STATISTICS WORKSHEET-1

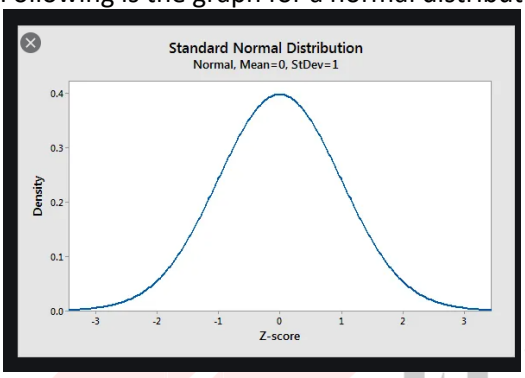
Q1 to Q9 have only one correct answer. Choose the correct option to answer your question.

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- normaldistribution
   b) Sums of normally distributed random variables are again normally distributed even if the variablesare dependent
   c) The square of a standard normal random variable follows what is called chi-squareddistribution
   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 theoriginal 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?

Ans10: The normal distribution is a probability function that describes how the values of a variable are distributed. It is a symmetric distribution where most of the observations cluster around the central peak and the probabilities for values further away from the mean taper off equally in both directions.
Following is the graph for a normal distribution:



Ans11: There are many ways one could handle missing data like:

1. Listwise or case deletion. ...
2. Pairwise deletion. ...
3. Mean substitution. ...
4. Regression imputation. ...
5. Last observation carried forward. ...
6. Maximum likelihood. ...
7. Expectation-Maximization. ...
8. Multiple imputations.

And the imputation techniques that I would recommend are as follows:-

1. Imputation with mean : Missing data is replaced by mean of the column
2. Imputation with median : Missing data is replaced by mean of the column
3. Imputation with Mode: Missing data is replaced with mode of the column
4. Imputation with linear regression: With real valued data, this is a common technique. The missing value is replaced by performing linear regression based on the other feature values.

Ans12: A/B testing also known as bucket testing or split-run testing is a user experience research methodology.[1] A/B tests consist of a randomized experiment with two variants, A and B.[2][3] 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.

Ans13: Yes it can be used. It is a non-standard, but a fairly flexible imputation algorithm. It uses RandomForest at its core to predict the missing data. It can be applied to both continuous and categorical variables which makes it advantageous over other imputation algorithms.

Ans14 : In statistics, linear regression is a linear approach to modelling the relationship between a scalar response and one or more explanatory variables (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 regressions.[1] This term is distinct from multivariate linear regression, where multiple correlated dependent variables are predicted, rather than a single scalar variable.
Linear regression was the first type of regression analysis to be studied rigorously, and to be used extensively in practical applications.[4] This is because models which depend linearly on their unknown parameters are easier to fit than models which are non-linearly related to their parameters and because the statistical properties of the resulting estimators are easier to determine.

Ans15: The two main branches of statistics are descriptive and inferential statistics.

They are both used as part of most statistical analyses. Descriptive statistics are first used to organize and present collected data into a coherent form that can be further analyzed through inferential statistics.

Examples of descriptive statistics include:

- Mean
- Median
- Variance
- Standard Deviation
- Correlation Coefficient

Inferential statistics are used to draw conclusions from descriptive statistics, and make predictions or assumptions about the entire population from which the data were sampled. Examples include (from simple to complex):

- Student’s T-test
- Regression analysis
- Analysis of Variance (ANOVA)
- Statistical Equation Modelling (SEM)