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

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

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

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

5. ______ random variables are used to model rates.

a) Empirical

b) Binomial

c) Poisson

d) All of the mentioned

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

a) True

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

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

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

Q10 and Q15 are subjective answer type questions, Answer them in your own words briefly.

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

It is a probability distribution graphically represented as a “bell curve” that is symmetric about the mean, i.e., where the data tends to be around a central value with no bias, left or right. And where the standard deviation determines the amount of dispersion away from the mean.

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

This will depend on the reasons behind these missing values and on the type of variable. If a column or a row has many missing values, it will be recommended to delete this entire row or column. But sometimes, imputation methods can deliver better results in the following situations:

For numerical variables:

If there´s a good knowledge of the topic and the dataset, one imputation technique I will recommend replacing the missing value with an arbitrary value like 0 or with a value that reflects an educated guess about the missing value.

If outliers are not present in the dataset, missing values can be replaced with the mean. Otherwise, an alternative is replacing missing values with the median.

Other alternatives are replacing the data with the previous value (forward fill) or with the next value (backard fill), or using interpolation.

For categorical variables:

Missing data can be replaced with the mode.

12. What is A/B testing?

It 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 not only which one performs better but also if there is a statistically significant relationship or not.

13. Is mean imputation of missing data acceptable practice?

It is not always a good practice, since it can ignore feature correlation, and also because it can decrease the variance of the imputed variables while increasing bias. As a result of the reduced variance, the model is less accurate and the confidence interval is narrower.

14. What is linear regression in statistics?

It is an approach to predict the value of a variable (dependent variable) based on the value of another variable (independent variable). In this way, it attempts to model the relationship between these two variables by fitting a linear equation to observed data.

15. What are the various branches of statistics?

1. Descriptive statistics: it describes the properties of sample and population data, it is mostly focused on the central tendency, variability, and distribution of sample data.
2. Inferential statistics: it uses those properties to test hypotheses and draw the right conclusions about the characteristics of a population from the characteristics of a sample. Inferential statistics are used to make generalizations about large groups.