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

ANSWER= TRUE

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

ANSWER= a central limit theorem.

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

ANSWER= Modeling bounded count data

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

ANSWER= The square of a standard normal random variable follows what is called chi-squared distribution

5. \_\_\_\_\_\_ random variables are used to model rates.

# a) Empirical b) Binomial

# c) Poisson d) All of the mentioned

ANSWER= poisson

6. 10. Usually replacing the standard error by its estimated value does change the CLT.

# a) True b) False

ANSWER= 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

ANSWER= HYPOTHESIS.

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

ANSWER= 0

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

ANSWER= c) Outliers cannot conform to the regression relationship

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

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

Answer

The term "Normal Distribution," also known as the Gaussian distribution or bell curve, refers to a fundamental and widely used probability distribution in statistics. It is a continuous probability distribution that is characterized by its symmetric, bell-shaped curve. The shape of the curve is determined by two parameters: the mean (μ) and the standard deviation (σ).

In a normal distribution:

The curve is symmetric around its mean, which is also its median and mode. This means that the data clusters around the mean, and the probability of observing values decreases as they deviate from the mean.

The standard deviation controls the spread or dispersion of the data. A smaller standard deviation results in a narrower curve, while a larger standard deviation leads to a wider curve.

The total area under the curve is equal to 1, representing the entire probability space.

The empirical rule, also known as the 68-95-99.7 rule, applies to the normal distribution. It states that approximately 68% of the data falls within one standard deviation of the mean, about 95% falls within two standard deviations, and roughly 99.7% falls within three standard deviations.

Many real-world phenomena can be approximated by a normal distribution, thanks to the Central Limit Theorem. This theorem states that the sum (or average) of a large number of Independent and identically distributed random variables will have an approximately normal distribution, regardless of the original distribution of the variables, as long as the sample size is sufficiently large.

The normal distribution is a fundamental concept in statistics and has numerous applications in various fields, including natural sciences, social sciences, finance, engineering, and more. It provides a foundation for understanding and analyzing data, estimating probabilities, and making statistical inferences.

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

Answer

Handling missing data is an important and often complex task in data analysis and statistical modeling. Different imputation techniques can be used to fill in missing values based on the characteristics of the data and the goals of the analysis. Here are some common approaches to handling missing data:

1. \*Complete Case Analysis (Listwise Deletion):\* This involves excluding observations with missing data from the analysis. While simple, it can lead to biased results if missingness is related to the outcome or other variables of interest.

2. \*Mean/Median Imputation:\* Replace missing values with the mean or median of the available data for that variable. This is a straightforward approach, but it can distort the distribution and reduce variability.

3. \*Mode Imputation:\* Replace missing categorical data with the mode (most frequent value) of the available data.

4. \*Interpolation/Extrapolation:\* Use nearby observations to estimate missing values based on trends or patterns in the data.

5. \*Regression Imputation:\* Predict missing values using regression models based on other variables. This can be especially useful if the missing data are related to other variables.

6. \*Multiple Imputation:\* Generate multiple plausible values for missing data and average the results. This technique accounts for uncertainty introduced by imputation.

7. \*K-Nearest Neighbors Imputation:\* Replace missing values with the average of the k-nearest neighbors in terms of other variables. Useful when data points with similar attributes are likely to have similar values.

8. \*Time-Series Imputation:\* For time-series data, missing values can be estimated using methods like linear interpolation or seasonally adjusted averages.

9. \*Domain-Specific Imputation:\* Some fields have domain-specific methods to impute missing data. For example, medical data might use clinical knowledge to impute missing values.

The choice of imputation technique depends on factors such as the amount of missing data, the nature of the data (continuous, categorical, etc.), the presence of relationships between variables, and the assumptions you're willing to make. It's important to carefully consider the potential impact of imputation on the results of your analysis and to perform sensitivity analyses to assess the robustness of your conclusions.

Always keep in mind that imputation introduces uncertainty, and the best approach depends on the specific context and goals of your analysis. It's also important to document the imputation process and any assumptions made to ensure the transparency and reproducibility of your results. Always keep in mind that imputation introduces uncertainty, and the best approach depends on the specific context and goals of your analysis. It's also important to document the imputation process and any assumptions made to ensure the transparency and reproducibility of your results.

12. What is A/B testing?

Answer

# A/B testing, also known as split testing, is a controlled experimentation method used to compare two versions of a web page, app, marketing campaign, or other digital assets to determine which one performs better in terms of a desired outcome. It is widely used in marketing, user experience (UX) design, and product development to make data-driven decisions and optimize various aspects of a business.

# Here's how A/B testing generally works:

# 1. \*Hypothesis:\* Start with a clear hypothesis about a change you want to make. For example, you might hypothesize that changing the color of a call-to-action button on a website will lead to a higher click-through rate.

# 2. \*Variations:\* Create two versions of the element you want to test – the original version (A) and the modified version (B), which includes the change you're testing.

# 3. \*Randomization:\* Randomly assign visitors or users to either version A or B. This helps ensure that the groups are statistically similar at the outset, reducing the likelihood of bias.

# 4. \*Data Collection:\* Collect data on user interactions or behavior for both versions. This could include metrics like clicks, conversions, sales, time spent, etc.

# 5. \*Statistical Analysis:\* Analyze the data to determine whether the differences in performance between the two versions are statistically significant. Statistical tests help determine whether any observed differences are likely due to the change you made and not just random variation.

# 6. \*Conclusion:\* Based on the analysis, you can conclude whether the change had a significant impact on the desired outcome. If one version clearly outperforms the other, you can adopt the winning version and implement it in your actual system.

# A/B testing allows you to make informed decisions by comparing the performance of different options directly, rather than relying on assumptions or intuition. It helps answer questions like:

# - Which headline generates more clicks?

# - Does changing the layout of a webpage increase engagement?

# - Which email subject line leads to higher open rates?

# - Is a new feature more effective at driving user conversions?

# A/B testing is a powerful tool, but it's important to ensure proper experimental design, sufficient sample sizes, and rigorous statistical analysis to draw reliable conclusions. It's also crucial to consider potential ethical concerns and the impact of testing on user experience.

13. Is mean imputation of missing data acceptable practice?

# Answer

# Mean imputation, where missing values are replaced with the mean of the available data, is a common method for handling missing data due to its simplicity. However, its acceptability as a practice depends on the context and the potential impact on the analysis. Here are some considerations to keep in mind:

# \*Advantages of Mean Imputation:\*

# 1. \*Ease of Implementation:\* Mean imputation is straightforward and easy to implement.

# 2. \*Preservation of Sample Size:\* It allows you to retain the original sample size, which can be important for maintaining statistical power.

# 3. \*Minimal Distortion:\* Mean imputation may work reasonably well when the proportion of missing values is small and missingness is unrelated to the variable being imputed.

# \*Disadvantages and Considerations:\*

# 1. \*Bias Introduction:\* Mean imputation assumes that missing values are missing completely at random (MCAR), meaning that the missingness is unrelated to the variable being imputed or any other variables. If the missingness is related to the variable itself or other variables, mean imputation can introduce bias in the analysis.

# 2. \*Loss of Variability:\* Mean imputation does not account for the variability of the data. It can lead to underestimation of standard errors, potentially affecting the significance of results and confidence intervals.

# 3. \*Distortion of Relationships:\* Mean imputation can artificially strengthen or weaken relationships between variables, especially in regression analysis.

# 4. \*Impact on Distribution:\* Mean imputation can impact the distribution of the imputed variable, making it more similar to the mean and potentially affecting downstream analyses.

# 5. \*Outliers:\* Mean imputation can be sensitive to outliers in the data, as they can disproportionately influence the mean.

# \*When to Use Mean Imputation:\*

# Mean imputation might be considered when:

# - The proportion of missing values is small.

# - Missingness is believed to be MCAR or missingness is not expected to introduce significant bias.

# - The variable with missing data has a relatively small impact on the analysis or is not a key focus.

# \*Alternatives to Mean Imputation:\*

# Consider more advanced imputation methods, such as multiple imputation or regression imputation, when:

# - The proportion of missing values is substantial.

# - Missingness is believed to be related to other variables.

# - You want to account for the variability of the data and preserve relationships.

# In summary, while mean imputation is a simple approach, its appropriateness depends on the specific situation and analysis goals. It's important to carefully consider the potential impact on your results, conduct sensitivity analyses, and document your imputation methods transparently.

14. What is linear regression in statistics?

# Linear regression is a fundamental statistical method used for modeling the relationship between a dependent variable (also known as the response variable) and one or more independent variables (also known as predictor variables or features). The primary goal of linear regression is to find the best-fitting linear relationship between these variables, allowing us to make predictions or understand the association between them.

# In its simplest form, linear regression assumes a linear relationship between the independent variables (X) and the dependent variable (Y), which can be represented by the equation of a straight line:

# Y = β₀ + β₁X + ε

# Where:

# - Y is the dependent variable (response).

# - X is the independent variable (predictor).

# - β₀ is the intercept (the value of Y when X = 0).

# - β₁ is the slope of the line (how much Y changes for a unit change in X).

# - ε represents the error term (the difference between the actual Y and the predicted Y).

# The goal of linear regression is to estimate the values of β₀ and β₁ that minimize the sum of squared differences between the observed Y values and the predicted Y values. This process is often carried out using a method called the Least Squares method, which aims to find the line that minimizes the sum of squared residuals (the differences between observed and predicted values).

# Linear regression can be extended to multiple independent variables, resulting in multiple linear regression. In this case, the equation becomes:

# Y = β₀ + β₁X₁ + β₂X₂ + ... + βₖXₖ + ε

# Where X₁, X₂, ..., Xₖ are the different independent variables, and β₁, β₂, ..., βₖ are their corresponding coefficients.

# Linear regression is widely used in various fields, including economics, social sciences, engineering, and machine learning, to analyze relationships between variables, make predictions, and understand underlying patterns in data.

15. What are the various branches of statistics

Answer

# Statistics is a diverse and multidisciplinary field that involves the collection, analysis, interpretation, and presentation of data. It has several branches that focus on different aspects of statistical analysis and application. Here are some of the main branches of statistics:

# 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.

# 2. \*Inferential Statistics\*: Inferential statistics is concerned with making predictions, generalizations, and inferences about a population based on a sample of data. It includes techniques like hypothesis testing, confidence intervals, and regression analysis.

# 3. \*Probability\*: Probability is the study of uncertainty and randomness. It provides a mathematical framework for dealing with uncertainty and is essential for many statistical techniques.

# 4. \*Biostatistics\*: Biostatistics applies statistical methods to biological and medical research. It involves designing experiments, analyzing clinical trials, and drawing conclusions from health-related data.

# 5. \*Econometrics\*: Econometrics applies statistical methods to economic data to test economic theories, estimate relationships, and forecast economic variables.

# 6. \*Social Statistics\*: Social statistics focuses on analyzing social phenomena, such as population demographics, crime rates, education levels, and social trends.

# 7. \*Business and Economic Statistics\*: This branch involves using statistical methods to analyze business and economic data, including market research, financial analysis, and business decision-making.

# 8. \*Psychological Statistics\*: Psychological statistics is used in psychological research to analyze and interpret data related to human behavior and mental processes.

# 9. \*Environmental Statistics\*: Environmental statistics involves the application of statistical methods to study environmental issues, such as pollution levels, climate change, and ecological modeling.

# 10. \*Statistical Computing\*: This branch focuses on developing and using software tools and algorithms for statistical analysis and data manipulation. It includes programming languages like R and Python.

# 11. \*Quality Control and Six Sigma\*: Quality control uses statistical techniques to monitor and improve the quality of products and processes. Six Sigma is a data-driven methodology aimed at process improvement.

# 12. \*Spatial Statistics\*: Spatial statistics deals with data that have a geographical or spatial component, such as analyzing patterns in geographic data and spatial relationships.

# 13. \*Time Series Analysis\*: Time series analysis involves studying data collected over time to identify patterns, trends, and seasonal variations. It's used in economics, finance, and other fields.

# 14. \*Bayesian Statistics\*: Bayesian statistics is a probabilistic approach that involves updating beliefs based on new evidence. It's used for making predictions and decisions under uncertainty.

# 15. \*Nonparametric Statistics\*: Nonparametric statistics involves methods that do not assume specific distributional properties of the data. They are often used when the data do not meet the assumptions of traditional parametric methods.

# These branches represent just a subset of the many areas within the field of statistics. Each branch has its own specialized techniques and applications, contributing to the overall understanding and utilization of data in various fields.