Vall d'Hebron
Institut de Recerca

Hypothesis Testing . Quantitative Variables

Curs d'Estadística Bàsica per a la Recerca Biomèdica

UEB – VHIR
Miriam Mota, Santiago Pérez-Hoyos and Alex Sánchez
miriam.mota@vhir.org santi.perezhoyos@vhir.org

Vall d'Hebron Hospital · Generalitat de Catalunya Departament de Salut · Institut Català de la Salut · UAB Universitat Autònoma de Barcelona · Generalitat de Catalunya Departament d'Economia i Coneixement · Institució CERCA Centres de Recerca de Catalunya · ISC Instituto de Salud Carlos III

Syllabus

1. INTRODUCTION
2. TYPE OF TEST
3. NORMALITY TEST
4. **ONE GROUP** COMPARISON
5. **TWO GROUPS** COMPARISON IN INDEPENDENT SAMPLES
6. **TWO GROUPS** COMPARISON IN DEPENDENT SAMPLES
7. **MORE THAN TWO GROUPS** COMPARISON IN INDEPENDENT SAMPLES
8. MULTIPLE COMPARISONS AND MULTIPLE TESTING

Example Data

- A study was designed to compare two distinct hypertension control programs.
- 60 individuals with HTA were randomly assigned to either one or the other group (30 per group)
- Blood pressure was measured each month during a year

A	B	C	D	E	F	G	H	I
numero	sexo	grupo	tas1	tad1	tas2	tad2	tas3	tad3
1	VARON	B	150	100	150	90	170	
2	MUJER	B	160	90	170	90	160	
3	MUJER	B	150	90	110	90	115	
4	VARON	A	120	80	140	90	140	
5	MUJER	A	150	85	145	85	160	
6	MUJER	B	140	75	160	70	135	
7	MUJER	A	150	100	140	90	130	
8	VARON	A	160	90	170	90	170	
9	MUJER	A	145	105	170	95	140	
10	MUJER	A	210	110				
11	MUJER	A	170	100	170	90	170	
12	MUJER	B	140	90	140	90	100	

Questions to solve

- Are samples “comparable” at baseline?
- Has there been a change in BP between month 1 (first measure) and month 12?

Syllabus

1. INTRODUCTION
2. **TYPE OF TEST**
3. NORMALITY TEST
4. **ONE GROUP** COMPARISONS
5. **TWO GROUPS** COMPARISON IN INDEPENDENT SAMPLES
6. **TWO GROUPS** COMPARISON IN DEPENDENT SAMPLES
7. **MORE THAN TWO GROUPS** COMPARISON IN INDEPENDENT SAMPLES
8. MULTIPLE COMPARISONS AND MULTIPLE TESTING

Type of Hypothesis

Confirmation Hypothesis

Aim is to confirm hypothesis about parameters or distributions.

Goodness of fit test to verify hypothesis about the distribution of variable in population

- Does arterial pressure in the population follow a normal distribution?

Test to verify values about a parameter.

- Is the average "bua" value in our population equal to 70?
- Is the proportion of lung cancer cases equal to 2.6%?

Type of Hypothesis

Independence Hypothesis

- Aim is to test hypothesis about
  - relation among variables in a population
    - Is CD4 lymphocytes count related with CD8 count in HIV positive?
  - differences in a variable among two or more populations.
    - Is the average "bua" value the same in menopausic and in no menopausic population?
    - Is the proportion of lung cancer cases the same in people with high or low fruit consumption?

Type of Test

Parametric Test

- Assumes that the variable under study follows a particular distribution and
  Values about parameters are tested
  - Distribution of proportion of lung cancer is binomial

$$H_0: p = 3\%$$

  - Mean BUA value is the same in menopausic and non menopausic (assuming variable is normal or symmetric)

$$H_0: \mu_{Menopausic} = \mu_{Non\ \ menopausic}$$

  - Proportion of lung cancer is the same in high and low fruit consummers (assuming distribution of counts is binomial).

$$H_0: p_{High\ fruit} = p_{Low\ \ fruit}$$

Type of Test

Non-Parametric Test

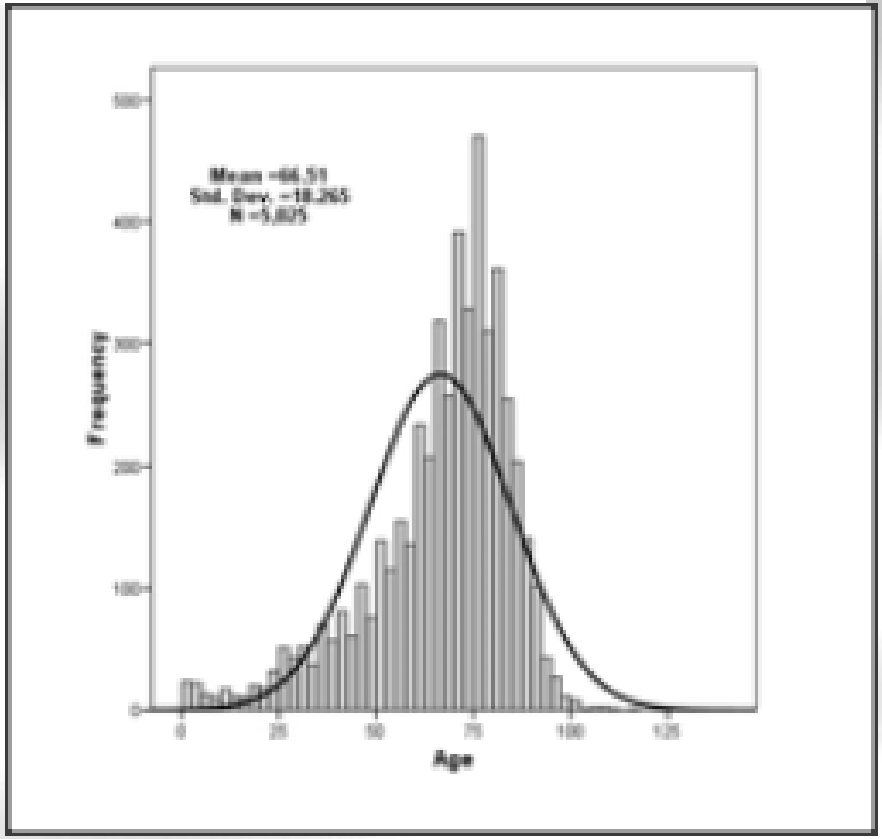
- No distribution is assumed for the data,
- Tests are about distribution not about is parameters
  - BUA distr. in Menopausic = BUA distr. in non-menopausic
  - Lung cancer is not related to fruit consumption. They are independent.

Syllabus

1. INTRODUCTION
2. TYPE OF TEST
3. **NORMALITY TEST**
4. **ONE GROUP** COMPARISON
5. **TWO GROUPS** COMPARISON IN INDEPENDENT SAMPLES
6. **TWO GROUPS** COMPARISON IN DEPENDENT SAMPLES
7. **MORE THAN TWO GROUPS** COMPARISON IN INDEPENDENT SAMPLES
8. MULTIPLE COMPARISONS AND MULTIPLE TESTING

Normality test

- Some parametrical test assume data come from a normal population
- How can we check this assumption?
- What can we do if assumption is false?

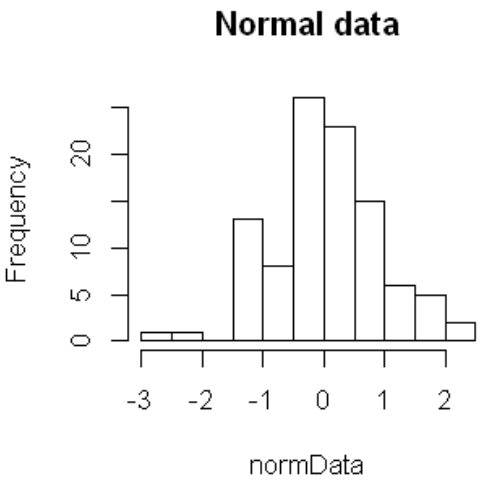


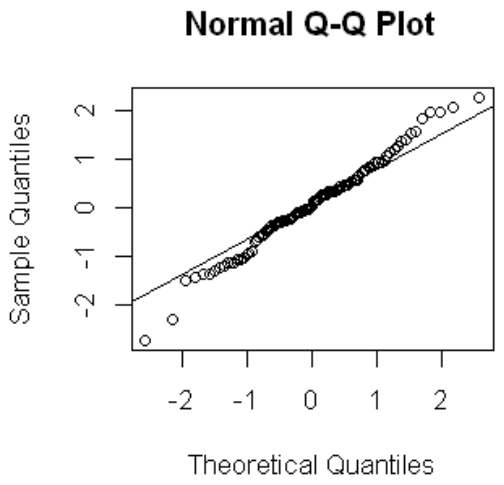
Testing normality

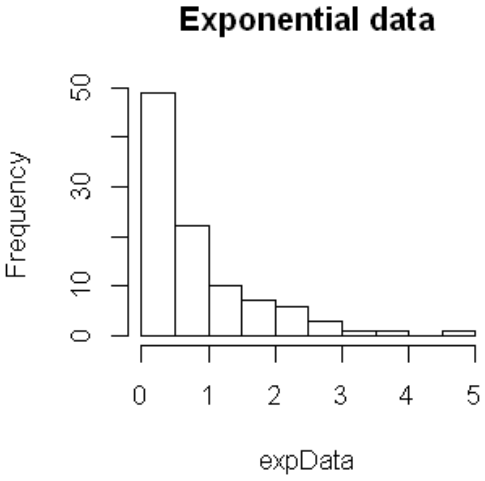
- We can use graphical methods or hypothesis tests
- Graphs
  - Check if it is a symmetric distribution
  - Probaility graphs (QQ-plots)
- Hypothesis test (Normality)
  - Kolmogorov-Smirnov test
  - Kolmogorov-Liliefors test
  - Shapiro-Wilks test

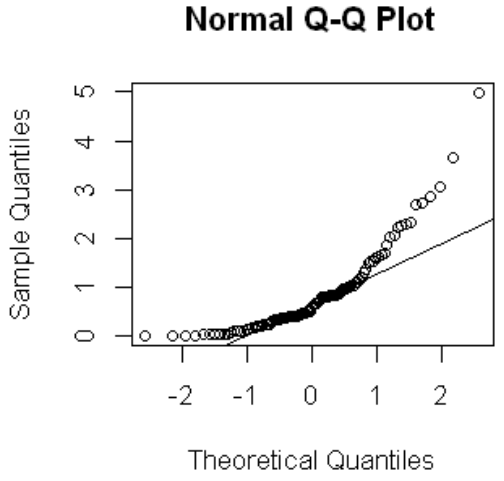
Histograms and QQ-plots

- Histogram
  - It should be symmetric with gaussian shape.
- QQ-plot
  - Dots should be over the diagonal line
- Non normal data deviate from normal patterns.
- Difficult to quantify if there are few data



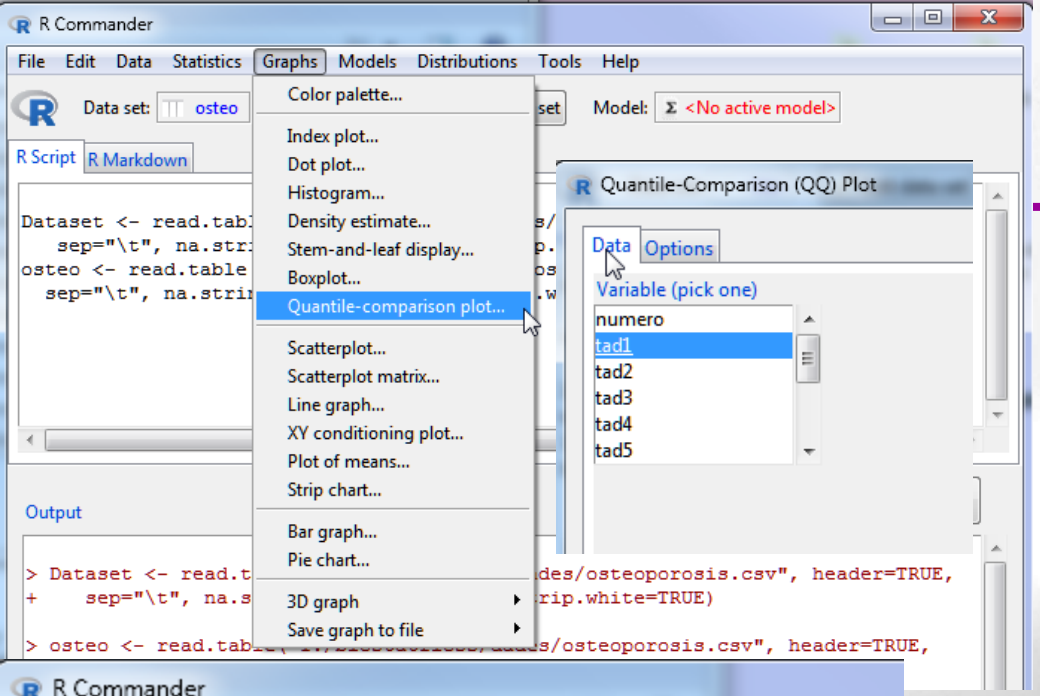


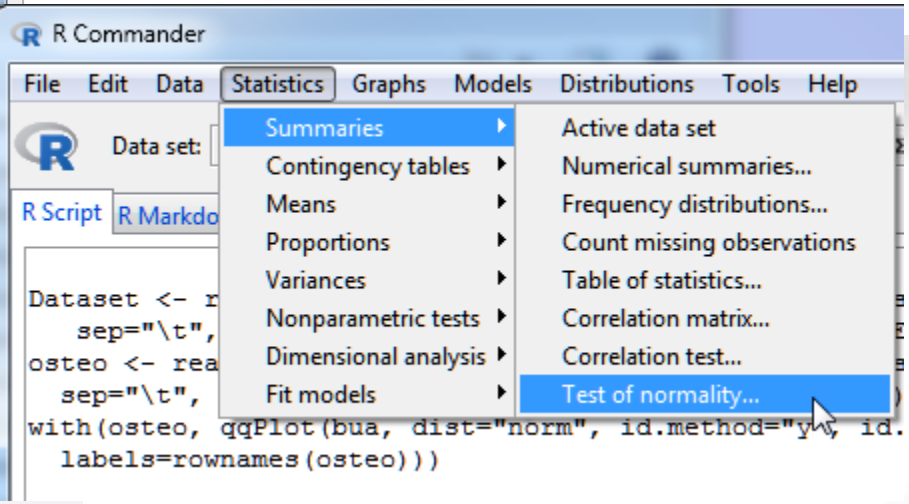


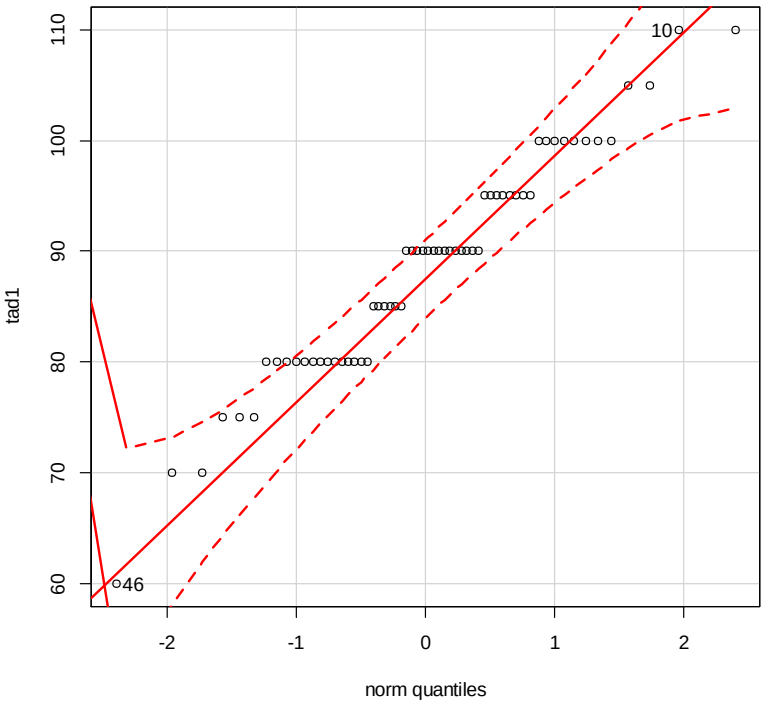


Normality test

- Statistical normality test are more precise than graphs. It is possible to calculate a p-value.
- The most used tests are Kolmogorov-Smirnov and Shapiro-Wilks test.
- The hypothesis to test are:
  - H_0: Data follow a normal distribution
  - H_1 : Data do not follow a normal distribution







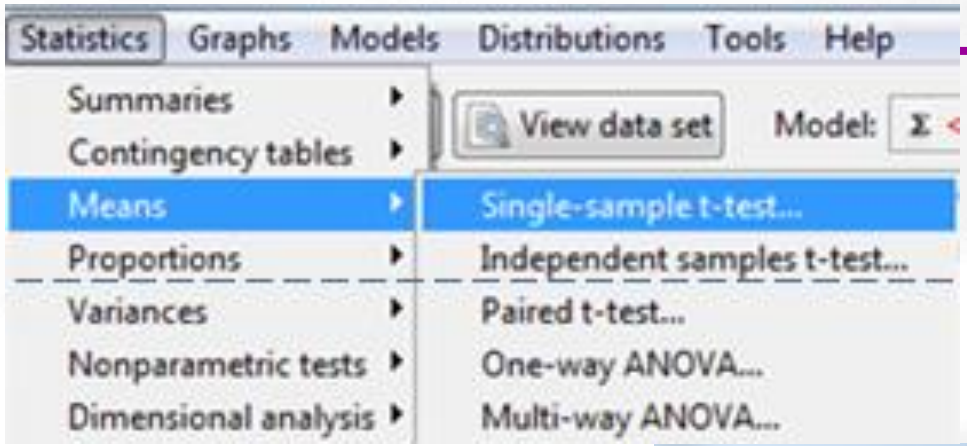

```
        Shapiro-Wilk normality test

data:  tad1
W = 0.96622, p-value = 0.09512
```

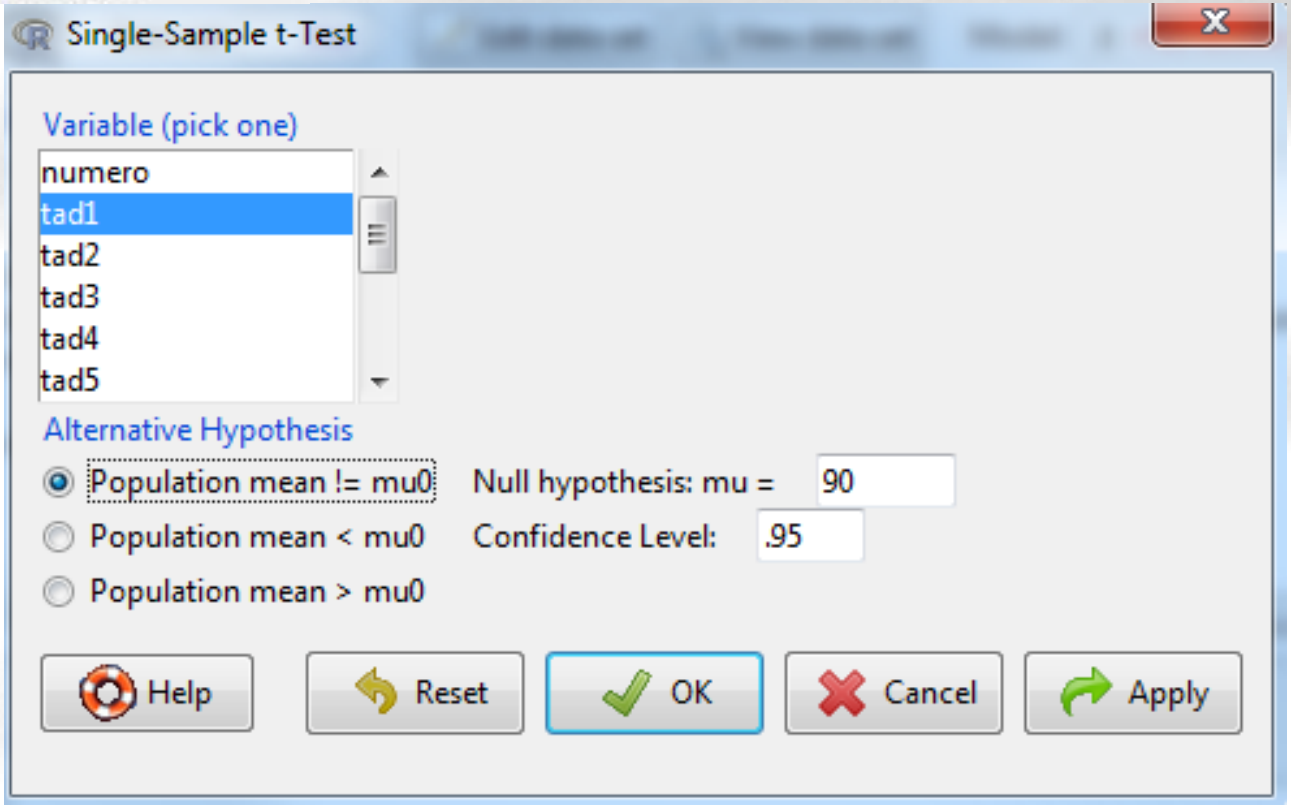
Syllabus

1. INTRODUCTION
2. TYPE OF TEST
3. NORMALITY TEST
4. **ONE GROUP** COMPARISON
5. **TWO GROUPS** COMPARISON IN INDEPENDENT SAMPLES
6. **TWO GROUPS** COMPARISON IN DEPENDENT SAMPLES
7. **MORE THAN TWO GROUPS** COMPARISON IN INDEPENDENT SAMPLES
8. MULTIPLE COMPARISONS AND MULTIPLE TESTING

One sample t-test

- We do not use it very often.
- Very similar to estimation questions. It can be solved calculalting a confidence itnerval
- Idea: We want to verify form a sample a previous hypothesis about the mean in a population
- *Can it be accepted that the initial TAD is 90 in Hipertensive patients?*


Statistics
Graphs
Models
Distributions
Tools
Help
Summaries
Contingency tables
Means
Proportions
Variances
Nonparametric tests
Dimensional analysis
View data set
Model:
Single-sample t-test...
Independent samples t-test...
Paired t-test...
One-way ANOVA...
Multi-way ANOVA...


Single-Sample t-Test
Variable (pick one)
numero
tad1
tad2
tad3
tad4
tad5
Alternative Hypothesis
Population mean != mu0
Population mean < mu0
Population mean > mu0
Null hypothesis: mu =
90
Confidence Level:
.95
Help
Reset
OK
Cancel
Apply

One Sample t-test

```
data:  tad1
t = -1.2137, df = 59, p-value = 0.2297
alternative hypothesis: true mean is not equal to 90
95 percent confidence interval:
 85.80626 91.02707
sample estimates:
mean of x
 88.41667
```

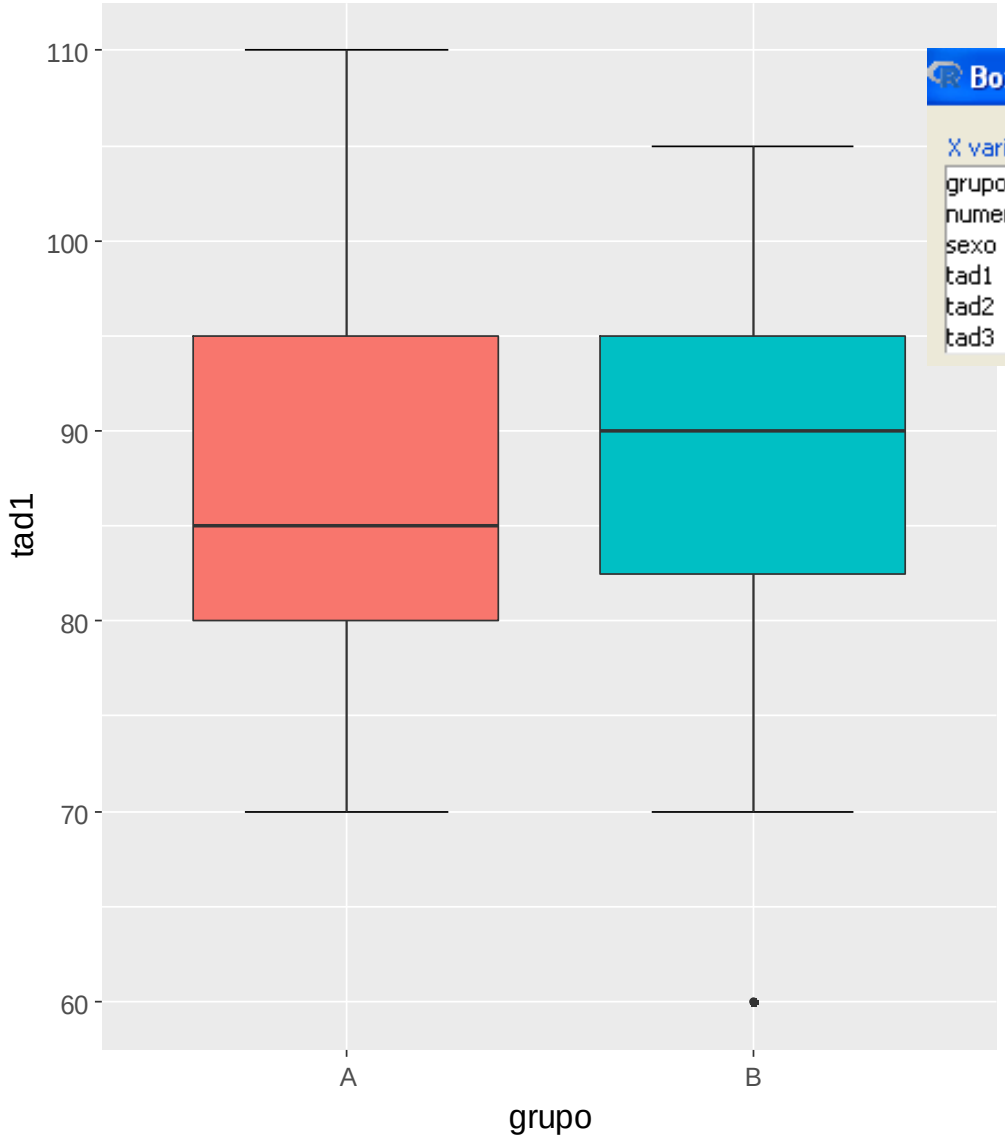
Syllabus

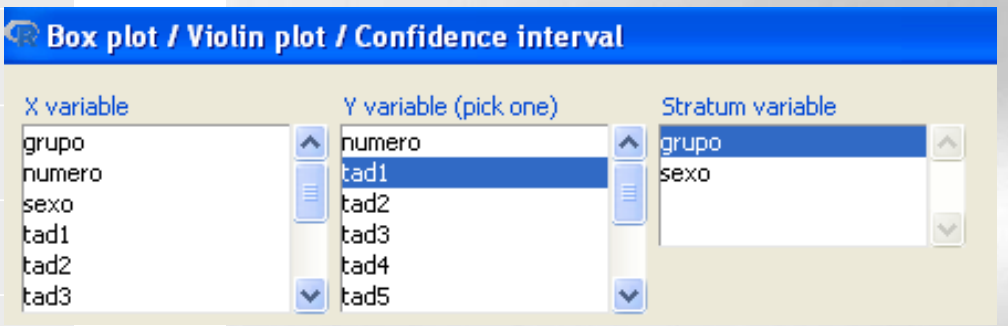
1. INTRODUCTION
2. TYPE OF TEST
3. NORMALITY TEST
4. **ONE GROUP** COMPARISON
5. **TWO GROUPS** COMPARISON IN INDEPENDENT SAMPLES
6. **TWO GROUPS** COMPARISON IN DEPENDENT SAMPLES
7. **MORE THAN TWO GROUPS** COMPARISON IN INDEPENDENT SAMPLES
8. MULTIPLE COMPARISONS AND MULTIPLE TESTING

Questions to answer

- Are samples comparable at baseline time
- Is blood pressure comparable between first and 12th measures

Boxplot TAD, by group





Compare a Quantitative variable in two groups

Null Hypothesis: There is not difference of the variable in two population or groups

Samples have been generated

INDEPENDENT

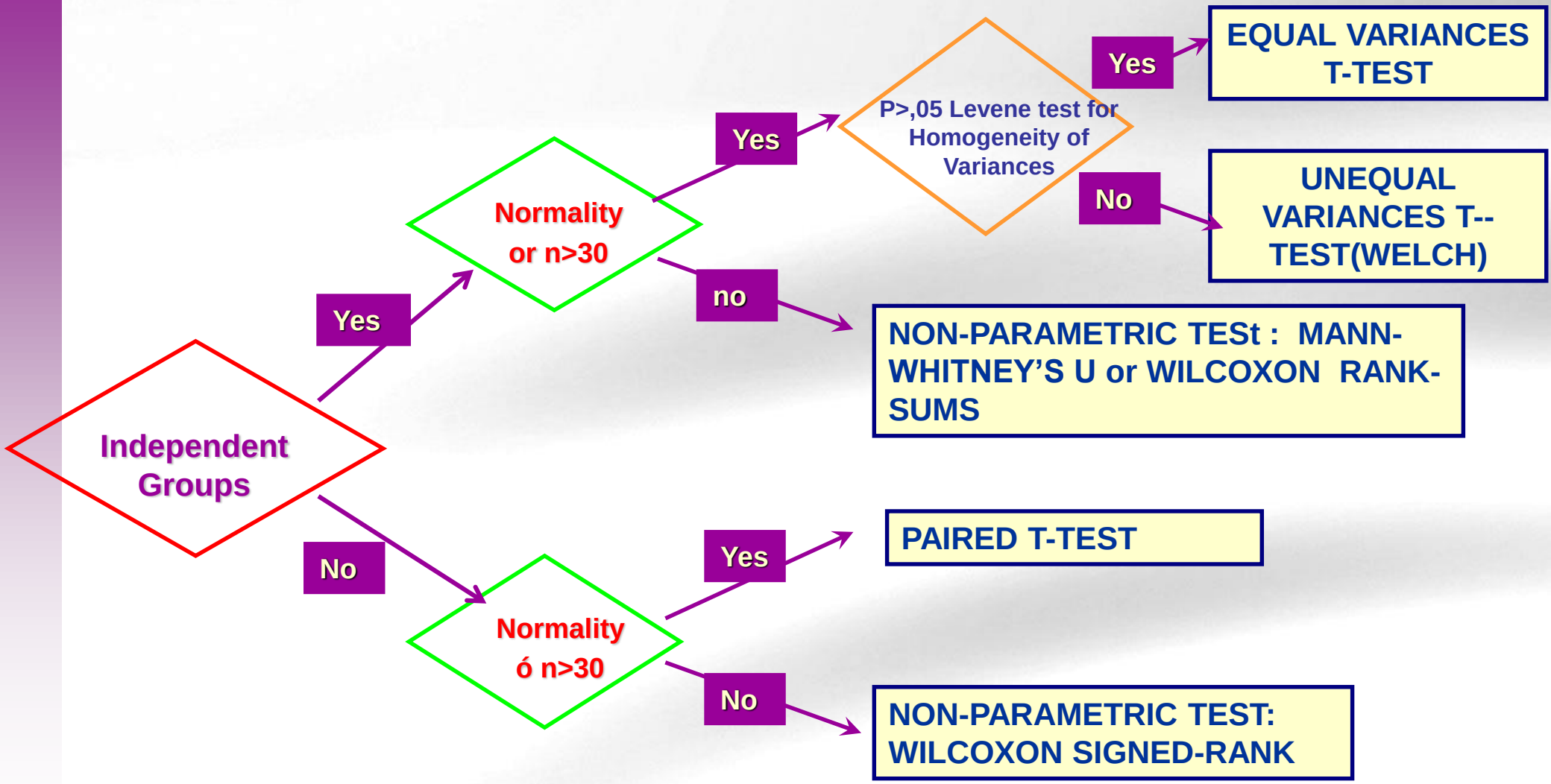
Selected individuals in a group have nothing to do with selected individuals in the other group.

DEPENDENT

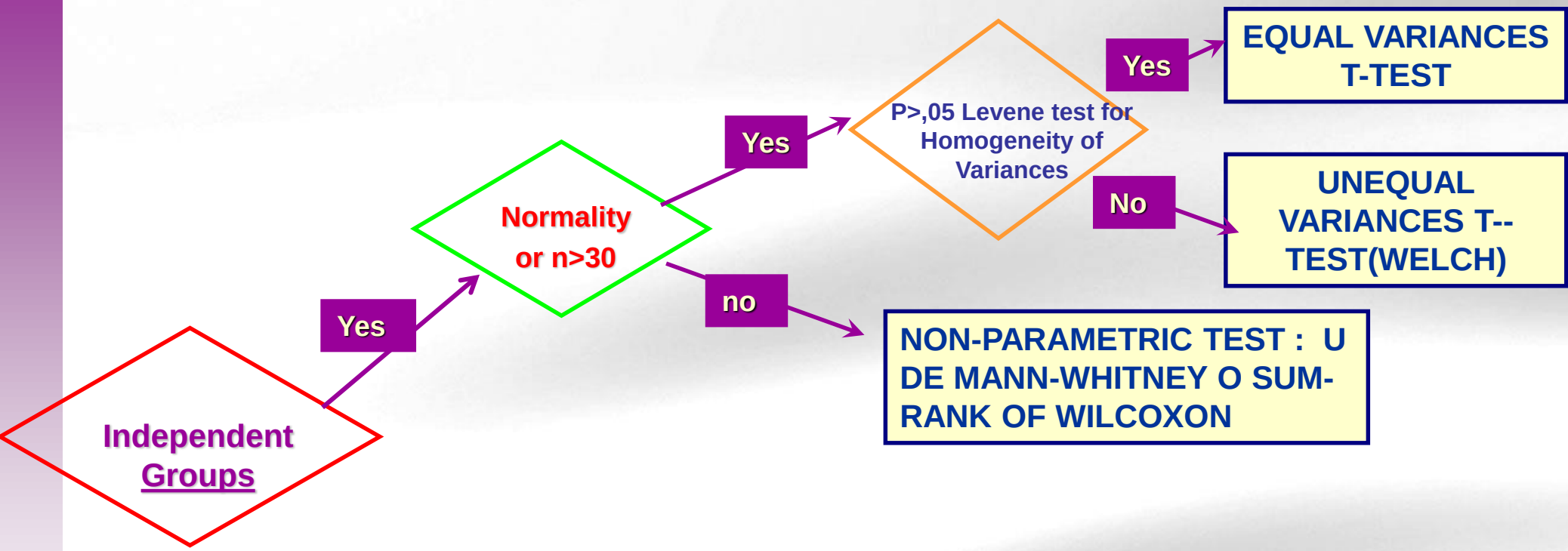
Each individual in a group has a correspondent in other group.

These are ***paired data***.

Two sample tests

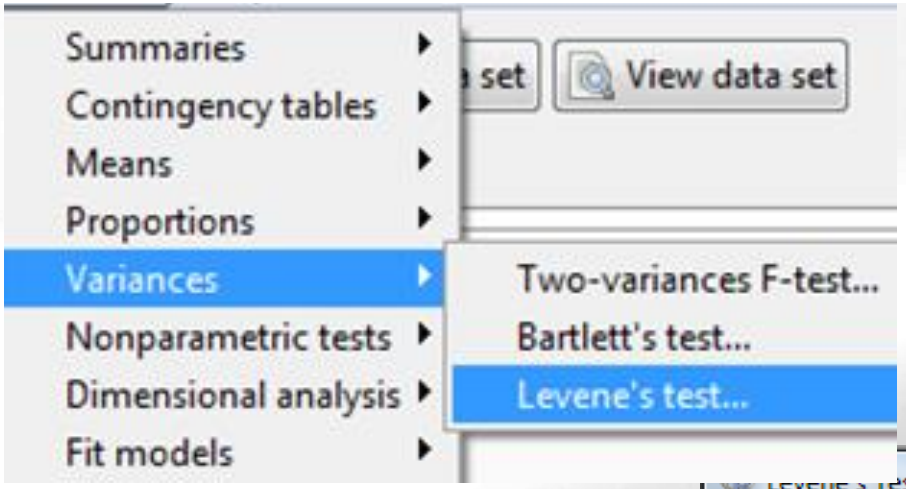


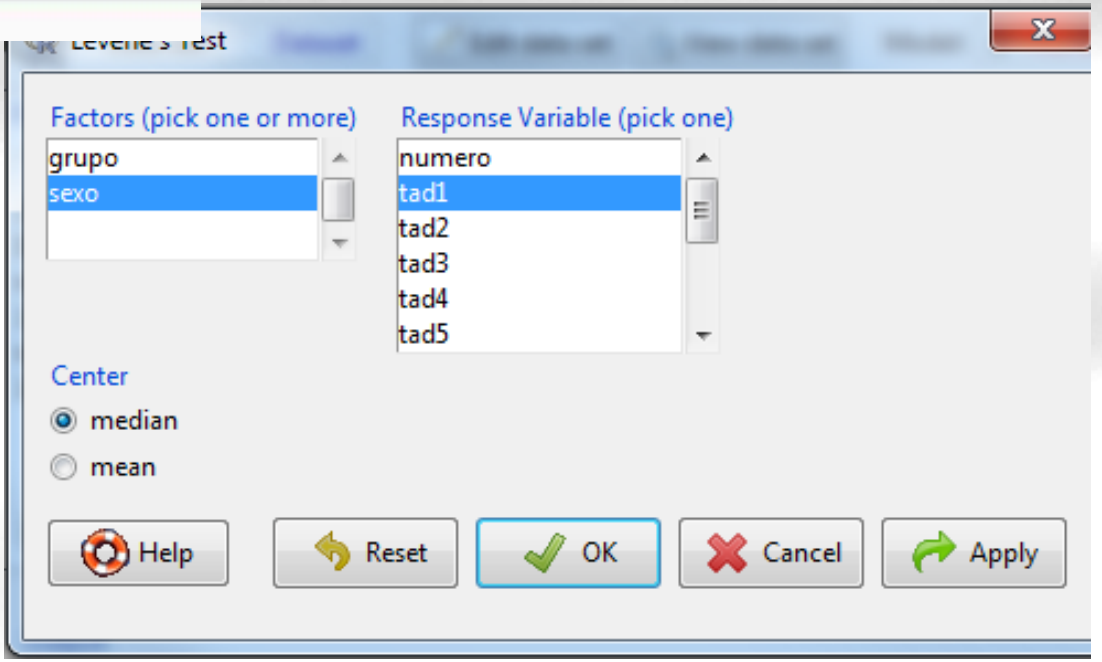
Two sample tests (1)



1. Data is normal (normality test) or sample size > 30.
2. Mean is a good summary statistic for this problem.
3. Test homogeneity of variances

Testing variance homogeneity

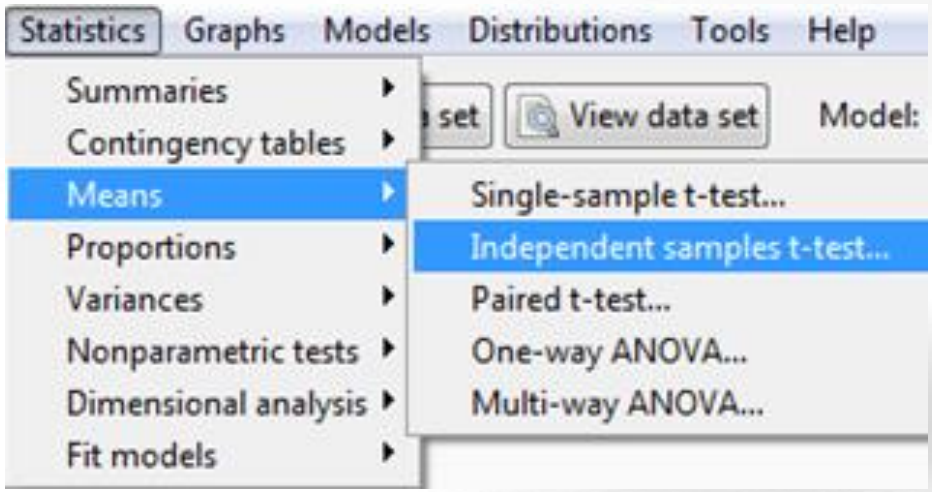


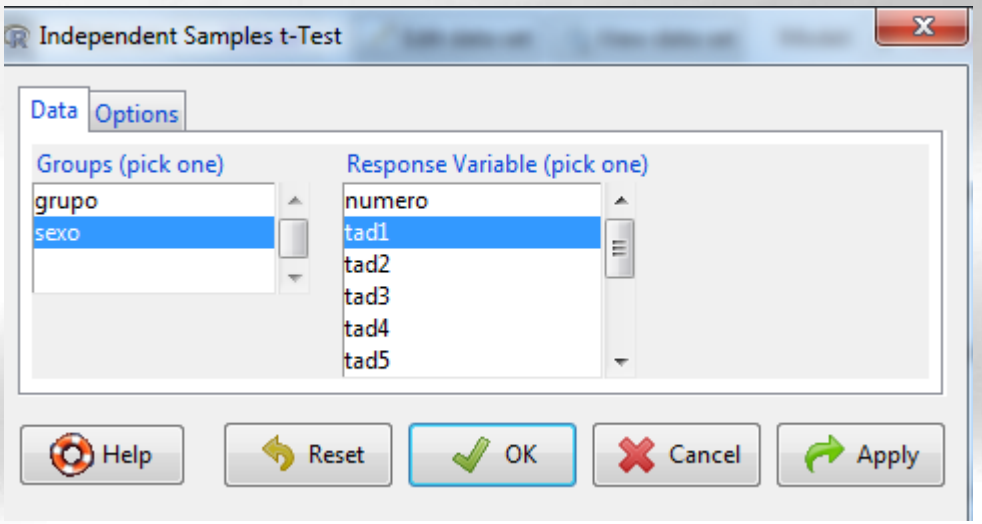


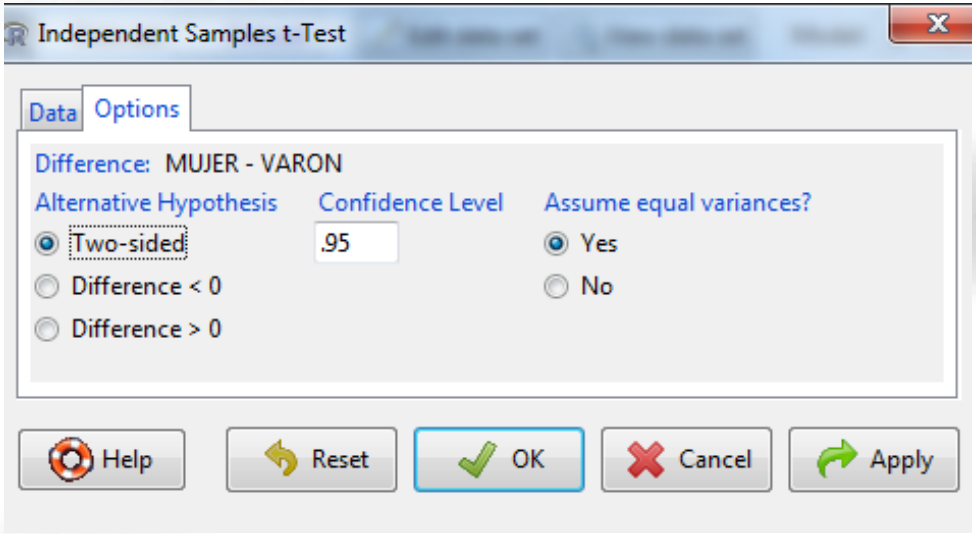
```
> with(Dataset, tapply(tad1, sexo, var, na.rm=TRUE))
    MUJER     VARON
 82.50751 138.24111

> leveneTest(tad1 ~ sexo, data=Dataset, center="median")
Levene's Test for Homogeneity of Variance (center = "median")
      Df F value Pr(>F)
group  1  1.3506 0.2499
      58
```

- P value is over 0.05
- We can assume homogeneity of variances

T-test when variances are equal







T-test when variances are equal


```
> t.test(tad1~sexo, alternative='two.sided', conf.level=.95, var.equal=TRUE,
+    data=Dataset)

        Two Sample t-test

data:  tad1 by sexo
t = 0.3543, df = 58, p-value = 0.7244
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -4.453505  6.368899
sample estimates:
mean in group MUJER mean in group VARON
           88.78378            87.82609
```

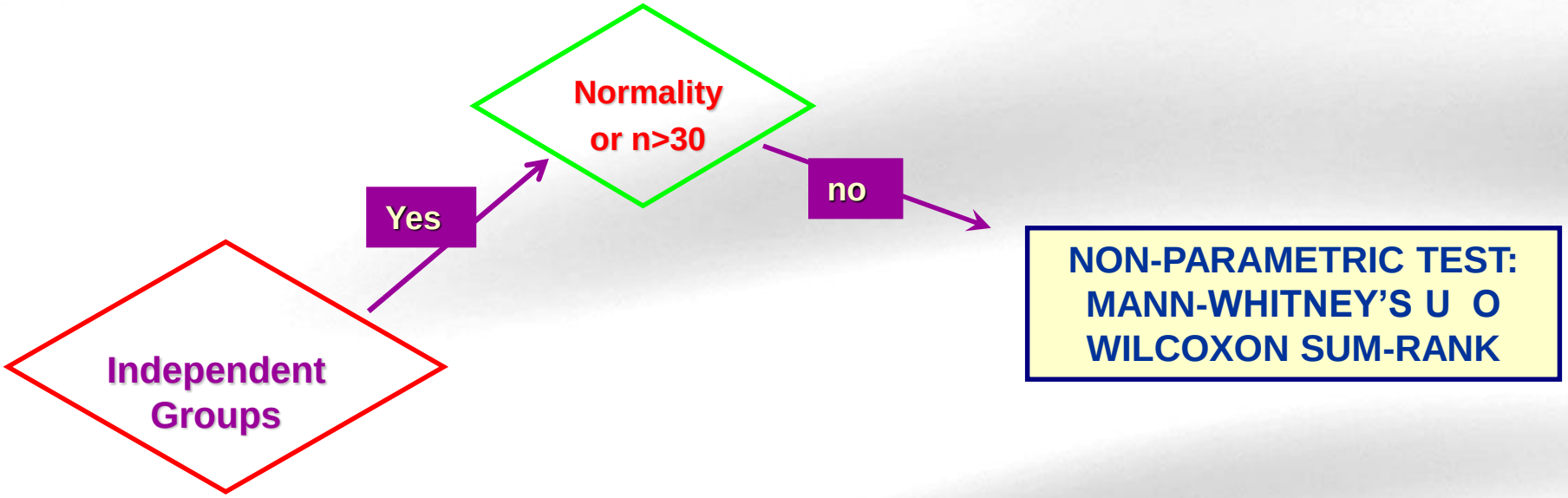
Differences are not statistically significant

T-test when variances are not equal (Welch)


```
> t.test(tad1~sexo, alternative='two.sided', conf.level=.95, var.equal=FALSE,
+    data=Dataset)

        Welch Two Sample t-test

data:  tad1 by sexo
t = 0.3336, df = 38.144, p-value = 0.7405
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -4.852834  6.768228
sample estimates:
mean in group MUJER mean in group VARON
           88.78378            87.82609
```

Differences are not statistically significant

Two groups, data non normal



Non parametric tests

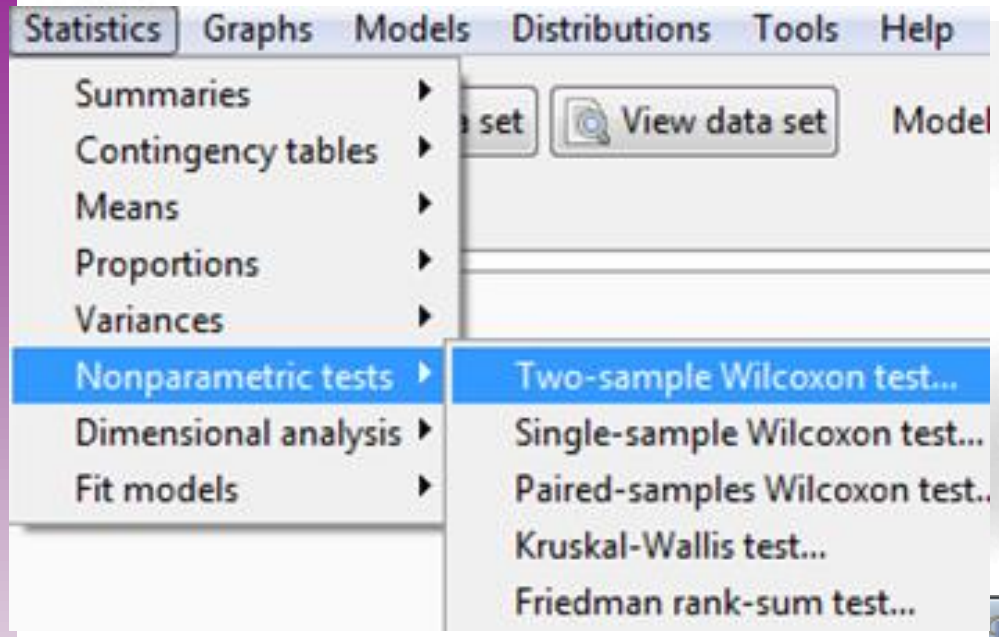
- If data distribution is unknown or mean is not the best way to summarize data …
  - Non parametric test are not based on the usual parameters from a distribution, such as μ or σ^2.
  - Instead they may be based …
    - On order statistics, such as median or percentiles
  - They take into account the whole distribution.

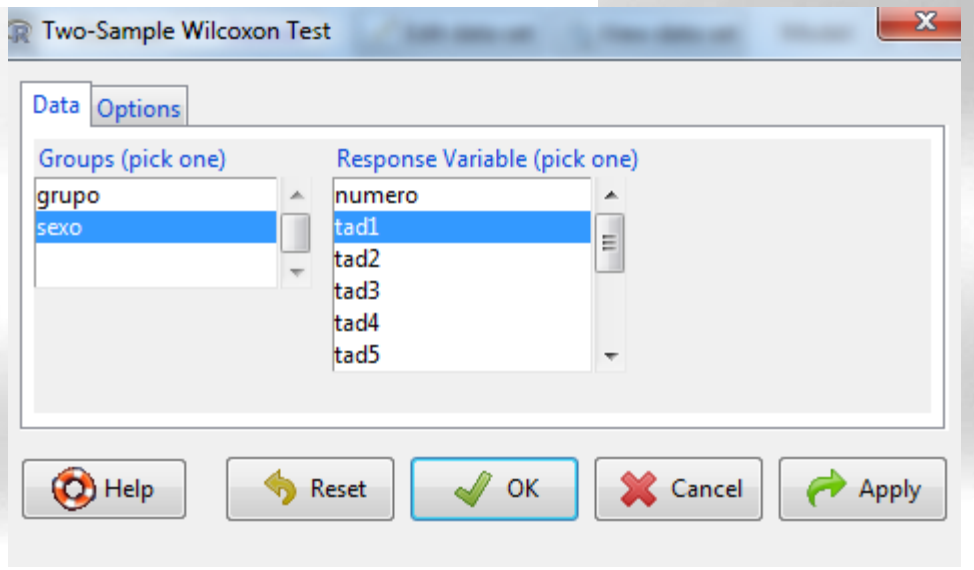
Test based on ranks(Wilcoxon)

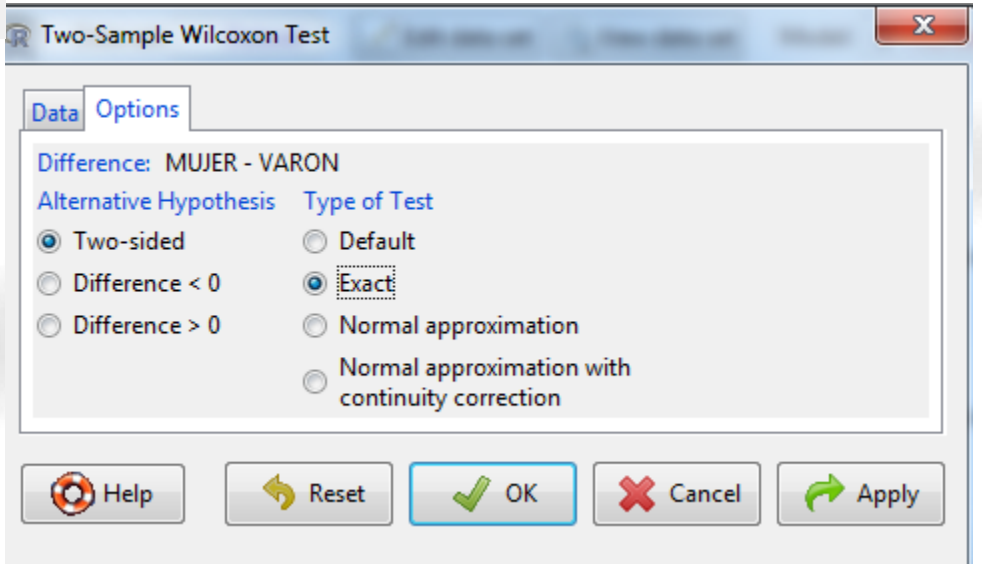
- Based on susbtituting original values by "ranks" in a joint sample
  - 12, 5, 14, 16, 3 → ranks are: 3, 2, 4, 5, 1
- Ranks only depend on the position of each value in the ordered sample.
  - 120, 95, 121, 130, 3 have the same ranks as values in the first sample

- ☺ NP test are more robust than parametrics ones
- ☹ In the ideal situation where parametric tests are valid they are considered to be preferable.

Mann Whitney's U (Wilcoxon Rank-sum)







Mann Whitney's U (Wilcoxon Rank-sum)


```
> wilcox.test(tad1 ~ sexo, alternative='two.sided', exact=TRUE, correct=FALSE, data=Dataset)

        Wilcoxon rank sum test

data:  tad1 by sexo
W = 434, p-value = 0.8955
alternative hypothesis: true location shift is not equal to 0


> with(Dataset, tapply(tad1, sexo, median, na.rm=TRUE))
MUJER VARON
   90    90
```

Differences are not significant

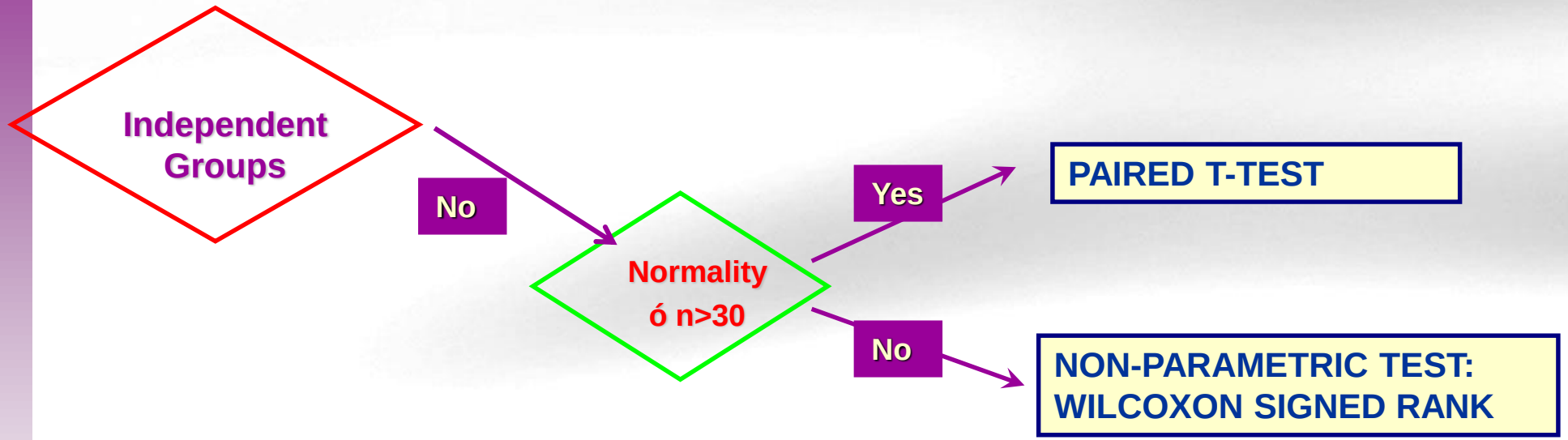
Questions to answer

- Are samples comparable at baseline time?
- Is blood pressure comparable between first and 12th measures?

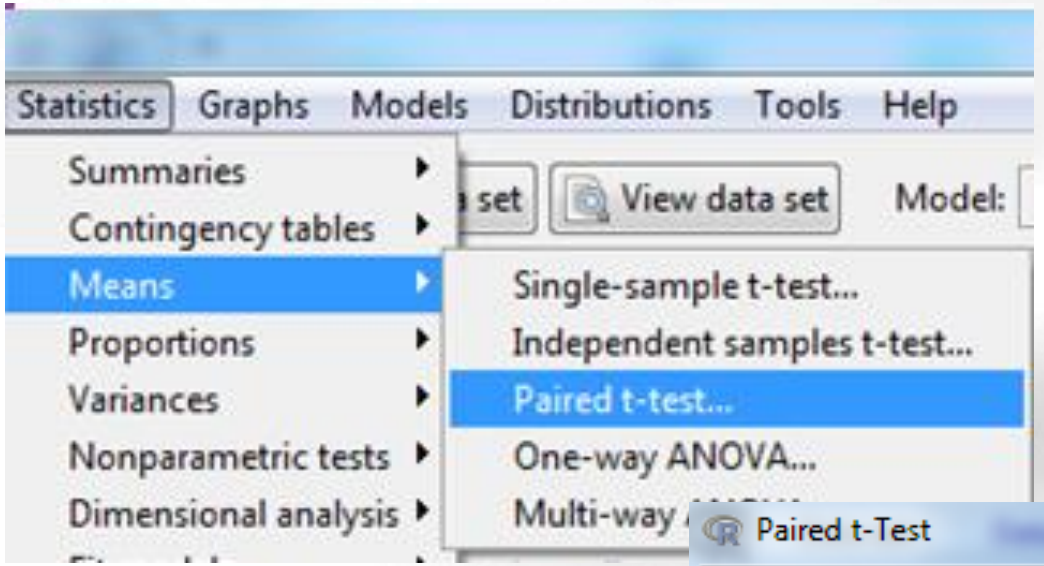
Syllabus

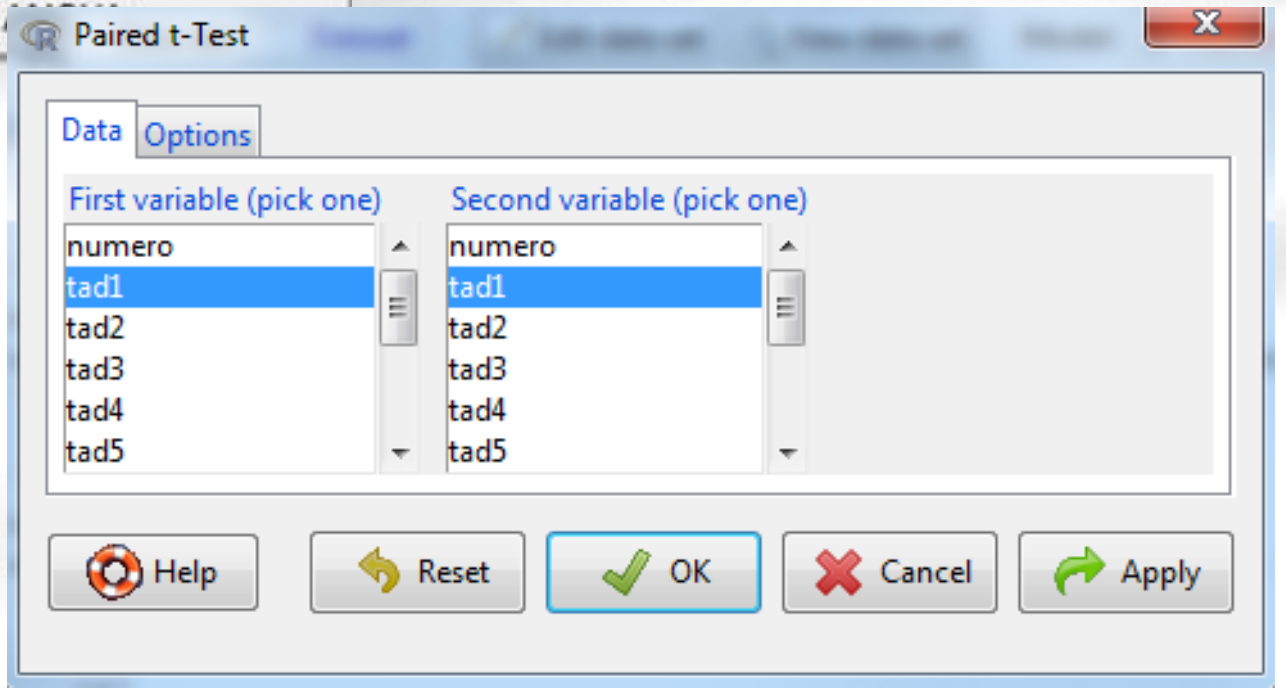
1. INTRODUCTION
2. TYPES OF TESTS
3. NORMALITY TESTS
4. **ONE GROUP** COMPARISON
5. **TWO GROUPS** COMPARISON IN INDEPENDENT SAMPLES
6. **TWO GROUPS** COMPARISON IN DEPENDENT SAMPLES
7. **MORE THAN TWO GROUPS** COMPARISON IN INDEPENDENT SAMPLES
8. MULTIPLE COMPARISONS AND MULTIPLE TESTING

Two dependent groups



Paired t-test



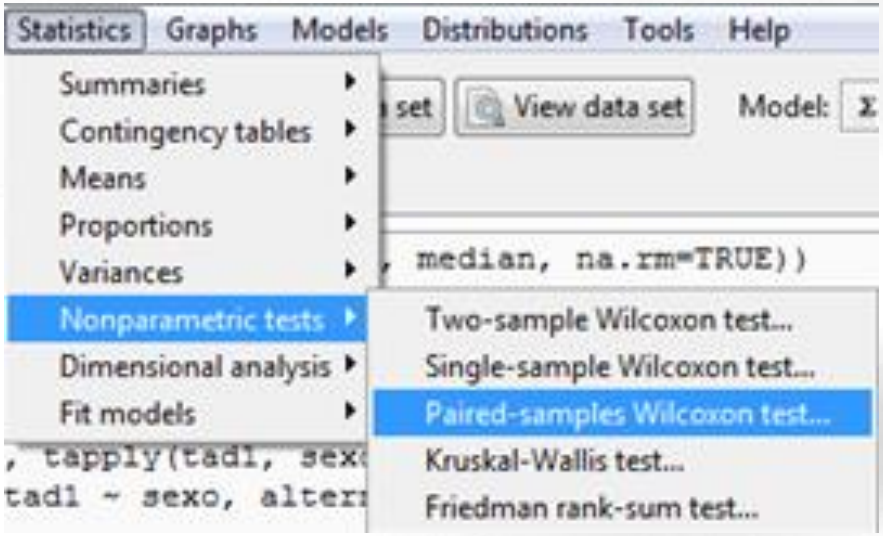


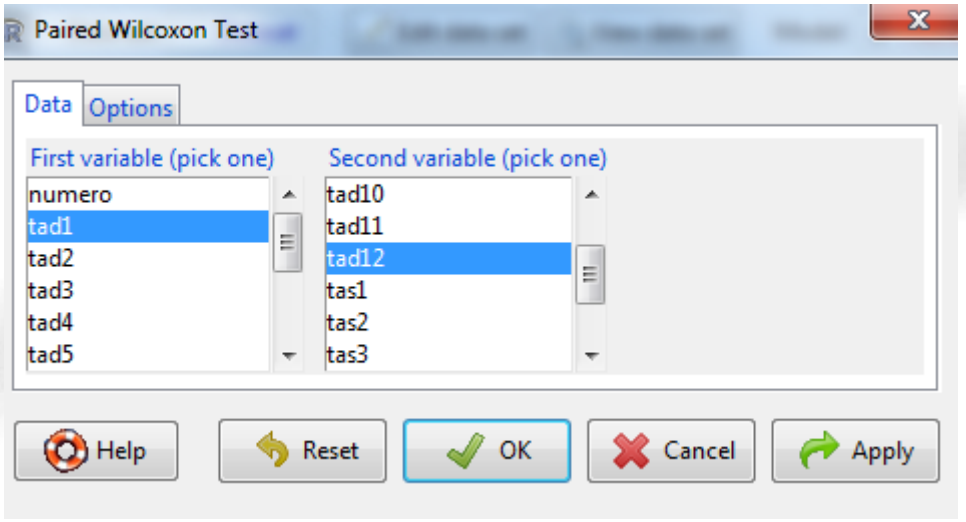
Paired t-test

```
> with(Dataset, (t.test(tad1, tad12, alternative='two.sided', conf.level=.95, paired=TRUE)))

        Paired t-test

data:  tad1 and tad12
t = 1.8507, df = 51, p-value = 0.07001
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.2364274  5.8133505
sample estimates:
mean of the differences
               2.788462
```

Differences are not significant

Wilcoxon signed-rank test





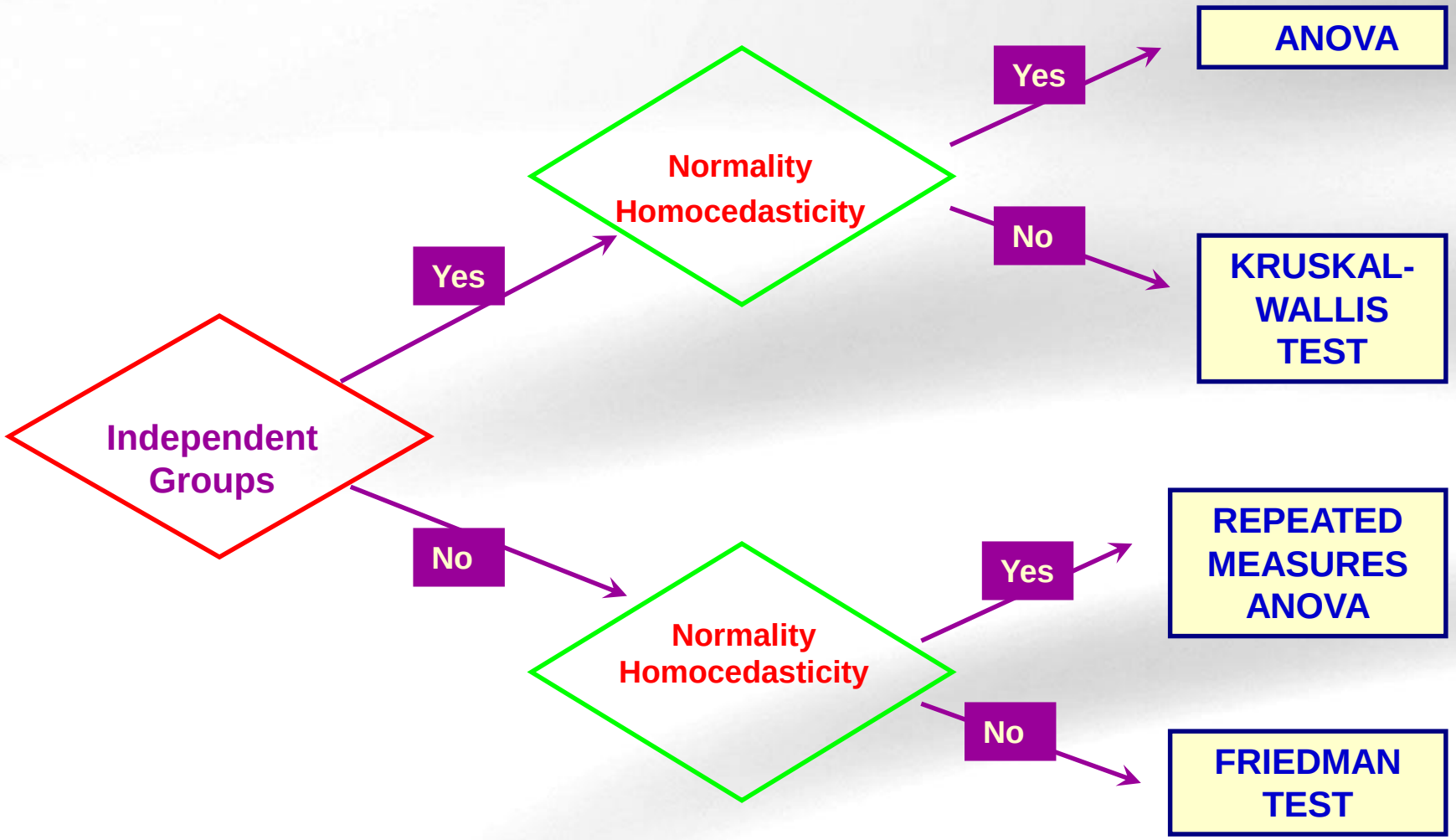
Wilcoxon signed-rank test


```
> with(Dataset, wilcox.test(tad1, tad12, alternative='two.sided', exact=TRUE, paired=TRUE))

        Wilcoxon signed rank test with continuity correction

data:  tad1 and tad12
V = 478.5, p-value = 0.05333
alternative hypothesis: true location shift is not equal to 0
```

Syllabus

1. INTRODUCTION
2. TYPE OF TEST
3. NORMALITY TEST
4. **ONE GROUP** COMPARISON
5. **TWO GROUPS** COMPARISON IN INDEPENDENT SAMPLES
6. **TWO GROUPS** COMPARISON IN DEPENDENT SAMPLES
7. **MORE THAN TWO GROUPS** COMPARISON IN INDEPENDENT SAMPLES
8. MULTIPLE COMPARISONS AND MULTIPLE TESTING

Three or more groups



Analysis of the variance

Null Hypothesis

The means of all population are equal

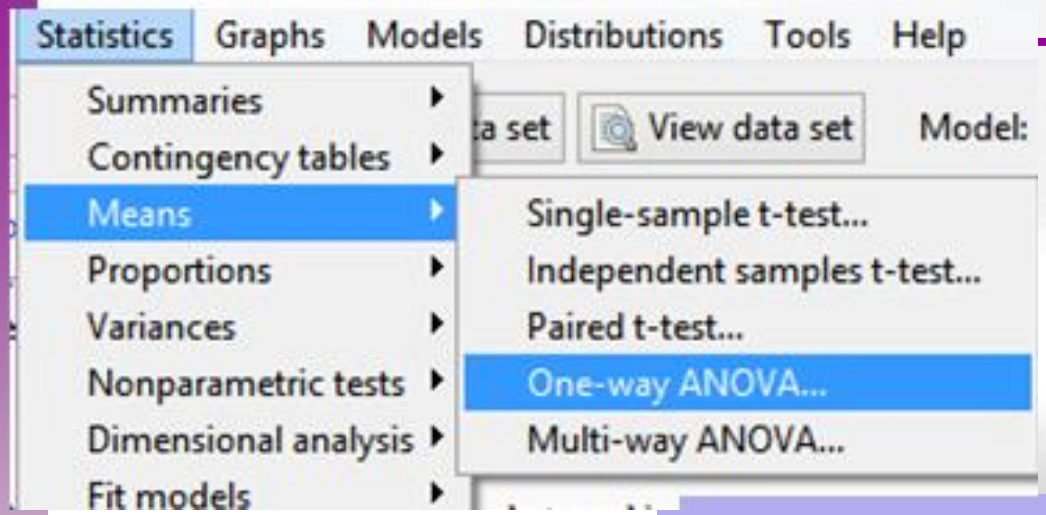
$$H_0 \quad \mu_1 = \mu_2 = \ldots = \mu_k$$

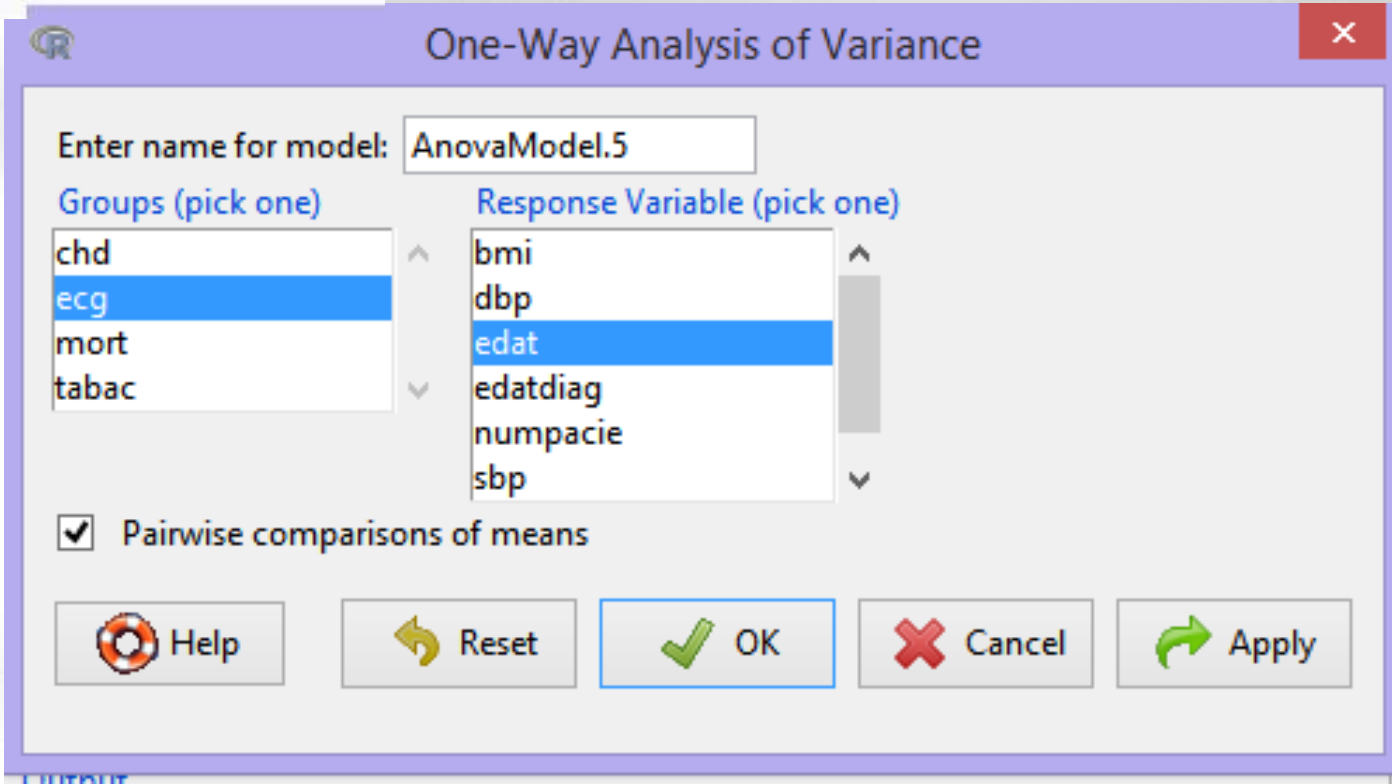
Alternative Hypothesis

Not all the means are equal. At least there are two different means

$$H_a \quad \exists\, i,j \;\; \mu_i \neq \mu_j$$

Anova in R Commander





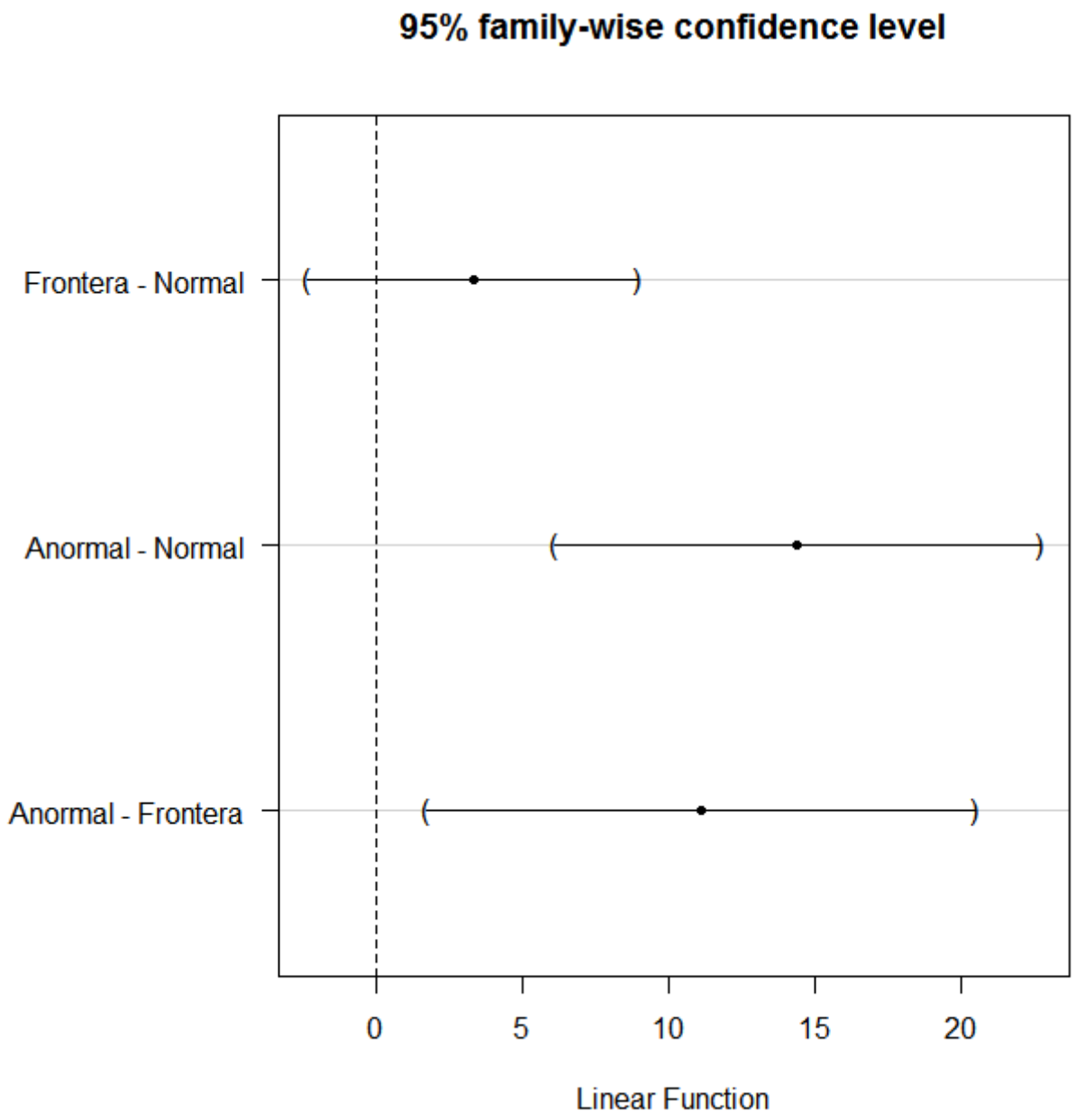
```
> AnovaModel.4 <- aov(edat ~ ecg, data=diabetes)
> summary(AnovaModel.4)
```

```
             Df Sum Sq Mean Sq F value  Pr(>F)
ecg           2   2166  1083.0   8.619 0.00029 ***
Residuals   146  18347   125.7
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
> with(diabetes, numSummary(edat, groups=ecg, statistics=c("mean", "sd")))
```

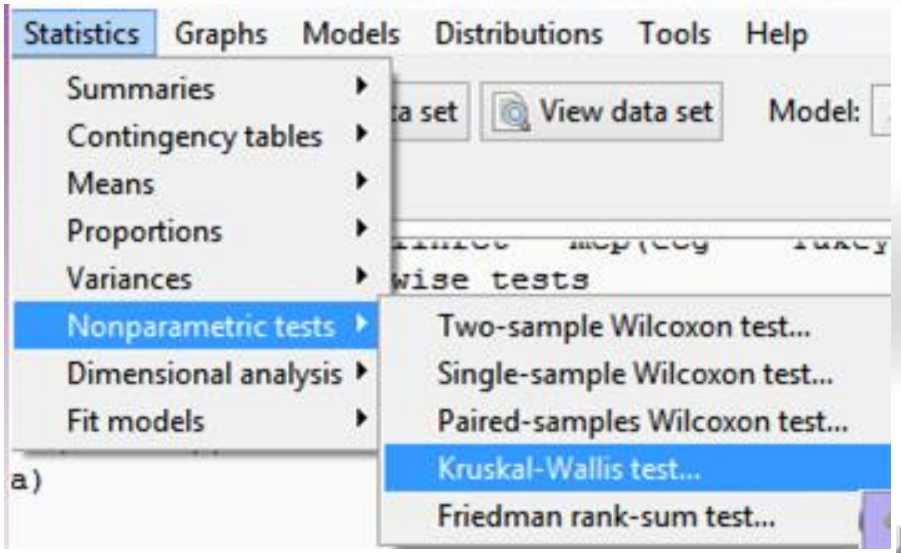
```
             mean       sd data:n
Normal   50.50450 11.492981    111
Frontera 53.81481 11.368097     27
Anormal  64.90909  6.759505     11
```

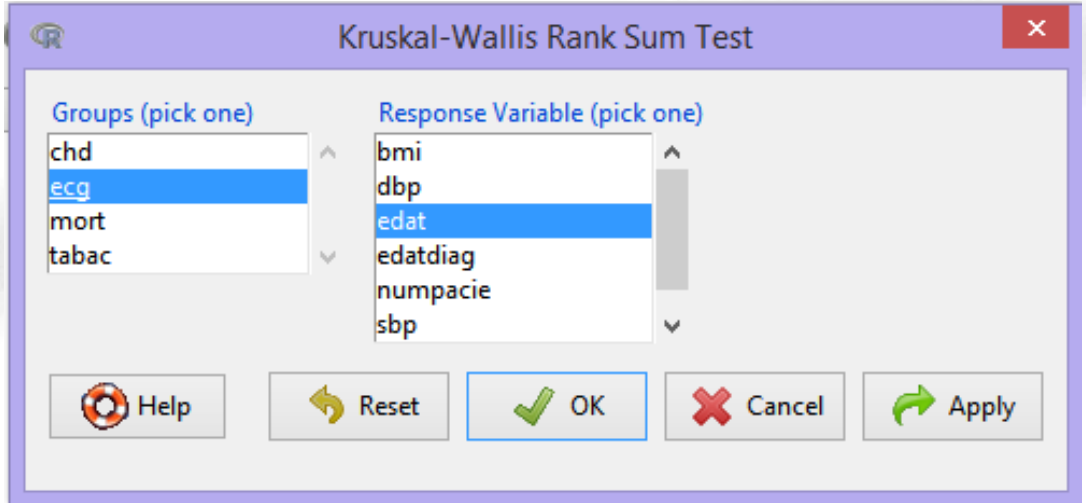
```
     Simultaneous Tests for General Linear Hypotheses

Multiple Comparisons of Means: Tukey Contrasts


Fit: aov(formula = edat ~ ecg, data = diabetes)

Linear Hypotheses:
                       Estimate Std. Error t value Pr(>|t|)
Frontera - Normal == 0    3.310      2.405   1.376 0.345713
Anormal - Normal == 0    14.405      3.543   4.065 0.000217 ***
Anormal - Frontera == 0  11.094      4.010   2.767 0.016472 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Adjusted p values reported -- single-step method)
```


95% family-wise confidence level
Frontera - Normal
Anormal - Normal
Anormal - Frontera
0
5
10
15
20
Linear Function

Kruskal-Wallis test

- Is the non-parametric versión of ANOVA based on ranks





Kruskal-Wallis Test

```
> with(diabetes, tapply(edat, ecg, median, na.rm=TRUE))
```

```
  Normal Frontera  Anormal
      49       53       64
```

```
> kruskal.test(edat ~ ecg, data=diabetes)
```

```
    Kruskal-Wallis rank sum test

data:  edat by ecg
Kruskal-Wallis chi-squared = 17.4826, df = 2, p-value = 0.0001598
```

Syllabus

1. INTRODUCTION
2. TYPE OF TEST
3. NORMALITY TEST
4. **ONE GROUP** COMPARISON
5. **TWO GROUPS** COMPARISON IN INDEPENDENT SAMPLES
6. **TWO GROUPS** COMPARISON IN DEPENDENT SAMPLES
7. **MORE THAN TWO GROUPS** COMPARISON IN INDEPENDENT SAMPLES
8. MULTIPLE COMPARISONS AND MULTIPLE TESTING

Multiple Comparisons and múltiple testing

Testing hypothesis repeatedly

- Every time we do a test there is a chance to take the wrong decision by rejecting the null hypothesis while it is TRUE.
- If, instead, we do many tests simultaneously the probability that there is, by chance, at least one false positive increases and does not match the type I error probability anymore.
- This increase in the probability of type I error has to be compensated in some way → **multiple testing adjustments**

To cross or not to cross?

The previous situation can be better understood with the “bridge analogy”.

Imagine you are an adventurer that has the option of to cross a bridge in order to escape from danger, find a treasure...

and that there is a post in front of the bridge stating:

“This bridge has broken only one out of 100 times"

To cross or not to cross?

Imagine you are an adventurer that has the option of to cross a bridge in order to escape from danger, find a treasure...

and that there is a post in front of the bridge stating:

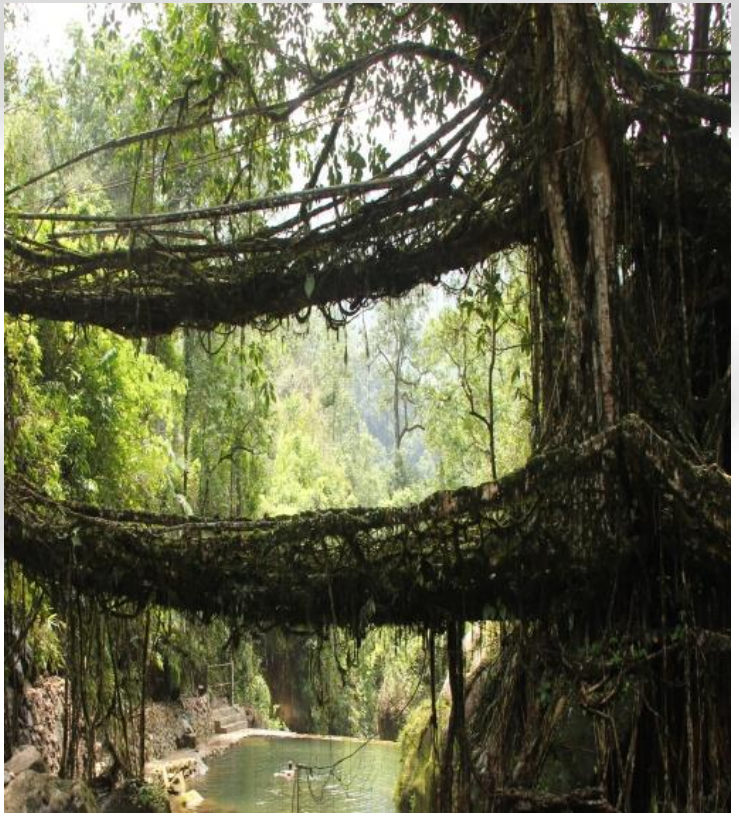
"This bridge has broken only one out of 100 times"

So, the p-value of our metaphor is 0.01

You could accept that **1%** is **a risk small** enough to **pass the bridge** and pursue your goal. OK

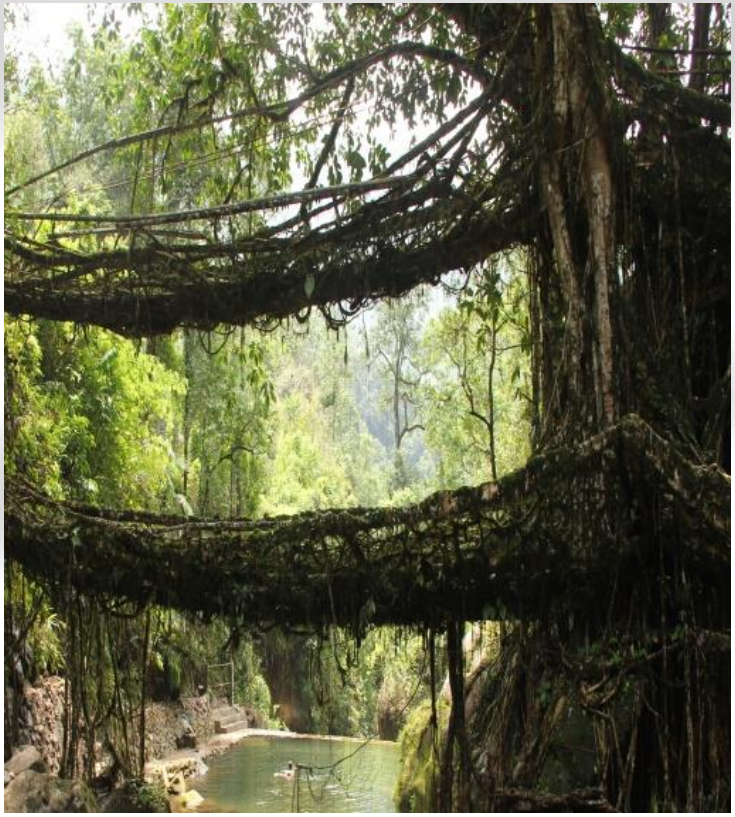
To cross or not to cross?

But... what do you decide if, in order to reach your goal, you have to cross hundreds of bridges of that kind?



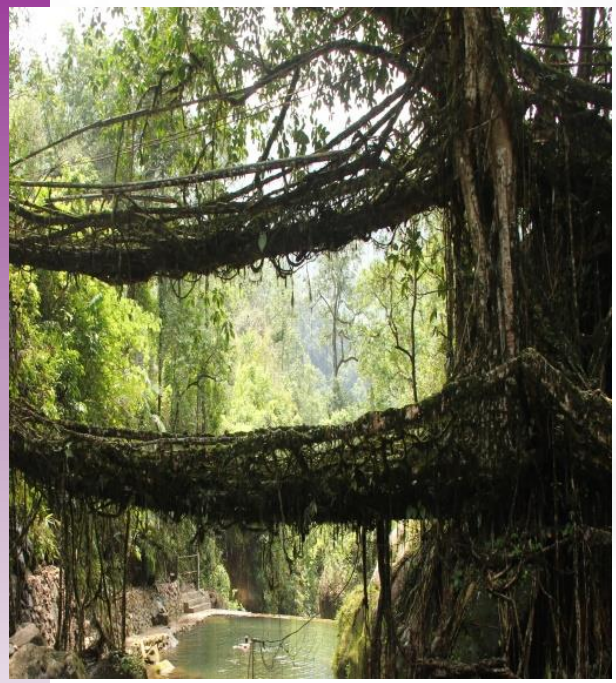
To cross or not to cross?

But... what do you decide if, in order to reach your goal, you have to cross hundreds of bridges of that kind?



•In this case, the probability of falling while crossing one of the bridges is obviously too high ('cause we have just one life).

To cross or not to cross?



Therefore, in this case (multiple testing), the p-value by itself is not a good reference for accepting or not statistical significance.

We must apply some type of adjustment to the p-values (allowing us to be safe in crossing all the bridges).

Some p-value adjustments

- Bonferroni (α/k)
- Post-Hoc test ANOVA (Tukey, Scheffe, Dunn-test)
- False Discovery rate
- Benjamini-Hochberg correction

Multiple comparisons vs multiple testing

- There are two distinct situations where p-value adjustment may be necessary:
  - Post-hoc tests in ANOVA:
    - This is usually called multiple comparisons and common methods of adjustment are Tukey, Fisher HSD.
  - Testing many variables in the same study
    - This is usually called multiple testing and common methods of adjustment are Bonferroni, Holm or Benjamini and Hochberg (False Discovery Rate).

Multiple testing

- When many variables are compared independently with the same test
  - Find differences between treated/untreated for a set of biomarkers such as cytokines.
    - Number of comparisons may be low ("dozens")
  - Find differentially expressed genes, i.e. genes whose expression may change between conditions.
    - Number of comparisons high ("hundreds" to "thousands")
- This is usually called multiple testing and common methods are Bonferroni, Holm or Benjamini and Hochberg (False Discovery Rate).

Post-hoc ANOVA tests

- If we wish to compare all means against all means he number of tests increases quickly (to compare all pairs of means if there are k groups (k*k-1)/2 tests are required).
- This is usually called **multiple comparisons** and common methods of adjustment are Tukey, Fisher HSD or Bonferroni.

Common misunderstandings about the p-value

Common misunderstandings about the p-value

- The p-value is **not** the probability that the null hypothesis is true, nor it is the probability that the alternative hypothesis is false (it is not connected to either of these).
- The p-value **cannot** be used to figure out the probability of a hypothesis being true.
- The p-value is **not** the probability of wrongly rejecting the null hypothesis.
- The p-value is **not** the probability that replicating the experiment would yield the same conclusion.
- The p-value does **not** indicate the size or importance of the observed effect. The two do vary together however: the larger the effect (effect size), the smaller sample size will be required to get a significant p-value.