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

ANS: A

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

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

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

5. _____ random variables are used to model rates.

a) Empirical b) Binomial c) Poisson d) All of the mentioned

ANS: C

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

a) True b) False

ANS: B

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

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

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

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 distributed is an arrangement of data set in which most values cluster in the middle of the range and the rest it taper off symmetrically toward either end. A graphical representation of a normal distribution is sometimes called a bell curve because of its flared shape. Normal distribution curves are sometimes designed with a histogram inside the curve. The graphs are commonly used in mathematics , statistics and corporate data analytics.

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

ANS: Missing data can be dealt with in a variety of ways. I believe the most common reaction is to ignore it. the application will remove things in a listwise sequence most of the time. Depending on why and how much data is gone. Another common strategy among those who pay attention is imputation. Imputation is a process of substituting an estimate for missing values and analysing the entire data set as if the imputed values were the true observed values.

Most imputation techniques used and they are mentioned as: mean imputation,substitution , hot deck imputation , cold deck imputation, regression imputation,stochastic regression imputation, interpolation and extrapolation, single and multiple imputation.

12. What is A/B testing?

ANS: A/B testing is a way to compare two versions of something to figure out which performs better. A/B in its current form, came into existence in 1990s . in 1920s statistician and biologist Ronald Fisher discovered the most important principle behind A/B testing and randomised controlled experiments in general. We run by deciding what it is you want to test. In its simplest form there are two treatments and one acts as the control for the other.

13. Is mean imputation of missing data acceptable practice?

ANS: the process of replacing null values in a data collection with data's mean is known as mean imputation. And it is not acceptable practice as it ignores feature correlation. Secondly it decreases the variance of our data while increasing bias. As a result of the reduced variance, the model is less accurate and confidence interval is narrower.

14. What is linear regression in statistics?

ANS: linear regression analysis is used to predict the value of a variable based on the value of another variable. The variable you want to predict is called the dependent variable . the variable you are using to predict the other variable's value is called the independent variable. This form of analysis estimates the coefficients of the linear equation, involving one or more independent variables that best predict the dependent variable . linear regression fits a straight line or surface that minimizes the descrepancies between predicted and actual output values.

15. What are the various branches of statistics?

ANS: the various branches of statistics are:

1- Data collection: it is all about how the actual data is collected but there are some significant issues to consider when actually collecting data.
2- Descriptive statistics: it deals with presenting the data we have. This can take two basic forms- presenting aspects of the data either visually or numerically.
3- Inferential statistics: it deals with making conclusions about the data. This is quite a wide area.
4- Discrete and continuous data: data comes into two distinct types. Discrete data can take distinct values, which can be clearly identified and separated.
5- Frequency distributions: sometimes the actual collection data isn't very meaningful and we wish to put into categories.