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]

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

10. What do you understand by the term Normal Distribution?
11. How do you handle missing data? What imputation techniques do you recommend?
12. What is A/B testing?
13. Is mean imputation of missing data acceptable practice?
14. What is linear regression in statistics?
15. What are the various branches of statistics?

Ans 10. The Normal Distribution is defined by the probability density function for a continuous random variable in a system. Let us say, f(x) is the probability density function and X is the random variable. Hence, it defines a function which is integrated between the range or interval (x to x + dx), giving the probability of random variable X, by considering the values between x and x+dx. The Normal Distribution is also called the Gaussian Distribution, is the most significant continuous probability distribution. Sometimes it is also called a bell curve.

The random variables following the normal distribution are those whose values can find any unknown value in a given range. For example, finding the height of the students in the school. Here, the distribution can consider any value, but it will be bounded in the range say, 0 to 6ft. This limitation is forced physically in our query.

Whereas the normal distribution doesn't even bother about the range. The range can also extend to – to + and still we can find a smooth curve. These random variables are called Continuous Variables, and the Normal Distribution then provides here probability of the value lying in a particular range for a given experiment. Also, use the normal distribution calculator to find the probability density function by just providing the mean and standard deviation value

Ans 11. Most common used techniques for handling missing data are:

Deleting Rows with missing values
Impute missing values for continuous variable
Impute missing values for categorical variable
Other Imputation Methods
Prediction of missing values

Various imputation techniques are:
1. For numerical variables:
Mean / Median Imputation.
Arbitrary value imputation
End of tail imputation
Mode imputation

2. For Categorical Variables:
Frequent category imputation
Adding missing category

3. Dropping null values.

Ans 12. A/B testing also known as bucket testing or split-run testing is a user experience research methodology. A/B tests consist of a randomized experiment with two variants, A and B. It includes application of statistical hypothesis testing or "two-sample hypothesis testing" as used in the field of statistics. A/B testing is a way to compare two versions of a single variable, typically by testing a subject's response to variant A against variant B and determining which of the two variants is more effective.
Various steps involved in A/B test are:
1. Formulate your Hypothesis
2.Deciding on Splitting and Evaluation Metrics
3.Create your Control group and Test group
4.Length of the A/B Test
5.Conduct the Test
6. Draw Conclusions

Ans 13. Mean imputation is typically considered terrible practice since it ignores feature correlation. Consider the following scenario: we have a table with age and fitness scores, and an eight-year-old has a missing fitness score. If we average the fitness scores of people between the ages of 15 and 80, the eighty-year-old will appear to have a significantly greater fitness level than he actually does.

Second, mean imputation decreases the variance of our data while increasing bias. As a result of the reduced variance, the model is less accurate and the confidence interval is narrower.

Ans 14. In statistics, linear regression is a linear approach for modelling the relationship between a scalar response/target and one or more explanatory variables/features (also known as dependent and independent variables). The case of one explanatory variable is called simple linear regression; for more than one, the process is called multiple linear regression.
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 value of the dependent variable. Linear regression fits a straight line or surface that minimizes the discrepancies between predicted and actual output values. There are simple linear regression calculators that use a "least squares" method to discover the best-fit line for a set of paired data. You then estimate the value of X (dependent variable) from Y (independent variable).

Ans 15. Three major branches of statistics are, data collection, descriptive and inferential statistics.