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
   Answer- 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
   Answer- 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
   Answer- 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
   Answer- d)

5. _____random variables are used to model rates.
   a) Empirical
   b) Binomial
   c) Poisson
   d) All of the mentioned
   Answer- c)

6. Usually replacing the standard error by its estimated value does change the CLT.
   a) True
   b) False
   Answer- False

7. Which of the following testing is concerned with making decisions using data?
   a) Probability
   b) Hypothesis
   c) Causal
   d) None of the mentioned
   Answer- b)

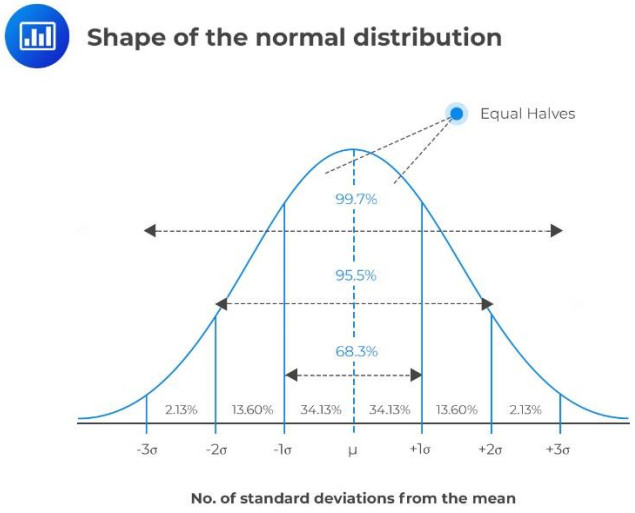
8. Normalized data are centered at_____and have units equal to standard deviations of the original data.

a) 0
b) 5
c) 1
d) 10
Answer- 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
   Answer- 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?
    Answer – Normal Distribution is the most widely known and used of all distribution. It is a term for proper bell curve.
    In this distribution, we have mean as zero and standard deviation as 1with zero skewness.
    For a normal distribution, 68% of the observations are within +/- one standard deviation of the mean, 95% are within +/- two standard deviations, and 99.7% are within +- three standard deviations.



11. How do you handle missing data? What imputation techniques do you recommend?
Answer – There are different ways to handle missing data:-
    1. Deleting Rows with missing values.
    2. Impute missing values
    3. Using Algorithms that support missing values.

Imputation is a technique used for replacing the missing data with some substitute value to retain most of the data/information of the dataset. These techniques are used because removing the data from the dataset is might lead to a reduction in the size of the dataset to a large extend, which can lead to incorrect analysis as it raises concerns for biasing the dataset..

Arbitrary Value Imputation: - This technique can handle both numerical and categorical variables and hence is the most important technique used in imputation. In this, we group the missing values in a column and assign them to a new value which is far away from the range of that column like 99999999 or -9999999 or "Missing" or "Not defined" for numerical & categorical variables.

When to use:-
a) Data is not Missing at Random(MAR)
b) Suitable for All

Frequent Category Imputation:- This technique says to replace the missing value with the variable with the highest frequency or in simple words replacing the values with the Mode of that column. This technique is also referred to as **Mode Imputation.**

When to use:-
c) Data is Missing at Random(MAR)
d) Missing data is not more than 5-6% of the dataset

12. What is A/B testing?
Answer- A/B testing is basically statistical hypothesis testing, also known as statistical inference. It is an analytical method for making decisions that estimates population parameters based on sample statistics.
The process is as follows:
    1. Make a null hypothesis and an alternative hypothesis.
    2. Start testing to gather statistical evidence to accept or reject the null hypothesis.
    3. With the help of final data, you can check whether your null hypothesis was correct, incorrect or inconclusive.

The null hypothesis states the default position to be tested i.e. the status quo. Whereas, the alternative hypothesis challenge the status quo (the null hypothesis). The alternative hypothesis is what you hope that you're A/B test might will prove to be true.

13. Is mean imputation of missing data acceptable practice?
Answer-
Mean imputation is a method in which the mean of the observed values for each variable is computed and then all the missing values for that variable are changed with this computed mean. However, this method can lead into severely biased estimates. If the number of missing values in a variable is large, and then we use this mean imputation method, then the resulting variance estimate can be severely underestimated for that variable.

14. What is linear regression in statistics?
Answer- It is a linear approach for showing the relationship between a dependent variable and one or more independent variables. If the independent or explanatory variable is one, it is called simple linear regression, for more than one, it is called multiple linear regression.
If we want to use a variable x to draw conclusions concerning another variable y, y is called dependent or response variable and x is an independent variable, predictor or explanatory variable. If the relationship between the variables is linear and can be summarized by a straight line.
The straight line can be described by an equation

$$Y=a+bx$$

Where a is called the intercept and b is the slope of the equation. The slope is the amount by which y increases when x is increased by 1.

15. What are the various branches of statistics?
Answer- There are four branches of statistics:-

1. Mathematical or theoretical statistics
   It helps in forming the experimental and statistical distribution.
   Mathematical statistics is the application of probability theory, a branch of mathematics, to statistics, as opposed to techniques for collecting statistical data.

2. Statistical methods or functions
   Statistical methods are mathematical formulas, models, and techniques that are used in statistical analysis of raw research data. The application of statistical methods extracts information from research data and provides different ways to assess the robustness of research outputs.
   It helps in the collection, tabulation and interpretation of the data. It helps in analyzing the data and return the insights from the data.

3. Descriptive statistics
   Descriptive statistics, in short, help describe and understand the features of a specific data set by giving short summaries about the sample and measures of the data. It helps in summarizing and organizing any data set characteristics. It also helps in the representation of data in both classification and diagrammatic way.

4. Inferential statistics
   Inferential statistics are used for hypothesis testing and include both parametric and nonparametric statistics such as ANOVA test. It helps in finding the conclusion regarding the population after analysis on a sample drawn from it.