# 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- The Normal Distribution, also known as the Gaussian distribution or bell curve, is a continuous probability distribution that is symmetric around its mean (average) value. It is characterized by its bell-shaped curve and is fully defined by its mean and standard deviation. The probability density function (pdf) of a normal distribution is given by the formula:

$$f(x|\mu,\sigma)=\sigma 2\pi 1 e-2\sigma 2(x-\mu)2$$

Where:

$\mu$ is the mean of the distribution

,$\sigma$ is the standard deviation,

$\pi$ is a mathematical constant (approximately 3.14159)

$e$ is the base of the natural logarithm (approximately 2.71828)

Properties of the normal distribution include:

1. **Symmetry:** The distribution is symmetric around its mean, meaning that the probability of an outcome being a certain distance from the mean is the same on both sides.

**2. Mean, Median, and Mode:** They are all equal and located at the center of the distribution.

3. **68-95-99.7 Rule:** Approximately 68% of the data falls within one standard deviation of the mean, 95% within two standard deviations, and 99.7% within three standard deviations.

4. **Standard Normal Distribution:** If a normal distribution has a mean of 0 and a standard deviation of 1, it is called a standard normal distribution. Z-scores, which represent the number of standard deviations a data point is from the mean, are often used to analyze data in the standard normal distribution.

The normal distribution is a fundamental concept in statistics and is applicable in various fields, including physics, biology, economics, and social sciences. Many natural phenomena, such as heights of individuals in a population or errors in measurements, tend to follow a normal distribution.

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

Ans- Handling missing data is an important aspect of data preprocessing in various data analysis and machine learning tasks. The approach to dealing with missing data depends on the nature of the data and the underlying reasons for the missing values. Here are some common strategies for handling missing data:

1. **Deletion of Missing Data:**

**Listwise Deletion:** Removing entire rows with missing values. This is simple but can lead to loss of valuable information.

**Pairwise Deletion:** Analyzing only the available pairs of variables for each analysis. This can be useful when missing values are spread across different variables.

2. **Imputation Techniques:**

**Mean, Median, or Mode Imputation:** Replace missing values with the mean, median, or mode of the observed values for that variable. This is a simple method but may not be suitable if data is not missing completely at random.

**Forward Fill or Backward Fill:** Propagate the last observed value forward or use the next observed value to fill missing data in time series or sequential data.

**Linear Regression Imputation:** Predict missing values based on a linear regression model using other variables.

**K-Nearest Neighbors (KNN) Imputation:** Replace missing values with the average of the k-nearest neighbors in the feature space.

**Multiple Imputation:** Generate multiple plausible values for each missing data point, creating multiple complete datasets. Analyze each dataset separately and combine results. This accounts for the uncertainty associated with imputed values.

3. **Domain-Specific Imputation:**

**Custom Imputation:** Use domain knowledge to impute missing values. For example, replacing missing income values with the average income for a specific occupation.

4. **Prediction Models:**

**Build Models with Missing Data:** Use machine learning algorithms that can handle missing values during training, such as XGBoost or Random Forests.

5. **Missing Data Indicators:**

**Create Indicator Variables:** Introduce binary indicator variables to signal whether a value is missing for a particular observation.

The choice of imputation technique depends on the context of the data and the reasons for missingness. It's crucial to carefully consider the implications of each method and recognize that imputation introduces some level of uncertainty. Additionally, it's important to evaluate the impact of missing data handling on the results of subsequent analyses.

## 12. What is A/B testing?

Ans-. A/B testing, also known as split testing, is a method used in marketing, product development, and user experience optimization to compare two versions of a product or webpage and determine which one performs better. The purpose of A/B testing is to make data-driven decisions about changes to a particular process or element, with the ultimate goal of improving performance.

A/B testing involves creating two (or more) versions of a webpage, email, advertisement, or any other item you want to test. These versions are referred to as the "A" and "B" variants.

The "A" variant is typically the existing or control version, while the "B" variant is the modified version with changes.

## 13. Is mean imputation of missing data acceptable practice?

Ans Mean imputation is a method of handling missing data by replacing missing values with the mean of the observed values for that variable. While mean imputation is a simple and

quick way to address missing data, it has both advantages and disadvantages, and its acceptability depends on the context and assumptions of the data.

Advantages of mean imputation:

1. **Preservation of Sample Size:** Mean imputation allows you to retain the entire sample size, which can be important, especially when dealing with small datasets.

**2. Easy Implementation:** It is a straightforward method to implement and understand, making it accessible for users without advanced statistical knowledge.

Disadvantages of mean imputation:

1. **Loss of Variability:** Mean imputation assumes that the missing values are missing completely at random (MCAR). If data are not missing completely at random, mean imputation can lead to biased estimates and a loss of variability in the imputed variable.

**2.Underestimation of Standard Errors:** Imputing missing values with the mean can lead to underestimation of standard errors, affecting the precision of statistical analyses

3. **Potential Distortion of Relationships:** Mean imputation can distort relationships between variables, especially if the missing data are related to specific subgroups or patterns in the data.

**4.Not Suitable for Categorical Data:** Mean imputation is typically used for continuous variables. For categorical variables, alternative imputation methods such as mode imputation or more sophisticated techniques may be more appropriate.

In summary, mean imputation can be acceptable in certain situations, especially when the missing data are believed to be missing completely at random, and the assumptions of normality hold. However, researchers and analysts should be aware of its limitations and consider alternative imputation methods, such as multiple imputation or model-based imputation, in more complex situations. The choice of imputation method should be based on the characteristics of the data and the research question at hand.

## 14. What is linear regression in statistics?.

Ans-Linear regression is a statistical method used to model the relationship between a dependent variable and one or more independent variables by fitting a linear equation to the observed data. The goal is to find the best-fitting line that describes the linear relationship between the variables. This line can then be used to make predictions about the dependent variable based on the values of the independent variables

The linear regression model aims to estimate the values of the coefficients ($\beta$) that minimize the sum of the squared differences between the observed values of the dependent variable and the values predicted by the model.

There are different types of linear regression, depending on the number of independent variables. Simple linear regression involves one independent variable, while multiple linear regression involves two or more independent variables.

Linear regression is widely used in various fields, including economics, finance, biology, and social sciences, for tasks such as predicting stock prices, understanding the relationship between variables, and making informed decisions based on data analysis. The assumptions of linear regression include linearity, independence, homoscedasticity, and normality of residuals, and these assumptions should be checked before interpreting the results of a linear regression 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, interpretation, and application. Some of the main branches of statistics include:

1. **Descriptive Statistics:** Descriptive statistics involve methods for summarizing and presenting data in a meaningful way. Measures such as mean, median, mode, range, and standard deviation fall under descriptive statistics.

**2. Inferential Statistics:** Inferential statistics involves making inferences and predictions about a population based on a sample of data. Common techniques include hypothesis testing, confidence intervals, and regression analysis.

3. **Probability Theory:** Probability theory is the mathematical foundation of statistics. It deals with the study of random events and the likelihood of their occurrence. Probability is fundamental to understanding uncertainty and randomness in statistical analyses.

4. **Biostatistics:** Biostatistics is the application of statistical methods to biology and related fields. It plays a crucial role in designing experiments, analyzing biological data, and drawing conclusions in areas such as medicine, genetics, and environmental science.

5. **Econometrics:** Econometrics is the application of statistical methods to economic data. It involves modeling and analyzing economic relationships, testing hypotheses, and making predictions in the field of economics

6. **Social Statistics:** Social statistics involves the application of statistical methods to social science research. It is used to analyze data in sociology, psychology, political science, and other social sciences.

7. **Statistical Computing:** Statistical computing involves the development and application of computational methods for statistical analysis. This includes programming languages and software used for data manipulation, analysis, and visualization.

8. **Multivariate Statistics:** Multivariate statistics deals with the analysis of datasets that involve multiple variables. Techniques such as factor analysis, principal component analysis, and multivariate regression are used to explore relationships among variables.

9. **Time Series Analysis:** Time series analysis focuses on analyzing data collected over time. It is commonly used in economics, finance, and other fields to study trends, seasonality, and other patterns in time-ordered data.

10. **Spatial Statistics:** Spatial statistics involves the analysis of data that has a spatial component. It is used in fields such as geography, ecology, and environmental science to analyze patterns and relationships in spatially distributed data.

11. **Quality Control and Six Sigma:** Quality control involves the application of statistical methods to monitor and improve the quality of processes and products. Six Sigma is a set of techniques and tools for process improvement that relies heavily on statistical methods.

these branches often overlap, and statisticians may specialize in one or more of these areas based on their interests and the specific needs of their field of study or industry.