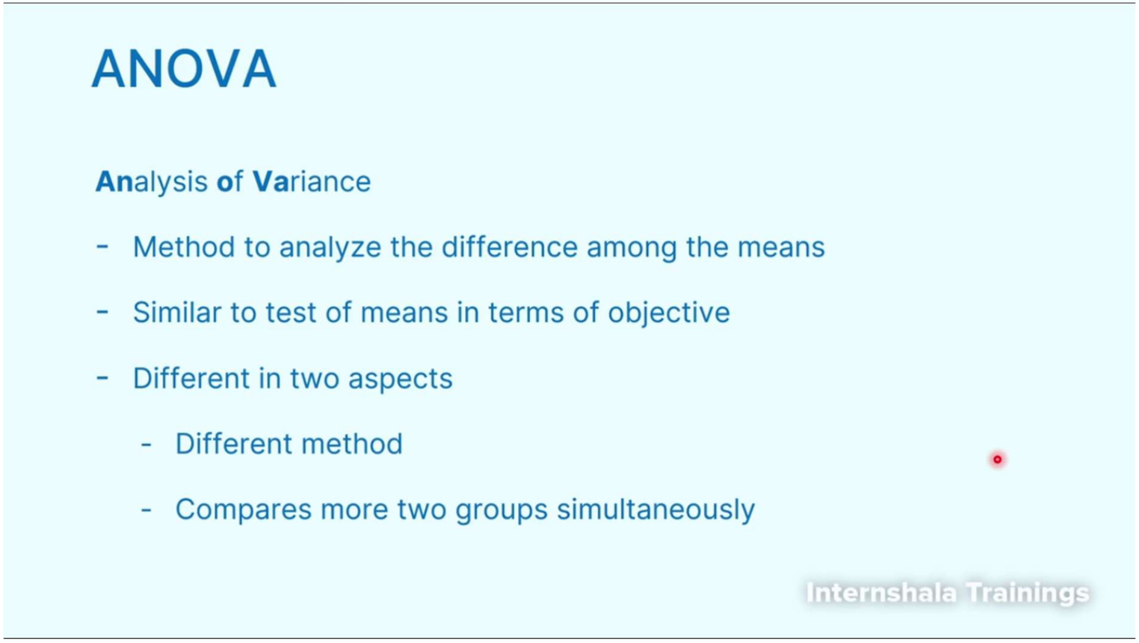
Module 5 Topic 1 Video 1

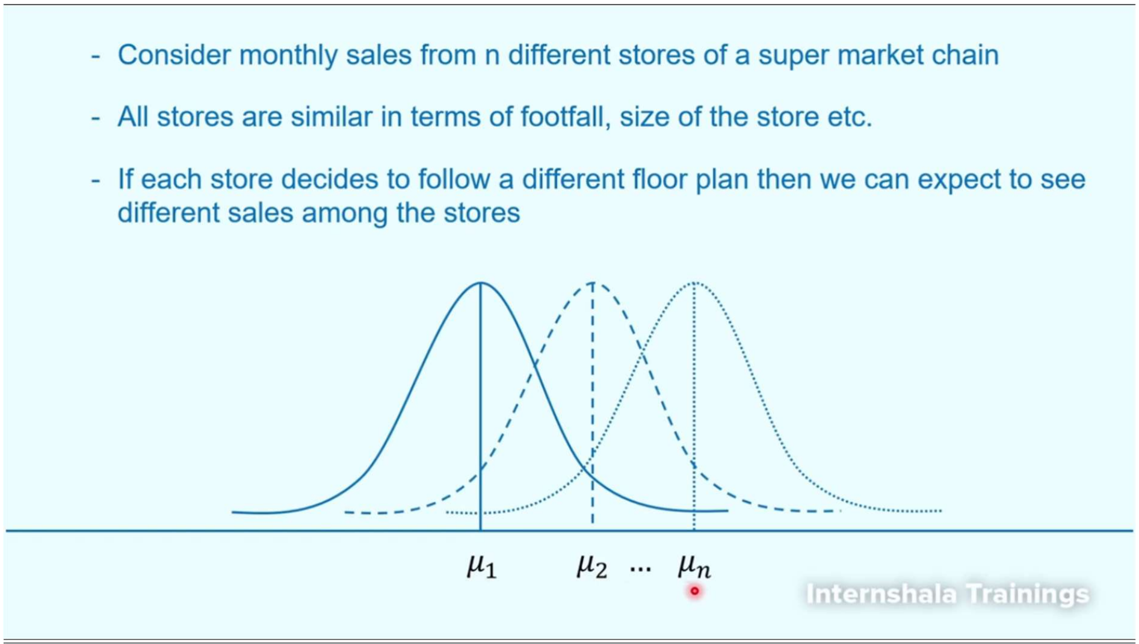
Introduction to ANOVA



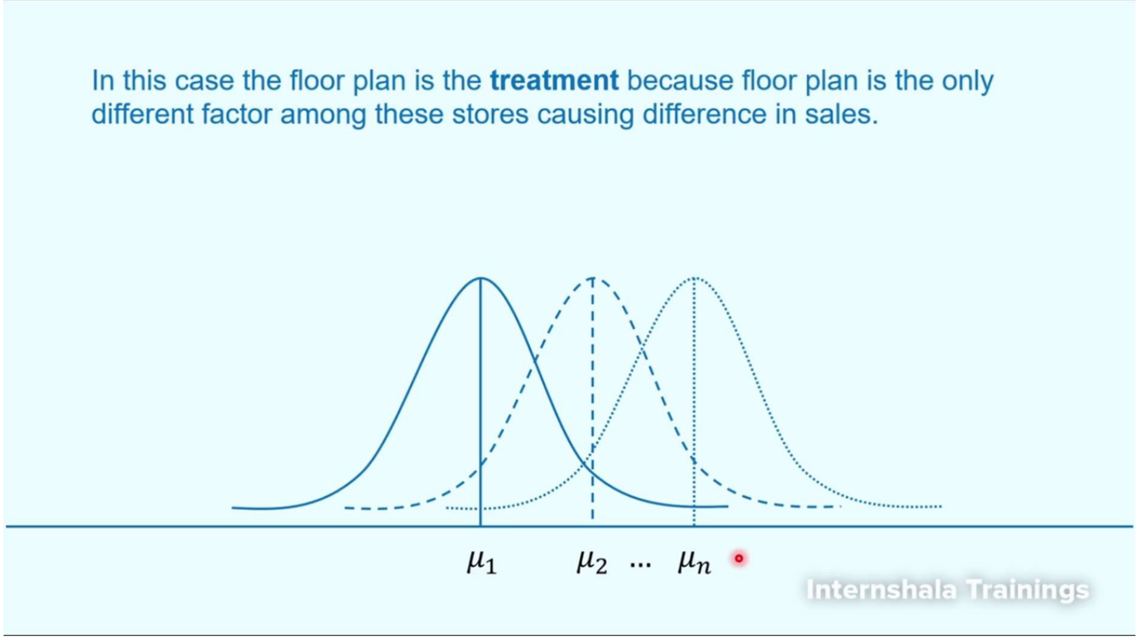
ANOVA

Analysis **o**f **Va**riance

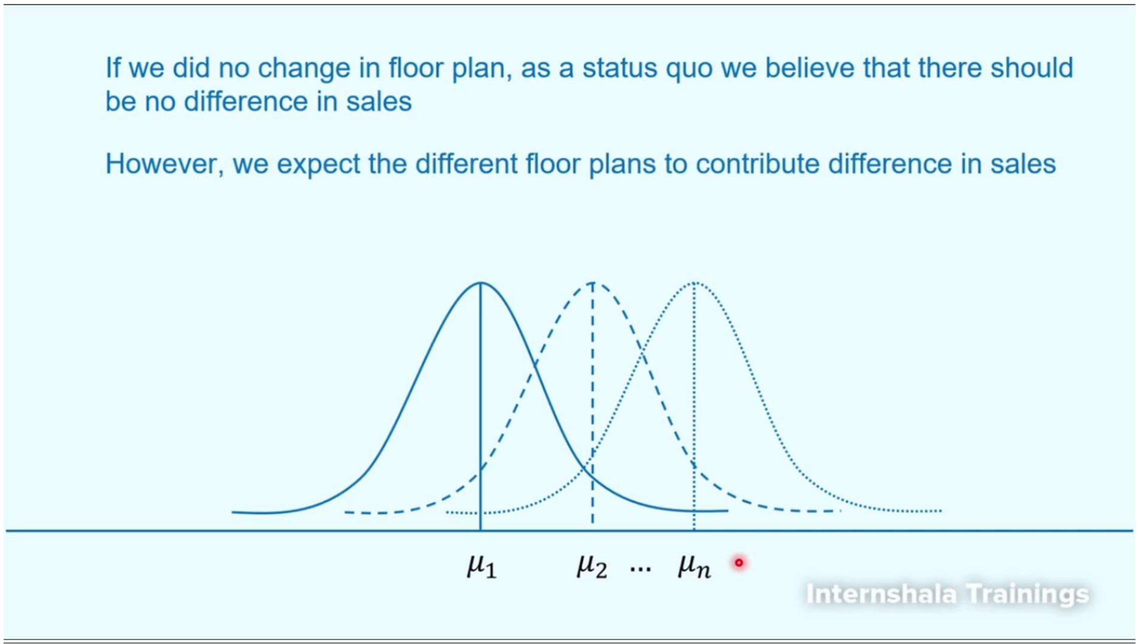
- Method to analyze the difference among the means
- Similar to test of means in terms of objective
- Different in two aspects
  - Different method
  - Compares more two groups simultaneously



- Consider monthly sales from n different stores of a super market chain
- All stores are similar in terms of footfall, size of the store etc.
- If each store decides to follow a different floor plan then we can expect to see different sales among the stores

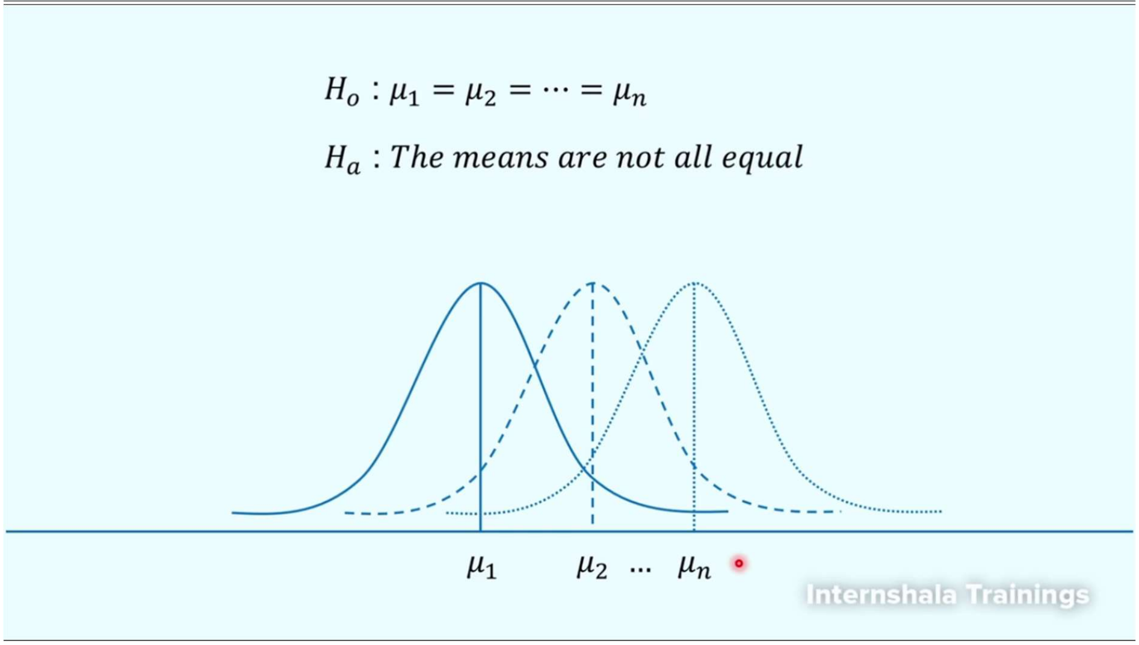


In this case the floor plan is the **treatment** because floor plan is the only different factor among these stores causing difference in sales.



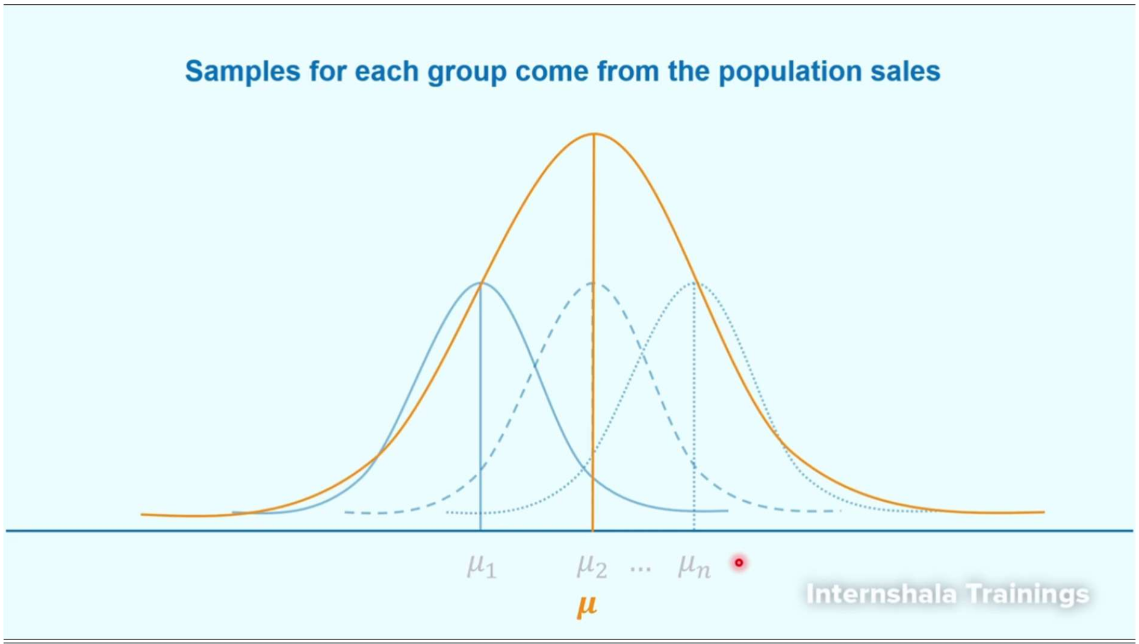
If we did no change in floor plan, as a status quo we believe that there should be no difference in sales

However, we expect the different floor plans to contribute difference in sales

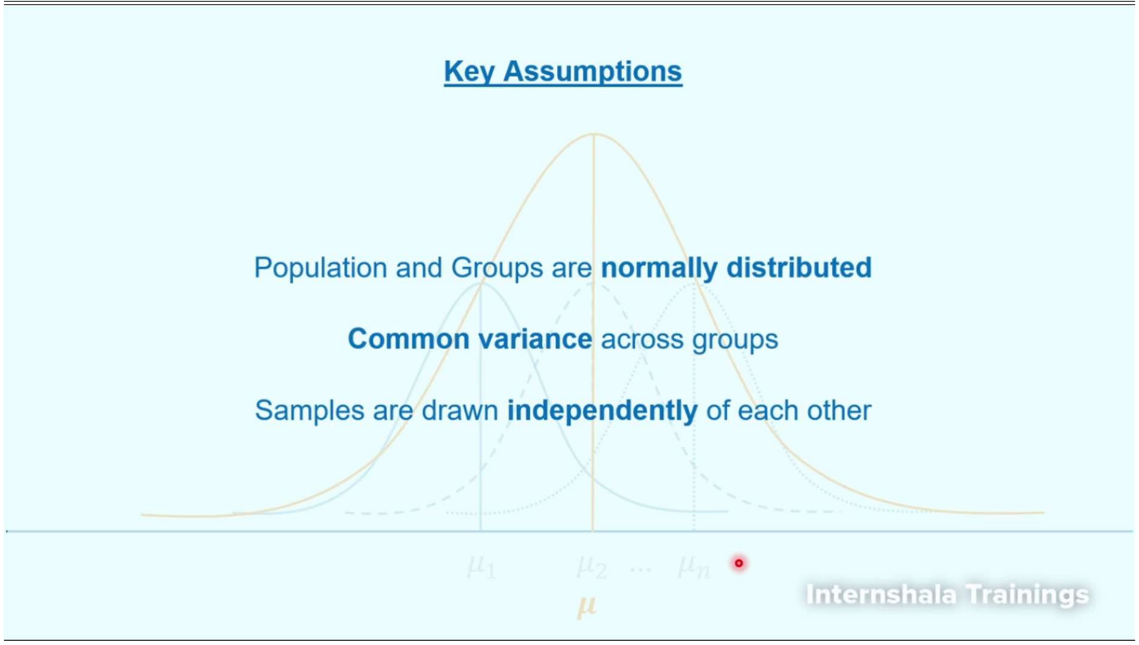


$H_o : \mu_1 = \mu_2 = \cdots = \mu_n$

$H_a : The\ means\ are\ not\ all\ equal$



Samples for each group come from the population sales



Key Assumptions

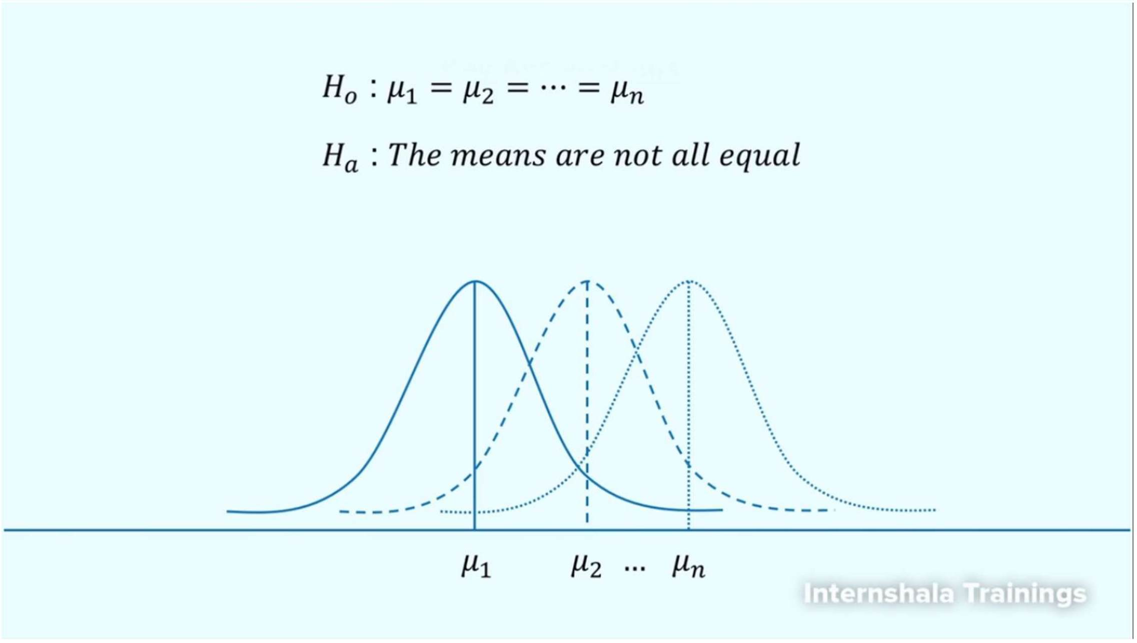
Population and Groups are **normally distributed**

Common variance across groups

Samples are drawn **independently** of each other

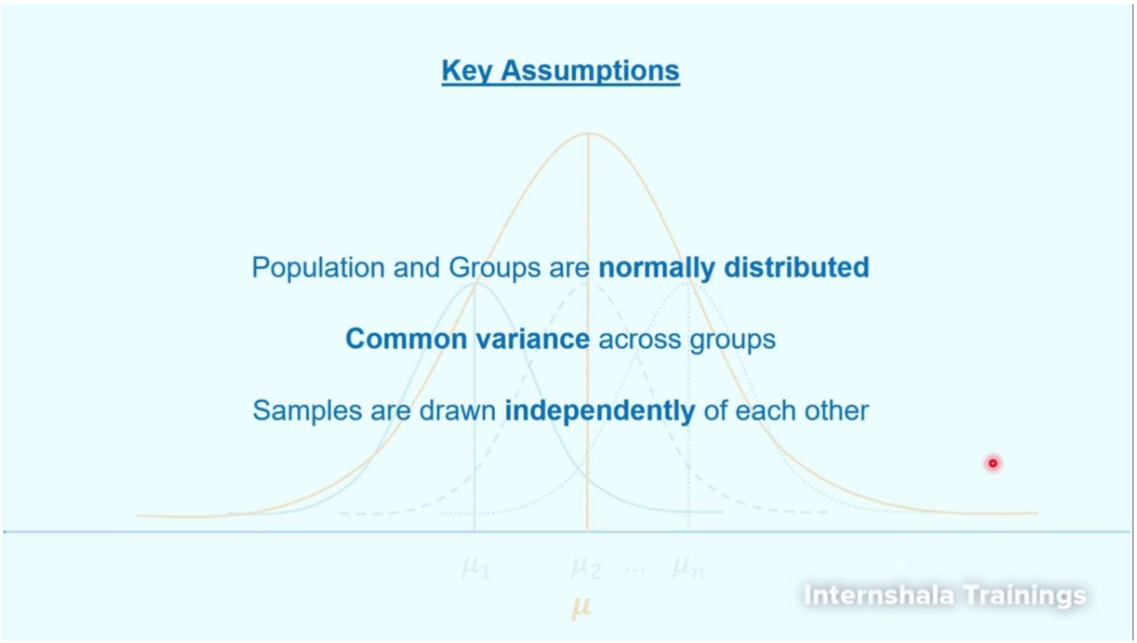
Module 5 Topic 1 Video 2

ANOVA Intuition



$H_o : \mu_1 = \mu_2 = \cdots = \mu_n$

$H_a : The\ means\ are\ not\ all\ equal$

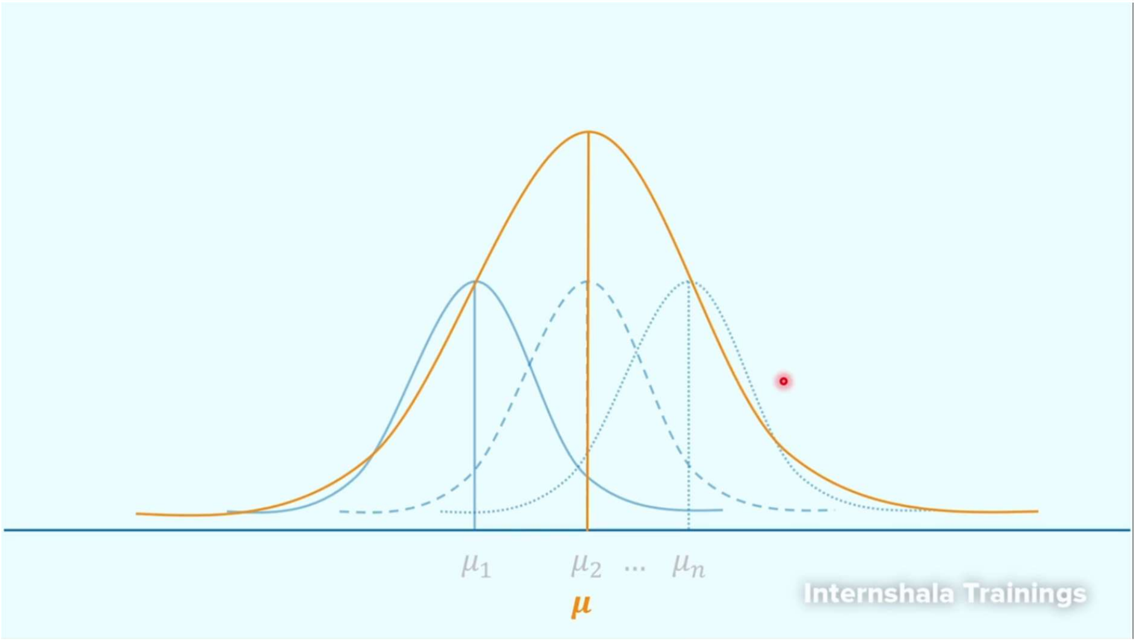


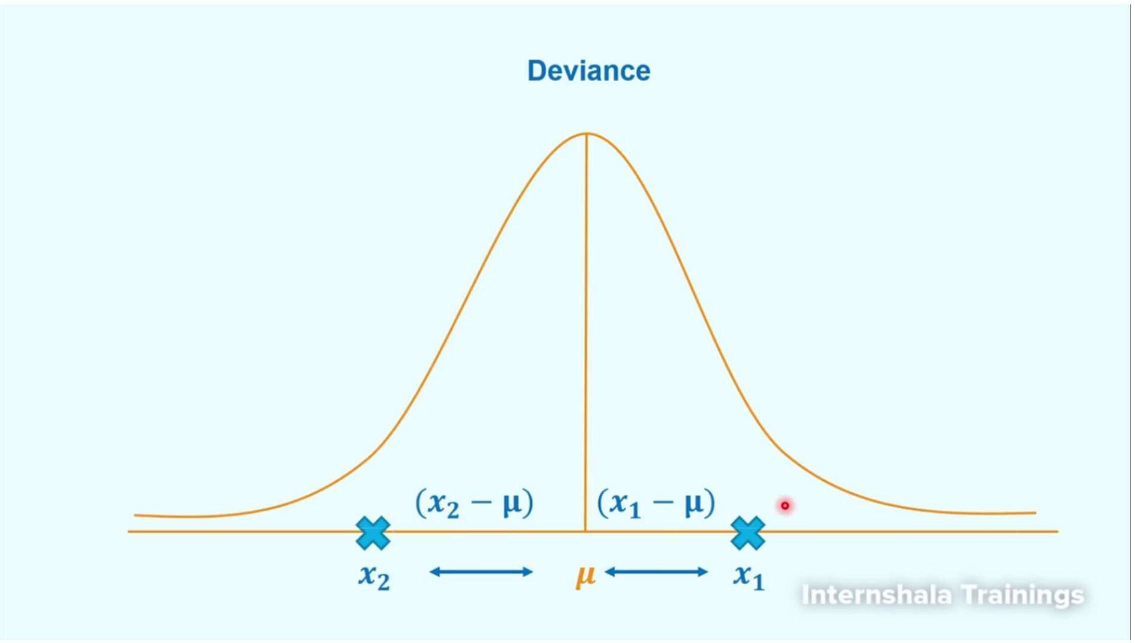
Key Assumptions

Population and Groups are **normally distributed**

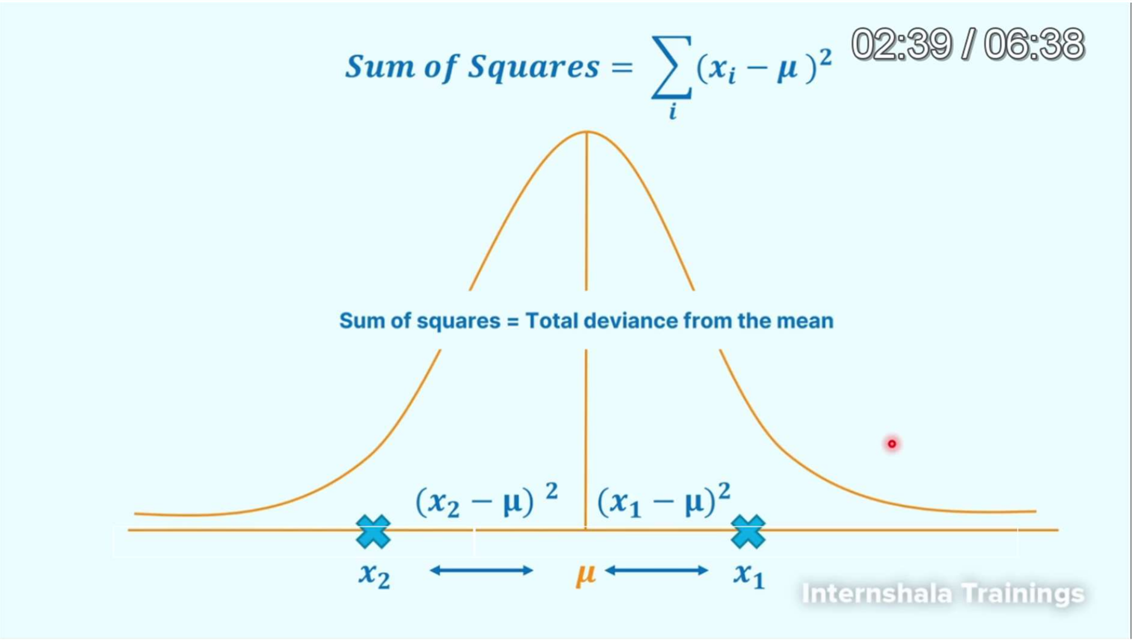
Common variance across groups

Samples are drawn **independently** of each other





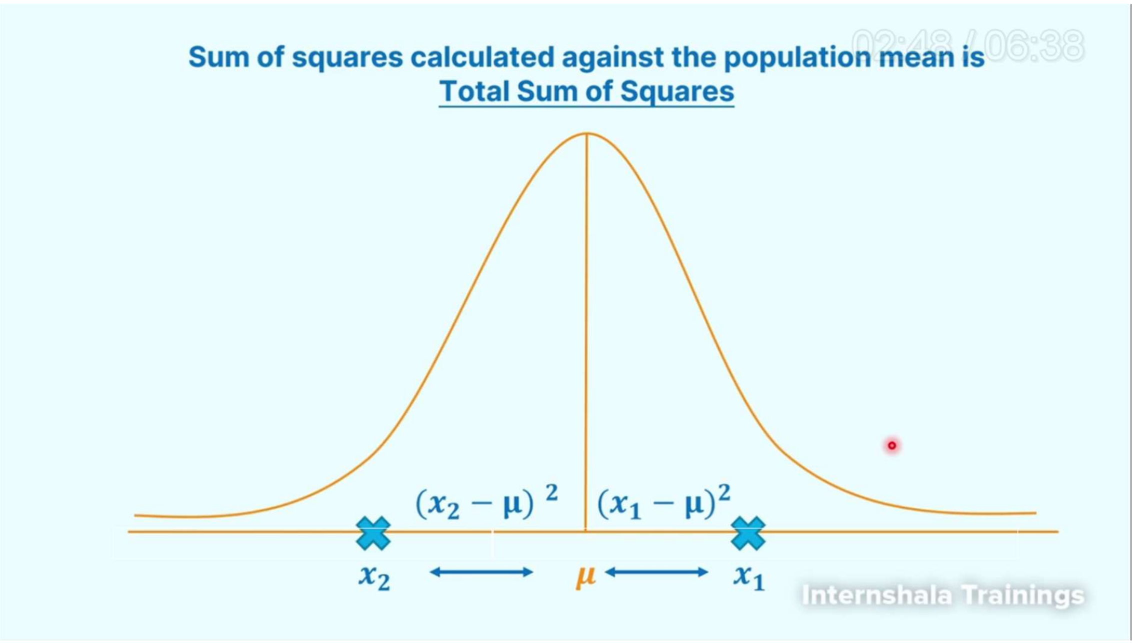
Deviance



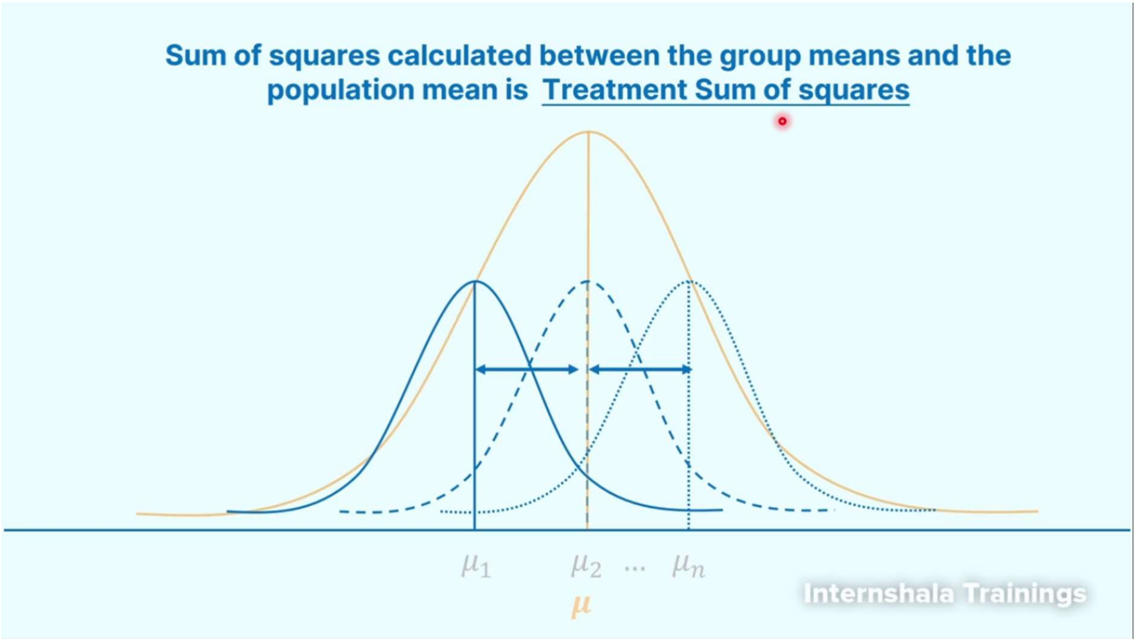
$$Sum\ of\ Squares = \sum_i (x_i - \mu)^2$$

02:39 / 06:38

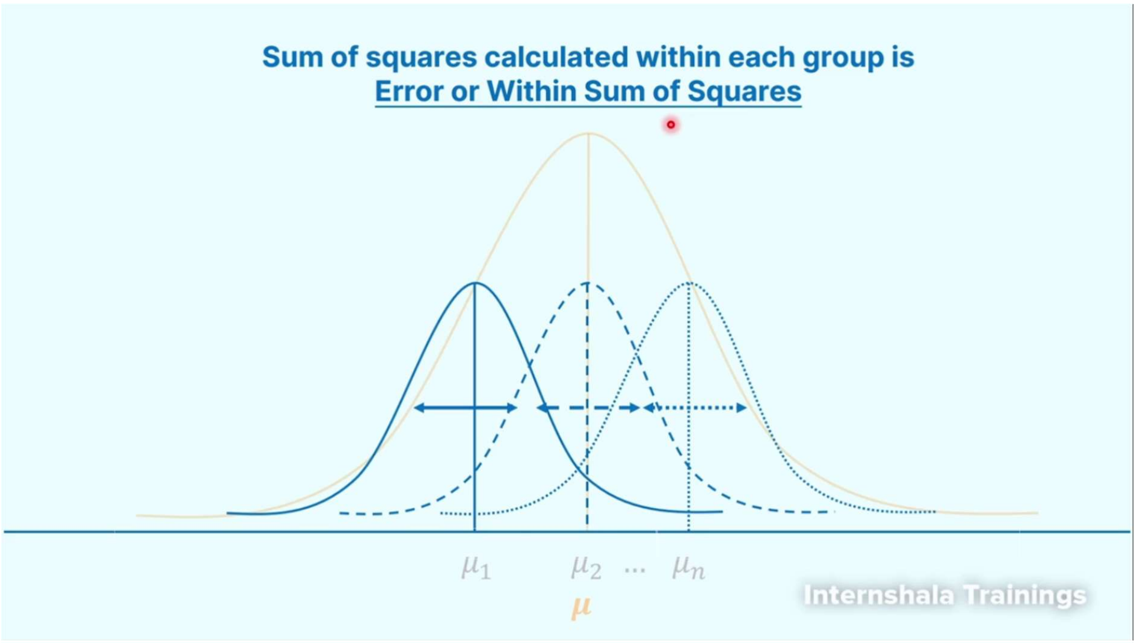
Sum of squares = Total deviance from the mean



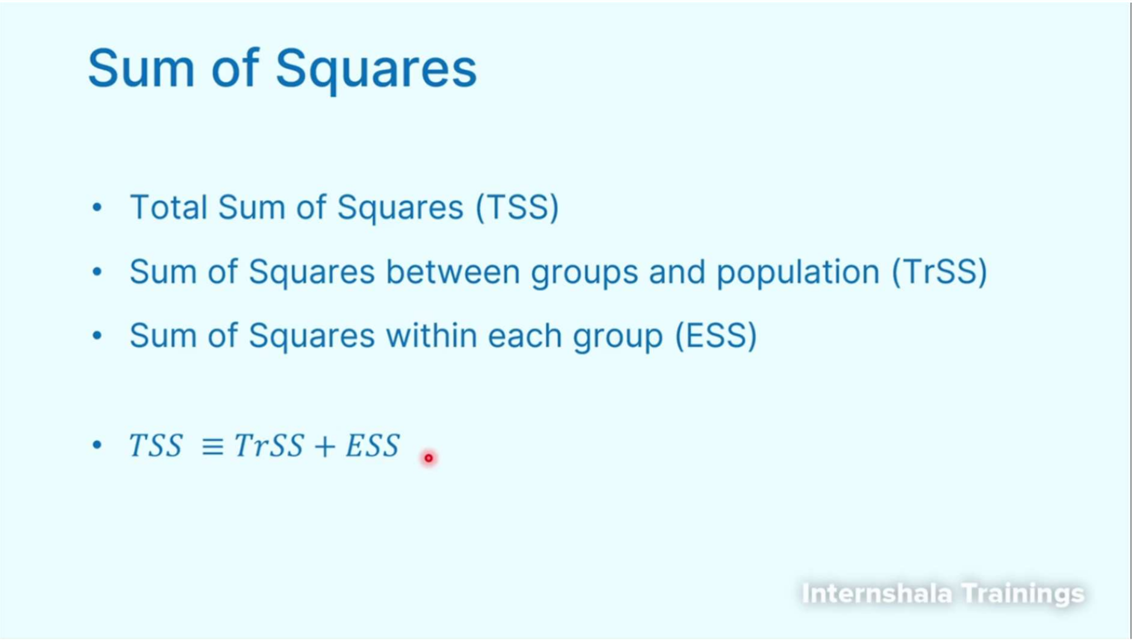
Sum of squares calculated against the population mean is **Total Sum of Squares**



Sum of squares calculated between the group means and the population mean is **Treatment Sum of squares**



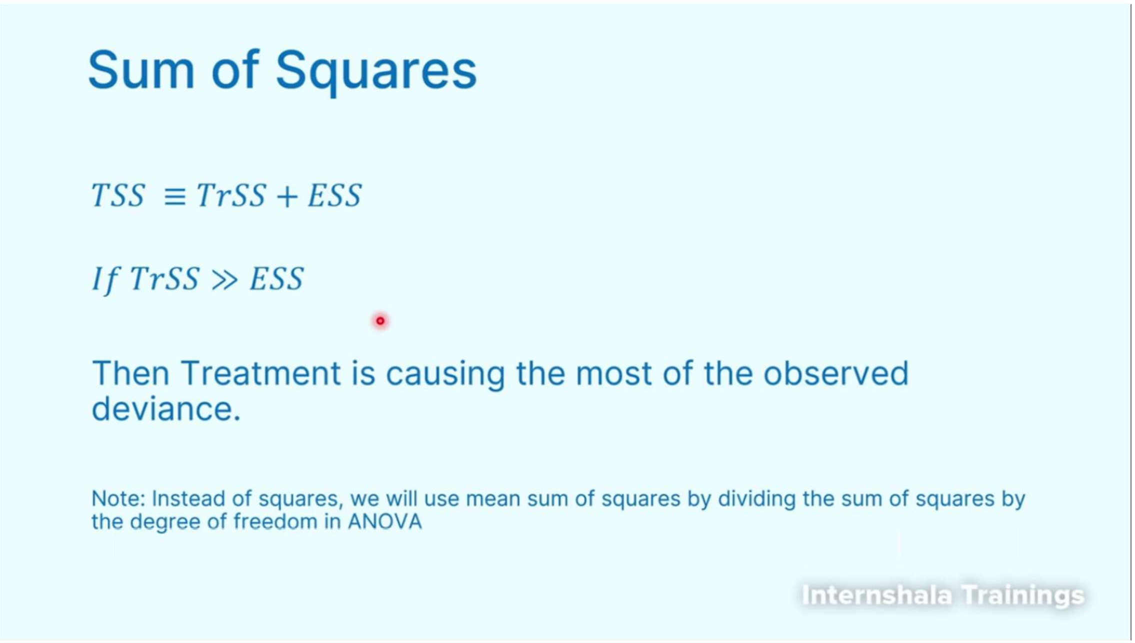
Sum of squares calculated within each group is
Error or Within Sum of Squares



Sum of Squares

- Total Sum of Squares (TSS)
- Sum of Squares between groups and population (TrSS)
- Sum of Squares within each group (ESS)

- $TSS \equiv TrSS + ESS$



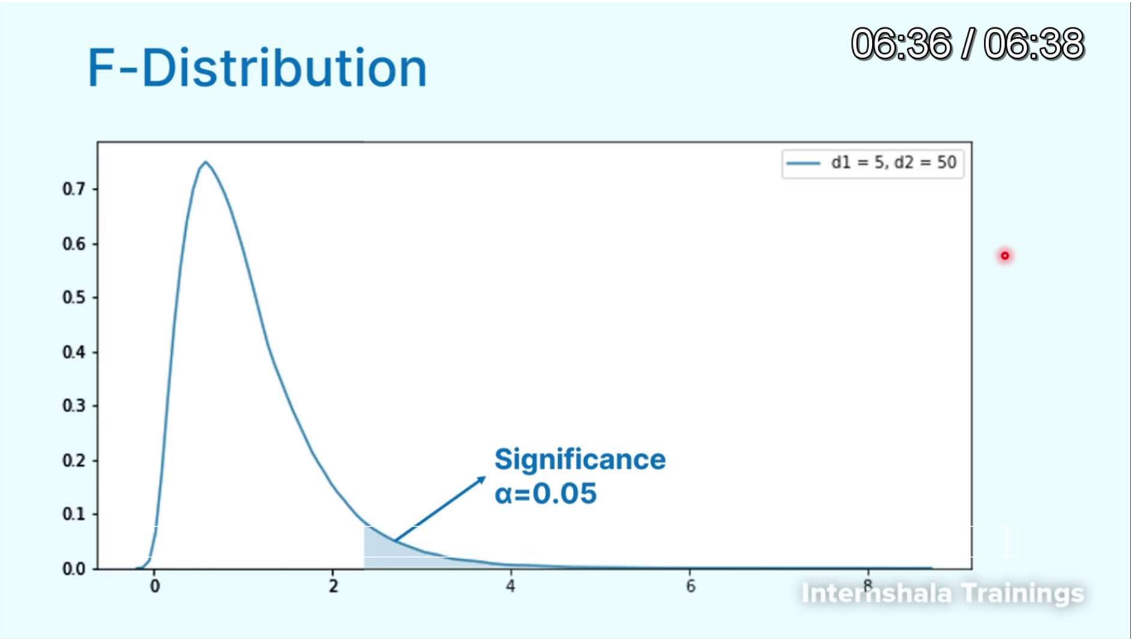
Sum of Squares

$TSS \equiv TrSS + ESS$

$If\ TrSS \gg ESS$

Then Treatment is causing the most of the observed deviance.

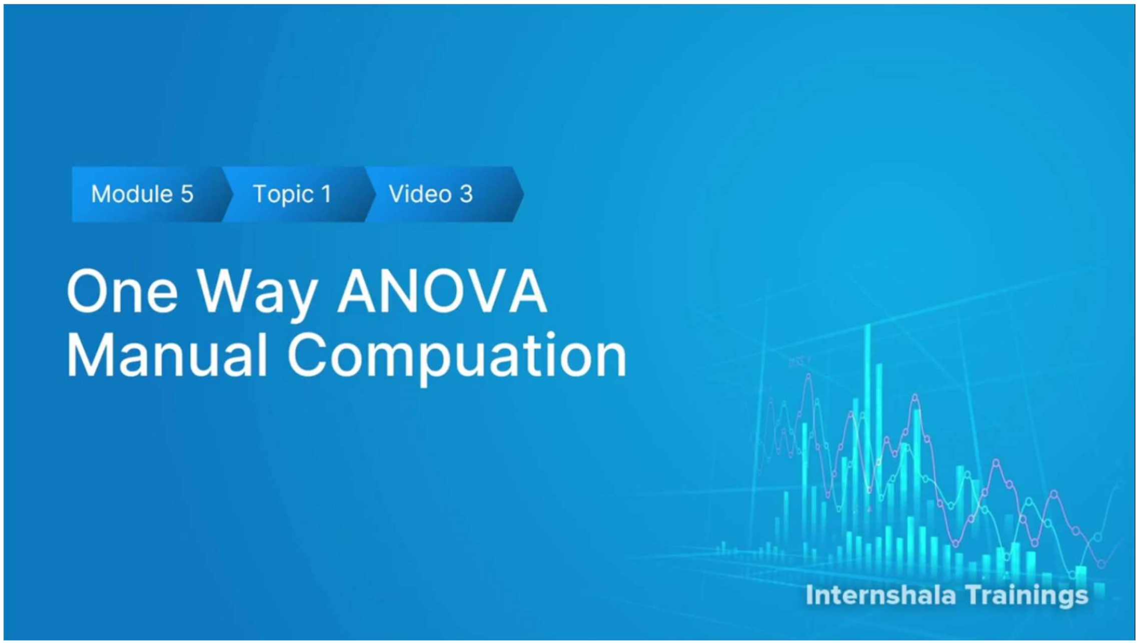
Note: Instead of squares, we will use mean sum of squares by dividing the sum of squares by the degree of freedom in ANOVA



F-Distribution

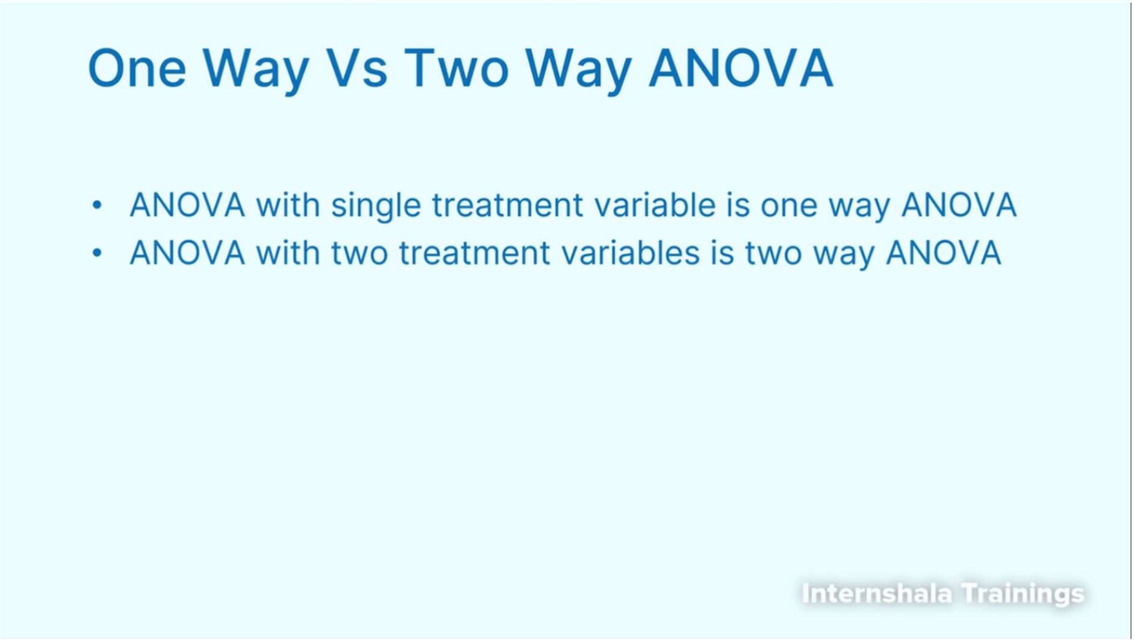
06:36 / 06:38

Significance
α=0.05



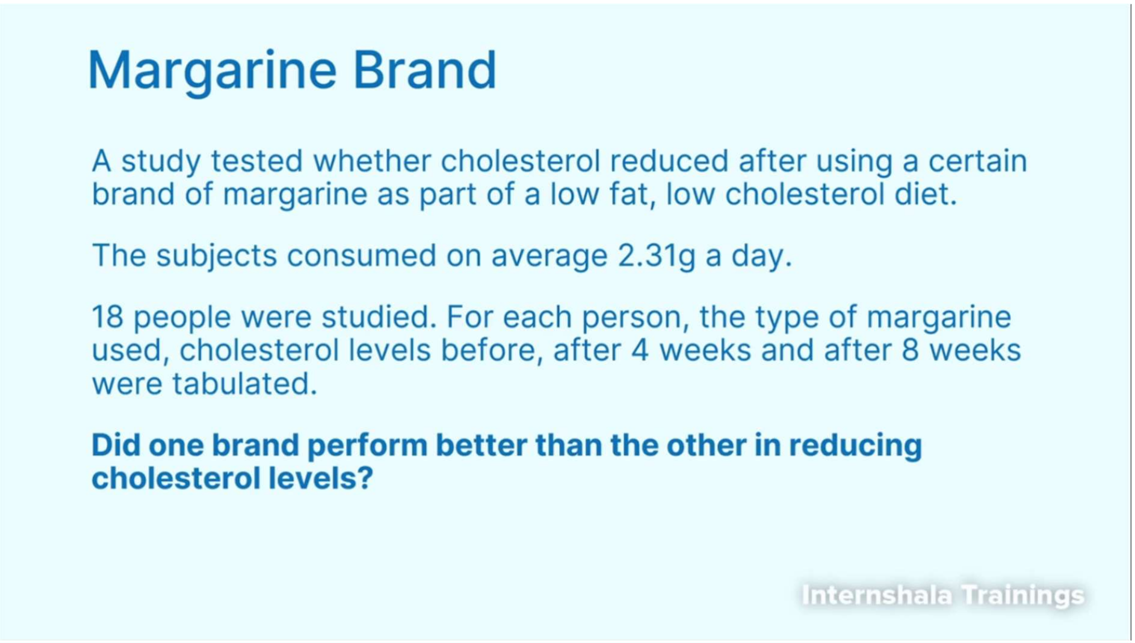
Module 5 Topic 1 Video 3

One Way ANOVA Manual Compuation



One Way Vs Two Way ANOVA

- ANOVA with single treatment variable is one way ANOVA
- ANOVA with two treatment variables is two way ANOVA



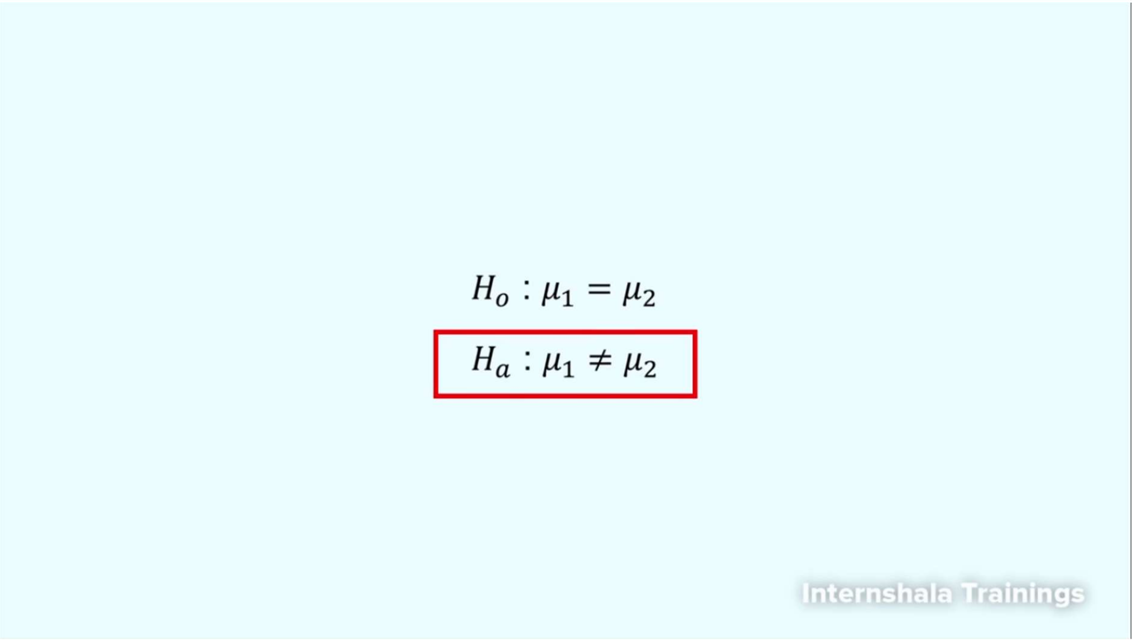
Margarine Brand

A study tested whether cholesterol reduced after using a certain brand of margarine as part of a low fat, low cholesterol diet.

The subjects consumed on average 2.31g a day.

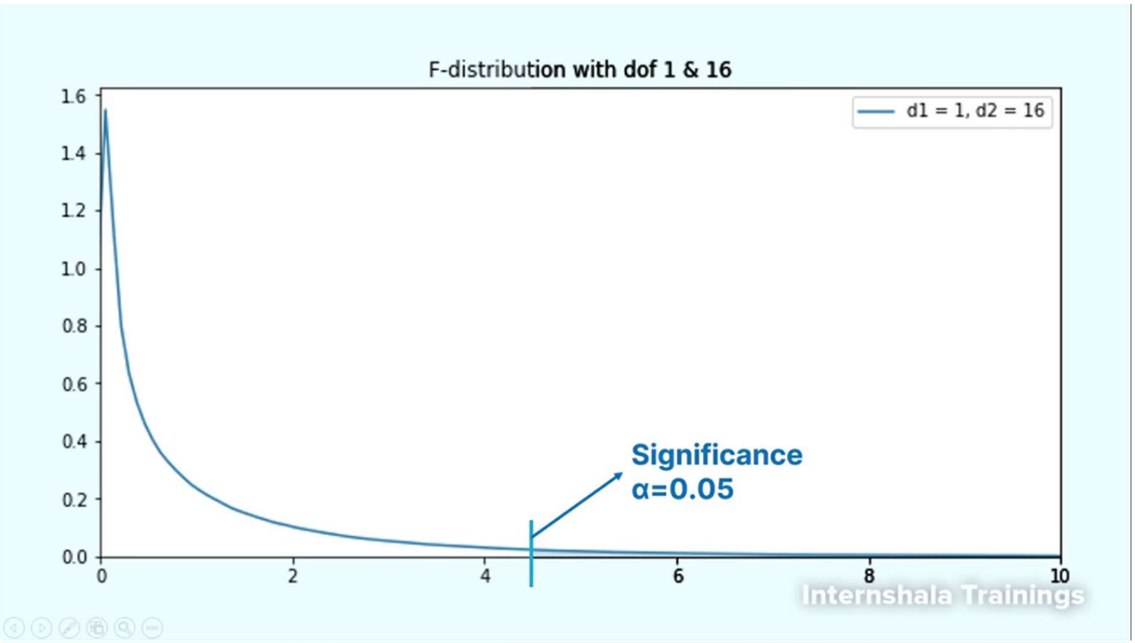
18 people were studied. For each person, the type of margarine used, cholesterol levels before, after 4 weeks and after 8 weeks were tabulated.

Did one brand perform better than the other in reducing cholesterol levels?



$$H_o : \mu_1 = \mu_2$$

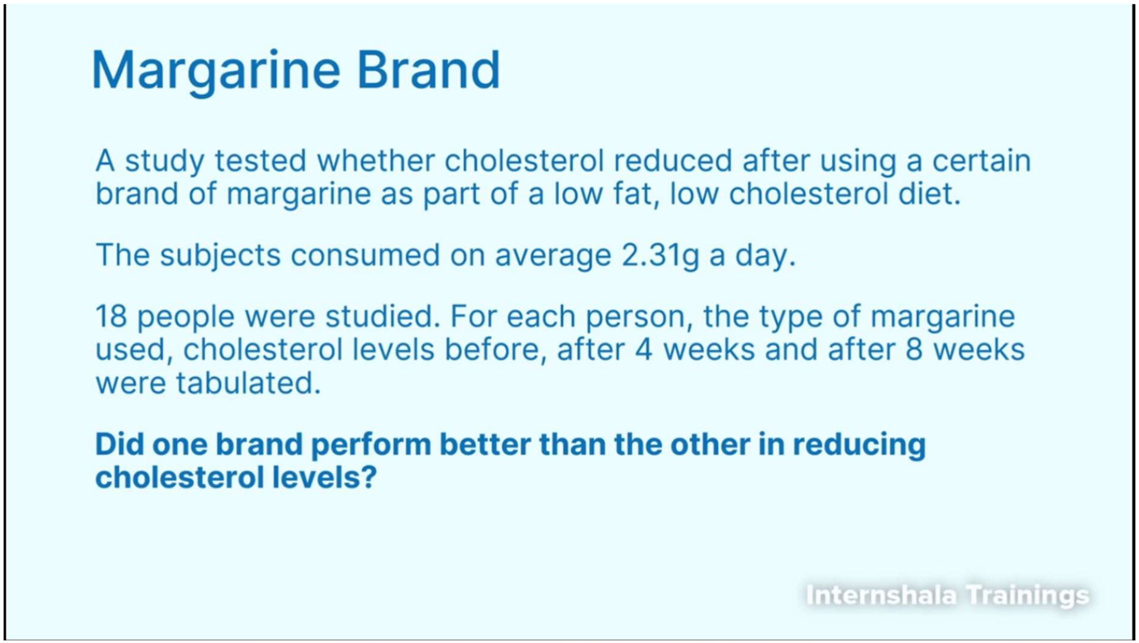
$$\boxed{H_a : \mu_1 \neq \mu_2}$$



Significance
α=0.05

Module 5 Topic 1 Video 4

One Way ANOVA using Python



Margarine Brand

A study tested whether cholesterol reduced after using a certain brand of margarine as part of a low fat, low cholesterol diet.

The subjects consumed on average 2.31g a day.

18 people were studied. For each person, the type of margarine used, cholesterol levels before, after 4 weeks and after 8 weeks were tabulated.

Did one brand perform better than the other in reducing cholesterol levels?

Module 5 Topic 2 Video 1

Two Way ANOVA
Diet Plan

Diet Plan

Weight loss due to diet plans were studied in a test in which 76 people participated.

Weight in kg before the study and 10 weeks after the start of the study were recorded.

3 diet plans were studied.

Participants gender are also given.

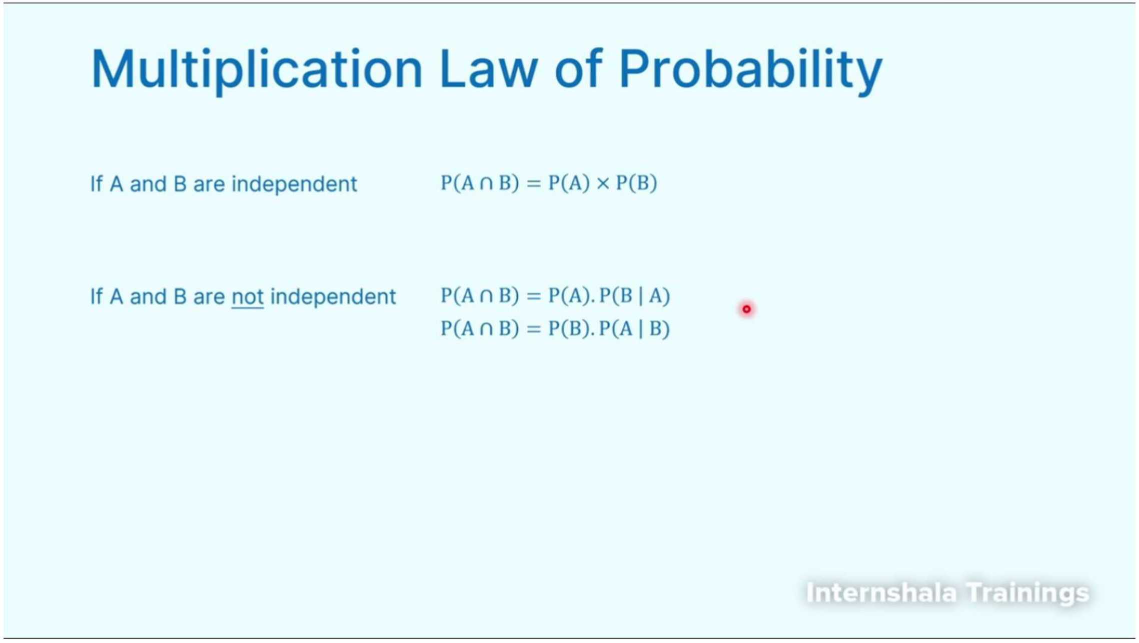
Does the mean weight loss differ among the groups by diet plan and gender.

$$H_o : \mu_1 = \mu_2 = \ \dots \mu_n$$

$$H_a : The\ group\ means\ are\ different$$

Module 5 Topic 3 Video 1

Introduction to Test of Independence / Chi-Square Test

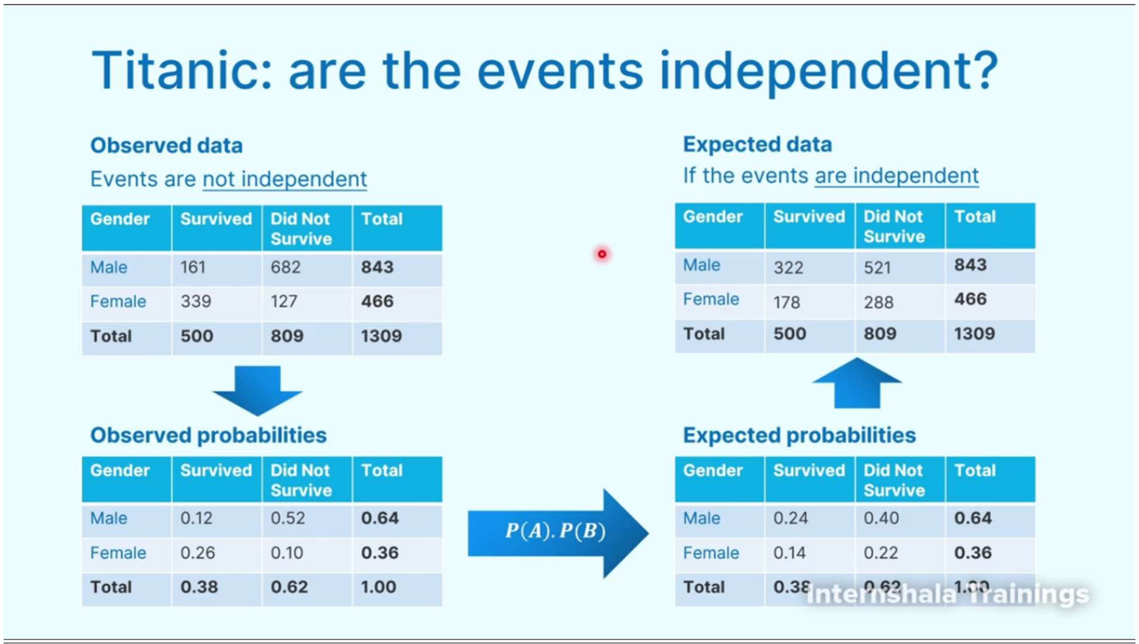


Multiplication Law of Probability

If A and B are independent $P(A \cap B) = P(A) \times P(B)$

If A and B are not independent $P(A \cap B) = P(A). P(B \mid A)$

$P(A \cap B) = P(B). P(A \mid B)$



Titanic: are the events independent?

Observed data

Events are not independent

Gender	Survived	Did Not Survive	Total
Male	161	682	**843**
Female	339	127	**466**
Total	**500**	**809**	**1309**

Observed probabilities

Gender	Survived	Did Not Survive	Total
Male	0.12	0.52	**0.64**
Female	0.26	0.10	**0.36**
Total	**0.38**	**0.62**	**1.00**

$P(A).P(B)$

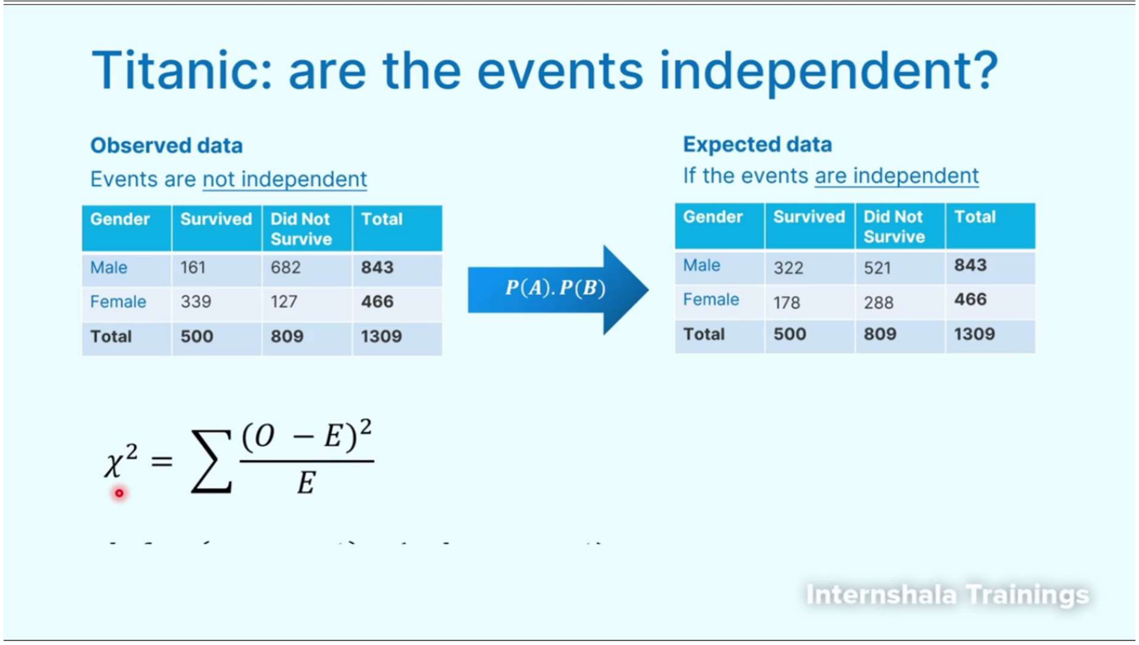
Expected probabilities

Gender	Survived	Did Not Survive	Total
Male	0.24	0.40	**0.64**
Female	0.14	0.22	**0.36**
Total	**0.38**	**0.62**	**1.00**

Expected data

If the events are independent

Gender	Survived	Did Not Survive	Total
Male	322	521	**843**
Female	178	288	**466**
Total	**500**	**809**	**1309**



Titanic: are the events independent?

Observed data

Events are not independent

Gender	Survived	Did Not Survive	Total
Male	161	682	**843**
Female	339	127	**466**
Total	**500**	**809**	**1309**

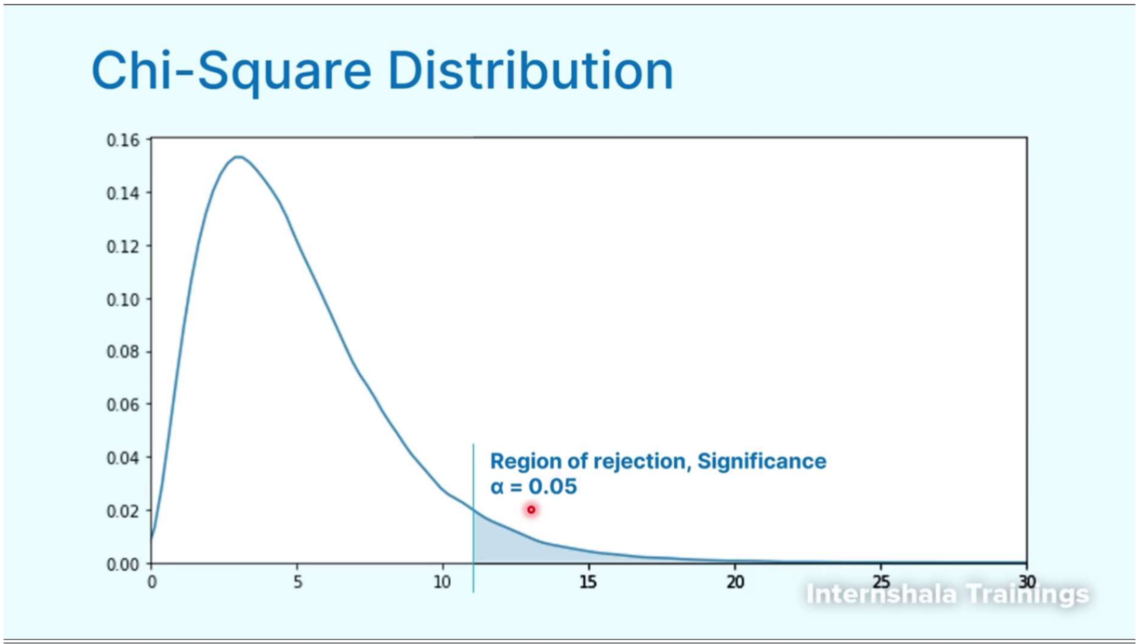
$P(A).P(B)$

Expected data

If the events are independent

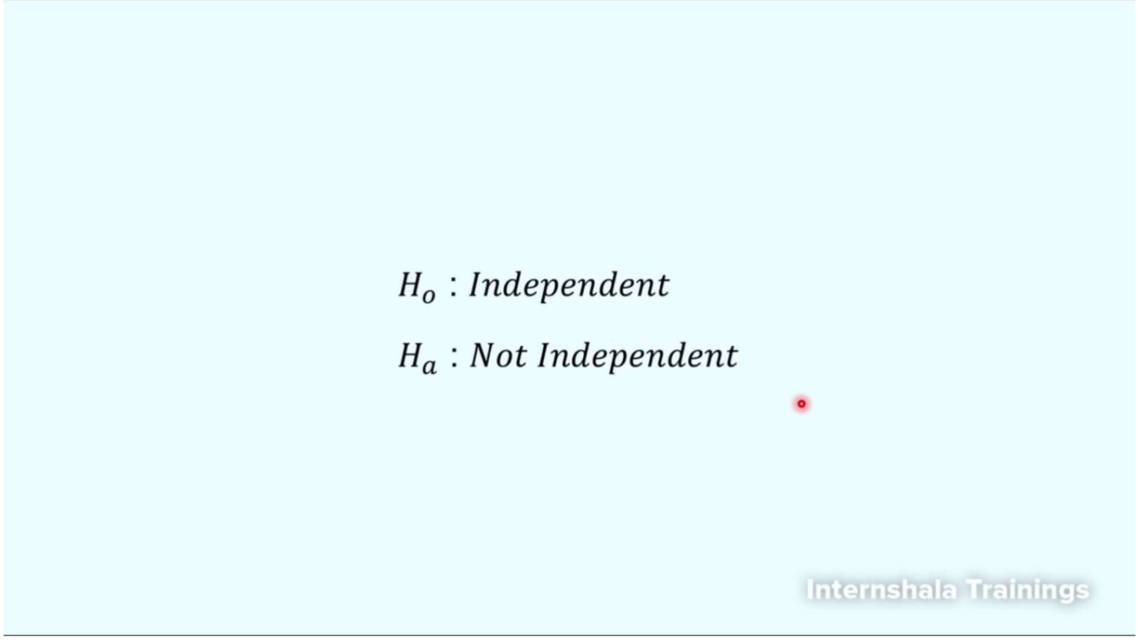
Gender	Survived	Did Not Survive	Total
Male	322	521	**843**
Female	178	288	**466**
Total	**500**	**809**	**1309**

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$



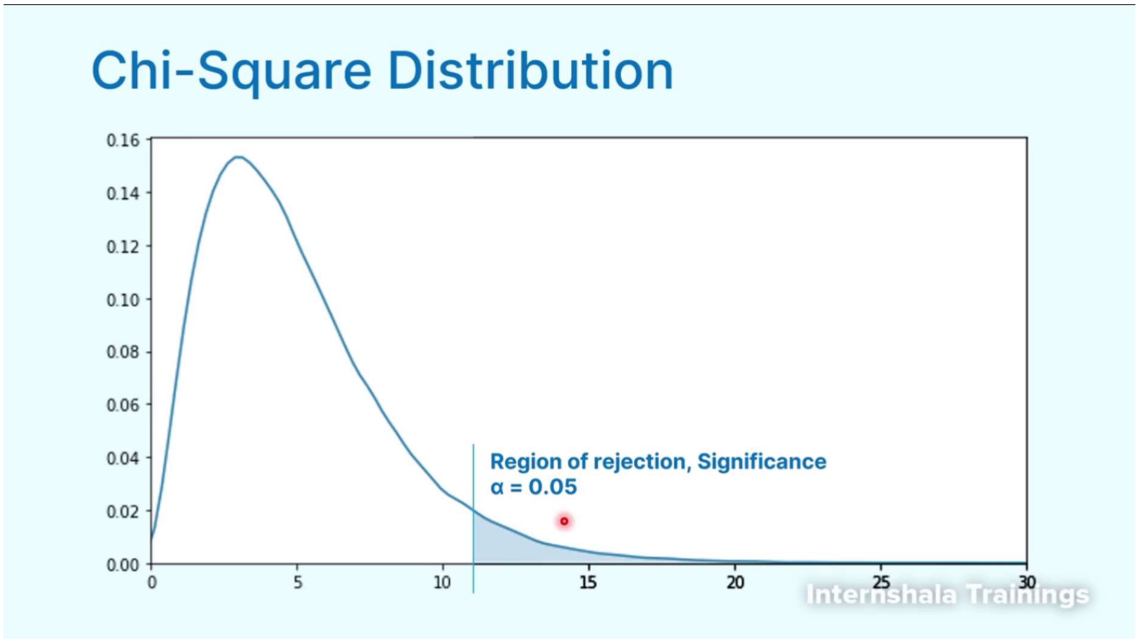
Chi-Square Distribution

Region of rejection, Significance
α = 0.05

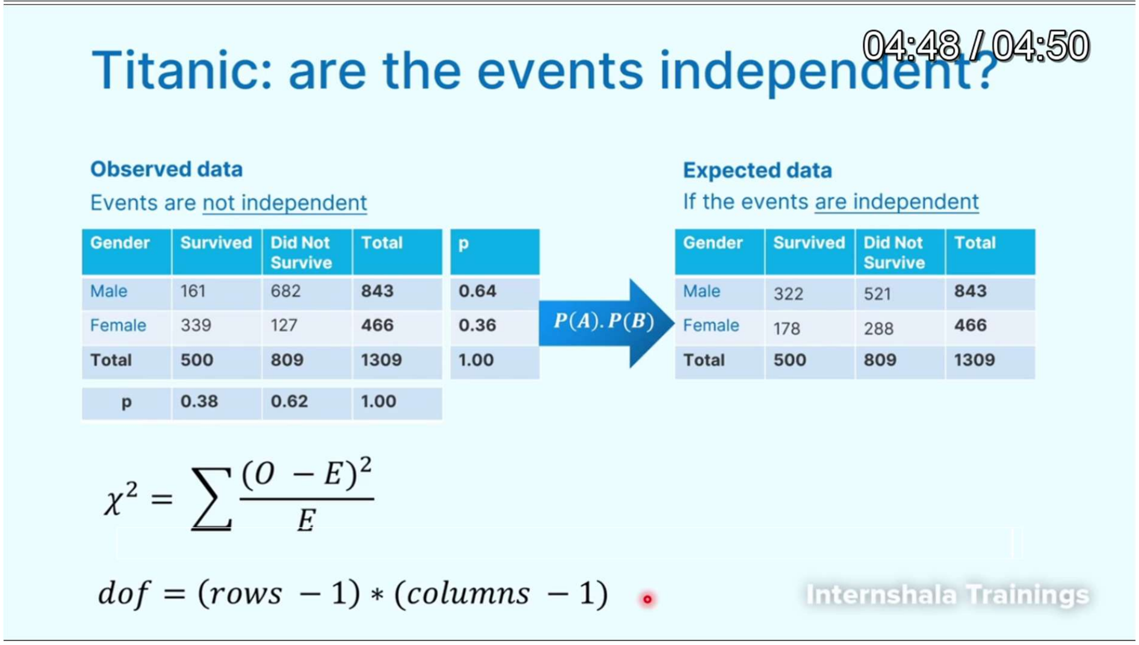


H_o : *Independent*

H_a : *Not Independent*



Chi-Square Distribution



04:48 / 04:50

Titanic: are the events independent?

Observed data

Events are not independent

Gender	Survived	Did Not Survive	Total
Male	161	682	**843**
Female	339	127	**466**
Total	**500**	**809**	**1309**
p	**0.38**	**0.62**	**1.00**

p
0.64
0.36
1.00

$P(A).P(B)$

Expected data

If the events are independent

Gender	Survived	Did Not Survive	Total
Male	322	521	**843**
Female	178	288	**466**
Total	**500**	**809**	**1309**

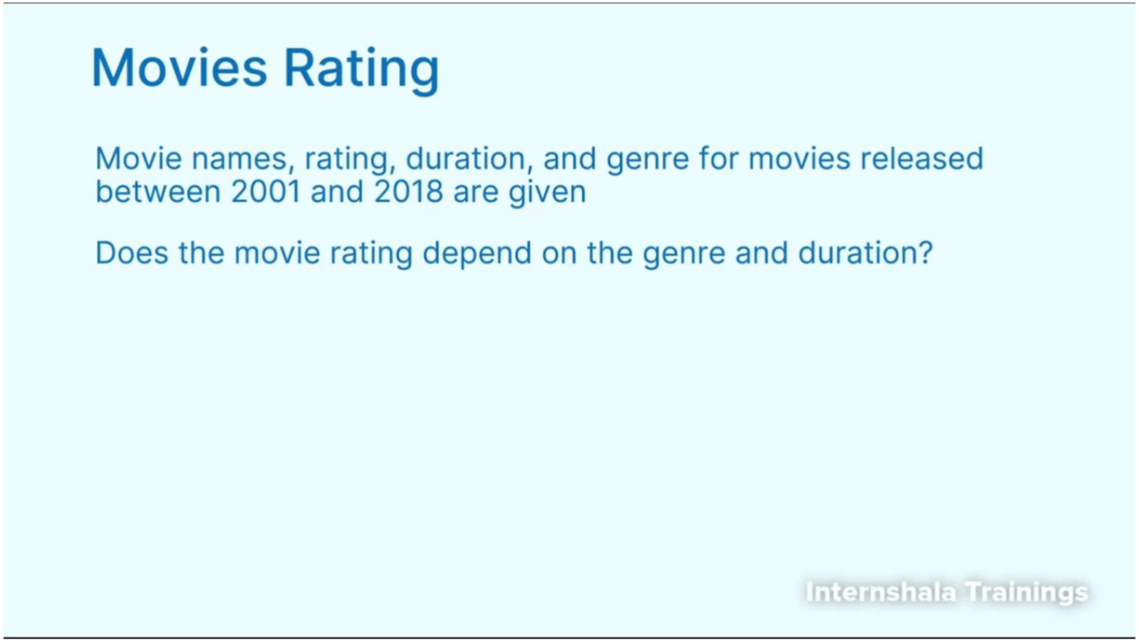
$$\chi^2 = \sum \frac{(O\ - E)^2}{E}$$

$$dof = (rows\ - 1) * (columns\ - 1)$$

Module 5 Topic 2 Video 2

Two Way ANOVA Movies Rating

Internshala Trainings

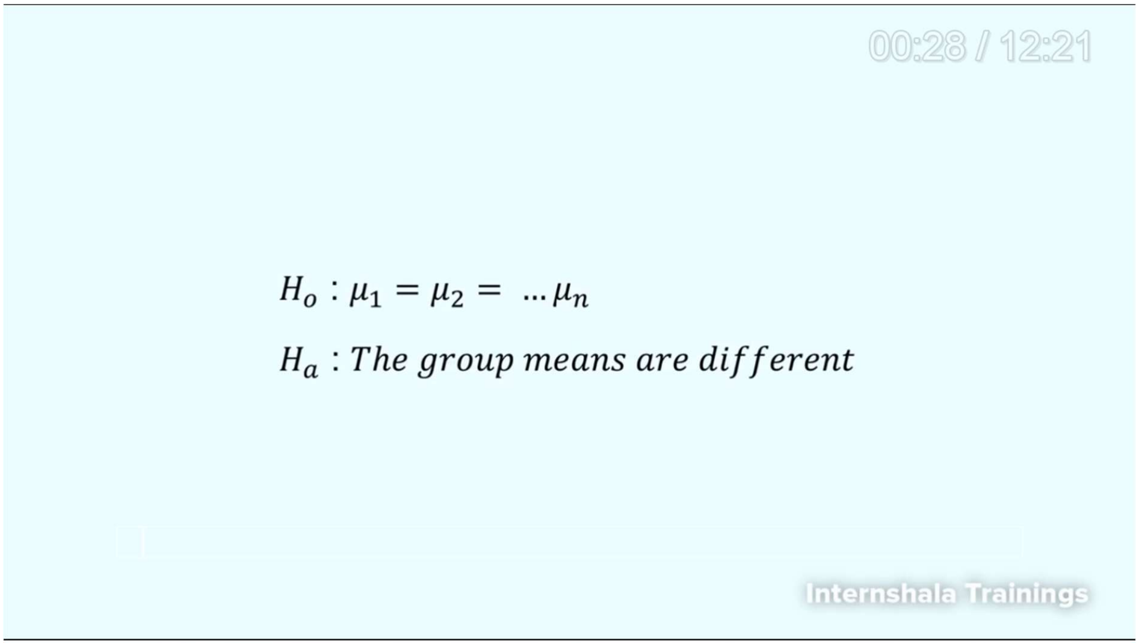


Movies Rating

Movie names, rating, duration, and genre for movies released between 2001 and 2018 are given

Does the movie rating depend on the genre and duration?

Internshala Trainings



00:28 / 12:21

$H_o : \mu_1 = \mu_2 = \ \dots \mu_n$

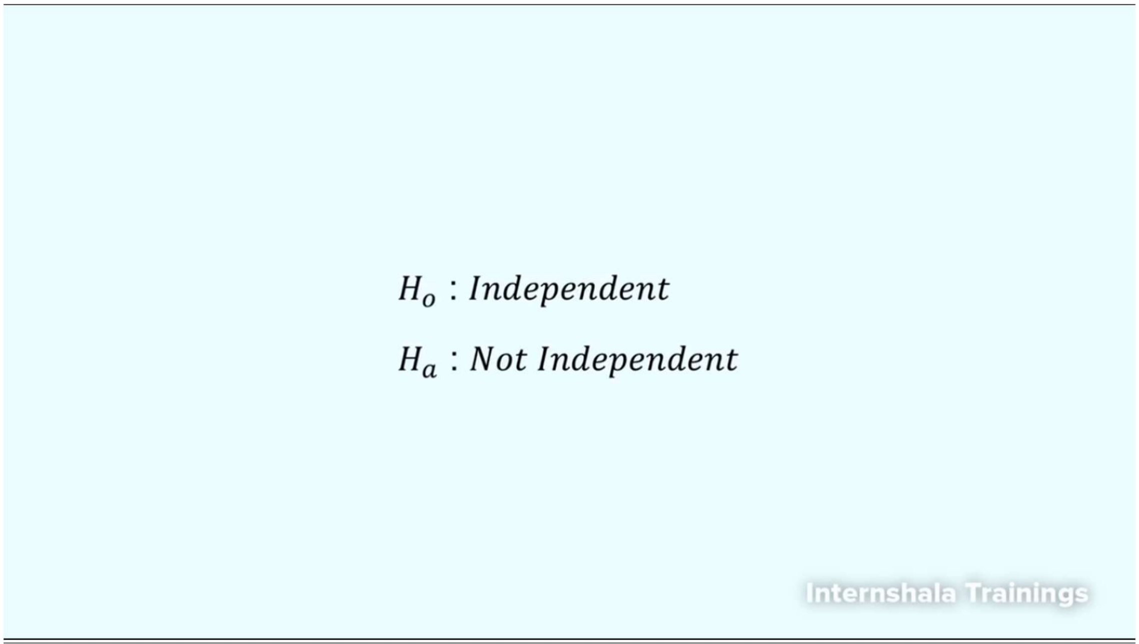
$H_a : The\ group\ means\ are\ different$

Internshala Trainings

Module 5 Topic 1 Video 2

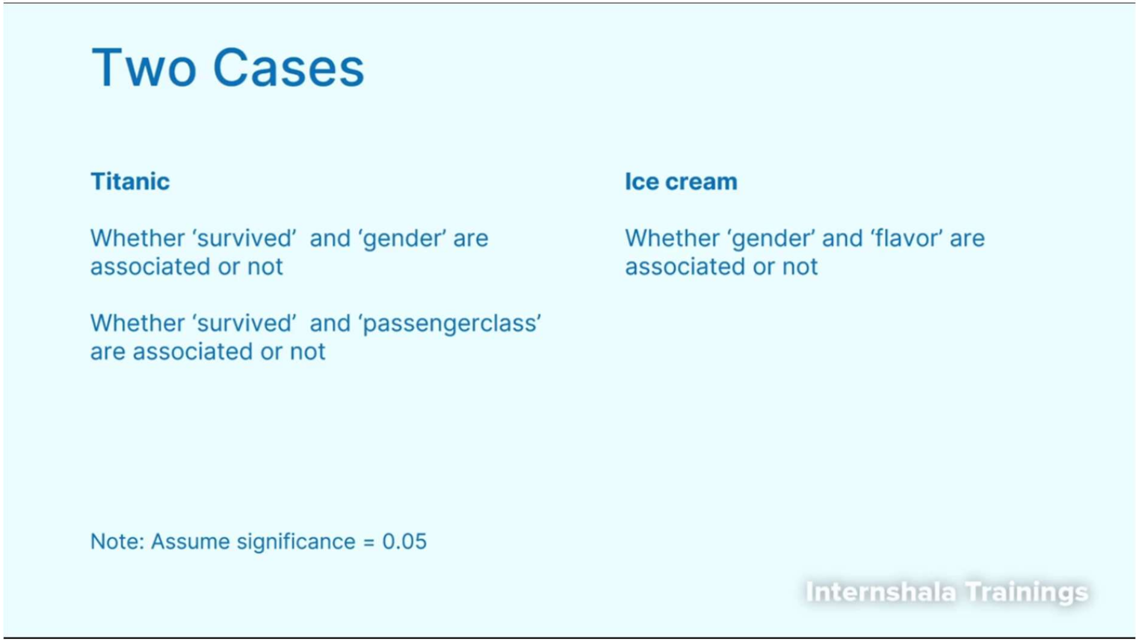
Test of Independence Exercises

Internshala Trainings



$$H_o : Independent$$

$$H_a : Not\ Independent$$



Two Cases

Titanic

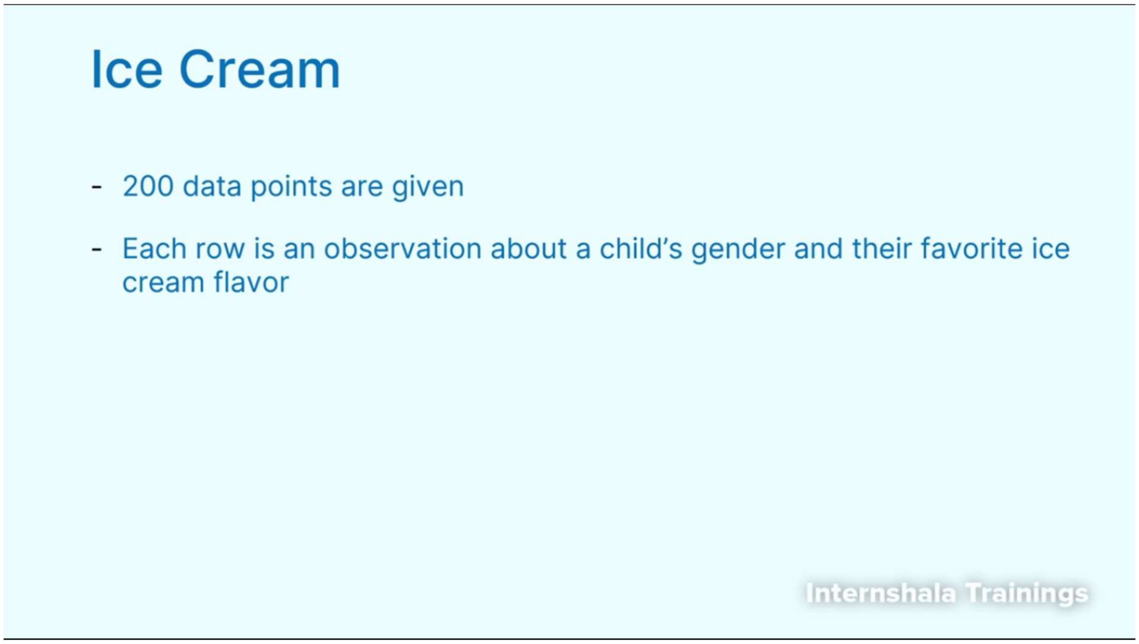
Whether 'survived' and 'gender' are associated or not

Whether 'survived' and 'passengerclass' are associated or not

Ice cream

Whether 'gender' and 'flavor' are associated or not

Note: Assume significance = 0.05



Ice Cream

- 200 data points are given
- Each row is an observation about a child's gender and their favorite ice cream flavor