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

Ans. A) True

2. Which of the following theorem states that the distribution of averages of iid variables, properlynormalized, 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- 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

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) Poission

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 theoriginal data.
   a) 0
   b) 5
   c) 1
   d) 10

Ans. A) 0

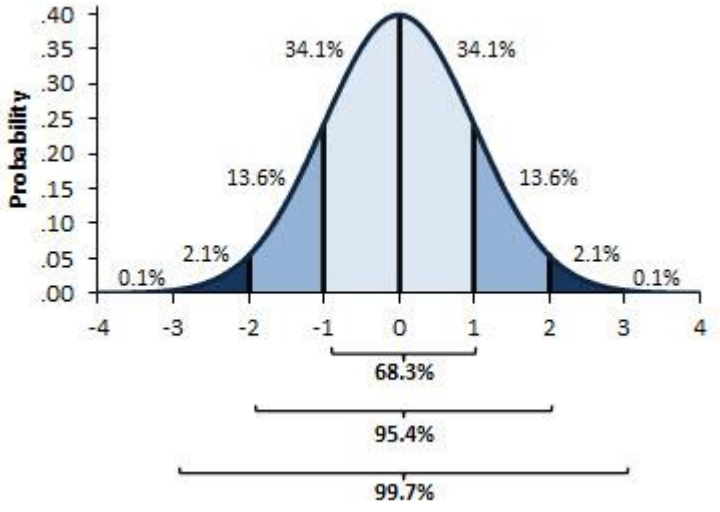
9. Which of the following statement is incorrect with respect to outliers?
   a) Outliers can have varying degrees of influence
   b) Outliers can be the result of spurious or real processes
   c) Outliers cannot conform to the regression relationship
   d) None of the mentioned

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 refers to 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. We can see it as a bell curve on a graph.



In this chart, the 0 is in middle of the x-axis, that's the mean of the data set. The data point in our dataset are distributed in a bell shaped curved that is centered at the mean, as we can see that our data are symmetrically distributed or mirrored each side of y-axis, we call this a normal distribution.

As we can notice at the bottom along the x-axis the positive numbers are moving to the right and negative numbers moving to the left. The distance between each number represents one standard deviation. According to Empirical Rule, It says when you have symmetrical distribution of the data, we can expect 68% of all our data points to be within one standard deviation So, 68% of all the data points are within negative one and the positive one on this graph.

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

Ans. There are several methods to handle missing data:

- **Delete the rows**
- **Replace with the most frequent values**
- **Apply classifier algorithm to predict**
- **Apply unsupervised machine learning**

First, we need to analyze and understand the nature of missing data, since it is critical in determining which method or treatment, we have to apply to impute the missing data. Data can be missing in various ways.

- **Missing Completely at Random: - When missing values are randomly distributed across all observations, then we consider the data to be missing completely random.**

- **Missing at Random: -The key difference between Missing completely at Random and Missing at Random (MAR) is that under MAR the data is not missing randomly across all observations but is missing randomly only with in sub-samples of data.**

- **Not Missing at Random: - When the missing data has a structure to it, we cannot treat it as missing at random.**

In first two cases, it is safe to remove the data with missing values depending upon their occurrences, while in the last one removing observations with missing values can produce a bias in the model. So, we have to be really careful before removing observations.

Note: - Imputation doesn't necessarily give better results.

12. What is A/B testing?

Ans. At a high-level A/B Testing is statistical way of comparing two or more version such as version A or version B, to determine only which version performs better but to also understand if it's difference between two version is statistically significant.

13. Is mean imputation of missing data acceptable practice?

Ans. It is a non-standard, but a fairly flexible imputation algorithm. It uses Random Forest 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.

14. What is linear regression in statistics?

Ans. Linear Regression allows us to model, mathematically, the relationship between two or more variables. For now, we will be working with just two variables: an independent variable and a dependent variable. The truth is, when we talk about how "good" a regression model is we are actually comparing it to another specific model.

15. What are the various branches of statistics?

Ans. The two main branches of statistics are:

- **Descriptive Statistics**
  - **It is a method of organizing, summarizing, and presenting data in an informative way.**
- **Inferential Statistics**
  - **It is a method which is used in determining something about a population on the basis of a sample.**
    - ✓ **Population: - The entire set of individuals or objects of interest or the measurements obtained from all individuals or objects of interest.**
    - ✓ **Sample: - A portion or part of the population of interest**