|  |  |
| --- | --- |
| **UNIT – 2 Statistics**  **Type of Data 1) Raw or Primary data**: when data collected having lot of unnecessary, irrelevant & un wanted information **2)** **Treated or Secondary data:** when we treat & remove this unnecessary, irrelevant & un wanted information  **3) Cooked data:** when data collected not genuinely and is false and fictitious **4) Ungrouped data:** when data presented or observed individually.  **5) Grouped data:** when we grouped the identical data by frequency.  **Variable**  A variable is something that can be changed, such as a characteristic or value. For example age, height, weight, blood pressure etc  **Types of Variables**  **Independent variable:** is typically the variable representing the value being manipulated or changed. • The independent variable is the cause. Its value is independent of other variables in study.  **Dependent variable:** is the observed result of the independent variable being manipulated. • The dependent variable is the effect. Its value depends on changes in the independent variable  **Confounding variables:** Are those that affect other variables in a way that produces spurious or distorted associations between two variables. | **Types of measurement**  **1) Discrete:** Quantitative data are called discrete if the sample space contains a finite or countably infinite number of values.  **2) Continuous:** Quantitative data are called continuous if the sample space contains an interval or continuous span of real numbers.  **3) Nominal:** Categorical variables. Numbers that are simply used as identifiers or names represent a nominal scale of measurement such as female vs. male.  **4) Ordinal:** An ordinal scale of measurement represents an ordered series of relationships or rank order. Likert-type scales (such as "On a scale of 1 to 10, with one being no pain and ten being high pain, how much pain are you in today?") represent ordinal data  **5) Qualitative vs. Quantitative variables**  –Qualitative variables: values are texts (e.g., Female, male), we also call them string variables.  –Quantitative variables: are numeric variables.  **Population:** any group of interest or any group that researchers want to learn more about. –Population parameters (unknown to us): characteristics of population  **Sample:** a group of individuals or data are drawn from population of interest. –Sample statistics: characteristics of sample |
| **Descriptive & Inferential Statistics**   |  |  |  | | --- | --- | --- | | Aspect | Descriptive Statistics | Inferential Statistics | | Purpose | Summarize and describe data | Draw conclusions or predictions | | Data Sample | Analyzes the entire dataset | Analyzes a sample of the data | | Examples | Mean, Median, Range, Variance | Hypothesis testing, Regression | | Scope | Focuses on data characteristics | Makes inferences about populations | | Goal | Provides insights and simplifies data | Generalizes findings to a larger population | | Assumptions | No assumptions about populations | Requires assumptions about populations | | Common Use Cases | Data visualization, data exploration | Scientific research, hypothesis testing | | **Key Aspects of Descriptive Statistics:**  **• Measures of Central Tendency:** Descriptive statistics include calculating the mean, median, and mode, which offer insights into the center of the data distribution.  **• Measures of Dispersion:** Variance, standard deviation, and range help us understand the spread or variability of the data.  **• Visualizations:** Creating graphs, histograms, bar charts, and pie charts visually represent the data’s distribution and  characteristics.  **Key Aspects of Inferential Statistics:**  **• Sampling Techniques:** Relies on carefully selecting representative samples from a population to make valid inferences.  **• Hypothesis Testing:** This process involves setting up hypotheses about population characteristics and using sample data to determine if these hypotheses are statistically significant.  **• Confidence Intervals:** These provide a range of values within which we’re confident a population parameter lies based on sample data.  **• Regression Analysis:** Inferential statistics also encompass  techniques like regression analysis to model relationships  between variables and predict outcomes. |

|  |  |
| --- | --- |
| **Measures of Central Tendency**  **1) Mean**  The mean, also known as the average, is a measure of central tendency in a dataset. It is calculated by summing up all the values in the dataset and then dividing by the total number of values. The formula for the mean  **Mean = sum of all observations / number of all observations**    **2) Median:**  The median is the middle value in a dataset when the values are arranged in ascending or descending order. If there is an odd number of values, the median is the middle value itself. If there is an even number of values, the median is the average of the two middle values. The median is useful because it is not influenced by extreme values (outliers) in the dataset.  **3) Mode** The mode is the value or values that appear most frequently in a dataset. A dataset can have one mode (unimodal), two modes (bimodal), or more than two modes (multimodal). If no value repeats, the dataset is said to have no mode. | **Dispersion (Variability):** a measure of the spread of scores in a distribution  Variability commonly measured with the following:  **– Range** Diff. between highest and lowest values  **– Inter Quartile Range (IQR):** Range of the middle half of a distribution  **– Standard deviation:** Average distance from the mean  **– Variance:** Average of squared distances from the mean  **4) Range** Range in statistics refers to the difference between the highest and lowest values in a dataset. It provides a measure of the spread or variability of the data. It is informative for data without outliers To calculate the range:  **Range = Highest Value − Lowest Value**  **Quartiles:** Quartiles are values that divide a dataset into four equal parts, each representing 25% of the data.  **There are three quartiles:**  **First Quartile (Q1):** The value that separates the lowest 25% of the data from the rest. It is also the 25th percentile.  **Second Quartile (Q2):** The median, which divides the data into two equal halves. It is also the 50th percentile.  **Third Quartile (Q3):** The value that separates the lowest 75% of the data from the highest 25%. It is also the 75th percentile.  **5) Inter Quartile Range (IQR)**  **Interquartile range (IQR) = Value of third**  **quartile (Q3) – Value of first quartile (Q1)** |
| **Outliers**  Outliers are data points that significantly differ from the rest of the dataset, often indicating potential errors, anomalies, or rare occurrences. They can skew statistical analyses and affect the accuracy of models if not properly addressed.  **When there are no outliers in a sample**, ̵ the mean and standard deviation are used to summarize a typical value and the variability in the sample, respectively.  **When there are outliers in a sample**, ̵ the median and interquartile range are used to summarize a typical value and the variability in the sample, respectively.  **Tukey's fences** offer a systematic approach to identifying outliers based on the distribution of the data, helping to ensure robust and accurate statistical analyses  Tukey's fences are based on the Interquartile Range (IQR) and are calculated as follows:  **Calculate the IQR**: IQR = Q3 - Q1 (where Q3 is the third quartile and Q1 is the first quartile).  **Lower Fence**: Q1 - 1.5 \* IQR  Any data point below this lower fence is considered an outlier.  **Upper Fence**: Q3 + 1.5 \* IQR  Any data point above this upper fence is considered an outlier. | **BOX – PLOT – refer (50 -61)**  A box plot, also known as a box-and-whisker plot, is a graphical representation of the distribution of a dataset. It displays key statistical measures such as the median, quartiles (Q1 and Q3), and potential outliers in a compact and informative way. Here's how a box plot is constructed and calculated:   1. **Median (Q2)**: The middle value of the dataset, represented by a line inside the box. 2. **First Quartile (Q1)**: The median of the lower half of the dataset, forming the lower boundary of the box. 3. **Third Quartile (Q3)**: The median of the upper half of the dataset, forming the upper boundary of the box. 4. **Interquartile Range (IQR)**: The range between Q1 and Q3, represented by the height of the box. 5. **Whiskers**: Lines extending from the top and bottom of the box to indicate the range of non-outlier data points.    * **Lower Whisker:** Typically extends to the lowest data point within 1.5 times the IQR below Q1, or to the minimum data point if no outliers are present.    * **Upper Whisker:** Extends to the highest data point within 1.5 times the IQR above Q3, or to the maximum data point if no outliers are present.   **Outliers**: Individual data points beyond the whiskers are considered potential outliers and are often marked separately on the plot. |

|  |  |
| --- | --- |
| **6) Percentiles:**  Values below which a given percentage of observations fall.  Percentiles are statistical measures that divide a dataset into hundred equal parts, each representing 1% of the data. They are useful for understanding the distribution of values within a dataset and identifying specific points relative to the entire dataset  **Formula:** p th percentile:  - p percent of observations below it  - (100 - p)% above it.  **Example:** Like 95% of CAT percentile means 95% are below it and 5% are above it | **7) Standard Deviation**  A measure of the amount of variation or dispersion of a set of values. Standard deviation is a measure of the dispersion or variability of a set of data points around their mean (average) value. It indicates how spread out the values in a dataset are from the mean, providing insight into the degree of consistency or deviation within the dataset |
| **8) Variance**  A measure of how spread out the values in a dataset are.  Variance is a statistical measure that quantifies the dispersion or spread of a set of data points around their mean (average) value. It is the average of the squared differences between each data point and the mean.  **Standard normal distribution or Z distribution** The standard normal distribution, also known as the Z distribution, is a specific type of normal distribution with a mean (average) of 0 and a standard deviation of 1. It is a fundamental concept in statistics and probability theory and serves as a reference for many statistical analyses.  **Z Score**  **z = (x – μ) / σ**  **– x observation**  **– μ mean**  **– σ standard deviation** | **Characteristics of normal distribution**  ▪ The normal distribution is mathematically defined  ▪ The normal distribution is theoretical.  ▪ The mean, median, and mode are all the same values at  the center of the distribution.  ▪ The normal distribution is symmetrical.  ▪ The form of a normal distribution is determined by its  mean and standard deviation.  ▪ Standard deviation can be any positive value.  ▪ The total area under the curve is equal to 1.  ▪ The tails of normal distribution are always approaching to  x-axis, but never touch it.  **Histograms**  Histograms are graphical representations of the distribution of a dataset. They display the frequencies or counts of data points falling within certain intervals or bins. Histograms are widely used in statistics and data analysis to visualize the distribution of continuous or discrete numerical data.  **Bins or Intervals**: The range of values is divided into intervals or bins along the horizontal axis (x-axis).  **Frequency or Count**: The vertical axis (y-axis) represents the frequency or count of data points falling within each bin.  **Bars**: Each bin is represented by a bar whose height corresponds to the frequency or count of data points in that  **No Gaps**: There are no gaps between the bars in a histogram, as it represents continuous data.  **Area Under the Curve**: The area of each bar is proportional to the frequency or count of data points in the corresponding interval. |

|  |  |
| --- | --- |
| **Normalization**:   * **Purpose**: The goal of normalization is to rescale the data so that it falls within a specific range, typically between 0 and 1. This process is helpful when the features (variables) in the dataset have different scales or units. * **Method**: The most common normalization technique is Min-Max normalization, which transforms each data point 𝑥*x* to a new value 𝑥′*x*′ using the formula * Effect: Normalization preserves the relative relationships and proportions between data points, but it may be sensitive to outliers, especially when using the Min-Max technique. | **Standardization**:   * **Purpose**: Standardization aims to center the data around a mean of 0 and a standard deviation of 1. It is particularly useful when the features in the dataset have different scales and follow a normal distribution or when performing certain statistical analyses like regression. * **Method**: The standardization process involves subtracting the mean (𝜇*μ*) from each data point and then dividing by the standard deviation (𝜎*σ*). The formula * Effect: Standardization results in a distribution with a mean of 0 and a standard deviation of 1, making it easier to compare and interpret the relative importance of different features in the dataset. It is less affected by outliers compared to normalization. |
| **9) Skewness**  A measure of the asymmetry of the distribution of values.  • Skewness is a number that indicates to what extent a  variable is asymmetrically distributed.  • It is the degree of distortion from the symmetrical bell  curve or the normal distribution.  • A symmetrical distribution will have a skewness of 0.  **Symmetric Distribution (No Skew):**  In a symmetric distribution, the data is evenly distributed around the mean, and the skewness value is close to 0.  The tails on the left and right sides of the distribution are balanced.  **Negative Skew (Left Skew):** Mode >= Median >=Mean  In a negatively skewed distribution, the tail on the left side of the distribution is longer or stretched out, indicating more low values. The mean is less than the median, and the skewness value is negative.  Example: Income distribution in a country where a few people have very high incomes (long left tail).  **Positive Skew (Right Skew**): Mean >= Median >= Mode  In a positively skewed distribution, the tail on the right side of the distribution is longer or stretched out, indicating more high values. The mean is greater than the median, and the skewness value is positive.  Example: Test scores distribution where most students score well but a few score very low (long right tail). | **10) Kurtosis**  Kurtosis is a statistical measure that quantifies the degree of peakedness or flatness of the distribution of data points in a dataset. It assesses whether the data is more or less peaked than a normal distribution and can provide insights into the presence of outliers or extreme values in the dataset.  • **High kurtosis** in a data set is an indicator that data has  heavy tails or outliers. investigate why do we have so many outliers.  **• Low kurtosis** in a data set is an indicator that data has  light tails or lack of outliers need to investigate and trim the dataset of unwanted results  **• Mesokurtic:** This distribution has kurtosis statistic similar to that of the normal distribution (Kurtosis = 0).  **• Leptokurtic (Kurtosis > 3):** Distribution is longer, tails are fatter. Peak is higher and sharper than Mesokurtic, which means that data are heavy-tailed or more outliers.  **• Platykurtic: (Kurtosis < 3):** Distribution is shorter, tails are thinner than the normal distribution. The peak is lower and broader than Mesokurtic, which means that data are lighttailed or lack of outliers. |

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **11) Correlation**  Correlation is a statistical measure that quantifies the relationship or association between two variables in a dataset. It indicates the extent to which changes in one variable are associated with changes in another variable.  Correlation describes strength of association between  two variables  • Falls between -1 and +1, with sign indicating direction  of association (formula & other details later )  • The larger the correlation in absolute value, the stronger  the association (in terms of a straight line trend)  **Formula:**  **12) Covariance** | **13) Regression**  Regression analysis is a statistical method used to model the relationship between a dependent variable (response variable) and one or more independent variables (predictor variables).  **Types of Regression:**  **Simple Linear Regression**: Involves one independent variable to predict the dependent variable.  **Multiple Linear Regression:** Includes multiple independent variables to predict the dependent variable.  **Polynomial Regression:** Fits a curve to the data by using polynomial functions.  **Logistic Regression:** Used for binary classification problems where the dependent variable is categorical.  **Covariance and Correlation**   |  |  |  | | --- | --- | --- | | Aspect | Covariance | Correlation | | Definition | Measures directional association. | Standardized linear relationship. | | Range of Values | -∞ to +∞ (unbounded) | -1 to +1 (bounded) | | Standardization | Not standardized | Standardized | | Strength Indicator | Directional association only | Indicates strength of relationship | | Application | Used in analysis, scale dependent | Widely used, scale independent | |
| **Quartile Vs Quantile**  **Quartiles**  **̵First quartile:** Also known as Q1 (the number halfway between the lowest number and the middle number).  **̵Second quartile**: Also known as Q2 or the median (the middle number halfway between the lowest number and the highest number).  **̵Third quartile:** Also known as Q3, or the upper quartile (the number halfway between the middle number and the highest number).  **Quantile**  ̵Are values that split sorted data or a probability distribution into equal parts (In general, q-quantile divides sorted data into q parts).  ̵A quartile is a type of quantile.  ▪ Quartiles (4-quantiles): Three quartiles split the data into four parts.  ▪ Deciles (10-quantiles): Nine deciles split the data into 10  ▪ Percentiles (100-quantiles): 99 percentiles split the data into 100 parts.  ̵There is always one fewer quantile than there are parts created by the quantiles  **Quantile-Quantile Plots (QQ Plots)**  The quantile-quantile (q-q plot) plot is a **graphical method for determining if a dataset follows a certain probability distribution** or whether two samples of data came from the same population or not. Q-Q plots are particularly useful for assessing whether a dataset is normally distributed or if it follows some other known distribution. | **Quantile-Quantile Plots (QQ Plots)**  **1.Collect the Data:** Gather the dataset for which you want to create the Q-Q plot. Ensure that the data  are numerical and represent a random sample from the population of interest.  **2.Sort the Data:** Arrange the data in either ascending or descending order. This step is essential for  computing quantiles accurately.  **3.Choose a Theoretical Distribution:** Determine the theoretical distribution against which you want to  compare your dataset. Common choices include the normal distribution, exponential distribution, or any  other distribution that fits your data well.  **4.Calculate Theoretical Quantiles:** Compute the quantiles for the chosen theoretical distribution. For  example, if you’re comparing against a normal distribution, you would use the inverse cumulative  distribution function (CDF) of the normal distribution to find the expected quantiles.  **5.Plotting:**  Plot the sorted dataset values on the x-axis.  Plot the corresponding theoretical quantiles on the y-axis.  Each data point (x, y) represents a pair of observed and expected values.  Connect the data points to visually inspect the relationship between the dataset and the  theoretical distribution |

|  |  |
| --- | --- |
| **Comparing the three visual techniques**  **Histograms**  • Advantages: With properly-sized bins, histograms can summarize any shape of the data (modes, skew, quantiles,  outliers)  • Disadvantages: Difficult to compare side-byside (takes up too much space in a plot) Depending on the size of the  bins, interpretation may be different  **Boxplots**  Advantages:  Can identify whether the data came from a certain distribution.  Don’t have to tweak with “graphical” parameters (i.e. bin size in histograms)  Summarize quantiles  Disadvantages:  Difficult to compare side by-side  Difficult to distinguish skews, modes, and outliers  **QQ Plots**  Advantages:  – Don’t have to tweak with “graphical” parameters (i.e. bin size in histograms)  – Summarize skew, quantiles, and outliers  – Can compare several measurements side-by-side Disadvantages:  – Cannot distinguish modes | **Binomial distribution**  • Binomial distribution is a type of discrete probability distribution representing probabilities of different values of the binomial random variable (X) in repeated independent N trials in an experiment.  • Thus, in an experiment comprising of tossing a coin 10 times (n), the binomial random variable (number of heads represented as successes) could take the value of 0-10.  • The binomial probability distribution is the probability distribution representing the probabilities of a random variable taking the value of 0-10  **• The necessary conditions and criteria to use binomial**  **distributions:**  **• Rule 1:** Situation where there are only two possible mutually exclusive outcomes (for example, yes/no survey questions).  **• Rule2:** A fixed number of repeated experiments and trials  are conducted (the process must have a clearly defined  number of trials).  **• Rule 3:** All trials are identical and independent (identical  means every trial must be performed the same way as the  others; independent means that the result of one trial does  not affect the results of the other subsequent trials).  **• Rule: 4:** The probability of success is the same in every one of the trials. |
| **Poisson distribution**   * The Poisson distribution is a type of discrete probability distribution used to model the number of events occurring within a fixed interval of time or space, given the average rate of occurrence (𝜆*λ*). * For example, in a Poisson process where events occur randomly but at a constant average rate of 5 events per hour (𝜆=5*λ*=5), the Poisson random variable (X) could take values 0, 1, 2, 3, and so on, representing the number of events in that interval. * The Poisson probability distribution represents the probabilities of X taking different values based on the Poisson parameter (𝜆*λ*). * **Conditions for Poisson Distribution**:   + Rule 1: Events occur randomly and independently of each other.   + Rule 2: The average rate of occurrence (𝜆*λ*) is constant throughout the interval.   Rule 3: The probability of more than one event occurring in an infinitesimally small interval is negligible  **Confidence Interval**  Confidence, in statistics, is another way to describe probability. For example, if you construct a confidence interval with a 95% confidence level, you are confident that 95 out of 100 times the estimate will fall between the upper and lower values specified by the confidence interval. • Your desired confidence level is usually one minus the alpha ( a ) value you used in your statistical test: • Confidence level = 1 − a | **Central Limit Theorem**  • The Central Limit Theorem states that the sampling distribution of the sampling means approaches a normal distribution as the sample size gets larger, no matter what the shape of the data distribution. • An essential component of the Central Limit Theorem is the average of sample means will be the population mean. • Similarly, if you find the average of all of the standard deviations in your sample, you will find the actual standard deviation for your population. • Mean of sample is same as the mean of the population. • The standard deviation of the sample is equal to the standard deviation of the population divided by the square root of the sample size. • Central limit theorem is applicable for sufficiently large sample sizes (n ≥ 30). The formula for central limit theorem can be stated as follows: |