Dimensionless Technologies Private Limited
Visit us at: www.dimensionless.in
info@dimensionless.in
9923170071, 8108094992

Notes for Students - Lesson 12

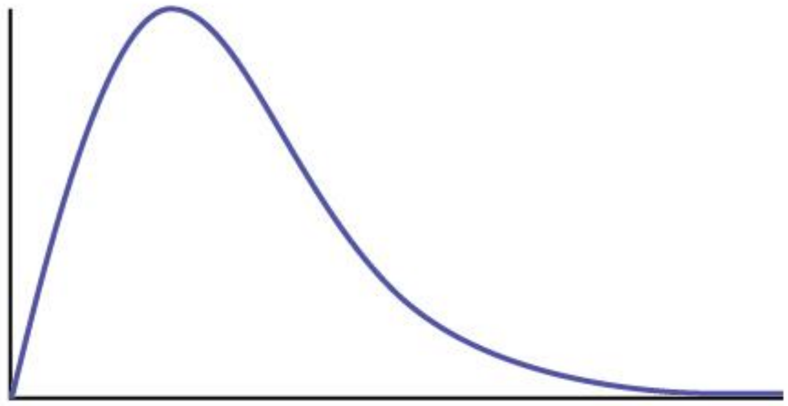
ANOVA

Till now, we saw how we can use t test to compare two sample means.

However, we may be interested to compare more than two sample means as well. In that scenario, we have to perform multiple t-test.

We can compare means of two or more samples with the help of Analysis of Variance (ANOVA). ANOVA can determine whether the means of two or more groups are different. ANOVA uses F test to statistically test equality of means.

F test is a ratio of two variances which represent measure of dispersion. Large F value represents greater dispersion. It uses Between Group Variability and Within Group Variability in it’s ratios.



F distribution

Assumptions of F Distribution

- The populations from which the samples are drawn are normally distributed
- The two populations are independent of each other.

Variation between samples / Between Group Variability

Q. Suppose you want to compare the prices of Sam's cafe and Bob's cafe.

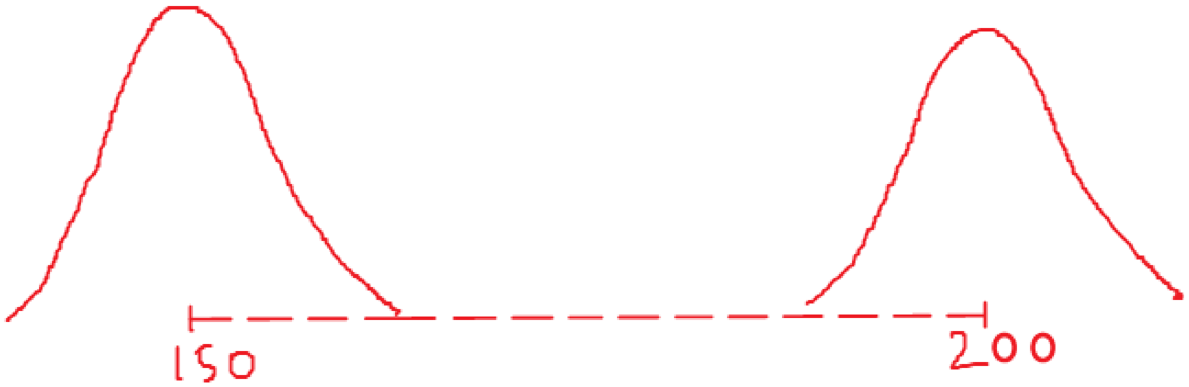
In what case, would you expect the price of both to differ significantly?

a) The greater the difference between their sample means.

b) The smaller the difference between their sample means.

The greater the difference between sample mean prices of Sam's cafe and Bob's cafe, the more likely the populations means prices in both outlets would differ significantly.

The smaller the difference between sample mean prices of Sam's cafe and Bob's cafe, the less likely the population means would differ significantly.



You have a sample mean price (in Rs.) for Sam's cafe as 150 and Bob's cafe as 200. Which café would you choose if you wish to decide on the basis of price?

a) **Sam's cafe**

b) **Bob's cafe**

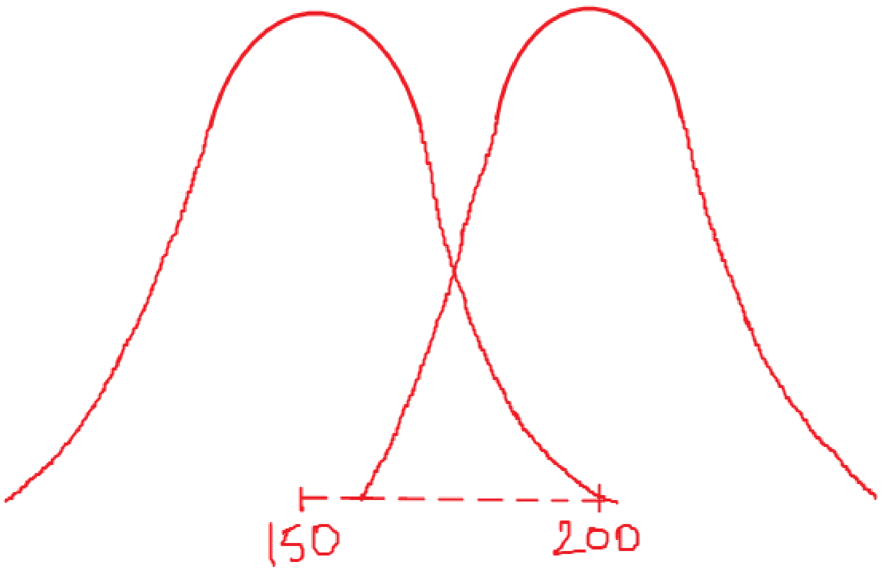
Variation within samples / Within Group Variability

You chose Sam's cafe as it costs low mean price. But your friend says that even Sam's cafe might cost you the same as Bob's cafe as there is lot of variation in it's price which ranges from as low as 30 to as high as 300, while the prices in Bob's cafe ranges between 150 to 250.

Q. In that case, would the prices in both cafes differ significantly according to you?

a) Yes

b) No



The greater the variability of prices within each sample, the lesser likely the population mean prices would differ significantly.

The lower the variability of prices within each sample, the more likely the population mean prices would differ significantly.

So, to test if the prices of both cafes differ significantly or not, we would conduct the Analysis of Variance (ANOVA).

ANOVA uses both between group variability and within group variability to test whether the population means are significantly different from each other or not.

Q. What will be the null hypothesis to test whether or not the population means are significantly different?

$$H_o : \mu_1 = \mu_2 = \mu_3$$

Q. What will be the alternate hypothesis?

H_A : All pairs of samples are significantly different

H_A : At least one pair of samples is significantly different

H_A : At most one pair of samples is significantly different

$$F = \frac{\text{between group variability}}{\text{within group varaibility}}$$

Q. Why do you think that between group variability is the numerator and within group variability is the denominator?

Between group variability

$$= n\sum (\overline{X}_K - \overline{X}_G)^2 / df$$

Q. What do you think will be the degrees of freedom for between group variability?

Within Group Variability

$$= \sum (\overline{X}_i - \overline{X}_K)^2 / df$$

Q. What do you think would be the degrees of freedom for within group variability?

Between group variability

$$= n\sum (\overline{X}_K - \overline{X}_G)^2 / (K\text{-}1)$$

(where K is the number of samples)

Within Group Variability

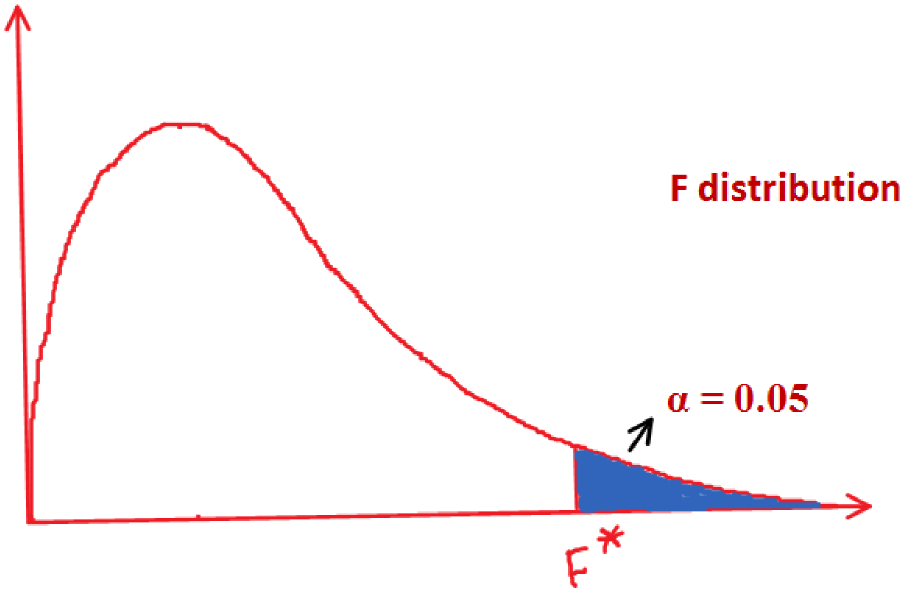
$$= \sum (\overline{X}_i - \overline{X}_K)^2 / (N-K)$$

(where N is the total number of values from all samples)

$$F = \frac{n \sum (\overline{X}_K - \overline{X}_G)^2 / (k-1)}{\sum (\overline{X}_i - \overline{X}_K)^2 / (N-K)}$$

$$F = \frac{SS_{between} / df_{between}}{SS_{within} / df_{within}}$$

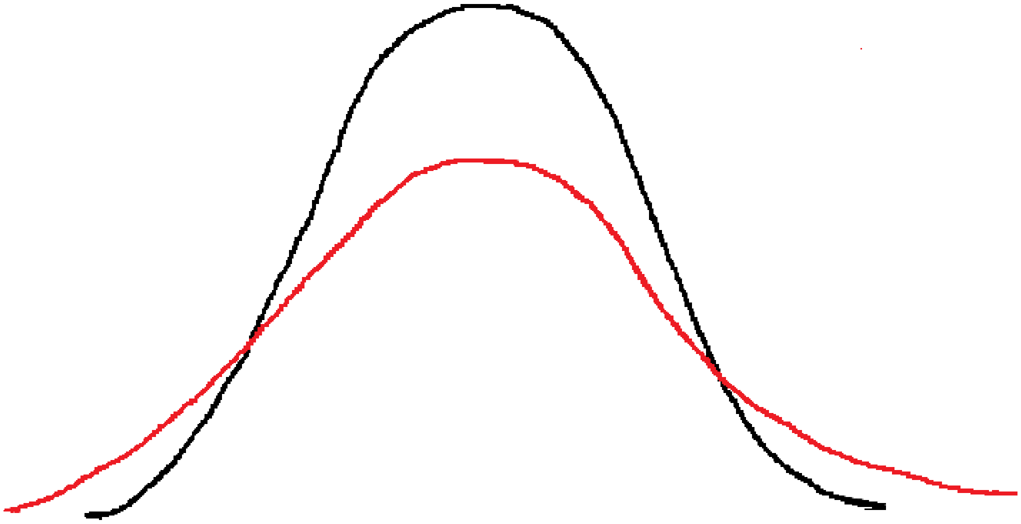
$$F = \frac{MS_{between}}{MS_{within}}$$



Q. Why do you think F distribution is one directional?

F statistic won’t take a negative value since both it’s numerator and denominator will be positive since they are variances.

Q. Would you expect the F value to be high or low?



$$SS_{total} = SS_{between} + SS_{within}$$

$$= \sum (X_i - \bar{X}_G)^2$$

Where $\bar{X}_G$ is the grand mean that is average of all the values taken together.

$$df_{total} = df_{between} + df_{within}$$

$$= N - 1$$

A study wants to check whether the time spent by people on mobile phones differ according to age groups.

10-20	20-30	30-40
5	1	2
6	1	1
4	2	0
3	2	2
5	1	1
4	3	2
2	1	0
4	1	1
5	2	3
6	1	2

Q. What would be the null hypothesis to test if the time spent differs across age groups?

$$H_o : \mu_1 = \mu_2 = \mu_3$$

Q. What would be the alternate hypothesis to test if the time spent differs across age groups?

H_A : At least one pair of age groups spend significantly different time on mobile phones.

Q. Calculate sample mean for all age goups.

$\bar{X}_1 = 4.4$

$\bar{X}_2 = 1.5$

$\bar{X}_3 = 1.4$

Q. Calculate grand mean.

$\bar{X}_4 = 2.43$

Q. Calculate $SS_{between}$

$$n\sum (\overline{X_K} - \overline{X_G})^2 = ?$$

=

Q. $\text{MS}_{\text{between}}$

$$n\sum (\overline{X_K} - \overline{X_G})^2 / (K\text{-}1) = ?$$

=

Q. Calculate $\text{SS}_{\text{within}}$

$$\sum (\overline{X_i} - \overline{X_K})^2 = ?$$

= 27.3

Q. $\text{MS}_{\text{within}}$?

$$\sum (\overline{X_i} - \overline{X_K})^2 / (N\text{-}K) = ?$$

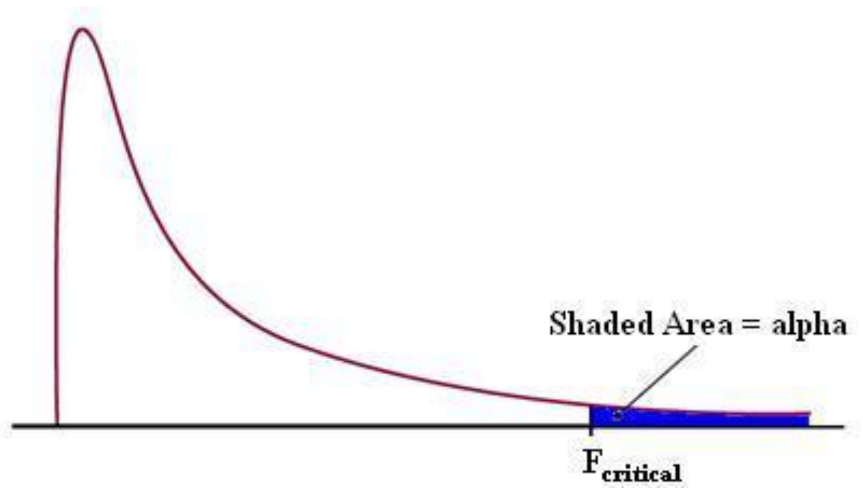
= 1.011

Q. F value?

$$= MS_{between} / MS_{within}$$

= 28.714

Q. What would be the F critical value for α = 0.05?



= 0.05

3.3541

Q. Would you reject or retain the null hypothesis?

Reject

Do Not Reject

What conclusion can be drawn from this observation?

Q. If the variance of an individual sample becomes bigger, all else held constant, does this lean more in favor of the null hypothesis or alternate hypothesis?

Q. As the between group variability increases meaning the sample means get further apart from each other all else held constant. Does this lean in favor of the null hypothesis or alternate hypothesis?

Let's perform ANOVA in Excel

Step 1 : Enter the data in Excel

Step 2 : Data > Data Analysis > Anova Single factor > Enter input range, output range and the value of α

Make an analysis of whether to reject or retain the null hypothesis by comparing the F critical and F value or α and P-value.

Q. 3 schools took part in a debate competition. The students were awarded marks out of 10. However, the judges are facing difficulty in choosing the winner as they feel all teams performed similar. Do you think the teams were really similar?

A	B	C
7	8	7.5
8	7.5	9
8.5	6.5	6
7.5	9.5	8.5

Anova: Single Factor

SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
Column 1	4	31	7.75	0.416667
Column 2	4	31.5	7.875	1.5625
Column 3	4	31	7.75	1.75

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	0.041667	2	0.020833	0.01676	0.983411	4.256495
Within Groups	11.1875	9	1.243056			
Total	11.22917	11				