Lecture 12:
One-Way Analysis of Variance (Anova) & Rank Correlation

Applied Statistics - STAN - 5.37 & 5.38
19 & 21 January 2021
Lecturer: Erika Siregar, SST, MS

ANOVA

About Anova

- Previous chapter: Comparing the means from two independent samples
- How about: comparing three or more population means? → **ANOVA**
- Anova → a hypothesis test to determine if the means of several population are the same or not.
- Anova works by **analyzing sample variances based on one treatment (or factor)**, which is a characteristic that allows us to distinguish the different populations from one another → karakter pembeda kelompok populasi/sampel
  - **One-way → based on 1 characteristic/variable**
  - Two-way → based on 2 characteristics/variables → not covered in this course.
  - Manova → >2 characteristics/variable → not covered in this course.
- What is characteristics?
  - The **distinguisher**
  - Anova → works with > 1 sample group
  - There must be a certain **characteristics** that make each group is different from another.

Group (Party ID)	Political Ideology							*n*	Mean	SD
	1	2	3	4	5	6	7			
Democrat	9	20	17	36	4	5	0	91	3.23	1.28
Independent	7	11	17	48	12	11	5	111	3.90	1.43
Republican	0	2	7	23	23	17	2	74	4.70	1.10

	fertilizer	weight
1	None	55
2	None	45
3	None	46
4	Biological	64
5	Biological	52
6	Biological	42
7	Chemical	65
8	Chemical	51
9	Chemical	66
10	Chemical	55

	A	B	C
1	economics	medicine	history
2	42	69	35
3	53	54	40
4	49	58	53
5	53	64	42
6	43	64	50
7	44	55	39
8	45	56	55
9	52		39
10	54		40
11			

How to do Anova Test?

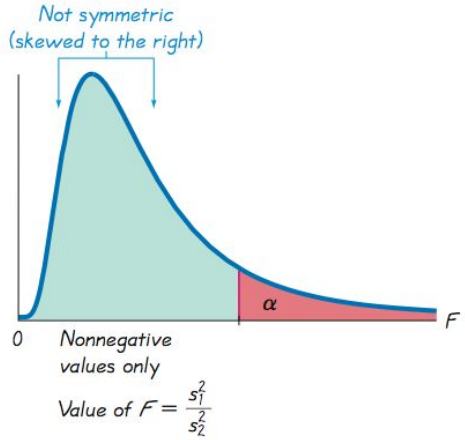
- Same steps as the other hypothesis tests.
- Hypothesis:
  H0: $\mu_1 = \mu_2 = \mu_3 = \ldots = \mu_k$.
  H1: At least one mean is different → **right-tailed test**
- Test statistics
  - Anova → comparing variance between samples and variance within samples
  - Comparing 2 variances == F distribution → refer back to slide lecture 09 (p.22).

$$F = \frac{\text{variance between samples}}{\text{variance within samples}} = \frac{\left[\frac{\Sigma n_i(\bar{x}_i - \bar{\bar{x}})^2}{k-1}\right]}{\left[\frac{\Sigma(n_i - 1)s_i^2}{\Sigma(n_i - 1)}\right]}$$

Sum of square for treatment → $SS_{treatment}$

Sum of square for error → SS_{error}

Df??

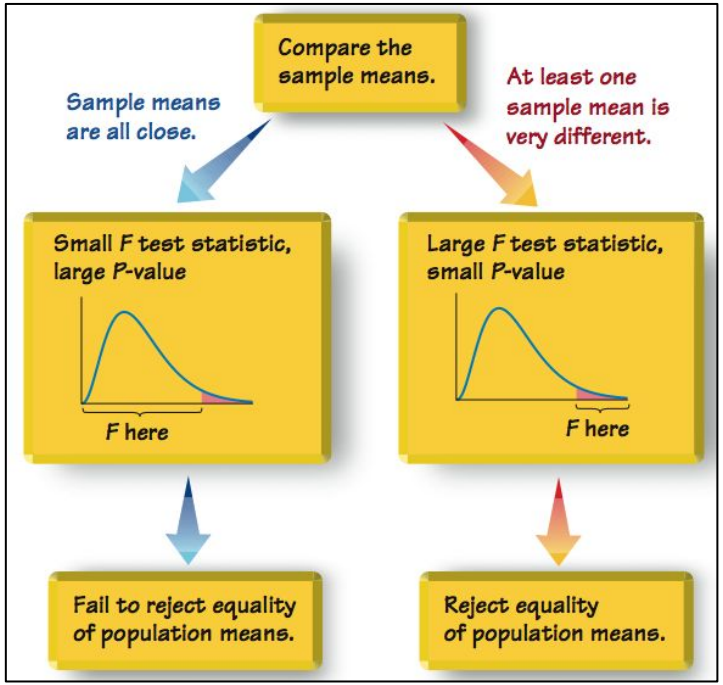


- Compute the p-value
- Decision:
  - Reject H0 → p-value ... α
  - Reject H0 → test statistics … critical value

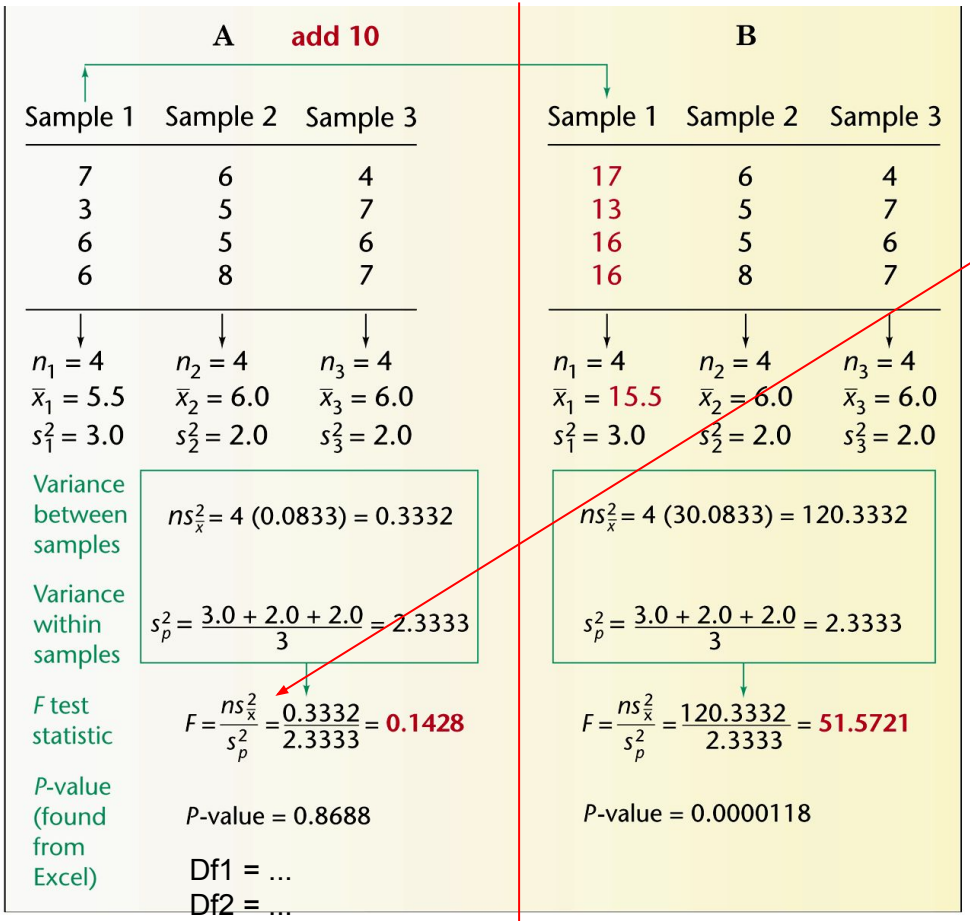
Why is Anova right-tailed?

$$F = \frac{\text{variance between samples}}{\text{variance within samples}}$$

- Goals: membuktikan bahwa semua μ sama.
- Jika semua μ sama, maka variance between samples akan semakin [besar/kecil]?
- Semakin berbeda nilai **μ** dari masing-masing kelompok → **variance** between samples akan semakin [besar/kecil]? → berakibat pada nilai **test statistics F** akan semakin [besar/kecil]?
- But how big is big enough?
  - Tidak semua nilai F besar akan otomatis kita tolak → mesti membuat **standar penolakan** → **tetapkan α**
  - Tolak H0 jika F terlalu besar sehingga jatuh di wilayah yang tidak bisa kita tolerir lagi → wilayah α (critical region) → So, it's intuitive to think that Anova must be a **right-tailed test.**
- Karena right-tailed → >> test statistics F → p-value akan semakin [besar/kecil]?
- α =5% in right-tailed test means we'll still **tolerate** anything that falls in the area between **0 - 0.95** and **reject** anything falls **after 0.95**. → critical region is on the right side.



Example 1 (Manual Way)



$$F = \frac{\text{variance between samples}}{\text{variance within samples}} = \frac{\left[\frac{\Sigma n_i(\bar{x}_i - \bar{\bar{x}})^2}{k-1}\right]}{\left[\frac{\Sigma(n_i - 1)s_i^2}{\Sigma(n_i - 1)}\right]}$$

For cases with unequal sample sizes → calculation is complicated → better use technology.

Example 2 (Using Technology)

Use the chest deceleration measurements listed in Table 12-1 and a significance level of 0.05 to test the claim that the three samples come from populations with means that are all equal.

Table 12-1 Chest Deceleration Measurements (in g) from Car Crash Tests

Small Cars	44	43	44	54	38	43	42	45	44	50	$\rightarrow \bar{x} = 44.7$ g
Medium Cars	41	49	43	41	47	42	37	43	44	34	$\rightarrow \bar{x} = 42.1$ g
Large Cars	32	37	38	45	37	33	38	45	43	42	$\rightarrow \bar{x} = 39.0$ g

Warning
Dengan menggunakan technology (e.g. R) tabel ini harus di reformat menjadi format panjang ke bawah

H0: μ1 = μ2 = μ3 = . . . = μk.
H1: At least one mean is different

```
> aov(value ~ gathercols, data = chest2)
Call:
   aov(formula = value ~ gathercols, data = chest2)

Terms:
                gathercols Residuals
Sum of Squares     162.8667  537.0000
Deg. of Freedom           2        27

Residual standard error: 4.459696
Estimated effects may be unbalanced
> chest_aov <- aov(value ~ gathercols, data = chest2)
> summary(chest_aov)
            Df Sum Sq Mean Sq F value Pr(>F)
gathercols   2  162.9   81.43   4.094 0.028 *
Residuals   27  537.0   19.89
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
```

- P-value of 0.028 < 0.05, we **REJECT the null hypothesis** of equal means.
- There is sufficient evidence to do not support the claim that the **three samples come from populations with means that are all equal.**
- Interpretation? → clue: check the data summary (use R, if necessary).

Dissecting F formula

$SS_{treatment} + SS_{error} = SS_{total}$

Sum of square for treatment → $SS_{treatment}$

$$F = \frac{\text{variance within samples}}{\text{variance between samples}} = \frac{\left[\frac{\Sigma n_i(\bar{x}_i - \bar{\bar{x}})^2}{k-1}\right]}{\left[\frac{\Sigma(n_i - 1)s_i^2}{\Sigma(n_i - 1)}\right]}$$

Sum of square for error → SS_{error}

Formula 12-8

$$F = \frac{MS\ (\text{treatment})}{MS\ (\text{error})}$$

***MS*(treatment) is a mean square for treatment, obtained as follows:**

Formula 12-5

$$MS(\text{treatment}) = \frac{SS\ (\text{treatment})}{k-1}$$

***MS*(error) is a mean square for error, obtained as follows:**

Formula 12-6

$$MS(\text{error}) = \frac{SS\ (\text{error})}{N-k}$$

***MS*(total) is a mean square for the total variation, obtained as follows:**

Formula 12-7

$$MS(\text{total}) = \frac{SS(\text{total})}{N-1}$$

Identifying Means That Are Different

Informal methods for comparing means

1. Use the same scale for constructing **boxplots** of the data sets to see if one or more of the data sets are very different from the others.
2. Construct confidence interval estimates of the means from the data sets, then compare those confidence intervals to **see if one or more of them do not overlap with the others**.

Rank Correlation

Rank correlation test

- The **rank correlation test** (or **Spearman's** rank correlation test) is a non-parametric test that **uses ranks of sample data** consisting of matched pairs.
  - It is used to test for an **association between two variables**.
  - Data values must be converted into ranks → **Bedanya dengan correlation biasa?**

Table 10-2 Calculating r

x (Shoe Print)	y (Height)
29.7	175.3
29.7	177.8
31.4	185.4
31.8	175.3
27.6	172.7
$\Sigma x = 150.2$	$\Sigma y = 886.5$

Konversi jadi rank

Overall Quality	8	2	1	7	5	4	3	6
Selectivity Rank	2	6	8	1	4	3	7	5

shoe					
height					

- Rank correlation can be used to **detect some (not all) relationships that are not linear**.
- How to test? → **Hypothesis Test**

Hypothesis Test for Rank Correlation

Marks in Commerce (X)	Rank (R_{1i})	Marks in Mathematics (Y)	Rank (R_{2i})
15	2	40	6
20	3.5	30	4
28	5	50	7
12	1	30	4
40	6	20	2
60	7	10	1
20	3.5	30	4
80	8	60	8

Steps

1. Hypothesis:
   H_0: $\rho_s = 0$ (There is no correlation between the two variables.)

   H_1: $\rho_s \neq 0$ (There is a correlation between the two variables.) → Two-tailed test
2. Convert the data into ranks
3. Compute the test statistics

Physics	Rank	Math	Rank
35	3	30	5
23	5	33	3
47	1	45	2
17	6	23	6
10	7	8	8
43	2	49	1
9	8	12	7
6	9	4	9
28	4	31	4

$$r_s = 1 - \frac{6\Sigma d^2}{n(n^2 - 1)}$$

Note:
d = **difference between ranks** for the two values within a pair

or

$$r_s = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{\sqrt{n(\Sigma x^2) - (\Sigma x)^2}\sqrt{n(\Sigma y^2) - (\Sigma y)^2}}$$

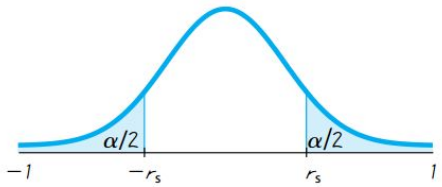
→ If there are ties among the ranks in a variable

Hypothesis Test for Rank Correlation

Steps

4. Critical values
   a. If n ≤ 30, critical values are found in Table A-6 (p.589)
   b. If n > 30, use Formula:

$$r_s = \frac{\pm z}{\sqrt{n - 1}}$$

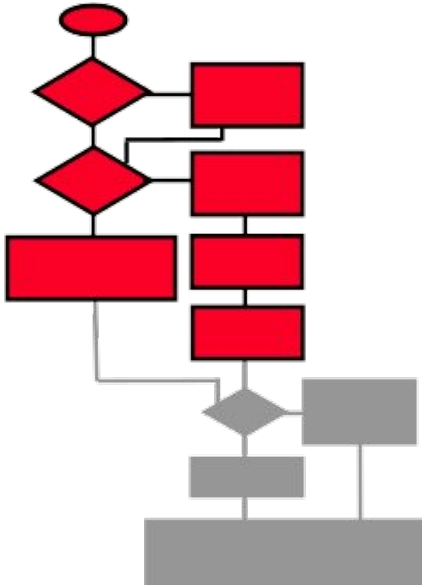


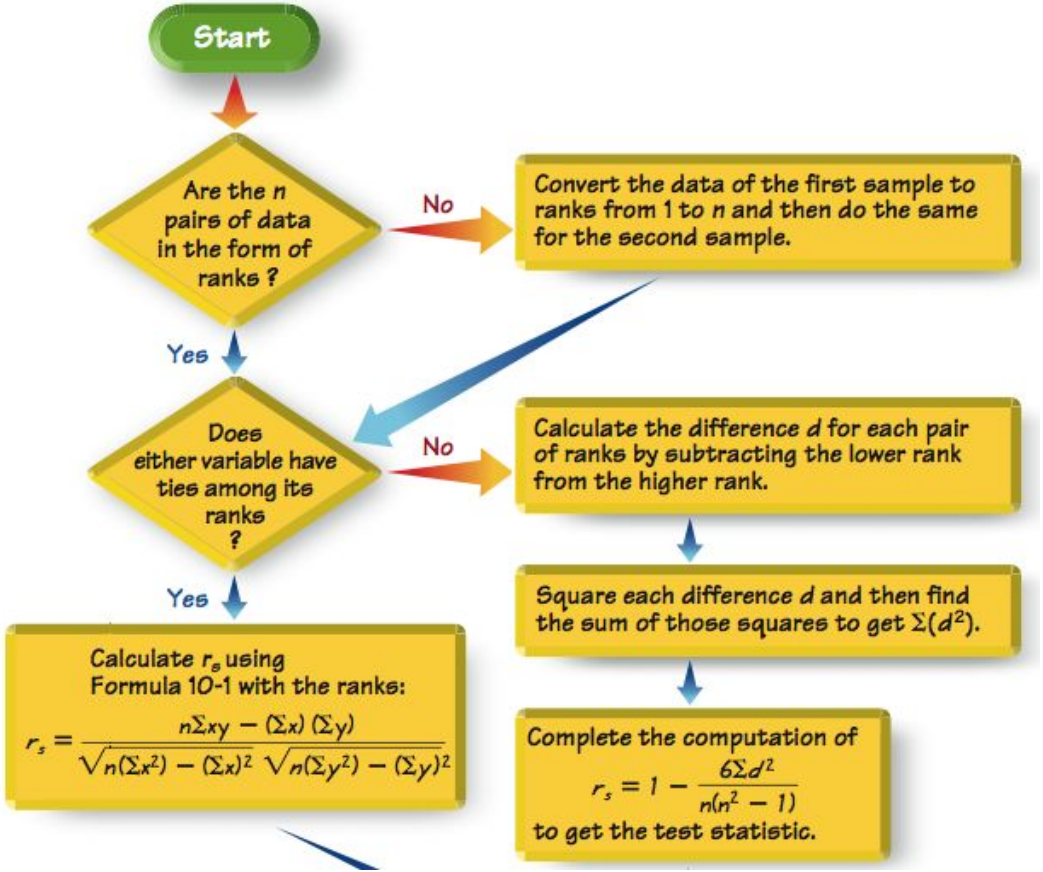
Notes:

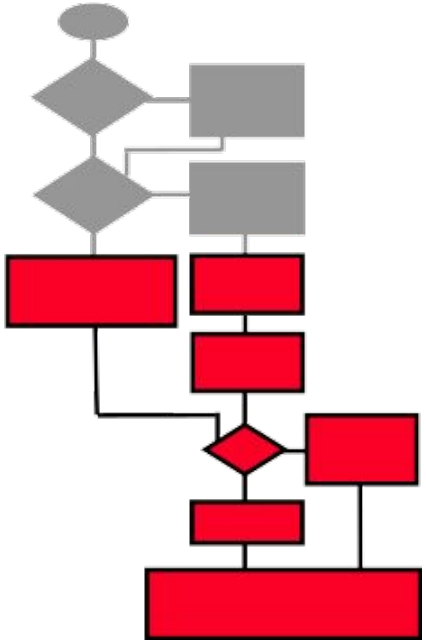
- the value of z corresponds to the significance level. For example, if α = 0.05 → z = qnorm(α/2) = 1.96.

5. Make decision:
   a. Reject H0 if the test statistics does not lie between the critical values interval.
   b. Reject H0 if the p-value < α → using technology

Steps




Start
Are the n pairs of data in the form of ranks ?
No
Convert the data of the first sample to ranks from 1 to n and then do the same for the second sample.
Yes
Does either variable have ties among its ranks ?
No
Calculate the difference d for each pair of ranks by subtracting the lower rank from the higher rank.
Square each difference d and then find the sum of those squares to get Σ(d²).
Yes
Calculate r_s using Formula 10-1 with the ranks:
$r_s = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{\sqrt{n(\Sigma x^2) - (\Sigma x)^2}\sqrt{n(\Sigma y^2) - (\Sigma y)^2}}$
Complete the computation of
$r_s = 1 - \frac{6\Sigma d^2}{n(n^2 - 1)}$
to get the test statistic.



Yes

Calculate r_s using Formula 10-1 with the ranks:

$$r_s = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{\sqrt{n(\Sigma x^2) - (\Sigma x)^2}\sqrt{n(\Sigma y^2) - (\Sigma y)^2}}$$

Square each difference d and then find the sum of those squares to get $\Sigma(d^2)$.

Complete the computation of

$$r_s = 1 - \frac{6\Sigma d^2}{n(n^2 - 1)}$$

to get the test statistic.

Yes

Is $n \leq 30$?

No

Calculate the critical values

$$r_s = \pm\frac{z}{\sqrt{n-1}}$$

where z corresponds to the significance level.

Yes

Find the negative and positive critical values of r_s from Table A-9.

- If r_s is between the negative and positive critical values, fail to reject the null hypothesis $\rho_s = 0$ (no correlation).
- If r_s is not between the negative and positive critical values, reject the null hypothesis $\rho_s = 0$ and conclude that there is sufficient evidence to support a claim of a correlation.

Example (no ties)

Table 13-1 lists **overall quality scores** and **selectivity rankings** of a sample of **national universities** (based on data from U.S. News and World Report). Find the value of the rank correlation coefficient and use it to determine **whether there is a correlation between the overall quality scores and the selectivity rankings**. Use a **0.05 significance level**. Based on the result, does it appear that national universities with higher overall quality scores are more difficult to get into?

Table 13-1 Overall Quality Scores and Selectivity Ranks of National Universities

Overall quality	95	63	55	90	74	70	69	86
Selectivity rank	2	6	8	1	4	3	7	5

Table 13-7 Ranks of Data from Table 13-1

Overall Quality	8	2	1	7	5	4	3	6
Selectivity Rank	2	6	8	1	4	3	7	5
Difference d	6	4	7	6	1	1	4	1
d^2	36	16	49	36	1	1	16	1 → $\Sigma d^2 = 156$

neither variable has ties in the ranks

Answer:

$H_0: \rho_s = 0$ $\quad H_1: \rho_s \neq 0$

$$r_s = 1 - \frac{6\Sigma d^2}{n(n^2 - 1)} = 1 - \frac{6(156)}{8(8^2 - 1)}$$

$$= 1 - \frac{936}{504} = -0.857$$

α =0.05, n = 8 → critical value = ± 0.738

Because the test statistic of **rs = –0.857** is **not between the critical values** of **–0.738 and 0.738**, we **reject** the null hypothesis.

There is **sufficient evidence** to support a claim of a **correlation between overall quality score and selectivity ranking**. It appears that **Universities with higher quality** scores are **more selective** and are more difficult to get into.

Example (with ties)

Below is the data of ranks and costs of LCD TV. Find the value of the rank correlation coefficient to determine if there is a correlation between quality and price. (sig. Level = 0.05)

quality_rank	1	2	3	4	5	6	7	8	9	10
cost	23	50	23	20	32	25	14	16	40	22

Answer:

$H_0: \rho_s = 0 \quad H_1: \rho_s \neq 0$

quality_rank	cost	qrank	cost_rank	xy	x2	y2
1	23	1	5.5	5.5	1	30.25
2	50	2	10	20	4	100
3	23	3	5.5	16.5	9	30.25
4	20	4	3	12	16	9
5	32	5	8	40	25	64
6	25	6	7	42	36	49
7	14	7	1	7	49	1
8	16	8	2	16	64	4
9	40	9	9	81	81	81
10	22	10	4	40	100	16
55	**265**	**55**	**55**	**280**	**385**	**384.5**

$$r_s = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{\sqrt{n(\Sigma x^2) - (\Sigma x)^2}\sqrt{n(\Sigma y^2) - (\Sigma y)^2}}$$

rs = ((10*280) - (55*55))/sqrt((10*385)-3025) * sqrt((10*384.5)-3025) = -0.2735

Exercise (α = 5%)

1. Dengan menggunakan data “IQ and Lead” (iqlead.csv), lakukanlah uji satu arah anova untuk membuktikan bahwa ketiga kelompok ‘blood lead level’ memiliki rata-rata yang sama. Gunakan cara manual dan R. **Apakah bisa menggunakan cara manual?**

2. Dengan menggunakan data english (english.csv), lakukanlah uji korelasi rank dengan cara menghitung manual. Lakukan juga dengan R, dan lihat bagaimana perbedaan hasilnya.

3. Dengan menggunakan data cigarette (cigar.xlsx), lakukanlah uji korelasi rank dengan cara menghitung manual. Lakukan juga dengan R, dan lihat bagaimana perbedaan hasilnya.

4. Dengan menggunakan data tentang daya tahan jam (watch.xlsx), lakukanlah uji satu arah anova untuk membuktikan bahwa ketiga merk jam memiliki rata-rata yang sama. Gunakan cara manual dan R. Apakah bisa menggunakan cara manual?

merk	daya_tahan_jam				
A	250	224	252	230	240
B	251	243	260	253	263
C	253	242	259	252	259

GCR

Tulis hal apa yang kamu belum paham (konsep & teori) dan berharap bisa dibahas di minggu depan.

Thanks!

Any questions?

You can find me at:

@erikaris