Analysis of Variance

Agenda

- Analysis of variance
  - One way ANOVA
    - Total Variation
    - Variation within treatment
    - Variation between treatment
  - Post-Hoc Test for ANOVA

Question

Ryan is a production manager at an industry manufacturing alloy wires. They have 4 machines - A, B, C and D.

Ryan wants to study whether all the machines have equal efficiency based on the tensile strength of the alloy wire.

Is it possible to test his claim?

Solution

The trivial solution is conducting multiple t tests. However performing multiple t tests has an effect on the type I error.

As the number of t-tests increases the probability of at least one type I error increases.

However, it is possible to test Ryans claim by using one way analysis of variance (one way ANOVA) where the probability of type I error does not change

Multiple t tests and type I error

- For a true null hypothesis, the probability of not obtaining a significant result is 0.95 if the $\alpha = 0.05$

- Say you conduct the t-test twice, the probability of not obtaining one or more significant result is $0.95 \times 0.95 = 0.9025$

- Thus the probability of at least one type error is $1-0.9025 = 0.0975$ (for two t-tests)

Multiple t tests and type I error

Number of t tests	Probability of not obtaining one or more significant result	Probability of at least one type I error
3 t tests	0.95 x 0.95 x 0.95 = 0.857	0.143
4 t tests	0.95 x 0.95 x 0.95 x 0.95 = 0.815	0.185
5 t tests	0.95 x 0.95 x 0.95 x 0.95 x 0.95 = 0.774	0.226

As the number of tests increase the probability of at least one type error also increases

ANOVA - History

- ANOVA was first introduced by Prof R. A. Fisher in 1920’s
- He developed ANOVA while dealing with agronomic data

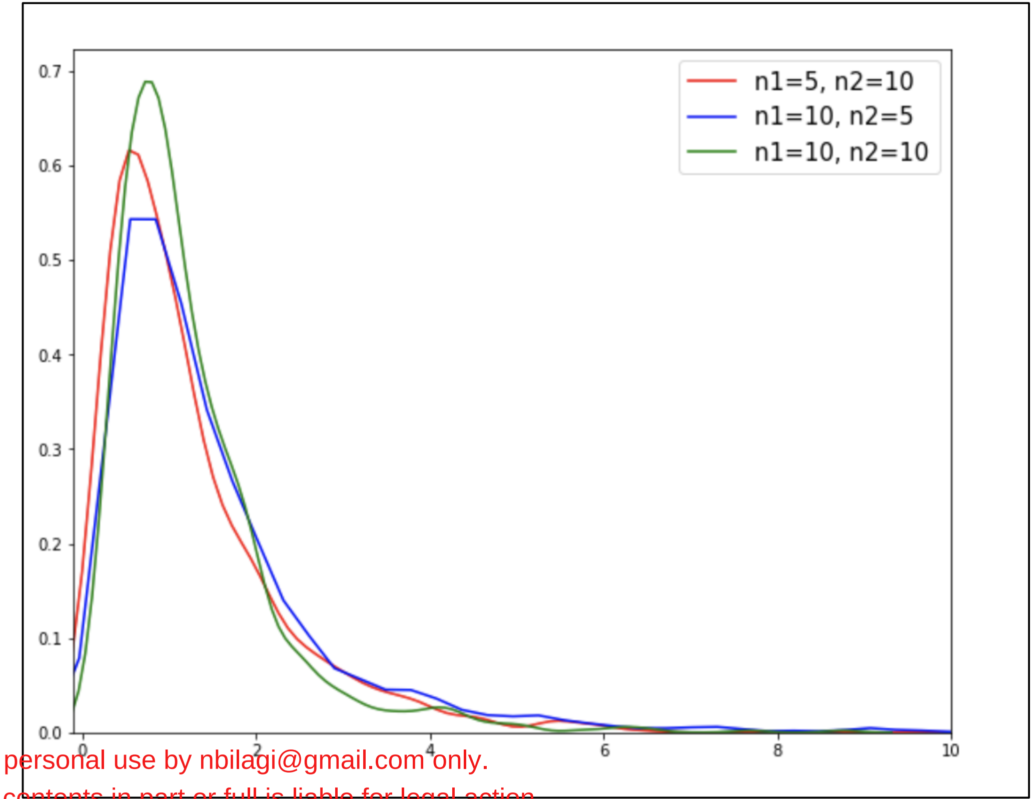
One way ANOVA

- A t-test is used when two unpaired data are compared
- To test the equality of population means for two or more unrelated samples ANOVA technique is used
- Each group is considered to be a treatment
- It is based on the F-distribution

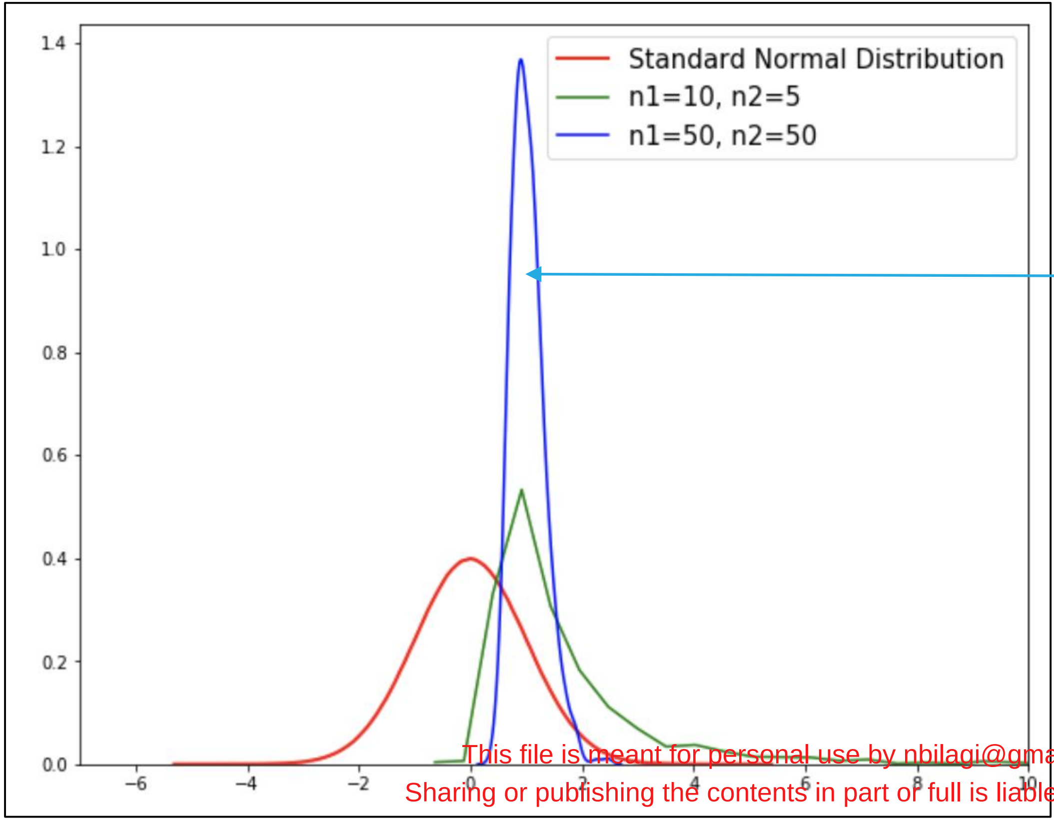
F distribution

- Let X be χ^2_m distribution and let Y be χ^2_n

- Then the ratio $\frac{X/m}{Y/n}$ follows F distribution with (m,n) df



F distribution



As n_1 and n_2 become large the F distribution becomes symmetric

One way ANOVA - assumption

- The samples should be independent
- Each sample should be from normally distributed population
- The population variance of the samples should be equal (homoscedastic)

One way ANOVA

- The null hypothesis to be tested is

H_0: The averages of all treatments are same.
i.e. $\mu_1 = \mu_2 = \ldots = \mu_n$

H_1: At the least one treatment has a different average

t

- Failing to reject H_0, implies that all treatments have the same average

One way ANOVA

- Suppose Ryan collects data for tensile strength of wires produced by each machine
- It is said there are 4 treatments (t = 4)
- Each treatment has 5 observations ($n_i = 5$) where i = 1, 2, …, t
- Total number of observations is given by N

$$N = \sum_i^t n_i$$

A	B	C	D
68.7	62.7	55.9	80.7
75.4	68.5	56.1	70.3
70.9	63.1	57.3	80.9
79.1	62.2	59.2	85.4
78.2	60.3	50.1	82.3

One way ANOVA

- Let μ_i (i=1, 2, …, t) denote the average strength due to each machine

- For our example, t = 4

- The test hypothesis can be written as

$H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4$ Against H_1: At least μ_i is different

A	B	C	D
68.7	62.7	55.9	80.7
75.4	68.5	56.1	70.3
70.9	63.1	57.3	80.9
79.1	62.2	59.2	85.4
78.2	60.3	50.1	82.3

One way ANOVA

- In one way ANOVA, the entire population variance is split into two component
    - Variation within treatment
    - Variation between treatment

- Total variation = Within treatment variation + Between treatment variation

Total variation

- It is the total sum of squares (TSS)
- Let x_{ij} be the observations in the i^{th} treatment and j^{th} row
- $\bar{x}_{..}$ is the grand mean, i.e. the mean of all observations
- The total variation is given by

$$TSS = \underbrace{\sum_i^t}_{\text{Summation over all treatments}} \underbrace{\sum_j^{n_i}}_{\text{Summation over all observation in the treatment}} (x_{ij} - \bar{x}_{..})^2$$

Within treatment variation

- It is the treatment sum of squares (TrSS)
- Let $x_{i.}$ be the observations in the i^{th} treatment with n_i in observation in each treatment and $\bar{x}_{i.}$ is the mean over i^{th} treatment
- $\bar{x}_{..}$ is the grand mean, i.e. the mean of all observations
- The treatment variation is given by

$$TrSS = \sum_i^t \sum_j^{n_i} n_i(\bar{x}_{i.} - \bar{x}_{..})^2$$

Summation over all treatments

Summation over all observation in the treatment

Between treatment variation

- It is the error sum of squares (ESS)
- Let $x_{i.}$ be the observations in the i^{th} treatment and is $\bar{x}_{.j}$ the mean over j^{th} row
- $\bar{x}_{..}$ is the grand mean, i.e. the mean of all observations
- The error sum of squares is given by

$$ESS = \sum_i^t \sum_j^{n_i} (x_{ij} - \bar{x}_{i.})^2$$

Summation over all treatments

Summation over all observation in the treatment

PLEASE!! NOTE

Error sum of squares

During problem solving, the error sum of squares is obtained as:

$$\mathrm{ESS} = \mathrm{TSS} - \mathrm{TrSS}$$

One way ANOVA

- The test statistic is given by

$$\text{F-ratio} = \frac{\frac{TrSS}{df_{Tr}}}{\frac{ESS}{df_e}} = \frac{\text{MTrSS}}{\text{MESS}}$$

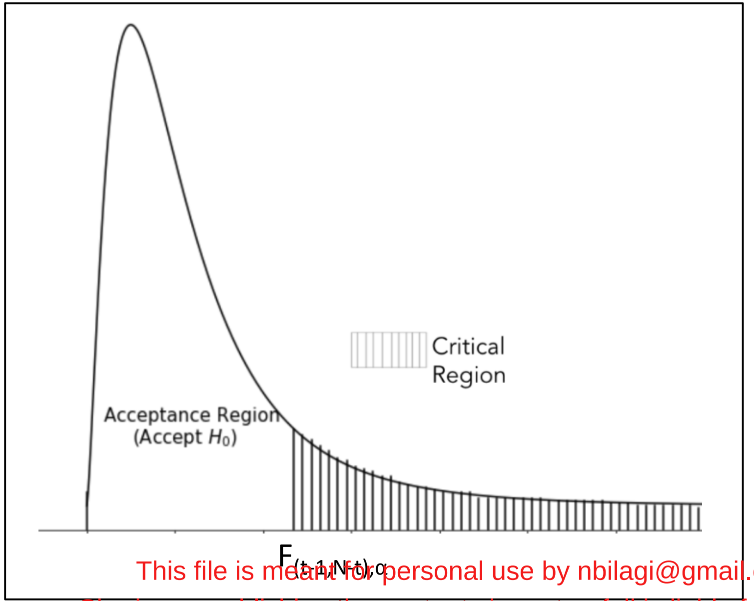
Mean Treatment Sum of Squares

Mean Error Sum of Squares

- Under H_0, the test statistic follows F-distribution with (df_{Tr}, df_e) degrees of freedom

One way ANOVA

Decision Rule: If $F_{cal} \geq F_{(t-1,N-t),\alpha}$ or p-value $\leq \alpha$, then we reject H_0 at α% level of significance



One way ANOVA

To ease the entire computational process, an ANOVA table is prepared as follows:

<table>
<tr><th>Source of variation</th><th>Degrees of freedom</th><th>Sum of Squares</th><th>Mean Sum of Squares</th><th>F-ratio</th></tr>
<tr><td>Treatment</td><td>t-1</td><td>TrSS</td><td>s^2_t</td><td rowspan="2">$\frac{s_t^2}{s_e^2}$</td></tr>
<tr><td>Error</td><td>N-t</td><td>ESS</td><td>s^2_e</td></tr>
<tr><td>Total</td><td>N-1</td><td>TSS</td><td></td><td></td></tr>
</table>

One way ANOVA - procedure

1. State the hypothesis to be tested
2. Compute the sum of squares
   a. The total sum of squares, TSS = $\sum^{t}_{j=1}\sum^{ni}_{i=1} (x_{ij} - \bar{x}_{..})^2$
   b. The treatment sum of squares TrSS = $\sum^{t}_{j=1}\sum^{ni}_{i=1} n_i(x_{ij} - \bar{x}_{i.})^2$
   c. The Error sum of squares, ESS = TSS - TrSS
3. Compute mean sum of squares
   a. s^2_t = Mean group sum of squares (MTrSS) = TrSS/(t-1)
   b. s^2_e = Mean error sum of squares (MESS) = ESS/(N-t)

One way ANOVA - procedure

4. Compute the F-ratio

$$\text{F-ratio} = \frac{\text{MTrSS}}{\text{MESS}} = \frac{s_t^2}{s_e^2}$$

4. Prepare the ANOVA table

5. Write the decision and conclusion accordingly

One way ANOVA

Question:

Ryan is a production manager at an industry manufacturing alloy seals. They have 4 machines - A, B, C and D. Ryan wants to study whether all the machines have equal efficiency.

Ryan collects data of tensile strength (in N/m^2) from all the 4 machines as given.

Test at 5% level of significance.

A	B	C	D
68.7	62.7	55.9	80.7
75.4	68.5	56.1	70.3
70.9	63.1	57.3	80.9
79.1	62.2	59.2	85.4
78.2	60.3	50.1	82.3

One way ANOVA

Solution:

Ryan is a production manager at an industry manufacturing alloy seals. They have 4 machines - A, B, C and D

Let μ_1 be the average tensile strength due to machine A

μ_2 be the average tensile strength due to machine B

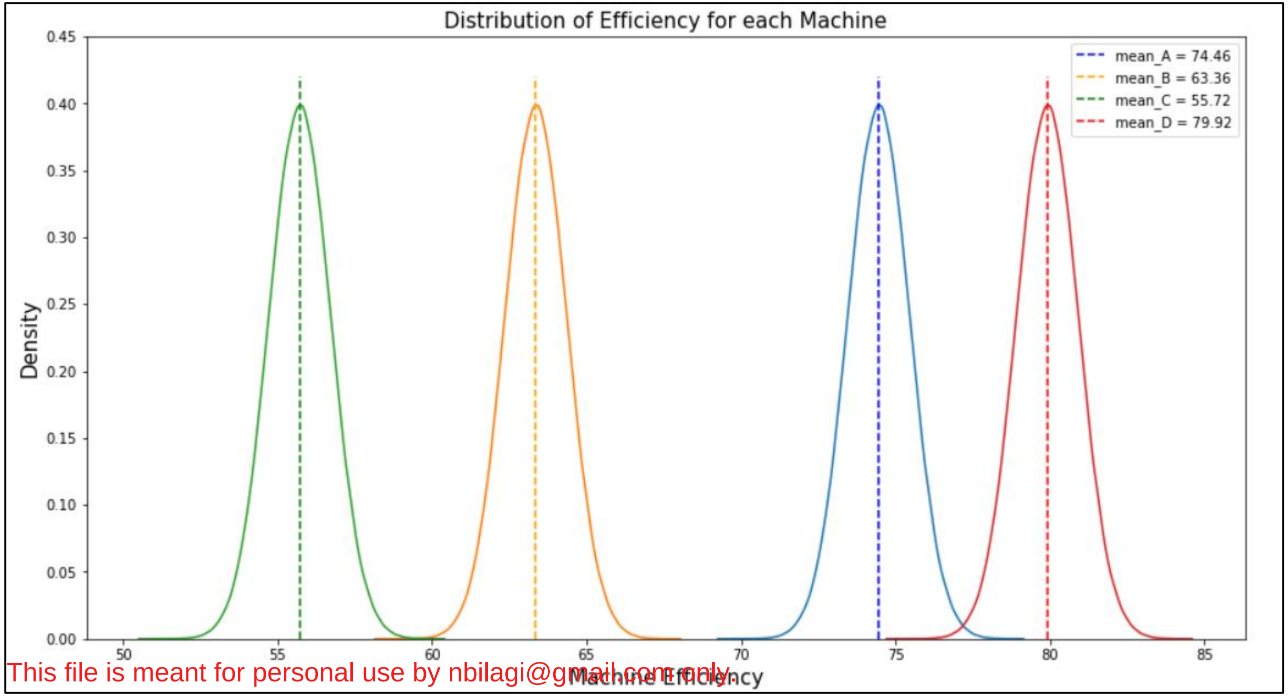
μ_3 be the average tensile strength due to machine C

μ_4 be the average tensile strength due to machine D

To test, $H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4$ Against H_1: At least one μ_i is different (i=1, 2, 3, 4)

One way ANOVA

The plot shows the difference between the average efficiency for each machine, which indicates the rejection of H_0.



One way ANOVA

Solution:

The grand mean:

$$\bar{x}_{..} = \frac{68.7+62.7+...+50.1+82.3}{20} = 68.365 \text{ N/m}^2$$

The total sum of squares:

$$TSS = \sum_i^t \sum_j^{n_i} (x_{ij} - \bar{x}_{..})^2$$

$$= (68.7 - 69.115)^2 + \ldots + (81.12 - 69.115)^2$$

$$= 2074.1255 \ (\text{N/m}^2)^2$$

A	B	C	D
68.7	62.7	55.9	80.7
75.4	68.5	56.1	70.3
70.9	63.1	57.3	80.9
79.1	62.2	59.2	85.4
78.2	60.3	50.1	82.3

One way ANOVA

Solution:

The treatment sum of squares is

$$TrSS = \sum_i^t \sum_j^{n_i} n_i(\overline{x}_{i.} - \overline{x}_{..})^2$$

$$= 5(74.46 - 68.365)^2 + \ldots + 5(79.92 - 68.365)^2$$

$$= 1778.0655\ (\mathrm{N/m}^2)^2$$

	A	B	C	D
	68.7	62.7	55.9	80.7
	75.4	68.5	56.1	70.3
	70.9	63.1	57.3	80.9
	79.1	62.2	59.2	85.4
	78.2	60.3	50.1	82.3
$\sum x_i$	372.3	316.8	278.6	399.6
$\bar{x}_{i.}$	74.46	63.36	55.72	79.92

One way ANOVA

Solution:

The error sum of squares can also be calculated as

$$ESS = \sum_i^t \sum_j^{n_i} (x_{ij} - \bar{x}_{i.})^2$$

$$= (68.7 - 74.46)^2 + \ldots + (82.3 - 79.92)^2$$

$$= 296.06$$

	A	B	C	D
	68.7	62.7	55.9	80.7
	75.4	64.5	56.1	80.3
	70.9	63.1	57.3	80.9
	79.1	59.2	55.2	81.4
	78.2	60.3	50.1	82.3
$\bar{x}_{i.}$	74.46	63.36	55.72	79.92

Or can be obtained as $\mathrm{ESS} = \mathrm{TSS} - \mathrm{TrSS} = 296.06$

One way ANOVA

Solution:

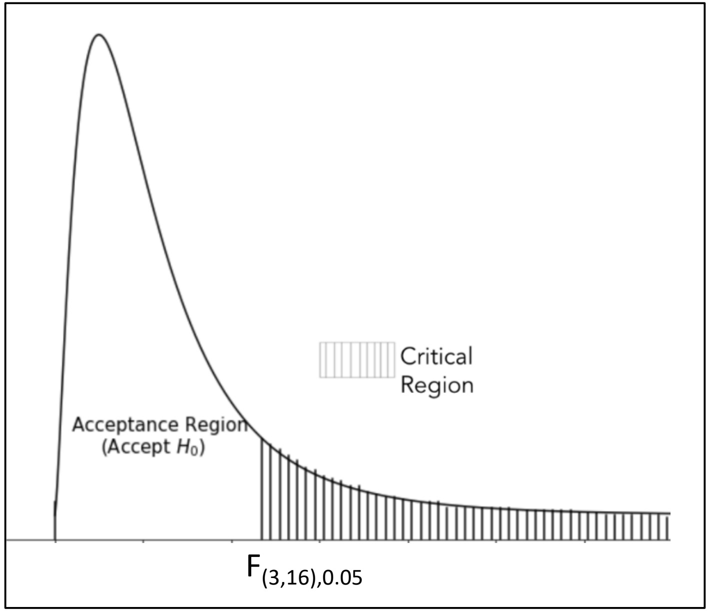
Source of variation	Degrees of freedom	Sum of Squares	Mean Sum of Squares	F-ratio
Treatment	t-1 = 4-1 =3	TrSS = 1778.0655	$s_t^2 = \frac{21778.0655}{3} = 592.6885$	$\frac{s_t^2}{s_e^2} = 32.031$
Error	N-t = 20-4 =16	ESS = 296.06	$s_e^2 = \frac{269.06}{16} = 18.50375$	
Total	N-1 = 20 -1 =19	TSS = 2241.5255		

One way ANOVA

Solution:

From the F-table we have $F_{(3,16),0.05} = 3.24$

Since $3.24 < 32.03$, we reject H_0.



One way ANOVA

Python solution:

```
# perform one-way ANOVA
# pass the given data using the dataframe 'df_machine'
test_stat, p_val = stats.f_oneway(df_machine[df_machine['machine'] == 'machine_A']['strength'],
                                  df_machine[df_machine['machine'] == 'machine_B']['strength'],
                                  df_machine[df_machine['machine'] == 'machine_C']['strength'],
                                  df_machine[df_machine['machine'] == 'machine_D']['strength'])

# print the test statistic and p-value
print('Test statistic:', test_stat)
print('p_value:', p_val)
```

```
Test statistic: 32.03072350199285
p_value: 5.375613532781072e-07
```

As p-value < 0.05, we reject H_0.

One way ANOVA can be said to check the effect of a nominal variable over a numerical variable.

Further analysis

- In the example, Ryan has tested for strength of materials due to 4 machines
- The null hypothesis for ANOVA was rejected
- Now it is of Ryan's interest to know which machine(s) has a different outcome

How would he find out?

Further analysis

- If we fail to reject the null hypothesis, it implies that all the treatments have the same effect
- However, if the null hypothesis is rejected, it implies that at least one treatment has a different average
- To know which treatment(s) has/have different outcome
- Can be found out using the post hoc tests

Post-Hoc Tests

Post-hoc test

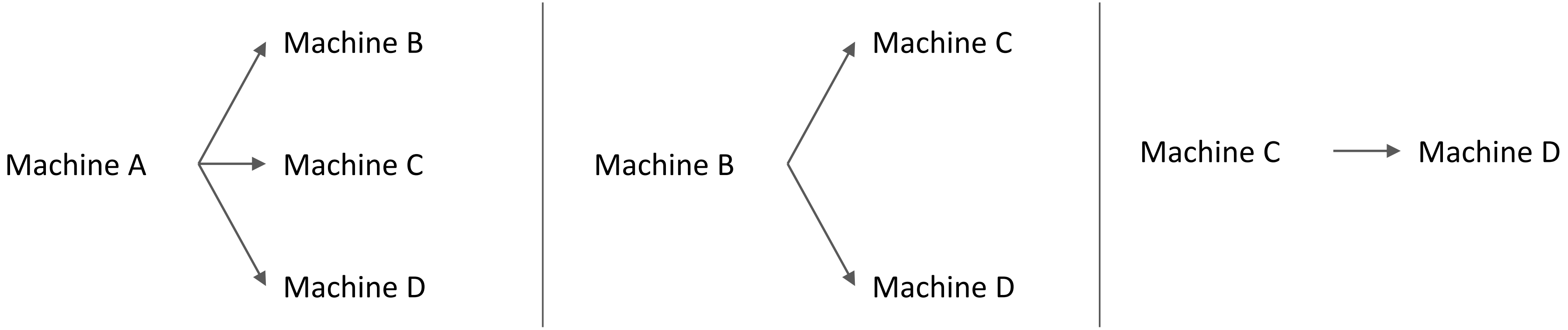
- A post hoc is conducted after the null hypothesis of ANOVA is rejected to determine the different treatments(s)

- There are various post hoc tests available such as:
    - Tukey’s HSD test (Tukey’s Honest(ly) Significant Difference test)
    - Scheffe test
    - Duncan's Multiple Range test
    - Fisher's’ LSD test (Fisher’s Least Significant Difference test)
    - Bonferroni test

- We will study the Tukey’s HSD test in detail

Post-hoc test

- Consider our example where Ryan wants to find out the which machines had different result
- Each pair of machines is tested for the statistical difference



Post-hoc test

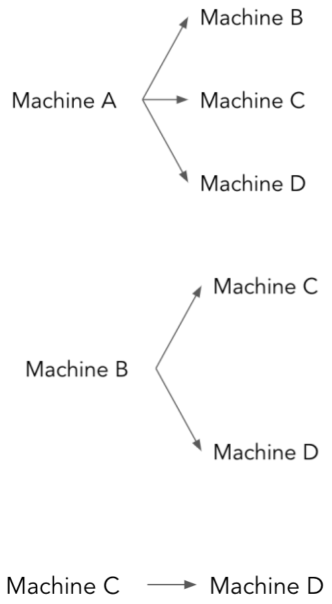
Thus the test hypothesis are

$H_{01}: \mu_{machine\_A} = \mu_{machine\_B}$ Against $H_{11}: \mu_{machine\_A} \neq \mu_{machine\_B}$

$H_{02}: \mu_{machine\_A} = \mu_{machine\_C}$ Against $H_{12}: \mu_{machine\_A} \neq \mu_{machine\_C}$

$H_{03}: \mu_{machine\_A} = \mu_{machine\_D}$ Against $H_{13}: \mu_{machine\_A} \neq \mu_{machine\_D}$

$H_{04}: \mu_{machine\_B} = \mu_{machine\_C}$ Against $H_{14}: \mu_{machine\_A} \neq \mu_{machine\_C}$



Post-hoc test

The test statistic is:

Obtained from the Tukey table

$$T_\alpha = q_{\alpha,(t,f)} \sqrt{\frac{MSE}{n}}$$

t: total treatments

f: error degrees of freedom

MSE: Mean error sum of squares (from ANOVA table)

n: number of observations in a group

Post-hoc test

- Consider the absolute difference between two treatments $|\bar{x}_i - \bar{x}_j|$
- The decision rule: Reject H_0, if the absolute difference $\geq T_\alpha$
- The python code:

  First create the DataFrame df_machine then use the following function

```
# perform tukey's HSD test to compare the mean efficiency for pair of machines
# pass the tensile strength to the parameter, 'data'
# pass the name of the machine to the parameter, 'groups'
comp = mc.MultiComparison(data = df_machine['strength'], groups = df_machine['machine'])

# tukey's HSD test
post_hoc = comp.tukeyhsd()

# print the summary table
post_hoc.summary()
```

Post-hoc test

The output is as follows:

```
    Multiple Comparison of Means - Tukey HSD, FWER=0.05
==============================================================
  group1    group2   meandiff p-adj   lower    upper   reject
--------------------------------------------------------------
machine_A machine_B    -11.1 0.0044 -18.8842  -3.3158   True
machine_A machine_C   -18.74  0.001 -26.5242 -10.9558   True
machine_A machine_D     5.46 0.2265  -2.3242  13.2442  False
machine_B machine_C    -7.64 0.0553 -15.4242   0.1442  False
machine_B machine_D    16.56  0.001   8.7758  24.3442   True
machine_C machine_D     24.2  0.001  16.4158  31.9842   True
--------------------------------------------------------------
```

True: reject H_0

False: fail to reject H_0 (accept H_0)

It can been seen that there is statistical difference between pairs of machines (A,B), (A,C), (B,D), and (C,D).

- For equal number of observations in each treatment, tukey HSD test can be used
- However when the data is unequal it is not efficient
- In such a scenario, one may use the Scheffe test