Analysing and presenting data: **practical hints**

Giorgio MATTEI

giorgio.mattei@centropiaggio.unipi.it

Course: Meccanica dei Tessuti Biologici
Date: 17 May 2016

Equal or different?

The case of ***two samples***

Independent two-sample *t-test*

Equal *sample sizes (**n**),* ***equal*** *variances ($S_{X_1X_2}$)*

The ***t* statistic** to test whether the **means** of **group 1 ($\overline{X_1}$)** and **group 2 ($\overline{X_2}$) are different** can be calculated as follows:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{S_{X_1X_2} \cdot \sqrt{\frac{2}{n}}}$$

$$S_{X_1X_2} = \sqrt{\frac{1}{2}(S_{X_1}^2 + S_{X_2}^2)}$$

«pooled» standard deviation

t-test DoFs = 2n - 2

H$_0$ is refused with a **significance level α if**
$t < -t_{DoF,\alpha}$ or $t > t_{DoF,\alpha}$

Independent two-sample *t-test*

__Unequal__ sample sizes (__n__$_1$ and __n__$_2$), __equal__ variances ($S_{X_1X_2}$)

The ***t* statistic** to test whether the **means** of **group 1** ($\overline{X_1}$) and **group 2** ($\overline{X_2}$) **are different** can be calculated as follows:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{S_{X_1X_2} \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \qquad S_{X_1X_2} = \sqrt{\frac{(n_1 - 1)S_{X_1}^2 + (n_2 - 1)S_{X_2}^2}{n_1 + n_2 - 2}}$$

«pooled» standard deviation

t-*test* DoFs = $n_1 + n_2 - 2$

H_0 is refused with a **significance level α if**
$t < -t_{DoF,\alpha}$ or $t > t_{DoF,\alpha}$

Independent two-sample *t-test*

Unequal *sample sizes (**n$_1$** and **n$_2$**),* ***unequal*** *variances ($S_{X_1X_2}$)*

The ***t*** **statistic** to test whether the **means** of **group 1** ($\overline{X_1}$) and **group 2** ($\overline{X_2}$) **are different** can be calculated as follows:

$$t = \frac{\overline{X}_1 - \overline{X}_2}{s_{\overline{X}_1 - \overline{X}_2}} \qquad s_{\overline{X}_1 - \overline{X}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

«unpooled» standard deviation

$$t\text{-}test \text{ DoFs} = \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{(s_1^2/n_1)^2/(n_1-1) + (s_2^2/n_2)^2/(n_2-1)}$$

Welch–Satterthwaite equation

H$_0$ is refused with a **significance level α if**

$t < -\, t_{DoF,\alpha}$ **or** $t > t_{DoF,\alpha}$

Independent two-sample *t-test* (*unequal sample sizes* and *equal variances*): an *example*

- Two groups of 10 *Dapnia magna* eggs, randomly extracted from the same clone, were reared in two different concentrations of hexavalent chromium
- After a month survived individuals were measured: 7 in group A and 8 in group B

	A	B
	2,7	2,2
	2,8	2,1
	2,9	2,2
	2,5	2,3
	2,6	2,1
	2,7	2,2
	2,8	2,3
		2,6
Mean	2.714	2.250

$$s_p^2 = \frac{0{,}10825 + 0{,}18000}{6 + 7} = 0{,}022173$$ «pooled» variance

$$t_{13} = \frac{2{,}714 - 2{,}250}{\sqrt{0{,}022173 \cdot \left(\frac{1}{7} + \frac{1}{8}\right)}} = 6{,}02$$ **t** with 13 DoF

Since $\boldsymbol{t_{13,\,0.05}}$ **= 2.160 what can we say?**

MATLAB

Independent two-sample t-test

H = 0, H_0 cannot be refused at **α**
H = 1, refuse H_0 at α

Confidence interval for the «true» **difference** of **population means**

Data **structure** containing **t-statistics value** and **number of DoF**

significance level

[H,P,CI,STATS] = TTEST2(X,Y,alpha,tail,vartype)

samples

p-value *(i.e. the probability of observing the given result, or one more extreme, by chance if the null hypothesis is true)*

'equal' or **'unequal'**

'both' → “means are not equal" (two-tailed test)
'right' → "$\overline{X}$ is greater than $\overline{Y}$" (right-tailed test)
'left' → "$\overline{X}$ is less than $\overline{Y}$" (left-tailed test)

MATLAB
Ind. 2-sample t-test: an example

```
>> X=[2.7 2.8 2.9 2.5 2.6 2.7 2.8]';
>> Y=[2.2 2.1 2.2 2.3 2.1 2.2 2.3 2.6]';
>> [H,P,CI,STATS] = ttest2(X,Y,0.05,'both','equal')
```

Dependent two-sample *t-test*

one sample tested twice or *two "paired" samples*

$$t = \frac{\overline{X}_D - \mu_0}{s_D / \sqrt{n}}$$

- ✓ Calculate the differences between all ***n*** pairs (X_D), then substitute their average ($\overline{X_D}$) and standard deviation (s_D) in the equation above to test if the average of the differences is significantly different from $\boldsymbol{\mu_0}$ ($\boldsymbol{\mu_0}$ **= 0** under $\mathbf{H_0}$**, DoFs =** ***n*** **- 1**)
- ✓ The "pairs" can be either one person's pre-test and post-test scores (repeated measures) or persons matched into meaningful groups (e.g. same age)

Example of repeated measures			
Number	**Name**	**Test 1**	**Test 2**
1	Mike	35%	67%
2	Melanie	50%	46%
3	Melissa	90%	86%
4	Mitchell	78%	91%

Example of matched pairs			
Pair	**Name**	**Age**	**Test**
1	John	35	250
1	Jane	36	340
2	Jimmy	22	460
2	Jessy	21	200

Dependent two-sample *t-test*: an example

Student	Pre-module score	Post-module score	Difference
1	18	22	+4
2	21	25	+4
3	16	17	+1
4	22	24	+2
5	19	16	-3
6	24	29	+5
7	17	20	+3
8	21	23	+2
9	23	19	-4
10	18	20	+2
11	14	15	+1
12	16	15	-1
13	16	18	+2
14	19	26	+7
15	18	18	0
16	20	24	+4
17	12	18	+6
18	22	25	+3
19	15	19	+4
20	17	16	-1

$$t = \frac{2.05}{0.634} = 3.231 \qquad \text{on 19 df}$$

Since $\boldsymbol{t_{19,\,0.05}}$ **= 2.093 what can we say?**

MATLAB

Dependent two-sample t-test

H = 0, H_0 cannot be refused at **α**
H = 1, refuse H_0 at α

Confidence interval for the «true» **difference** of **population means**

Data **structure** containing **t-statistics value** and **number of DoF**

significance level

[H,P,CI,STATS] = TTEST(X,Y,alpha,tail)

samples

p-value *(i.e. the probability of observing the given result, or one more extreme, by chance if the null hypothesis is true)*

'both' → “means are not equal" (two-tailed test)
'right' → "$\overline{X}$ is greater than $\overline{Y}$" (right-tailed test)
'left' → "$\overline{X}$ is less than $\overline{Y}$" (left-tailed test)

MATLAB

Dep. 2-sample t-test: an example

```
>> X=[22 25 17 24 16 29 20 23 19 20 15 15 18 26 18 24 18 25 19 16]';
>> Y=[18 21 16 22 19 24 17 21 23 18 14 16 16 19 18 20 12 22 15 17]';
>> [H,P,CI,STATS] = ttest(X,Y,0.05,'both')
```

Equal or different?
more than ***two samples***

ANalysis Of VAriance (ANOVA)

- **More than 2** groups: **NO pairwise comparisons** (*t-test*)

  ↑ groups → ↑ overall probability that at least one of them is significant

  (e.g. α=0.05 and n=20 → in average 1 group will be significantly different for the case, even if H_0 is true)

$$H_0: \mu_1 = \mu_2 = \mu_3 = \ldots = \mu_k$$

H_1: not all means are equal

all means **are different**

...

one mean **is different** from the others, which are all equals

- **ANOVA**
  - uses **Fisher's distribution** (**F-**distribution)
  - the **sources of variations** on observed values of **two or more groups** can be **decomposed** and **accurately measured**
  - the **source of variation** is called **EXPERIMENTAL FACTOR** (or **TREATMENT**) and can be multi-levelled
  - each **unit or observation** of the experimental factor is called **REPLICATION**

one-way ANOVA: an example
The problem

- Content of iron in air in 3 different zones (A, B, C) of a city (μg/N mc at 0 °C and 1013 mbar)

	EXPERIMENTAL FACTOR				
	A	B	C		
	2,71	1,75	2,22		
	2,06	2,19	2,38		
	2,84	2,09	2,56		
	2,97	2,75	2,60		
	2,55		2,72		
	2,78				
$\sum X_j$	15,91	8,78	12,48	$\sum X$	37,17
$\mathbf{n_i}$	6	4	5	**n**	15
$\overline{X}_{\cdot j}$	2,652	2,195	2,496	$\overline{X}_{\cdot\cdot}$	2,478

MATLAB
one-way ANOVA

***p*-value for H_0** (means of the groups are equal)

ANOVA table values

Structure of statistics useful **for performing a multiple comparison of means** with the **MULTCOMPARE** function

[P,ANOVATAB,STATS] = anova1(X,GROUP,DISPLAYOPT)

X:

Matrix with **1 group per column** (requires equal-sized samples)

Vector of data

GROUP:

Character array: one row per column of X, containing the **group names**

Vector: one group name for **each element of X**

DISPLAYOPT:

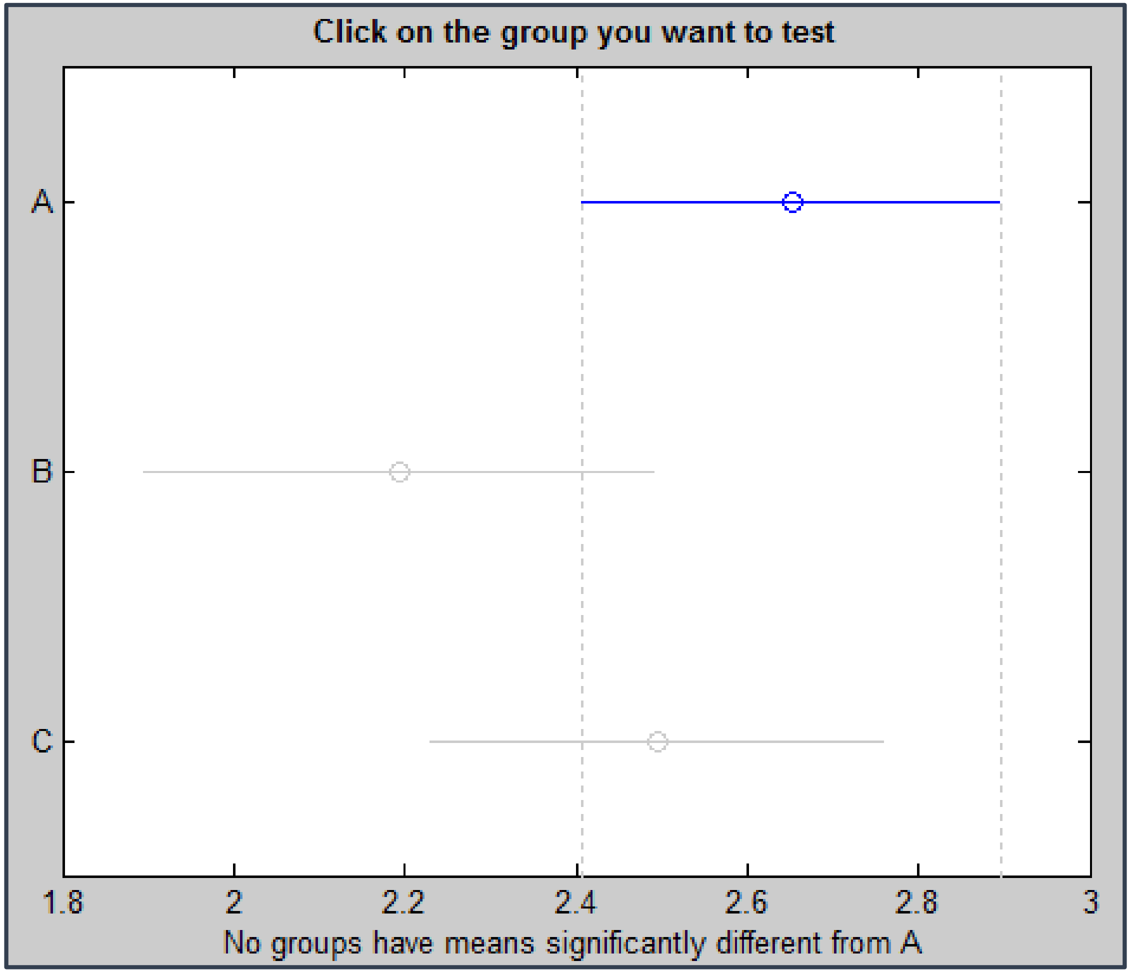
'on' (the default) to **display figures containing a standard one-way anova table and a boxplot**, or 'off' to omit these displays

MATLAB
one-way ANOVA: example

```
>> X=[2.71,2.06,2.84,2.97,2.55,2.78,1.75,2.19,2.09,2.75,2.22,2.38,2.56,2.6,2.72]';
>> GROUP=['A','A','A','A','A','A','B','B','B','B','C','C','C','C','C']';
>> [P,ANOVATAB,STATS] = anova1(X,GROUP)
```

MATLAB
one-way ANOVA: example

COMPARISON = multcompare(STATS)



two-way ANOVA: an example
The problem

- **Content** of **Pb** in air in **5 different urban zones** revealed every **6 hours** during the day

	TREATMENTS (urban zone)					X_{ij}	
BLOCKS (time)	**1**	**2**	**3**	**4**	**5**	**sums**	**means**
6 am	28	25	30	22	26	131	26,2
12 am	34	32	37	31	30	164	32,8
6 pm	22	21	24	20	19	106	21,2
12 pm	36	31	40	33	29	169	33,8
sums	120	109	131	106	104	570	
means	30,00	27,25	32,75	26,50	26,00		28,50

MATLAB
two-way ANOVA

***p*-value for H_0** (means of the groups are equal)

ANOVA table values

Structure of statistics useful **for performing a multiple comparison of means** with the **MULTCOMPARE** function

[P,ANOVATAB,STATS] = anova2(X,REPS,DISPLAYOPT)

Matrix of data (balanced ANOVA → equal number of repetitions)

Columns: 1st factor
Rows: 2nd factor

REPS indicates the **number** of **observations per "cell"**

A **"cell" contains REPS number** of **rows**

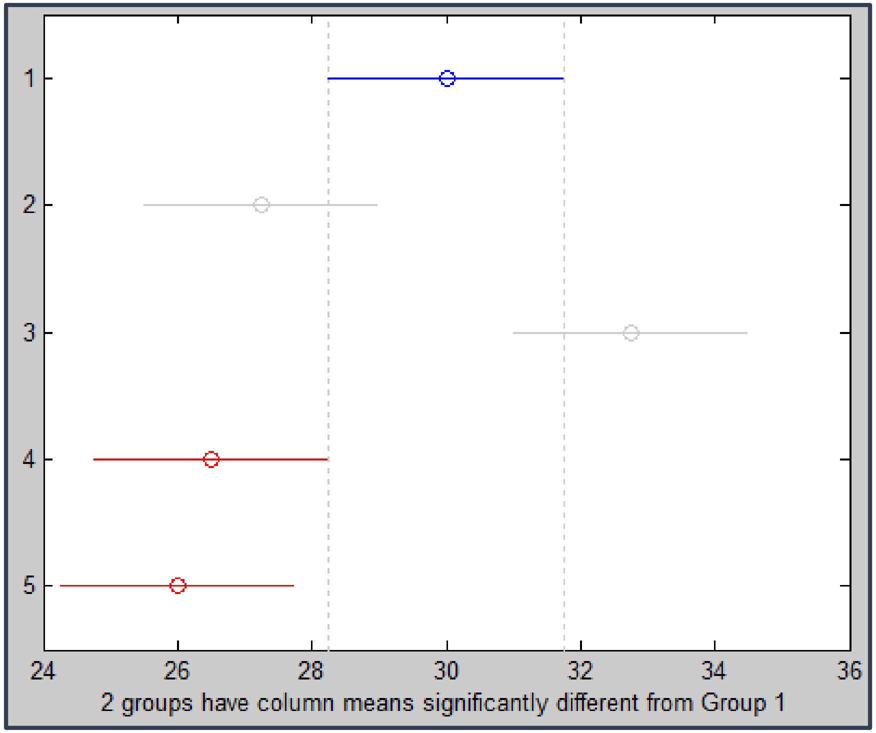
'on' (the default) to **display a standard two-way anova table**, or 'off' to skip the display

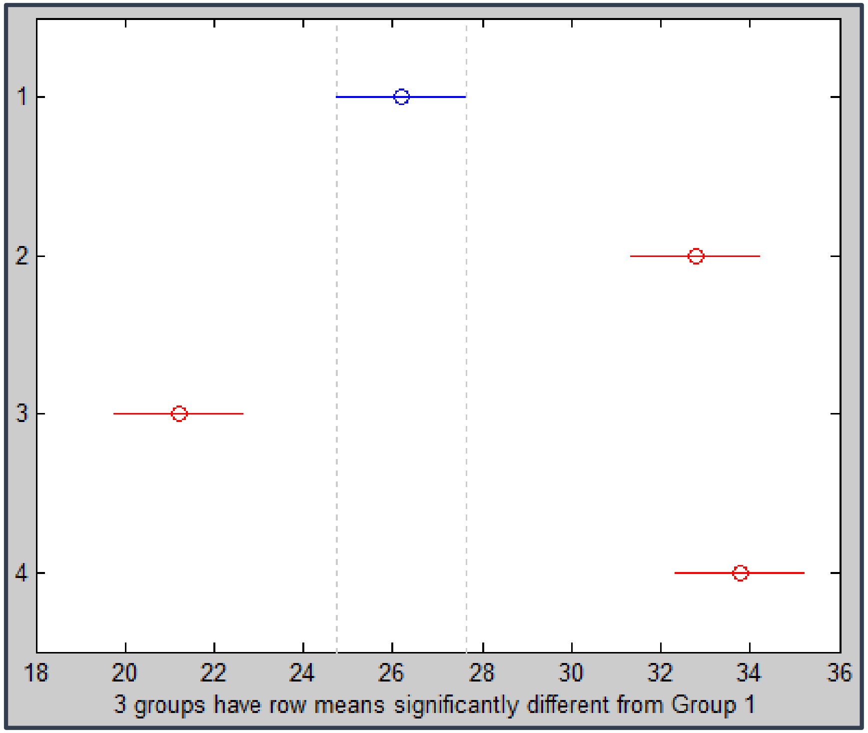
MATLAB
two-way ANOVA: example

```
>> X=[28 25 30 22 26;
34 32 37 31 30;
22 21 24 20 19;
36 31 40 33 29];
>> [P,ANOVATAB,STATS] = anova2(X)
```

MATLAB
one-way ANOVA: example

COMPARISON = multcompare(STATS, 'estimate', 'column' (default) or 'row')





Columns (i.e. **urban zones**)

Rows (i.e. **times**)

MATLAB
anovan: N-way analysis of variance