothesis in either direction.

A two-tailed test is used when there is no prior expectation about the direction of the difference or relationship between two variables.

For example, a researcher may want to test whether a new drug has an effect on reducing blood pressure. The null hypothesis would be that the drug has no effect on blood pressure, and the alternative hypothesis would be that the drug has an effect on blood pressure, either increasing or decreasing it. In this case, a two-tailed test is appropriate because the direction of the effect is not known in advance.

40. Power analysis

Power analysis is a statistical technique used to determine the probability of correctly rejecting a null hypothesis when it is in fact false. In other words, it determines the probability of detecting an effect or relationship between variables if it actually exists.

Suppose we want to test the hypothesis that the mean weight of a population of chickens is 2.5 kg, with a significance level of 0.05. We plan to take a random sample of 50 chickens from the population and test their weights. We want to determine the statistical power of this test, which is the probability of correctly rejecting the null hypothesis if the true mean weight is actually different from 2.5 kg.

Use cases of power analysis include:

Determining the sample size needed to achieve a desired level of statistical power Evaluating the sensitivity of a statistical test to detect an effect or relationship between variables Comparing the statistical power of different study designs or analysis methods Optimizing experimental design to maximize the statistical power of a study.

41. Critical value

The critical value is a value of a test statistic that is used to determine whether to reject or fail to reject the null hypothesis in a hypothesis test. It is typically based on the significance level of the test and the degrees of freedom.

42. Randomization

Randomization is the process of randomly assigning subjects or treatments to groups or conditions in a study, to control for extraneous variables and ensure that the groups are comparable. Randomization is a key principle in experimental design and is used to minimize bias and increase the generalizability of study results.

43. Effect size

Effect size is a measure of the magnitude of the difference between two groups or populations. It is an important concept in inferential statistics as it helps to understand the practical significance of statistical results.

There are several ways to calculate effect size for different types of data. One commonly used

measure is Cohen's d, which is defined as the difference between the means of two groups divided by the pooled standard deviation.

44. Point estimation

Point estimation is a method of estimating a population parameter (such as the mean, standard deviation, or proportion) based on a single value, called a point estimate, obtained from a sample. The point estimate is calculated from the sample data and is used as an approximation of the unknown population parameter.

Example: Let's say we want to estimate the average height of students in a college. We can collect a random sample of 50 students from the college and measure their heights. The sample mean height is calculated as X cm. Based on this sample mean, we can make a point estimate of the population mean height.

Use cases of point estimation:

Point estimation is widely used in statistical inference, where we want to estimate a population parameter based on a sample.

It can be used in quality control to estimate the mean or standard deviation of a production process based on a sample of products.