STATISTICS WORKSHEET-1

1. **Bernoulli random variables take (only) the values 1 and 0**.
   Answer: - 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?**
   Answer: - a) Central Limit Theorem

3. **Which of the following is incorrect with respect to use of Poisson distribution?**
   Answer: - b) Modelling bounded count data.
   Poisson distribution use for modelling unbounded count data.

4. **Point out the correct statement**.
   Answer: - d) All of the mentioned.
   Many random variables, properly normalized, limit to a normal distribution.

5. **_____ random variables are used to model rates.**
   Answer: - C) Poisson.
   Poisson random variables are used to model count.

6. **Usually replacing the standard error by its estimated value does change the CLT**.
   Answer:-b) False
   Usually replacing the standard error by its estimated value doesn't change the CLT.

7. **Which of the following testing is concerned with making decisions using data?**
   Answer: - b) Hypothesis
   The null hypothesis is assumed true and statistical evidence is required to reject it in favour of a research or alternative hypothesis.

8. **Normalized data are cantered at_____and have units equal to standard deviations of the original data.**
   Answer: - a) 0
   In statistics and applications of statistics, normalization can have a range of meaning.

9. **Which of the following statement is incorrect with respect to outliers?**
   Answer:- c) Outliers cannot conform to the regression relationship.
   Outliers can conform to the regression relationship.

10. **What do you understand by the term Normal Distribution?**
    Answer:- 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.
    A normal distribution is an arrangement of a data set in which most values cluster in the middle of the range and the rest taper off symmetrically toward either extreme.

Height is one simple example of something that follows a normal distribution pattern: Most people are of average height, the numbers of people that are taller and shorter than average are fairly equal and a very small (and still roughly equivalent) number of people are either extremely tall or extremely short.
Here's an example of a normal distribution curve:
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. In a normal distribution, the mean, mode and median are all the same.
Normal distribution curves are sometimes designed with a histogram inside the curve. The graphs are commonly used in mathematics, statistics and corporate data analytics.

11. **How do you handle missing data? What imputation techniques do you recommend?**
    Answer:-Use deletion methods to eliminate missing data. The deletion methods only work for certain datasets where participants have missing fields.
    Use regression analysis to systematically eliminate data.
    Data scientists can use data imputation techniques.

12. **What is A/B testing?**
    Answer:- **A/B testing** is one of the most popular controlled experiments used to optimize web marketing strategies. It allows decision makers to choose the best design for a website by looking at the analytics results obtained with two possible alternatives A and B.
    A/B testing is about, let's consider two alternative designs: A and B. Visitors of a website are randomly served with one of the two. Then, data about their activity is collected by web analytics. Given this data, one can apply statistical tests to determine whether one of the two designs has better efficacy.
    A/B testing also called **split testing,** originated from the **randomized control trials** in Statistics, is one of the most popular ways for Businesses to test new UX features, new versions of a product, or an algorithm to decide whether your business should launch that new product/feature or not.
    Benefits of A/B testing:-:-
    Allows to learn what works and what doesn't in a quick manner
    You get feedback directly from actual/real product customers
    Since the users are not aware that they are being tested, the results will be unbiased.
    Demerits of A/B testing
    Presenting different content/price/features to different customers especially in the same geolocation might potentially be dangerous resulting in **Change Aversion** (we will discuss how this can be addressed later on)
    Requires a significant amount of Product, Engineering, and Data Science resources
    Might lead to wrong conclusions if not conducted properly.

13. **Is mean imputation of missing data acceptable practice?**
    Answer:- The process of replacing null values in a data collection with the data's mean is known as mean imputation.
    Mean imputation is typically considered terrible practice since it ignores feature correlation.
    Consider the following scenario: we have a table with age and fitness scores, and an eight-year-old has a missing fitness score. If we average the fitness scores of people between the

ages of 15 and 80, the eighty-year-old will appear to have a significantly greater fitness level than he actually does.

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?**

Answer: -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.

This form of analysis estimates the coefficients of the linear equation, involving one or more independent variables that best predict the value of the dependent variable.

Linear regression fits a straight line or surface that minimizes the discrepancies between predicted and actual output values. There are simple linear regression calculators that use 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).

Linear-regression models are relatively simple and provide an easy-to-interpret mathematical formula that can generate predictions. Linear regression can be applied to various areas in business and academic study.

linear regression is used in everything from biological, behavioral, environmental and social sciences to business. Linear-regression models have become a proven way to scientifically and reliably predict the future. Because linear regression is a long-established statistical procedure, the properties of linear-regression models are well understood and can be trained very quickly.

15. **What are the various branches of statistics?**

Answer: -