STATISTICS WORKSHEET -1

Q1 . Bernoulli random variables take (only) the values 1 and 0.

ANS:- True

Q2. 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:- Central Limit Theorem

Q3:- Which of the following is incorrect with respect to use of Poisson distribution?

ANS:- Modeling bounded count data

Q4:- Point out the correct statement.

ANS:- All of the mentioned

Q5:- ___ random variables are used to model rates.

ANS:- Poisson

Q6:- Usually replacing the standard error by its estimated value does change the CLT.

ANS:-False

Q7:- Which of the following testing is concerned with making decisions using data?

ANS:- Hypothesis

Q8:- . Normalized data are centered at_____and have units equal to standard deviations of the original data.

ANS:- 0

Q9:- Which of the following statement is incorrect with respect to outliers?

ANS:- Outliers cannot conform to the regression relationship

Q10:- . What do you understand by the term Normal Distribution?

ANS:- 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.In probability and statistics, the normal distribution or Gaussian distribution or bell curve is one of the most important continuous probability distributions.

Normal distributions are symmetric, unimodal and the mean, median, and mode are all equal. A normal distribution is perfectly symmetrical around its center. That is, the right side of the center is a mirror image of the left side. There is also only one mode, or peak, in a normal distribution.

Q11:- How do you handle missing data? What imputation techniques do you recommend?

ANS:- The problem of missing value is quite common in many real-life datasets. Missing value can bias the results of the machine learning models and/or reduce the accuracy of the model.

Missing data appear when no value is available in one or more variables of an individual.

We can simply handle the missing data or values by these ways:-

- Deletions. Pair-wise Deletion. List-wise Deletion/ Dropping rows. Dropping complete columns.
- Basic Imputation Techniques. Imputation with a constant value. Imputation using the statistics (mean, median, mode)
- K-Nearest Neighbor Imputation.

We can Imputing the Missing Value by the following ways:-

- Replacing With Arbitrary Value.
- Replacing With Mode.
- Replacing With Median.
- Replacing with previous value – Forward fill.
- Replacing with next value – Backward fill.
- Interpolation.
- Impute the Most Frequent Value.

Q12:- . What is A/B testing?

ANS:- A/B testing (also known as bucket testing or split-run testing) is a user experience research methodology. A/B tests consist of a randomized experiment with two variants, A and B. It includes application of statistical hypothesis testing or "two-sample hypothesis testing" as used in the field of statistics. A/B testing is a way to compare two versions of a single variable typically by testing a subject's response to variant A against variant B, and determining which of the two variants is more effective.

A/B testing is a shorthand for a simple randomized controlled experiment, in which two samples (A and B) of a single vector-variable are compared. These values are similar except for one variation which might affect a user's behavior. A/B tests are widely considered the simplest form of controlled experiment. However, by adding more variants to the test, its complexity grows.

"Two-sample hypothesis tests" are appropriate for comparing the two samples where the samples are divided by the two control cases in the experiment. Z-test are appropriate for comparing means

under stringent conditions regarding normality and a known standard deviation. Student's t-test are appropriate for comparing means under relaxed conditions when less is assumed.

PROS AND CONS OF A/B TEST:-

When conducting A/B testing, the user should evaluate the pros and cons of it to see if it aligns best with the results that they're hoping for.

Pros: Through A/B testing, it's easy to get a clear idea of what users prefer, since it's directly testing one thing over the other. It's based on real user behavior so the data can be very helpful especially when determining what works better between two options. In addition, it can also provide answers to very specific design questions. One example of this is Google's A/B testing with hyperlink colors. In order to optimize revenue, they tested dozens of different hyperlink hues to see which color the users tend to click more on.

Cons: However, there are a couple cons to A/B testing. Like mentioned above, A/B testing is good for specific design questions but it can also be a downside since it's mostly only good for specific design problems with very measurable outcomes. It could also be a very costly and timely process. Depending on the size of the company and/or team, there could be a lot of meetings and discussions about what exactly to test and what the impact of the A/B test is. If there's not a significant impact, it could end up as a waste of time and resources.

Q13:- Is mean imputation of missing data acceptable practice?

ANS:- Outliers data points will have a significant impact on the mean and hence, in such cases, **it is not recommended to use the mean for replacing the missing values**. Using mean values for replacing missing values may not create a great model and hence gets ruled out.Mean imputation: So simple. And yet, so dangerous.

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.

On the other hand, there are many alternatives to mean imputation that provide much more accurate estimates and standard errors, so there really is no excuse to use it.

reason is applies to any type of single imputation. Any statistic that uses the imputed data will have a standard error that's too low.

There are some methods of imputation other than mean imputation:-

- Substitution. ...
- Hot deck imputation. ...
- Cold deck imputation. ...
- Regression imputation. ...
- Stochastic regression imputation. ...
- Interpolation and extrapolation.

Q14:- What is linear regression in statistics?

ANS:- 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.

Linear regression is a kind of statistical analysis that attempts to show a relationship between two variables

There are two kinds of Linear Regression Model:-

- Simple Linear Regression: A linear regression model with one independent and one dependent variable.
- Multiple Linear Regression: A linear regression model with more than one independent variable and one dependent variable

SIMPLE REGRESSION

Simple linear regression is a type of regression analysis where the number of independent variables is one and there is a linear relationship between the independent(x) and dependent(y) variable.

MULTIPLE LINEAR REGRESSION

Multiple regression is a statistical technique that can be used to analyze the relationship between a single dependent variable and several independent variables. The objective of multiple regression analysis is to use the independent variables whose values are known to predict the value of the single dependent value. Multiple regression works by considering the values of the available multiple independent variables and predicting the value of one dependent variable. Example: A researcher decides to study students' performance from a school over a period of time.

Q15:- What are the various branches of statistics?

ANS:- Statistics is a study of presentation, analysis, collection, interpretation and organization of data.

Statistics is the discipline that concerns the collection, organization, analysis, interpretation, and presentation of data. In applying statistics to a scientific, industrial, or social problem, it is conventional to begin with a statistical population or a statistical model to be studied.

There are **two main branches** of statistics
- Inferential Statistic.
- Descriptive Statistic.

Inferential Statistics:
Inferential statistics used to make inference and describe about the population. These stats

are more useful when it's not easy or possible to examine each member of the population. Inferential statistics describe the many ways in which statistics derived from observations on samples from study populations can be used to deduce whether or not those populations are truly different.

Descriptive Statistics:

Descriptive statistics are use to get a brief summary of data. You can have the summary of data in numerical or graphycal form. A descriptive statistic is a summary statistic that quantitatively describes or summarizes features from a collection of information, while descriptive statistics is the process of using and analysing those statistics.