DATA SCIENCE INTERVIEW QUESTIONS
TOPIC: INTRODUCTION TO STATISTICS

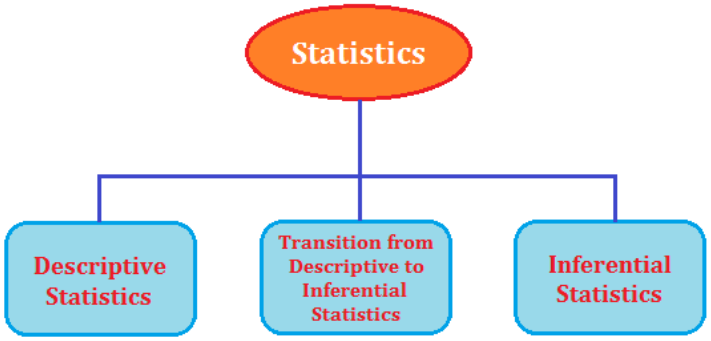
Q.1) What is the definition of statistics?

A.1) Statistics – is a branch of applied mathematics dealing with data collection, organization, analysis, interpretation and presentation. It refers to numbers that are used to describe data or relationships. Statistics involves the process of gathering and evaluating data and then summarizing in mathematical form. It studies methodologies to gather, review, analyze and draw conclusions from data some statistical methods include the following:

- Mean
- Regression analysis
- Skewness
- Kurtosis
- Variance
- Analysis of Variance

Q.2) What are types of statistics?

A.2) Statistics can be categorized into 2 types:

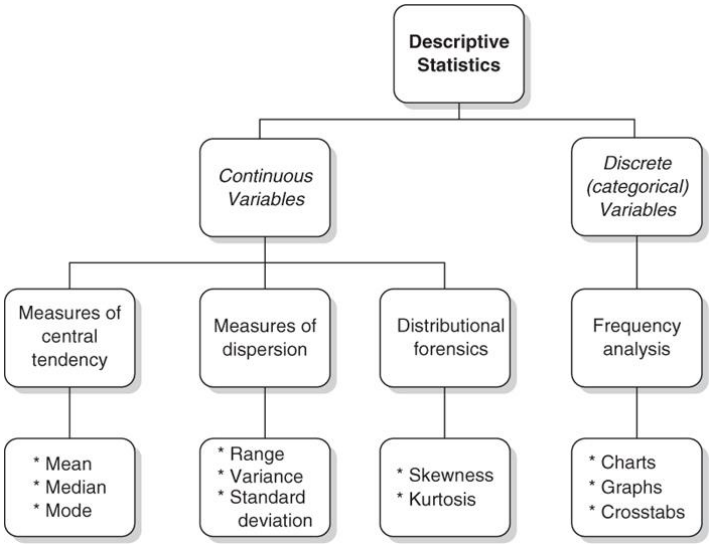


a.) Descriptive Statistics - used to describe a specific data or a population and summarizing observations. The results are specific to a single population. Include techniques to organize, display and describe data using tables, graphs, summary measures. These are used to see observable patterns in the data belongs to
   - avg
   - range
   - std deviation
   - frequency distribution
   - quartiles
   - variance
   - mean

b.) Inferential statistics - makes predictions about a population based on a sample of data taken from the population and draws conclusions from the data that are subject to random variation (e.g.: observational errors, sampling variation).

Q.3) What is Descriptive statistics?
A.3) Descriptive Statistics:

a.) It is the summarization of collection of data in clear and understandable way.
b.) Used to describe a specific data or a population. The results are specific to a single population. These are used to see observable patterns in the data.



There are four major types of descriptive statistics:

- Measures of Frequency: Count, Percent, Frequency.
- Measures of Central Tendency: Mean, Median, and Mode.
- Measures of Dispersion or Variation: Range, Variance, Standard Deviation.
- Measures of Position: Percentile Ranks, Quartile Ranks.

a.) Inferential Statistics – include techniques that enable us to make use of information that gathered from a sample to make decisions, inferences and predictions. Belongs to
    - Intervals
    - Hypothesis testing

There are two major types of inferential statistics:

- Estimating parameters: Taking a statistic from sample data and using it to say something about a population parameter.
- Hypothesis tests: Use sample data to answer research questions.

Q.4) When to use GM, HM, AM?
A.4) Relation between AM, GM, HM can be derived with the basic knowledge of progressions or mathematical sequences. A collection of objects in a specified pattern in mathematics is called a "Mathematical Sequence" – also called as "Progression". It includes 3 types of sequences:

a.) AM [Arithmetic Mean] : is the mean/avg of the set of numbers which is computed by adding all the terms in the set of numbers and dividing the sum by total number of terms.
Ex : Sequence of 'n' terms as {a1, a2, a3,….an}

$$AM = \frac{a1+a2+a3\ldots+an}{n}$$

Used : 1. Am is useful in Machine Learning to summarize a variable

2. Can be calculated using the mean () Numpy function

3. Can be easily distorted if the sample of observations contains outliers or data has a non-Gaussian distance.

4. Used in situations where mean/avg of any statistical data to be determined, used on data like Prices.

b.) GM [Geometric Mean] : is the mean value or central term in the set of numbers in geometric progression. This is computed as nth root of the product of all terms in sequence.
Ex: GM = $n\sqrt{a1, a2, a3,\ldots an}$

Used: 1. GM is used in computation of stock indexes.

2. Used to calculate the annual returns of the portfolio.

3. Used in studying biological process such as cell division and bacterial growth.

4. GM is appropriate when data contains values with different units of measure, like height, dollars.

5. All values should be +ve,-ve or zero values.

6. In ML, used to calculate the model evaluation metric.

7. GM can be used when dealing with ratios.

c.) HM [Harmonic Mean] : one type of determining avg. It is computed by dividing the number of values in sequence by sum of reciprocals of the terms in the sequence.
Ex: HM = $n/ (\sum 1/x\_i)$

Used: 1. HM is used to determine the price earnings ratio and other avg multiples in finance.

2. Used to specify the ration b/n 2 quantities with different measures like speed, acceleration, frequency etc..

3. No –ve or zero values, all rates must be +ve.

4. In ML, calculate the F-Measure [F1-Measure]

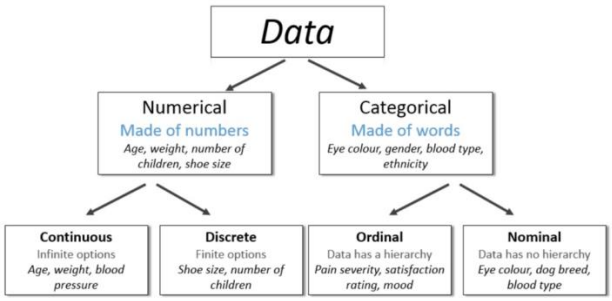
5. HM can be used when working with rates like mileage.

- ✓ F1-Measure: is the model evaluation metric i.e. used to calculate the HM of the precision and recall metrics.
- ✓ Which mean to choose?
  a. AM – if values have same units
  b. GM – if values have differing units
  c. HM – if values are rates.

Q.5) When to use mean and median ?

A.5) There are 3 types of data:

a) Numerical data
b) Categorical data
c) Ordinal data –amalgamation of numerical and categorical data



- Mean – is another name for average. Is used to calculate the avg/mean of a given list of numbers. Mainly used to calculate continuous variables. It returns the mean of the dataset passed as parameter. Mean is used when the data is symmetrical meaning with no extreme high or low in values of data. The numbers to the extreme will skew the data so that the mean becomes much higher or lower. Mean gives equal weight to every value.

  Formula: x=np.mean(data)

Mean Formula

$$\text{Mean} = \frac{\text{Sum of All Data Points}}{\text{Number of Data Points}}$$

$$\text{Mean} = \text{Assumed Mean} + \frac{\text{Sum of All Deviations}}{\text{Number of Data Points}}$$

- Median – is the midpoint value. Is used to calculate the median value of the dataset. It takes one parameter: data list. Median is used if the data is asymmetrical. That means if the numbers in the data are very low or higher than most of the numbers. Median gives less weight to the values at the extreme numbers (outliers).

  Advantages of the median: Extreme values (outliers) do not affect the median as strongly as they do the mean.

  Formula: x=np.median(data)

- Mode – is the most common value, highest frequency in the data set. Mode can be used when data is not numbers and want to find the most occurring event. Like a result from a survey.

  Formula: x=np.mode (data)

Q.6) What is percentile?

A.6) Percentile in python –used to compute the nth percentile of the given data along the specified axis. Percentile is defined as the percentage of total observations present below a specified value.Percentile is a number where certain input of scores fall below that number.

Example: if we say 50 marks out of 100 is at 60^{th} percentile, it explains that 60% per total students secured less or equal to 50 marks.

Formula: n = (P/100) * N, where P = Percentile,

N = number of values in a data set (sorted from smallest to largest)

n = ordinal rank of a given value.

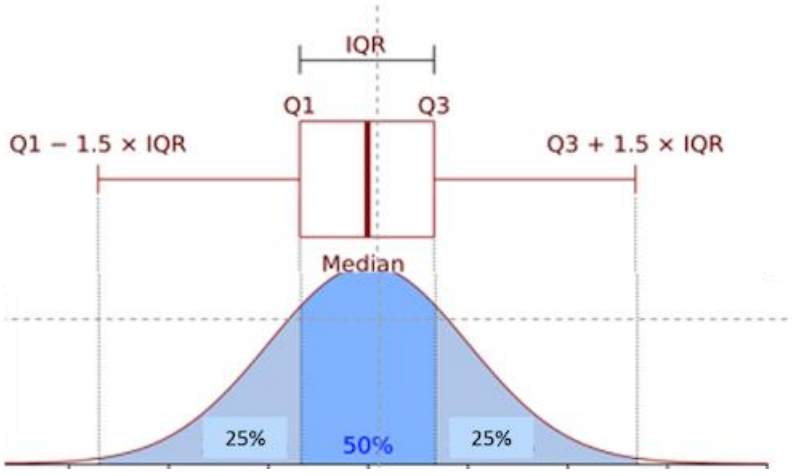
- Percentiles are frequently used to understand test scores and biometric measurements.

    - 25^{th} percentile is called "First Quartile"
    - 50^{th} percentile is called "Median"
    - 75^{th} percentile is called "Third Quartile"
    - Difference between First and Third Quartile is IQR [Inter Quartile Range]
    - Makes data easy to read
    - Helps in getting an idea on outliers
    - Used to report scores in GRE,SAT

Q.7) What is Quantile?

A.7) Quantile in python – The word "Quantile" comes from the word Quantity. **quantile** () function used to get values at the given quantile over the requested axis. Plays an imp role in statistics with nd. In each of any set of values of a variate which divide a frequency distribution into equal groups, each containing the same fraction of the total population. Quantile is something that divides the data set into equal parts. It divides the dataset into 4 equal parts is called quartile. Median divides the whole data in 2 equal parts. Quartile divides the data in 4 equal parts.

Use of Quantiles: Besides specifying the position of a set of data, quantiles are helpful in other ways. Suppose we have a simple random sample from a population, and the distribution of the population is unknown. To help determine if a model, such as a normal distribution or Weibull distribution is a good fit for the population we sampled from, we can look at the quantiles of our data and the model. By matching the quantiles from our sample data to the quantiles from a particular probability distribution, the result is a collection of paired data. We plot these data in a scatterplot, known as a quantile-quantile plot or q-q plot. If the resulting scatterplot is roughly linear, then the model is a good fit for our data.



Common Quantiles: Certain types of quantiles are used commonly enough to have specific names. Below is a list of these:

- The 2 quantile is called the median
- The 3 quantiles are called terciles
- The 4 quantiles are called quartiles
- The 5 quantiles are called quintiles
- The 6 quantiles are called sextiles

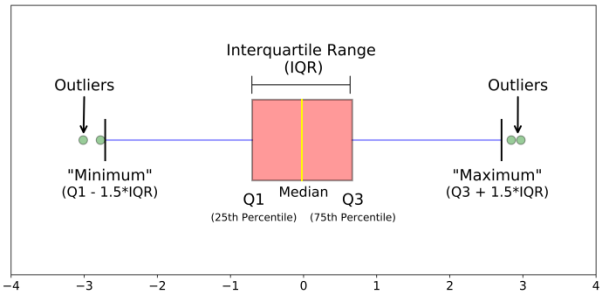
- The 7 quantiles are called septiles
- The 8 quantiles are called octiles
- The 10 quantiles are called deciles
- The 12 quantiles are called duodeciles
- The 20 quantiles are called vigintiles
- The 100 quantiles are called percentiles
- The 1000 quantiles are called permilles.

Q.8) What is IQR?

A.8) IQR [Inter Quartile Range] – is the difference between the 75^{th} and 25^{th} percentile of the data.

IQR = Q3 – Q1

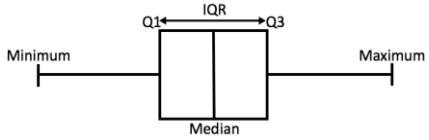
- Used to build box plots, simple graphical representation of a probability distribution.
- Identify outliers in the given dataset.
- Gives the central tendency of the data.



- median (Q2/50th Percentile): the middle value of the dataset.
- first quartile (Q1/25th Percentile): the middle number between the smallest number (not the "minimum") and the median of the dataset.
- third quartile (Q3/75th Percentile): the middle value between the median and the highest value (not the "maximum") of the dataset.
- interquartile range (IQR): 25th to the 75th percentile (75%-25%)
- whiskers (shown in blue)
- outliers (shown as green circles)
- "maximum": Q3 + 1.5*IQR
- "minimum": Q1 -1.5*IQR
- The median is the median (or centre point), also called second quartile, of the data (resulting from the fact that the data is ordered).
- Q1is the first quartile of the data, i.e., to say 25% of the data lies between minimum and Q1
- Q3is the third quartile of the data, i.e., to say 75% of the data lies between minimum and Q3
- The difference between Q3 and Q1 is called the Inter-Quartile Range or IQR.

Steps: i. Arrange data in ascending order.
ii. Calculate Q1
iii. Calculate Q3
iv. Find IQR=Q3-Q1
v. Find lower range = Q1-(1.5 * IQR)
vi. Find upper range=Q3+ (1.5 * IQR)
vii. Anything that lies outside of lower and upper range is outlier.

Q.9) Why should you consider 1.5 times of IQR for finding outliers?
A.9) An outlier is a data point which differs significantly from other observations. To explain outlier we use box plot. A box plot tells us more or less about the distribution of the data. It gives a sense of how much the data is actually spread about, what's its range, and about its skewness.



In the above fig: minimum - minimum value in the dataset
maximum - maximum value in the dataset
median - median (or center point), also called second quartile, of the data.
Q1 - is the first quartile of the data, i.e., to say 25% of the data lies between minimum and Q1.
Q3 - is the third quartile of the data, i.e., to say 75% of the data lies between minimum and Q3.
IQR = Q3 - Q1
Lower Bound: (Q1 - 1.5 * IQR)
Upper Bound: (Q3 + 1.5 * IQR)
"Any data point less than the Lower Bound or more than the Upper Bound is considered as an outlier".

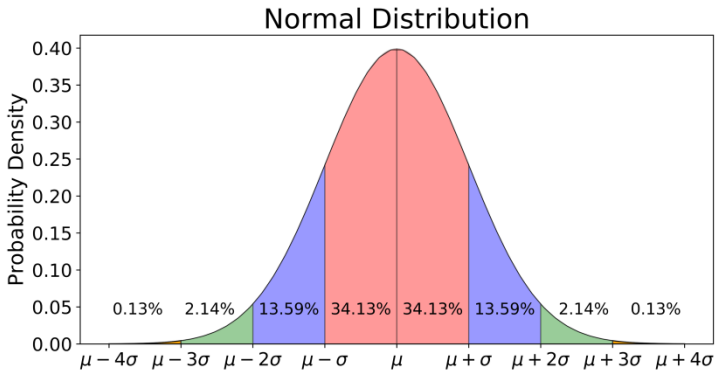
Q: Why only 1.5 times the IQR? Why not any other number?
A: About 68.26% of the whole data lies within one standard deviation (<σ) of the mean (μ)
About 95.44% of the whole data lies within two standard deviations (2σ) of the mean (μ)
About 99.72% of the whole data lies within three standard deviations (<3σ) of the mean (μ)
And the rest 0.28% of the whole data lies outside three standard deviations (>3σ) of the mean (μ).
And this part of the data is considered as outliers.
The first and the third quartiles, Q1and Q3, lies at - 0.675σ and +0.675σ from the mean, respectively.

- Taking scale= 1:
  Lower Bound:
  = Q1 - 1 * IQR
  = Q1 - 1 * (Q3 - Q1)
  = -0.675σ - 1 * (0.675 - [-0.675])σ
  = -0.675σ - 1 * 1.35σ
  = -2.025σ
  Upper Bound:
  = Q3 + 1 * IQR
  = Q3 + 1 * (Q3 - Q1)
  = 0.675σ + 1 * (0.675 - [-0.675])σ
  = 0.675σ + 1 * 1.35σ
  = 2.025σ
- Taking scale= 1.5:
  Lower Bound:
  = Q1 - 1.5 * IQR
  = Q1 - 1.5 * (Q3 - Q1)
  = -0.675σ - 1.5 * (0.675 - [-0.675])σ
  = -0.675σ - 1.5 * 1.35σ
  = -2.7σ
  Upper Bound:
  = Q3 + 1.5 * IQR
  = Q3 + 1.5 * (Q3 - Q1)

$= 0.675\sigma + 1.5 * (0.675 - [-0.675])\sigma$
$= 0.675\sigma + 1.5 * 1.35\sigma$
$= 2.7\sigma$

When scale is taken as 1.5, then according to IQR Method any data which lies beyond 2.7σ from the mean (μ).
To get exactly 3σ, we need to take the scale =1.7, but then 1.5 is more "symmetrical" than 1.7.

Q.10) What is Gaussian Distribution?
A.10) Gaussian Distribution also called as Normal Distribution. It is a form of presenting data by arranging the probability distribution of each value in the data. In this most values remains around 'Mean'. Histograms are used to plot the probability distribution curve. It is also called as "Bell Curve", because of its bell-curved shape.

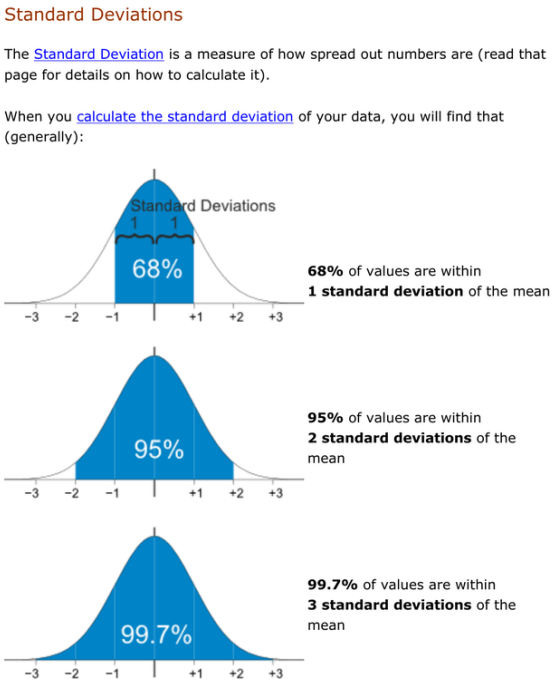


Normal Distribution – A ND random variable might have mean as 0 and std deviation as 1.

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

$\mu =$ Mean
$\sigma =$ Standard Deviation
$\pi \approx 3.14159\cdots$
$e \approx 2.71828\cdots$

Properties of a Standard ND:
1) The normal curve is symmetric about mean and bell shaped.
2) Mean, Median, Mode is 0 which is center of the curve.
3) App 68% of data will be b/n -1 and +1 – w/n std deviation from mean
4) 95% of data will be b/n -2 and +2 – w/n 2 std deviation from mean
5) 99.7% of data will be w/n -3 and +3 – w/n 3 std deviation from mean.



If a distribution is normal, then the values of the mean, median and mode are the same. However, the value of the mean, median, and mode may be different if the distribution is skewed (not Gaussian distribution). Other characteristics of Gaussian distributions are as follows:
Mean±1 SD contains 68.2% of all values.
Mean±2 SD contains 95.5% of all values.
Mean±3 SD contains 99.7% of all values.

Normal distributions do not necessarily have the same mean and standard deviation. A normal distribution with a mean of 0 and a standard deviation of 1 is called a "standard normal distribution".

Q.11) What is Variance?

A.11) Variance – is the squared deviation of a variable from its mean. It measures the spread of random data in a set from its mean or median value. Variance is a calculation that results in a statistical measure of distance that considers random variables in terms of its relationship to the mean of its data set. Variance is the average of squared differences from the mean.

$$\mathrm{Var}(X) = E\left[(X-\mu)^2\right];$$

$$\sigma^2 = \frac{\sum (X-\mu)^2}{N}$$

To compute variance, we use some steps:

1) Find the mean of the set of data.
2) Subtract each number from mean.
3) Square the result.
4) Add the results together.
5) Divide the result by the total number of numbers in the dataset.
6) Can calculate using numpy var () function.

Q.12) What is covariance?

A.12) Covariance is a measure or degree to which 2 variables are linearly associated, like any change in one variable changes in another variable. Can be computed using numpy function – cov ().

Formula: $\text{cov}(x, y) = \frac{\sum (X_i - \bar{X})(Y_j - \bar{Y})}{n-1}$

Where, X_i - the values of X variable
Y_j - the values of Y variable
$\bar{X}$ - mean (avg) of X variable
$\bar{Y}$ - mean (avg) of Y variable
N – No of data points

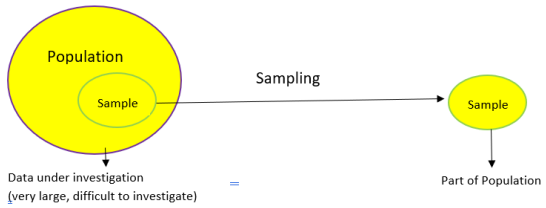
Covariance provides a measure of strength of correlation b/n 2 variables or more set of variables. The covariance matrix element C_{ij} is the covariance of X_i and Y_j. The element C_{ii} is the variance of X_i.

- If cov (X_i, X_j) = 0, then variables are uncorrelated.
- If cov (X_i, X_j) > 0, then variables are positively correlated, indicates 2 variables tend to move in same direction.
- If cov (X_i, X_j) < 0, then variables are negatively correlated, indicates 2 variables tend to move in inverse direction.

Note : If you want to understand the relationship between two variables, then we can use co-variance metric to measure the same and it will helps us to understand whether it has +ve related or -ve related. If co-variance is zero, then both variables are independent to each other.

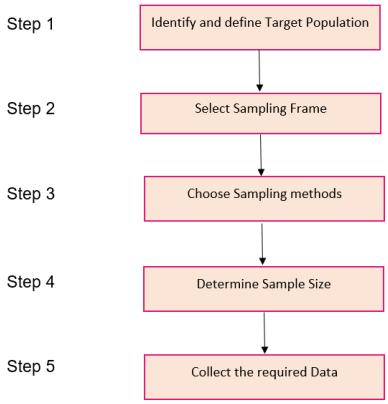
Q.13) What are the sampling techniques?

A.13) Sampling – Sampling is a process used in statistical analysis in which a predetermined number of observations are taken from a larger population. Sample is the subset of population. A well-chosen sample should contain most of the information about population.

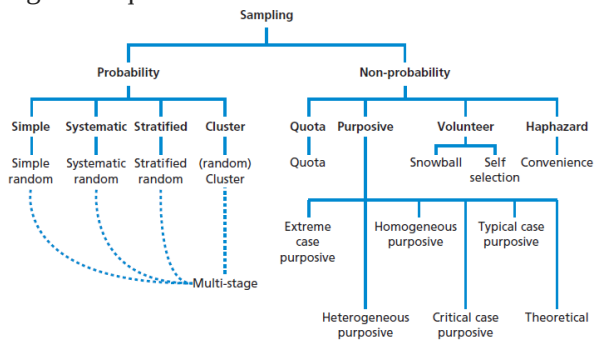


i.) Why to do sampling: it is done to draw conclusions about populations from sample, and it enables us to determine population's characteristics by directly observing only a portion (sample) of population. By doing sampling, requires less time, cost efficient method and analysis is very efficient.

ii.) Steps involved in sampling :



iii.) Different types of sampling techniques :



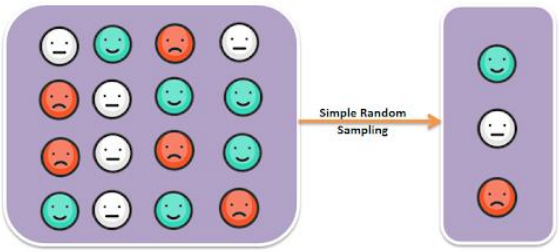
Sampling techniques are of two types:

a. Probability Sampling
b. Non-probability Sampling

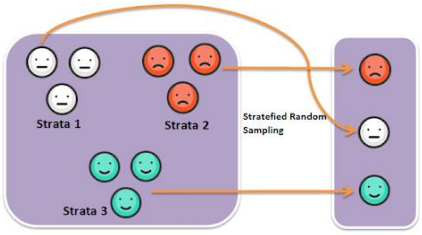
Probability Sampling: uses randomization that every element of population gets an equal chance to get selected through a sample. Also called as Random Sampling. There are 4 types in this:

i. Simple Random Sampling – sample () is an inbuilt function of random module in Python that returns a particular length list of items chosen from the sequence i.e. list, tuple, string or set. Used for random sampling without replacement.

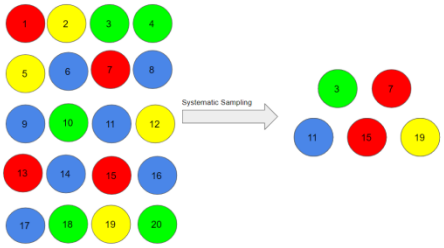
Syntax: random.sample(sequence, k)



ii. Stratified Sampling – divides the elements into subgroups, based on similarity its elements w/n the group belongs to homogenous or heterogeneous among other subgroups. Need to have prior knowledge /information about the population to create subgroups.

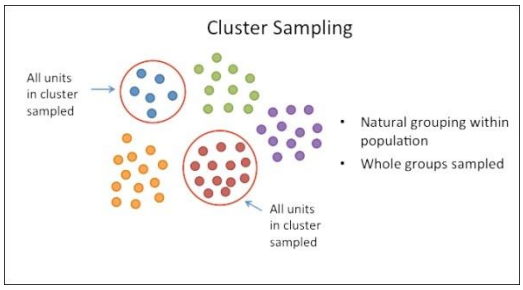


iii. Systematic Sampling – here selection of elements is systematic and not random except the first element. Elements of a sample are chosen at regular intervals of population. All the elements are put together in a sequence first and each element has equal chance of being selected.



iv. Cluster Sampling –entire population is divided into clusters/sections and then these clusters are randomly selected. Clusters identified as age, sex, location. It is done in 2 ways:

a. Single stage clustering sampling : entire cluster is selected randomly for sampling
b. Two stage clustering sampling: randomly select clusters and then form selected clusters.



v. Multi Stage Sampling – combination of one or more methods. Population divided into multiple clusters and these are divided and grouped into various subgroups based on similarity.

Non-probability Sampling: No randomization, it just relies on researcher's ability to select elements for a sample. There are 3types in this:

i. Convenience Sampling - here samples are selected based on availability
ii. Purposive Sampling – it is based on the intention or the purpose of study, only those elements will be selected from the population.
iii. Quota Sampling – depends on some preset- standards.
iv. Referral/Snowball Sampling – used in the situations where the population is completely unknown and rare.

Q.14) What is mse (assume predicted values is an average)

A.14) MSE - Mean Squared Error is metric to measure the average of error squares. It calculates the mean of square of error between actual and predicted values. It's an error calculation metric.

$$MSE = \frac{1}{n}\Sigma\underbrace{\left(y - \widehat{y}\right)^2}_{\text{The square of the difference between actual and predicted}}$$

Q.15) What is probability?
A.15) Measure of likelihood that an event will occur. It stands for the chance that something will happen and calculates how likely it is for that event to happen. It's an intuitive concept that we use on a regular basis without actually realizing that we're speaking and implementing **probability** at work.

It is a measure of how likely an event is. So, if it is 60% chance that it will rain tomorrow, the probability of Outcome "it rained" for tomorrow is 0.6.

Learning of probability helps you in making informed decisions about likelihood of events, based on a pattern of collected data. In the context of data science. There are few terminologies used in probability:

- **Experiment** – are the uncertain situations, which could have multiple outcomes. Whether it rains on a daily basis is an experiment.
- **Outcome** is the result of a single trial. So, if it rains today, the outcome of today's trial from the experiment is "It rained"
- **Event** is one or more outcome from an experiment. "It rained" is one of the possible event for this experiment.

Formula of probability is: **P(A) = n(E)/n(S)**
Where,

- P(A) is the probability of an event "A"
- n(E) is the number of favorable outcomes
- n(S) is the total number of events in the sample space.

Basic Probability Formulas:
Let A and B are two events. The probability formulas are listed below:

All Probability Formulas List in Maths	
Probability Range	$0 \leq P(A) \leq 1$
Rule of Addition	$P(A \cup B) = P(A) + P(B) - P(A \cap B)$
Rule of Complementary Events	$P(A') + P(A) = 1$
Disjoint Events	$P(A \cap B) = 0$
Independent Events	$P(A \cap B) = P(A) \cdot P(B)$

All Probability Formulas List in Maths	
Conditional Probability	$P(A \mid B) = P(A \cap B) / P(B)$
Bayes Formula	$P(A \mid B) = P(B \mid A) \cdot P(A) / P(B)$

Q.16) What is discrete continuous random variable?

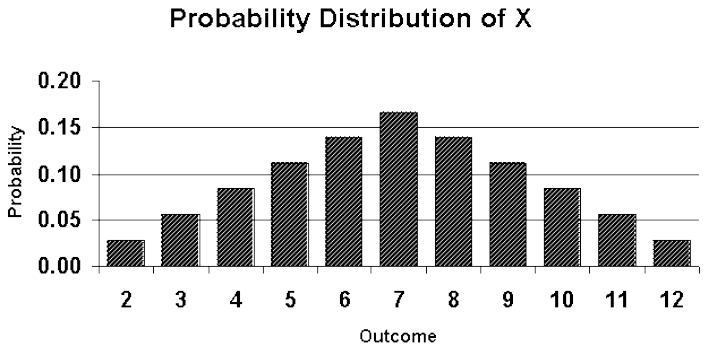
A.16) Discrete variable is a variable whose value is obtained by counting and it is measured by PMF.
Examples: number of student's present, number of red marbles in a jar, number of heads when flipping three coins
A continuous variable is a variable whose value is obtained by measuring and it is measured by PDF
Examples: height of students in class, weight of students in class, time it takes to get to school, distance traveled between classes.
A random variable is a variable whose value is a numerical outcome of a random phenomenon.

- A random variable is denoted with a capital letter
- The probability distribution of a random variable X tells what the possible values of X are and how probabilities are assigned to those values
- A random variable can be discrete or continuous

A discrete random variable X has a countable number of possible values.
Example: Let X represent the sum of two dice. Then the probability distribution of X is as follows:

X	2	3	4	5	6	7	8	9	10	11	12
P(X)	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

To graph the probability distribution of a discrete random variable, construct a probability histogram.



A continuous random variable X takes all values in a given interval of numbers.

- The probability distribution of a continuous random variable is shown by a density curve.

- The probability that X is between an intervals of numbers is the area under the density curve between the interval endpoints.
- The probability that a continuous random variable X is exactly equal to a number is zero.

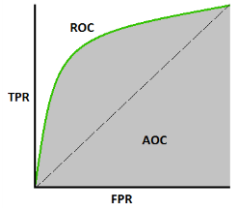
Q.17) What is continuous variable?

A.17) A continuous variable is a variable whose value is obtained by measuring. Continuous variables are also considered metric or quantitative variables, where the variable can have an infinite number or value between two given points. Continuous variables are often measured in infinitely small units.

Examples: Height of students in class
Weight of students in class
Time it takes to get to school
Distance traveled between classes.

Q.18) What is area under the curve?

A.18) An ROC curve (receiver operating characteristic curve) is a graph showing the performance of a classification model at all classification thresholds. ROC is a probability curve and AUC represent degree or measure of separability. Higher the AUC, better the model is at predicting 0s as 0s and 1s as 1s.



This curve plots two parameters:

- True Positive Rate
- False Positive Rate

True Positive Rate (**TPR**)/ **Recall /Sensitivity** is a synonym for recall and is therefore defined as follows:

$$\text{TPR /Recall / Sensitivity} = \frac{TP}{TP + FN}$$

Sensitivity:

$$\text{Specificity} = \frac{TN}{TN + FP}$$

False Positive Rate (**FPR**) is defined as follows:

$$\text{FPR} = 1 - \text{Specificity}$$

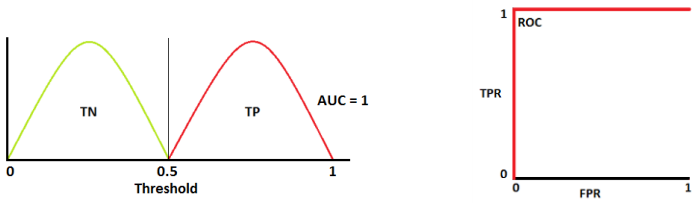
$$= \frac{FP}{TN + FP}$$

How to speculate the performance of the model?

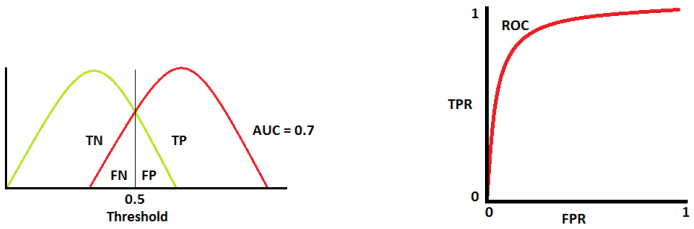
An excellent model has AUC near to the 1 which means it has good measure of separability. A poor model has AUC near to the 0 which means it has worst measure of separability. In fact it means it is reciprocating the result. It is predicting 0s as 1s and 1s as 0s. And when AUC is 0.5, it means model has no class

separation capacity whatsoever. As we know, ROC is a curve of probability. So let's plot the distributions of those probabilities:

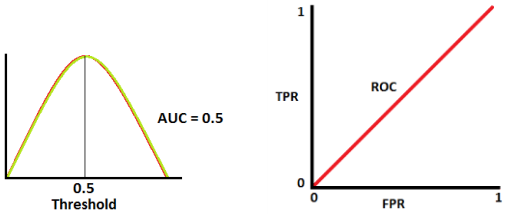
Note: Red distribution curve is of the positive class (patients with disease) and green distribution curve is of negative class (patients with no disease).



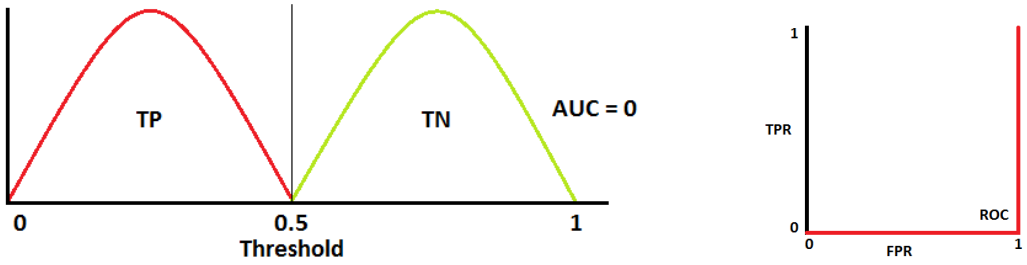
This is an ideal situation. When two curves don't overlap at all means model has an ideal measure of separability. It is perfectly able to distinguish between positive class and negative class.



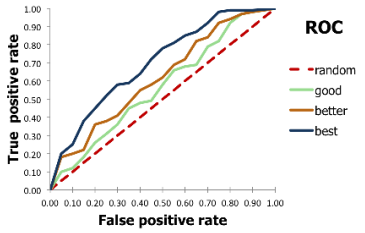
When two distributions overlap, we introduce type 1 and type 2 error. Depending upon the threshold, we can minimize or maximize them. When AUC is 0.7, it means there is 70% chance that model will be able to distinguish between positive class and negative class.



This is the worst situation. When AUC is approximately 0.5, model has no discrimination capacity to distinguish between positive class and negative class.



When AUC is approximately 0, model is actually reciprocating the classes. It means, model is predicting negative class as a positive class and vice versa.



Relation between Sensitivity, Specificity, FPR and Threshold:

Sensitivity and Specificity are inversely proportional to each other. So when we increase Sensitivity, Specificity decreases and vice versa. When we decrease the threshold, we get more positive values thus it increases the sensitivity and decreasing the specificity. Similarly, when we increase the threshold, we get more negative values thus we get higher specificity and lower sensitivity. As we know FPR is 1 - specificity. So when we increase TPR, FPR also increases and vice versa.

How to use AUC ROC curve for multi-class model?

In multi-class model, we can plot N number of AUC ROC Curves for N number classes using One vs ALL methodology. So for Example, If you have **three** classes named **X, Y** and **Z**, you will have one ROC for X classified against Y and Z, another ROC for Y classified against X and Z, and a third one of Z classified against Y and X.

Q.19) What is the necessity for integration in continuous variables?

A.19) Continuous integration is a coding philosophy and set of practices that drive development teams to implement small changes and check in code to version control repositories frequently. Because most modern applications require developing code in different platforms and tools, the team needs a mechanism to integrate and validate its changes. Computationally, to go from discrete to continuous we simply replace sums by integrals. It will help you to keep in mind that (informally) an integral is just a continuous sum.

Example 1. Since time is continuous, the amount of time Jon is early (or late) for class is a continuous random variable. Let's go over this example in some detail. Suppose you measure how early Jon arrives to class each day (in units of minutes). That is, the outcome of one trial in our experiment is a time in minutes. We'll assume there are random fluctuations in the exact time he shows up. Since in principle Jon could arrive, say,3.43 minutes early, or 2.7 minutes late (corresponding to the outcome -2.7), or at any other time, the sample space consists of all real numbers. So the random variable which gives the outcome itself has a continuous range of possible values.

Q.20) What is PMF?

A.20) Probability mass function (PMF) is a function that gives the probability that a discrete random variable is exactly equal to some value. Sometimes it is also known as the discrete density function. The probability mass function is often the primary means of defining a discrete probability distribution, and such functions exist for either scalar or multivariate random variables whose domain is discrete. A probability mass function differs from a probability density function (PDF) in that the latter is associated with continuous rather than discrete random variables. A PDF must be integrated over an interval to yield a probability. The value of the random variable having the largest probability mass is called the mode.

$$p(x) = P(X = x)$$

In probability and statistics, a probability mass function is a function that gives the probability that a discrete random variable is exactly equal to some value. Sometimes it is also known as the discrete density function

> Let X be a discrete random variable with range $R_X = \{x_1, x_2, x_3, \dots\}$ (finite or countably infinite). The function
>
> $$P_X(x_k) = P(X = x_k), \text{ for } k = 1, 2, 3, \dots,$$
>
> is called the *probability mass function (PMF)* of X.

A Probability Mass Function is also termed as a frequency function and is a vital part of statistics. Probability Mass Function integrates that any given variable has the probability that the random number will be equal to that variable. All the probabilities for the given discrete random variables provided by Probability Mass Function. Here discrete essentially means that there are a set number of outcomes for the variables. For understanding discrete variables better, the set number of outcomes in a die can only be 1, 2, 3, 4, 5 or 6. Here a discrete random value when considering a die is a set of random variables which are finite.

Q.21) What is PDF ?

A.21) A Probability Density Function (PDF), or density of a continuous random variable, is a function whose value at any given sample (or point) in the sample space (the set of possible values taken by the random variable) can be interpreted as providing a relative likelihood that the value of the random variable would equal that sample. Probability density function (PDF) is a statistical expression that defines a probability distribution (the likelihood of an outcome) for a discrete random variable (e.g., a stock or ETF) as opposed to a continuous random variable. The difference between a discrete random variable is that you can identify an exact value of the variable.

For instance, the value for the variable, e.g., a stock price, only goes two decimal points beyond the decimal (e.g. 52.55), while a continuous variable could have an infinite number of values (e.g. 52.5572389658...).When the PDF is graphically portrayed, the area under the curve will indicate the interval in which the variable will fall. The total area in this interval of the graph equals the probability of a discrete random variable occurring. More precisely, since the absolute likelihood of a continuous random variable taking on any specific value is zero due to the infinite set of possible values available, the value of a PDF can be used to determine the likelihood of a random variable falling within a specific range of values.

KEY TAKEAWAYS:

- Probability Density Functions are a statistical measure used to gauge the likely outcome of a discrete value, e.g., the price of a stock or ETF.
- PDFs are plotted on a graph typically resembling a bell curve, with the probability of the outcomes lying below the curve.
- A discrete variable can be measured exactly, while a continuous variable can have infinite values.
- PDFs can be used to gauge the potential risk/reward of including a particular security/fund in a portfolio.

The Basics of Probability Density Functions (PDFs):

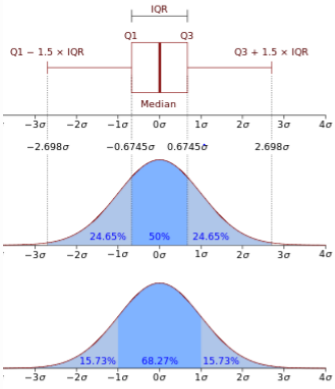
PDFs are used to gauge the risk of a particular security, such as an individual stock or ETF. They are typically depicted on a graph, with a normal bell curve indicating neutral market risk, and a bell at either end indicating greater or lesser risk/reward. A bell at the right side of the curve suggests greater reward, but with lesser likelihood, while a bell on the left indicates lower risk and lower reward. Investors should use PDFs as one of many tools to calculate the overall risk/reward in play in their portfolios.

An Example of a Probability Density Function (PDF):

As indicated previously, PDFs are a visual tool depicted on a graph based on historical data. A neutral PDF is the most common visualization, where risk is equal to reward across a spectrum. Someone willing to take limited risk will only be looking to expect a limited return and would fall on the left side of the bell curve below. An investor willing to take higher risk looking for higher rewards would be on the right side of the bell curve. Most of us, looking for average returns and average risk would be at the center of the bell curve.

$$\int p(x)\,dx=1$$

For the probability mass function, we have seen that the sum of the probabilities has to be equal to 1. This is not the case for probability density functions since the probability corresponds to the area under the curve and not directly to y values. However, this means that **the area under the curve has to be equal to 1**.
We saw in the last example, that the area was actually equal to 1. It can be easily obtained and visualized because of the squared shape of the uniform distribution. It is thus possible to multiply the height by the width: $2\times0.5=12\times0.5=1$.
However, in many cases, the shape is not a square and we still need to calculate the area under the curve.



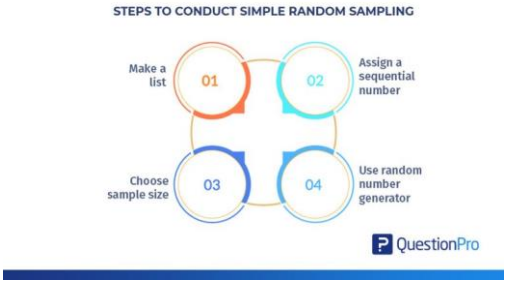
Boxplot and probability density function of a normal distribution $N(0, \sigma^2)$.

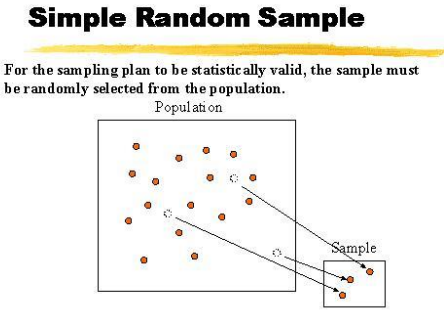
Q.22) What is CMF?

A.22) (CMF) is sometimes shortened as "distribution function", its $F(x) = Pr(X\leq x)$ the definition is the same for both discrete and continuous random variables. In dice case it's probability that the outcome of your roll will be x or smaller.

Q.23) What is sampling?
A.23) Collecting information about an entire large population costs too much. Instead we use a sample of population which is subset of entire data set. The process which provides a means of gaining information about the population without the need to examine the population in its entirety called Sampling. Statisticians attempt for the samples to represent the population in question. Two advantages of sampling are lower cost and faster data collection than measuring the entire population.

Q.24) What is random sampling?
A.24) A random sample is a subset of a statistical population in which each member of the subset has an equal probability of being chosen. A random sample is meant to be an unbiased representation of a group. An example of a random sample would be the names of 25 employees being chosen out of a hat from a company of 250 employees. In this case, the population is all 250 employees, and the sample is random because each employee has an equal chance of being chosen. Random sampling is used in science to conduct randomized control tests or for blinded experiments.





sample () is an inbuilt function of random module in Python that returns a particular length list of items chosen from the sequence i.e. list, tuple, string or set. Used for random sampling without replacement.

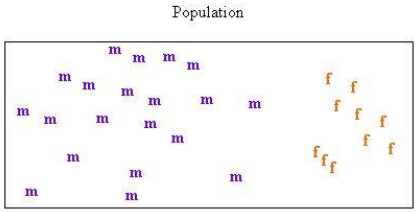
Syntax: random.sample(sequence, k)

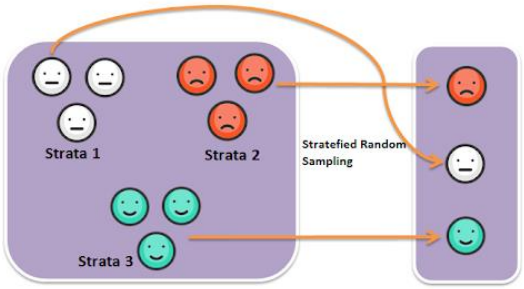
Q.25) What is stratified sampling ?
A.25) Stratified sampling is a type of sampling method in which the total population is divided into smaller groups or strata to complete the sampling process. The strata are formed based on some common characteristics in the population data. After dividing the population into strata, the researcher randomly selects the sample proportionally.

Description: Stratified sampling is a common sampling technique used by researchers when trying to draw conclusions from different sub-groups or strata. The strata or sub-groups should be different and the data should not overlap. While using stratified sampling, the researcher should use simple probability sampling. The population is divided into various subgroups such as age, gender, nationality, job profile, educational level etc. Stratified sampling is used when the researcher wants to understand the existing relationship between two groups.

Stratified Random Sample

When the population consists of a mixture of more than one element, stratified random sampling can assure that the sample is representative of the population.





Q.26) What are random variables ?

A.26) A random variable is a variable which contains numeric values i.e., it can take on different values and each value of a random variable have a probability associated with it. It is denoted with 'x'. Random variable is divided into two types: Discrete and Continuous

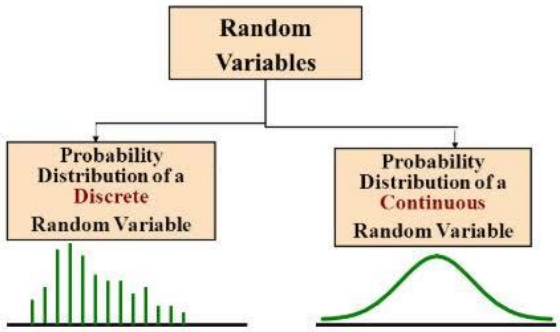
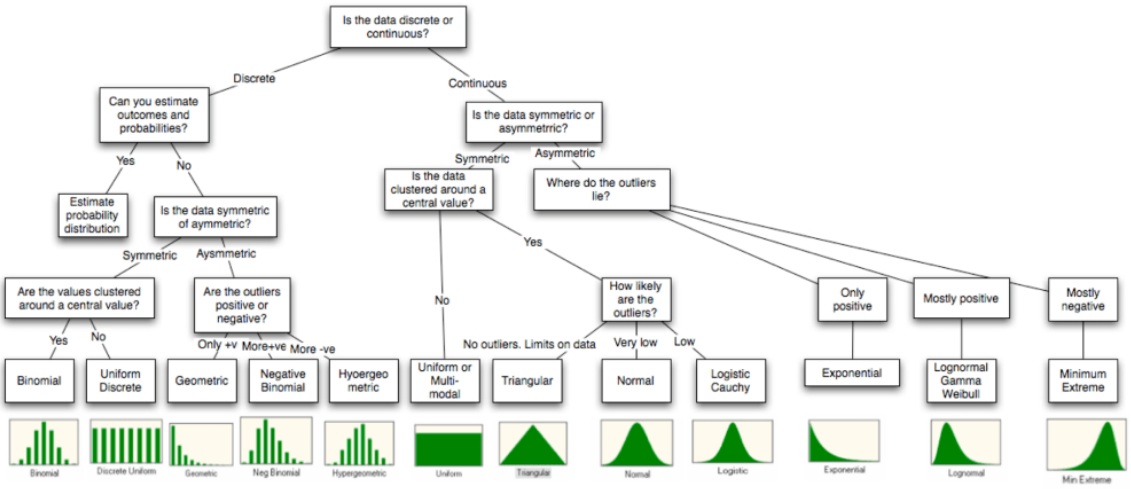


Figure 6A.15: Distributional Choices



Q.27) What is central limit theorem?

A.27) Central limit theorem states that when we add large number of independent random variables, irrespective of the original distribution of these variables, their normalized sum tends towards a Gaussian distribution. The central limit theorem states that the sampling distribution of the mean approaches a normal distribution, as the sample size increases. This fact holds especially true for sample sizes over 30.

Therefore, as a sample size increases, the sample mean and standard deviation will be closer in value to the population mean and standard deviation. A sufficiently large sample can predict the parameters of a population such as the mean and standard deviation.

Aspect 1: The sample distribution will be less spread than the population from which it is drawn.

Aspect 2: The sample distribution will be well-modeled by a normal distribution (bell shape). if we take larger samples take larger samples the distribution gets closer in shape to normal distribution.

Aspect 3: The spread of the sampling distribution is related to the spread of the values in the population

Aspect 4: Bigger samples lead to a smaller spread in the sampling distribution.

The spread reduces as the sample size increases. The spread of sampling distribution is related to the sqrt of the sample size (n) $s/\sqrt{n}$.

Q.28) What is coefficient of variation?

A.28) The coefficient of variation (CV) is a statistical measure of the dispersion of data points in a data series around the mean. The coefficient of variation represents the ratio of the standard deviation to the mean, and it is a useful statistic for comparing the degree of variation from one data series to another, even if the means are drastically different from one another.

- The coefficient of variation shows the extent of variability of data in a sample in relation to the mean of the population. In finance, the coefficient of variation allows investors to determine how much volatility, or risk, is assumed in comparison to the amount of return expected from investments. Ideally, the coefficient of variation formula should result in a lower ratio of the standard deviation to mean return, meaning the better risk-return trade-off. Note that if the expected return in the denominator is negative or zero, the coefficient of variation could be misleading.
- The coefficient of variation is helpful when using the risk/reward ratio to select investments. For example, an investor who is risk-averse may want to consider assets with a historically low degree of volatility and a high degree of return, in relation to the overall market or its industry. Conversely, risk-seeking investors may look to invest in assets with a historically high degree of volatility.
- The lower the ratio of the standard deviation to mean return, the better risk-return trade-off.
- **Coefficient of Variation Formula**
- **Population Cv:**

$CV = \sigma/\mu$ where: σ=standard deviation

μ=mean

- **Sample Cv:**

  Cv=s/x where: s=sample standard deviation
  x= sample mean.

Q.29) What is covariance?

A.29) In mathematics and statistics, covariance is a measure of the relationship between two random variables. The metric evaluates how much – to what extent – the variables change together. In other words, it is essentially a measure of the variance between two variables. However, the metric does not assess the dependency between variables.

Unlike the correlation coefficient, covariance is measured in units. The units are computed by multiplying the units of the two variables. The variance can take any positive or negative values. The values are interpreted as follows:

Positive covariance: Indicates that two variables tend to move in the same direction.

Negative covariance: Reveals that two variables tend to move in inverse directions.

In finance, the concept is primarily used in portfolio theory. One of its most common applications in portfolio theory is the diversification method, using the covariance between assets in a portfolio. By choosing assets that do not exhibit a high positive covariance with each other, the unsystematic risk can be partially eliminated.

Formula for Covariance:

Sample Covariance:

$$\text{Cov}(X, Y) = \frac{\sum(X_i - \bar{X})(Y_j - \bar{Y})}{n - 1}$$

Population Covariance:

Cov(X,Y)= Σ(Xi- μx)(Yj- μy)/N

Q.30) What is Pearson Correlation?

A.30) Pearson's Correlation Coefficient helps you find out the relationship between two quantities. It gives you the measure of the strength of association between two variables.

- The value of Correlation will always lie between 1 and -1
- Correlation=0, it means there is absolutely **no** relationship between the selected feature value and the target value.
- Correlation=1, it means that there is a **perfect** relationship between the selected feature value and the target value and this would mean that the selected feature is appropriate for our model to learn.
- Correlation= -1, it means that there exists a **negative** relationship between the selected feature value and the target value, generally, the use of the feature value having a negative value of low magnitude is discouraged for e.g. -0.1 0r -0.2.

Mathematically **Pearson's correlation** is calculated as:

$$r = \frac{\sum(X-\overline{X})(Y-\overline{Y})}{\sqrt{\sum(X-\overline{X})^2}\sqrt{(Y-\overline{Y})^2}}$$

Where, $\overline{X}$ = mean of X variable

$\overline{Y}$ = mean of Y variable

We generally store the feature values in X and target value in the Y. The formula written above will tell us whether there exists any correlation between the selected feature value and the target value.

Q.31) What is spearman rank correlation?

A.31) The Spearman rank correlation coefficient, r_s, is the nonparametric version of the Pearson correlation coefficient. Your data must be ordinal, interval or ratio. Spearman's returns a value from -1 to 1,where:+1=a +ve relation between ranks

-1=a perfect negative correlation between ranks

0 = no correlation between ranks.

Use Spearman rank correlation when you have two ranked variables, and you want to see whether the two variables covary; whether, as one variable increases, the other variable tends to increase or decrease. You also use Spearman rank correlation if you have one measurement variable and one ranked variable; in this case, you convert the measurement variable to ranks and use Spearman rank correlation on the two sets of ranks.

The formula for the Spearman rank correlation coefficient when there are no tied ranks is:

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$$

Q.32) What is the difference between covariance and correlation?

A.32) Covariance: It is used to determine how much two random variables vary together and by how much.

A covariance of '0' indicates the variables are totally unrelated. If covariance is positive the variable increases in same direction, if negative the variable change in opposite direction.

Correlation: It is used to determine when change in one variable can result a change in another.

Key Differences between Covariance and Correlation

1. A measure used to indicate the extent to which two random variables change in tandem is known as covariance. A measure used to represent how strongly two random variables are related known as correlation.
2. Covariance is nothing but a measure of correlation. On the contrary, correlation refers to the scaled form of covariance. Correlation is obtained by dividing the covariance of the two variables by the product of their standard deviations.
3. The value of correlation takes place between -1 and +1. Conversely, the value of covariance lies between $-\infty$ and $+\infty$.
4. Covariance is affected by the change in scale, i.e. if all the value of one variable is multiplied by a constant and all the value of another variable are multiplied, by a similar or different constant, then the covariance is changed. As against this, correlation is not influenced by the change in scale.

5. Correlation is dimensionless, i.e. it is a unit-free measure of the relationship between variables. Unlike covariance, where the value is obtained by the product of the units of the two variables.

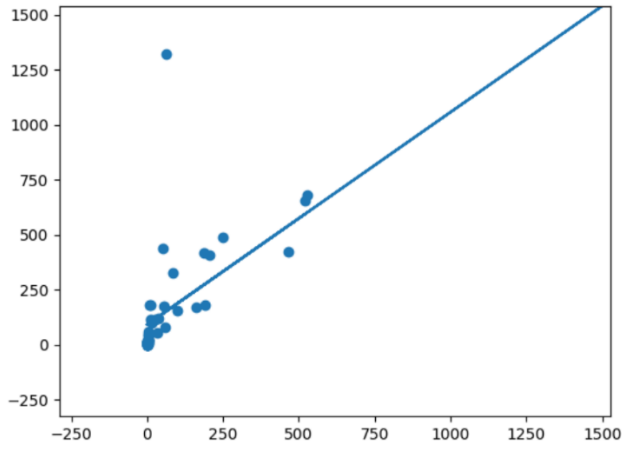
Q.33) How to interpret correlation relation ?
A.33) Correlation coefficient is a single number that measures both the strength and direction of the linear relationship between two continuous variables. Values can range from -1 to +1.

Data correlation is the way in which one set of data may correspond to another set. In ML, think of how your features correspond with your output.

For example, the image below visualizes a dataset of brain size versus body size. Notice that as the body size increases, so does the brain size. This is known as a linear correlation. In a simple definition of linear correlation, the data follows a straight line.

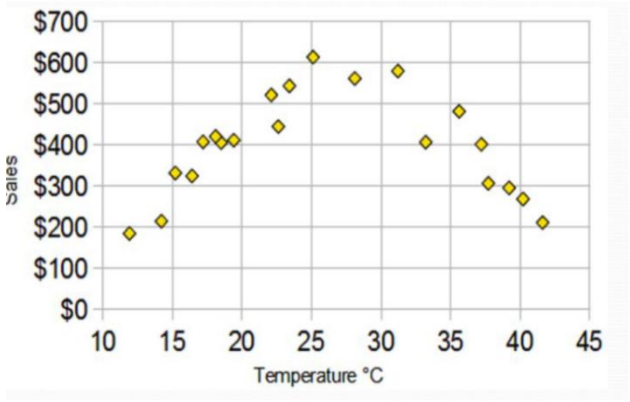
Not all data is linearly correlated. The following image shows a curve of ice cream sales versus temperature. It has an inverted U shaped graph.It could mean that if it's hot enough, people might not want to leave their homes to go buy ice cream. Or there could be another reason. So using Linear Correlation doesn't make sense for this dataset.

Not all data is linearly correlated. The following image shows a curve of ice cream sales versus temperature. It has an inverted U shaped graph.It could mean that if it's hot enough, people might not want to leave their homes to go buy ice cream. Or there could be another reason. So using Linear Correlation doesn't make sense for this dataset.



Brain Weight Vs. Body Weight

Not all data is linearly correlated. The following image shows a curve of ice cream sales versus temperature. It has an inverted U shaped graph. It could mean that if it's hot enough, people might not want to leave their homes to go buy ice cream. Or there could be another reason. So using Linear Correlation doesn't make sense for this dataset.



Ice Cream Sales vs. Temperature

It becomes very hard to figure out how data correlates if you have more than two features. Data visualization can help find how individual features may correlate with the output.
For example the below figure explains:

```
import pandas as pd
import numpy as np

rs = np.random.RandomState(0)
df = pd.DataFrame(rs.rand(10, 10))
corr = df.corr()
corr.style.background_gradient(cmap='coolwarm')
# 'RdBu_r' & 'BrBG' are other good diverging colormaps
```

	0	1	2	3	4	5	6	7	8	9
0	1	0.347533	0.398948	0.455743	0.0729144	-0.233402	-0.731222	0.477978	-0.442621	0.0151847
1	0.347533	1	-0.284056	0.571003	-0.285483	0.38248	-0.362842	0.642578	0.252556	0.190047
2	0.398948	-0.284056	1	-0.523649	0.152937	-0.139176	-0.0928948	0.0162655	-0.434016	-0.383585
3	0.455743	0.571003	-0.523649	1	-0.225343	-0.227577	-0.481548	0.473286	0.279258	0.44665
4	0.0729144	-0.285483	0.152937	-0.225343	1	-0.104438	-0.147477	-0.523283	-0.614603	-0.189916
5	-0.233402	0.38248	-0.139176	-0.227577	-0.104438	1	-0.0302517	0.41764	0.205851	0.0950844
6	-0.731222	-0.362842	-0.0928948	-0.481548	-0.147477	-0.0302517	1	-0.49444	0.381407	-0.353652
7	0.477978	0.642578	0.0162655	0.473286	-0.523283	0.41764	-0.49444	1	0.375873	0.417863
8	-0.442621	0.252556	-0.434016	0.279258	-0.614603	0.205851	0.381407	0.375873	1	0.150421
9	0.0151847	0.190047	-0.383585	0.44665	-0.189916	0.0950844	-0.353652	0.417863	0.150421	1

Note that if you have a large dataset and if you get a small coefficient, say 0.4, then it's not necessarily bad. The dataset might have a large statistically significant correlation.
Also note that correlation may not mean causation.

Because two variables are related, does not mean that one directly caused the other. In Titanic dataset, people aboard the Titanic did not die because they were male and aged 28. Rather, a large number of them died because officers were saving "Women and Children First".

Q.34) What is the best way to test a strategy, promotion, campaign?
A.34)

Q.35) What hypothesis testing?

A.35) First let us understand what a hypothesis is. A statistical hypothesis is an assumption about a population parameter. A parameter is a measurable characteristic of a population or a sample, like mean, standard deviation, etc., this assumption may or may not be true.

Hypothesis testing refers to the formal procedures used by statisticians to accept or reject the statistical hypothesis. In simple terms, it is a logical statement which derives True or False. The statement needs to be proven based on the historical data.

The following **Hypothesis Testing Procedure** is followed to test the assumption made:

1. Set up a Hypothesis
2. Set up a suitable Significance Level
3. Determining a suitable Test Statistic
4. Determining the Critical Region
5. Performing computations
6. Decision-making

Q.36) **Types of hypothesis testing?**

A.36)There are 2 types of hypothesis:

- Null Hypothesis: There is no effect or relationship between the variables. This is denoted by $\mathbf{H_0}$.
- Alternative Hypothesis: Effect or relationship exists. This is denoted by $\mathbf{H_1}$ or $\mathbf{H_a}$.

We assume that the null hypothesis is correct until we have enough evidence to suggest otherwise.

For example, suppose we wanted to determine whether a coin was fair and balanced. A null hypothesis might be that half the flips would result in Heads and half, in Tails. The alternative hypothesis might be that the number of Heads and Tails would be very different. Symbolically, these hypotheses would be expressed as

H_0:	P	=	0.5
H_a:	P	≠	0.5

E.g. 36:1

Suppose we flipped the coin 50 times, resulting in 40 Heads and 10 Tails. Given this result, we would be inclined to reject the null hypothesis. We would conclude, based on the evidence, that the coin was probably not fair and balanced.

Can We Accept the Null Hypothesis?

Some researchers say that a hypothesis test can have one of two outcomes: you accept the null hypothesis, or you reject the null hypothesis. Many statisticians, however, take issue with the notion of "accepting the null hypothesis." Instead, they say: you reject the null hypothesis, or you fail to reject the null hypothesis.

Why the distinction between "acceptance" and "failure to reject?" Acceptance implies that the null hypothesis is true. Failure to reject implies that the data are not sufficiently persuasive for us to prefer the alternative hypothesis over the null hypothesis.

In hypothesis testing there are 2 types of errors:

- Type I: Rejection of a true Null Hypothesis (False Positive). The probability of committing this type of error is called Significance level or alpha. This is denoted by α.
- Type II: Non-Rejection of a false Null Hypothesis (False Negative). The probability of committing this type of error is called beta. This is denoted by β.

One-Tailed and Two-Tailed Tests:

A test of a statistical hypothesis,

- Where the region of rejection is on only one side of the sampling distribution, is called a **one-tailed test**.
  For example, the claim is, Average earning is more than 20$
  H_0: μ (mean) > 20$
  H_a: μ (mean) <= 20$
  E.g. ***36:2***
  Here, the region of rejection would consist of a range of numbers located on the right side of sampling distribution; i.e.; a set of earning greater than 20$.
- Where the region of rejection is on both sides of the sampling distribution, is called a **two-tailed test**. For example, the claim is, Average earning is 20$.
  H_0: μ (mean) = 20$
  H_a: μ (mean) ≠ 20$
  E.g. ***36:3***
  Here, the region of rejection would consist of a range of earning located on both sides of the sampling distribution; i.e.; the region of rejection would consist partly of numbers that were less than 20$ and partly of numbers that were greater than 20$.

Confusion matrix: Once we are done with model building and deploy this model to Test Data. This matrix will help us to identify the model accuracy.
Let's take a scenario where we say customer paid an EMI or not
Size of the confusion matrix depends on how many values we want to predict for example if we want to predict 2 values (customer paid an EMI, Not paid an EMI) hence matrix will be 2*2, or we can say for n values confusion matrix will be n*n

	Actual Values	
Predicted values	**TP**	**FP**
	FN	**TN**

- Rows Indicates predicted values (our model)
- Column indicates Actual values (test data)

We will use these values for model selection in case we wants to test our data against multiple algorithms ((Logistic regression, KNN, Random forest) and compare the values.

Q.37) What is false positive**?**
A.37) Where model predicted wrong for positive scenario, other words in actual customer did not paid an EMI but we predicted he has paid.

Q.38) What is false negative?
A.39) Where model predicted wrong for Negative scenario, other words in actual customer paid an EMI but our model predicted he did not.

Q.39) What is true positive?
A.39) Where our model predicted correct values for Positive scenario, other words in actual customer paid an EMI and our model said yes, he has paid.

Q.40) What is true negative?

A.40) Where our model predicted correct values for Negative scenario, other words in actual customer did not paid an EMI and our model also said he did not.

n=165	**Predicted: NO**	**Predicted: YES**	
Actual: NO	TN = 50	FP = 10	60
Actual: YES	FN = 5	TP = 100	105
	55	110	

This is a list of rates that are often computed from a confusion matrix for a binary classifier:

- **Accuracy:** Overall, how often is the classifier correct?
  - (TP+TN)/total = (100+50)/165 = 0.91
- **Misclassification Rate:** Overall, how often is it wrong?
  - (FP+FN)/total = (10+5)/165 = 0.09
  - equivalent to 1 minus Accuracy
  - also known as "Error Rate"
- **True Positive Rate:** When it's actually yes, how often does it predict yes?
  - TP/actual yes = 100/105 = 0.95
  - also known as "Sensitivity" or "Recall"
- **False Positive Rate:** When it's actually no, how often does it predict yes?
  - FP/actual no = 10/60 = 0.17
- **True Negative Rate:** When it's actually no, how often does it predict no?
  - TN/actual no = 50/60 = 0.83
  - equivalent to 1 minus False Positive Rate
  - also known as "Specificity"
- **Precision:** When it predicts yes, how often is it correct?
  - TP/predicted yes = 100/110 = 0.91
- **Prevalence:** How often does the yes condition actually occur in our sample?
  - actual yes/total = 105/165 = 0.64
- **ROC Curve:**
  A Receiver Operator Characteristic (ROC) curve is a graphical plot used to show the diagnostic ability of binary classifiers.
- This is a commonly used graph that summarizes the performance of a classifier over all possible thresholds. It is generated by plotting the True Positive Rate (y-axis) against the False Positive Rate (x-axis) as you vary the threshold for assigning observations to a given class.

Q.41) What is accuracy score?

A.41) Accuracy is one metric for evaluating classification models. Informally, accuracy is the fraction of predictions our model got right. Accuracy is the most intuitive performance measure and it is simply a ratio of correctly predicted observation to the total observations.

One may think that, if we have high accuracy then our model is best. Yes, accuracy is a great measure but only when you have symmetric datasets where values of false positive and false negatives are almost same. Therefore, you have to look at other parameters to evaluate the performance of your model.

Formally, accuracy has the following definition:

Accuracy = Number of correct predictions/Total number of predictions

Classification Accuracy is the number of correct predictions made divided by the total number of predictions made, multiplied by 100 to turn it into a percentage.

For binary classification, accuracy can also be calculated in terms of positives and negatives as follows:
Accuracy=TP+TN/TP+TN+FP+FN

Q.42) What is type 1 error?

A.42) Hypothesis Testing helps in determining if a specific treatment has the effect on the population. There are 5 steps involved in any Hypothesis testing.

- State the hypothesis. (both H_O – Null Hypothesis and H_A – Alternative Hypothesis)
- Set up the suitable significant level i.e. $\alpha = 0.05 \; or \; 0.01$
- Setting a test criterion (i.e. Z-test, t-test, f-test, Chi-square test)
- Doing computation
- Making Decision

- During decision making 2 types of error may occurs.
- Hypothesis test depends on sample data, there is a chance of misleading data will cause the hypothesis test to conclude a wrong answer.
- Those 2 types of errors are:
  - ✓ Type-I error
  - ✓ Type – II error

		Actual Situation	
		No Effect, H_0 True	Effect Exists, H_0 False
EXPERIMENTER'S DECISION	Reject H_0	Type I error	Decision correct
	Retain H_0	Decision correct	Type II error

Based on our sample if we reject the true null hypothesis then it is called as Type-I error.

When the Null hypothesis is false and we fail to reject this hypothesis, then it is known as Type-II error.

A type 1 error is also known as false positive and occurs when we incorrectly reject a true null hypothesis.

This means that we report that our findings are significant when in fact they have occurred by chance.

The probability of making a type I error is represented by your alpha level (α), which is the p-value below which you reject the null hypothesis. A p-value of 0.05 indicates that you are willing to accept a 5% chance that you are wrong when you reject the null hypothesis.

You can reduce your risk of committing a type I error by using a lower value for p. For example, a p-value of 0.01 would mean there is a 1% chance of committing a Type I error.

However, using a lower value for alpha means that you will be less likely to detect a true difference if one really exists (thus risking a type II error).

Q.43) Why type 1 error is critical?

A.43) Type 1 error control is more important than Type 2 error control, because inflating Type 1 errors will very quickly leave you with evidence that is too weak to be convincing support for your hypothesis, while inflating Type 2 errors will do so more slowly.

- A Type I error, on the other hand, is an error in every sense of the word. A conclusion is drawn that the null hypothesis is false when, in fact, it is true. Therefore, Type I errors are generally considered more serious than Type II errors.
- The probability of a Type I error (α) is called the significance level and is set by the experimenter. There is a trade-off between Type I and Type II errors. The more an experimenter protects himself or herself against Type I errors by choosing a low level, the greater the chance of a Type II error.
- Requiring very strong evidence to reject the null hypothesis makes it very unlikely that a true null hypothesis will be rejected. However, it increases the chance that a false null hypothesis will not be rejected, thus lowering power.
- The Type I error rate is almost always set at .05 or at .01, the latter being more conservative since it requires stronger evidence to reject the null hypothesis at the .01 level then at the .05 level.

Q.44) Why should you perform hypothesis testing?

A.44) In data science, when we try to arrive at a conclusion/solution for problem statement, it is necessary that we make assumptions on the outcomes, which we term as hypothesis –null hypothesis and alternate hypothesis. Through hypothesis testing, we can conclude if the null hypothesis is significant statistically i.e., through repeated tests on different samples and eventually repeated hypothesis testing, we will be able to confirm if a certain hypothesis is significant or not. Also, through repeated tests and repeated hypothesis testing, we can assess the probability of a certain outcome.

Q.45) What are variable types ?

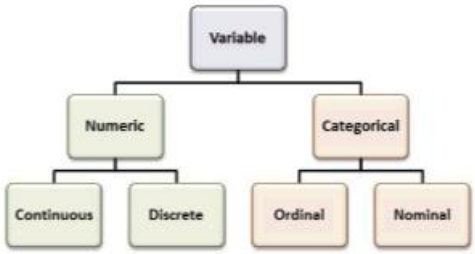
A.45) Variable: A variable is any characteristics, number, or quantity that can be measured or counted. A variable may also be called a data item.

Ex: Age, sex, business income and expenses, country of birth, capital expenditure, class grades...etc

Types of variables: Variables can be divided into different types major they are classified into two types

1. Numeric variables
2. Categorical variables

Flow chart:



Numeric Variable: Numeric variables have values that describe a measurable quantity as a number. Therefore, numeric variables are quantitative variables. Numeric variables classified into two types.

i. Discrete
ii. Continuous

i.Discrete variable: A discrete variable cannot take the value of a fraction between one value and the next closest value. It takes only Integer numbers.
Ex: Number of registered cars, number of business locations, and number of children in a family (i.e.,1,2,3,4...)

ii. Continuous variable: The value given to an observation for a continuous variable can include values as small as the instrument of measurement allows.
Ex: It measured continuous scale like height, weight, time, temperature (i.e., 2.3,
60.5, 3.54...)

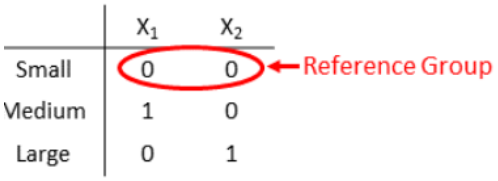
Categorical Variable: Categorical variables have place people into groups/categories. Therefore, categorical variables are qualitative variables and tend to be represented by a non-numeric value. Categorical variables may be further described as ordinal or nominal:

i. Ordinal
ii. Nominal

i. Ordinal variable: It can be logically ordered or ranked. The categories associated with ordinal variables can be ranked higher or lower than another, but do not necessarily establish a numeric difference between each category.
Ex: Academic grades (i.e. A, B, C), clothing size (i.e. small, medium, large, extra-large) and attitudes (i.e. strongly agree, agree, disagree, and strongly disagree).
ii. Nominal variable: Observations can take a value that is not able to be organized in a logical sequence.
Ex: Gender, business type, eye colour, religion and brand.

Q.46) What is class variable ?
A.46) Class variables − Class variables also known as static variables are declared with the static keyword in a class, but outside a method, constructor or a block. When space is allocated for an object in the heap, a slot for each instance variable value is created. The variables that are defined inside the class but outside the method can be accessed within the class (all methods included) using the instance of a class. For Example – self. var_name. If you want to use that variable even outside the class, you must declare that variable as a global.

Q.47) What are dummy variables ?
A.47) A dummy variables represents in regression analysis which 1 takes only the value 0 or 1. i.e., to indicate if its presence or absent. Of some categorical effect.



Q.48) What is label Encoders ?
A.48) In the Datasets few of the columns contain categorical data (like for the column "Country Name" the values can be as Spain, Germany, India etc). The machine cannot understand categorical data and the data has to be converted to the numerical data. For this Label Encoders is used.

- The categorical data like Spain, Germany and India is converted to the numerical data as 0, 1, and 2. So in the column "Country Name" the categorical data (Spain, Germany and India) gets replaced by the numerical data (0, 1 and 2) after encoding. This process is called Label Encoding.
- It is an important pre-processing step for the structured dataset in supervised learning. Label Encoding is implemented on the training dataset. Label Encoder is a utility class to help normalize labels such that they contain only values between 0 and n_classes-1 where n is the number of distinct labels.
- Scikit Learn Preprocessing library is used for encoding in data preprocessing steps.

```
# Import label encoder
from
sklearn import preprocessing
# label_encoder object knows how to understand word labels.
LEncoder=LabelEncoder()
# Encode labels in column
x.iloc[:,0]=lEncoder.fit_transform(x.iloc[:,0])
```

Disadvantages:

The Machine Learning models are based on the equations so when the Label Encoders are used and each of the categorical values like Spain, Germany, India gets converted to the numerical data 0,1,2. As 2 is greater than 0,1 the equation in the model thinks that India had greater value than Spain and Germany. To overcome this Dummy Variable method is used.

Q.49) What is contingency table ?
A.49) Contingency table is essentially a matrix that gives information on the frequency distribution of variables, usually bivariate or multivariate. The table represents the interrelation between variables. In python, contingency tables are represented by using the function crosstab.

- Contingency table is a part of Descriptive Statistics. This is also known as cross tabulation or crosstab or 2-way frequency table. This is used to summarize the relationship between 2 or more categorical variables by presenting the frequency distribution.

- Contingency tables are constructed by listing all the categories of one variable as rows in a table and the categories of the other variables as columns. Contingency table summarizes 3 probability distributions:

1. Joint Distribution
2. Marginal Distribution
3. Conditional Distribution

gender	cup	cone	sundae	sandwich	other	total
male	592	300	204	24	80	1200
female	410	335	180	20	55	1000
total	1002	635	384	44	135	2200

Rows (R) are categories of a variable gender
(C) are categories of a preferred way of eating ice creams
on of a row and a column of a contingency table is called a cell.

i. Joint Distribution: This is used to find the relationship between a category of two variables. The cells of the contingency table divided by the total provide the joint distribution.

E.g., the probability that a female person prefers ice cream in a cone is 335 / 2200 ≈ 15.23%.

ii. Marginal Distribution: This describes the distribution of the R (row) or C (column) variable alone. The row and column totals of the contingency table provide the marginal distribution.

E.g., the probability that a person prefers ice cream in a cup is 1002 / 2200 ≈45.54%, while the probability of random female participant is 1000 / 2200 ≈ 45.45%

iii. Conditional Distribution: This describes the distribution of one variable given the category of the other variable. The cells of the contingency table divided by the row or column totals provide the conditional distributions.

E.g., the probability that a person prefers ice cream sandwiches, given that the person is male is 24 / 1200 = 2%, while the conditional probability that a person is a male, given that ice cream sandwiches are preferred is 24/44 ≈ 54.54%

Note: Hypothesis tests on contingency may be performed based on a statistic called chi-square.

Q.50) What is linear and non-linear relation ?

A.50) Linear Relation : A linear relationship (or linear association) is a statistical term used to describe a straight-line relationship between a variable and a constant. Linear relationships can be expressed either in a graphical format where the variable and the constant are connected via a straight line or in a mathematical format where the independent variable is multiplied by the slope coefficient, added by a constant, which determines the dependent variable.

The Linear Equation Is: Mathematically, a linear relationship is one that satisfies the equation:

$y=mx+b$ where: m=slope , b=y-intercept

- An equation expressing a linear relationship can't consist of more than two variables, all of the variables in an equation must be to the first power, and the equation must graph as a straight line.
- A linear function in mathematics is one that satisfies the properties of additivity and homogeneity. Linear functions also observe the superposition principle, which states that the net output of two or more inputs equals the sum of the outputs of the individual inputs.
- A commonly used linear relationship is a *correlation,* which describes how one variable changes in a linear fashion to changes in another variable.

Example:
1) A linear relationship can also be found in the equation distance = rate x time.
2) In order to convert Celsius to Fahrenheit

$$C=5/9(F-32)$$

Nonlinear relation : Nonlinear relation in equations, such as conic sections, include at least one equation that is nonlinear. A nonlinear equation is defined as an equation possessing at least one term that is raised to a power of 2 or more. When graphed, these equations produce curved lines.

- Since at least one function has curvature, it is possible for nonlinear equations to contain multiple solutions. As with linear equations, substitution can be used to solve nonlinear relation for one variable and then the other.
- Solving nonlinear equations algebraically is similar to doing the same for linear of equations. However, subtraction of one equation from another can become impractical if the two equations have different terms, which is more commonly the case in nonlinear systems.

Example:
1>GPS signal
2>Earth orbit rotation
3>Radar signal

Differential view : Linear functions have a constant slope, so nonlinear functions have a slope that varies between points. Algebraically, linear functions are polynomials with highest exponent equal to 1 or of the form y = c where c is constant. Nonlinear functions are all other functions. An example of a nonlinear function is y = x^2.

In real life example, children are often prescribed in proportion to weight. This is an example of a linear relationship where as a drug may be ineffective up until a certain threshold and then become effective is a nonlinear relationship. Linear relation always follows below property but a nonlinear cannot:

a. Additivity: f(x + y) = f(x) + f(y)
b. Homogeneity of degree 1: f(αx) = αf(x) for all α

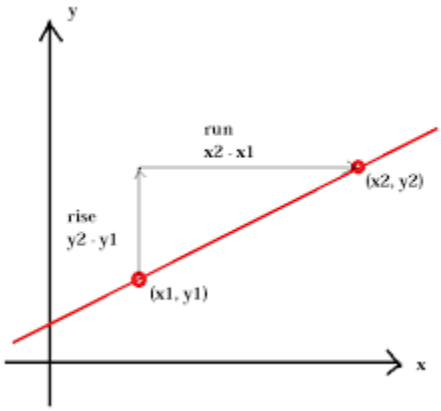
A linear relation as functions are those who can be brought to the form **‘ f(x) = ax + b ’** and but non-linear equations are those who can not be brought to the for form ‘**ax + b = 0 ’.**

The graph of a linear equation forms a straight line (all data points make a line as unidirectional), whereas the graph for a non-linear relationship is curved (all data points make a non-uniform line graph in linear relation, we can predict other points based on Y=MX+C, but in nonlinear, it is not predictable.

Graphical view -

Graphs :

Linear



Nonlinear