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

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

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

WORKSHEET

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

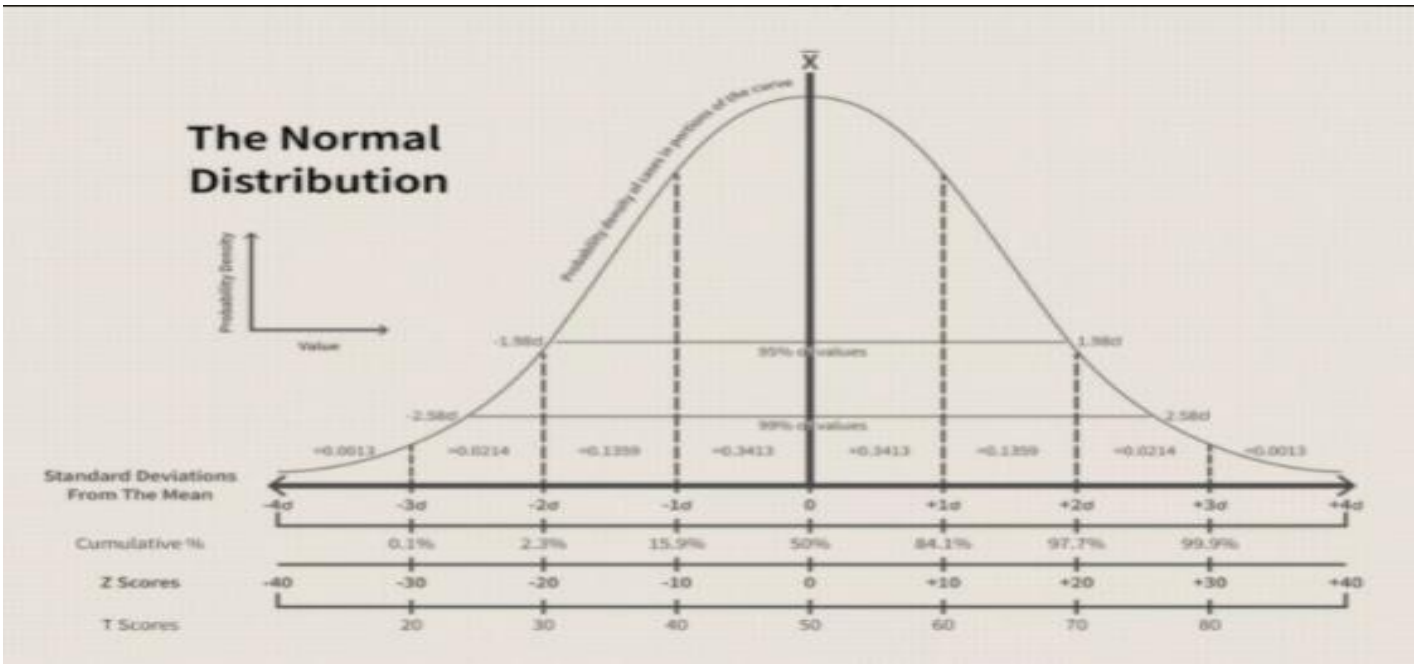
10. What do you understand by the term Normal Distribution?

Ans. A normal distribution is an arrangement of a data in which most of the data values are in the middle range and rest will be symmetric towards either extreme.

For example: Height is one of the simplest example of normal distribution. Most of the people are of average height, however very few people are having very short and very long height.

A graphical representation of a normal distribution is sometimes called a bell curve because of its flared shape. The precise shape can vary according to the distribution of the population but the peak is always in the middle and the curve is always symmetrical.

Its probability distribution is symmetric about the mean. This shows that the data near the mean are more frequent in occurrence than the data far from the mean. Its standard deviation depicts the bell curve's relative width around the mean.



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

Ans: When no value is stored for a variable (feature) in an observation is called Missing value. It could be represented as "?", "N/A",Nan, 0 or just a blank cell.

How to deal with missing data?

There are many ways to deal with missing values. Each situation is different so it should be judged differently. One of the possibility is just to remove the data if it doesn't effect your result much. When we drop the data we can drop the variable or the data entry. If we don't have lot of observations with missing data then in that case dropping the particular entry is best. Replacing data is another option since no data is wasted. However, it is less accurate we need to replace the data with value it should be.

Technique 1: Replace missing value by the average value i.e. mean of the entire variable.

Technique 2: If incase the value is not average i.e variable is categorical. So, in this case we try using mode of the most frequent data.

Mean imputation. Perhaps the easiest way to impute is to replace each missing value with the mean of the observed values for that variable.

12. What is A/B testing?

Ans. A/B testing is one of the components of the overarching process of conversion rate optimization (CRO) using which you can gather both qualitative and quantitative user insights. You can further use this collected data to understand user behavior, engagement rate, pain points, and even satisfaction with website features.

13. Is mean imputation of missing data acceptable practice?

Ans. . Perhaps mean imputation is the easiest way to impute is to replace each missing value with the mean of the observed values for that variable. It is consider as terrible practice since it ignores the feature correlation. Let us suppose

an example in which we are having a data of health score of the person age ranging between 15 to 80. If we are replace missing value of higher age group then it is replace by average score which is pretty much higher for a age group of 80 years.
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.

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. Simple linear regression uses 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).

15. What are the various branches of statistics?
Ans. Statistics is the branch of science which deals with the collection, classification and tabulation of numerical facts as the basis for explanations, description and comparison of phenomenon.

There are basically four division onto which statistics is divided.

a. **Mathematical or theoretical statistics**: It helps in forming the experimental and statistical distribution.
b. **Statistical methods or functions**: It helps in the collection, tabulation and interpretation of the data. It helps in analysing the data and returns insight from the data.
c. **Descriptive statistics:** It helps in summarizing and organizing any data set characteristics. It also helps in the representation of data in both classification and diagrammatic way.
d. **Inferential statistics:** It helps in finding the conclusion regarding the population after analytics on the sample drawn from it.