STATISTICS WORKSHEET-1

Q1 to Q9 have only one correct answer. Choose the correct option to answer your question.

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

a) True

b) False

Answers: 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

Answers: 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

Answers: 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

Answers: d) All of the mentioned

5. _____ random variables are used to model rates.

a) Empirical

b) Binomial

c) Poisson

d) All of the mentioned

Answers: d) Poisson

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

a) True

b) False

Answers: 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

Answers: 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

Answers: 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 WORKSHEET

Answers: 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?

Answers: The Normal distribution also called Gaussian distribution. The normal distribution most commonly seen continuous distribution in nature. Just as the binomial distribution, every event is independent from one another. In the normal distribution the mean, median, and mode all line up such that the centre of the distribution is mean. Because of this exactly half of the results fall to either side of the mean. The normal distribution is also identifiable by its bell shape and may sometimes be referred to as a bell curve.

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

Answers: The concept of missing data is important to comprehend in order to efficiently manage data. If the researcher, programmer, or academician does not properly handle the missing figures, then they may get to the wrong conclusion about the data, which will have a significant impact on the modelling phase.

There are mainly four imputation techniques

1. **Univariate imputation, or mean imputation, is when values are imputed using only the target variable.**
2. **Multivariate imputation: Impute values depending on other factors, such as estimating missing values based on other variables using linear regression.**
3. **Single imputation: To construct a single imputed dataset, only impute any missing values once inside the dataset.**
4. **Numerous imputations: imputation of the same missing values multiple times inside the dataset. This essentially entails repeating a single imputation to obtain numerous imputed datasets.**

12. What is A/B testing?

Answers: A/B testing in its simplest sense is an experiment on two variants to see which performs better based on a given metric. A/B testing is a form of statistical and two-sample hypothesis testing. Statistical hypothesis testing is a method in which a sample dataset is compared against the population data. Two-sample hypothesis testing is a method in determining whether the differences between the two samples are statistically significant or not.

13. Is mean imputation of missing data acceptable practice?

Answers: No, mean imputation is not a good practice by below reasons:

1. **Mean imputation reduces the variance of the imputed variables.**
2. **Mean imputation shrinks standard errors, which invalidates most hypothesis tests and the calculation of confidence interval.**
3. **Mean imputation does not preserve relationships between variables such as correlations.**

14. What is linear regression in statistics?

Answers: Simple linear regression is used to model the relationship between two continuous variables. Often, the objective is to predict the value of an output variable (or response) based on the value of an input (or predictor) variable. The simplest form of the regression equation with one dependent and one independent variable is defined by the formula y = a + b*x, where y = estimated dependent variable score, a = constant/intercept, b = regression coefficient, and x = score on the independent variable.

15. What are the various branches of statistics?

Answers: There are two branches of statistics:

1. **Descriptive Statistics**
2. **Inferential Statistics**

Descriptive Statistics: deals with the presentation and collection of data. Data can be described without any statistical tools .

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.