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:- d) All of the mentioned

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

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

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

Ans. A normal distribution or Gaussian distribution refers to a probability distribution where the values of a random variable are distributed symmetrically. These values are equally distributed on the left and the right side of the central tendency. Thus, a bell-shaped curve is formed.

It has two key parameters: the mean (μ) and the standard deviation (σ). This probability method plays a crucial role in asset return calculation and risk management strategy decisions.

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

Ans. When dealing with missing data, data scientists can use 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.

Before deciding which approach to employ, data scientists must understand why the data is missing.

In missing data research literature, these three methods are highly respected for their ability to improve data quality (Learn more: regression imputation; predictive mean matching; hot deck imputation). Regression imputation and hot deck imputation seem to have increased their popularity until 2013.

12. What is A/B testing?

Ans. A/B tests, also known as split tests, allow you to compare 2 versions of something to learn which is more effective. Simply put, do your users like version A or version B better?

The concept is similar to the scientific method. If you want to find out what happens when you change one thing, you have to create a situation where only that one thing changes.

Think about the experiments you conducted in elementary school. If you put 2 seeds in 2 cups of dirt and put one in the closet and the other by the window, you'll see different results. This kind of experimental setup is A/B testing.

13. Is mean imputation of missing data acceptable practice?

Ans. Mean imputation preserves the mean of the dataset with missing values, as can be seen in our example above. This, however, is only appropriate if we assume that our data is normally distributed where it is common to assume that most observations are around the mean anyway. It also is substantially helpful, for small missing data cases.

14. What is linear regression in statistics?

Ans. Linear regression is a basic and commonly used type of predictive analysis. The overall idea of regression is to examine two things: (1) does a set of predictor variables do a good job in predicting an outcome (dependent) variable? (2) Which variables in particular are significant predictors of the outcome variable, and in what way do they–indicated by the magnitude and sign of the beta estimates–impact the outcome variable? These regression estimates are used to explain the relationship between one dependent variable and one or more independent variables. 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.

Naming the Variables. There are many names for a regression's dependent variable. It may be called an outcome variable, criterion variable, endogenous variable, or regressand. The independent variables can be called exogenous variables, predictor variables, or regressors.

Three major uses for regression analysis are (1) determining the strength of predictors, (2) forecasting an effect, and (3) trend forecasting.

15. What are the various branches of statistics?

Ans. The two main branches of statistics are **descriptive statistics and inferential statistics.**

Descriptive statistics deals with the presentation and collection of data. This is usually the first part of a statistical analysis. It is usually not as simple as it sounds, and the statistician needs to be aware of designing experiments, choosing the right focus group and avoid biases that are so easy to creep into the experiment.

Inferential statistics, as the name suggests, involves drawing the right conclusions from the statistical analysis that has been performed using descriptive statistics. In the end, it is the inferences that make studies important and this aspect is dealt with in inferential statistics.