# DATA SCIENCE INTERVIEW PREPARATION
# (30 Days of Interview Preparation)
# # Day-16

## Q1.What is Statistics Learning?

**Answer:**

**Statistical learning**: It is the framework for understanding data based on the statistics, which can be classified as the supervised or unsupervised. Supervised statistical learning involves building the statistical model for predicting, or estimating, an output based on one or more inputs, while in unsupervised statistical learning, there are inputs but no supervising output, but we can learn relationships and structure from such data.

$Y = f(X) + \varepsilon$ , $X = (X1, X2, \ldots, Xp)$,

$f$: It is an unknown function & $\varepsilon$ is random error (reducible & irreducible).

**Prediction & Inference:**

In the situations , where the set of inputs X are readily available, but the output Y is not known, we often treat f as the black box (not concerned with the exact form of "f"), as long as it yields the accurate predictions for Y. This is the *prediction*.

There are the situations where we are interested in understanding the way that Y is affected as X change. In this type of situation, we wish to estimate f, but our goal is not necessarily to make the predictions for Y. Here we are more interested in understanding the relationship between the X and Y. Now f cannot be treated as the black box, because we need to know it's exact form. This is *inference*.

## Machine Learning             Statistics

Prediction                    Inference

Inference                     Prediction
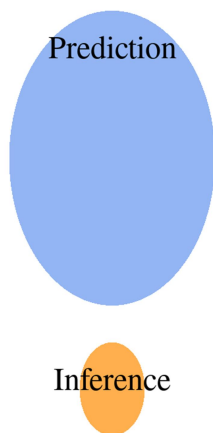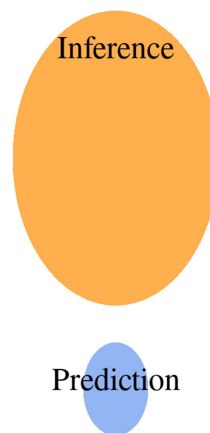
**Parametric & Non-parametric methods**

**Parametric statistics**: This statistical tests based on underlying the assumptions about data's distribution. In other words, It is based on the parameters of the normal curve. Because parametric statistics are based on the normal curve, data must meet certain assumptions, or parametric statistics cannot be calculated. Before running any parametric statistics, you should always be sure to test the assumptions for the tests that you are planning to run.

$f(X) = β0 + β1X1 + β2X2 + . . . + βpXp$

As by the name, nonparametric statistics are not based on parameters of the normal curve. Therefore, if our data violate the assumptions of a usual parametric and nonparametric statistics might better define the data, try running the nonparametric equivalent of the parametric test. We should also consider using nonparametric equivalent tests when we have limited sample sizes (e.g., n < 30). Though the nonparametric statistical tests have more flexibility than do parametric statistical tests, nonparametric tests are not as robust; therefore, most statisticians recommend that when appropriate, parametric statistics are preferred.
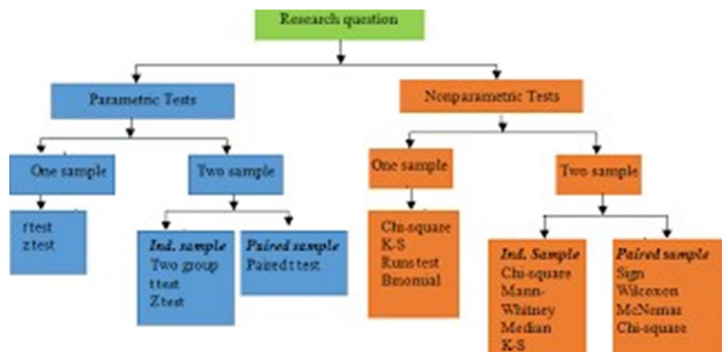


**Prediction Accuracy and Model Interpretability:**

Out of many methods that we use for the statistical learning, some are less flexible and more restrictive . When inference is the goal, then there are clear advantages of using the simple and relatively inflexible statistical learning methods. When we are only interested in the prediction, we use flexible models available.
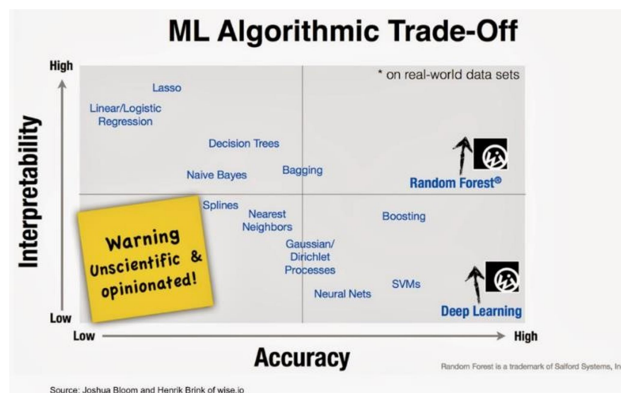
## Q2. What is ANOVA?

**Answer:**

**ANOVA**: it stands for " Analysis of Variance " is an extremely important tool for analysis of data (both One Way and Two Way ANOVA is used). It is a statistical method to compare the population means of two or more groups by analyzing variance. The variance would differ only when the means are significantly different.

ANOVA test is the way to find out if survey or experiment results are significant. In other words, It helps us to figure out if we need to reject the null hypothesis or accept the alternate hypothesis. We are testing groups to see if there's a difference between them. Examples of when we might want to test different groups:

- The group of psychiatric patients are trying three different therapies: counseling, medication, and biofeedback. We want to see if one therapy is better than the others.

- The manufacturer has two different processes to make light bulbs if they want to know which one is better.

- Students from the different colleges take the same exam. We want to see if one college outperforms the other.
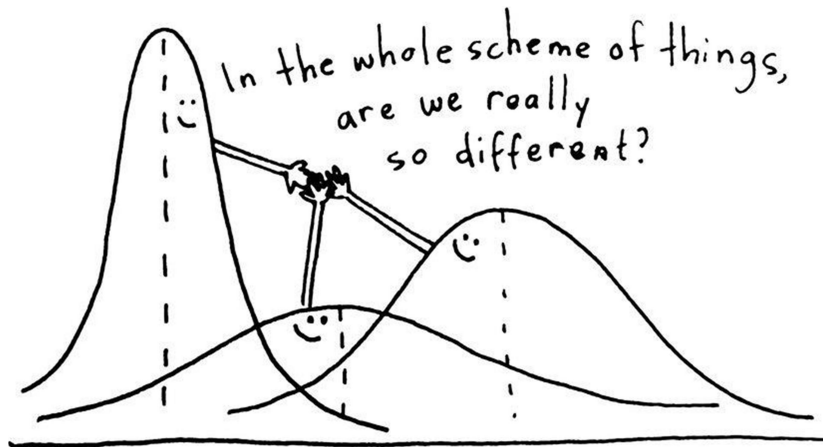


**Types of ANOVA:**
- One-way ANOVA
- Two-way ANOVA

One-way ANOVA is the hypothesis test in which only one categorical variable or the single factor is taken into consideration. With the help of F-distribution, it enables us to compare means of three or more samples. The Null hypothesis (H0) is the equity in all population means while an Alternative hypothesis is the difference in at least one mean.
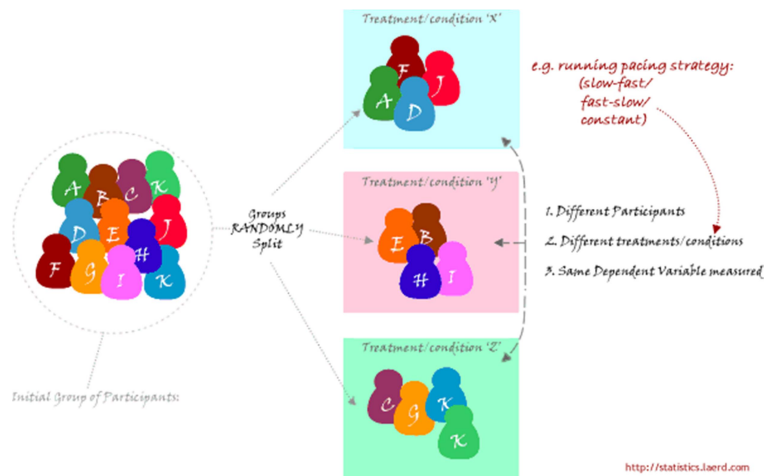
There are two-ways ANOVA examines the effect of two independent factors on a dependent variable. It also studies the inter-relationship between independent variables influencing the values of the dependent variable, if any.



## Q3. What is ANCOVA?

**Answer:**

**Analysis of Covariance (ANCOVA):** It is the inclusion of the continuous variable in addition to the variables of interest ( the dependent and independent variable) as means for the control. Because the ANCOVA is the extension of the ANOVA, the researcher can still assess main effects and the interactions to answer their research hypotheses. The difference between ANCOVA and an ANOVA is that an ANCOVA model includes the "covariate" that is correlated with dependent variable and means on dependent variable are adjusted due to effects the covariate has on it. Covariates can also

be used in many ANOVA based designs: such as between-subjects, within-subjects (repeated measures), mixed (between – and within – designs), etc. Thus, this technique answers the question

In simple terms, The difference between ANOVA and the ANCOVA is the letter "C", which stands for 'covariance'. Like ANOVA, "Analysis of Covariance" (ANCOVA) has the single continuous response variable. Unlike ANOVA, ANCOVA compares the response variable by both the factor and a continuous independent variable (example comparing test score by both 'level of education' and the 'number of hours spent in studying'). The terms for the continuous independent variable (IV) used in the ANCOVA is "covariate".

Example of ANCOVA

## ANCOVA EXAMPLE

### Independent Variables

(Factor)

Level of Education
(High School, College Degree,
or Graduate Degree)

(Covariate)

Number of Hours
Spent Studying

### Dependent Variable

(Response)

Test Score

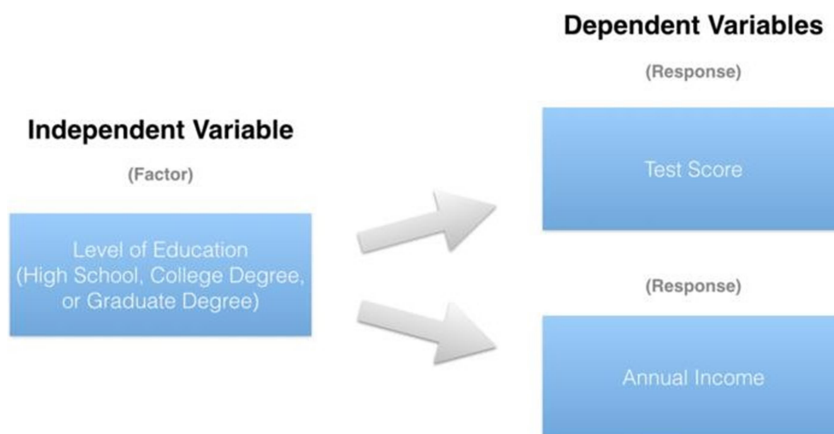Between-group variance → Group | Error | Cov ← Within-group variance
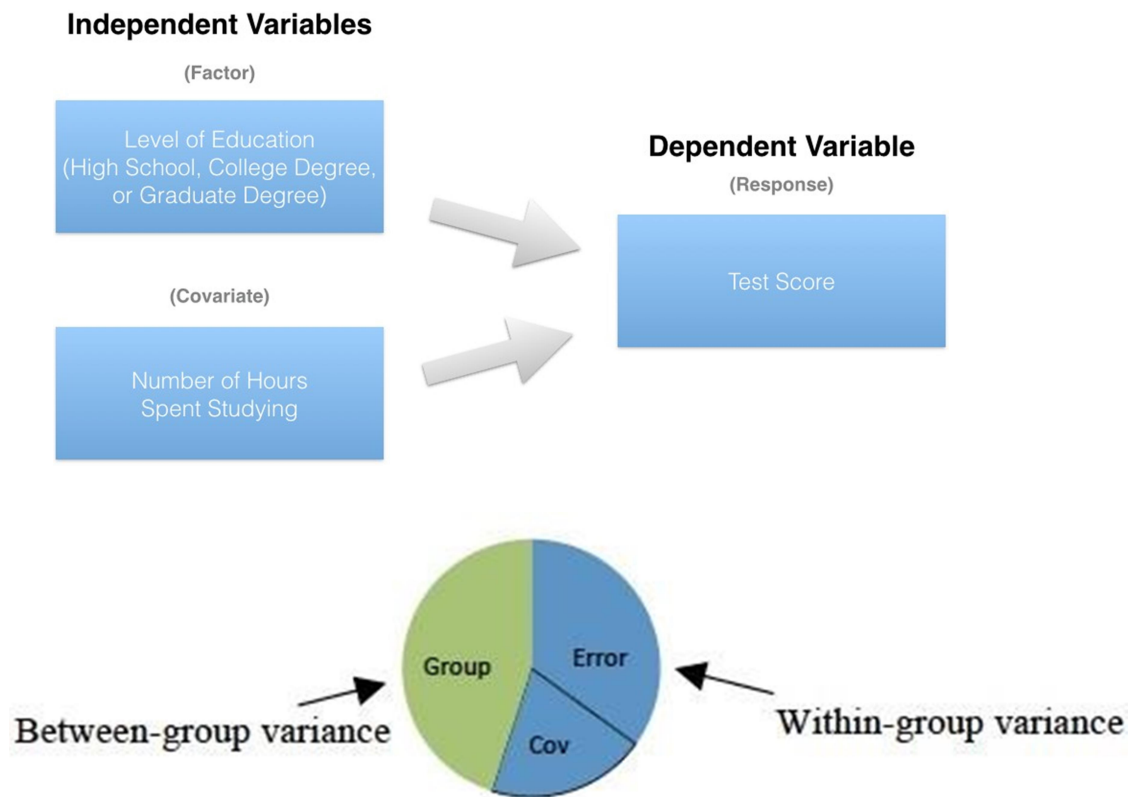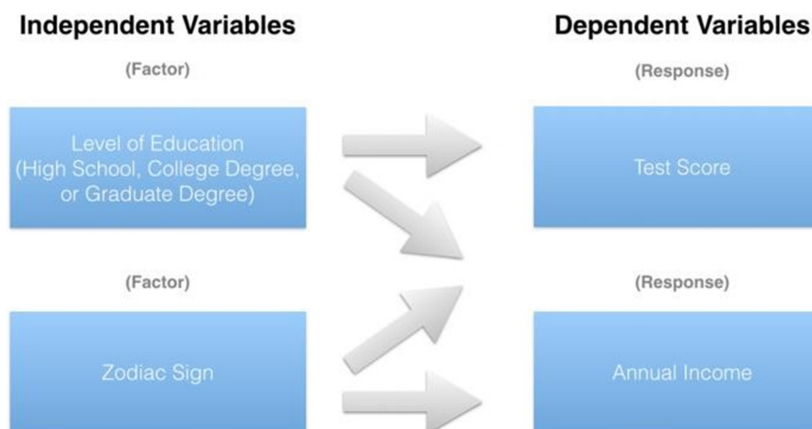
## Q4. What is MANOVA?

**Answer:**

**MANOVA (multivariate analysis of variance)**: It is a type of multivariate analysis used to analyze data that involves more than one dependent variable at a time. MANOVA allows us to test hypotheses regarding the effect of one or more independent variables on two or more dependent variables.

The obvious difference between ANOVA and the "Multivariate Analysis of Variance" (MANOVA) is the "M", which stands for multivariate. In basic terms, MANOVA is an ANOVA with two or more continuous response variables. Like ANOVA, MANOVA has both the one-way flavor and a two-way flavor. The number of factor variables involved distinguish the one-way MANOVA from a two-way MANOVA.

### ONE-WAY MANOVA EXAMPLE

**Dependent Variables**
(Response)

**Independent Variable**
(Factor)

Level of Education
(High School, College Degree, or Graduate Degree)

Test Score

(Response)

Annual Income

### TWO-WAY MANOVA EXAMPLE

**Independent Variables**
(Factor)

**Dependent Variables**
(Response)

Level of Education
(High School, College Degree, or Graduate Degree)

Test Score

(Factor)

(Response)

Zodiac Sign

Annual Income

When comparing the two or more continuous response variables by the single factor, a one-way MANOVA is appropriate (e.g. comparing 'test score' and 'annual income' together by 'level of

education'). The two-way MANOVA also entails two or more continuous response variables, but compares them by at least two factors (e.g. comparing 'test score' and 'annual income' together by both 'level of education' and 'zodiac sign').
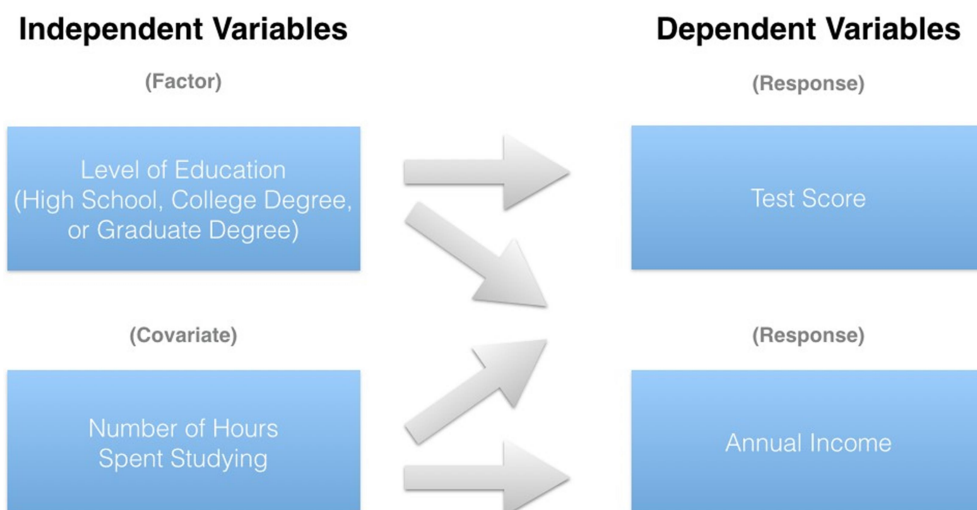
## Q5. What is MANCOVA?

**Answer:**

**Multivariate analysis of covariance (MANCOVA)**: It is a statistical technique that is the extension of analysis of covariance (ANCOVA). It is the multivariate analysis of variance (MANOVA) with a covariate(s).). In MANCOVA, we assess for statistical differences on multiple continuous dependent variables by an independent grouping variable, while controlling for a third variable called the covariate; multiple covariates can be used, depending on the sample size. Covariates are added so that it can reduce error terms and so that the analysis eliminates the covariates' effect on the relationship between the independent grouping variable and the continuous dependent variables.

ANOVA and ANCOVA, the main difference between the MANOVA and MANCOVA, is the "C," which again stands for the "covariance." Both the MANOVA and MANCOVA feature two or more response variables, but the key difference between the two is the nature of the IVs. While the MANOVA can include only factors, an analysis evolves from MANOVA to MANCOVA when one or more covariates are added to the mix.

## MANCOVA EXAMPLE

**Independent Variables**

(Factor)

Level of Education
(High School, College Degree,
or Graduate Degree)

(Covariate)

Number of Hours
Spent Studying

**Dependent Variables**

(Response)

Test Score

(Response)

Annual Income

## Q6. Explain the differences between KNN classifier and KNN regression methods.

**Answer:**

They are quite similar. Given a value for $KK$ and a prediction point $x0x0$, KNN regression first identifies the $KK$ training observations that are closes to $x0x0$, represented by N0. It then estimates f(x0) using the average of all the training responses in N0. In other words,

$$\hat{f}(x_0) = \frac{1}{K} \sum_{x_i \in N_0} y_i$$

So the main difference is the fact that for the classifier approach, the algorithm assumes the outcome as the class of more presence, and on the regression approach, the response is the average value of the nearest neighbors.
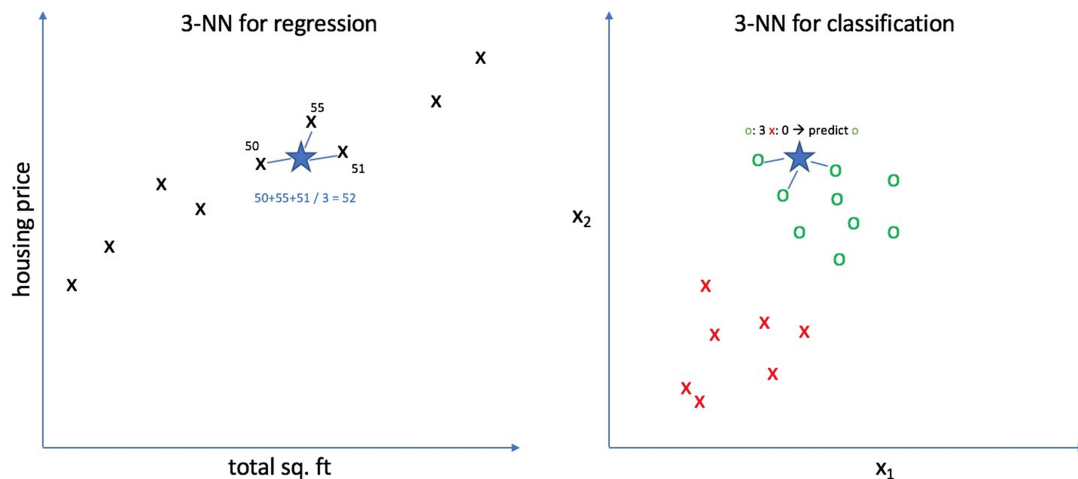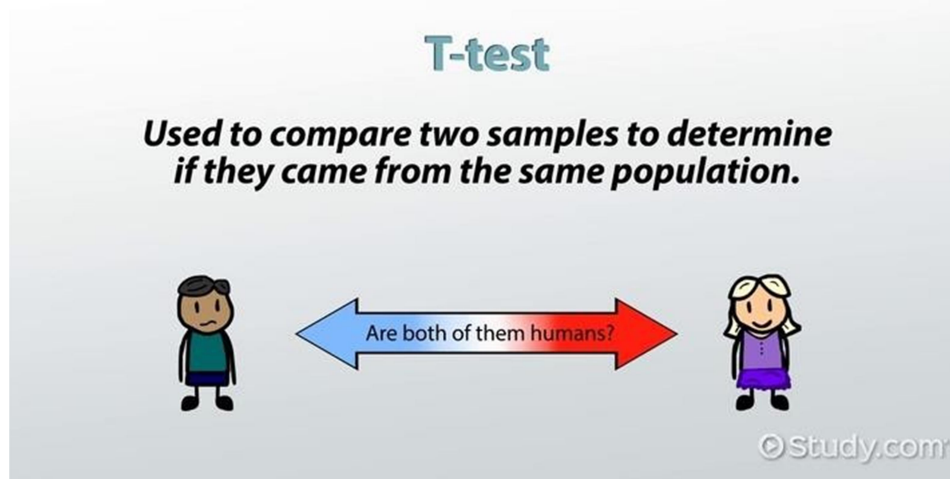


## Q7. What is t-test?

**Answer:**

To understand T-Test Distribution, Consider the situation, you want to compare the performance of two workers of your company by checking the average sales done by each of them, or to compare the performance of a worker by comparing the average sales done by him with the standard value. In such situations of daily life, t distribution is applicable.

A t-test is the type of inferential statistic used to determine if there is a significant difference between the means of two groups, which may be related in certain features. It is mostly used when the data sets, like the data set recorded as the outcome from flipping a coin 100 times, would follow a normal distribution and may have unknown variances. A t-test is used as a hypothesis testing tool, which allows testing of an assumption applicable to a population.

**Understand t-test with Example:** Let's say you have a cold, and you try a naturopathic remedy. Your cold lasts a couple of days. The next time when you have a cold, you buy an over-the-counter pharmaceutical, and the cold lasts a week. You survey your friends, and they all tell you that their colds were of a shorter duration (an average of 3 days) when they took the homeopathic remedy. What you want to know is, are these results repeatable? A t-test can tell you by comparing the means of the two groups and letting you know the probability of those results happening by chance.

| Type | T-statistic | Degrees of freedom |
|------|-------------|--------------------|
| One-sample t-test | $t = \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}}$ | df = n – 1 |
| Paired t-test | $t = \dfrac{\bar{X}_D - \mu_0}{s_D/\sqrt{n}}$ . | df = n – 1 |

OVERVIEW OF T-TESTS

# T-test

**Used to compare two samples to determine if they came from the same population.**

Are both of them humans?

©Study.com

## Q8. What is Z-test?

**Answer:**

**z-test:** It is a statistical test used to determine whether the two population means are different when the variances are known, and the sample size is large. The test statistic is assumed to have the normal distribution, and nuisance parameters such as standard deviation should be known for an accurate z-test to be performed.

Another definition of Z-test: A Z-test is a type of hypothesis test. Hypothesis testing is just the way for you to figure out if results from a test are valid or repeatable. Example, if someone said they had found the new drug that cures cancer, you would want to be sure it was probably true. Hypothesis test will tell you if it's probably true or probably not true. A Z test is used when your data is approximately normally distributed.

**Z-Tests Working :**

Tests that can be conducted as the z-tests include one-sample location test, a two-sample location test, a paired difference test, and a maximum likelihood estimate. Z-tests are related to t-tests, but t-tests are best performed when an experiment has the small sample size. Also, T-tests assumes the standard deviation is unknown, while z-tests assumes that it is known. If the standard deviation of the population is unknown, then the assumption of the sample variance equaling the population variance is made.

**When we can run the Z-test :**

Different types of tests are used in the statistics (i.e., f test, chi-square test, t-test). You would use a Z test if:

- Your sample size is greater than 30. Otherwise, use a t-test.
- Data points should be independent from each other. Some other words, one data point is not related or doesn't affect another data point.
- Your data should be normally distributed. However, for large sample sizes (over 30), this doesn't always matter.
- Your data should be randomly selected from a population, where each item has an equal chance of being selected.
- Sample sizes should be equal, if at all possible.

Z-TEST

Formula to find the value of Z (z-test) Is:

$$Z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

$\bar{x}$ = mean of sample
$\mu_0$ = mean of population
$\sigma$ = standard deviation of population
n = no. of observations

## Q9. What is Chi-Square test?

**Answer:**

**Chi-square ($\chi^2$) statistic**: It is a test that measures how expectations compare to actual observed data (or model results). The data used in calculating a chi-square statistic must be random, raw, mutually exclusive, drawn from independent variables, and drawn from a large enough sample. For example, the results of tossing a coin 100 times meet these criteria.

Chi-square test is intended to test how it is that an observed distribution is due to chance. It is also called the **"goodness of fit"** statistic because it measures how well the observed distribution of the data fits with the distribution that is expected if the variables are independent.

Chi-square test is designed to analyze the **categorical** data. That means that the data has been counted and divided into categories. It will not work with parametric or continuous data (such as height in inches). For example, if you want to test whether attending class influences how students perform on an exam, using test scores (from 0-100) as data would not be appropriate for a Chi-square test. However, arranging students into the categories "Pass" and "Fail" would. Additionally, the data in a Chi-square grid should not be in the form of percentages, or anything other than frequency (count) data.

$$\chi^2_c = \sum \frac{(O_i - E_i)^2}{E_i}$$
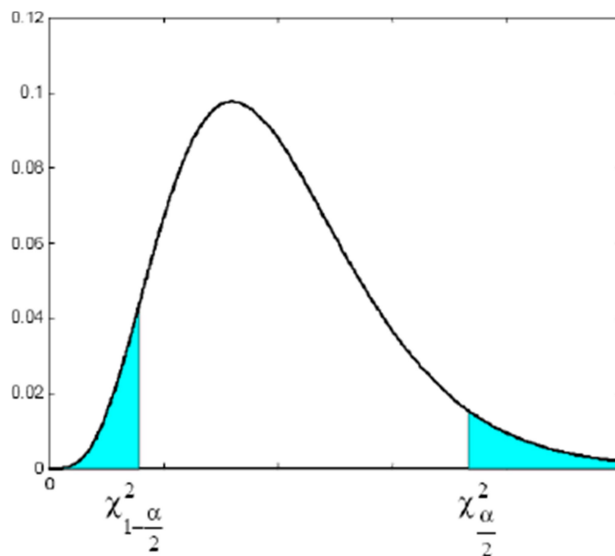
## Q10. What is correlation and the covariance in the statistics?

**Answer:**

The Covariance and Correlation are two mathematical concepts; these two approaches are widely used in the statistics. Both Correlation and the Covariance establish the relationship and also measures the dependency between the two random variables, the work is similar between these two, in the mathematical terms, they are different from each other.

**Correlation:** It is the 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 relationship isn't perfect. People of the same height vary in weight, and you can easily think of two people you know where the shorter one is heavier than the taller one. Nonetheless, the average weight of people 5'5" is less than the average weight of people 5'6", and their average weight is less than that of people 5'7", etc. Correlation can tell you just how much of the variation in peoples' weights is related to their heights.

### Correlation Formula

$$\rho_{xy} = \frac{Con(r_x, r_y)}{\sigma_x \sigma_y}$$

**Covariance:** It measures the directional relationship between the returns on two assets. The positive covariance means that asset returns move together while a negative covariance means they move inversely. Covariance is calculated by analyzing at-return surprises (standard deviations from the expected return) or by multiplying the correlation between the two variables by the standard deviation of each variable.

### Covariance Formula

**For Population**

$$Cov(x,y) = \frac{\Sigma (x_i - \bar{x}) * (y_i - \bar{y})}{N}$$

**For Sample**

$$Cov(x,y) = \frac{\Sigma (x_i - \bar{x}) * (y_i - \bar{y})}{(N-1)}$$

---------------------------------------------------------------------------------------------------------------------