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

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

b) False

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

Ans = b) Hypothesis

8. 4. 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 are subjective answer type questions, Answer them in your own words briefly

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

Ans = A Normal distribution is the proper terms of a probability bell curve . in a normal distribution the mean is zero and the standard deviation is 1) Skewed Right 2) Normal distribution 3)Skewed Left . Normal distribution is also knows 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 then data far from then mean. In graphical form the normal distribution appears as a bell curve the normal distribution is the proper term for a probability bell curve in a normal distribution is a the mean is zero and and the standard deviation one . normal distribution are symmetrical but not all symmetrical distribution are normal . many naturally occurring phenomena tend to approximate the normal distribution. In finance most pricing distribution are not however perfectly normal. A normal distribution is most common type of distribution assumed in technical stock market analysis and in other type of statistical analyses.the standard normal distribution has two parameters the mean of the standard deviation first its mean average ,median is midpoint , and mode most frequent observation are all equal to one another moreover these values are represent the peck or highest point of the distribution . the distribution are fall symmetrically around the mean the width of this is define by the standard deviation

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. I believe the most common reaction is to ignore it. Choosing to make no decision, on the other hand, indicates that your statistical programme will make the decision for you. Your application will remove things in a listwise sequence most of the time. Depending on why and how much data is gone, listwise deletion may or may not be a good idea. Another common strategy among those who pay attention is imputation. Imputation is the process of substituting an estimate for missing values and analysing the entire data set as if the imputed values were the true observed values. Single and multiple imputation are the two forms of imputation. When people say imputation, they usually mean single. The term single refers to the fact that you only use one of the seven methods to estimate the missing number outlined above. When listwise deletion eliminates a considerable amount of the data set, single imputation appears to be a tempting option. It does, however, have certain restrictions. Unless the data is Missing Completely at Random, certain imputation processes, such as means, correlations, and regression coefficients, result in skewed parameter estimations. The bias is frequently worse than with listwise deletion, which is most software's default The level of the bias is determined by a number of factors, including the imputation technique, the missing data mechanism, the fraction of missing data, and the information in the data set.

12. What is A/B testing?

Ans = A/B testing , also know as split testing , refer to a randomized experimentation wherein two or more version of a variable web page , page element, etc are shows to different segments of website visitors at the same time to determine which version leaves the maximum impute and drives business metrics. An AB test is an example of statistical hypothesis testing a process whereby a hypothesis is mode about the relationship between two data sets each those data sets are then compared against each other to determine if there is a statistically significant relationship or not. Am A/B test is used to determine which version or variant of something will perform more effectively is the market the strategy is commonly used by marketing and advertising professionals who show multiple version of an ad marketing users and then the results.

13. Is mean imputation of missing data acceptable practice?

Ans = Mean imputation So simple And yet, so dangerous that’s a bit dramatic, but mean imputation (also called mean substitution) really ought to be a last resort.It’s a popular solution to missing data, despite its drawbacks. Mainly because it’s easy. It can be really painful to lose a large part of the sample you so carefully collected, only to have little power But that doesn’t make it a good solution, and it may not help you find relationships with strong parameter estimates. Even if they exist in the population. On the other hand, there are that provide much more accurate estimates and standard errors, so there really is no excuse to use it. This post is the first explaining the many reasons not to use mean imputation (and to be fair, its advantages).First, a definition: mean imputation is the replacement of a missing observation with the mean of the non-missing observations for that variable. This is the original logic involved in mean imputation. If all you are doing is estimating mean(which is rarely the point of research studies), and if the data are missing completely at random, mean imputation will not bias your parameter estim ate.It *will* still bias your standard error, but I will get to that in .Since most research studies are interested in the relationship among variables, mean imputation is not a good solution. The following graph illustrates this well Although imputing missing values by using the mean is a popular imputation technique, there are serious problems with mean imputation. The variance of a mean-imputed variable is always biased downward from the variance of the un-imputed variable.

14. What is linear regression in statistics?

Ans = In statistics linear regression is a linear approach for modelling the relationship between a scalar response and one or more explanatory variable also knows as dependent and independent variable the case of one explanatory variable is called simple linear regression for more than one the process is called multiple linear is this page helpful linear regression is used to predict the relationship between two variable by applying a linear equation to observed data there are two types of variable one variable ic called an independent variable and the other dependent variable linear regression is commonly used for predictive analysis

15. What are the various branches of statistics?

- Ans = Descriptive statistics :- Descriptive statistics are brief informational coefficients that summarize a given data set, which can be either a representation of the entire population or a sample of a population. Descriptive statistics are broken down into measures of central tendency and measures of variability (spread). Measures of central tendency include the mean, median, and mode, while measures of variability include standard deviation, variance, minimum and maximum variables, Descriptive statistics summarizes or describes the characteristics of a data set. Descriptive statistics consists of three basic categories of measures: measures of central tendency, measures of variability (or spread), and frequency distribution. Measures of central tendency describe the center of the data set (mean, median, mode). Measures of variability describe the dispersion of the data set (variance, standard deviation). Measures of frequency distribution describe the occurrence of data within the data set (count).

Inferential statistics :- Inferential statistics is a branch of statistics that makes the use of various analytical tools to draw inferences about the population data from sample data. Apart from inferential statistics, descriptive statistics forms another branch of statistics. Inferential statistics help to draw conclusions about the population while descriptive statistics summarizes the features of the data set. There are two main types of inferential statistics - hypothesis testing and regression analysis. The samples chosen in inferential statistics need to be representative of the entire population. In this article, we will learn more about inferential statistics, its types, examples, and see the important formulas. Inferential statistics helps to develop a good understanding of the population data by analyzing the samples obtained from it. It helps in making generalizations about the population by using various analytical tests and tools. In order to pick out random samples that will represent the population accurately many sampling techniques are used. Some of the important methods are simple random sampling, stratified sampling, cluster sampling, and systematic sampling techniques.