# ANOVA

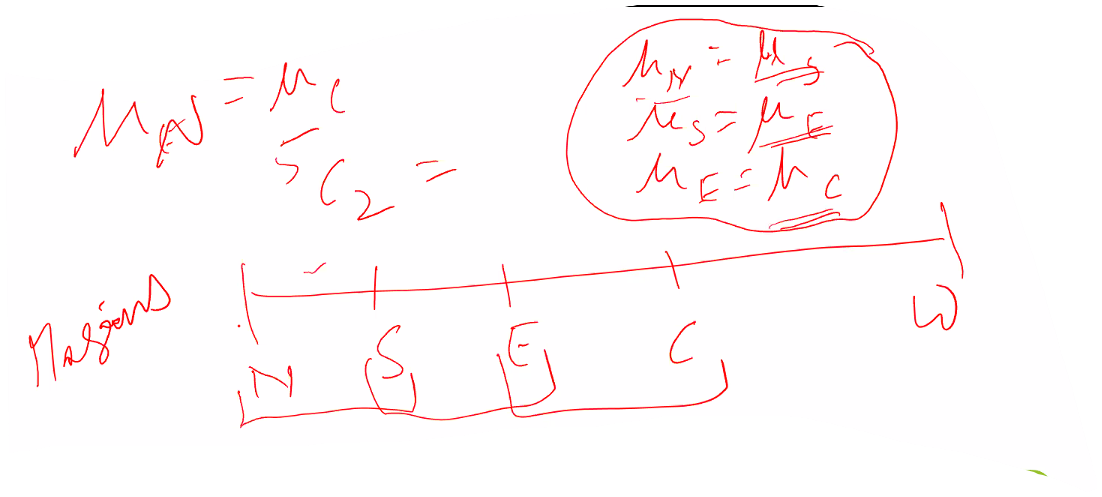
Why ANOVA is needed?

If we do a pairwise test as below it causes the below problem:

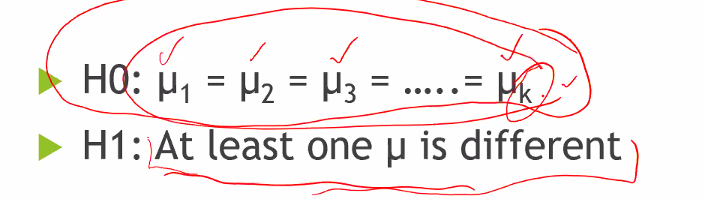
Suppose we have 5 populations. If we do a pair wise test for means we might end up in a situation as below:

1. µ1 = µ2
2. µ2 = µ3
3. µ3 = µ4
4. µ4 = µ5

Which finally may result is a contention that µ1 = µ5 which might not be true if plotted on a number line.



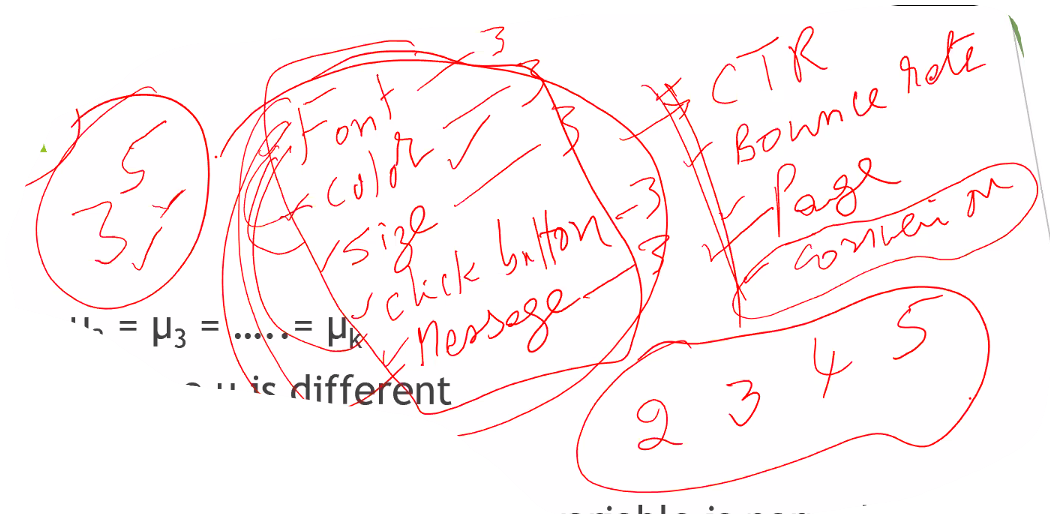
So we need something like a one-shot test to do this for multiple variables.



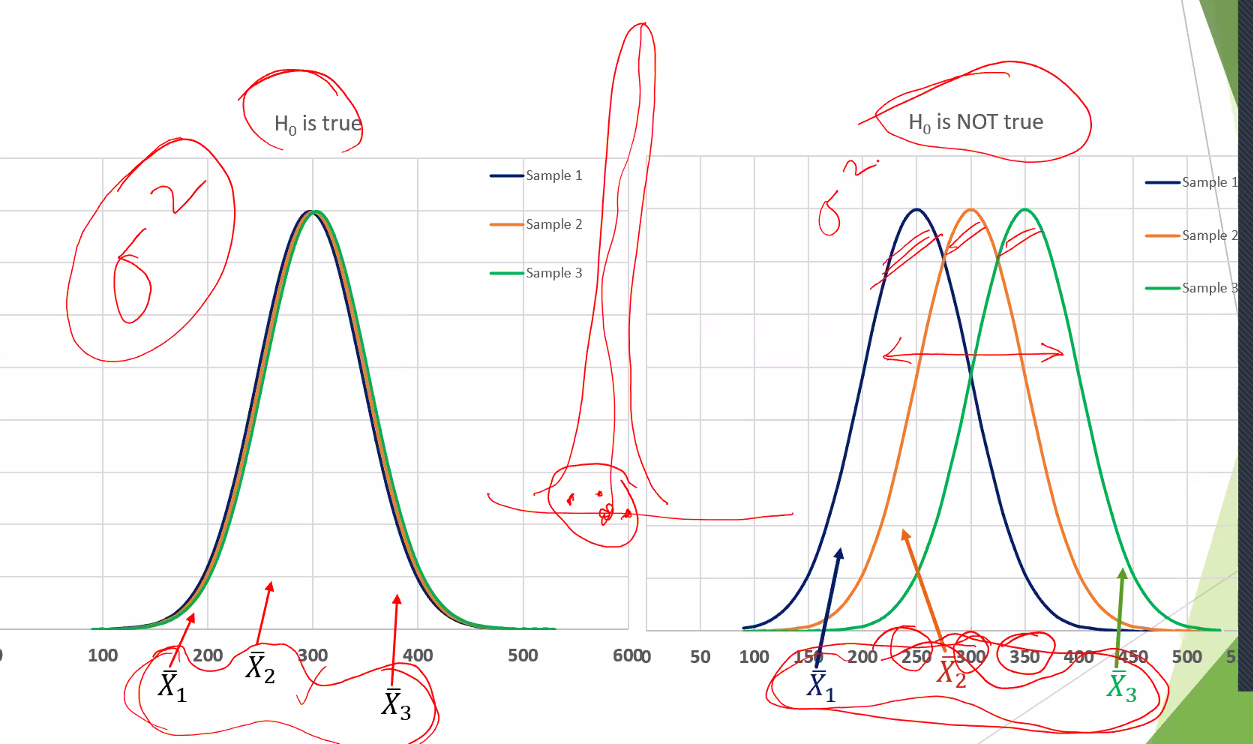
H0 is presumed that all the samples come from same population.

H1: That at least one of the samples come from a different population.

**What is MANOVA?**

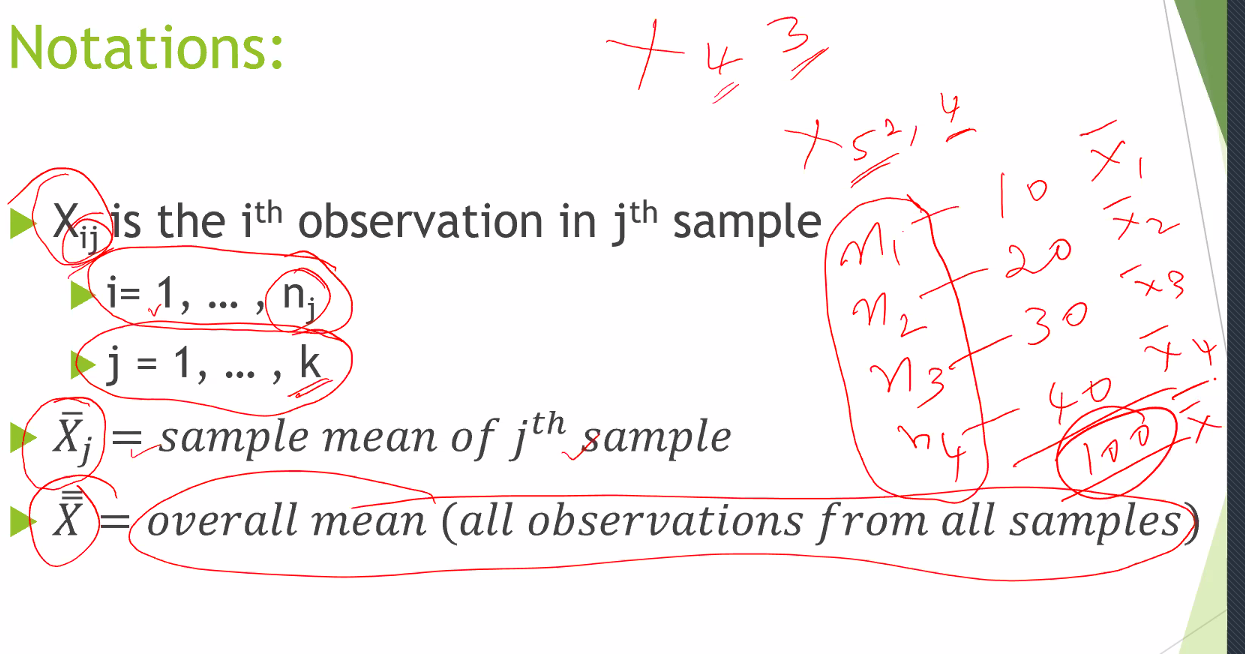
****

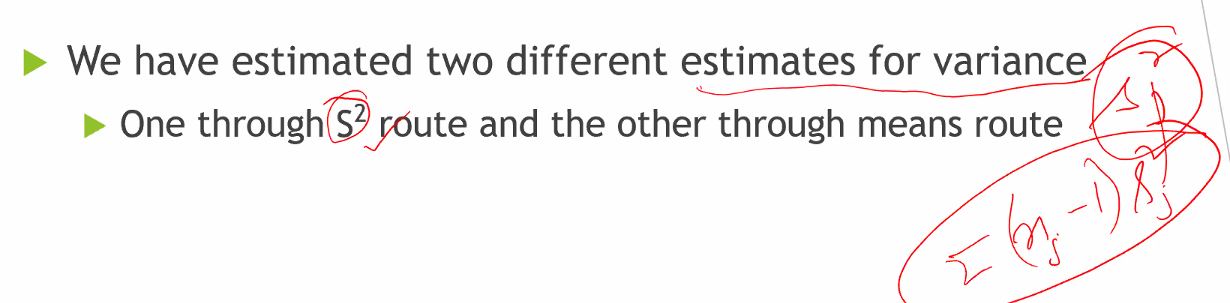
Multiple dependent variables make it MANOVA.

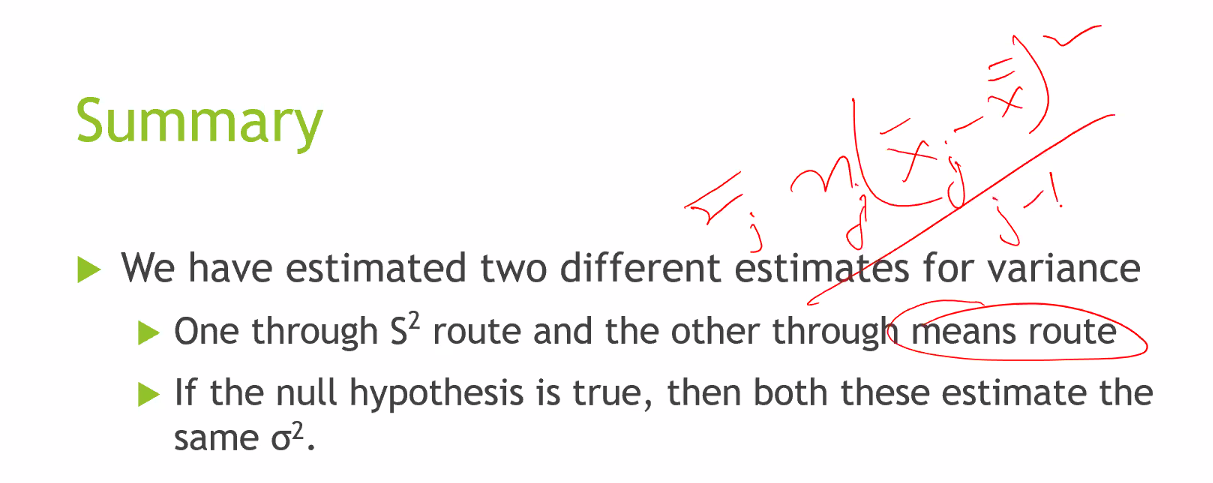
****

If I am operating on the left side H0: The differences between X\_bar1,2,3 the variation is attributable only to the inherent variation of the population.

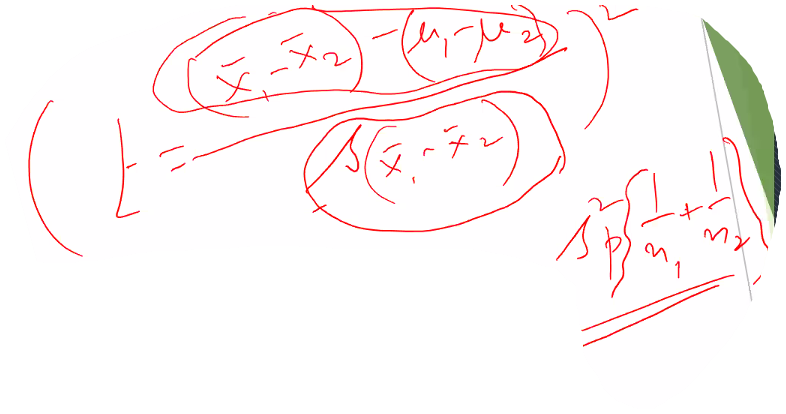
If H1: The differences is due to the variations of both the population variance and also the shifting of the means.

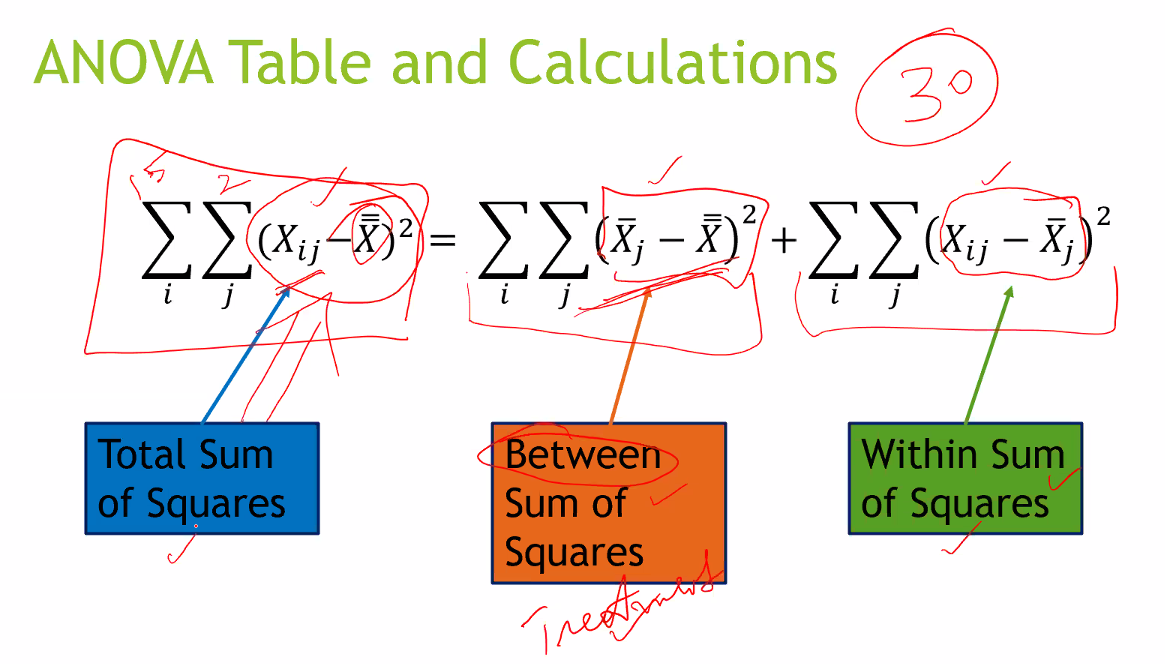






F = t2 if n = 2. (E.g. Chubby Chunky)





Between = Treatment SS2

Within = Error SS2.

Closer between SS2 to 0’s more reason not to reject the H0.

By default it is one sided test.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | Sample 1 | Sample 2 | | … | | Sample k | |
|  | x(1,1) | x(1,2) | |  | | x(1,k) | |
|  | x(2,1) | x(2,2) | |  | | x(2,k) | |
|  | … | … | |  | | … | |
|  | x(n1,1) | x(n2,2) | |  | | x(nk,k) | |
| Sample mean of jth sample |  |  | | … | |  | |
| Overall mean of all observations from all samples or the **Grand mean** |  | | | | | | |
| Sample variance of each sample w.r.t individual means or within **group variances**. |  | |  | |  | |  |
| Total variance of all sample means w.r.t grand mean mean (X̄̄) or **Between group variances.** |  | | | | | | |

## Main Principle

The main principle behind ANOVA is as follows:

1. If we have k samples then we can set the H0 and H1 as follows:
   1. H0: µ1 = µ2 = µ3 =… = µk
   2. H1: at least 2 of the population means from which the samples were drawn is different. There can be many possibilities by which it can happen. ANOVA tests this hypothesis only. How?
2. How can we understand that we have to reject H0 because H1 is true? The following will happen if H1 is true:
   1. Each sample values will have their individual variances ( w.r.t individual means . The window is defined by .
   2. Each of the k samples means will have variances w.r.t the grand mean defined by window and is denoted by
   3. Then we can tell that if H1 is true then average of (within group variances) will be small than (between group variance).

Not necessary that n1 = n2 = … = nk.. They are expected to be unequal.

ANOVA principle says that when the means are not equal, then the **average** **error** **deviation** is likely to be relatively smaller as compared to the **average treatment deviation.** Page 404 of the book **Complete Business Statistics** shows the idea using a small sample list. Now the main problem with average error deviations is that it sums up to 0. So similar to variance we will square them up and take the average.

## Definitions

1. **Error deviation e(i,j)** = Difference between a datapoint x(i,j) from the respective sample mean where 1 ≤ j ≤ k. So e(i,j) = x(i,j) - .
2. **Treatment deviation tj =** Difference between a given sample mean from the grand mean =
3. **Total deviation t(i,j) =** Deviation of the value of a given sample from the grand mean =
4. It can be shown that t(i,j) = tj + e(i,j). Note here we will take algebraic deviations and not absolute values. Book page 407
5. Thus **Total deviation = Treatment deviation + Error deviation**.
6. Now tj + e(i,j) = + {x(i,j) - }. Now term cancels out resulting in only

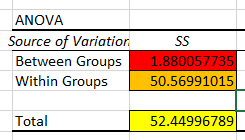
. In other words the total deviation of a point **t(i, j)** is nothing but the total deviation of the point from the grand mean.

1. Now .
2. Also . Now this is not equal to 7 if a single point is considered.
3. We can simulate using any random values on excel that:

Sum of total deviation of all observations2 = . View Calculation templates.xlsx ANOVA tab “Proof of relationship between Total deviation, Treatment deviation and Error deviation”

Thus **Sum of squares total (SST) = Sum of squares for treatment (SSTR) (or between groups) + Sum of squares for errors within (SSE) or within groups**

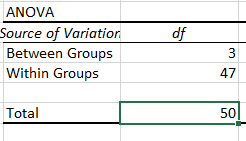
Or the more complicated way of writing to give a professional look :

as shown here: 

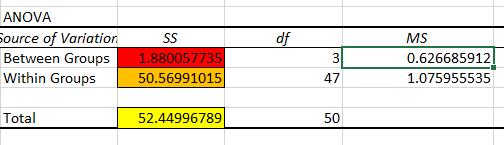
1. What are the terms mean in the data:
   1. SSTR: Explained variation. This is the part of the total variation that is explained by the fact that the data points belong to **several different groups.** The term is present because sum of square deviations of the treatment means **tj** from the grand mean is to be counted for all the observations in a particular sample.
   2. SSE: **Sum of squares within** i.e. the within group variations. SSE is that part of the variation within the data that is due to error - the part that cannot be explained by differences among the groups.
   3. SST: Total amount of variations within the whole dataset.

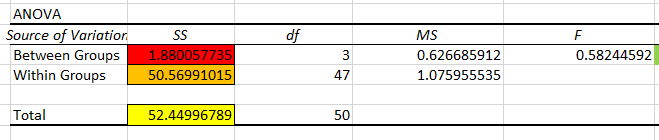
## Interpretation of the columns in ANOVA table

1. The degrees of freedom column:
   1. To know the variations between groups where we had k groups (or k samples) the degrees of freedom = k-1
   2. To know the variations of data within group we required n1 + n2 + … + nk = n amount of data but computed separately for each group. Thus the degrees of freedom will be 1 lost from each of the k groups. Thus the df = n - k.
   3. To know the variations of all the data points from the grand mean X̄̄ required the usage of all the n data points in our data set thereby losing 1 degree of freedom. So df = n - 1.
   4. Thus dftotal = dftreatment + dferror as shown here:



1. Mean squares: This is totally straight forward once we understood the rationale behind calculations of between groups SS and within groups SS.
   1. MSTR (Mean square treatment or Mean square Between group) = SSTR/dftreatment
   2. MSE (Mean square error or mean square error within group) = SSE/dferror. This is shown below:



1. F value = MSTR/MSE 
2. P value = F.DIST.RT(F value, dftreatment, dferror). Why right tailed rejection region? Let’s go back to the quadrants of operations for hypothesis testing.

If H0: Samples are from same population is true, Mean square error between groups <= Mean square error within groups (because all the means of each sample will lie closer to one another around the grand mean X̄̄). Thus F value becomes <= 1. So H0 will be rejected if and only if we get a sample whose F value i.e. the ratio of MSTR/MSE is > 1 as determined by the confidence interval. Hence right hand side rejection region.