Statistics for Data Science - Interview Questions

1. **What is Inferential Statistics?**
   Inferential statistics is a statistical method that deduces from a small but representative sample the characteristics of a bigger population. In other words, it allows the researcher to make assumptions about a wider group, using a smaller portion of that group as a guideline.

2. **What is the difference between Population and Sample?**
   From the population we take a sample. We cannot work on the population either due to computational costs or due to availability of all data points for the population. From the sample we calculate the statistics. From the sample statistics we conclude about the population.

3. **What is the relationship between mean and median in normal distribution?**
   In the normal distribution mean is equal to median

4. **What is an outlier?**
   An outlier is an abnormal value (It is at an abnormal distance from the rest of the data points).

5. **What can I do with Outliers?**
   - **Remove outlier**
     When we know the data-point is wrong (negative age of a person)
     When we have lots of data
     We should provide two analyses. One with outliers and another without.
   - **Keep outlier**
     When there are lot of outliers (skewed data)
     When results are critical
     When outliers have meaning (fraud data)

6. **What is the difference between population parameters and sample statistics?**
   - Population parameters are:
     Mean = μ
     Standard deviation = σ
   - Sample statistics are:
     Mean = x (bar)
     Standard deviation = s

7. **What is the difference between inferential statistics and descriptive statistics?**
   - Descriptive statistics – provides exact and accurate information.

   - Inferential statistics – provides information of a sample and we need inferential statistics to reach to a conclusion about the population.

8. What is the difference between population and sample in inferential statistics?

- From the population we take a sample. We cannot work on the population either due to computational costs or due to availability of all data points for the population.
- From the sample we calculate the statistics
- From the sample statistics we conclude about the population

9. What are descriptive statistics?

- Descriptive statistic is used to describe the data (data properties)
- 5-number summary is the most commonly used descriptive statistics

10. Most common characteristics used in descriptive statistics?

- *Central Tendency*– middle of the data. Mean / Median / Mode are the most commonly used as measures.

Mean – average of all the numbers

Median – the number in the middle

Mode – the number that occurs the most. The disadvantage of using Mode is that there may be more than one mode.

- *Spread* – How the data is dispersed. Range / IQR / Standard Deviation / Variance are the most commonly used as measures.

Range = Max – Min

Inter Quartile Range (IQR) = Q3 – Q1

Standard Deviation (σ) = $\sqrt{(\sum(x-\mu)2 / n)}$

Variance = $\sigma 2$

- *Shape* – the shape of the data can be symmetric or skewed

Symmetric – the part of the distribution that is on the left side of the median is same as the part of the distribution that is on the right side of the median

Left skewed – the left tail is longer than the right side

Right skewed – the right tail is longer than the left side

- *Outlier* – An outlier is an abnormal value

  Keep the outlier based on judgement

  Remove the outlier based on judgement

11. What are 5 Point Summary (or 5 number summary)?

- Low extreme (minimum)
- Lower quartile (Q1)
- Median
- Upper quartile (Q3)
- Upper extreme (maximum)

12. What are the measures of central tendency? When to use which one?

Measures of central Tendency are – Mean, Median & Mode.

a. Mean is the most frequently used measure of central tendency and generally considered the best measure of it. However, there are some situations where either median or mode are preferred.

b. Median is the preferred measure of central tendency when:
   - There are a few extreme scores in the distribution of the data. (NOTE: Remember that a single outlier can have a great effect on the mean). b.
   - There are some missing or undetermined values in your data. c.
   - There is an open ended distribution (For example, if you have a data field which measures number of children and your options are 0, 1, 2, 3, 4, 5 or "6 or more," then the "6 or more field" is open ended and makes calculating the mean impossible, since we do not know exact values for this field) d.
   - You have data measured on an ordinal scale.

c. Mode is the preferred measure when data are measured in a nominal (and even sometimes ordinal) scale.

13. What is CLT? What is its significance?

"The central limit theorem states that the sampling distribution of the sample mean approaches a normal distribution as the sample size gets larger no matter what the shape of the population distribution."

The central limit theorem is important because it is used in hypothesis testing and also to calculate confidence intervals.

(Please follow for more reference: https://towardsdatascience.com/understanding-the-central-limit-theorem-e598158cc5da)

14. What is p-value?

The P value, or calculated probability, is the probability of finding the observed, or more extreme, results when the null hypothesis (H0) of a study question is true – the definition of 'extreme' depends on how the hypothesis is being tested. P is also described in terms of rejecting H0 when it is actually true, however, it is not a direct probability of this state.

When you perform a hypothesis test in statistics, a p-value helps you determine the significance of your results. Hypothesis tests are used to test the validity of a claim that is made about a population. This claim that's on trial, in essence, is called the null hypothesis.

The alternative hypothesis is the one you would believe if the null hypothesis is concluded to be untrue. The evidence in the trial is your data and the statistics that go along with it. All hypothesis tests ultimately use a p-value to weigh the strength of the evidence (what the data are telling you about the population). The p-value is a number between 0 and 1 and interpreted in the following way:

- A small p-value (typically ≤ 0.05) indicates strong evidence against the null hypothesis, so you reject the null hypothesis.

- A large p-value (> 0.05) indicates weak evidence against the null hypothesis, so you fail to reject the null hypothesis.

- p-values very close to the cut-off (0.05) are considered to be marginal (could go either way). Always report the p-value so your readers can draw their own conclusions.

15. What are z and t stats?

Z-test:

In a z-test, the sample is assumed to be normally distributed. A z-score is calculated with population parameters such as "population mean" and "population standard deviation" and is used to validate a hypothesis that the sample drawn belongs to the same population.

Null: Sample mean is same as the population mean
Alternate: Sample mean is not same as the population mean

The statistics used for this hypothesis testing is called z-statistic, the score for which is calculated as

- $z = (x - \mu) / (\sigma / \sqrt{n})$, where
- x= sample mean
- μ = population mean
- $\sigma / \sqrt{n}$ = population standard deviation

If the test statistic is lower than the critical value, accept the hypothesis or else reject the hypothesis

T-test:

A t-test is used to compare the mean of two given samples. Like a z-test, a t-test also assumes a normal distribution of the sample. A t-test is used when the population parameters (mean and standard deviation) are not known.
There are three versions of t-test
1. Independent samples t-test which compares mean for two groups
2. Paired sample t-test which compares means from the same group at different times
3. One sample t-test which tests the mean of a single group against a known mean.

The statistic for this hypothesis testing is called t-statistic, the score for which is calculated as

- $t = (x1 - x2) / (\sigma / \sqrt{n1} + \sigma / \sqrt{n2})$, where
- x1 = mean of sample 1
- x2 = mean of sample 2
- n1 = size of sample 1
- n2 = size of sample 2

16. What are hypothesis testing?

A statistical hypothesis is an assumption about a population parameter. This assumption may or may not be true. **Hypothesis testing** refers to the formal procedures used by statisticians to accept or reject statistical hypotheses.

Statistical Hypotheses
The best way to determine whether a statistical hypothesis is true would be to examine the entire population. Since that is often impractical, researchers typically examine a random sample from the population. If sample data are not consistent with the statistical hypothesis, the hypothesis is rejected.

There are two types of statistical hypotheses.

Null hypothesis. The null hypothesis, denoted by Ho, is usually the hypothesis that sample observations result purely from chance.

Alternative hypothesis. The alternative hypothesis, denoted by H1 or Ha, is the hypothesis that sample observations are influenced by some non-random cause. 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

- Ho: P = 0.5
- Ha: P ≠ 0.5

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.

Other terms related to Hypothesis Testing are:

Level of significance: Refers to the degree of significance in which we accept or reject the null-hypothesis. 100% accuracy is not possible for accepting or rejecting a hypothesis, so we therefore select a level of significance that is usually 5%.

Type I error: When we reject the null hypothesis, although that hypothesis was true. Type I error is denoted by alpha. In hypothesis testing, the normal curve that shows the critical region is called the alpha region.

Type II errors: When we accept the null hypothesis but it is false. Type II errors are denoted by beta. In Hypothesis testing, the normal curve that shows the acceptance region is called the beta region.

Power: Usually known as the probability of correctly accepting the null hypothesis. 1-beta is called power of the analysis.

One-tailed test: When the given statistical hypothesis is one value like H0: $\mu 1 = \mu 2$, it is called the one-tailed test.

Two-tailed test: When the given statistics hypothesis assumes a less than or greater than value, it is called the two-tailed test

17. How do you determine outliers?

The different ways of determining the outliers:

a. **Simply sort your data** sheet for each variable and then look for unusually high or low values.
b. **Graphing your data**: Using Boxplots, Histogram and Scatterplots.

Boxplot: We use boxplots when we have groups in our data. Boxplots display asterisks or other symbols on the graph to indicate explicitly when datasets contain outliers.

Histogram: Histograms also emphasize the existence of outliers. Look for isolated bars in the plot.

Scatterplot: We can use Scatterplot to determine outliers in a multivariate setting. Isolated points from regression line gives us the outliers.

c. **Using Z-Score**: Z-scores can quantify the unusualness of an observation when your data follow the normal distribution. Z-scores are the number of standard deviations above and below the mean that each value falls. For example, a Z-score of 2 indicates that an observation is two standard deviations above the average while a Z-score of -2 signifies it is two standard deviations below the mean. A Z-score of zero represents a value that equals the mean. Statistical rule of thumb is if Z-Score for a specific data point is less than -4 or greater than 4 is a suspected outlier.

d. **Using IQR:** You can use the interquartile range (IQR), several quartile values, and an adjustment factor to calculate boundaries for what constitutes minor and major outliers. Minor and major denote the unusualness of the outlier relative to the overall distribution of values. Major outliers are more extreme. Analysts also refer to these categorizations as mild and extreme outliers.

The IQR is the middle 50% of the dataset. It's the range of values between the third quartile and the first quartile (Q3 – Q1). We can take the IQR, Q1, and Q3 values to calculate the following outlier fences for our dataset:
lower outer, lower inner, upper inner, and upper outer.

These fences determine whether data points are outliers and whether they are mild or extreme.

Values that fall inside the two inner fences are not outliers.

Execute | Share main.py STDIN Login ×

```python
#Created By: Dhrub Satyam

# identify outliers with interquartile range
from numpy.random import seed
from numpy.random import randn
from numpy import percentile

# seed the random number generator
seed(1)

# generate univariate observations
data = 5 * randn(10000) + 50

# calculate interquartile range
q25, q75 = percentile(data, 25), percentile(data, 75)
iqr = q75 - q25
print('Percentiles: 25th=%.3f, 75th=%.3f, IQR=%.3f' % (q25, q75, iqr))

# calculate the outlier cutoff
cut_off = iqr * 1.5
lower, upper = q25 - cut_off, q75 + cut_off

# identify outliers
outliers = [x for x in data if x < lower or x > upper]
print('Identified outliers: %d' % len(outliers))

# remove outliers
outliers_removed = [x for x in data if x >= lower and x <= upper]
print('Non-outlier observations: %d' % len(outliers_removed))
```

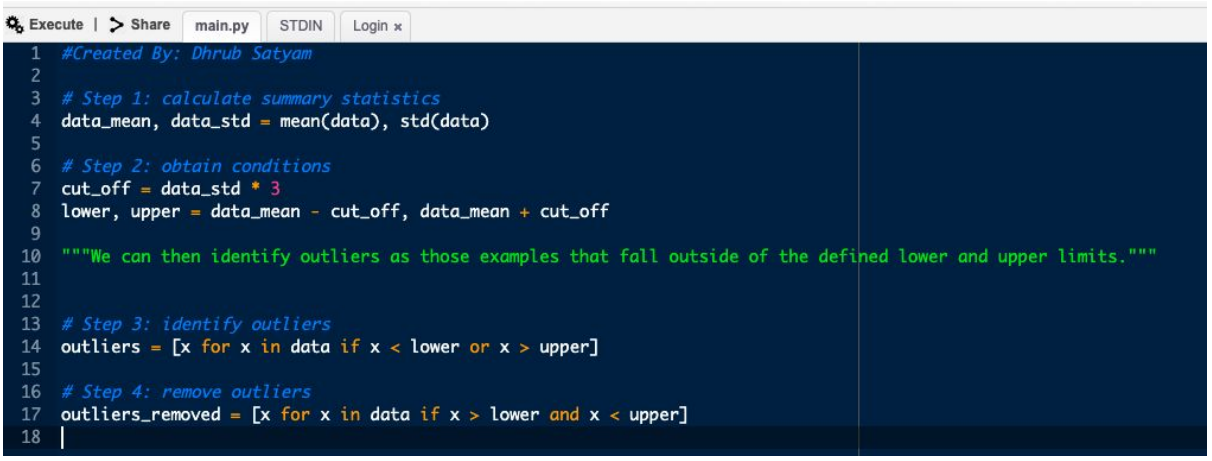
e. **<u>Using Hypothesis Testing:</u>** You can use hypothesis tests to find outliers. Many outlier tests exist, but I'll focus on one to illustrate how they work. In this post, I demonstrate Grubbs' test, which tests the following hypotheses:

Null: All values in the sample were drawn from a single population that follows the same normal distribution.

Alternative: One value in the sample was not drawn from the same normally distributed population as the other values.

If the p-value for this test is less than your significance level, you can reject the null and conclude that one of the values is an outlier.

f. Standard Deviation Method: We know that if the data follow normal distribution then the data covers 99.7% of the points up to 3 standard deviation. We can have our outliers calculated beyond that on both sides. Python Code for this is as follows:



18. When do you reject or accept null hypothesis? List Steps.

Step 1: State the null hypothesis. When you state the null hypothesis, you also have to state the alternate hypothesis. Sometimes it is easier to state the alternate hypothesis first, because that's the researcher's thoughts about the experiment.

Step 2: Support or reject the null hypothesis. Several methods exist, depending on what kind of sample data you have. For example, you can use the P-value method.

If you are able to reject the null hypothesis in Step 2, you can replace it with the alternate hypothesis.

That's it!

When to Reject the Null hypothesis?

Basically, you reject the null hypothesis when your test value falls into the rejection region. There are four main ways you'll compute test values and either support or reject your null hypothesis. Which method you choose depends mainly on if you have a proportion or a p-value.

(For details refer :
https://www.sagepub.com/sites/default/files/upmbinaries/40007_Chapter8.pdf)

19. What are Type I and Type II errors? How would you proceed to minimise them?
No hypothesis test is 100% certain. Because the test is based on probabilities, there is always a chance of making an incorrect conclusion. When you do a hypothesis test, two types of errors are possible: type I and type II. The risks of these two errors are inversely related and determined by the level of significance and the power for the test. Therefore, you should determine which error has more severe consequences for your situation before you define their risks.

Type I error:
When the null hypothesis is true and you reject it, you make a type I error. The probability of making a type I error is α, which is the level of significance you set for your hypothesis test. An α of 0.05 indicates that you are willing to accept a 5% chance that you are wrong when you reject the null hypothesis. To lower this risk, you must use a lower value for α. However, using a lower value for alpha means that you will be less likely to detect a true difference if one really exists.

Type II error:
When the null hypothesis is false and you fail to reject it, you make a type II error. The probability of making a type II error is β, which depends on the power of the test. You can decrease your risk of committing a type II error by ensuring your test has enough power. You can do this by ensuring your sample size is large enough to detect a practical difference when one truly exists.

The probability of rejecting the null hypothesis when it is false is equal to $1-\beta$. This value is the power of the test.

	Truth about the population	
Decision based on sample	H_0 is true	H_0 is false
Fail to reject H_0	Correct Decision (probability = 1 - α)	**Type II Error** - fail to reject H_0 when it is false (probability = β)
Reject H_0	**Type I Error** - rejecting H_0 when it is true (probability = α)	Correct Decision (probability = 1 - β)

Avoiding Type-I Error:
Minimize the significance level (α) of Hypothesis Test: since we choose the significance level. However, lowering the significance level may lead to a situation wherein the results of the hypothesis test may not capture the true parameter or the true difference of the test.

Avoiding Type-II Error:
a. Increase the sample size
b. Increase the significance level of Hypothesis Test.

20. What is Bias and Variance? How can we have optimum of both?
Bias is the difference between the average prediction of our model and the correct value which we are trying to predict. Model with high bias pays very little attention to the training data and oversimplifies the model. It always leads to high error on training and test data.

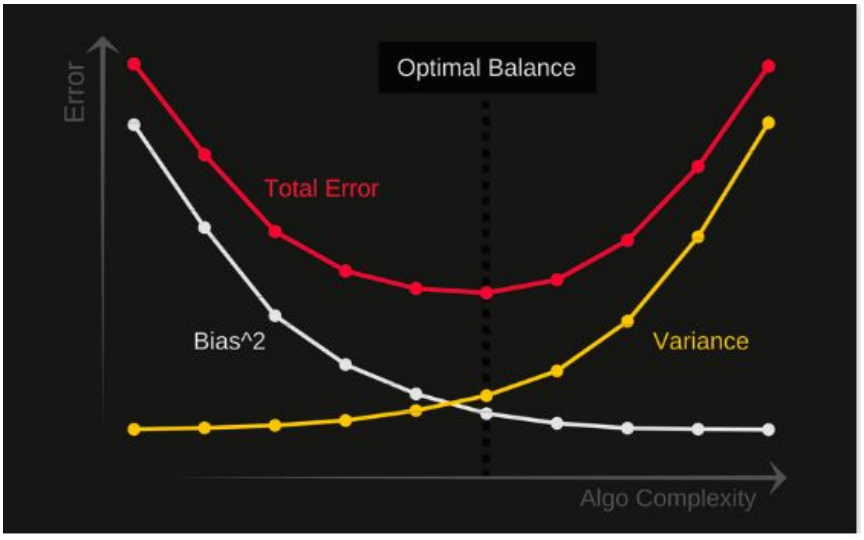
Variance is the variability of model prediction for a given data point or a value which tells us spread of our data. Model with high variance pays a lot of attention to training data and does not generalize on the data which it hasn't seen before. As a

result, such models perform very well on training data but has high error rates on test data.

In supervised learning, **underfitting** happens when a model unable to capture the underlying pattern of the data. These models usually have high bias and low variance. It happens when we have very less amount of data to build an accurate model or when we try to build a linear model with a nonlinear data. Also, these kind of models are very simple to capture the complex patterns in data like Linear and logistic regression.

In supervised learning, **overfitting** happens when our model captures the noise along with the underlying pattern in data. It happens when we train our model a lot over noisy dataset. These models have low bias and high variance. These models are very complex like Decision trees which are prone to overfitting.

Bias Variance Tradeoff:



For any supervised algorithm, having a high bias error usually means it has low variance error and vice versa. To be more specific, parametric or linear ML algorithms often have a high bias but low variance. On the other hand, non-parametric or non-linear algorithms have vice versa.

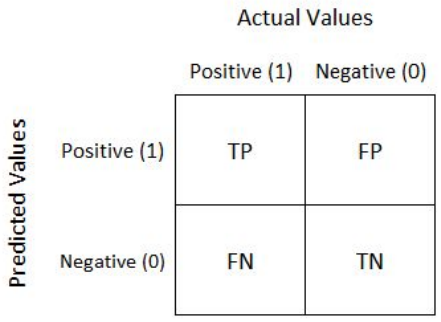
The goal of any ML model is to obtain a low variance and a low bias state, which is often a task due to the parametrization of machine learning algorithms. Common ways to achieve optimum Bias and Variance are:

a. *By minimising total error*
b. *Using Bagging and resampling techniques*
c. *Adjusting minor values in Algorithms*

(Refer : https://hub.packtpub.com/heres-how-you-can-handle-the-bias-variance-trade-off-in-your-ml-models/)

21. What is Confusion Matrix?

it is a performance measurement for machine learning classification problem where output can be two or more classes. It is a table with 4 different combinations of predicted and actual values.



Definition of the terms:

- Positive (P) : Observation is positive (for example: is an apple).
- Negative (N) : Observation is not positive (for example: is not an apple).
- True Positive (TP) : Observation is positive, and is predicted to be positive.
- False Negative (FN) : Observation is positive, but is predicted negative.
- True Negative (TN) : Observation is negative, and is predicted to be negative.
- False Positive (FP) : Observation is negative, but is predicted positive.

Classification Rate/Accuracy: Classification Rate or Accuracy is given by the relation:

$$\textbf{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

However, there are problems with accuracy. It assumes equal costs for both kinds of errors. A 99% accuracy can be excellent, good, mediocre, poor or terrible depending upon the problem.

Recall/Sensitivity/True Positive Rate: Recall can be defined as the ratio of the total number of correctly classified positive examples divide to the total number of positive examples. High Recall indicates the class is correctly recognized (a small number of FN).

$$\textbf{\textit{Recall}} = \frac{TP}{TP + FN}$$

Precision: To get the value of precision we divide the total number of correctly classified positive examples by the total number of predicted positive examples. High Precision indicates an example labelled as positive is indeed positive (a small number of FP).

$$Precision = \frac{TP}{TP + FP}$$

F-measure/F-stats/F1 Score: Since we have two measures (Precision and Recall) it helps to have a measurement that represents both of them. We calculate an F-measure which uses Harmonic Mean in place of Arithmetic Mean as it punishes the extreme values more.
The F-Measure will always be nearer to the smaller value of Precision or Recall.

$$F - measure = \frac{2*Recall*Precision}{Recall + Precision}$$

Specificity: Percentage of negative instances out of the total actual negative instances. Therefore, denominator (TN + FP) here is the actual number of negative instances present in the dataset. It is similar to recall but the shift is on the negative instances. Like finding out how many healthy patients were not having cancer and were told they don't have cancer. Kind of a measure to see how separate the classes are.

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

22. What are different ways of measuring the performance of different models?

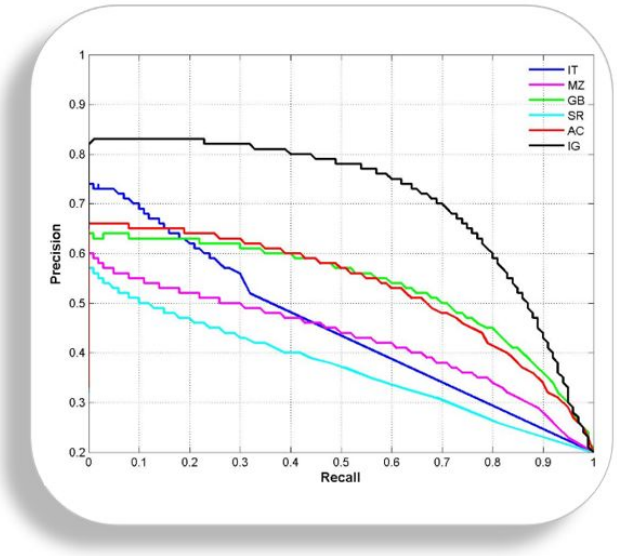
The different ways are as follows:

- Confusion matrix
- Accuracy
- Precision
- Recall
- Specificity
- F1 score
- Precision-Recall or PR curve
- ROC (Receiver Operating Characteristics) curve
- PR vs ROC curve

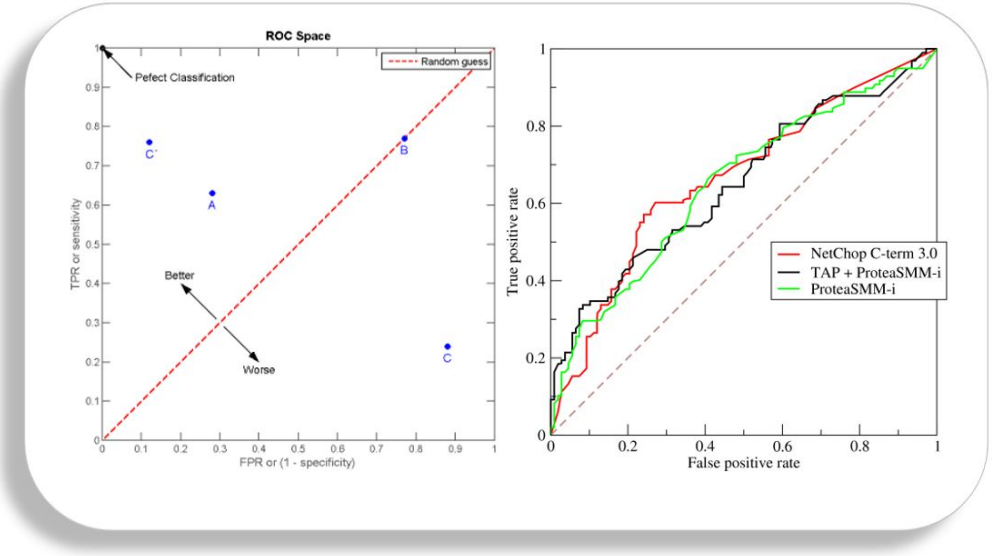
Most of the ways we have already seen in the above question. We will discuss only which are left.

PR Curve: It is the curve between precision and recall for various threshold values. In the figure below we have 6 predictors showing their respective precision-recall curve

for various threshold values. The top right part of the graph is the ideal space where we get high precision and recall. Based on our application we can choose the predictor and the threshold value. PR AUC is just the area under the curve. The higher its numerical value the better.



ROC Curve: ROC stands for receiver operating characteristic and the graph is plotted against TPR and FPR for various threshold values. As TPR increases FPR also increases. As you can see in the first figure, we have four categories and we want the threshold value that leads us closer to the top left corner. Comparing different predictors (here 3) on a given dataset also becomes easy as you can see in figure 2, one can choose the threshold according to the application at hand. ROC AUC is just the area under the curve, the higher its numerical value the better.



PR vs ROC Curve:
Both the metrics are widely used to judge a models performance.
Which one to use PR or ROC?

The answer lies in ***TRUE NEGATIVES***.

Due to the absence of TN in the precision-recall equation, they are useful in imbalanced classes. In the case of class imbalance when there is a majority of the negative class. The metric doesn't take much into consideration the high number of TRUE NEGATIVES of the negative class which is in majority, giving better resistance to the imbalance. This is important when the detection of the positive class is very important.
Like to detect cancer patients, which has a high class imbalance because very few have it out of all the diagnosed. We certainly don't want to miss on a person having cancer and going undetected (recall) and be sure the detected one is having it (precision).
Due to the consideration of TN or the negative class in the ROC equation, it is useful when both the classes are important to us. Like the detection of cats and dog. The importance of true negatives makes sure that both the classes are given importance, like the output of a CNN model in determining the image is of a cat or a dog.

23. What is correlation? What is the range that a correlation coefficient can have?

Correlation is a statistical technique that can show whether and how strongly pairs of variables are related. For example, height and weight are related; taller people tend to be heavier than shorter people.
The main result of a correlation is called the correlation coefficient (or "r"). It ranges from -1.0 to +1.0. The closer r is to +1 or -1, the more closely the two variables are related.

24. What is ROC Curve? What is AUC?

Refer to Question 16 for ROC Curve.
AUC: ROC curve is a performance measurement for classification problem at various thresholds settings. ROC is a probability curve and AUC represents degree or measure of separability. It tells how much model is capable of distinguishing between classes. Higher the AUC, better the model is at predicting 0s as 0s and 1s as 1s. By analogy, Higher the AUC, better the model is at distinguishing between patients with disease and no disease.
The ROC curve is plotted with TPR against the FPR where TPR is on y-axis and FPR is on the x-axis.

25. What is Collinearity and Correlation?

Correlation measures the strength of linear relationship between two variables.

Collinearity: If in multiple regression analysis, one of the predictors is linearly associated/dependent on other predictor, then this issue is known as collinearity.

For example, let's consider the linear model

$Y = \alpha x1 + \beta 1x1 + \beta 2x2 \ldots (1)$

If predictor x1 can be expressed as linear combination of x2, say, x1 = 3*x2

Then this is known as collinearity among the predictors. Note that there will be perfect (or very high) correlation between the predictors as opposed to the assumption of linear regression model (All predictors are assumed to be independent).

Essentially it means that one of the independent variables is not really necessary to the model because its effect/impact on the model is already captured by some of the other variables. This variable is not contributing anything extra to the predictions and can be removed. If we have true collinearity (perfect correlation as in the example above), the one of the predictor is automatically deleted by some of the software's like R, other shows an error or warning for the same.

The effects of collinearity are seen in the variances of the parameter estimates, not in the parameter estimates themselves.

Note: Collinearity implies correlation but not vice- versa.

26. What are the different types of Sampling?

Some of the Common sampling ways are as follows:

- **Simple random sample**: Every member and set of members has an equal chance of being included in the sample. Technology, random number generators, or some other sort of chance process is needed to get a simple random sample.

Example—A teacher puts students' names in a hat and chooses without looking to get a sample of students.

Why it's good: Random samples are usually fairly representative since they don't favour certain members.

- **Stratified random sample**: The population is first split into groups. The overall sample consists of some members from every group. The members from each group are chosen randomly.

Example—A student council surveys 100100100 students by getting random samples of 252525 freshmen, 252525 sophomores, 252525 juniors, and 252525 seniors.

Why it's good: A stratified sample guarantees that members from each group will be represented in the sample, so this sampling method is good when we want some members from every group.

- **Cluster random sample**: The population is first split into groups. The overall sample consists of every member from some of the groups. The groups are selected at random.

Example—An airline company wants to survey its customers one day, so they randomly select 555 flights that day and survey every passenger on those flights.

Why it's good: A cluster sample gets every member from some of the groups, so it's good when each group reflects the population as a whole.

- **Systematic random sample**: Members of the population are put in some order. A starting point is selected at random, and every nth member is selected to be in the sample.

Example—A principal takes an alphabetized list of student names and picks a random starting point. Every 20^{th} student is selected to take a survey.

27. What is confidence interval? What is its significance?

A confidence interval, in statistics, refers to the probability that a population parameter will fall between two set values for a certain proportion of times. Confidence intervals measure the degree of uncertainty or certainty in a sampling method. A confidence interval can take any number of probabilities, with the most common being a 95% or 99% confidence level.

28. What are the effects of the width of confidence interval?

- Confidence interval is used for decision making
- As the confidence level increases the width of the confidence interval also increases
- As the width of the confidence interval increases, we tend to get useless information also.
- Useless information – wide CI
- High risk – narrow CI

29. What is level of significance (Alpha)?

The significance level, also denoted as alpha or α, is a measure of the strength of the evidence that must be present in your sample before you will reject the null hypothesis and conclude that the effect is statistically significant. The researcher determines the significance level before conducting the experiment.

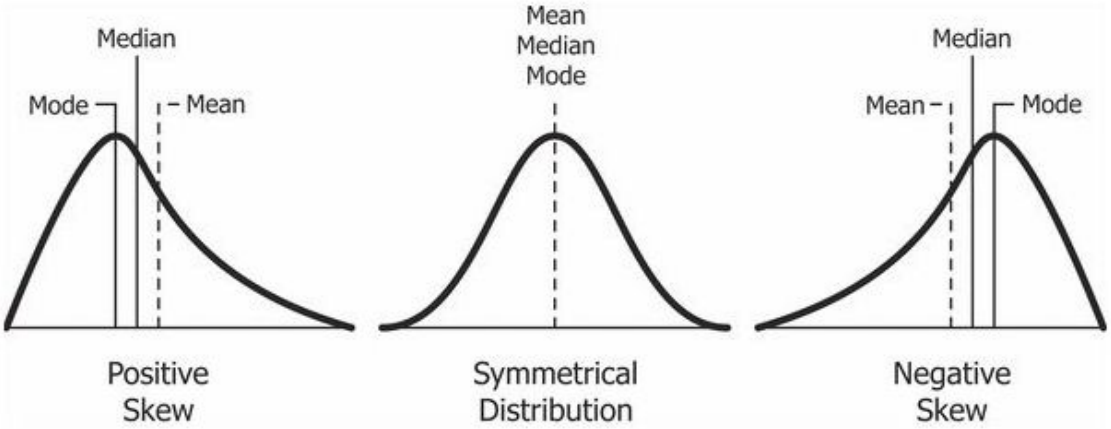
The significance level is the probability of rejecting the null hypothesis when it is true. For example, a significance level of 0.05 indicates a 5% risk of concluding that a difference exists when there is no actual difference. Lower significance levels indicate that you require stronger evidence before you will reject the null hypothesis.

Use significance levels during hypothesis testing to help you determine which hypothesis the data support. Compare your p-value to your significance level. If the p-value is less than your significance level, you can reject the null hypothesis and conclude that the effect is statistically significant. In other words, the evidence in your sample is strong enough to be able to reject the null hypothesis at the population level.

30. What is Skewness and Kurtosis? What does it signify?

Skewness: It is the degree of distortion from the symmetrical bell curve or the normal distribution. It measures the lack of symmetry in data distribution.
It differentiates extreme values in one versus the other tail. A symmetrical distribution will have a skewness of 0.

There are two types of Skewness: Positive and Negative



Positive Skewness means when the tail on the right side of the distribution is longer or fatter. The mean and median will be greater than the mode.
Negative Skewness is when the tail of the left side of the distribution is longer or fatter than the tail on the right side. The mean and median will be less than the mode.

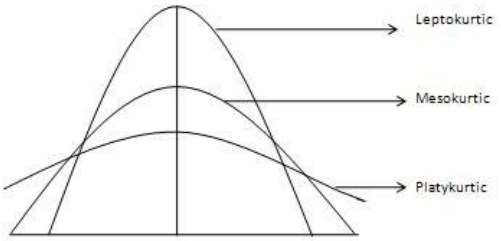
So, when is the skewness too much?

The rule of thumb seems to be:

- If the skewness is between -0.5 and 0.5, the data are fairly symmetrical.
- If the skewness is between -1 and -0.5(negatively skewed) or between 0.5 and 1(positively skewed), the data are moderately skewed.
- If the skewness is less than -1(negatively skewed) or greater than 1(positively skewed), the data are highly skewed.

Example
Let us take a very common example of house prices. Suppose we have house values ranging from $100k to $1,000,000 with the average being $500,000.
If the peak of the distribution was left of the average value, portraying a positive skewness in the distribution. It would mean that many houses were being sold for less than the average value, i.e. $500k. This could be for many reasons, but we are not going to interpret those reasons here.
If the peak of the distributed data was right of the average value, that would mean a negative skew. This would mean that the houses were being sold for more than the average value.

Kurtosis: Kurtosis is all about the tails of the distribution — not the peakedness or flatness. It is used to describe the extreme values in one versus the other tail. It is actually the measure of outliers present in the distribution.



High kurtosis in a data set is an indicator that data has heavy tails or outliers. If there is a high kurtosis, then, we need to investigate why do we have so many outliers. It indicates a lot of things, maybe wrong data entry or other things. Investigate!
Low kurtosis in a data set is an indicator that data has light tails or lack of outliers. If we get low kurtosis(too good to be true), then also we need to investigate and trim the dataset of unwanted results.

Mesokurtic: This distribution has kurtosis statistic similar to that of the normal distribution. It means that the extreme values of the distribution are similar to that of a normal distribution characteristic. This definition is used so that the standard normal distribution has a kurtosis of three.

Leptokurtic (Kurtosis > 3): Distribution is longer, tails are fatter. Peak is higher and sharper than Mesokurtic, which means that data are heavy-tailed or profusion of outliers.
Outliers stretch the horizontal axis of the histogram graph, which makes the bulk of the data appear in a narrow ("skinny") vertical range, thereby giving the "skinniness" of a leptokurtic distribution.

Platykurtic: **(Kurtosis < 3):** Distribution is shorter; tails are thinner than the normal distribution. The peak is lower and broader than Mesokurtic, which means that data are light-tailed or lack of outliers. The reason for this is because the extreme values are less than that of the normal distribution.

31. What is Range and IQR? What does it signify?

Range: The range of a set of data is the difference between the highest and lowest values in the set.
IQR(Inter Quartile Range): The interquartile range (IQR) is the difference between the first quartile and third quartile. The formula for this is:

$$IQR = Q3 - Q1$$

The range gives us a measurement of how spread out the entirety of our data set is. The interquartile range, which tells us how far apart the first and third quartile are, indicates how spread out the middle 50% of our set of data is.

32. What is difference between Variance and Standard Deviation? What is its significance?

The central tendency mean gives you the idea of average of the data points(i.e centre location of the distribution)

And now you want to know how far are your data points from mean

So, here comes the concept of variance to calculate how far are your data points from mean (in simple terms, it is to calculate the variation of your data points from mean)

$$\text{Population variance} : \sum_{i=1}^{n} \frac{(x_i-\mu)^2}{N}$$

$$\text{Sample variance} : \sum_{i=1}^{n} \frac{(x_i-\bar{x})^2}{n-1}$$

Standard deviation is simply the square root of variance . And standard deviation is also used to calculate the variation of your data points

(And you may be asking, why do we use standard deviation , when we have variance. Because, in order to maintain the calculations in same units i.e suppose mean is in cm/m, then variance is in $cm2/m2$, whereas standard deviation is in cm/m , so we use standard deviation most)

$$\text{population standard deviation:} \sqrt{\sum_{i=1}^{n} \frac{(x_i-\mu)^2}{N}}$$

$$\text{sample standard deviation} : \sqrt{\sum_{i=1}^{n} \frac{(x_i-\bar{x})^2}{n-1}}$$

33. What is selection Bias? Types of Selection Bias?

Selection bias is the phenomenon of selecting individuals, groups or data for analysis in such a way that proper randomization is not achieved, ultimately resulting in a sample that is not representative of the population.

Understanding and identifying selection bias is important because it can significantly skew results and provide false insights about a particular population group.
Types of selection bias include:

- **Sampling bias**: a biased sample caused by non-random sampling
- **Time interval**: selecting a specific time frame that supports the desired conclusion. e.g. conducting a sales analysis near Christmas.
- **Exposure**: includes clinical susceptibility bias, protopathic bias, indication bias. Read more here.

- **Data**: includes cherry-picking, suppressing evidence, and the fallacy of incomplete evidence.
- **Attrition**: attrition bias is similar to survivorship bias, where only those that 'survived' a long process are included in an analysis, or failure bias, where those that 'failed' are only included
- **Observer selection**: related to the Anthropic principle, which is a philosophical consideration that any data we collect about the universe is filtered by the fact that, in order for it to be observable, it must be compatible with the conscious and sapient life that observes it.

Handling missing data can make selection bias worse because different methods impact the data in different ways. For example, if you replace null values with the mean of the data, you adding bias in the sense that you're assuming that the data is not as spread out as it might actually be.

34. What are the ways of handling missing Data?

- Delete rows with missing data
- Mean/Median/Mode imputation
- Assigning a unique value
- Predicting the missing values using Machine Learning Models
- Using an algorithm which supports missing values, like random forests

35. What are the different types of probability distribution? Explain with example?

The common Probability Distribution are as follows:

1. Bernoulli Distribution
2. Uniform Distribution
3. Binomial Distribution
4. Normal Distribution
5. Poisson Distribution

1. Bernoulli Distribution: A Bernoulli distribution has only two possible outcomes, namely 1 (success) and 0 (failure), and a single trial. So the random variable X which has a Bernoulli distribution can take value 1 with the probability of success, say p, and the value 0 with the probability of failure, say q or 1-p.

Exmple: whether it's going to rain tomorrow or not where rain denotes success and no rain denotes failure and Winning (success) or losing (failure) the game.

2. Uniform Distribution: When you roll a fair die, the outcomes are 1 to 6. The probabilities of getting these outcomes are equally likely and that is the basis of a uniform distribution. Unlike Bernoulli Distribution, all the n number of possible outcomes of a uniform distribution are equally likely.

Example: Rolling a fair dice.

3. Binomial Distribution: A distribution where only two outcomes are possible, such as success or failure, gain or loss, win or lose and where the probability of success and failure is same for all the trials is called a Binomial Distribution.
    - Each trial is independent.

- There are only two possible outcomes in a trial- either a success or a failure.
- A total number of n identical trials are conducted.
- The probability of success and failure is same for all trials. (Trials are identical.)

Example: Tossing a coin.

4. Normal Distribution: Normal distribution represents the behaviour of most of the situations in the universe (That is why it's called a "normal" distribution. I guess!). The large sum of (small) random variables often turns out to be normally distributed, contributing to its widespread application. Any distribution is known as Normal distribution if it has the following characteristics:
    - The mean, median and mode of the distribution coincide.
    - The curve of the distribution is bell-shaped and symmetrical about the line $x=\mu$.
    - The total area under the curve is 1.
    - Exactly half of the values are to the left of the center and the other half to the right.

5. Poisson Distribution: A distribution is called **Poisson distribution** when the following assumptions are valid:
    - Any successful event should not influence the outcome of another successful event.
    - The probability of success over a short interval must equal the probability of success over a longer interval.
    - The probability of success in an interval approaches zero as the interval becomes smaller.

    Example: The number of emergency calls recorded at a hospital in a day.

36. What are statistical Tests? List Them.

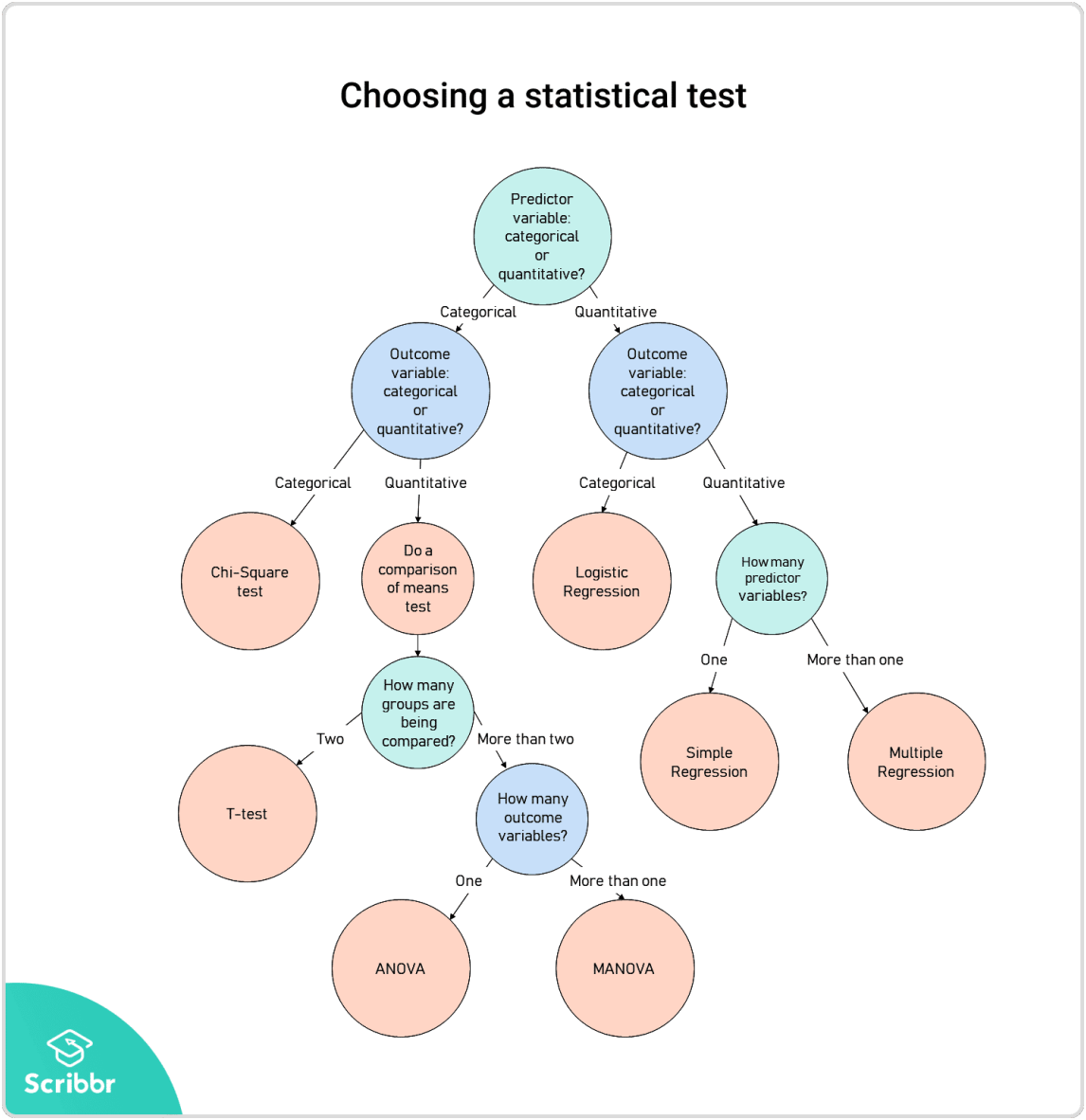
Statistical tests are used in hypothesis testing. They can be used to:

- determine whether a predictor variable has a statistically significant relationship with an outcome variable.
- estimate the difference between two or more groups.

Statistical tests assume a null hypothesis of no relationship or no difference between groups. Then they determine whether the observed data fall outside of the range of values predicted by the null hypothesis.

Common Tests in Statistics:

a. T-Test/Z-Test
b. ANOVA
c. Chi-Square Test
d. MANOVA



37. How do you calculate the sample size required?

You can use the margin of error (ME) formula to determine the desired sample size.

$$ME = t * \frac{S}{\sqrt{n}} \quad -or- \quad ME = z * \frac{\sigma}{\sqrt{n}}$$

Formula for margin of error

- t/z = t/z score used to calculate the confidence interval
- ME = the desired margin of error
- S = sample standard deviation

38. What are the different Biases associated when we sample?

Potential biases include the following:

- Sampling bias: a biased sample caused by non-random sampling
- Under coverage bias: sampling too few observations
- Survivorship bias: error of overlooking observations that did not make it past a form of selection process.

39. How to convert normal distribution to standard normal distribution?

Standardized normal distribution has mean = 0 and standard deviation = 1

To convert normal distribution to standard normal distribution we can use the formula:

$$X \text{ (standardized)} = (x-\mu) / \sigma$$

40. How to find the mean length of all fishes in a river?

- Define the confidence level (most common is 95%)
- Take a sample of fishes from the river (to get better results the number of fishes > 30)
- Calculate the mean length and standard deviation of the lengths
- Calculate t-statistics
- Get the confidence interval in which the mean length of all the fishes should be.

41. What do you mean by degree of freedom?

- DF is defined as the number of options we have
- DF is used with t-distribution and not with Z-distribution
- For a series, DF = n-1 (where n is the number of observations in the series)

42. What do you think if DF is more than 30?

- As DF increases the t-distribution reaches closer to the normal distribution
- At low DF, we have fat tails
- If DF > 30, then t-distribution is as good as normal distribution

43. When to use t distribution and when to use z distribution?

- The following conditions must be satisfied to use Z-distribution
- Do we know the population standard deviation?
- Is the sample size > 30?
- CI = x (bar) – $Z^* \sigma / \sqrt{n}$ to x (bar) + $Z^* \sigma / \sqrt{n}$
- Else we should use t-distribution
- CI = x (bar) – $t^*s/\sqrt{n}$ to x (bar) + $t^*s/\sqrt{n}$

44. What is H0 and H1? What is H0 and H1 for two-tail test?

- H0 is known as null hypothesis. It is the normal case / default case.

For one tail test $x <= \mu$

For two-tail test $x = \mu$

- H1 is known as alternate hypothesis. It is the other case.

For one tail test $x > \mu$
For two-tail test $x <> \mu$

45. What is Degree of Freedom?

DF is defined as the number of options we have
DF is used with t-distribution and not with Z-distribution
For a series, DF = n-1 (where n is the number of observations in the series)

46. How to calculate p-Value?

Calculating p-value:

a. Using Excel:
   I. Go to Data tab
   II. Click on Data Analysis
   III. Select Descriptive Statistics
   IV. Choose the column
   V. Select summary statistics and confidence level (0.95)

b. By Manual Method:
   I. Find H0 and H1
   II. Find n, x(bar) and s
   III. Find DF for t-distribution
   IV. Find the type of distribution – t or z distribution
   V. Find t or z value (using the look-up table)
   VI. Compute the p-value to critical value

47. What is ANOVA?

ANOVA expands to the analysis of variance, is described as a statistical technique used to determine the difference in the means of two or more populations, by examining the amount of variation within the samples corresponding to the amount of variation between the samples. It bifurcates the total amount of variation in the dataset into two parts, i.e. the amount ascribed to chance and the amount ascribed to specific causes.

It is a method of analysing the factors which are hypothesised or affect the dependent variable. It can also be used to study the variations amongst different categories, within the factors, that consist of numerous possible values. It is of two types:

One way ANOVA: When one factor is used to investigate the difference amongst different categories, having many possible values.
Two way ANOVA: When two factors are investigated simultaneously to measure the interaction of the two factors influencing the values of a variable.

48. What is ANCOVA?

ANCOVA stands for Analysis of Covariance, is an extended form of ANOVA, that eliminates the effect of one or more interval-scaled extraneous variable, from the dependent variable before carrying out research. It is the midpoint between ANOVA and regression analysis, wherein one variable in two or more population can be compared while considering the variability of other variables.

When in a set of independent variable consist of both factor (categorical independent variable) and covariate (metric independent variable), the technique used is known as ANCOVA. The difference in dependent variables because of the covariate is taken off by an adjustment of the dependent variable's mean value within each treatment condition.

This technique is appropriate when the metric independent variable is linearly associated with the dependent variable and not to the other factors. It is based on certain assumptions which are:

There is some relationship between dependent and uncontrolled variable.
The relationship is linear and is identical from one group to another.
Various treatment groups are picked up at random from the population.
Groups are homogeneous in variability.

49. What are difference between ANOVA and ANCOVA?

The points given below are substantial so far as the difference between AOVA and ANCOVA is concerned:

- The technique of identifying the variance among the means of multiple groups for homogeneity is known as Analysis of Variance or ANOVA. A statistical process which is used to take off the impact of one or more metric-scaled undesirable variable from dependent variable before undertaking research is known as ANCOVA.
- While ANOVA uses both linear and non-linear model. On the contrary, ANCOVA uses only linear model.
- ANOVA entails only categorical independent variable, i.e. factor. As against this, ANCOVA encompasses a categorical and a metric independent variable.
- A covariate is not taken into account, in ANOVA, but considered in ANCOVA.
- ANOVA characterises between group variations, exclusively to treatment. In contrast, ANCOVA divides between group variations to treatment and covariate.
- ANOVA exhibits within group variations, particularly individual differences. Unlike ANCOVA, that bifurcates within group variance in individual differences and covariate.

50. What are t and z scores? Give Details.

<u>T-Score vs. Z-Score: Overview</u>

A z-score and a t score are both used in hypothesis testing.

<u>T-score vs. z-score: When to use a t score</u>

The general rule of thumb for when to use a t score is when your sample:

Has a sample size below 30,
Has an unknown population standard deviation.
You must know the standard deviation of the population and your sample size should be above 30 in order for you to be able to use the z-score. Otherwise, use the t-score.

Z-score

Technically, z-scores are a conversion of individual scores into a standard form. The conversion allows you to more easily compare different data. A z-score tells you how many standard deviations from the mean your result is. You can use your knowledge of normal distributions (like the 68 95 and 99.7 rule) or the z-table to determine what percentage of the population will fall below or above your result.

The z-score is calculated using the formula:

- $z = (X-\mu)/\sigma$

Where:

- σ is the population standard deviation and
- μ is the population mean.
- The z-score formula doesn't say anything about sample size; The rule of thumb applies that your sample size should be above 30 to use it.

T-score

Like z-scores, t-scores are also a conversion of individual scores into a standard form. However, t-scores are used when you don't know the population standard deviation; You make an estimate by using your sample.

- $T = (X - \mu) / [s/\sqrt{(n)}]$

Where:

- s is the standard deviation of the sample.

If you have a larger sample (over 30), the t-distribution and z-distribution look pretty much the same.

To know more about Data Science, Artificial Intelligence, Machine Learning and Deep Learning programs visit our website www.learnbay.co

Follow us on:
[LinkedIn](#)
[Facebook](#)
[Twitter](#)

Watch our Live Session Recordings to precisely understand statistics, probability, calculus, linear algebra and other math concepts used in data science.
[Youtube](#)
To get updates on Data Science and AI Seminars/Webinars - Follow our [Meetup](#) group.