Statistics

Q.1)

Option A - True

Q.2)

Option A - True

Q.3)

Option B- Central Mean Theorem

Q.4)

Option D - All of the mentioned

Q.5)

Option C) Poisson

Q.6)

Option A - True

Q.7)

Option B - Hypothesis testing is concerned with making decisions using data.

Q.8)

Option A – Normalized data are centered at 0 and have units equal to standard deviations of the original data.

Q.9)

Option C- Outliers cannot conform to the regression relationship

Answer 10)

Normal distribution is also known Gaussian distribution and bell curve as its curve looks like bell shape. It shows frequent data near the men and then the data far away from the mean.

The following are some important concept for Normal distribution

1. Normal distribution is most common type of distribution used.
2. Normal distribution comes up frequently again and again.
3. In Normal distribution mean, mode and median all are equal.
4. In this 68 % of data falls within 1 standard deviation.

5. It describes that how the values of the variables are equally distributed.
6. Some examples for normal distribution blood pressure , height , IQ score,

Following is the formulae to calculate Normal distribution

Formula

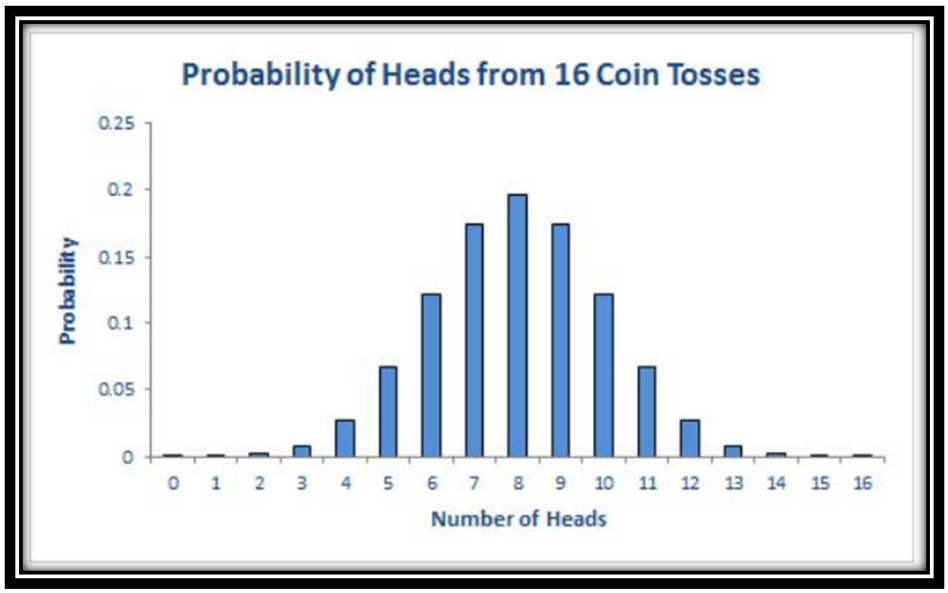
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

Where

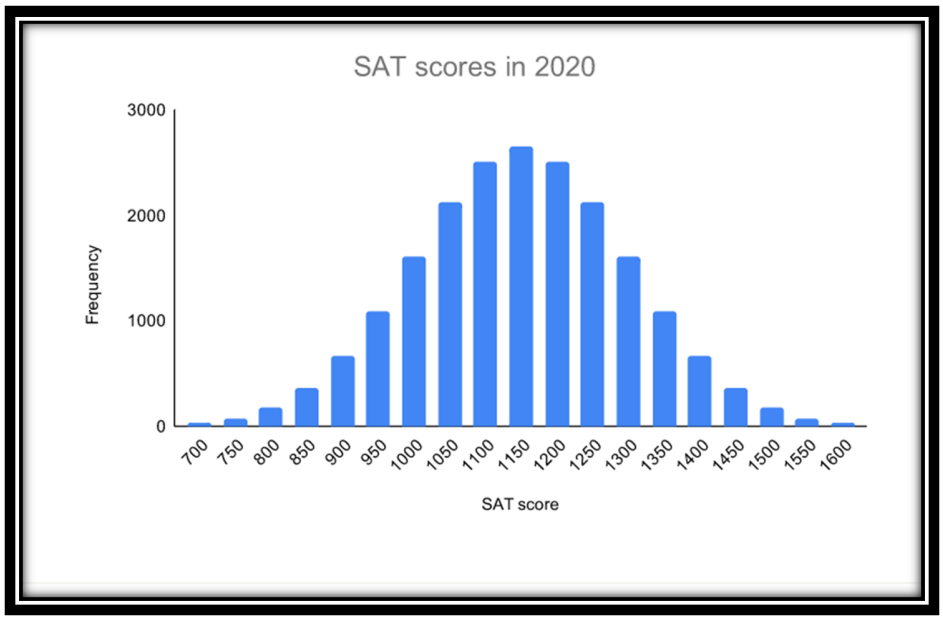
f(x) = probability density function
σ = standard deviation
μ = mean

Some examples for Normal distribution curve in daily life

1)



2)

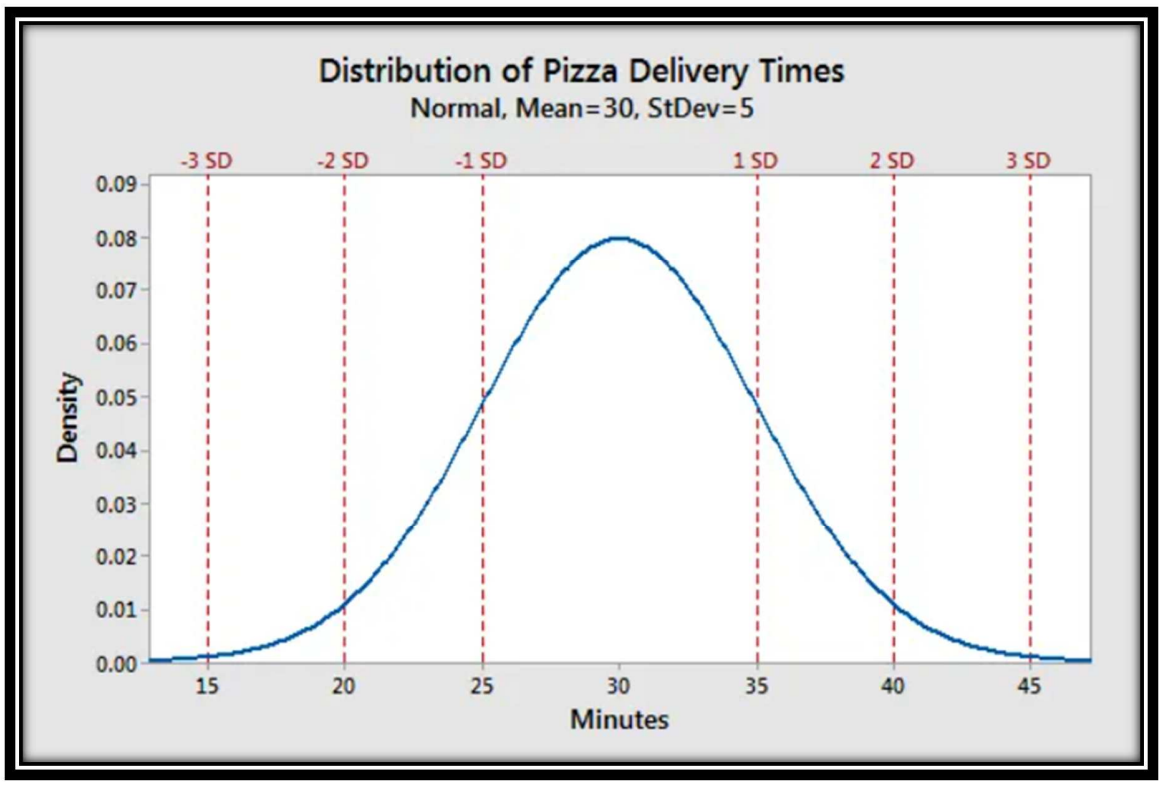


The following is the proportion value for the Normal distribution which describes the percentage of data contained.

Mean Standard deviation	Percentage of data
1	68 %
2	95%
3	99.7%

Example.

Assume that a pizza restaurant has a mean delivery time of 30 minutes and a standard deviation of 5 minutes. The chart below illustrates this property graphically.



Answer 11)

As Data scientist we put lot of efforts, resource ,time and energy to make the data set as perfect as possible. But sometimes the data and undergo missing .It is the most unwanted part dealing with data but it may happens which can be frustrating but not anyone's fault .Sometimes data comes short no matter how many time Data scientist cleans it and ten prepare it . To overcome this situation we use some techniques to minimize this damage and get the missing data.

The following is the reason why missing data is important

1. It can distort the finding in the dataset.
2. It decreases confidence to work on the missed data.

There are 3 types of missing data as

1. MCAR (Missing Completely at Random)
   When data is completely missing randomly.

2. MAR(Missing At Random)
   When data is not missing randomly, but only within sub-samples of data.

3. NMAR(Not Missing at Random)
   When there is a noticeable trend in the way data is missing.

There are some best techniques to eliminate the missing data

Deletion method –

The deletion methods only work for certain datasets where participants have missing fields. There are common method used in deletion method as

1. List wise Deletion
   List wise deletion (complete-case analysis) removes all data for a case that has one or more missing values.
2. Pair wise Deletion
   Pair wise deletion (available-case analysis) attempts to minimize the loss that occurs in list wise deletion.

Properties:

- It deletes the data entries with missing value.
- This method is particularly advantageous to samples where there is a large volume of data
- This method is not suitable where the dataset is too large.
- Pair wise deletion saves more data than list wise deletion

Substitute the value as mean :

This method can be used when the percentage is too high by substituting the mean , medium value .This method can create complication buy causing variance.

Logistic regression:

In this method the following steps are carried out

1. First several predictors of variables with missing values are identified using correlation matrix.
2. Best Predictors are selected and are used as independent variable in the regression equation.
3. The variables with missing data are used as dependent variables.
4. Cases with competed data are used to get the regression equation
5. The equation is then used to predict missing values for incomplete cases and the process is repeated until there is little difference between the Predicted values.

Answer no 12

A/B Testing

A/B Testing is random testing model which is used to compare versions of variables to identify which performs better .It compares the two versions of variables.

For example there is a sales company and they want to increase their sale . Hence they need to compare the products . In this they will use Random experiment or scientific testing mode.
A/B is the scientific testing model and is more prominent and widely used statistic tool.

Working:

1. In this model we divide the product in two parts as A and B.
2. A will remain unchanged and we make changes only on B packaging.
3. We will observe the customer response on both the packages A and B and then will decide which is performing better on the basis of the response.
4. It has two attributes as **population** refers the all the customer buying the product and **sample** refers the number of customers that participate in the test.

For example – There is a news company where they want to determine which news format works better and is more popular .It takes the original format of the news and name it as A and makes some changes in same A format and renames it to B .Finally depending on the popularity and response from the customers identifies which works more better.

This A/B testing it is tested using hypothesis testing to experiment its fact .
Two types of Hypothesis.

1. **Null hypothesis or H0 –**It means there is no change in conversion rate both the variable variance A and B

2. **Alternative hypothesis –**This is the hypothesis which researchers believe it to be true almost

Two type of Errors –

1. Type 1 error – In this we accept variant B when A is not working.
2. Type 2 error – In this we conclude Variant B when A is not better

A/B testing should be avoided in following circumstances –

1. Invalid hypothesis
2. Testing too many elements together.
3. Not considering external factor

Answer 13 –

Mean imputation – The process of replacing Null values in the data set with its mean value is known as mean imputation .

As according mean imputation is not an effective method for implementing to missing data .As it ignores feature correlation.

For example consider a fitness score table

The table has two sets as age and fitness score and an nine year old child had missing fitness score .If we find the average of fitness score of people from 15 to 90 years .The 90year old will appear greater fitness level what he

actually has. Thus here mean imputation method is not effective to be used .This will increase the bias and will reduce the variance

Answer 14

Linear regression in statistics

If we want to draw the conclusion using variable x concerning variable y .In this linear model , it assumes linear relation between input variable (x) and output variable (y) In this case

y is called dependent variable or response variable

x is called as independent variable or predictor

When there is a single input variable (x), the method is referred to as **simple linear regression**. When there are **multiple input variables**, literature from statistics often refers to the method as multiple linear regression.

This can be summarized by using a straight line if the relationship between two variables is linear.

A straight line can be represented using equation as

$$y= a + bx$$

Where a = intercept

b = slope of the equation

Techniques to train the model

There are several techniques used to train the model

1. Simple Linear Regression

With simple linear regression when we have a single input, we can use statistics to estimate the coefficients.

2. Ordinary Least Squares

When we have more than one input we can use Ordinary Least Squares to estimate the values of the coefficients.

3. Gradient Descent

This operation is called Gradient Descent and works by starting with random values for each coefficient.

4. Regularization

There are extensions of the training of the linear model called regularization methods. These seek to both minimize the sum of the squared error of the model on the training data (using ordinary least squares) but also reduce the complexity of the model

Two popular examples of regularization procedures for linear regression are:

- **Lasso Regression**: where Ordinary Least Squares is modified to also minimize the absolute sum of the coefficients (called L1 regularization).
- **Ridge Regression:** where Ordinary Least Squares is modified to also minimize the squared absolute sum of the coefficients (called L2 regularization).

Answer 15

Statistics is the main branch of mathematics .It is a type of analyses where data is experimented and tested and can be used .It is used for various operation like data collection, data analysis and organizing the data.

The following are measures used in statistics:

1. Mean
2. Regression analysis
3. Skewness
4. Variance
5. Kurtosis

There are two branches of statistics

Descriptive statistics

Descriptive statistics help to describe and understand the data set .It has measures like

1. Mean – mean is calculated by adding all the figures in the data set and dividing by number of data set.
   Eg
   A (2, 3, 4, 5, 6)
   The mean is 20/5
   Mean=4
2. Median –It is the figure in center of the data set.
3. Mode – The values which are appearing most in the data set

Random variables or variants are the attributes whose value can be changed with time.

There are various types of variants as

Qualitative or nominal –This is described as word
Quantitative – This is described as numbers
Ordinal – This are in between like much, same , worse

Types of Descriptive statistics

1) Center tendency

   This type focuses on mean and median of the data set .This focus on dispersion of data.

It uses graphs, tables and general discussion. This measures the most common pattern of the analyzed data.

2) Measures of variability

It analysis the data that how much it is dispersed in the data set. It describes the spread of data.

Measures of variability-

- Absolute deviation
- Range
- Quartiles

Inferential Statistics –

We have understood about descriptive statistics as it calculates the mean, median and standard deviation of the population .Consider an example we need to calculate the mean, median and standard deviation for the group of 100 students .In this we are thus interested in all 100 students data set .This is known as population .If we want to calculate the mean, median and standard deviation of only those students marks which are from Canada. So we are interested only in those part of data set .This is known as sample.

Inferential Statistics is about calculation of the sample data set instead of the population