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?

a) Modeling event/time 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

5. _____ random variables are used to model rates.

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?

b) Hypothesis

8. 4. Normalized data are centered at_____and have units equal to standard deviations of the

original data.

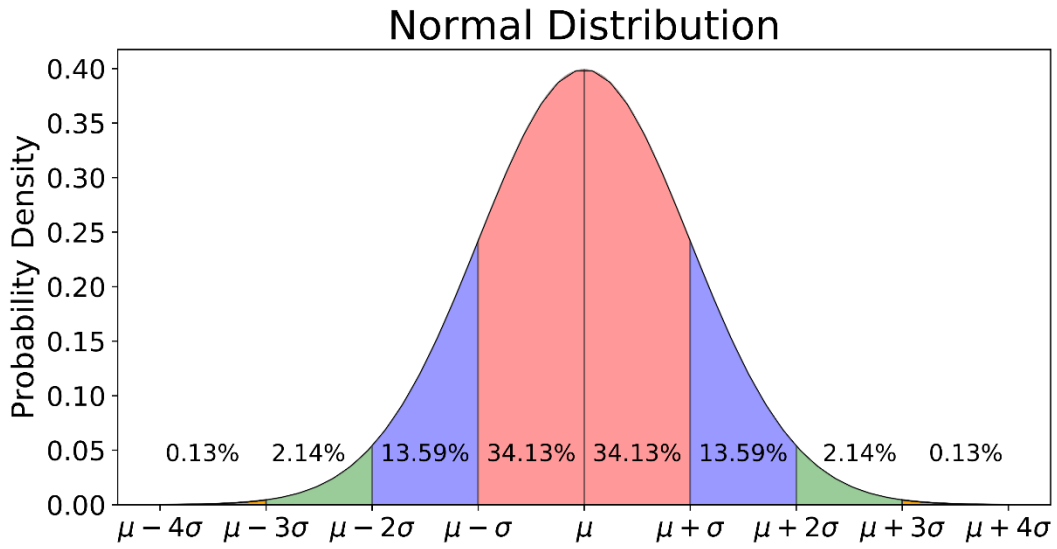
a) 0

9. Which of the following statement is incorrect with respect to outliers?

a) Outliers can have varying degrees of influence

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

A common random continuous distribution you will encounter in statistics is the **normal distribution**. It is **bell-shaped** and it is sometimes called the **bell-curve**. It is a continuous probability distribution relating to the mean and standard deviation. The normal distribution plays an important role in inferential statistics.



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.

Followings are the imputation techniques that can be implemented :-

K Nearest Neighbors

In this method, data scientists choose a distance measure for k neighbors, and the average is used to impute an estimate. The data scientist 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

Linear Interpolation

Linear interpolation is often used to approximate a value of some function by using two known values of that function at other points. This formula can also be understood as a weighted average. The weights are inversely related to the distance from the end points to the unknown point. The closer point has more influence than the farther point.

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, data scientists 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. This method does not use time-series characteristics or depend on the relationship between the variables

Q 12 What is A/B testing?

A/B testing in its simplest sense is an experiment on two variants to see which performs better based on a given metric. Typically, two consumer groups are exposed to two different versions of the same thing to see if there is a significant difference in metrics like sessions, click-through rate, and/or conversions.

Using the visual above as an example, we could randomly split our customer base into two groups, a control group and a variant group. Then, we can expose our variant group with a red website banner and see if we get a significant increase in conversions. It's important to note that all other variables need to be held constant when performing an A/B test.

Getting more technical, A/B testing is a form of statistical and two-sample hypothesis testing. Statistical hypothesis testing is a method in which a sample dataset is compared against the population data. Two-sample hypothesis testing is a method in determining whether the differences between the two samples are statistically significant or not.

Q13. 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 typically considered terrible practice since it ignores feature correlation. Consider the following scenario: we have a table with age and fitness scores, and an eight-year-old has a missing fitness score. If we average the fitness scores of people between the ages of 15 and 80, the eighty-year-old will appear to have a significantly greater fitness level than he actually does.

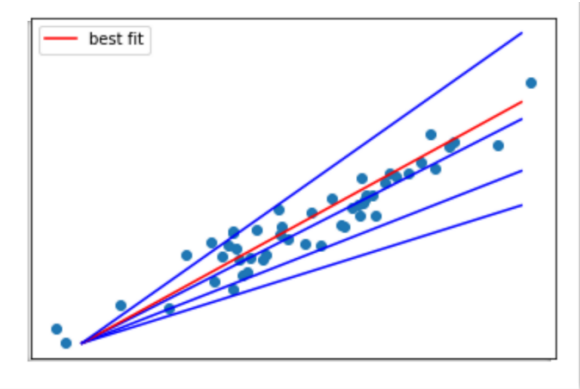
Second, mean imputation decreases the variance of our data while increasing bias. As a result of the reduced variance, the model is less accurate and the confidence interval is narrower.

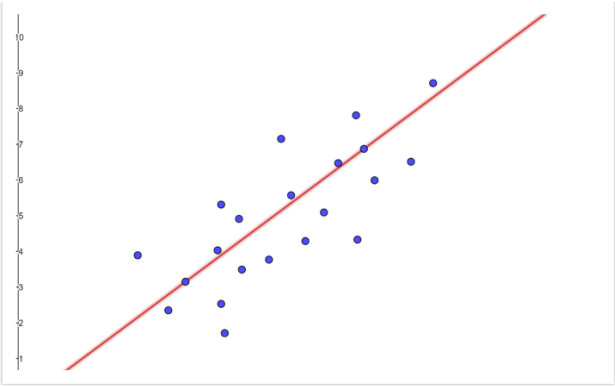
Q14 What is linear regression in statistics?

It tries to find out the best possible linear relationship between the input features and the target variable(y).

That's it! This is what Linear Regression does. Pretty simple right?

In machine learning jargon the above can be stated as "It is a supervised machine learning algorithm that best fits the data which has the target variable (dependent variable) as a linear combination of the input features (independent variables)."





Simple linear regression is a statistical method that allows us to summarize and study relationships between two continuous (quantitative) variables:

- One variable, denoted x, is regarded as the predictor, explanatory, or independent variable.
- The other variable, denoted y, is regarded as the response, outcome, or dependent variable

Q15. What are the various branches of statistics?

Two branches, **descriptive statistics and inferential statistics**, comprise the field of statistics.

Statistics

Descriptive Statistics

It describes the important characteristics/ properties of the data using the measures the central tendency like mean/ median/mode and the measures of dispersion like range, standard deviation, variance etc.

Data can be summarized and represented in an accurate way using charts, tables and graphs.

Inferential Statistics

It is about using data from sample and then making inferences about the larger population from which the sample is drawn. The goal of the inferential statistics is to draw conclusions from a sample and generalize them to the population. It determines the probability of the characteristics of the sample using probability theory. The most common methodologies used are hypothesis tests, Analysis of variance etc