STATISTICS WORKSHEET-1

1. Bernoulli random variables take (only) the values 1 and 0.
   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
3. Which of the following is incorrect with respect to use of Poisson distribution?
   b) Modeling bounded count data
4. Point out the correct statement
   d) All of the mentioned
5. _____ random variables are used to model rates.
   c) Poisson
6. Usually replacing the standard error by its estimated value does change the CLT
   b) False
7. Which of the following testing is concerned with making decisions using data?
   b) Hypothesis
8. Normalized data are centered at_____and have units equal to standard deviations of the original data
   a) 0
9. Which of the following statement is incorrect with respect to outliers?
   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?

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 a normal distribution the mean is zero and the standard deviation is 1. It has zero skew and a kurtosis of 3. Normal distributions are symmetrical, but not all symmetrical distributions are normal. The normal distribution has two parameters, the mean and standard deviation. The Gaussian distribution does not have just one form. Instead, the shape changes based on the parameter values.

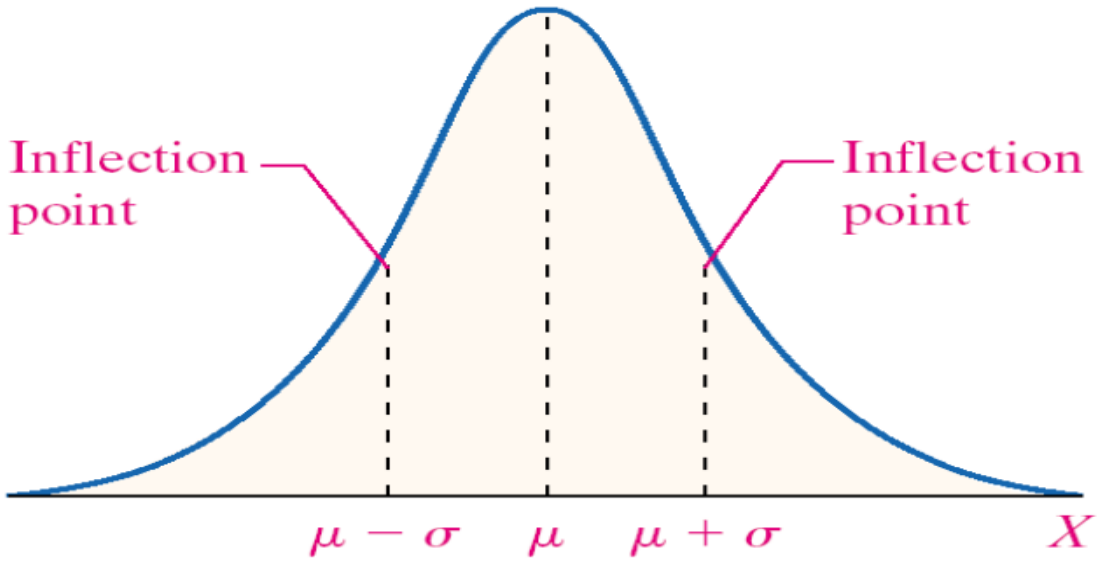
Mean:

The mean is the central tendency of the normal distribution. It defines the location of the peak for the bell curve. Most values cluster around the mean. On a graph, changing the mean shifts the entire curve left or right on the X-axis.

Standard deviation:

The standard deviation is a measure of variability. It defines the width of the normal distribution. The standard deviation determines how far away from the mean the values tend to fall. It represents the typical distance between the observations and the average.

On a graph, changing the standard deviation either tightens or spreads out the width of the distribution along the X-axis. Larger standard deviations produce wider distributions.



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

When dealing with missing data, we can use two primary methods to solve the error: imputation or the removal of data.

The imputation method develops reasonable guesses for missing data. It's most useful when the percentage of missing data is low. If the portion of missing data is too high, the results lack natural variation that could result in an effective model.

The other option is to remove data. When dealing with data that is missing at random, related data can be deleted to reduce bias. Removing data may not be the best option if there are not enough observations to result in a reliable analysis. In some situations, observation of specific events or factors may be required

There are various imputation techniques which can help with missing data:

Mean, Median and Mode:

This is one of the most common methods of imputing values when dealing with missing data. In cases where there are a small number of missing observations, we can calculate the mean or median of the existing observations. However, when there are many missing variables, mean or median results can result in a loss of variation in the data.

K Nearest Neighbors :

In this method, data scientists choose a distance measure for k neighbors, and the average is used to impute an estimate. We must select the number of nearest neighbors and the distance metric. KNN can identify the most frequent value among the neighbors and the mean among the nearest neighbors.

Random Forest:

Random forest is a non-parametric imputation method applicable to various variable types that works well with both data missing at random and not missing at random. Random forest uses multiple decision trees to estimate missing values and outputs OOB (out of bag) imputation error estimates.

Dropping Variables:

If data is missing for more than 60% of the observations, it may be wise to discard it if the variable is insignificant.

12. What is A/B testing?

A/B testing is a basic randomized control experiment. It is a way to compare the two versions of a variable to find out which performs better in a controlled environment.

For example, there is an e-commerce company XYZ. It wants to make some changes in its newsletter format to increase the traffic on its website. It takes the original newsletter and marks it A and makes some changes in the language of A and calls it B. Both newsletters are otherwise the same in color, headlines, and format.

We will use A/B testing and collect data to analyze which newsletter performs better. In hypothesis testing, we have to make two hypotheses i.e Null hypothesis and the alternative hypothesis.
Null hypothesis or H0:
Alternative Hypothesis or $H_{a:}$

Here, Ha is- "the conversion rate of newsletter B is higher than those who receive newsletter A". There are two types of errors that may occur in our hypothesis testing:

Type I error: We reject the null hypothesis when it is true. That is we accept the variant B when it is not performing better than A

Type II error: We failed to reject the null hypothesis when it is false. It means we conclude variant B is not good when it performs better than A.

The rejection of the hypothesis is based on the P-value. It is the probability that the difference between the two values is just because of random chance. P-value is evidence against the null hypothesis. The smaller the P-value stronger the chances to reject the H0. For the significance level of 0.05, if the p-value is lesser than it, we can reject the null hypothesis.

In this case, the p-value is less than the significance level i.e 0.05. Therefore, we can reject the null hypothesis. This means that in our A/B testing, newsletter B is performing better than newsletter A.

13. Is mean imputation of missing data acceptable practice?

The process of replacing null values in a data collection with the data's mean is known as mean imputation.

Mean imputation is mostly not acceptable as it ignores the feature correlation. For example, if we have a table which says if the person has diabetes or no based on their regular checkups, now suppose a few of them missed their regular check ups, and it has the missing checkup values. If we average the medical reports, a person who doesn't have diabetes would be considered diabetic.

Second, mean imputation decreases the variance of our data while increasing bias. As a result of the reduced variance, the model is less accurate and the confidence interval is narrower.

14. What is linear regression in statistics?

In a cause and effect relationship, the independent variable is the cause, and the dependent variable is the effect. Least squares linear regression is a method for predicting the value of a dependent variable Y, based on the value of an independent variable X**.**

The Least Squares Regression Line:

Linear regression finds the straight line, called the least squares regression line or, that best represents observations in a bivariate data set. Suppose Y is a dependent variable, and X is an independent variable. The population regression line is:

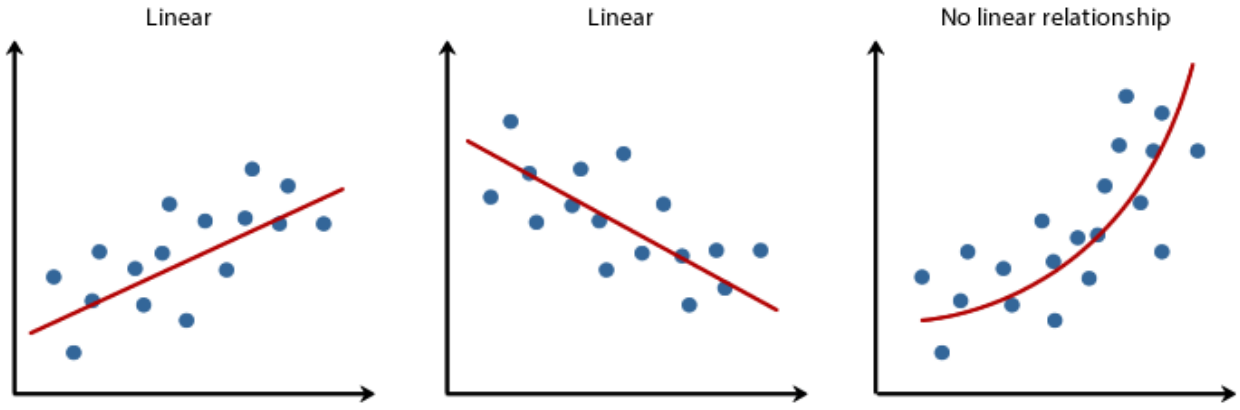
$$Y = B0 + B1X$$

where B0 is a constant, B1 is the regression coefficient, X is the value of the independent variable, and Y is the value of the dependent variable.

Given a random sample of observations, the population regression line is estimated by a sample regression line. The sample regression line is:

$$\hat{y} = b0 + b1x$$

where b0 is a constant, b1 is the regression coefficient, x is the value of the independent variable, and $\hat{y}$ is the predicted value of the dependent variable**.**



15. What are the various branches of statistics?

Statistics is a branch of applied mathematics that involves the collection, description, analysis, and inference of conclusions from quantitative data. Statistics is the study and manipulation of data, including ways to gather, review, analyze, and draw conclusions from data.

The two major areas of statistics are known as descriptive statistics, which describes the properties of sample and population data, and inferential statistics, which uses those properties to test hypotheses and draw conclusions.

Descriptive Statistics:

Descriptive statistics mostly focus on the central tendency, variability, and distribution of sample data. Central tendency means the estimate of the characteristics, sample or population, and includes descriptive statistics such as mean, median, and mode. Variability refers to a set of statistics that show how much difference there is among the elements of a sample or population along the characteristics measured, and includes metrics such as range, variance, and standard deviation.

The distribution refers to the overall "shape" of the data, which can be depicted on a chart such as a histogram or dot plot, and includes properties such as the probability distribution function, skewness, and kurtosis. Descriptive statistics can also describe differences between observed characteristics of the elements of a data set.

Inferential Statistics:

It is used to draw conclusions about the characteristics of a population, drawn from the characteristics of a sample, and to decide how certain they can be of the reliability of those conclusions. Based on the sample size and distribution, one can calculate the probability that statistics, which measure the central tendency, variability, distribution, and relationships between characteristics within a data sample, provide an accurate picture of the corresponding parameters of the whole population from which the sample is drawn.

Inferential statistics are used to make generalizations about large groups, such as estimating average demand for a product by surveying a sample of consumers' buying habits or to attempt to predict future events, such as projecting the future return of a security or asset class based on returns in a sample period.