Analiza wariancji i metody klasyfikacyjne

Analysis of variance and classification methods

lecture 4

28 October 2019

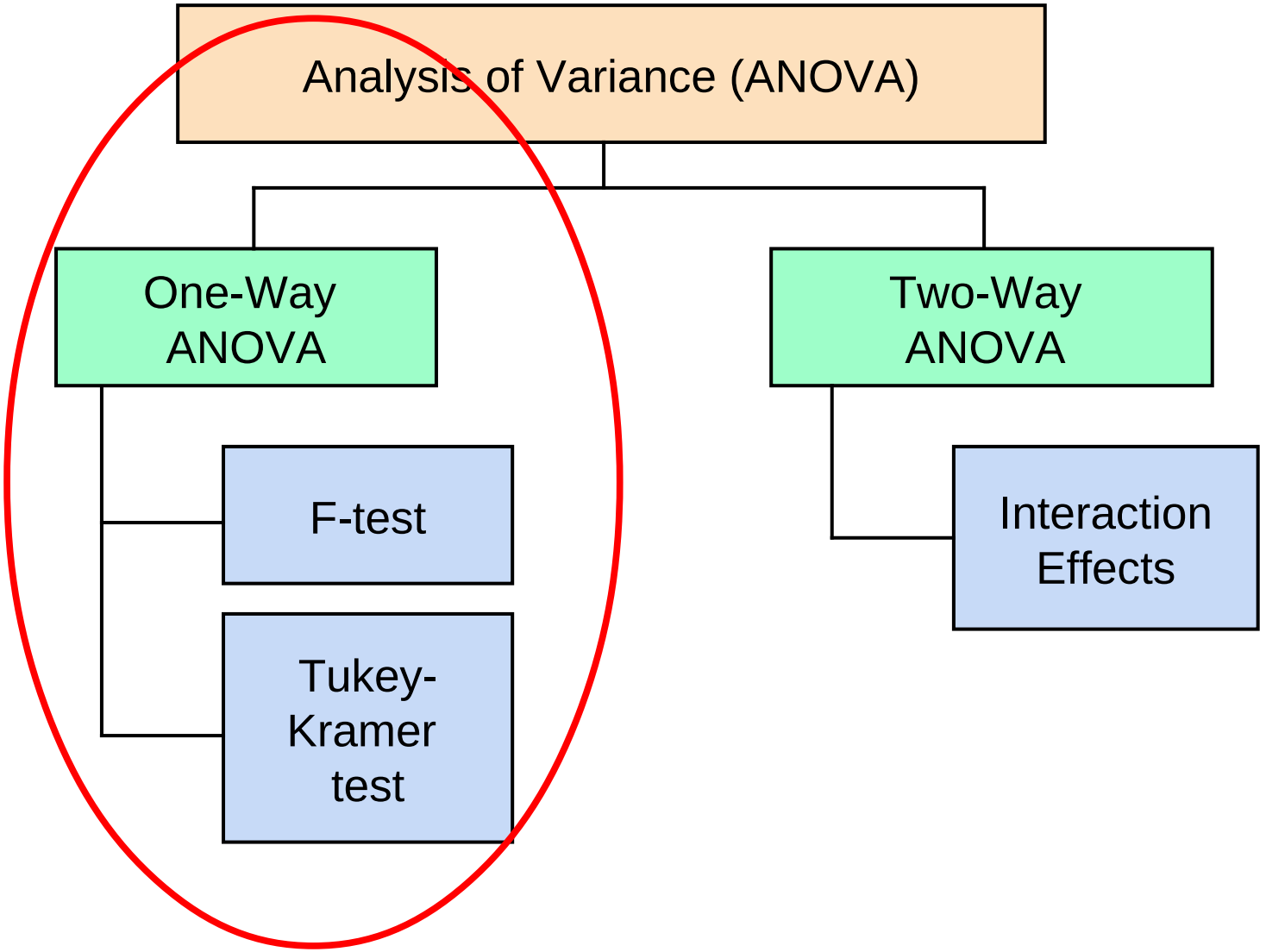
Ilona Anna Urbaniak (PK)

Marcin Wolter (IFJ PAN)

e-mail: marcin.wolter@ifj.edu.pl, phone: 12 662 8024

Slides: https://indico.ifj.edu.pl/event/271/

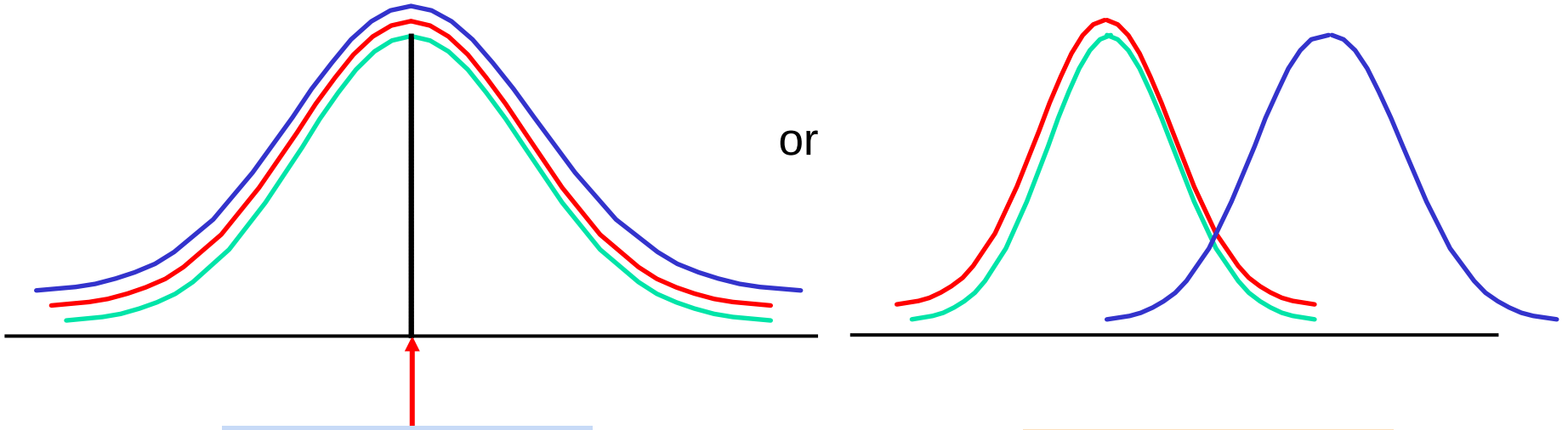
What we have learned?



One-Factor ANOVA

$H_0 : \mu_1 = \mu_2 = \mu_3 = \cdots = \mu_c$

$H_1 :$ Not all μ_i are the same



$\mu_1 = \mu_2 = \mu_3$

$\mu_1 = \mu_2 \neq \mu_3$

One-factor ANOVA

$$MSA = \frac{SSA}{c - 1}$$

Mean Square Among

$$MSW = \frac{SSW}{n - c}$$

Mean Square Within

c = number of groups
n = sum of the sample sizes from all groups

$$F\ ratio = \frac{MSA}{MSW}$$

F ratio – test statistic (a measure of "difference" between distributions).

F follows the Fisher–Snedecor distribution

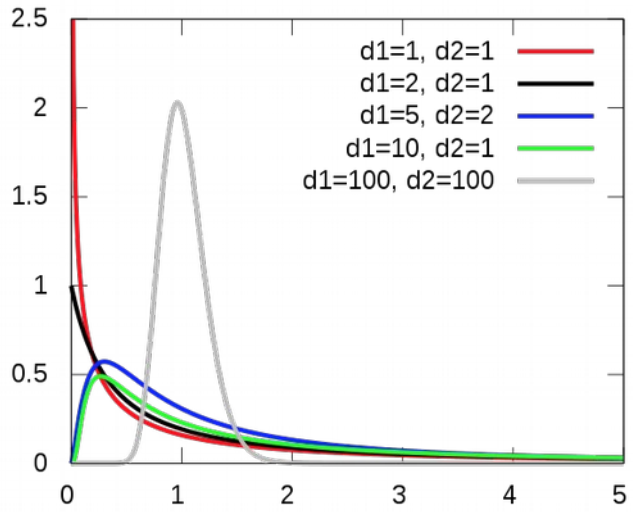
- If a random variable X has an F-distribution with parameters d_1 and d_2 (degrees of freedom), we write $X \sim F(d_1, d_2)$. Then the probability density function for X is given by:

$$f(x; d_1, d_2) = \frac{\sqrt{\frac{(d_1\,x)^{d_1}\;d_2^{d_2}}{(d_1\,x+d_2)^{d_1+d_2}}}}{x\,\mathrm{B}\left(\frac{d_1}{2}, \frac{d_2}{2}\right)}$$

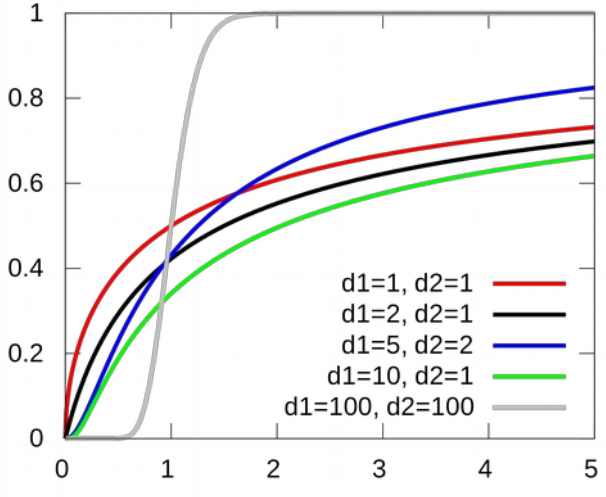
where B is a beta function:

$$\mathrm{B}(x, y) = \int_0^1 t^{x-1}(1-t)^{y-1}\,dt$$

Probability density function



Cumulative

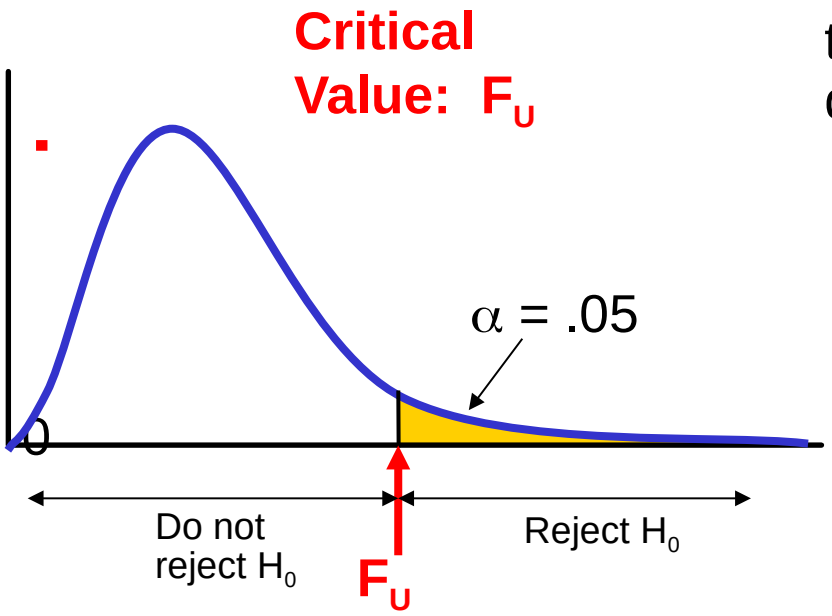


F-test function:
https://www.stat.purdue.edu/~jtroisi/STAT350Spring2015/tables/FTable.pdf

How to accept or reject H_0?

- Calculate the value of F
- For a given probability (typically $\alpha = 0.05$) find a critical value of F called F_U. Remember, the Fisher-Snedecor distribution depends on the number of classes and the number of events:
  - df1 = c – 1 (c – number of groups)
  - Df2 = n - c (n – number of events)
- If $F > F_U$ – reject H_0

  If $F < F_U$ – accept H_0



If $F > F_U$ we know, that the means of at least two groups are significantly different, bu we do not know which ones!

F-test function:
https://www.stat.purdue.edu/~jtroisi/STAT350Spring2015/tables/FTable.pdf

How to check which groups are different? POST-HOC tests

- LSD (Least Significant Difference) – proposed by Ronald Fisher: student test t performed for each pair.
- If we compare group A with B, A with C and B with C, and later we do it once more in the “opposite direction”, I.e B with A, C with A and C with B the probability of type I error cumulates (we can reject H0 hypothesis by mistake).
- Other tests: for example Tuckey-Kramer

Student t-test

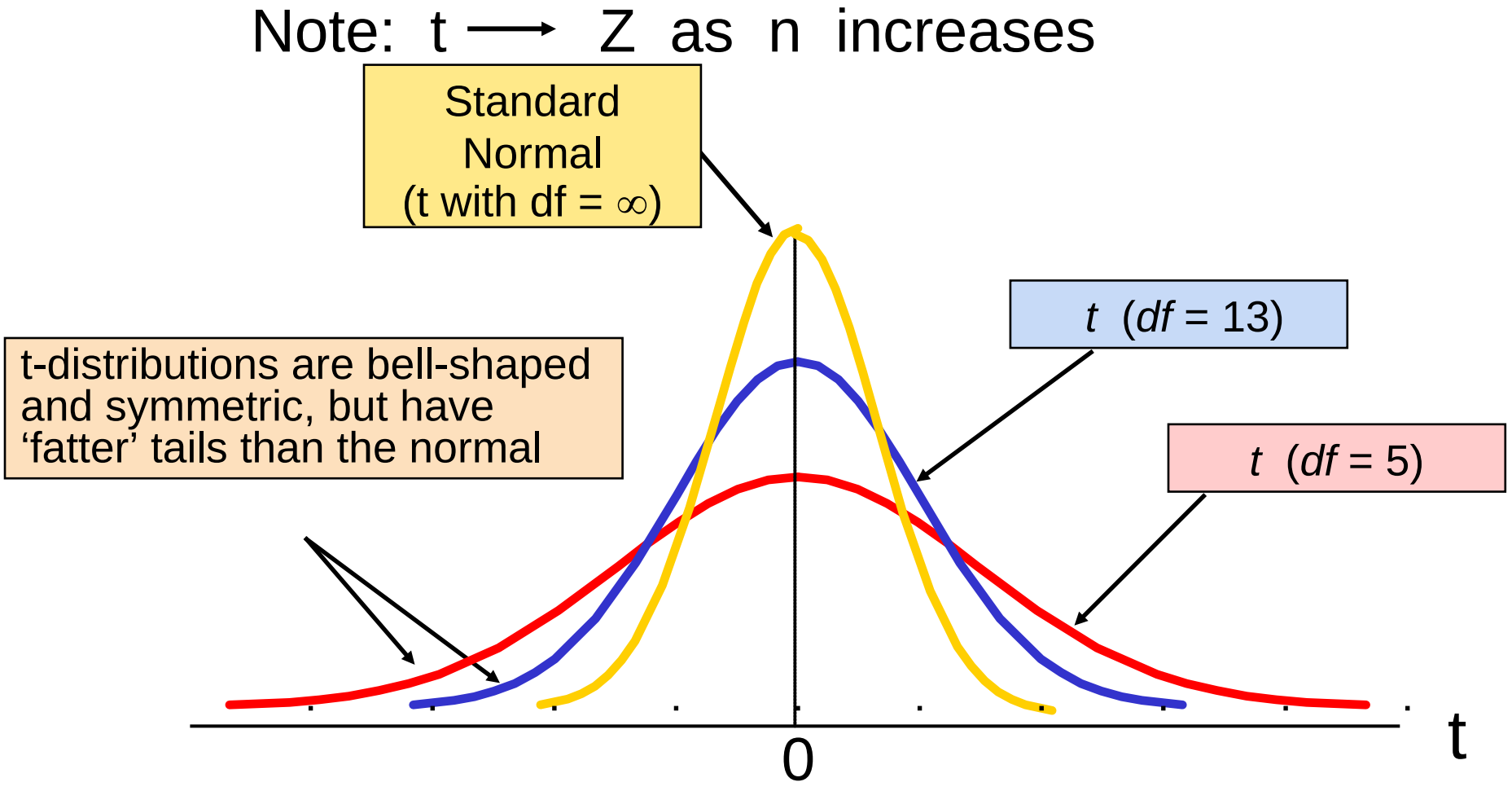
- Introduced in 1908 by William Sealy Gosset, a chemist working for the Guinness brewery.
- For equal sample sizes, equal variance.
  - The t statistic to test whether the means are different can be calculated as follows:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{s_p \sqrt{\frac{2}{n}}} \qquad s_p = \sqrt{\frac{s_{X_1}^2 + s_{X_2}^2}{2}}.$$

  - For non-eaqual sample sizes:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{s_p \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \qquad s_p = \sqrt{\frac{(n_1 - 1)\, s_{X_1}^2 + (n_2 - 1)\, s_{X_2}^2}{n_1 + n_2 - 2}}$$

Student's t Distribution



https://www.stat.purdue.edu/~jtroisi/STAT350Spring2015/tables/TTable.pdf

Example

- Suppose we have to compare the mean value of two groups, one with 7 subjects and the other with 5 subjects .
- These were their scores:

Case	Group 1	Group 2
1	78	87
2	82	92
3	87	86
4	65	95
5	75	73
6	82	
7	71	

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\left[\frac{SS_1 + SS_2}{n_1 + n_2 - 2}\right]\left[\frac{1}{n_1} + \frac{1}{n_2}\right]}} = \frac{77.14 - 86.60}{\sqrt{\left[\frac{334.86 + 285.20}{7 + 5 - 2}\right]\left[\frac{1}{7} + \frac{1}{5}\right]}}$$

$$= \frac{-9.46}{\sqrt{\left(\frac{620.06}{10}\right)\left(\frac{12}{35}\right)}} = \frac{-9.46}{\sqrt{21.26}} = -0.44$$

For an independent or between subjects' t test: df = n1+ n2 - 2

•Now, take the absolute value of this, which is 0.44.

•Now, for the .05 probability level with 10 degrees of freedom, we see from the table that the critical t score is 2.228 for a two-tailed test.

•Since the calculated t score is lower than the critical t score, the results are not significant at the .05 probability level.

The Tukey-Kramer Procedure

- Compare the difference between means divided by the standard error of the sum of the means SE

$$q_s = \frac{|\overline{X_1} - \overline{X_2}|}{SE}$$

to the critical value q obtained from the Studentized Range Distributionfor a given α

$$Critical\ Range = Q_U \sqrt{\frac{MSW}{2}\left(\frac{1}{n_{j'}} + \frac{1}{n_j}\right)}$$

The Tukey-Kramer Procedure: Example

Club 1	Club 2	Club 3
254	234	200
263	218	222
241	235	197
237	227	206
251	216	204

1. Compute absolute mean differences:

$$|\bar{x}_1 - \bar{x}_2| = |249.2 - 226.0| = 23.2$$

$$|\bar{x}_1 - \bar{x}_3| = |249.2 - 205.8| = 43.4$$

$$|\bar{x}_2 - \bar{x}_3| = |226.0 - 205.8| = 20.2$$

2. Find the Q_U value from the table with c = 3 and (n – c) = (15 – 3) = 12 degrees of freedom for the desired level of α (α = .05 used here):

$$Q_U = 3.773$$

See table:
https://www.stat.purdue.edu/~xbw/courses/stat512/q-table.pdf

The Tukey-Kramer Procedure: Example

(continued)

3. Compute Critical Range:

$$\text{Critical Range} = Q_U \sqrt{\frac{MSW}{2}\left(\frac{1}{n_j} + \frac{1}{n_{j'}}\right)} = 3.77 \sqrt{\frac{93.3}{2}\left(\frac{1}{5} + \frac{1}{5}\right)} = 16.285$$

4. Compare:

$$\left|\bar{x}_1 - \bar{x}_2\right| = 23.2$$

$$\left|\bar{x}_1 - \bar{x}_3\right| = 43.4$$

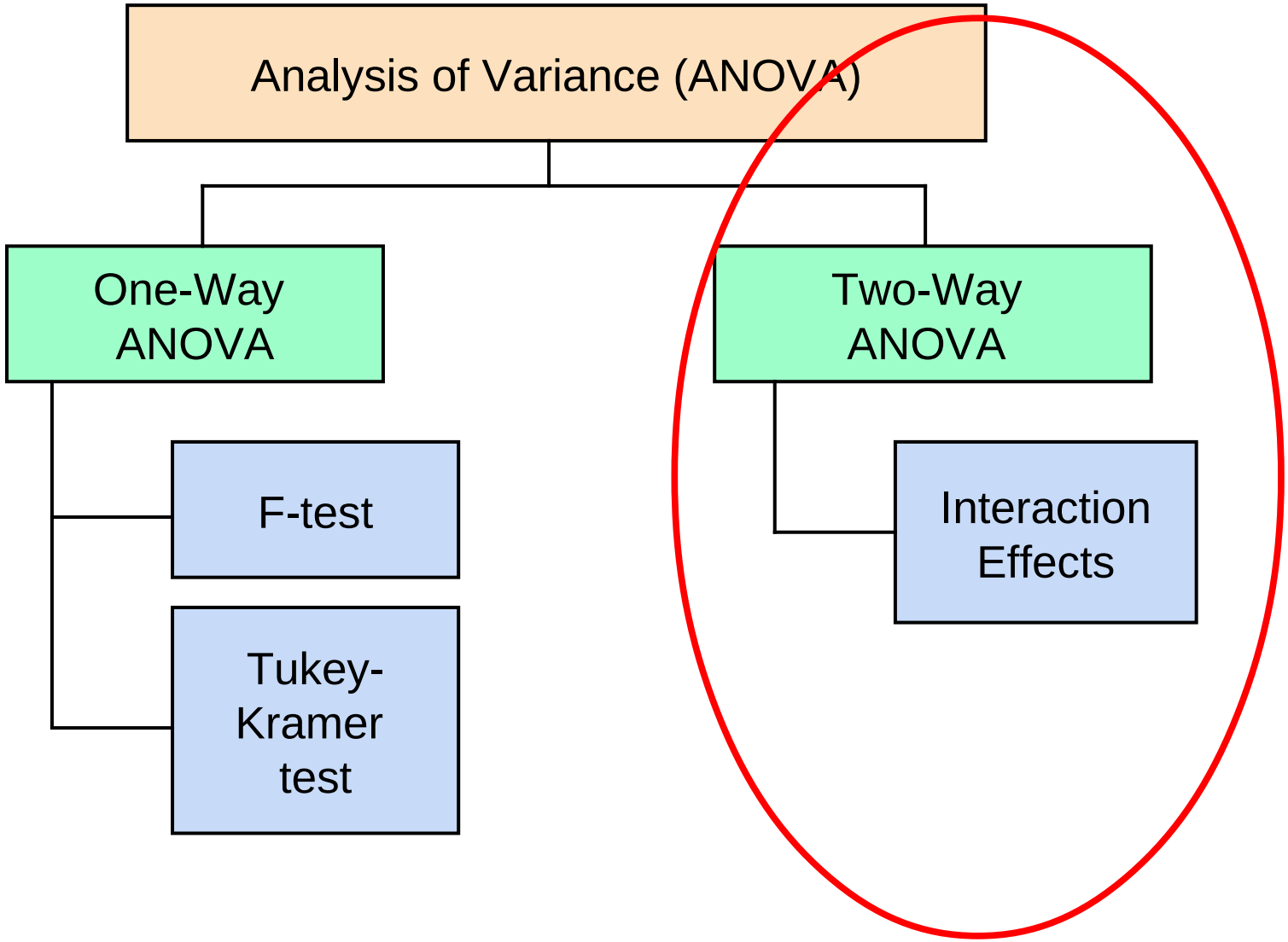
$$\left|\bar{x}_2 - \bar{x}_3\right| = 20.2$$

5. All of the absolute mean differences are greater than critical range. Therefore there is a significant difference between each pair of means at 5% level of significance.

Why Tukey-Kramer?

- It’s easier to find a significant difference using Fisher’s LSD due to fluctuations.

Chapter Overview



Two-Way ANOVA

- Examines the effect of
  - Two factors of interest on the dependent variable
    - e.g., Percent carbonation and line speed on soft drink bottling process
  - Interaction between the different levels of these two factors
    - e.g., Does the effect of one particular carbonation level depend on which level the line speed is set?

Two-Way ANOVA

- **Assumptions**
  - Populations are normally distributed
  - Populations have equal variances
  - Independent random samples are drawn

Two-Way ANOVA
Sources of Variation

Two Factors of interest: A and B

r = number of levels of factor A

c = number of levels of factor B

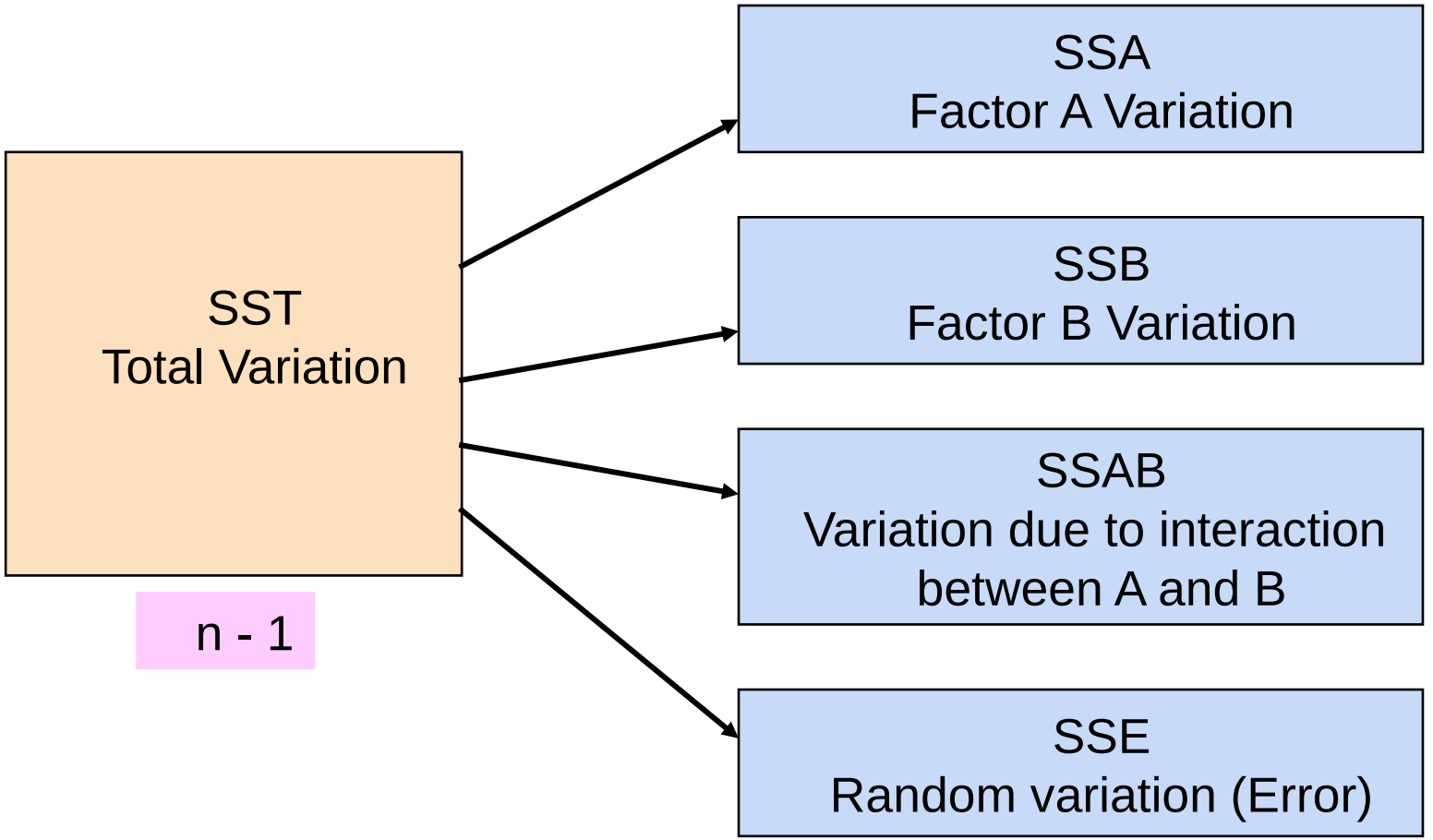
n’ = number of replications for each cell

n = total number of observations in all cells (n = rcn’)

X_{ijk} = value of the k^{th} observation of level i of factor A and level j of factor B

Two-Way ANOVA
Sources of Variation

SST = SSA + SSB + SSAB + SSE



Degrees of Freedom:

r – 1

c – 1

(r – 1)(c – 1)

rc(n' – 1)

Error decomposition

SST = SSA + SSB + SSAB + SSE

$$\underbrace{\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{r}\left(Y_{ijk}-\bar{Y}_{...}\right)^2}_{SS_{Total}} = \underbrace{r\cdot b\cdot\sum_{i=1}^{a}\left(\bar{Y}_{i..}-\bar{Y}_{...}\right)^2}_{SS_A} + \underbrace{r\cdot a\cdot\sum_{j=1}^{3}\left(\bar{Y}_{.j.}-\bar{Y}_{...}\right)^2}_{SS_B}$$

$$+\underbrace{r\times\sum_{i=1}^{a}\sum_{j=1}^{b}\left(\bar{Y}_{ij.}-\bar{Y}_{i..}-\bar{Y}_{.j.}+\bar{Y}_{...}\right)^2}_{SS_{A\times B}} + \underbrace{\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{r}\left(Y_{ijk}-\bar{Y}_{ij.}\right)^2}_{SS_{within}}$$

Two Factor ANOVA Equations

Total Variation:

$$SST = \sum_{i=1}^{r} \sum_{j=1}^{c} \sum_{k=1}^{n'} (X_{ijk} - \overline{\overline{X}})^2$$

Factor A Variation:

$$SSA = cn' \sum_{i=1}^{r} (\overline{X}_{i..} - \overline{\overline{X}})^2$$

Factor B Variation:

$$SSB = rn' \sum_{j=1}^{c} (\overline{X}_{.j.} - \overline{\overline{X}})^2$$

Two Factor ANOVA Equations

Interaction Variation:

$$SSAB = n' \sum_{i=1}^{r} \sum_{j=1}^{c} (\bar{X}_{ij.} - \bar{X}_{i..} - \bar{X}_{.j.} + \bar{X})^2$$

Sum of Squares Error:

$$SSE = \sum_{i=1}^{r} \sum_{j=1}^{c} \sum_{k=1}^{n'} (X_{ijk} - \bar{X}_{ij.})^2$$

Two Factor ANOVA Equations

where:

$$\overline{\overline{X}} = \frac{\sum_{i=1}^{r}\sum_{j=1}^{c}\sum_{k=1}^{n'} X_{ijk}}{rcn'} = \text{Grand Mean}$$

$$\overline{X}_{i..} = \frac{\sum_{j=1}^{c}\sum_{k=1}^{n'} X_{ijk}}{cn'} = \text{Mean of } i^{th} \text{ level of factor A} \quad (i = 1, 2, \ldots, r)$$

$$\overline{X}_{.j.} = \frac{\sum_{i=1}^{r}\sum_{k=1}^{n'} X_{ijk}}{rn'} = \text{Mean of } j^{th} \text{ level of factor B} \quad (j = 1, 2, \ldots, c)$$

$$\overline{X}_{ij.} = \sum_{k=1}^{n'} \frac{X_{ijk}}{n'} = \text{Mean of cell ij}$$

r = number of levels of factor A
c = number of levels of factor B
n′ = number of replications in each cell

Mean Square Calculations

$$\text{MSA} = \text{Mean square factor A} = \frac{\text{SSA}}{r - 1}$$

$$\text{MSB} = \text{Mean square factor B} = \frac{\text{SSB}}{c - 1}$$

$$\text{MSAB} = \text{Mean square interaction} = \frac{\text{SSAB}}{(r - 1)(c - 1)}$$

$$\text{MSE} = \text{Mean square error} = \frac{\text{SSE}}{rc(n' - 1)}$$

Two-Way ANOVA: The F Test Statistic

F Test for Factor A Effect

H_0: $\mu_{1..} = \mu_{2..} = \mu_{3..} = \bullet\bullet\bullet$

H_1: Not all $\mu_{i..}$ are equal

$$F = \frac{MSA}{MSE}$$

Reject H_0 if $F > F_U$

F Test for Factor B Effect

H_0: $\mu_{.1.} = \mu_{.2.} = \mu_{.3.} = \bullet\bullet\bullet$

H_1: Not all $\mu_{.j.}$ are equal

$$F = \frac{MSB}{MSE}$$

Reject H_0 if $F > F_U$

F Test for Interaction Effect

H_0: the interaction of A and B is equal to zero

H_1: interaction of A and B is not zero

$$F = \frac{MSAB}{MSE}$$

Reject H_0 if $F > F_U$

Two-Way ANOVA Summary Table

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Squares	F Statistic
Factor A	**SSA**	**r – 1**	**MSA** = SSA /(r – 1)	$\frac{\textbf{MSA}}{\textbf{MSE}}$
Factor B	**SSB**	**c – 1**	**MSB** = SSB /(c – 1)	$\frac{\textbf{MSB}}{\textbf{MSE}}$
AB (Interaction)	**SSAB**	**(r – 1)(c – 1)**	**MSAB** = SSAB / (r – 1)(c – 1)	$\frac{\textbf{MSAB}}{\textbf{MSE}}$
Error	**SSE**	**rc(n' – 1)**	**MSE =** SSE/rc(n' – 1)	
Total	**SST**	**n – 1**		

Two-Factor ANOVA
With Replication

- As production manager, you want to see if 3 filling machines have different mean filling times when used with 5 types of boxes. At the .05 level, is there a difference in machines, in boxes? Is there an interaction?

Box	Machine1	Machine2	Machine3
1	25.40	23.40	20.00
	26.40	24.40	21.00
2	26.31	21.80	22.20
	25.90	23.00	22.00
3	24.10	23.50	19.75
	24.40	22.40	19.00
4	23.74	22.75	20.60
	25.40	23.40	20.00
5	25.10	21.60	20.40
	26.20	22.90	21.90

Summary Table

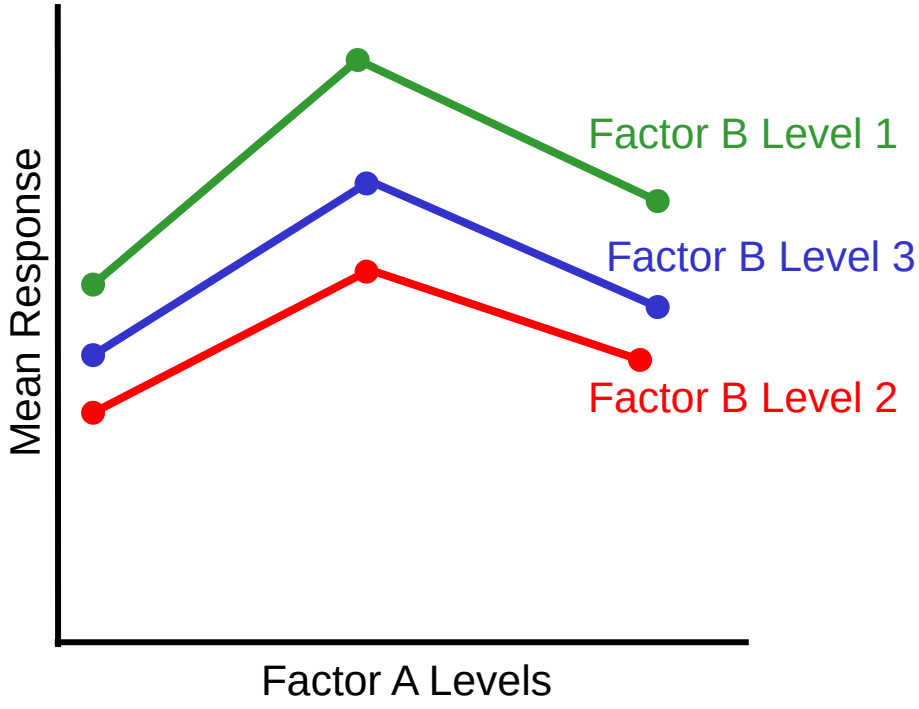
Source of Variation	Degrees of Freedom	Sum of Squares	Mean Square	F	P-Value
Sample (Boxes)	5 - 1 = 4	7.4714	1.8678	3.6868	.0277
Columns (Machines)	3 - 1 = 2	106.298	53.149	104.908	1.52E-09
Interaction	(5-1)(3-1) = 8	9.7032	1.2129	2.3941	.0690
Within (Error)	5·3·(2-1)=15	7.5994	0.5066		
Total	3·5·2 -1 = 29	131.0720			

Features of Two-Way ANOVA F-test

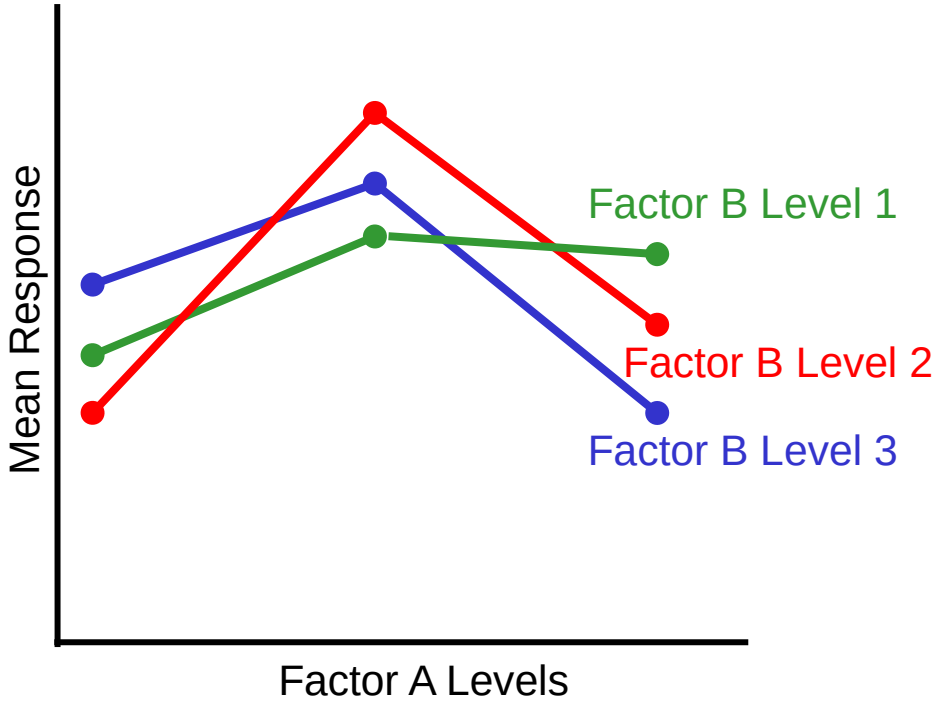
- Degrees of freedom always add up
  - n-1 = rc(n’-1) + (r-1) + (c-1) + (r-1)(c-1)
  - Total = error + factor A + factor B + interaction
- The denominator of the F-test is always the same but the numerator is different
- The sums of squares always add up
  - SST = SSE + SSA + SSB + SSAB
  - Total = error + factor A + factor B + interaction

Examples: Interaction vs. No Interaction

- No interaction:



- Interaction is present:



Chapter Summary

- Described one-way analysis of variance
  - The logic of ANOVA
  - ANOVA assumptions
  - F test for difference in c means
  - The Tukey-Kramer procedure for multiple comparisons
- Described two-way analysis of variance
  - Examined effects of multiple factors
  - Examined interaction between factors

Lab exercises

- Two way ANOVA:
- A consumer research firm wants to compare three brands of radial tires (X, Y, and Z) in terms of tread life over different road surfaces. Random samples of four tires of each brand are selected for each of three surfaces (asphalt, concrete, gravel). A machine that can simulate road conditions for each of the road surfaces is used to find the tread life (in thousands of miles) of each tire. Construct an ANOVA table and conduct F-tests for the presence of nonzero brand effects, road surface effects, and interaction effects.

Surface/ Brand	X	Y	Z
Asphalt	36, 39, 39, 38	42, 40, 39, 42	32, 36, 35, 34
Concrete	38, 40, 41, 40	42, 45, 48, 47	37, 33, 33, 34
Gravel	34, 32, 34, 35	34, 34, 30, 31	36, 35, 35, 33

	sum_sq	df	F	PR(>F)
C(surface)	241.722222	2.0	40.663551	7.152528e-09
C(manufacture)	155.388889	2.0	26.140187	4.838091e-07
C(surface):C(manufacture)	195.611111	4.0	16.453271	6.093609e-07
Residual	80.250000	27.0	NaN	NaN

$(1) = (\Sigma\Sigma\Sigma y_{ijk})^2/n = 253^2/36 = 1778.03$
$(2) = \Sigma\Sigma\Sigma y_{ijk}^2 = 6^2 + 9^2 + 12^2 + \ldots + 3^2 = 2451$
$(3) = \Sigma\, T_{Aj}^2/n_{Aj} = 92^2/12 + 118^2/12 + 43^2/12 = 2019.75$
$(4) = \Sigma\, T_{Bk}^2/n_{Bk} = 86^2/12 + 114^2/12 + 53^2/12 = 1933.42$
$(5) = \Sigma\Sigma\, T_{AjBk}^2/n_{AjBk} = 32^2/4 + 39^2/4 + \ldots + 19^2/4 = 2370.75$

SS Total = (2) - (1) = 2451 - 1778.03 = 672.97
SS Rows = (3) - (1) = 2019.75 - 1778.03 = 241.72
SS Columns = (4) - (1) = 1933.42 - 1778.03 = 155.39
SS Interaction = (5) + (1) - (3) - (4) =
2370.75 + 1778.03 - 2019.75 - 1933.42 = 195.61
SS Main = SS Rows + SS Columns = 397.11
SS Cells = (5) - (1) = 592.72
SS Error = (2) - (5) = 80.25

Python codes

- Python – one-way ANOVA:

  One_Way_Python_ANOVA.ipynb

  https://indico.ifj.edu.pl/event/271/attachments/1139/1679/One_Way_Python_ANOVA.ipynb

- Python examples (two-way ANOVA):
  - Two_Way_ANOVA_in_Python_Tutorial.ipynb
  - Two_way_ANOVA_statsmodel_tyres.ipynb
  - tyres2.csv
  - https://indico.ifj.edu.pl/event/271/attachments/1139/1681/Two_Way_ANOVA_in_Python_Tutorial.ipynb
  - https://indico.ifj.edu.pl/event/271/attachments/1139/1682/Two_way_ANOVA_statsmodel_tyres.ipynb
  - https://indico.ifj.edu.pl/event/271/attachments/1139/1683/tyres2.csv

- **EXERCISE:**

  Please run the codes for 2-way ANOVA and check, whether we calculated correctly our TYRES ANOVA during our LAB