Unit 5
ANOVA

ESIGELEC

Instructor: Federico Perea

1

Analysis of Variance (ANOVA)

- In the previous unit, you have studied techniques to compare two groups (paired data).
- ANOVA allows you to compare more than two groups, all of them coming from normal distributions with (possibly) different means and equal variances.
- ANOVA involves methods for testing equality of sample means
- Applications
  - When cars are arranged into different groups according to the type of transmission they have, we can test for equality of their mean fuel consumption.
  - When three different bank offices offer the same products, we can test for a difference in the mean sales.
  - …
- Comparing more than two groups! (If two groups are to be compared, you can use hypothesis tests).
- The populations under study have normal distributions with the same variance.
- The samples are randomly selected from each population and independent of one another.

2

One-way ANOVA

- Also called single-factor ANOVA.
- It allows you to test hypotheses that assert equality of more than two means. You have *I>2* groups coming from different populations. Population i is distributed as a normal distribution $N(\mu_i, \sigma)$. In each sample there are have n_i elements.

Therefore the test is:

$H_o: \mu_1 = \mu_2 = \mu_3 \ldots = \mu_I = \mu$

$H_1: \exists\ i,j: \mu_i \neq \mu_j$

A **factor (or treatment)** is a property or characteristic that allows you to distinguish the different populations from one another. Each treatment has different values or groups. In One-Way ANOVA, only one factor is considered.

3

Formulae and Notation

$$\sum_{ij}(X_{ij} - X..)^2 = \sum_i n_i(X_i. - X..)^2 + \sum_{ij}(X_{ij} - X_i.)^2$$

TSS = FSS + RSS

- TSS= Total Sum of Squares, measures the total variation around X..
- FSS (Factor Sum of Squares) is a measure of the variability among the sample means. Factor = Treatment RSS (residual sum of squares) represents the variability that is assumed to be common to all the populations being considered (the part of the total variability that is not explained by differences between sample means)
- X_{ij}= j-th element of group *i*
- $X_i.$ = average of group *i*
- X.. = average of all data

4

F-test

- If you divide TSS, FSS and RSS by their corresponding number of **degrees of freedom**, you get mean squares, TMS, FMS and RMS, respectively.

$$F = \frac{FMS}{RMS} = \frac{FSS/(I-1)}{RSS/(N-I)} \sim F_{I-1,N-I}$$
$$F > F^{\alpha}_{I-1,N-I} \Leftrightarrow Reject\ the\ null\ hypothesis$$

- Alternatively, if p-value < significance level, reject H_o .
  p-value is given by most statistical software, and defined as:

$$p-value = P[F_{I-1,I(J-1)} \geq F]$$

5

ANOVA Table

Source	SS	Df	MS	F	p
Factor	FSS	I-1	FSS/(I-1)	FMS/RMS	P-value
Residual	RSS	I(J-1)	RSS/I(J-1)		
Total	TSS	IJ-1	TMS/(IJ-1)		

The ANOVA table in R looks a bit different!

6

Example

The same product can be obtained form 4 different resources. In the following table, the price in the market of five different units made out of the four different resources is shown. Do units produced with different resources have significantly different prices?

Resource 1	Resource 2	Resource 3	Resource 4
6.0	6.2	5.9	5.0
6.2	6.1	6.0	5.1
6.5	5.9	6.0	4.2
6.8	6.0	6.2	4.6
6.0	6.0	5.8	4.5

7

ANOVA table

In this case we have I = 4 (four groups for the factor) and J=5 (5 elements in each group). The ANOVA table results in:

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Resource_Factor	3	7.922	2.641	36.17	2.33e-07
Residuals	16	1.168	0.073		

Since p-value=2.33e-07 < 0.05, there is enough evidence to reject the null hypothesis (mean equality), and we therefore state that the different resources yield different market price for the studied product. In R:
summary(aov(lm(Price~Resource_Factor)))

8

ANOVA TABLE IN R

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Resource_Factor	3	7.922	2.641	36.17	2.33e-07
Residuals	16	1.168	0.073		

Source	SS	Df	MS	F	p
Factor	7.922	3	2.64067	36.17	0.000
Residual	1.168	16	0.073		
Total	9.09	19			

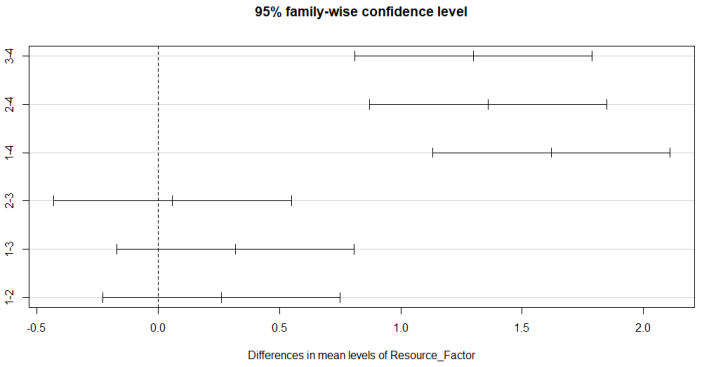
9

Which levels have different means?

- The previous ANOVA table helps you decide if the groups have different means or not.
- But, if they have different means, which of the groups do actually have different means?
- Tukey HSD intervals give confidence intervals for the difference in means between any pair of groups ($\mu_1 - \mu_2$, $\mu_1 - \mu_3$, …)
- If 0 is in the Tukey HSD interval of groups i_1 and i_2, then you can consider that $\mu_{i_1} = \mu_{i_2}$. Otherwise, these two means can be considered different.

10

Tukey HSD in the example



Group 4 is different from the other three. Groups 1,2,3 can be considered equal.

Example

An electronic company is considering the purchase of a machine to reduce CO2 emissions (in tons per working day) in their factory. Three different machines (*A,B,C)* are used in a sample of 12 working days. In four of them machine *A* was installed, five used machine *B*, and 3 machine *C*. The CO2 emissions of each sample is:

A	B	C
23	27	24
23	29	26
20	25	24
21	23	
	24	

At the 0.05 significance level, test the claim that the mean CO2 emission values are the same for all three plants. Complete the ANOVA able, and draw a conclusion!

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	34,3	??	??	??	??	??
Within Groups	??	??	??			
Total	66,9166667	??				

Two-factor ANOVA

- We now see what to do when more than one factor (treatment) is considered
- Other than the same questions as for the single-factor case, we now wonder whether the way in which one factor affects the dependent variable varies with the values of the other factor: **interaction**
- Example: the graphic card and the speed of the processor are suspected to affect the performance of certain computers. For this aim, 4 different speeds (0.75, 1, 1.25, 1.5) and three different cards (A,B,C) were combined and tested on 36 computers (randomly chosen) in groups of 3. The results obtained are shown in the table in the next slide. Study how the two factors affect the performance of the computers.

13

Example

	Card A	Card B	Card C
Speed 0.75	68	52	66
	60	55	72
	62	61	68
Speed 1	91	62	83
	75	67	82
	86	60	78
Speed 1.25	90	64	72
	98	75	66
	94	74	74
Speed 1.5	105	68	61
	95	85	58
	99	83	58

14

ANOVA table in the example

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
CARD_FACTOR	2	2287.2	1143.6	42.93	1.18e-08
SPEED_FACTOR	3	1613.6	537.9	20.19	9.44e-07
CARD_FACTOR:SPEED_FACTOR	6	2284.6	380.8	14.29	6.93e-07
Residuals	24	639.3	26.6		

The p-value of factor Speed is less than 0.05, then this factor is significant: the average performance is not the same for different values of Speed

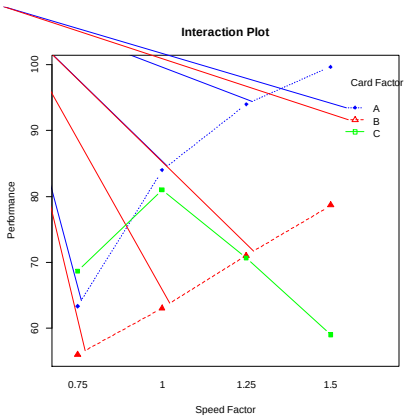
The p-value of factor Card is less than 0.05, then this factor is significant: the average performance is not the same for different values of Card

The p-value of the interaction is less than 0.05, then the interaction is significant: the way in which the factor Card affects the average performance is different for different values of factor Speed (and viceversa)

15

Interaction in the example: table of means and means plot

Speed	Card A	Card B	Card C
0,75	63,3333333	56	68,6666667
1	84	63	81
1,25	94	71	70,6666667
1,5	99,6666667	78,6666667	59



In the table you see, for each value of Speed, the average performance for each value of Card. *aggregate()*

The more parallel the lines are in the means plot, the less interaction. *interaction.plot()*

In this case, lines cross each other (specially the averages with Card C with the other two), this is a graphical indication of interaction between factors.

16

Exercise

The resistance of certain electronic system is suspected to depend on two of its components: A and B. Several values for these two components have been tested, and the resistances obtained are given in the following table. Study the effect of components A and B over the resistance as well as the interaction between them. Interpret the results.

Resistance	A	B
6045	0	1
6030	0	1
6090	5	1
6095	5	1
6110	10	1
6120	10	1
5425	0	2
5455	0	2
5490	5	2
5500	5	2
5530	10	2
5535	10	2
4460	0	3
4490	0	3
4520	5	3
4525	5	3
4550	10	3
4560	10	3

17