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.
   Ans: (a) True

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
   Ans: (a) Central Limit Theorem

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
   Ans: (b) Modeling bounded count data

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
   Ans: (d) All of the mentioned

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

6. 10. Usually replacing the standard error by its estimated value does change the CLT.
   a) True
   b) False
   Ans: (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
   Ans: (b) Hypothesis

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
   Ans: (a) 0

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
   Ans: (c) Outliers cannot conform to the regression relationship

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: Normal distribution, also known as the Gaussian distribution, is a probability distribution that is symmetric about the mean, showing that data near the mean are more frequent in occurrence than data far from the mean. In graph form, normal distribution will appear as a bell curve.

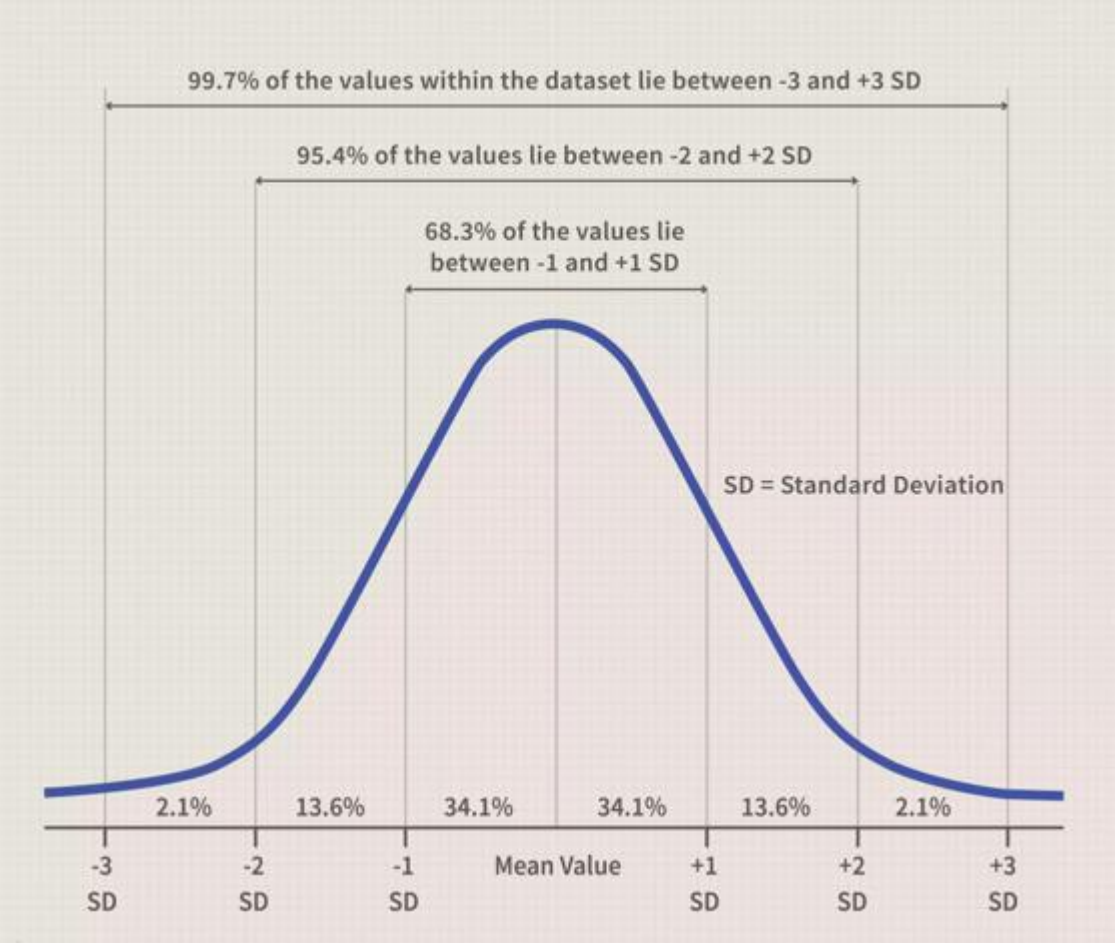

Fig: Normal Distribution

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

Ans: We can handle missing data using various techniques. Some of these imputation techniques are as follows:

- Deleting Rows/Columns: This method commonly used to handle the null values. Here, we either delete a particular row if it has a null value for a particular feature and a particular column if it has more than 70-75% of missing values. This method is advised only when there are enough samples in the data set.
- Replacing with Mean/Median/Mode.
- Assigning a Unique Category: We can replace unknown categorical variables with another unique category such as 'unknown' or 'U'.
- Predicting the Missing Values: Using the features which do not have missing values, we can predict the nulls with the help of a machine learning algorithm.
- Using Algorithms Which Support Missing Values: e.g. KNN – K Nearest Neighbor Algorithm.

12. What is A/B testing?

Ans: 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.

13. Is mean imputation of missing data acceptable practice?

Ans: We can fill in missing values with the mean of the variable over the time period of observation. Pros: Easy to compute and understand. Decent option if the variables are distributed normally. Cons: If our data has a trend (if the rolling-mean is increasing over time) our added values may make our charting look odd. Also, this is not acceptable if our variables have an odd distribution that makes the mean value meaningless.

14. What is linear regression in statistics?

Ans: In statistics, linear regression is a linear approach to modelling the relationship between a scalar response and one or more explanatory variables. The case of one explanatory variable is called simple linear regression; for more than one, the process is called multiple linear regression.

15. What are the various branches of statistics?

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

a) Descriptive Statistics: Descriptive statistics deals with the presentation and collection of data. This is usually the first part of a statistical analysis.
b) Inferential Statistics: Inferential statistics involves drawing the right conclusions from the statistical analysis that has been performed using descriptive statistics. In the end, it is the inferences that make studies important and this aspect is dealt with in inferential statistics.