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
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
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
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
5. _____ random variables are used to model rates.
   a) Empirical
   b) Binomial
   c) Poisson
   d) All of the mentioned
6. 10. Usually replacing the standard error by its estimated value does change the CLT.
   a) True
   b) False
7. 1. Which of the following testing is concerned with making decisions using data?
   a) Probability
   b) Hypothesis
   c) Causal
   d) None of the mentioned
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
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

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

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

Normal distribution also known as the Gaussian distribution. The normal distribution is a core concept in statistics. It helps performing exploratory data analysis. It can simplify the normal distributions probability density by using only two parameters: μ mean and σ2. In Simple Language, the normal distribution is an arrangement of a data set which values cluster in the middle of the range and the rest taper off symmetrically towards either end. A graphical representation of a normal distribution is sometimes called a bell curve because of its flared shape.

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

Data can be missing in the following ways. Missing completely at random(MCAR) when missing values are randomly distributed across all observations, then we consider the data to be missing completely at random. So there are two types of observations are done, first with missing observations and the other without missing observations, if we find any difference between the two observations, so it is assumed that the data is MCAR.
Second-Missing at random(MAR) There is little difference between MCAR and MAR is that under MAR the data is not missing randomly across observations but missing randomly in sub samples.
Third-not missing at random(NMAR).when missing data has a structure of its own, its not consider at missing at random.

For imputations techniques
Complete case analysis(CCA)- this is straightforward method of dealing with missing data, which directly removes rows that have missing data. This method is also known as listwise deletion. It has its pros and cons. Advantages- it is easy to implement and no data manipulation is required. Limitation- Deleted data can be informative. Can lead to the delete of large part of data. The production model will not know what to do with missing data.
Arbitrary value imputation- This technique can handle both numerical and categorical variables.This technique states that we can group the missing values in a column and assign them to a new value that is far away from the range of that column. It has its pros and cons. Advantage- easy to implement and we can use it in production. It retains the importance of missing values, if there are any. Disadvantage- can distort original variable distribution. Arbitrary values can create outliers and need extra caution while selecting arbitrary value.
Frequent category Imputation- This technique is to replacing the values with the mode of that column. This technique is also referred to as Mode imputation. It also have its pros and cons. Advantage- implementation is easy, we can obtain a complete data set in very little time.we can use this technique in production model. Disadvantage- The higher the percentage of missing values, the higher will be the distortion. Can distort original variable distribution.

12. What is A/B testing?
A/B testing is a basic randomized control experiment. It's a way to compare the two versions of a variable to find out which performs better in a controlled environment. It's a hypothetical testing methodology for making decision that estimate population parameters based on simple statistics.

13. Is mean imputation of missing data acceptable practice?
Mean imputation of missing data is a acceptable practice. Imputing the mean preserves the mean of the observerd data. So if the data are missing completely at random, the estimate of the mean remains unbiased and you are able to keep your sample size up to the full size. And it if you doing is estimating means,and if the data are missing completely at random, mean imputation will not bias your parameter estimate.

14. What is linear regression in statistics?
Linear regression is a basic and commonly used type of predictive analysis. It does a set of predictor variables do a good job in predicting the outcome variables. And which variable is particular are significant predictors of the outcome variable, in what way do they indicate the impact the outcome variable. These regression estimates are explain the

relationship between one dependent variable and one or more independent variable. there are many name for a regression's dependent variable. It may be called an outcome variable, criterion variable ,endogenous variables. The independent variables can be called exogenous variables, predictor variables. The three major uses for regression analysis are – determining the strength of predictors. – forecasting an effect.-trend forecasting.

15. What are the various branches of statistics?

There are 5 branches of statistics.
1-Data collection.- its all about how the data is collected. Values are simply collected together to make a data set. If we are collecting aa data, we have to be careful where we get it from. There are issues in the collection of data, need to make sure that the data has been collected fairly before you go on a deal with it.the words population and sample are used in general in statistics. The population is the entire set of data and the sample is a subset of the population.
2-Descriptive statistics- it's a part of statistics that deals with presenting the data we have. It has two basic forms. Presenting data either visually or numerically. Visual techniques can be used in graphs, bar charts and pie charts. Numerical techniques are such as averages and spreads. Its main aim is to present the data in an understandable way.
3-Inferential Statistics – its aspect that deals with making conclusions about the data. Its simple language. Take the data you have and make an inference or conclusion from it
4-Discrete and continuous data-This data comes in two distinct types. Discrete data can take the distinct values which can be clearly identified and separated and continuous data can take any value.
5-Frequency distributions- the actual collection of data is not very meaningful and we have to put the data into categories defining the degrees in classifications by standardization. Instead of allocation the data everywhere, we are pacing the data in appropriate category to get more concise view of the results.