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

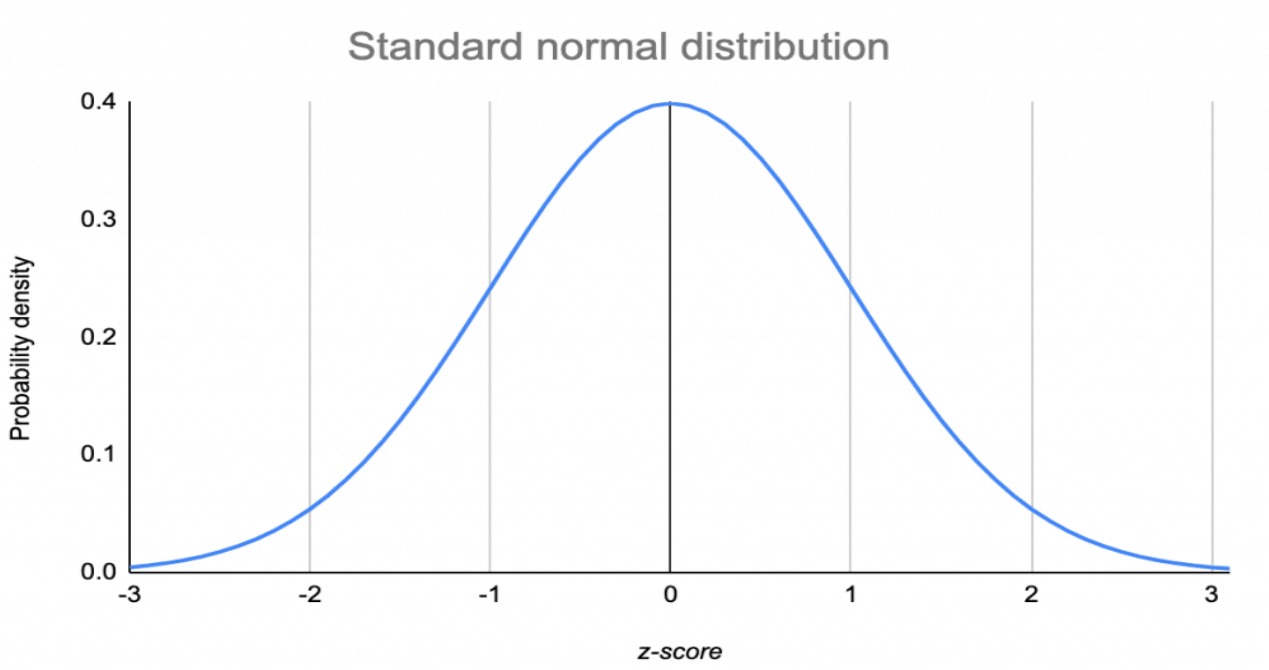
Q10and Q15 are subjective answer type questions, Answer them in your own words briefly.

1. What do you understand by the term Normal Distribution?
2. How do you handle missing data? What imputation techniques do you recommend?
3. What is A/B testing?
4. Is mean imputation of missing data acceptable practice?
5. What is linear regression in statistics?
6. What are the various branches of statistics?

ANSWERS

1.a
2.a
3.b
4.d
5.c
6.b
7.b
8.a
9.c

10. Normal Distribution is the probability distribution curve in which mean is zero and standard deviation is 1. It is symmetrical in nature with zero skew. It is also called as Bell curve.



11. Imputation with constant value:

it replaces the missing values with either zero or any constant value.

Imputation using statistics:

Mean" will replace missing values using the mean in each column. It is preferred if data is numeric and not skewed.

“Median” will replace missing values using the median in each column. It is preferred if data is numeric and skewed.

“Most frequent” will replace missing values using the most_frequent in each column. It is preferred if data is a string(object) or numeric.

K_Nearest Neighbor Imputation:

The KNN algorithm helps to impute missing data by finding the closest neighbors using the Euclidean distance metric to the observation with missing data and imputing them based on the non-missing values in the neighbors.

12. A/B testing is also known as split testing or bucket testing. It is a user experience research methodology. A/B tests consist of a randomized experiment with two variants, A and B. It is a method of comparing two versions of a webpage or app against each other to determine which one performs better.

13. Mean Imputation is not a good practice because it does improve power, but the results will be biased. Mean imputation reduces the variance of the imputed variables. Mean imputation shrinks standard errors, which invalidates most hypothesis tests and the calculation of confidence interval. Mean imputation does not preserve relationships between variables such as correlations.

14. 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. It estimates the relationship between one independent variable and one dependent variable using a straight line.

15.The two major areas of statistics are known as descriptive statistics, which describes the properties of sample and population data, and inferential statistics, which uses those properties to test hypotheses and draw conclusions.