STATISTICS WORKSHEET-1

Q1 to Q9 have only one correct answer. Choose the correct option to answer your question.

1. Bernoulli random variables take (only) the values 1 and 0.
   a) True b) False
2. Which of the following theorem states that the distribution of averages of iid variables, properly normalized, becomes that of a standard normal as the sample size increases?
   a) Central Limit Theorem
   b) Central Mean Theorem
   c) Centroid Limit Theorem
   d) All of the mentioned
3. Which of the following is incorrect with respect to use of Poisson distribution?
   a) Modeling event/time data
   b) Modeling bounded count data
   c) Modeling contingency tables
   d) All of the mentioned
4. Point out the correct statement.
   a) The exponent of a normally distributed random variables follows what is called the log- normal distribution
   b) Sums of normally distributed random variables are again normally distributed even if the variables are dependent
   c) The square of a standard normal random variable follows what is called chi-squared distribution
   d) All of the mentioned
5. _____ random variables are used to model rates.
   a) Empirical
   b) Binomial
   c) Poisson
   d) All of the mentioned
6. 10. Usually replacing the standard error by its estimated value does change the CLT.
   a) True
   b) False
7. 1. Which of the following testing is concerned with making decisions using data?
   a) Probability
   b) Hypothesis
   c) Causal
   d) None of the mentioned
8. 4. Normalized data are centered at_____and have units equal to standard deviations of the original data.
   a) 0
   b) 5
   c) 1
   d) 10
9. Which of the following statement is incorrect with respect to outliers?
   a) Outliers can have varying degrees of influence
   b) Outliers can be the result of spurious or real processes
   c) Outliers cannot conform to the regression relationship
   d) None of the mentioned

Q10and Q15 are subjective answer type questions, Answer them in your own words briefly.

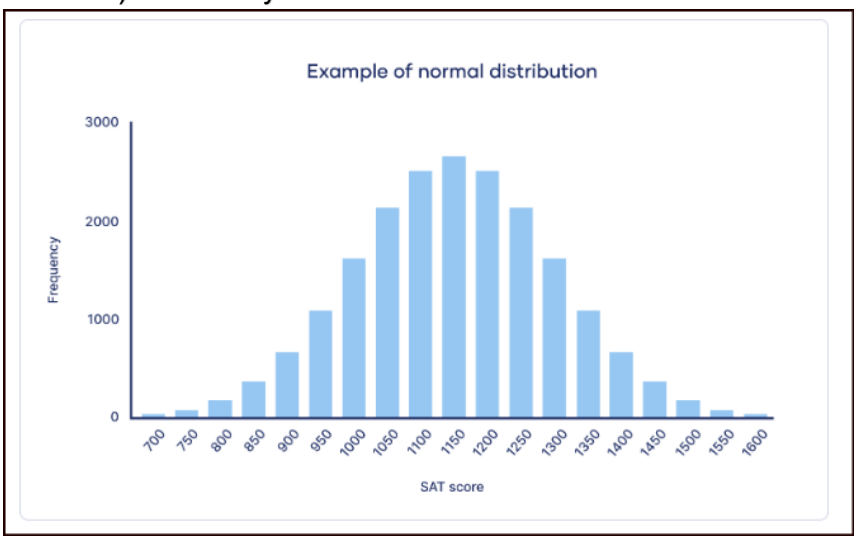
10.What do you understand by the term Normal Distribution?

A normal distribution is a type of continuous probability distribution in which most data points cluster towards the middle of the range, while the rest Tapper off symmetrically towards either extreme. The middle of the range is also known as mean of distribution.

Normal distributions are also called Gaussian distributions or bell curves because of their shape.

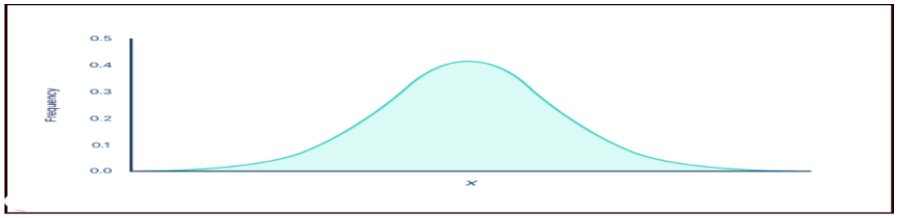
The normal distribution is produced by the normal density function $p(x) = e^{-(x-\mu)2}/2\sigma^2/\sigma\sqrt{2\pi}$.
In this exponential function e is the constant 2.71828…, is the mean, and σ is the standard deviation

In a normal distribution, data are symmetrically distributed with no skew. Most values clusters around a central region, the values tampering off as they go further allow from the centre. The measures of central tendency (mean, mode and medium) are exactly the same in normal distribution



Normal distributions have key characteristics that are easy to spot in graphs:

- The mean, median and mode are exactly the same.
- The distribution is symmetric about the mean—half the values fall below the mean and half above the mean.
- The distribution can be described by two values: the mean and the standard deviation.



11.How do you handle missing data what imputation techniques do you recommend?

In statistics, missing data can be a common challenge when analysing datasets. Imputation techniques are used to estimate or fill in missing values based on the available data. Here are some commonly used imputation techniques:

1. Mean/median imputation: In this method, missing values are replaced with the mean or median value of the variable. This approach assumes that the missing values are roughly similar to the observed values.
2. Mode imputation: This technique is used for categorical variables. The missing values are replaced with the mode (most frequently occurring value) of the variable.
3. Last observation carried forward (LOCF): This approach is often used in time series or longitudinal data. Missing values are replaced with the most recent observed value for that variable.
4. Multiple imputation: Multiple imputation involves creating several imputed datasets, where missing values are replaced with plausible values based on statistical models. The analyses are then performed on each imputed dataset, and the results are combined to obtain valid inferences.
5. Regression imputation: This method involves using regression models to predict missing values based on other variables. A regression model is built using the observed values, and then missing values are estimated using the model's predictions.
6. K-nearest neighbours (KNN) imputation: KNN imputation is a non-parametric method where missing values are imputed based on the values of k nearest neighbors in the dataset. The distance between observations is calculated, and the missing values are filled in based on the values of the nearest neighbours.
7. Expectation-maximization (EM) algorithm: The EM algorithm is an iterative procedure that estimates missing values by maximizing the likelihood of the observed data. It is particularly useful for data with a complex pattern of missingness.

The choice of imputation technique depends on the characteristics of the data, the amount of missingness, and the specific analysis being performed. Each technique has its own assumptions and limitations, so it's important to carefully consider the context and implications of imputing missing values.

12. What is A/B testing?

A/B testing, also known as split testing, is a statistical method used in marketing and product development to compare two or more versions of a webpage, advertisement, or other elements to determine which one performs better. It is a controlled experiment where two or more variations, often labeled as A and B, are presented to different groups of users or customers. The goal is to measure the impact of changes and determine which variation leads to a desired outcome.

The process of A/B testing typically involves the following steps:

1. Hypothesis: Formulate a hypothesis about the changes or variations that are expected to have an impact on the desired outcome. For example, changing the color of a call-to-action button might increase click-through rates.
2. Control and Variation: Create two or more versions (A and B) of the element being tested. One version, called the control, remains unchanged, while the other version, called the variation, incorporates the proposed changes.
3. Randomization: Randomly assign users or customers into groups, ensuring that each group represents a similar demographic or characteristic. The control and variation versions are then presented to the different groups simultaneously.
4. Data Collection: Track and collect data on user interactions, behaviors, or conversions associated with each variation. This may include metrics such as click-through rates, conversion rates, engagement time, or revenue generated.
5. Statistical Analysis: Analyze the collected data using statistical methods to determine if there are statistically significant differences between the variations. Statistical significance helps establish if the observed differences are not due to chance.
6. Conclusion: Based on the analysis, draw conclusions about the effectiveness of each variation in achieving the desired outcome. Determine which version performs better and decide whether to implement the changes permanently.

A/B testing allows businesses to make data-driven decisions and optimize their strategies based on empirical evidence rather than assumptions. It is widely used in digital marketing, website optimization, user interface design, and other areas where changes can be tested and measured. By continuously testing and refining different elements, organizations can improve their conversion rates, user experiences, and overall performance.

13. Is mean imputation of missing data acceptable practice?

Mean imputation of missing data is a commonly used technique in data analysis when dealing with missing values. However, its acceptability as a practice depends on the context and the nature of the data being analysed. While mean imputation is simple to implement and can preserve the sample size, it has certain limitations and potential drawbacks that should be considered.

One major concern with mean imputation is that it can introduce bias into the data. By replacing missing values with the mean, the imputed values may not accurately reflect the true values that were missing. This can distort the relationships between variables and lead to incorrect conclusions in statistical analyses.

Additionally, mean imputation assumes that the missing data are missing completely at random (MCAR). This means that the probability of data being missing is unrelated to both observed and unobserved data. If the missing data are not MCAR, meaning that the missingness is related to the value of the variable or other factors, mean imputation can introduce further bias and distort the results.

There are alternative methods available for handling missing data, such as multiple imputation, maximum likelihood estimation, or using advanced machine learning techniques. These methods can provide more robust and accurate results by taking into account the uncertainty associated with the missing values.

In summary, while mean imputation is a simple and commonly used approach for handling missing data, it should be applied with caution. It may be acceptable in certain situations, especially when the missingness is minimal and the assumptions of MCAR are reasonable. However, for more accurate and reliable results, it is often recommended to explore alternative approaches that better account for the missing data patterns and potential biases.

14. What is linear regression in statistics?

Linear regression is a statistical technique used to model the relationship between a dependent variable and one or more independent variables. It aims to find the best-fitting linear equation that describes the relationship between these variables.

In linear regression, the dependent variable (also known as the response or outcome variable) is assumed to be a linear combination of the independent variables (also known as predictor or explanatory variables). The goal is to estimate the coefficients of the linear equation that minimize the difference between the observed values of the dependent variable and the predicted values based on the independent variables.

The linear equation takes the form:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \varepsilon$$

where:

- Y represents the dependent variable.
- $X_1, X_2, \ldots, X_p$ represent the independent variables.
- $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$ are the coefficients (also known as regression coefficients or parameters) that quantify the relationship between the independent variables and the dependent variable.
- ε represents the error term, which captures the discrepancy between the observed and predicted values of the dependent variable.

The coefficients ($\beta_0, \beta_1, \beta_2, \ldots, \beta_p$) are estimated using a method called Ordinary Least Squares (OLS), which minimizes the sum of the squared differences between the observed and predicted values. The estimated coefficients provide information about the magnitude and direction of the relationships between the variables.

Linear regression can be used for various purposes, such as predicting future values of the dependent variable, understanding the impact of the independent variables on the dependent variable, and assessing the statistical significance of the relationships. It assumes linearity, independence, homoscedasticity (constant variance of errors), and normally distributed errors in the data.

15. What are the various branches of statistics?

Statistics is a broad field that encompasses several branches or subfields, each focusing on different aspects of data analysis, inference, and modelling. Some of the main branches of statistics include:

1. Descriptive Statistics: This branch involves summarizing and describing data using measures such as mean, median, mode, variance, standard deviation, and graphical representations like histograms and box plots. Descriptive statistics provide an overview of the data's central tendency, dispersion, and distribution.
2. Inferential Statistics: Inferential statistics deals with making inferences and drawing conclusions about populations based on sample data. It involves hypothesis testing, estimation, and determining the confidence intervals to make predictions and generalizations.
3. Probability Theory: Probability theory is the foundation of statistical inference. It deals with quantifying uncertainty and randomness. It includes concepts like probability distributions, random variables, independence, conditional probability, and Bayes' theorem.
4. Biostatistics: Biostatistics applies statistical methods to analyse data in the field of biology, medicine, and public health. It includes designing clinical trials, analysing epidemiological data, and studying the relationship between risk factors and diseases.
5. Econometrics: Econometrics focuses on the application of statistical methods to economic data. It involves analysing economic relationships, estimating economic models, and testing hypotheses in areas such as finance, macroeconomics, and microeconomics.
6. Time Series Analysis: Time series analysis deals with analysing and modelling data that is collected over time. It includes techniques for detecting patterns, trends, seasonality, and forecasting future values in areas such as economics, finance, and signal processing.
7. Multivariate Analysis: Multivariate analysis deals with analysing data that involves multiple variables simultaneously. It includes techniques such as multivariate regression, principal component analysis (PCA), factor analysis, and cluster analysis.
8. Experimental Design: Experimental design focuses on designing and analysing controlled experiments to determine cause-and-effect relationships between variables. It involves techniques for randomization, blocking, factorial designs, and analysis of variance (ANOVA).
9. Spatial Statistics: Spatial statistics deals with analysing data that has a spatial or geographic component. It includes techniques for spatial interpolation, spatial autocorrelation, and spatial regression to understand patterns and relationships in spatial data.
10. Data Mining and Machine Learning: Data mining and machine learning involve developing algorithms and techniques to extract knowledge and make predictions from large and complex datasets. It includes methods such as decision trees, random forests, support vector machines, and neural networks.

These are just a few examples of the branches of statistics, and the field continues to evolve with advancements in technology and new applications emerging across various domains.