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

A Poisson distribution is a discrete probability distribution. It gives the probability of an event happening a certain number of times (k) within a given interval of time or space.

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-

In a normal distribution, data is symmetrically distributed with no skew. When plotted on a graph, the data follows a bell shape, with most values clustering around a central region and tapering off as they go further away from the center. Normal distributions are also called Gaussian distributions or bell curves because of their shape.

The properties of normal distribution is that the curve mean, median and mode are exactly same to eachother

Actually the normal distribution is defined by two parameters: the mean (μ), which locates the center of the distribution, and the standard deviation (σ), which determines the spread or width of the distribution.

The distribution is symmetric around its mean, with equal areas within the curves on both side.

The mean is the location parameter while the standard deviation is the scale parameter.

The standard deviation stretches or squeezes the curve. A small standard deviation results in a narrow curve, while a large standard deviation leads to a wide curve.

The mean determines where the peak of the curve is centered. Increasing the mean moves the curve right, while decreasing it moves the curve left.

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

Ans-

Handling missing data is vital in data analysis and machine learning, as it can lead to bias and diminish the quality of insights or predictive models. Various techniques exist for managing missing data, and the best approach depends on the data characteristics and the analysis context.

- Mean/Median/Mode Imputation: Fill in missing values with the mean (for continuous data), median (for continuous data with outliers), or mode (for categorical data) of the column. This method is straightforward but can reduce data variability.
- Forward Fill/Backward Fill: Use the previous (forward fill) or next (backward fill) data point to fill in missing values, which is particularly useful for time series data.
- K-Nearest Neighbors (KNN) Imputation: Replace missing values using the values of the k-nearest neighbors, taking into account the similarity between observations.
- Regression Imputation: Predict missing values using a regression model based on other variables in the dataset.

12. What is A/B testing?

A/B testing, also referred to as split testing, is a statistical technique used to compare two variations of a variable (A and B) to see which one performs better. This approach is widely utilized across different fields such as marketing, web development, and product development to make informed decisions based on data.

A/B testing involves comparing two iterations of an email, website, or other marketing assets to evaluate their performance differences. This is done by distributing one version to one group and the other version to a different group. The performance of each version is then observed and analyzed. Think of it as a competition where two versions of your assets compete against each other to determine the more effective one.

13. Is mean imputation of missing data acceptable practice?

Ans-

Mean imputation, which replaces missing data with the average of the observed values, is a straightforward and widely used method. However, its appropriateness depends on the context and specific analysis requirements.

Acceptability of Mean Imputation:

Small Proportion of Missing Data: It may be acceptable if the missing data is minimal and doesn't significantly affect the overall analysis.

Preliminary Analysis: Useful for preliminary or exploratory data analysis when the primary goal is to quickly understand the dataset.

Non-Critical Applications: Suitable for non-critical applications where high precision is not essential.

15. What are the various branches of statistics?

Ans:

Descriptive statistics utilize data to describe the characteristics of a population, either through numerical calculations, graphs, or tables, providing a graphical summary of the data. Its primary purpose is to summarize objects, among other things. This branch is divided into two main categories:

Measures of Central Tendency

Measures of Variability

Measures of Central Tendency

Measures of central tendency, also known as summary statistics, represent the central point or a typical value within a dataset or sample. The three common measures of central tendency in statistics are:

Mean: The average of the data points.

Median: The middle value when the data points are ordered.

Mode: The most frequently occurring value in the dataset.

Inferential Statistics involves making inferences and predictions about a population based on a sample of data drawn from that population. It extends findings from a smaller dataset to a larger population, applying probabilities to derive conclusions.

This branch of statistics is employed to interpret and analyze results, drawing meaningful conclusions from descriptive statistics. It is closely linked with hypothesis testing, focusing on rejecting or failing to reject the null hypothesis.

14. What is linear regression in statistics?

Linear Regression is a machine learning algorithm based on supervised learning. It performs a regression task. Regression models a target prediction value based on independent variables.

It is mostly used for finding out the relationship between variables and forecasting.

Different regression models differ based on - the kind of relationship between dependent and independent variables they are considering, and the number of independent variables getting used.

Performs the task to predict a dependent variable value (y) based on a given independent variable (x).

So, this regression technique finds out a linear relationship between x (input) and y(output).

Hence, the name is Linear Regression.

In the figure above, X (input) is the work experience and Y (output) is the salary of a person.

The regression line is the best fit line for our model.

Example: To calculate the sales records and affecting factors by dependent variable and independent variable.

Two types of equations:

1.Simplelar equation: y=mx+c

2.Multi linear equation: y= |x