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

Ans:- a)True

2) **Which of the following theorem states that distribution of averages of iid variables, properly normalized become of that standard normal as the sample size increases?**

Ans:- a) Central Limit Theorem

3) **Which of the following is incorrect with respect to 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) **Usually replacing the standard error by its estimated value does change the CLT**

Ans:- b) False

7) **Which of the following testing is concerned with making decisions using data?**

Ans:- b)Hypothesis

8) **Normalized data are centered at_____ have units equal to standard deviations of the normal data?**

Ans:- a) 0

9) **Which of the following statement is incorrect with respect to the outliers**

Ans:- c) Outliers can conform to the regression relationship

10) **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.

KEY TAKEAWAYS

- A normal distribution is the proper term for a probability 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.
- In reality, most pricing distributions are not perfectly normal.

11) **How to you handle missing data? What imputation technique do you recommend?**

Ans:- Missing data is an inevitable part of the process. As data researchers, we pour a lot of resources, time and energy into making sure the data set is as accurate as possible. However, data inevitably goes missing. As someone

who has been handling data analytics and overseen dozens of research projects for several years, missing data is just one of those “It sucks, but it’s no one’s fault” scenarios. Sometimes, data sets come up short, no matter how many times data scientists clean and prepare it. The best way to handle such situations is to develop contingency plans to minimise the damage.

Missing data – Why does it matter so much?

Missing data is a huge problem for data analysis because it distorts findings. It’s difficult to be fully confident in the insights when you know that some entries are missing values. Hence, why they must be addressed. According to data scientists, there are three types of missing data. These are Missing Completely at Random (MCAR) – when data is completely missing at random across the dataset with no discernable pattern. There is also Missing At Random (MAR) – when data is not missing randomly, but only within sub-samples of data. Finally, there is Not Missing at Random (NMAR), when there is a noticeable trend in the way data is missing.

Best techniques to handle missing data

Use deletion methods to eliminate missing data

The deletion methods only work for certain datasets where participants have missing fields. There are several deleting methods – two common ones include Listwise Deletion and Pairwise Deletion. It means deleting any participants or data entries with missing values. This method is particularly advantageous to samples where there is a large volume of data because values can be deleted without significantly distorting readings. Alternatively, data scientists can fill out the missing values by contacting the participants in question. The problem with this method is that it may not be practical for large datasets. Furthermore, some corporations obtain their information from third-party sources, which only makes it unlikely that organizations can fill out the gaps manually. Pairwise deletion is the process of eliminating information when a particular data point, vital for testing, is missing. Pairwise deletion saves more data compared to likewise deletion because the former only deletes entries where variables were necessary for testing, while the latter deletes entire entries if any data is missing, regardless of its importance.

Use regression analysis to systematically eliminate data

Regression is useful for handling missing data because it can be used to predict the null value using other information from the dataset. There are several methods of regression analysis, like Stochastic regression. Regression methods can be successful in finding the missing data, but this largely depends on how well connected the remaining data is. Of course, the one drawback with regression analysis is that it requires significant computing power, which could be a problem if data scientists are dealing with a large dataset.

Data scientists can use data imputation techniques

Data scientists use two data imputation techniques to handle missing data: Average imputation and common-point imputation. Average imputation uses the average value of the responses from other data entries to fill out missing values. However, a word of caution when using this method – it can artificially reduce the variability of the dataset. Common-point imputation, on the other hand, is when the data scientists utilise the middle point or the most commonly chosen value. For example, on a five-point scale, the substitute value will be 3. Something to keep in mind when utilising this method is the three types of middle values: mean, median and mode, which is valid for numerical data (it should be noted that for non-numerical data only the median and mean are relevant).

Keeping things under control

Missing data is a sad fact of life when it comes to data analytics. We cannot avoid situations like these entirely because there are several remedial steps data scientists need to take to make sure it doesn’t adversely affect the analytics process. While these methods are helpful, they are not foolproof because they are contentious, meaning,

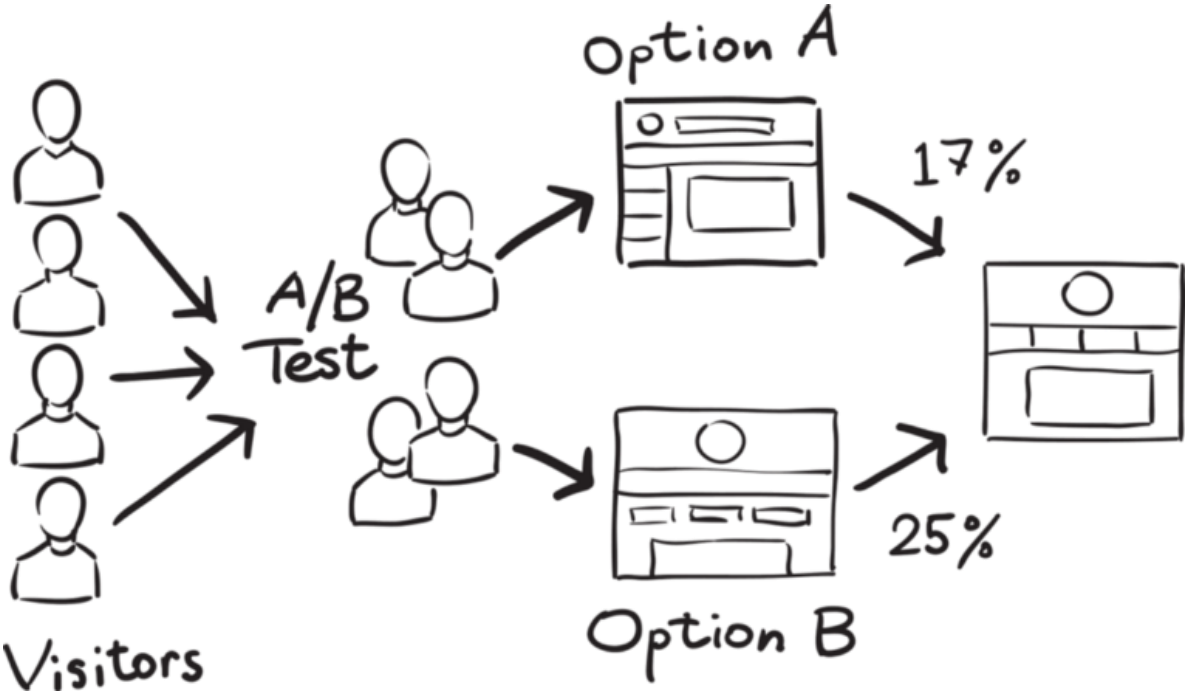
their effectiveness depends heavily on circumstances. The best option available to data scientists is to work with powerful, processing tools that can make the data capturing and analysis process significantly easier. It is the best way to handle missing data**.**

12) What is A/B testing?

Ans:- 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 instance, let's say you own a company and want to increase the sales of your product. Here, either you can use random experiments, or you can apply scientific and statistical methods. A/B testing is one of the most prominent and widely used statistical tools.

In the above scenario, you may divide the products into two parts – A and B. Here A will remain unchanged while you make significant changes in B's packaging. Now, on the basis of the response from customer groups who used A and B respectively, you try to decide which is performing better.



It is a hypothetical testing methodology for making decisions that estimate population parameters based on sample statistics. The population refers to all the customers buying your product, while the sample refers to the number of customers that participated in the test.

13) Is mean imputation of missing data acceptable practice?

Ans:- 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.

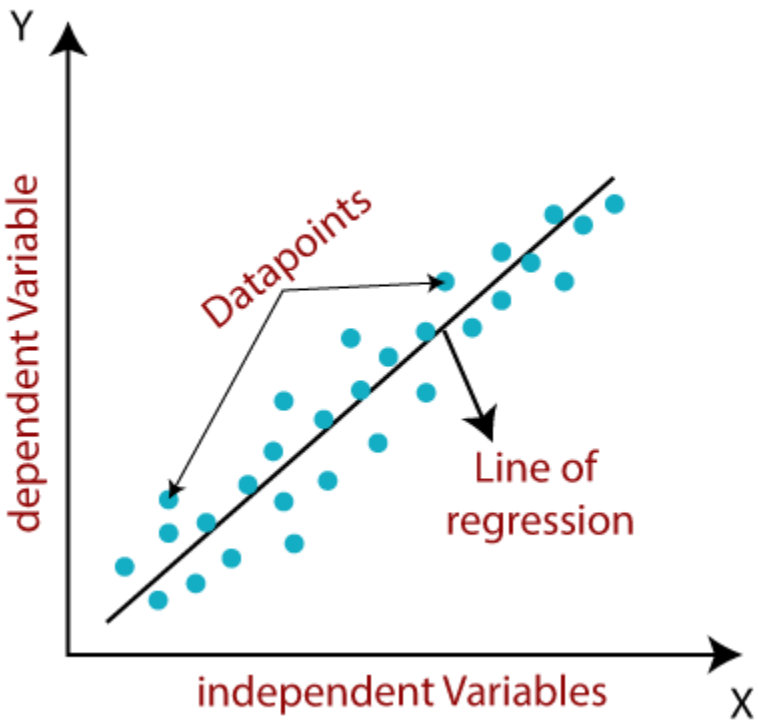
Second, mean imputation decreases the variance of our data while increasing bias. As a result of the reduced variance, the model is less accurate and the confidence interval is narrower.

14) What is linear regression in statistics?

Ans:- Linear regression is one of the easiest and most popular Machine Learning algorithms. It is a statistical method that is used for predictive analysis. Linear regression makes predictions for continuous/real or numeric variables such as **sales, salary, age, product price,** etc.

Linear regression algorithm shows a linear relationship between a dependent (y) and one or more independent (y) variables, hence called as linear regression. Since linear regression shows the linear relationship, which means it finds how the value of the dependent variable is changing according to the value of the independent variable

The linear regression model provides a sloped straight line representing the relationship between the variables. Consider the below image:



Mathematically, we can represent a linear regression as:

```
1.  y= a0+a1x+ ε
```

Here,

Y= Dependent Variable (Target Variable)
X= Independent Variable (predictor Variable)

a0= intercept of the line (Gives an additional degree of freedom)
a1 = Linear regression coefficient (scale factor to each input value).
ε = random error

The values for x and y variables are training datasets for Linear Regression model representation

Types of Linear Regression

Linear regression can be further divided into two types of the algorithm:

- **Simple Linear Regression:**
  If a single independent variable is used to predict the value of a numerical dependent variable, then such a Linear Regression algorithm is called Simple Linear Regression.
- **Multiple Linear regression:**
  If more than one independent variable is used to predict the value of a numerical dependent variable, then such a Linear Regression algorithm is called Multiple Linear Regression

15 **) What are the various branches of statistics?**

Ans:- **Statistics:**

Statistics is a study of presentation, analysis, collection, interpretation and organization of data

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 its not easy or possible to examine each member of the population.

- **Descriptive Statistics:**
  Descriptive statistics are use to get a brief summary of data. You can have the summary of data in numerical or graphical form.