STATISTICS WORKSHEET-1(With Answers)

Q1 to Q9

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?

Ans- a) Central Limit Theorem

3. Which of the following is incorrect with respect to use of Poisson distribution?

Ans- b) Modeling bounded count data

4. Point out the correct statement.

Ans- d) All of the mentioned

5. _____ random variables are used to model rates.

Ans- c) Poisson

6. Usually replacing the standard error by its estimated value does change the CLT.

Ans- b) false

7. Which of the following testing is concerned with making decisions using data?

Ans- b) Hypothesis

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

Ans- a) 0

9. Which of the following statement is incorrect with respect to outliers?

Ans- c) Outliers cannot conform to the regression relationship

Q10and Q15

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

Ans- Normal Distribution is also called 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.

- Commonly seen as continuous distribution in nature with every event is independent from one another.

- In a normal distribution the mean is zero and the standard deviation is 1. It has zero skewness.

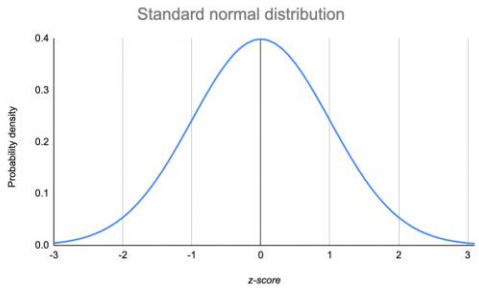
- A normal distribution is the proper term for a probability bell curve.

- The area under the curve is 1.

- Exactly half of the data are to the left side of the Centre/mean and other half to the right side of the center.

- Normal distributions are symmetrical, but not all symmetrical distributions are normal.

- Mean=Median=Mode



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

Ans- Missing data can be dealt with in a variety of ways. If the missing data are less or random, the most common reaction is to ignore it or to drop it. Choosing to make no decision, on the other hand, indicates that your statistical programme will make the decision for you.

But if the missing data are huge, it creates a huge difference in the programme if NaN values are dropped. In that case we use another strategy called imputation.

Imputation is the process of replacing missing values with substituted data. It is done as a preprocessing step.

There are so many methods of imputation.

1.SIMPLE IMPUTATION|: We can replace the missing values with the below methods depending on the data type

Mean and Median: If the data is numerical

Mode: if the data is categorical

2.SUBSTITUTION: pick a new subject and employ their worth instead.

3. REGRESSION IMPUTATION: The result of regressing the missing variable on other factors to get a predicted value. Instead of utilizing the mean, you're relying on the anticipated value, which is influenced by other factors.

4. MULTIPLE IMPUTATION

12.What is A/B testing?

Ans- A/B testing, also known as split testing is a user experience research methodology. If we have two options A and B and we don't know which option is performing better, we use this technique.

A/B testing is a randomized experimentation process to compare two versions of a single variable (web page, page element, etc.), typically by testing a subject's response to variant A against variant B, and determining which of the two variants is more effective and leaves the maximum impact.

In A/B testing, A refers to 'control' or the original testing variable. Whereas B refers to 'variation' or a new version of the original testing variable. It includes application of statistical hypothesis testing or "two-sample hypothesis testing" as used in the field of statistics.

13. Is mean imputation of missing data acceptable practice?

Ans- The process of replacing null values in a data collection with the data's mean is known as mean imputation.

True, imputing the mean preserves the mean of the observed data but it cannot be applied to all type of data. If the missing data is a string, we cannot use mean imputation. Also, if there are outliers, it ignores feature correlation. So, it is not a good practice.

14. What is linear regression in statistics?

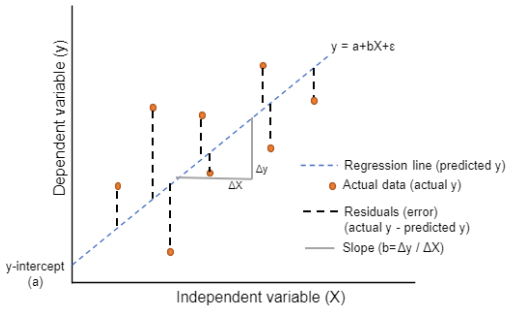
Ans- Linear regression is a statistical method that is used for predictive analysis. Linear regression makes predictions for continuous/real or numeric variables such as sales, salary, age, product price, etc.

Linear regression algorithm shows a linear relationship between a dependent (y) and one or more independent (y) variables, hence called as linear regression. Since linear regression shows the linear relationship, which means it finds how the value of the dependent variable is changing according to the value of the independent variable.

Mathematically, we can represent a simple linear regression as:

$$\mathbf{y= a+bx+ \varepsilon}$$

Y= Dependent Variable (Target Variable)
X= Independent Variable (predictor Variable)
a= intercept of the line
b = Linear regression coefficient.
ε = random error



The linear regression model provides a sloped straight line representing the relationship between the variables**.**

There are two types of linear regression-

a) Single linear regression- Where there is only on independent variable.

b) Multiple linear regression- Where there is more than one independent variables.

15. What are the various branches of statistics?

Ans- There are mainly two branches of statistics-

1)Descriptive Statistica

2)Inferential Statistics

1)Descriptive Statistics: The branch of statistics that focuses on collecting, summarizing, and presenting a set of data. Descriptive statistics can be categorized into

a) Measure of central tendency- Single value that attempts to describe a set of data by identifying the central position within that set of data. Measure of central tendency are **mean, median and mode.**

b) Measure of dispersion spread- It helps to understand the distribution of the data or the extent to which a numerical data is likely to vary about an average value. Measure of dispersion spread method includes **range, standard deviation, mean deviation, variance, percentile, quartile and quartile deviation, mean deviation.**

2)Inferential Statistics: Inferential Statistics means sampling the data of a population, utilizing the data from the sample and inferring the result to conclude and predict the behavior of a given population.

This technique is mainly used for data analysis, writing, and drawing conclusions from the limited data.

There are different types of inferential statistics which includes the following:

Regression analysis

Analysis of variance (ANOVA)

Analysis of covariance (ANCOVA)

Statistical significance (t-test)

Central limit theorem

Correlation analysis