STATISTICS WORKSHEET-1

1. Bernoulli random variables take (only) the values 1 and 0.
   a) True
   b) False
   Ans:- a) 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
   Ans:- 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
   Ans:- b) 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
   Ans:- c) 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
Ans:- c) Poisson

6. 10. Usually replacing the standard error by its estimated value does change the CLT.
   a) True
   b) False
   Ans:- 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
   Ans:- b) 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
   Ans:- a) 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
   Ans:- c) Outliers cannot conform to the regression relationship

10. What do you understand by the term Normal Distribution?
    Ans:- Normal Distribution is a fundamental concept in statistics and probability theory. It is a continuous probability distribution characterized by its bell-shaped curve, which is symmetric about the mean. Here are the key features and properties of a normal distribution:

a) **Symmetry:-** The normal distribution is perfectly symmetrical around its mean, meaning that the left and right halves of the curve are mirror images of each other.
b) **Mean, Median, and Mode:-** In a normal distribution, the mean, median, and mode all occur at the same point, which is the center of the distribution.
c) **Bell-Shaped Curve:-** The graph of a normal distribution has a distinctive bell shape, with the highest point at the mean and tails that extend infinitely in both directions, approaching but never touching the horizontal axis.
d) **Standard Deviation:-** The spread of the distribution is determined by the standard deviation (σ). About 68% of the data falls within one standard deviation (σ) of the mean, approximately 95% within two standard deviations, and about 99.7% within three standard deviations. This is known as the Empirical Rule or the 68-95-99.7 rule.
e) **Applications:-** Normal distribution is widely used in various fields, including psychology, finance, biology, and social sciences, as many natural phenomena tend to follow this distribution. It serves as a foundation for various statistical methods, including hypothesis testing and confidence intervals.
f) **Central Limit Theorem (CLT):-** The normal distribution is crucial in statistics because of the CLT, which states that the distribution of the sample means of a sufficiently large number of independent random variables will be approximately normally distributed, regardless of the original distribution of the variables.

11. How do you handle missing data? What imputation techniques do you recommend?
Ans:- Handling missing data is a critical step in data analysis and can significantly impact the results of your analysis. There are several techniques for dealing with missing data, including imputation methods. Here are some common strategies:
1. Identify the Nature of Missing Data:-
Missing Completely at Random (MCAR):- The missingness is unrelated to the observed or unobserved data. In this case, analyses can be conducted without any special treatment for missing values.
Missing at Random (MAR):- The missingness is related to the observed data but not the missing data itself. Imputation can be used effectively here.
Missing Not at Random (MNAR):- The missingness is related to the unobserved data. Special modeling techniques may be required.
2. Imputation Techniques:-
Here are some recommended imputation techniques based on the nature of the data:
Mean/Median/Mode Imputation:- Replace missing values with the mean (for continuous data), median (for skewed continuous data), or mode (for categorical data). This is a simple and quick method but can reduce variability in the data.
K-Nearest Neighbors (KNN) Imputation:- Uses the K-nearest neighbors algorithm to impute missing values based on the similarity of other observations. This technique can preserve the relationships in the data but can be computationally expensive.

Regression Imputation:- Predicts missing values using regression models based on other available data. This method can provide more accurate estimates than mean imputation but may lead to underestimating variability.

Multiple Imputation:- Involves creating multiple datasets with different imputed values, analyzing each dataset separately, and then combining the results. This technique accounts for uncertainty in the imputed values and is suitable for MAR data.

Last Observation Carried Forward (LOCF):- Used primarily in time-series data, this method fills missing values with the last observed value. It can introduce bias if the data is not stationary.

Interpolation/Extrapolation:- Suitable for time-series data, it estimates missing values by interpolating or extrapolating based on surrounding data points.

Machine Learning Algorithms:- Advanced methods, such as Random Forest or other predictive models, can be employed to predict and impute missing values based on patterns in the data.

3. Considerations:-

Understand the impact of missing data:- Before imputation, analyze the extent and pattern of missingness to inform your choice of method.

Evaluate the assumptions:- Different imputation methods come with assumptions. Ensure they align with your data characteristics.

Assess the results:- After imputation, check for biases or changes in distributions to ensure that the imputation has not distorted the data.

Sensitivity Analysis:- Conduct sensitivity analyses to determine how the imputation impacts your results.

12. What is A/B testing?

    Ans:- A/B testing**,** also known as split testing or bucket testing, is a statistical method used to compare two versions of a variable to determine which one performs better in achieving a specific outcome. It is widely used in marketing, product development, and web design to optimize user experience and improve conversion rates.

13. Is mean imputation of missing data acceptable practice?

    Ans:- Mean imputation is a commonly used method for handling missing data, but its acceptability depends on the context and the characteristics of the data. Here are some considerations regarding mean imputation:

    Pros of Mean Imputation

    a) Simplicity: Mean imputation is straightforward to implement. It involves calculating the mean of the observed values and replacing missing values with this mean, making it easy to understand and apply.

    b) Preservation of Data Size: This method allows you to retain all observations in your dataset, which can be useful in analyses that require complete datasets.

c) Minimal Impact on Distribution: For symmetric distributions, mean imputation may not drastically distort the overall distribution of the data.

Cons of Mean Imputation

a) Loss of Variability: Mean imputation reduces variability in the data because it replaces missing values with the same mean. This can lead to an underestimation of the standard deviation and skew the results of statistical analyses.

b) Bias in Estimates: If the missing data are not missing completely at random (MCAR), mean imputation can introduce bias into the analysis. It does not account for the relationships between variables or the potential impact of the missing values.

c) Invalid Inference: Mean imputation may lead to misleading conclusions in statistical tests, as it does not reflect the uncertainty associated with the missing values. This can affect confidence intervals and p-values.

d) Assumption of Normality: Mean imputation assumes that the data are normally distributed, which may not be the case for many real-world datasets. This assumption can lead to inaccurate estimates when applied to non-normal data.

When to Use Mean Imputation

a) When Missing Data are MCAR: If the missing data are completely random and not related to any other variables, mean imputation might be acceptable.

b) In Preliminary Analyses: It can be used as a quick fix in exploratory analyses where the goal is to get a general sense of the data, but it should not be the final method for handling missing data.

c) As Part of a Larger Strategy: Mean imputation can be used in combination with other methods, such as predictive modeling, to better account for missingness.

14. What is linear regression in statistics?

Ans:- Linear regression is a fundamental statistical method used to model the relationship between one or more independent (predictor) variables and a dependent (response) variable by fitting a linear equation to the observed data. It is widely used in various fields, including economics, biology, engineering, and social sciences, to understand relationships and make predictions. Linear regression is a powerful and widely used statistical tool that allows researchers and analysts to model and understand relationships between variables. Its simplicity, interpretability, and applicability to a variety of problems make it a foundational technique in statistics and data analysis.

15. What are the various branches of statistics?

Ans:- Statistics is a broad field that encompasses various branches, each focusing on different aspects of data analysis and interpretation. Here are the main branches of statistics:

1. Descriptive Statistics

Purpose: To summarize and describe the main features of a dataset.

Techniques: Includes measures of central tendency (mean, median, mode) and measures of dispersion (range, variance, standard deviation).
Visualizations: Utilizes charts, graphs, and tables (e.g., histograms, pie charts) to present data in a meaningful way.

2. Inferential Statistics

Purpose: To make inferences or generalizations about a population based on a sample of data.
Techniques: Involves hypothesis testing, confidence intervals, and regression analysis.
Applications: Used to draw conclusions and make predictions about a larger group from which the sample was taken.

3. Probability Theory

Purpose: To study and quantify uncertainty and randomness.
Concepts: Includes probability distributions (e.g., normal, binomial, Poisson), random variables, and theorems (e.g., Central Limit Theorem).
Applications: Forms the foundation for inferential statistics and is used in various fields such as finance, gambling, and risk assessment.

4. Bayesian Statistics

Purpose: To update the probability of a hypothesis as more evidence becomes available.
Concepts: Incorporates Bayes' theorem to combine prior knowledge with new data.
Applications: Used in various fields, including machine learning, genetics, and decision-making.

5. Non-parametric Statistics

Purpose: To analyze data without assuming a specific distribution.
Techniques: Includes methods such as the Mann-Whitney U test, Kruskal-Wallis test, and Wilcoxon signed-rank test.
Applications: Useful for small sample sizes or when data do not meet the assumptions of parametric tests.

6. Multivariate Statistics

Purpose: To analyze and interpret data involving multiple variables simultaneously.
Techniques: Includes multiple regression, factor analysis, cluster analysis, and principal component analysis (PCA).
Applications: Useful in fields like marketing, psychology, and biology, where multiple factors may influence outcomes.

7. Quality Control and Six Sigma

Purpose: To monitor and improve processes and products in manufacturing and service industries.
Techniques: Includes control charts, process capability analysis, and the use of statistical tools for quality improvement.
Applications: Widely used in production, healthcare, and service delivery to ensure quality standards.

8. Time Series Analysis

Purpose: To analyze data points collected or recorded at specific time intervals.
Techniques: Includes methods such as moving averages, exponential smoothing, and ARIMA models.

Applications: Used in economics, finance, and environmental studies to forecast future values based on historical data.

9. Survival Analysis

Purpose: To analyze time-to-event data, often used in medical research.

Techniques: Includes Kaplan-Meier estimators and Cox proportional hazards models.

Applications: Useful for studying the duration until an event occurs, such as death, disease recurrence, or failure of equipment.

10. Experimental Design

Purpose: To plan experiments and studies to ensure that valid and reliable conclusions can be drawn.

Techniques: Includes randomization, replication, and blocking to control for confounding variables.

Applications: Used in clinical trials, agricultural studies, and industrial experiments.