Statistics Worksheet – 01

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

4. Point out the correct statement.
Answer - (D) All of the mentioned

5. _____ random variables are used to model rates.
Answer - (C) Poisson

6. Usually replacing the standard error by its estimated value does change the CLT.
Answer - (B) False

7. Which of the following testing is concerned with making decisions using data?
Answer - (B) Hypothesis

8. Normalized data are centered at_____and have units equal to standard deviations of the original data.
Answer - (A) 0

9. Which of the following statement is incorrect with respect to outliers?
Answer - (C) Outliers cannot conform to the regression relationship

10. What do you understand by the term Normal Distribution?
Answer - The probability density function for a continuous random variable in a system defines the Normal Distribution. Let's assume that X is the random variable and that f(x) is the probability density function. In order to determine the probability of the random variable X, it specifies a function that is integrated across the range or the interval (x to x + dx) while taking the values between x and x+dx into account.

$f(x) \geq 0 \; \forall \; x \in (-\infty, +\infty)$ and $\int_{-\infty}^{+\infty} f(x) = 1$

Normal Distribution Formula - The probability density function of normal or gaussian distribution is given by;

$$f(x, \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

x is the variable
μ is the mean
σ is the standard deviation

11. How do you handle missing data? What imputation techniques do you recommend?
Answer - There are many approaches to handle missing data. The most typical response is to disregard it. On the other hand, choosing to make no decision means that the statistical software will decide. Most of the time, the programme will remove items in a listwise order. Listwise deletion may or may not be a wise choice, depending on why and how much data is lost.

Imputation is another tactic that is used frequently. Imputation involves replacing missing values with an estimate and analysing the complete set of data. The most popular techniques include mean imputation, hot deck imputation,

substitution, cold deck imputation, regression imputation, stochastic regression imputation, interpolation and extrapolation.

12. What is A/B testing?
Answer - A/B testing is an elementary randomised control experiment. It is a method for contrasting two variations of a variable to see which performs better in a regulated setting.

One of the most well-known and often employed statistical tools is A/B testing.

Making a claim is how the A/B testing process is initiated (hypothesis). The test is used to gather the statistical data to support or refute the hypothesis. The hypothesis' final results let us know if it was accurate, false, or inconclusive.

13. Is mean imputation of missing data acceptable practice?
Answer - Mean imputation is the process of replacing null values in a data collection with the mean of the data. Mean imputation is frequently seen as a bad approach since it disregards feature correlation.

Mean imputation increases bias while reducing the variance of our data. The model is less accurate and the confidence interval is smaller as a result of the lower variance.

14. What is linear regression in statistics?
Answer - A fundamental and widely used form of predictive analysis is linear regression. The link between one dependent variable and one or more independent variables is explained using the regression estimations.

The formula y = c + b*x,
Where,
y is the estimated score of the dependent variable
c is a constant, b is the regression coefficient
x is the score on the independent variable, defines the simplest form of the regression equation with one dependent and one independent variable.

The dependant variable in a regression has many different names. It can be referred to as an endogenous variable, criteria variable, or outcome variable. The independent variables is referred to as predictor variables, regressors, or exogenous variables. Regression analysis has three main applications such as assessing predictor power, predicting an effect, and predicting trends.

15. What are the various branches of statistics?
Answer - Data collection, descriptive statistics, and inferential statistics are the three main subfields of statistics.

A) Data Collection - The process of gathering, measuring, and analysing precise insights for research using accepted, established methods is known as data collection. A researcher can assess their hypothesis using the data that they have gathered. Regardless of the subject of study, gathering data is typically the first and most crucial phase in the research process. Depending on the type of data needed, different disciplines of research require different approaches to data gathering.

B) Descriptive Statistics - The properties of a data set are collected and presented using descriptive statistics. A data set is a compilation of observations or responses from a sample of a population or the complete population. The initial stage in statistical analysis in quantitative research is to define the features of the data, such as the average of one variable (for example, height) or the relationship between two variables (e.g., height and gender)

C) Inferential Statistics - Inferential statistics are the next step, which aid in determining whether your data supports or contradicts your hypothesis and whether it can be applied to a broader population.
Two purposes dominate the use of inferential statistics are estimating population numbers and testing hypotheses to make population-level judgments.