STATISTICS

1. Bernoulli random variables take (only) the values 1 and 0.

a) True

b) False

Answer: - **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

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: - **b) modeling bounded count data**

4. Point out the correct statement.

a) The exponent of 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: - **d) All of the mentioned**

5. ______ random variables are used to model rates.

a) Empirical

b) Binomial

c) Poisson

d) All of the mentioned

Answer: - **c) Poisson**

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

a) True

b) False

Answer: - **b) False**

7. Which of the following testing is concerned with making decisions using data?

a) Probability

b) Hypothesis

c) Causal

d) None of the mentioned

Answer: - **b) Hypothesis**

8. 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: - **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

Answer: - **c) Outliers cannot conform to the regression relationship.**

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

\: - **Normal Distribution is also called as bell curve. It is a probability distribution that is symmetric about the mean, showing that data near the mean are more frequent in occurrence than data far from the mean.**

In a normal distribution the mean is zero and the standard deviation is 1. It has zero skew. The normal distribution is the most important probability distribution in statistics because it fits many natural phenomena. For example, heights, blood pressure, measurement error, and IQ scores follow the normal distribution. The normal distribution model is motivated by the Central Limit Theorem.

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

Answer: - **There are two primary methods to solve the error: imputation or the removal of data.**

The imputation method develops reasonable guesses for missing data. It's most useful when the percentage of missing data is low. If the portion of missing data is too high, the results lack natural variation that could result in an effective model.

The other option is to remove data. When dealing with data that is missing at random, related data can be deleted to reduce bias. Removing data may not be the best option if there are not enough observations to result in a reliable analysis. In some situations, observation of specific events or factors may be required.

Data can be missing in the following ways:-

- **Missing Completely At Random (MCAR):** When missing values are randomly distributed across all observations, then we consider the data to be missing completely at random.
- **Missing At Random (MAR):** The key difference between MCAR and MAR is that under MAR the data is not missing randomly across all observations, but is missing randomly only within sub-samples of data.
- **Not Missing At Random (NMAR):** When the missing data has a structure to it, we cannot treat it as missing at random.

Imputation Techniques: -

1. Mean or Median Imputation

2. Multivariate Imputation by Chained Equations (MICE)

3. Random Forest

You could find missing/corrupted data in a dataset and either drop those rows or columns, or decide to replace them with another value.
In Pandas, there are two very useful methods: isnull () and dropna() that will help you find columns of data with missing or corrupted data \nd drop those values. If you want to fill the invalid values with a placeholder value (for example, \), you could use the fillna() method.

12. What is A/B testing?

Answer: - **A/B testing also known as split testing. An AB test is an example of statistical hypothesis testing, a process whereby a hypothesis is made about the relationship between two data sets and those data sets are then compared against each other to determine if there is a statistically significant relationship or not.**

Essentially, A/B testing eliminates all the guesswork out of

and enables experience optimizers to make data-backed decisions. In A/B testing, A refers to 'control' or the original testing variable. Whereas B refers to 'variation' or a new version of the original testing variable.

13. Is mean imputation of missing data acceptable practice?

A\swer: - **It is a non-standard, it uses Random Forest. It is use to predict the missing data. It also can be used for both i.e. continuous as well as categorical data and so it makes advantageous over other imputations.**

- **Bad practice in general**
- **If just estimating means: mean imputation preserves the mean of the observed data**
- **Leads to an underestimate of the standard deviation**
- **Distorts relationships between variables by "pulling" estimates of the correlation toward zero**

There are some limitations too: -

1. Mean imputation does not preserve the relationship among variables. It preserves the mean of observed data. If data is missing completely at random, the estimate of the mean remains unbiased.

2. Mean Imputation leads to an underestimate of standard errors.

14. What is linear regression in statistics?

Answer:- ***Linear regression*** **quantifies the relationship between one or more** ***predictor variable(s)*** **and one** ***outcome variable.*** **Linear regression is commonly used for predictive analysis and modelling. Linear regression attempts to model the relationship between two variables by fitting a linear equation to observed data. One variable is considered to be an explanatory variable, and the other is considered to be a dependent variable. For example, a modeller might want to relate the weights of individuals to their heights using a linear regression model.**

A linear regression line has an equation of the form $Y = mx + c$, where X is the explanatory variable and Y is the dependent variable. The slope of the line is m, and c is the intercept (the value of y when $x = 0$).

The simplest form of the regression equation with one dependent and one independent variable is defined by the formula y = c + b*x, where y = estimated dependent variable score, c = constant, b = regression coefficient, and x = score on the independent variable.

Types of linear regression: -

1. Simple linear regression

2. Multiple linear regressions

3. Logistic regression

4. Ordinal regression

5. Multinomial regression

15. What are the various branches of statistics?

Answer: - Various branches of statistics are given below: -

There are two main branches of statistics :-

1. Descriptive Statistics

Descriptive statistics is the first part of statistics that deals with the collection of data. People seem it too easy, but it is not that easy. The statisticians need to be aware of the designing and experiments. They also need to choose the right focus group and avoid biases. In contrast, Descriptive statistics are used in use to do various kinds of analysis

on different studies**.**

Descriptive statistics have two parts

- Central tendency measures
- Variability measures

Measures of Central Tendency

Central tendency measures specifically help statisticians evaluate the distribution center of values. These tendency measures are:

Mean

Mean is a conventional method used to describe the central tendency. Typically, to calculate the average of values, count all values, and then divide them with the number of available values.

Median

It is the result that is in the middle of a set of values. An easy way to calculate the median is to edit the results in numerical journals and locate the result that is in the center of the distributed sample.

Mode

The mode is the frequently occurring value in the given data set.

Measures of Variability

The variability measure helps statisticians to analyze the distribution that is spreading from a specific data set. Some of the variables of variability include quartiles, ranges, variances, and standard deviation.

2. Inferential Statistics

The inference statistics are techniques that enable statisticians to use the information collected from the sample to conclude, bring decisions, or predict a defined population.

Inference statistics often speak in terms of probability by using descriptive statistics. Besides, these techniques are used primarily by a statistician for data analysis, drafting, and making conclusions from limited information. That is obtained by taking samples and testing how reliable they are.
Most predictions of the future and generalization on a population study of a smaller specimen are in the scope of the inference statistics. Besides, most of the social sciences experiments deal with the study of a small sample population that helps determine the behavior of the community.

Different types of inferential statistics include:

- Regression analysis
- Analysis of variance (ANOVA)
- Analysis of covariance (ANCOVA)
- Statistical significance (t-test)
- Correlation analysis