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

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?

a) Central Limit Theorem b) Central Mean Theorem c) Centroid Limit Theorem d) All of the mentioned

Answer: 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

Answer: 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

Answer: d) All of the mentioned

5. _____ random variables are used to model rates.

a) Empirical b) Binomial c) Poisson d) All of the mentioned

Answer: c) Poisson

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

a) True b) False

Answer : a) True

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

a) Probability b) Hypothesis c) Causal d) None of the mentioned

Answer: 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

Answer: 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 WORKSHEET

Q10and Q15 are subjective answer type questions, Answer them in your own words briefly.

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

Answer: -Normal distribution is a probability distribution that is symmetric about the mean, showing that the data near the mean are more frequent in occurrence than data far from the mean. In graph form the Normal distribution will look like a bell curve.

-The four characteristics of Normal distribution are Symmetric, Unimodal, Asymptotic and the mean .

-It can be used to determine the proration of values that fall within a specified number of standard derivations from the mean

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

Answer: Popular strategies to handle missing data are as follows:

a) Deleting rows with missing values.

b) Impute missing values for continuous and categorical variables.

c) Using algorithms that support missing values.

d) Prediction of missing values.

- Imputation is a technique to replace missing data with another substitute value to retain most of the data/information in the dataset.

The most popular widely used imputation techniques I would recommend would be KNN (K Nearest Neighbour) method for the first model to check the accuracy else also try with PCA(Principal Component Analysis).But in case of complex missing data and values I would recommend Random Forests.

12. What is A/B testing?

Answer: A/B testing is basically statistical hypothesis testing or also called statistical inference. It is an analytical method for making decisions that estimates population parameters based on sample statistics. It starts by making a claim(hypothesis).

-Here we model the metric for each variant as a random variable with some probability distribution, by calculating the posterior distribution for each variant, we can express the uncertainty of our beliefs with probability statements.

13. Is mean imputation of missing data acceptable practice?

Answer: True, imputing the mean preserves the mean of the observed data. So if the data are missing completely at random, the estimate of the mean remains unbiased Since most research studies are interested in the relationship among variables, mean imputation is not a good solution.

Mean imputation reduces the variance of the imputed variables. ... Mean imputation shrinks standard errors, which invalidates most hypothesis

tests and the calculation of confidence interval. Mean imputation does not preserve relationships between variables such as correlations.

14. What is linear regression in statistics?

-Linear regression is a linear approach for modelling the relationship between a scalar response and one or more explanatory variables, also known as dependent and independent variables.

-Linear Regression analysis allows to determine the strength of relationships between variables using statistical measurements like R squared, adjusted R squared regression analysis which tell the total variability of data in the model.

The types of Linear Regression are as follows Multiple, Logistic, Ordinal, Multinomial, Discriminant

15. What are the various branches of statistics?

Answer: There are three main real branches in statistics they are

-Data collection:-

Data collection is described as the procedure of collecting, measuring analysing accurate insights for research using validated techniques. The four types of data collection include Observational, Experimental, Simulation, and , derived

-Observational:- Data collected by direct observation without any alteration or processing and originates from a valid source .

-Experimental:-Data experimentation is described as the data which results from testing, measurement and experimental methods.it is used by the researcher to produce and measure change or create difference when variable is altered.

The 3 types of experimental data are pre-experimental, quasi experimental, and , true experimental data.

#Descriptive Statistics:-

Descriptive Statistics summarizes or describes the characteristics of the data set. Descriptive statistics consists of two basic categories of measures: measures of central tendency which include the mean, median, and, mode and measures of variability (or spread) wherein the measures of variability or spread describe the dispersion of data within the set which include variance, standard deviation, minimum and maximum variables

Inferential Statistics:- It is described as when a random sample of dataset or data taken from a population to describe or make inferences about the dataset or population. The 3 main types of Inferential statistics are Hypothesis tests, Confidence Intervals, and Regression Analysis.

-Hypothesis Tests: It is used to access the plausibility of a hypothesis data given a sample data.

-Confidence Intervals: It describes the probability a parameter will fall between a pair of values that falls around the mean. Confidence levels measures the degree of certainty or uncertainty in a given sampling method. The confidence intervals are constructed mostly using 95% or 99% levels.

-Regression Analysis: It is described as a set of statistical methods for estimation of the relationship between dependent and one or more independent variables. It can be used to assess the strength of the variables and for modelling the future relationship between them. The different types of Regression are as follows Linear Regression, logistic Regression, Ridge Regression, Lasso Regression, Polynomial Regression and Bayesian Linear Regression.