Data Analytics

Mrinal Das

Example

Line A	Line B	Line C
30	35	28
32	37	29
31	34	30
29	36	27
30	38	28

In a manufacturing company

There are 3 lines

Their production times are different

Their averages are also different

Are the differences between the average production times real, or could they simply be due to random variation?

Analysis of Variance (ANOVA)

ANOVA is used to compare means through variation

$$\boxed{\text{Total variation} = \text{Between-group variation} + \text{Within-group variation}}$$

```
                Total Variation
                       |
         ┌─────────────┴─────────────┐
         ↓                           ↓
 Between-group                Within-group
   variation                    variation
         |                           |
  Differences                   Natural
  between lines                variation
```

Hypothesis Testing through ANOVA

We are concerned about following Hypotheses :

Null hypothesis: all groups are similar to each other

Alternative hypothesis: all groups are not similar to each other

A group is collection of values which are related to each other

E.g. time taken by a machine across several runs

Basic Idea

Basic idea is to compare the averages across groups

But only average may not be a proper metric to check always

E.g. averages may be different, but question is that a slight difference can be due to some random condition

How can we get confident about the differences ?

Simple Rule

Simple rule is to consider variance of spread of values also

ANOVA does exactly that

$$\boxed{F = \frac{MS_{\text{between}}}{MS_{\text{within}}}} \qquad F = \frac{\text{between-group variation}}{\text{within-group variation}}$$

If F is near 1 then all groups are similar, null hypothesis is valid.

Simple Rule

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

$$MS_{\text{between}} = \frac{SS_{\text{between}}}{k-1}$$

$$MS_{\text{within}} = \frac{SS_{\text{within}}}{N-k}$$

- N = total number of observations
- k = number of groups

$$\boxed{SS_{\text{between}} = \sum_{g=1}^{k} n_g(\bar{X}_g - \bar{X})^2}$$

- k = number of groups
- n_g = number of observations in group g
- $\bar{X}_g$ = group mean
- $\bar{X}$ = grand mean

If F is near 1 then all groups are similar, null hypothesis is valid.

Group Means

Line A	Line B	Line C
30	35	28
32	37	29
31	34	30
29	36	27
30	38	28

$$\bar{X}_A = \frac{30 + 32 + 31 + 29 + 30}{5}$$

$$\boxed{\bar{X}_A = 30.4}$$

$$\bar{X}_B = \frac{35 + 37 + 34 + 36 + 38}{5}$$

$$\boxed{\bar{X}_B = 36.0}$$

$$\bar{X}_C = \frac{28 + 29 + 30 + 27 + 28}{5}$$

$$\boxed{\bar{X}_C = 28.4}$$

Grand Mean

Line A	Line B	Line C
30	35	28
32	37	29
31	34	30
29	36	27
30	38	28

$$\bar{X} = \frac{474}{15}$$

$$\boxed{\bar{X} = 31.6}$$

$$30 + 32 + 31 + 29 + 30 + 35 + 37 + 34 + 36 + 38 + 28 + 29 + 30 + 27 + 28 = 474$$

Within Group Variation

Line A	Line B	Line C
30	35	28
32	37	29
31	34	30
29	36	27
30	38	28

Group A $\boxed{\bar{X}_A = 30.4}$

Observation	Deviation	Squared deviation
30	-0.4	0.16
32	1.6	2.56
31	0.6	0.36
29	-1.4	1.96
30	-0.4	0.16

$$SS_A = 0.16 + 2.56 + 0.36 + 1.96 + 0.16$$

$$\boxed{SS_A = 5.20}$$

Within Group Variation

Line A	Line B	Line C
30	35	28
32	37	29
31	34	30
29	36	27
30	38	28

Group B $\boxed{\bar{X}_B = 36.0}$

Observation	Deviation	Squared deviation
35	-1	1
37	1	1
34	-2	4
36	0	0
38	2	4

$$SS_B^{within} = 1 + 1 + 4 + 0 + 4$$

$$\boxed{SS_B^{within} = 10}$$

Within Group Variation

Group C

$$\boxed{\bar{X}_C = 28.4}$$

Line A	Line B	Line C
30	35	28
32	37	29
31	34	30
29	36	27
30	38	28

Observation	Deviation	Squared deviation
28	-0.4	0.16
29	0.6	0.36
30	1.6	2.56
27	-1.4	1.96
28	-0.4	0.16

$$SS_C = 0.16 + 0.36 + 2.56 + 1.96 + 0.16$$

$$\boxed{SS_C = 5.20}$$

Total Within Group Variation

$$SS_W = SS_A + SS_B + SS_C$$

$$= 5.2 + 10 + 5.2$$

$$\boxed{SS_W = 20.4}$$

$$df_W = N - k$$

$$df_W = 15 - 3 = \boxed{12}$$

Within Group Mean Square

$$MS_W = \frac{SS_W}{df_W}$$

$$= \frac{20.4}{12}$$

$$\boxed{MS_W = 1.7}$$

Between Group Variation

Line	Mean
A	30.4
B	36
C	28.4

$$SS_B = \sum_g n_g(\bar{X}_g - \bar{X})^2$$

$$n_g = 5.$$

$$\boxed{\bar{X} = 31.6}$$

Group A

$$5(30.4 - 31.6)^2$$
$$= 5(-1.2)^2$$
$$= 5(1.44) = 7.2$$

Group B

$$5(36 - 31.6)^2$$
$$= 5(4.4)^2$$
$$= 5(19.36) = 96.8$$

Group B

$$5(28.4 - 31.6)^2$$
$$= 5(-3.2)^2$$
$$= 5(10.24) = 51.2$$

$$SS_B = 7.2 + 96.8 + 51.2$$

$$\boxed{SS_B = 155.2}$$

Within and Between Mean Squares

Between Group Mean Square

$$MS_B = \frac{SS_B}{df_B}$$

$$= \frac{155.2}{2}$$

$$\boxed{MS_B = 77.6}$$

$$df_B = k - 1$$

$$df_B = 3 - 1 = \boxed{2}$$

Within Group Mean Square

$$MS_W = \frac{SS_W}{df_W}$$

$$= \frac{20.4}{12}$$

$$\boxed{MS_W = 1.7}$$

$$df_W = N - k$$

$$df_W = 15 - 3 = \boxed{12}$$

F Statistic

$$F = \frac{MS_B}{MS_W}$$

$$F = \frac{77.6}{1.7}$$

$$\boxed{F \approx 45.65}$$

$$F = \frac{\text{between-group variation}}{\text{within-group variation}}$$

$$\boxed{\text{Between variation} \gg \text{Within variation}}$$

F is quite high, so groups are not similar.

Conclusion

F is quite high, so groups are not similar.

Line B > Line A > Line C

Line	Mean
A	30.4
B	36
C	28.4

$$\boxed{\text{Between variation} \gg \text{Within variation}}$$

Conclusion

F is quite high, so groups are not similar.

$$\text{Line B} > \text{Line A} > \text{Line C}$$

Line	Mean
A	30.4
B	36
C	28.4

So Line B is considerably slower.

But **ANOVA alone does not establish why** Line B is slower.

Management should investigate possible causes:

- machine configuration,
- operator procedures,
- material handling,
- machine age,
- maintenance,

Octane vs Horse Power

87 Octane	89 Octane	91 Octane	93 Octane
168.8	178.0	179.9	185.3
171.5	178.0	183.4	189.7
172.6	173.6	180.1	187.1
168.1	173.7	181.8	187.7
167.1	172.8	179.4	186.2
170.9	176.4	183.9	188.3
168.3	172.9	184.2	186.6
169.2	178.5	184.0	191.2
168.4	176.8	183.3	187.4
171.8	172.3	182.5	185.4
169.7	**175.3**	**182.2**	**187.5**

The F-Value

$$F - value = \frac{MST}{MSE} = \frac{Mean\ Square\ of\ the\ Treatment}{Mean\ Square\ of\ the\ Error} = \frac{610.2}{4.1} = 148.8$$

Variation Source	Sum of Squares	Degrees of freedom	Mean Squares	F-Value
Treatment (Between)	1831	3	610.2	$\mathbf{F} = \frac{MS_T}{MS_E}$
Error (Within)	148	36	4.1	
Total	1978	39		

ANOVA vs Significance Test

A **statistical significance test** is a general framework for deciding whether the observed data provide enough evidence against a null hypothesis.

ANOVA is also a significance test

ANOVA is a significance test for more than two groups

Design of Experiments

Experiments are any activity, physical, chemical, computational that given some input produces some output

Design of Experiments is systematic approach to plan experiments given the goal of the experiment

Revisiting DoE

Suppliers	Inputs	Process	Outputs	Customers
Providers of Resources	*Resources Required for Process*	*How we Transform inputs into outputs*	*Deliverables of Process*	*Receivers of Process Outputs*
Ingredient	**Cake Ingredients**	1. Preheat Oven	Cake **flavor**	Bakery Customers
Vendors	**Mixing Equipment**	**2. Mix Dry Ingredients**	Cake **texture**	Party Guests
	Mixing Bowls	**3. Mix Wet Ingredients**	Cake **color**	
Grocery Store	**Cake Pan**	4. Combine Wet and Dry	Cake **size**	
	The Oven	5. Pour into Pan	Cake **shape**	
Home Goods	**Baking Temperature**	**6. Bake the cake**		
Store	**Baking Time**	7. Remove and cool	**Yield,**	
(Equipment)	**Measuring Cup**	8. Decorate	**Lead Time,**	
	The Bakers	9. Package and Ship	**Product Cost**	

Factors,
Controllable Variables,
Independent Variables

Response Variables,
Dependent Variables

Design the Experiment

What combination of **input factors** (**Temperature, Time** and **Sugar**) produces the **output (best tasting cake)**?

Factor	High Level (+)	Low Level (-)
Temperature	400° F	350° F
Time	20 Min.	15 Min.
Sugar	3 Cups	2.5 Cups

A *level* refers to a specific *setting* (Set point) of a **factor (Input)**.

Treatments

What combination of **input factors** (**Temperature, Time** and **Sugar**) produces the **output (best tasting cake)**?

Factor	High Level (+)	Low Level (-)
Temperature	400° F	350° F
Time	20 Min.	15 Min.
Sugar	3 Cups	2.5 Cups

	Factors		
	Temperature	Time	Sugar
Treatment 1	+	+	+
Treatment 2	+	+	-

A *treatment* is a unique combination of *factors and levels* within an experiment.

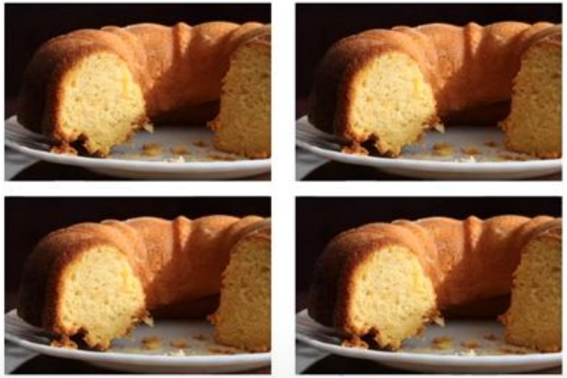
Factor	High Level (+)	Low Level (-)
Temperature	400° F	350° F
Time	20 Min.	15 Min.
Sugar	3 Cups	2.5 Cups

We're looking for the **best tasting cake** across all these unique combinations!

	Factors		
	Temperature	Time	Sugar
Treatment 1	+	+	+
Treatment 2	+	+	-
Treatment 3	+	-	+
Treatment 4	+	-	-
Treatment 5	-	+	+
Treatment 6	-	+	-
Treatment 7	-	-	+
Treatment 8	-	-	-

Replications

Replication is the act of performing an **experiment all over again.**



Replicates

Replication provides **degrees of freedom** for **ANOVA Analysis**.

Replications!

87 Octane	89 Octane	91 Octane	93 Octane
168.8	178.0	179.9	185.3
171.5	178.0	183.4	189.7
172.6	173.6	180.1	187.1
168.1	173.7	181.8	187.7
167.1	172.8	179.4	186.2
170.9	176.4	183.9	188.3
168.3	172.9	184.2	186.6
169.2	178.5	184.0	191.2
168.4	176.8	183.3	187.4
171.8	172.3	182.5	185.4
169.7	175.3	182.2	187.5

Variation **WITHIN** a Treatment (MS_E)

ANOVA with DoE

Collect data from repeated experiments

Conclude from the observations

Use ANOVA to confirm or reject

Quiz 10

1. Write 2 things that you liked/understood today.
2. Write 2 things that you did not like/understand today.

3. When you should use ANOVA?
4. When you should use DoE ?

5. Pick any process from your experience, and describe how you can improve it through DoE and ANOVA