Vall d'Hebron Institut de Recerca

Ricardo Gonzalo
ricardo.gonzalo@vhir.org

13/02/2019

Module 2

Session 3.1: Principles of Experimental Design. Analysis of Variance

Table of contents

Table of contents

Experimental design should be a mandatory step in every experiment

- Experimental design is a structured, organized method for **determining the relationship** between the different **factors** affecting an **experimental** process, and the **output** of that process.

Sir Ronald A. Fisher

Father of modern Mathematical Statistics and Developer of Experimental Design and ANOVA

"To consult the statistician after an experiment is finished, is often merely to ask him to conduct a post mortem examination. He can perhaps say what the experiment died of. "

Types of variability that play role in an experiment:

- **Planned variability**: These are the differences in response between treatments applied.

- **Noise variability:** Differences between two consecutives measures. We cannot avoid that.

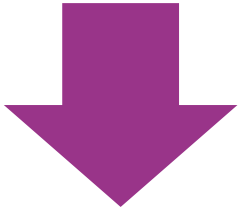
- **Variability not planned:** Produce a systematic variation in the results. A priori the reason is not known.

A good experimental design should…

- Avoid systematic error: e.g. samples from one group processed with instrument A and samples from the other group processed with instrument B.
- Be precise: try to maintain the random error as low as possible
- Allow estimation of error: enough replicates in each treatment
- Have broad validity: our experimental units should reflect the population about which we wish to draw inference

A good experimental design should…

- Avoid systematic error: e.g. samples from one group processed with instrument A and samples from the other group processed with instrument B.
- Be precise: try to maintain the random error as low as possible
- Allow estimation of error: enough replicates in each treatment
- Have broad validity: our experimental units should reflect the population about which we wish to draw inference



Randomization **Replication** **Local Control**

Table of contents

Randomization

- The method for assigning treatments to units involves *randomization ("any individual experimental subject has the same chance as any other individual of finding itself in each experimental group")*
- It is one of the most important elements of a well-designed experiment
- Made valid most of the statistical analysis usually performed

e.g. 4 treatments to be assigned to 16 units

1. Use sixteen identical slips of paper, 4 marked with A, 4 with B, and so on to D. Put the slips of paper into a basket and mix them thoroughly. For each unit, we draw a slip of paper from the basket and use the treatment marked on the slip

2. Treatment A is assigned to the first four units we have encounter, treatment B to next four units, and so on.

Saying “randomly assign…” is sometimes easier to say than to do, especially in complex designs.

Some tools may help

- Research Randomizer
  http://www.randomizer.org/

- Interactive Statistical Calculation pages
  http://statpages.org/ (look por “Experimental design”)

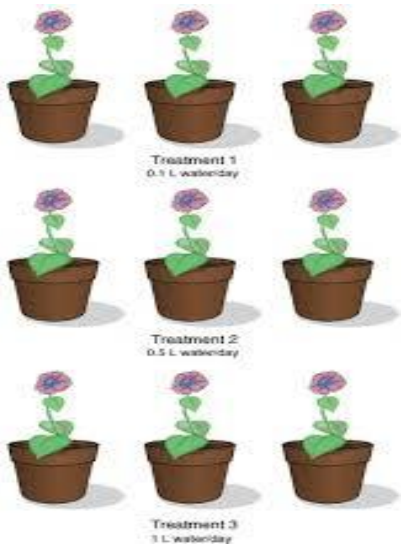
Replication

- It is the basis of all experimental design
- It is the repetition of the basic experiment with another experimental units

How many replicates I need?

the more replicates we have, the more confident we can be that differences between groups are real and not simply due to chance effects

More replicates increase in time/money cost



- To be in mind:
    1. To Know the variability of the technology used
    2. Previous works with similar technology
    3. Directly correlated with the precision of the experiment

Local control

- When the experimental units are not homogeneous or the process to analyze them neither are (Kits lot numbers, batch,…)

We are not interested in to find out the differences between the levels of the blocks

Effect of treatment A and B?

Sample	Treatment	Sex	Batch
1	A	Male	1
2	A	Male	1
3	A	Male	1
4	A	Male	1
5	B	Female	2
6	B	Female	2
7	B	Female	2
8	B	Female	2

Sample	Treatment	Sex	Batch
1	A	Male	1
2	A	Female	2
3	A	Male	2
4	A	Female	1
5	B	Male	1
6	B	Female	2
7	B	Male	2
8	B	Female	1

Effect of treatment A and B?

Sample	Treatment	Sex	Batch
1	A	Male	1
2	A	Male	1
3	A	Male	1
4	A	Male	1
5	B	Female	2
6	B	Female	2
7	B	Female	2
8	B	Female	2

Treatment is confounded between sex and batch

Sample	Treatment	Sex	Batch
1	A	Male	1
2	A	Female	2
3	A	Male	2
4	A	Female	1
5	B	Male	1
6	B	Female	2
7	B	Male	2
8	B	Female	1

Treatment is well balanced

Local control or randomize?

- Local control assure you that differences are not due to blocks in the sample
- Local control eliminate the noise due to differences among blocks
- Randomization is good for balance effects from variables not taken into account from the beginning.

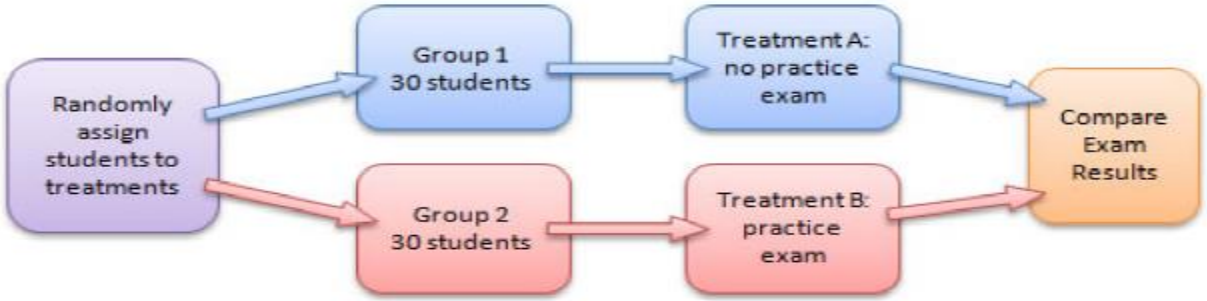
"Block what you can, randomize what you cannot" (George Box, 1978)

Table of contents

COMPLETE RANDOMIZED DESIGN (CRD)

- Simplest of all designs
- Uses randomization and replication.

✓Treatments are allocated at random to experimental units



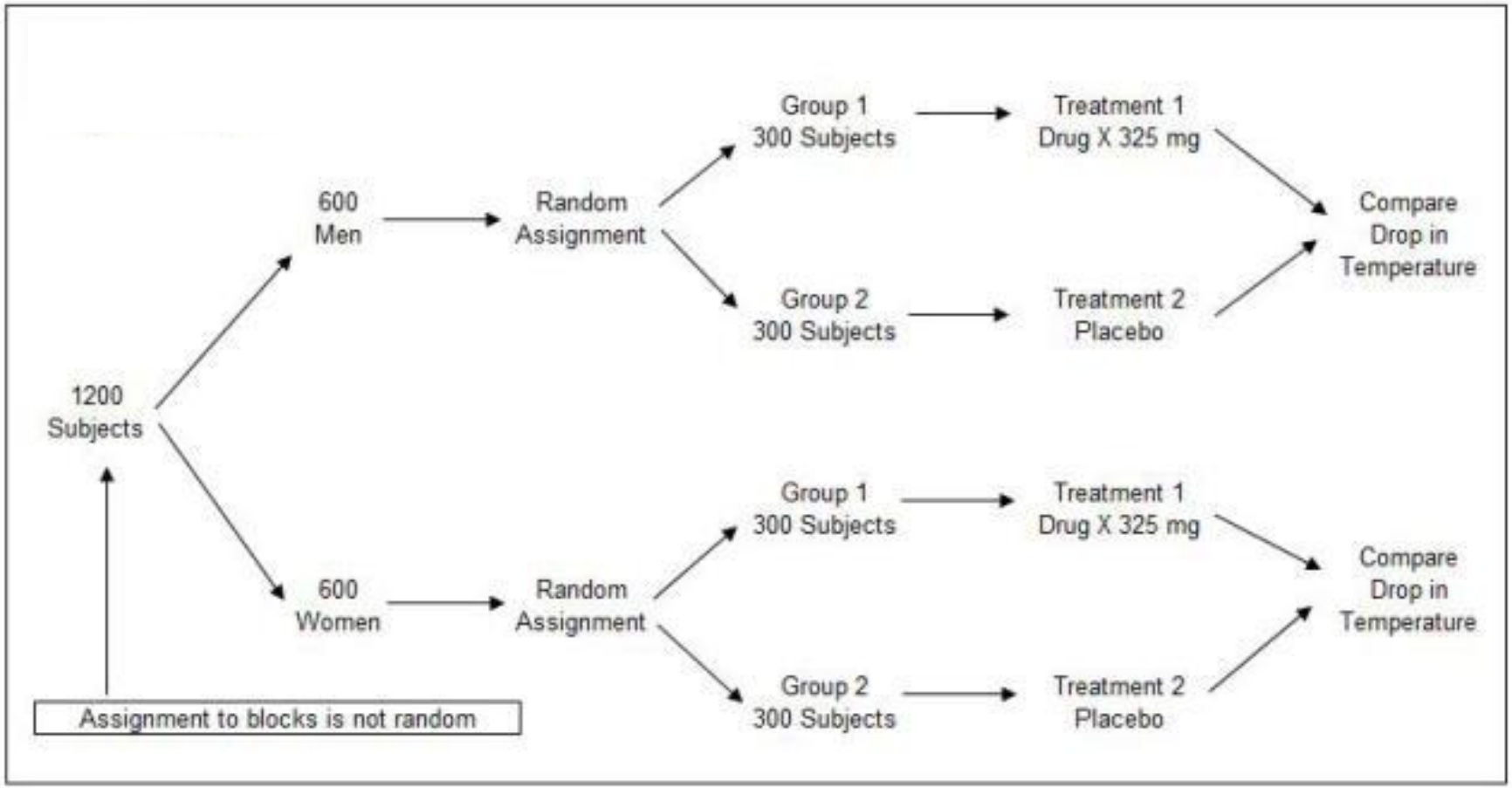
Response = mean + treatment effect + error

RANDOMIZED BLOCK DESIGN (RBD)

- Blocking may be used to divide the whole experimental material into homogeneous strata or sub-groups known as blocks.
- Blocking to "remove" the effect of nuisance factors

Then the experimental units are randomly assigned treatments.

Response = mean + treatment effect + block effect + error


1200
Subjects
600
Men
Random
Assignment
Group 1
300 Subjects
Treatment 1
Drug X 325 mg
Group 2
300 Subjects
Treatment 2
Placebo
Compare
Drop in
Temperature
600
Women
Random
Assignment
Group 1
300 Subjects
Treatment 1
Drug X 325 mg
Group 2
300 Subjects
Treatment 2
Placebo
Compare
Drop in
Temperature
Assignment to blocks is not random

FACTORIAL DESIGN

- Effects of several factors of variation are studied simultaneously.
- The treatments are all the combinations of different factors under study.
- The effects of each of the factors and the *interaction* effects, which are the variations in the effect of one factor as a result to different levels of other factors, are studied.

Response = mean+ treat-1 effect + treat-2 effect +treat-1:2-Interaction + error

The simplest factorial experiment contains two levels for each of two factors.

Study the effects of **two drugs** at **two** different administration **doses**

200 mg/kg

400 mg/kg

200 mg/kg

400 mg/kg

REPEATED MEASURES

When **more than one measure** is taken **on each experimental unit** one has a within subjects design:

- Measures on the same individual are correlated
- New source of variability to be accounted for:
  variability within subjects
- Same possibilities as with other designs
- Problems:
  - some patients can pull out before complete the whole study
  - Individuals become better/worse at a task over time

Repeated measures Example:

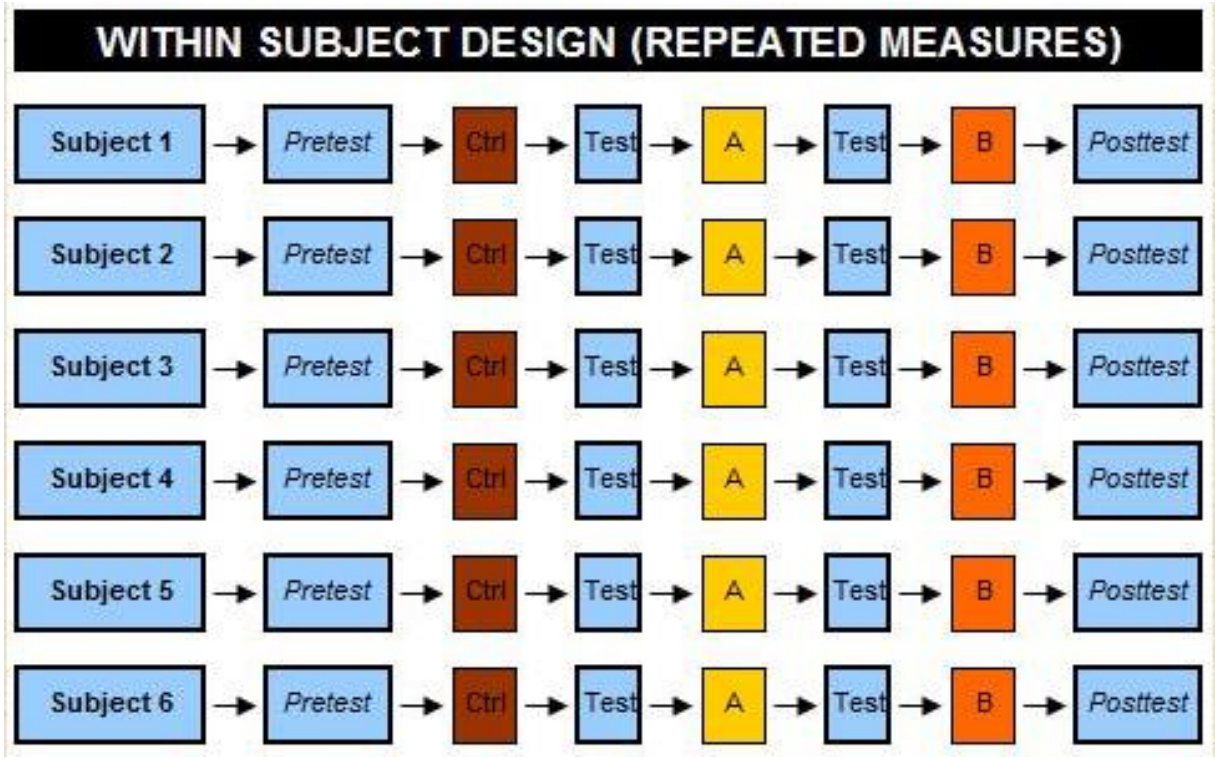


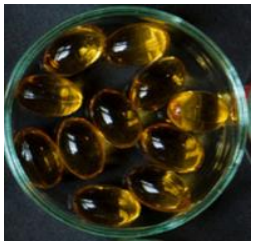
Table of contents

Until now, we were comparing two samples, but what happens it we have more to compare? Could we still use Student's t test?

A pharmaceutical laboratory wants to test which of three drugs are the best:

drug 1

drug 2



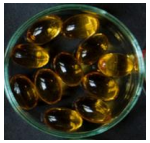
drug 3

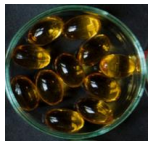
To know which of the drugs is the best one, one could think to perform the following comparison using a t test:

Comparisons with t test would be:

Chance of Type I error

 vs  → $\alpha = 5\%$

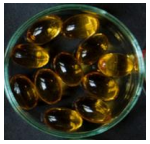
 vs  → $\alpha = 5\%$

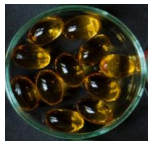
 vs  → $\alpha = 5\%$

Comparisons with t test would be:

Chance of Type I error

 vs  → $\alpha = 5\%$

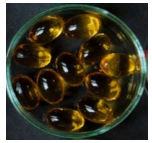
 vs  → $\alpha = 10\%$

 vs  → $\alpha = 15\%$

Comparisons with t test would be:

Chance of Type I error

 vs  → $\alpha = 5\%$

 vs  → $\alpha \approx 10\%$

Would be easier to accept the null hypothesis when it was wrong (more false positives)

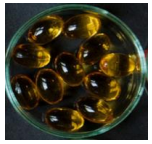
 vs  → $\alpha \approx 15\%$

Table of contents

ANOVA tests the following hypotheses:

From linear model of ANOVA: $Y_{ij} = \mu + \alpha + \epsilon_{ij}$

H_0:

The means of all the groups are equal:

$$H_0: \quad \mu_1 = \mu_2 = \cdots = \mu_a$$

No effects among the different treatments:

$$H_0: \quad \alpha_1 = \alpha_2 = \cdots = \alpha_a = 0$$

H_1: Not all the means are equal

ANOVA tests the following hypotheses:

H_1: Not all the means are equal

If we reject H_0 we only could say that there is a difference among the groups, but we couldn't say between which of them

Can follow up with "**multiple comparisons**"

Example. Complete Randomized Design

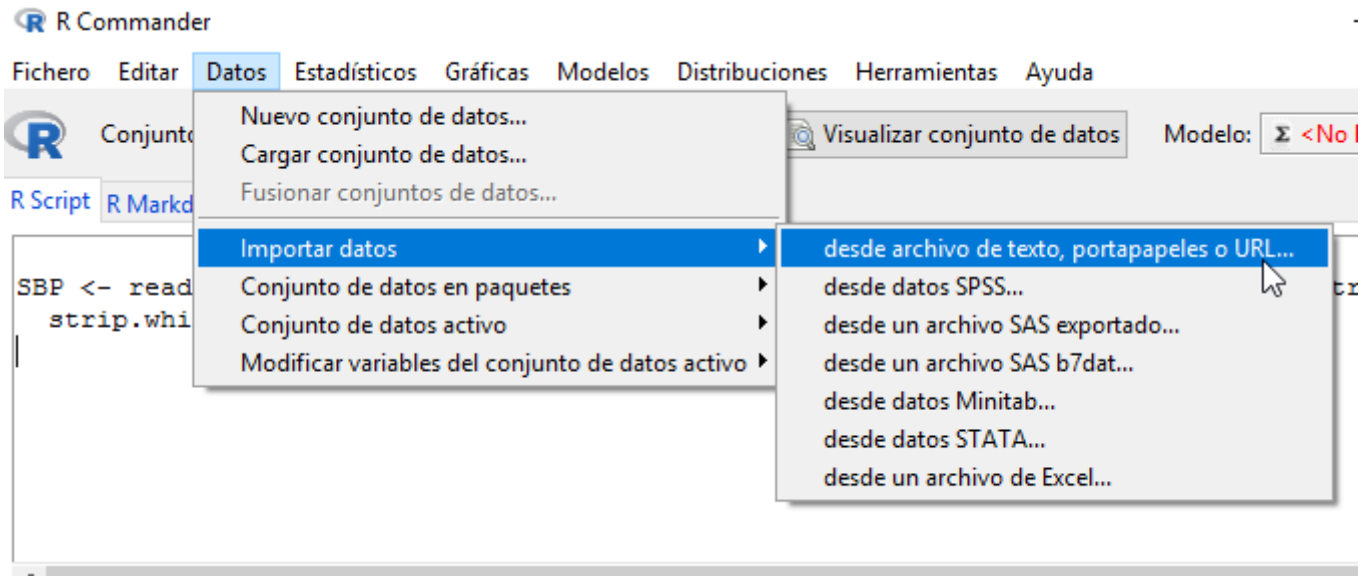
Which of the three drugs is the best to reduce the SBP? Three drugs are randomly assigned to 24 patients (same characteristics), and SBP is monitored after a month.

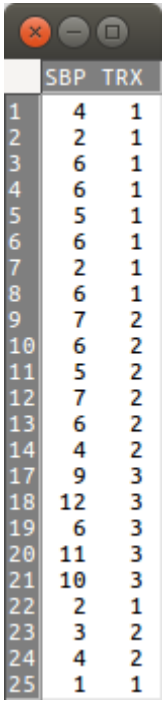
Sample	Drug1	Drug2	Drug3
1	4	7	9
2	2	6	12
3	6	5	6
4	6	7	11
5	5	6	10
6	6	4	11
7	2	7	9
8	6	5	10

→ One factor

Three levels

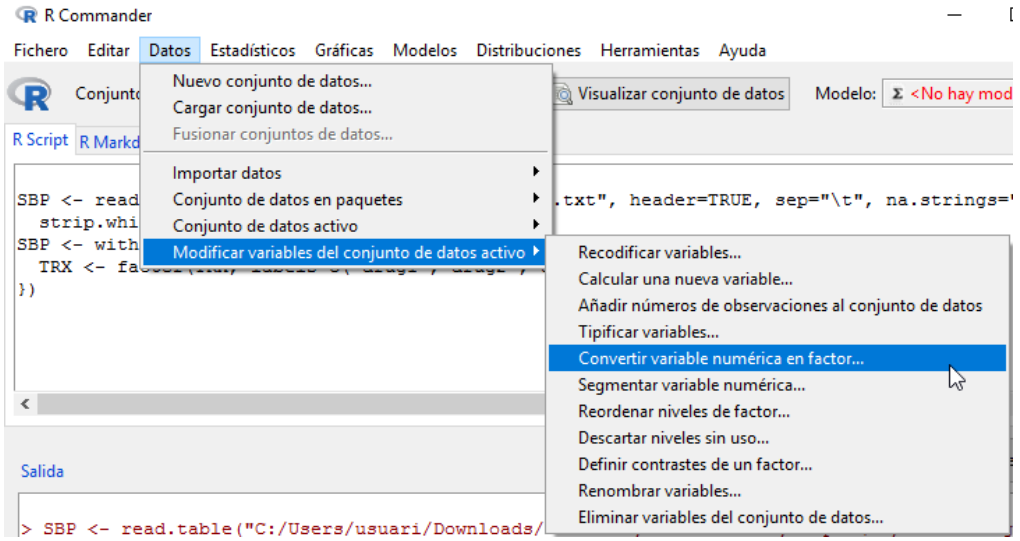
Always is good to “look” graphically the data. Let’s try with R-commander:

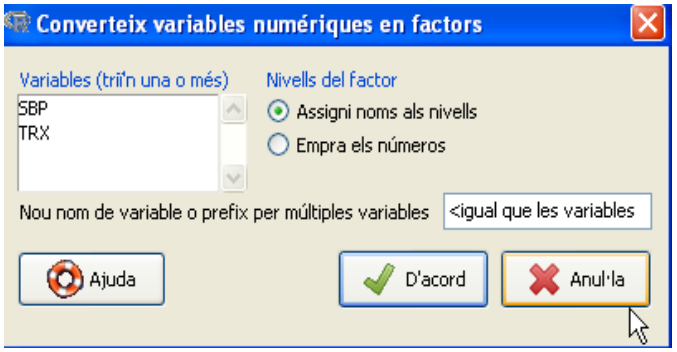


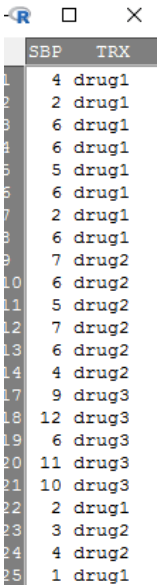


	SBP	TRX
1	4	1
2	2	1
3	6	1
4	6	1
5	5	1
6	6	1
7	2	1
8	6	1
9	7	2
10	6	2
11	5	2
12	7	2
13	6	2
14	4	2
17	9	3
18	12	3
19	6	3
20	11	3
21	10	3
22	2	1
23	3	2
24	4	2
25	1	1

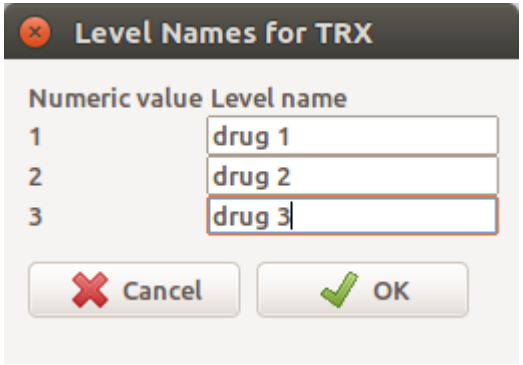
Recode Trx (numeric) variable to a factor:







	SBP	TRX
1	4	drug1
2	2	drug1
3	6	drug1
4	6	drug1
5	5	drug1
6	6	drug1
7	2	drug1
8	6	drug1
9	7	drug2
10	6	drug2
11	5	drug2
12	7	drug2
13	6	drug2
14	4	drug2
17	9	drug3
18	12	drug3
19	6	drug3
20	11	drug3
21	10	drug3
22	2	drug1
23	3	drug2
24	4	drug2
25	1	drug1



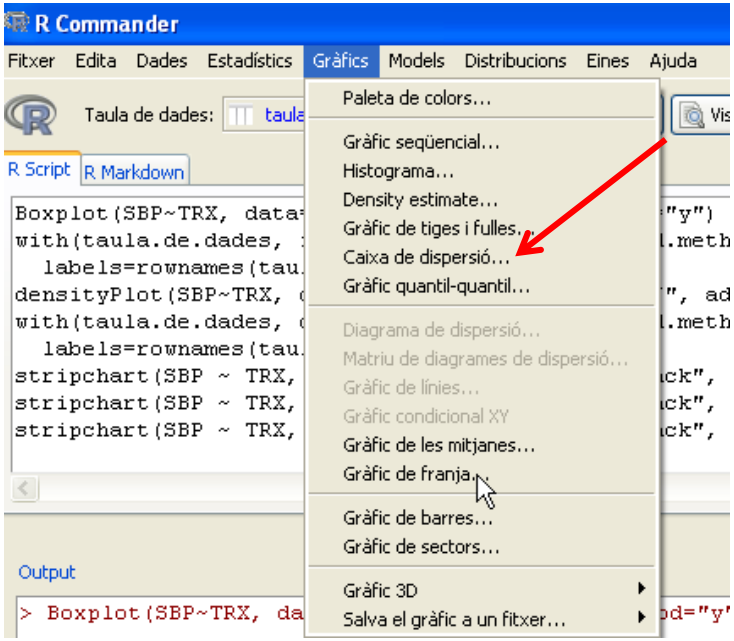
–Set up the null hypothesis: $H_0 : \mu_A = \mu_B = \mu_C$

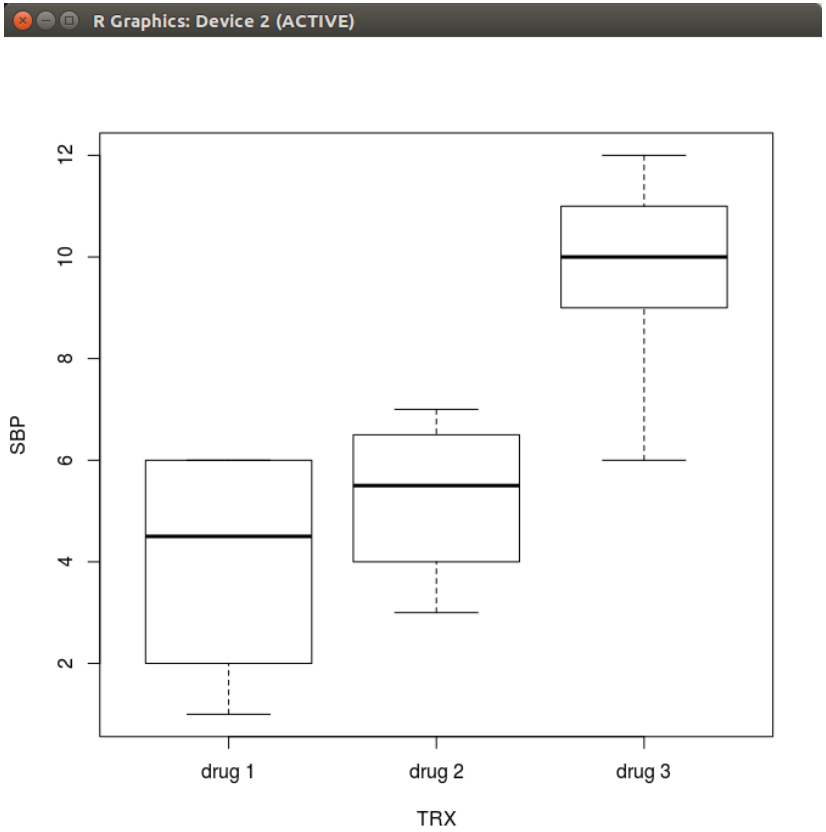
–Load and look the data:

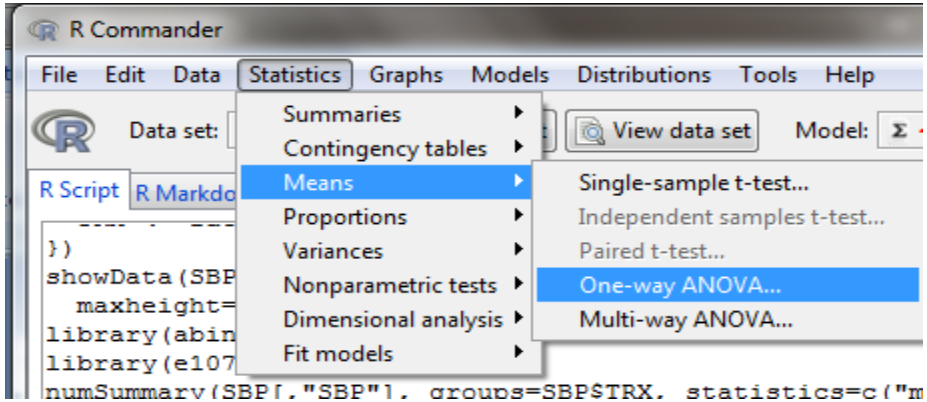
Numerical summaries:

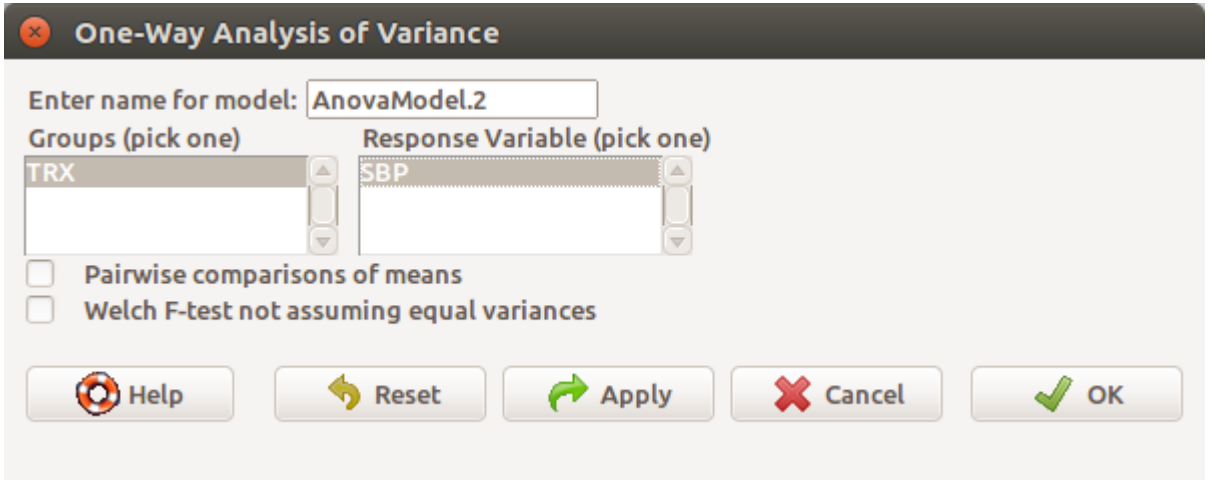
	mean	sd	IQR	0%	25%	50%	75%	100%	data:n
drug 1	4.00	2.054805	4.00	1	2	4.5	6.00	6	10
drug 2	5.25	1.488048	2.25	3	4	5.5	6.25	7	8
drug 3	9.60	2.302173	2.00	6	9	10.0	11.00	12	5

Let's see graphically:






R Commander
File Edit Data Statistics Graphs Models Distributions Tools Help
Data set:
View data set
Model:
Summaries
Contingency tables
Means
Proportions
Variances
Nonparametric tests
Dimensional analysis
Fit models
Single-sample t-test...
Independent samples t-test...
Paired t-test...
One-way ANOVA...
Multi-way ANOVA...
R Script
R Markdo
})
showData(SBP
maxheight=
library(abin
library(e107
numSummary(SBP[,"SBP"], groups=SBP$TRX, statistics=c("m


One-Way Analysis of Variance
Enter name for model: AnovaModel.2
Groups (pick one)
Response Variable (pick one)
TRX
SBP
Pairwise comparisons of means
Welch F-test not assuming equal variances
Help
Reset
Apply
Cancel
OK

```
> AnovaModel.2 <- aov(SBP ~ TRX, data=sbp)
```

```
> summary(AnovaModel.2)
             Df Sum Sq Mean Sq F value   Pr(>F)
TRX           2  106.5   53.26   14.26 0.000142 ***
Residuals    20   74.7    3.73
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
> with(sbp, numSummary(SBP, groups=TRX, statistics=c("mean", "sd")))
       mean       sd data:n
drug 1 4.00 2.054805     10
drug 2 5.25 1.488048      8
drug 3 9.60 2.302173      5
```

If we use ANOVA to compare means….why is it called “analysis of variance”?

H_0: The means of all the groups are equal

H_1: Not all the means are equal

ANOVA analyzes if there are differences between two variances to deduce if the are differences between the means

ANOVA partitions the total variability in the sample data into two component parts.

- Due to treatments: variation **BETWEEN** groups
- Due to the observations: variation **WITHIN** groups

Types of analysed variability by ANOVA

Sample	Drug1	Drug2	Drug3
1	4	7	9
2	2	6	12
3	6	5	6
4	6	7	11
5	5	6	10
6	6	4	11
7	2	7	9
8	6	5	10
Mean	**4.6**	**5.9**	**9.8**

$$\text{Total variability} = V_{\text{Between groups}} + V_{\text{Within groups}}$$

Vall d'Hebron Institut de Recerca

Types of analysed variability by ANOVA

Sample	Drug1	Drug2	Drug3
1	4	7	9
2	2	6	12
3	6	5	6
4	6	7	11
5	5	6	10
6	6	4	11
7	2	7	9
8	6	5	10
Mean	**4.6**	**5.9**	**9.8**

Variations among the conditions represents variation due to "treatment effects"

Total variability = $V_{\text{Between groups}} + V_{\text{Within groups}}$

Vall d'Hebron Institut de Recerca

Types of analysed variability by ANOVA

Sample	Drug1	Drug2	Drug3
1	4	7	9
2	2	6	12
3	6	5	6
4	6	7	11
5	5	6	10
6	6	4	11
7	2	7	9
8	6	5	10
Mean	**4.6**	**5.9**	**9.8**

Variation among participants within each condition represents "individual differences"

$$\text{Total variability} = V_{\text{Between groups}} + V_{\text{Within groups}}$$

Types of variability analysed by ANOVA

- Variation **BETWEEN** groups (SS_{Effect}): for each data value looks at the difference between its group mean and the overall mean

- Variation **WITHIN** groups (SS_{Error}): for each data value looks at the difference between that value and the mean of its group

Differences will be found between groups when:

Variation between groups > **Variation within groups**

- Strictly speaking to compare the two sum of squares (SS) they must be standardized.

dividing each sum of squares by their *degrees of freedom*

- The Table of ANOVA

Source	DF	SS	MS	F
Treatments	$g-1$	SS_{Trt}	$SS_{\text{Trt}}/(g-1)$	MS_{Trt}/MS_E
Error	$N-g$	SS_E	$SS_E/(N-g)$	

The ANOVA F-statistic is a ratio of the **Between group variation divided by the Within group variation**.

F = Between / Within

✓A **large F** is evidence against H_0, since it indicates that there are more difference between groups than within groups.

Source	DF	SS	MS	F
Treatments	$g-1$	SS_{Trt}	$SS_{\text{Trt}}/(g-1)$	MS_{Trt}/MS_E
Error	$N-g$	SS_E	$SS_E/(N-g)$	

```
> summary(AnovaModel.2)
            Df Sum Sq Mean Sq F value   Pr(>F)
TRX          2  106.5   53.26   14.26 0.000142 ***
Residuals   20   74.7    3.73
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

ANOVA's Table for two factor design:

Two *F* tests:

1. "Age" effective?

$$F = \frac{\text{variation from block}}{\text{variation due to error}} = \frac{MSB}{MSE}$$

2. "Dose" effective?

$$F = \frac{\text{variation from treatment}}{\text{variation due to error}} = \frac{MSTr}{MSE}$$

EXAMPLE

ANOVA's Table for two factor design:

Source of Variation	SS	df	MS	F	p-Value
Age Group	793	5	158.6	4.21	**0.0136**
Dose Level	217	3	72.3	1.92	**0.169**
Error	564	15	37.6		
TOTAL	1574	23			

We fail to reject the claim that there is no significant effect due to dose level on gene expression measurements

Table of contents

ANOVA's assumptions have to be checked in the data that they are true before run the model:

1. Each group is approximately normal (normally distributed).
2. The variances of different samples are homogeneous (homocedasticity)
3. The errors are independent from observation to observation.

How to check them?

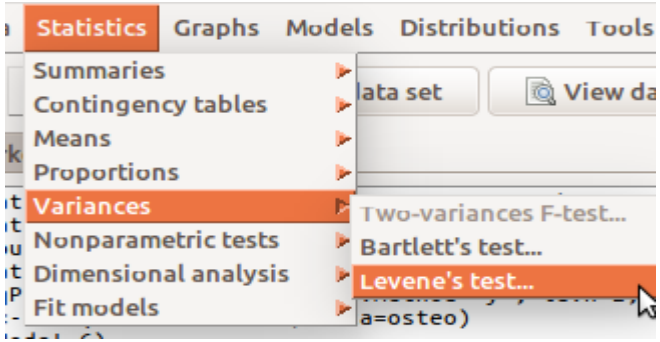
Normality of the data.

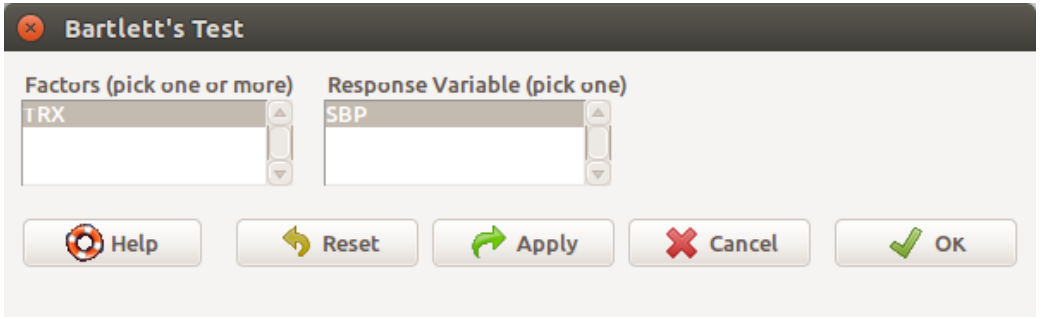
- Generally deviations from the normality do not seriously affect the validity of the assumptions. Not seriously outliers.
- F-test is very robust against non-normal data.
- The best way to check the normality of the data is a QQ plot (Shapiro-Wilk test is also valid)

Homocedasticity of the variances.

- F-test is very robust against heterogeneity of variances, especially with fixed factors and equal sample sizes (*balanced* designs)
- There are some statistical test, like Breusch-Pagan and Levene's test, to check that

Homocedasticity of the variances (in Rcommander)





H_0: error variances are equal

```
> leveneTest(SBP ~ TRX, data=sbp, center="median")

Levene's Test for Homogeneity of Variance (center = "median")
      Df F value Pr(>F)
group  2  0.6501 0.5327
      20

> bartlett.test(SBP ~ TRX, data=sbp)

        Bartlett test of homogeneity of variances

data:  SBP by TRX
Bartlett's K-squared = 1.075, df = 2, p-value = 0.5842
```

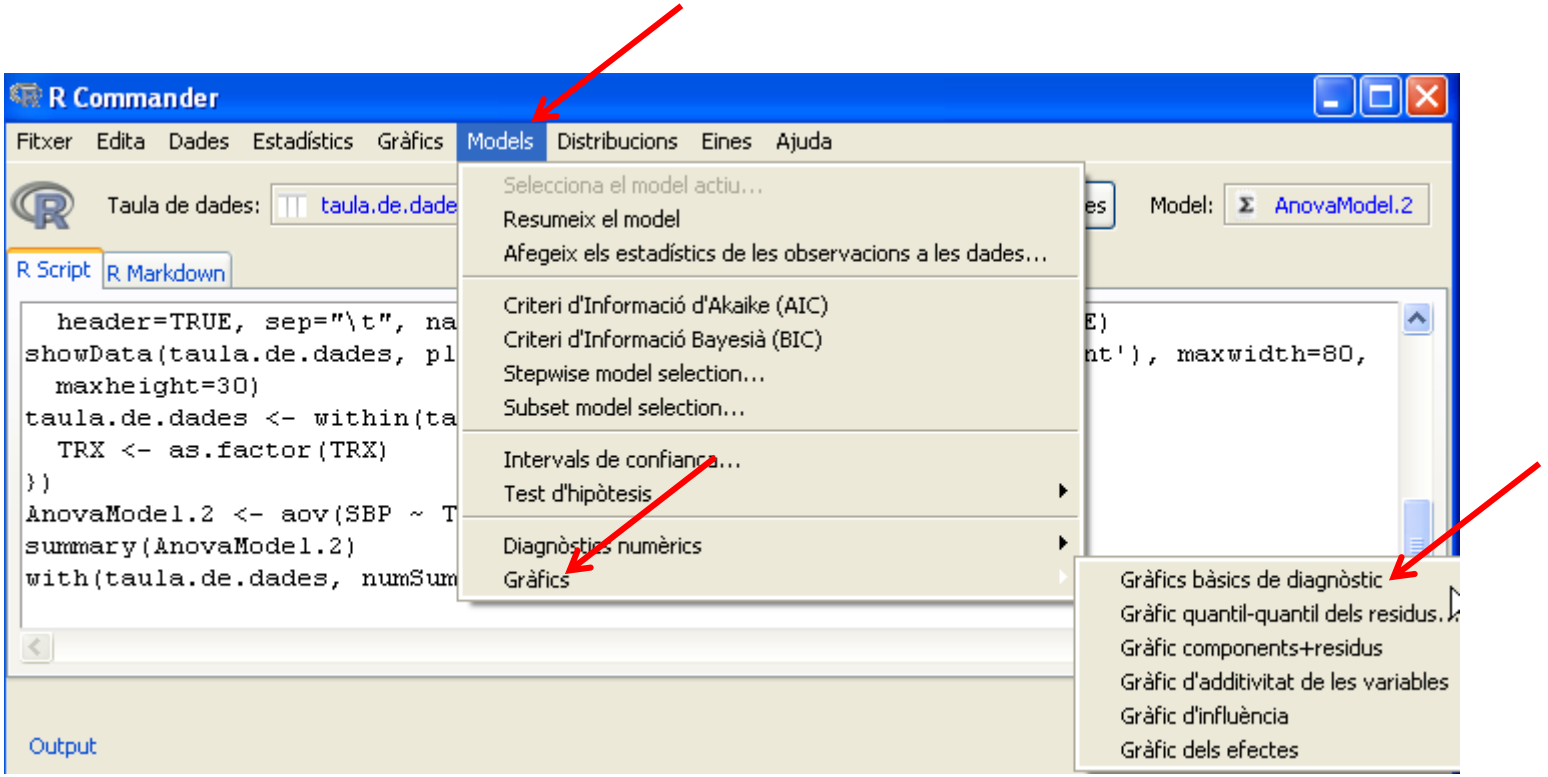
Independence of the residues.

- Independent observations
  - ✓No correlation between error terms
  - ✓No correlation between independent variables and error

- If the residues are independent they won't have to follow any clear pattern when we observe them in a plot.

We observe a plot of the residues vs estimates values

- How to check in R-Commander



- How to check in R-Commander.

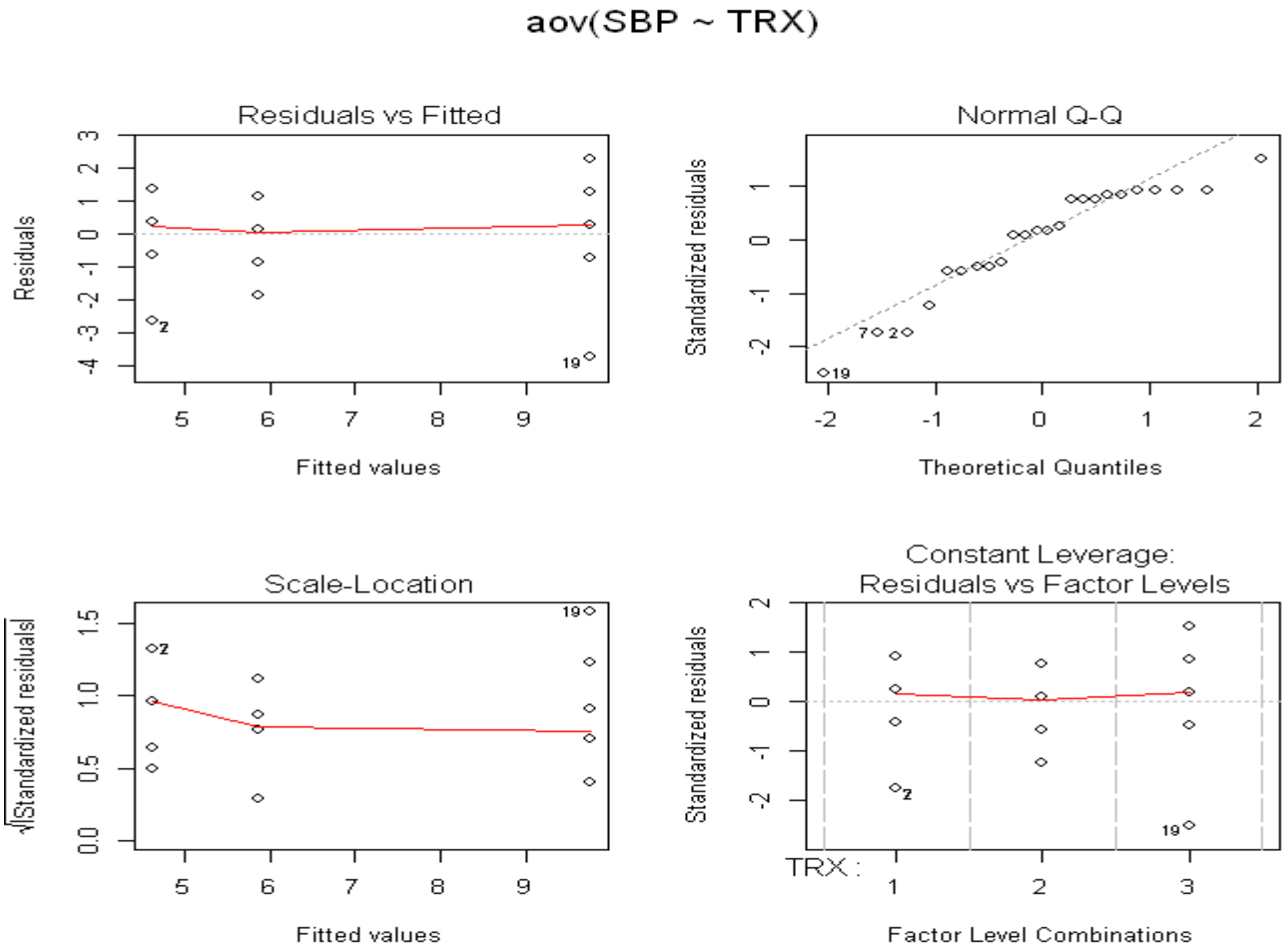


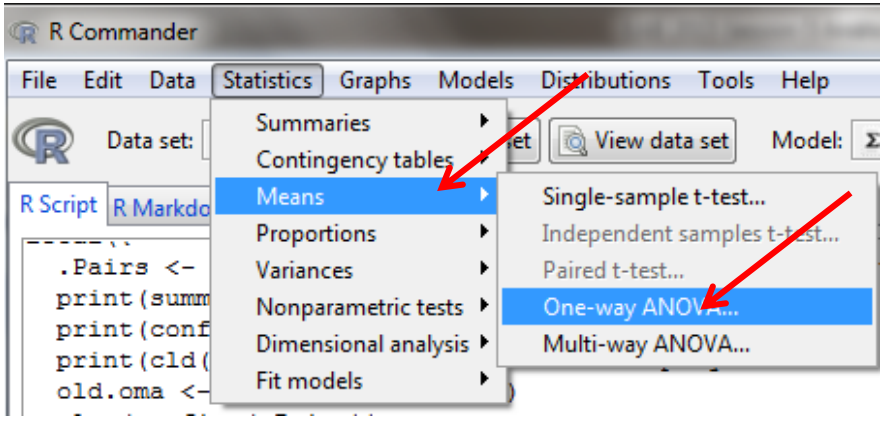
Table of contents

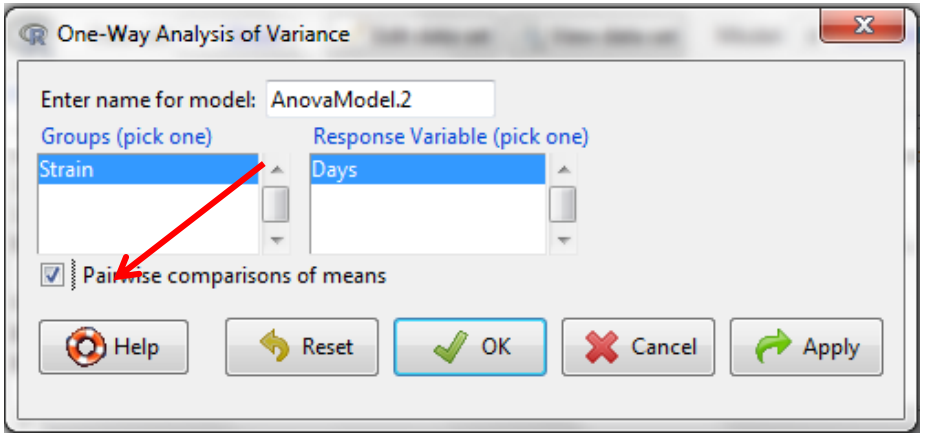
Until now we only can say if there differences among the groups compared, but we don't now between which groups.

- Usually we are interested in the comparison of the samples or treatments two by two (remember we couldn't uses a t-test!!!)

- We need to **adjust our p-value** threshold because we are doing multiple tests with the same data (type I error probability increases). There some test that take into account this.

Tukey-Kramer in R-Commander





A priori comparisons result is displayed again, plus…

```
> summary(AnovaModel.2)
            Df Sum Sq Mean Sq F value    Pr(>F)
Strain       4  148.2   37.05   29.88 3.39e-09 ***
Residuals   25   31.0    1.24
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> with(Strain, numSummary(Days, groups=Strain, statistics=c("mean", "sd")))
             mean        sd data:n
StrainA 10.000000 1.4142136      6
StrainB  6.166667 1.1690452      6
StrainC  8.666667 1.0327956      6
StrainD 11.500000 1.0488088      6
StrainE  5.666667 0.8164966      6
```

Tukey-Kramer in R-Commander

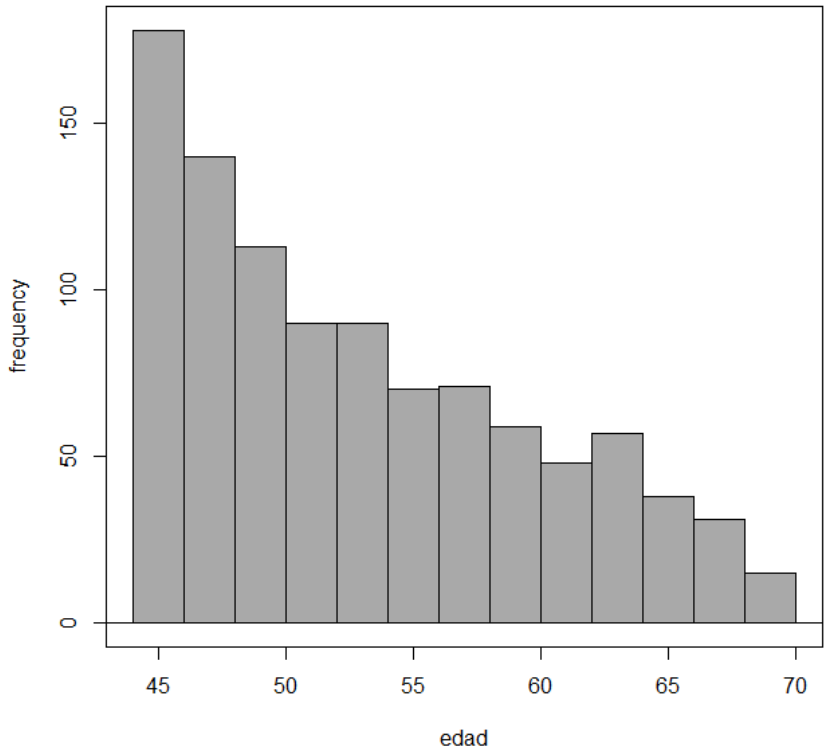
```
           Simultaneous Tests for General Linear Hypotheses

Multiple Comparisons of Means: Tukey Contrasts


Fit: aov(formula = Days ~ Strain, data = Strain)

Linear Hypotheses:
                       Estimate Std. Error t value Pr(>|t|)
StrainB - StrainA == 0  -3.8333     0.6429  -5.962  < 1e-04 ***
StrainC - StrainA == 0  -1.3333     0.6429  -2.074 0.262331
StrainD - StrainA == 0   1.5000     0.6429   2.333 0.167875
StrainE - StrainA == 0  -4.3333     0.6429  -6.740  < 1e-04 ***
StrainC - StrainB == 0   2.5000     0.6429   3.889 0.005444 **
StrainD - StrainB == 0   5.3333     0.6429   8.296  < 1e-04 ***
StrainE - StrainB == 0  -0.5000     0.6429  -0.778 0.934683
StrainD - StrainC == 0   2.8333     0.6429   4.407 0.001467 **
StrainE - StrainC == 0  -3.0000     0.6429  -4.666 0.000793 ***
StrainE - StrainD == 0  -5.8333     0.6429  -9.073  < 1e-04 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Adjusted p values reported -- single-step method)
```

Kruskal-Wallis Test (Non parametric ANOVA)

- Kruskal-Wallis test is used to find differences among *k* experimental samples or treatments
- This test don't assume the normality of the variables.
- It is similar to ANOVA, but it uses the ranges
- Used when number of observations are little and unbalanced designs.

Kruskal-Wallis Test in Rcommander

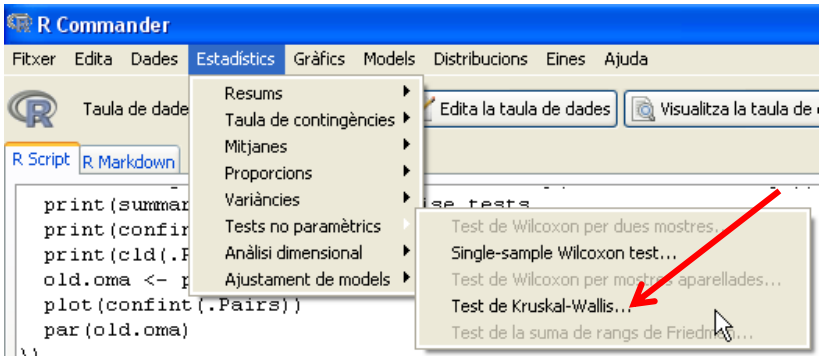
"Osteo" data set. Check if "age" is the same in the women of the study (classified by degree of illness)

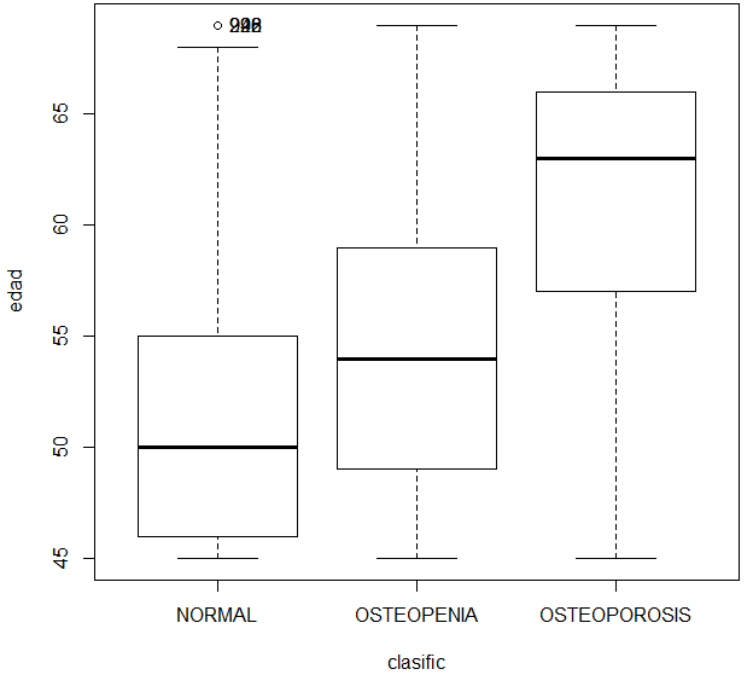


7. Beyond ANOVA

Kruskal-Wallis Test in Rcommander

Perform the test.






```
> kruskal.test(edad ~ clasific, data=osteo)

        Kruskal-Wallis rank sum test

data:  edad by clasific
Kruskal-Wallis chi-squared = 112.0766, df = 2, p-value < 2.2e-16
```

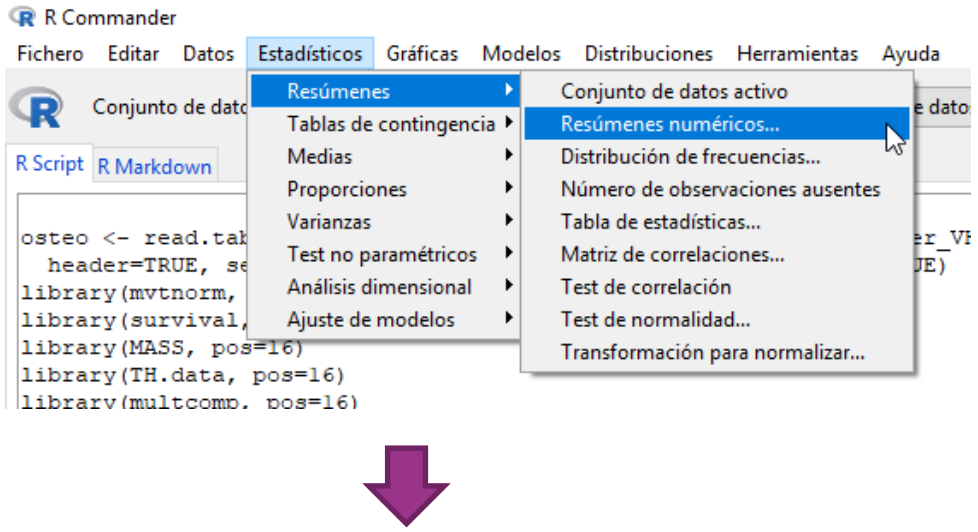
Table of contents

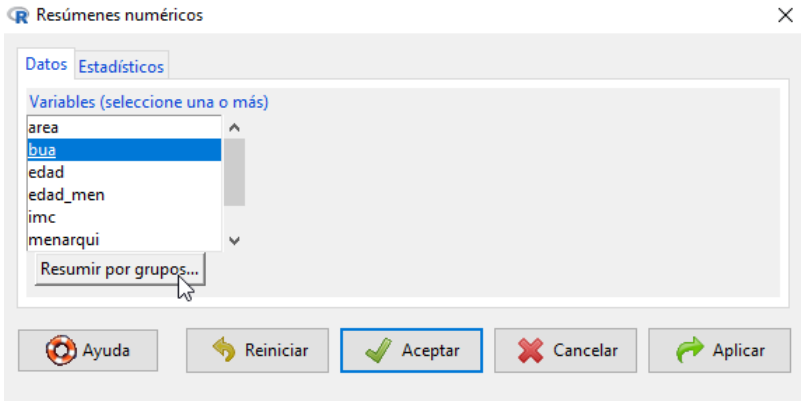
Exercise

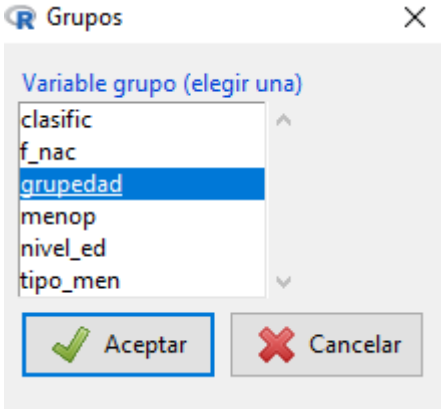
With the dataset osteo, find out if there are some relation between the *age of the women (**grupedad**)* and *the broadband ultrasound attenuation* (***bua***). After that try to analyze among which groups. The desirable steps to follow would be:

1. Exploratory data analysis (numerically and graphically)
2. Run the model.
3. Check for assumptions of the model (graphically)
4. Run post hoc comparison if necessary

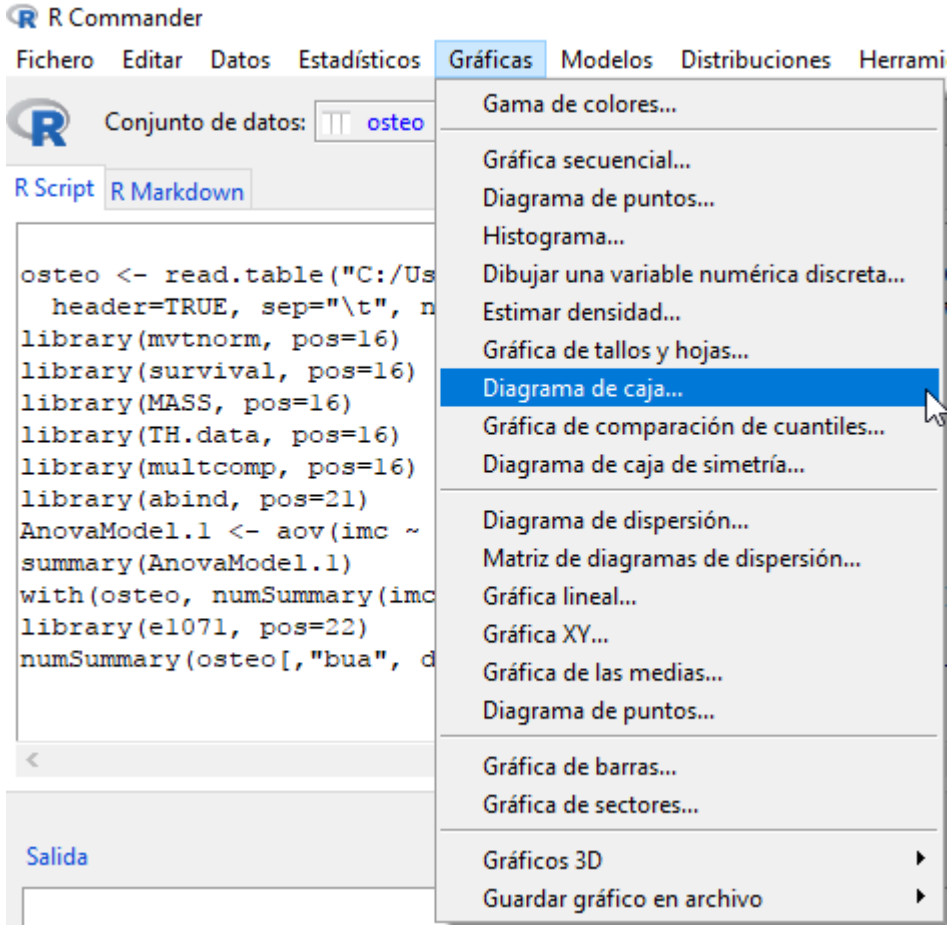
1. Exploratory data analysis (numerically and graphically)





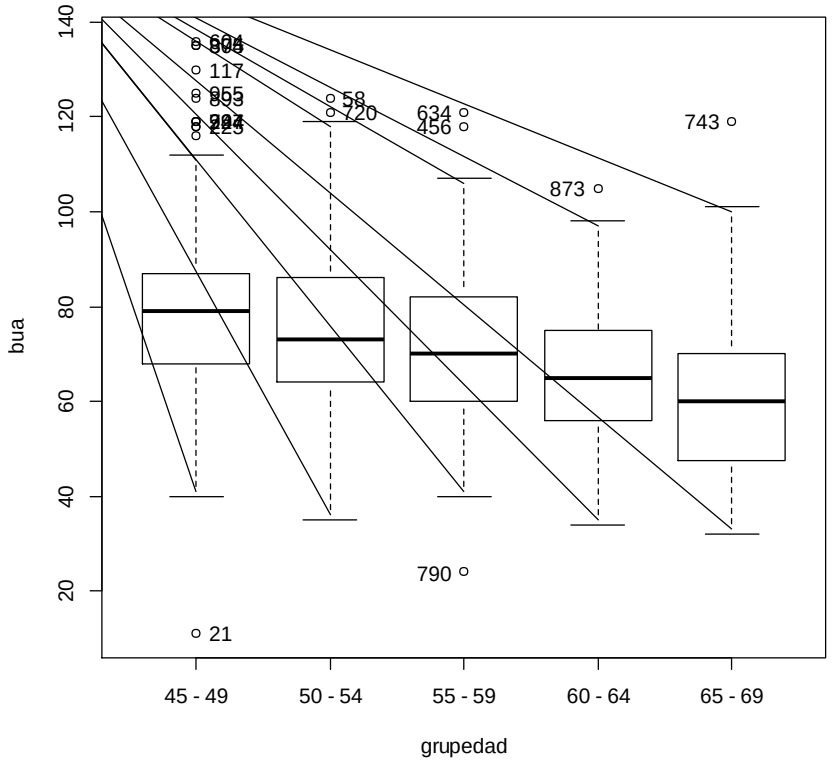


1. Exploratory data analysis (numerically and graphically)

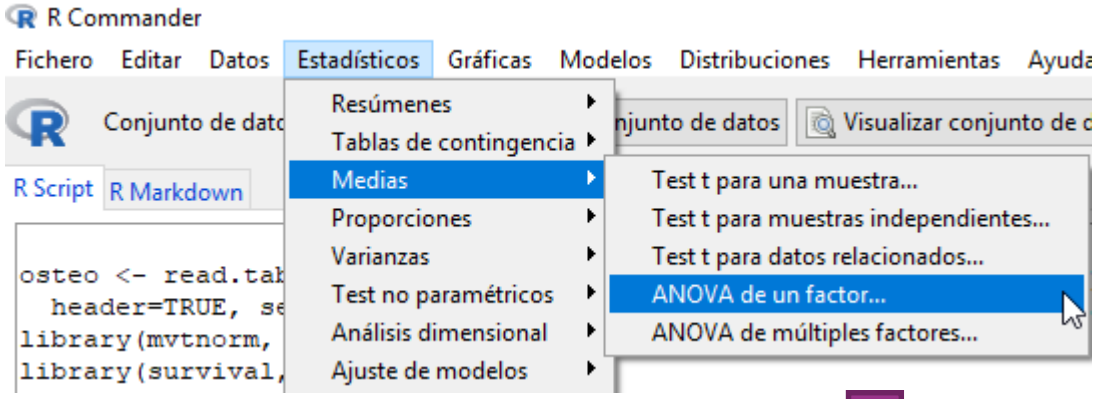


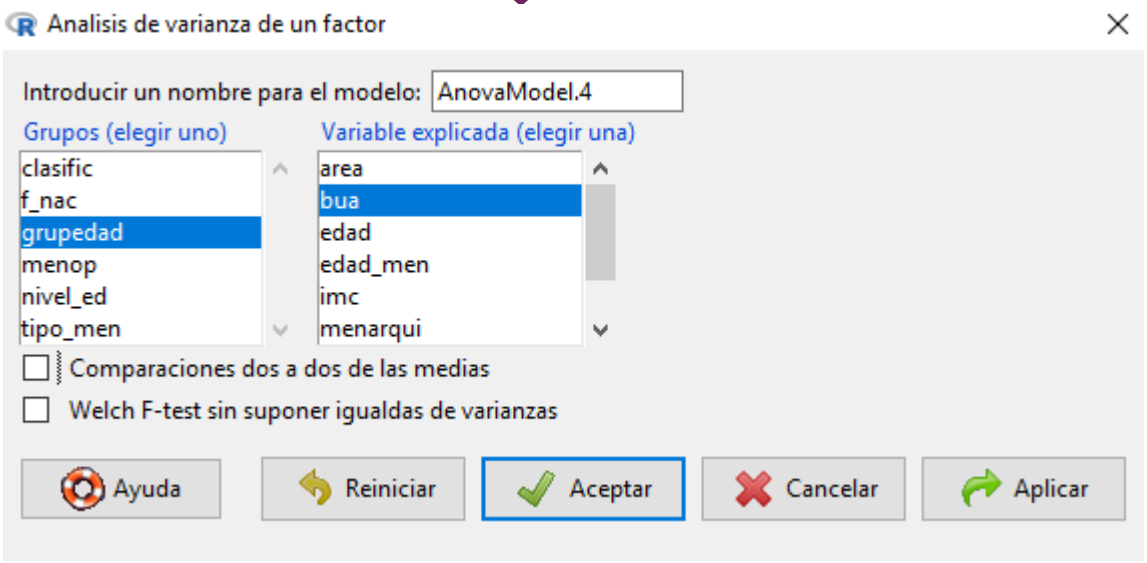
1. Exploratory data analysis (numerically and graphically)

```
> numSummary(osteo[,"bua", drop=FALSE], groups=osteo$grupedad, statistics=c("mean", "sd", "IQR", "quantiles"), quantiles=c(0,.25,.5,.75,1))
          mean       sd   IQR 0%   25% 50% 75% 100% bua:n
45 - 49 78.75926 16.24163 19.00 11 68.00  79  87  136   378
50 - 54 75.05150 15.78921 22.00 35 64.00  73  86  124   233
55 - 59 71.43182 15.86248 22.00 24 60.00  70  82  121   176
60 - 64 64.89147 13.91225 19.00 34 56.00  65  75  105   129
65 - 69 60.66667 15.92704 22.25 32 47.75  60  70  119    84
```



2. Run the model

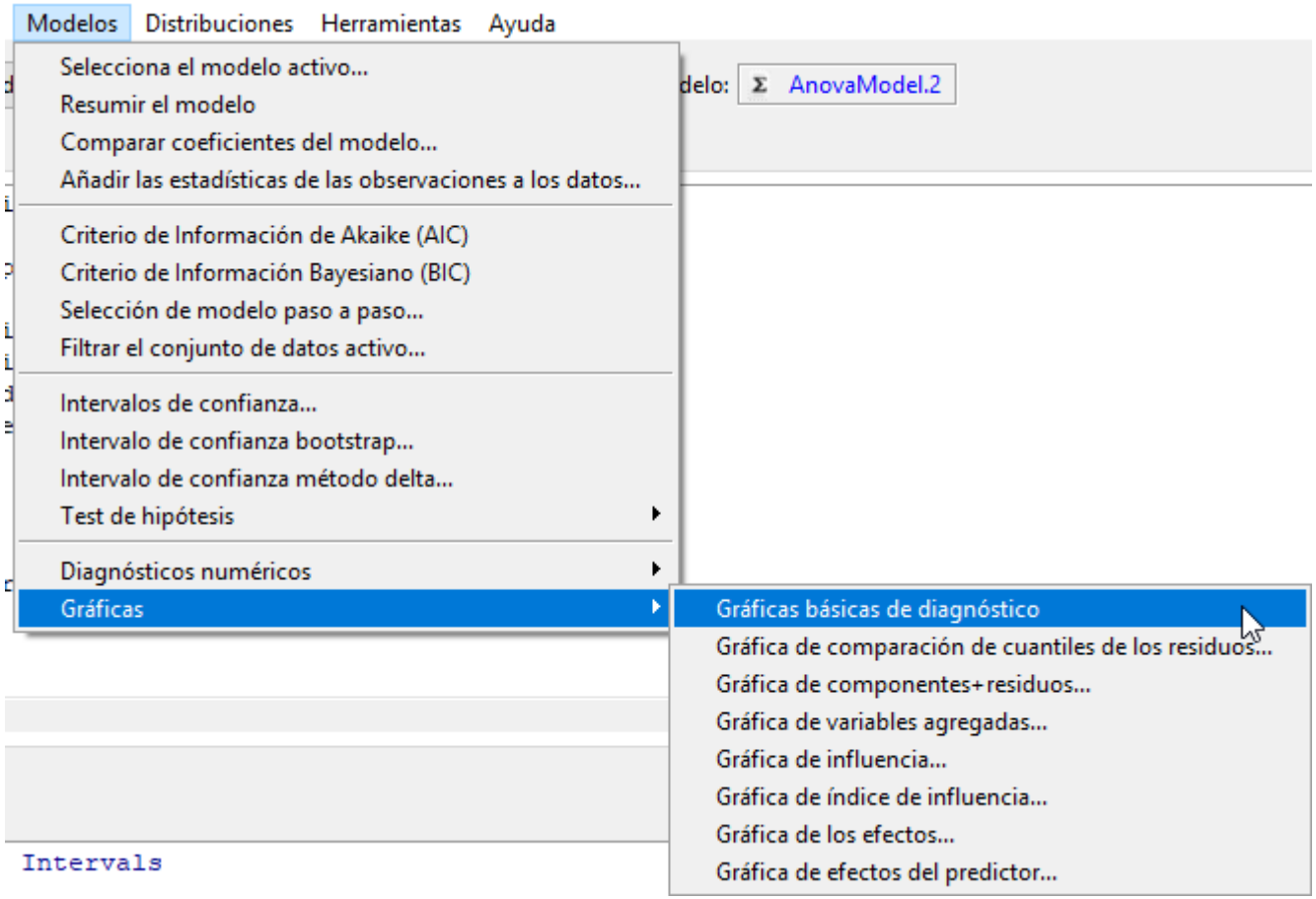




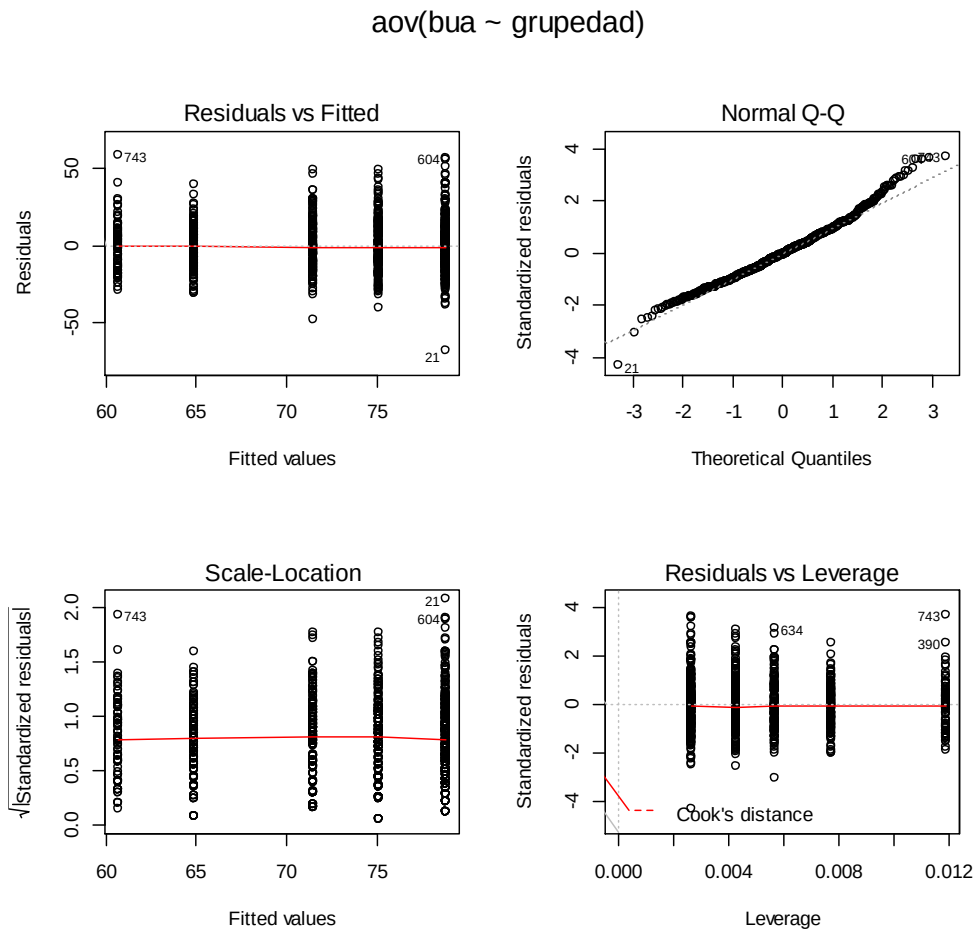
2. Run the model

```
> AnovaModel.5 <- aov(bua ~ grupedad, data=osteo)

> summary(AnovaModel.5)
             Df Sum Sq Mean Sq F value Pr(>F)
grupedad      4  35122    8780   35.35 <2e-16 ***
Residuals   995 247149     248
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

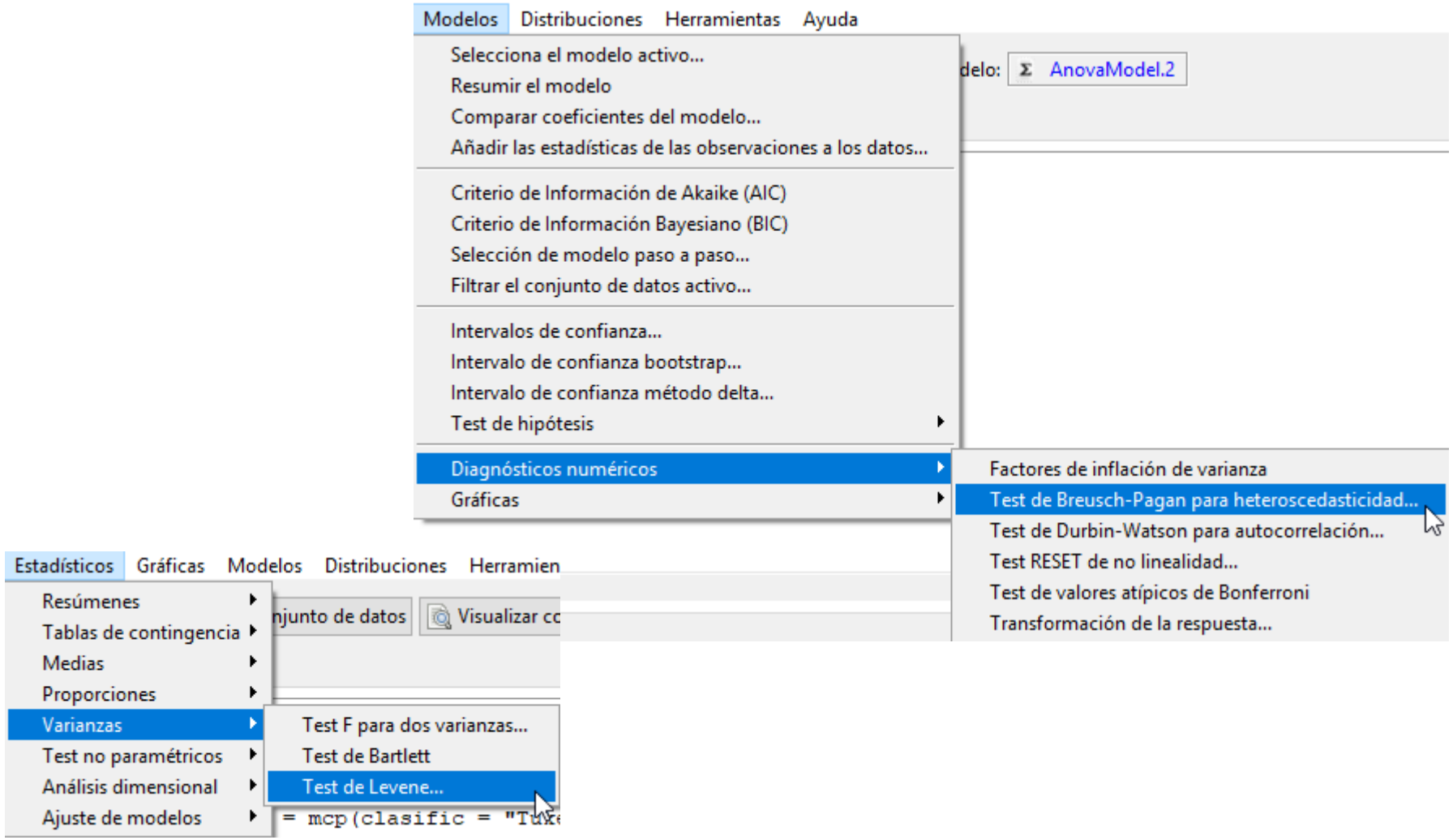
3. Check for assumptions of the model (graphically)



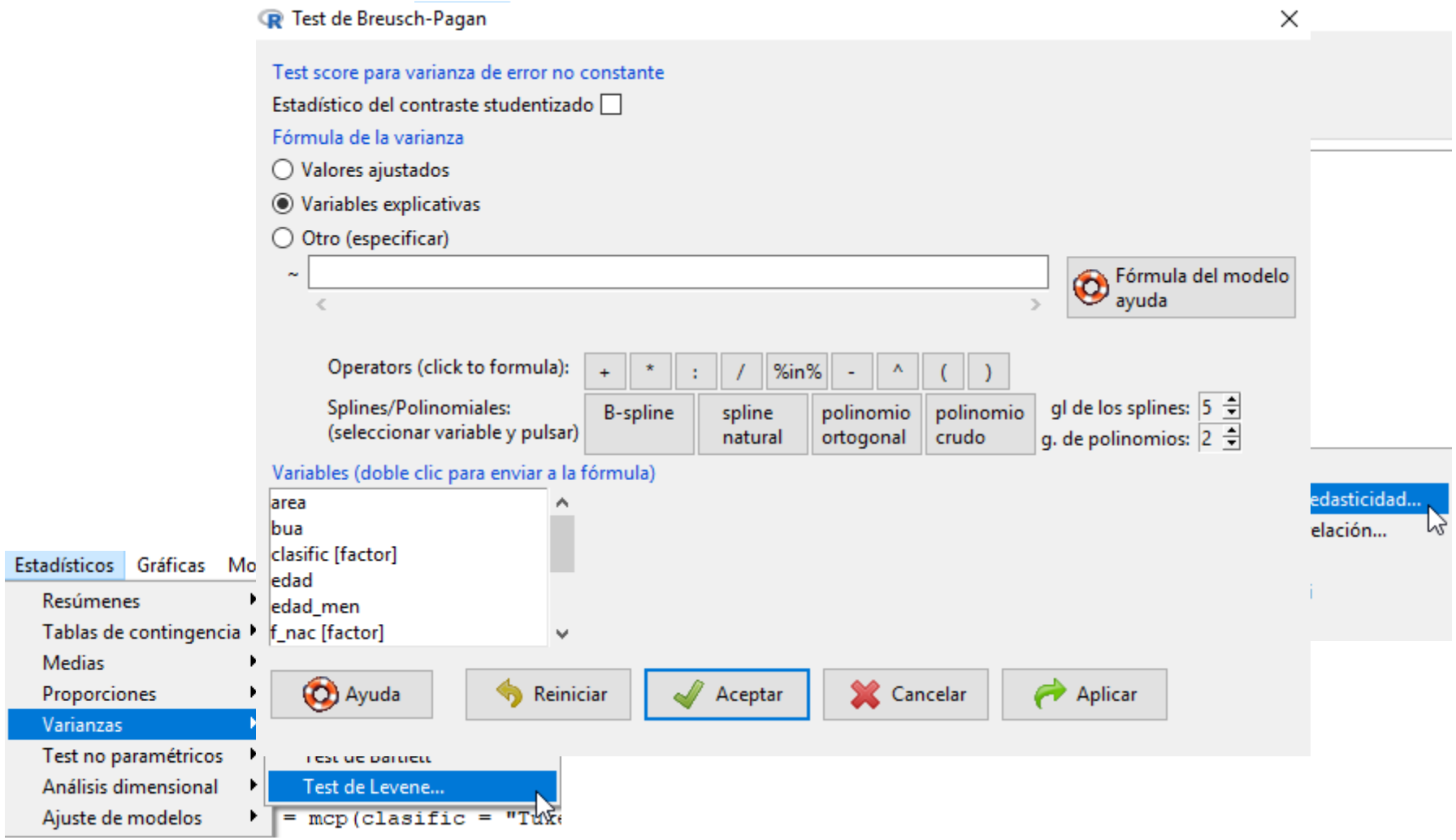
3. Check for assumptions of the model (graphically)



3. Check for assumptions of the model (graphically)



3. Check for assumptions of the model (graphically)



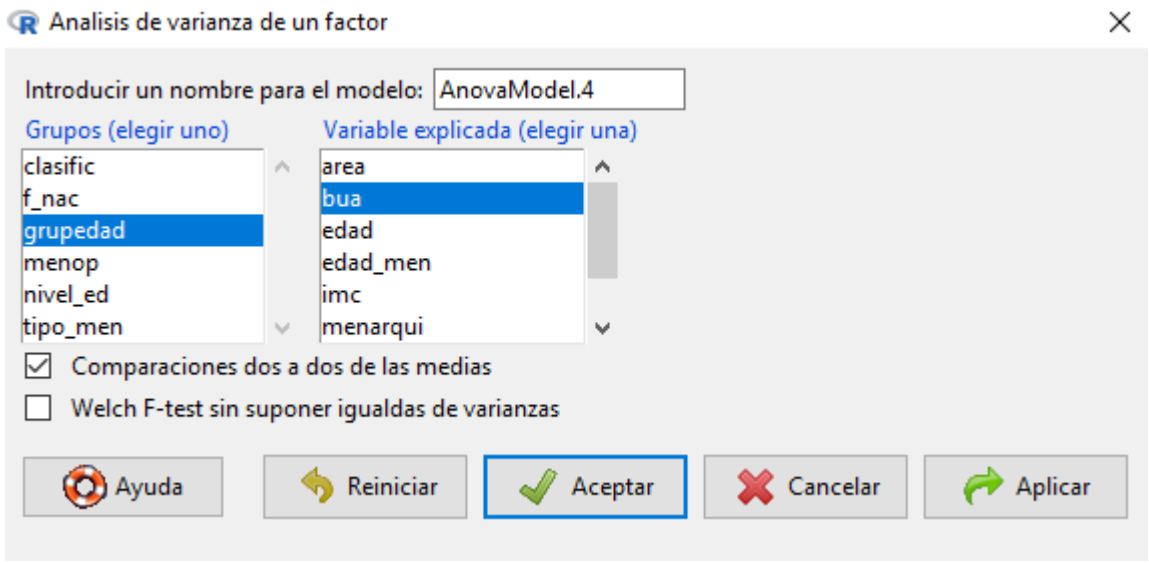
Vall d'Hebron
Institut de Recerca

3. Check for assumptions of the model (graphically)

```
> leveneTest(bua ~ grupedad, data=osteo, center="median")
Levene's Test for Homogeneity of Variance (center = "median")
       Df F value Pr(>F)
group   4  0.4487 0.7734
      995
```

```
> bptest(bua ~ grupedad, studentize=FALSE, data=osteo)

        Breusch-Pagan test

data:  bua ~ grupedad
BP = 4.0163, df = 4, p-value = 0.4038
```

4. Run post hoc comparison if necessary



4. Run post hoc comparison if necessary

```
         Simultaneous Tests for General Linear Hypotheses

Multiple Comparisons of Means: Tukey Contrasts


Fit: aov(formula = bua ~ clasific, data = osteo)

Linear Hypotheses:
                             Estimate Std. Error t value Pr(>|t|)
OSTEOPENIA - NORMAL == 0       -23.4279     0.6185  -37.88   <2e-16 ***
OSTEOPOROSIS - NORMAL == 0     -44.5249     1.2608  -35.31   <2e-16 ***
OSTEOPOROSIS - OSTEOPENIA == 0 -21.0970     1.2611  -16.73   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Adjusted p values reported -- single-step method)
```

Practical exercise of Biostatistics. Master VHIR 2018-19

In the database information of 149 diabetics patients followed up during 17 years is available. On the table below time to follow up after a baseline exam, vital status and other potential prognostic factors are provide. Age at study entry(EDAT), Body Mass Index (BMI), Age at diagnosis(EDATDIAGNO), Smoking habits(TABAC), systolic (SBP) and diastolic(DBP) blood pressure, Electrocardiogram result (ECG) and Coronary Heart Disease(CHD) are those prognosis factors.

- Is there a relationship among blood pressure and the result of ECG