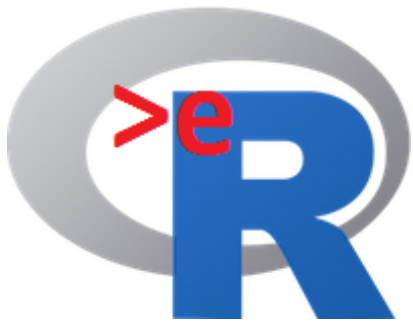




This course was developed as a part of the VLIR-UOS Cross-Cutting project s:

- Statistics: 2011-2016, 2017.
- Statistics: 2017.
- Statistics for development : 2018-2020.



The >eR-Biostat initiative  
Making R based education materials in  
statistics accessible for all

# Basic concepts in statistical modeling using R: The One-way ANOVA model

Developed by

Legesse Kassa Debusho (UNISA), Ziv Shkedy (Hasselt University)



ER-BioStat

Email: [erbiostat@gmail.com](mailto:erbiostat@gmail.com)



<https://github.com/eR-Biostat>



@erbiostat

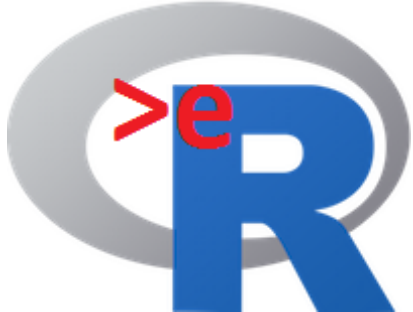


## contents

- One-way ANOVA model:
  - The one-way ANOVA model.
  - Sources of Variability.
  - One-way ANOVA using R: the `aov ( )` function.
  - Model formulation and hypotheses testing.
  - Analysis of the pharmaceutical experiment.
  - Model diagnostic in R: normal probability plot.
  - Multiple testing.

# YouTube tutorials

- YouTube tutorials are available for:
  - Statistics with R: One way ANOVA example (host by Phil Chan): <https://www.youtube.com/watch?v=T0Lmyz6EscY>.
  - ANOVA in R (host by Ed Boone): <https://www.youtube.com/watch?v=Dwd3ha0P8uw>.



# Introduction: The one-way ANOVA model

# A Biopharmaceutical Problem

- A group of 24 rats were randomized into two treatment groups: active drug and placebo
- After the administration of the drug, the rat was placed on a surface, and the distanced traveled by the rat (in meters) was measured.

## The data

```
> print(Biophar)
```

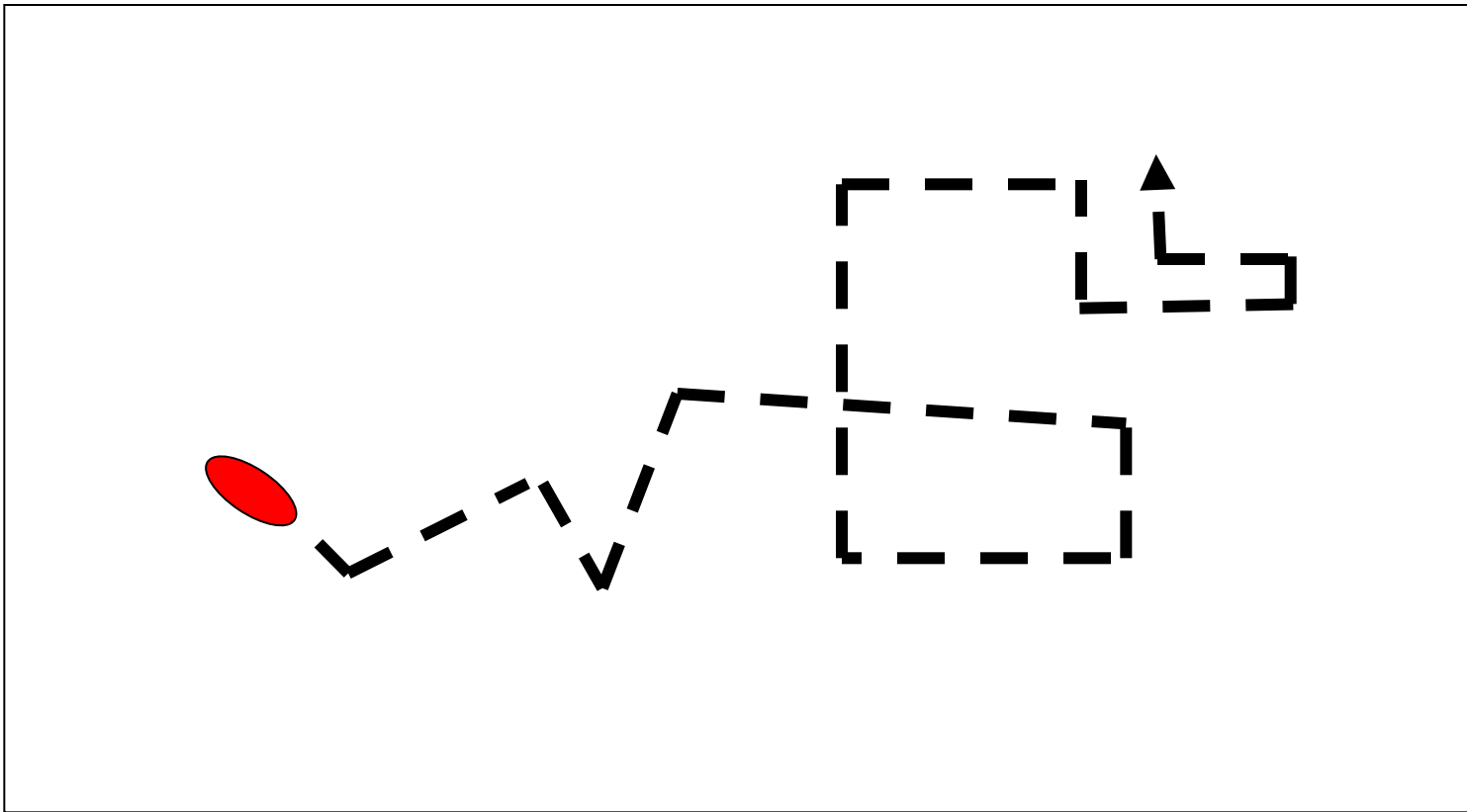
Obs	Treat	Dist
22	QNP	186.6145
11	QNP	103.3529
4	QNP	191.3850
16	QNP	334.9845
7	QNP	89.2831
13	QNP	345.5070
2	QNP	169.5161
20	QNP	173.1491
19	QNP	130.9634
8	QNP	363.4392
10	QNP	76.5340
24	QNP	202.1145
1	SALINE	12.8458
17	SALINE	44.3092
15	SALINE	41.3581
6	SALINE	24.5560
23	SALINE	61.5525
18	SALINE	38.8464
5	SALINE	27.0107
12	SALINE	45.9960
21	SALINE	13.7927
14	SALINE	42.4009
3	SALINE	17.5861
9	SALINE	11.7937

Response

Treatment group

# The Evaluation of the Rat

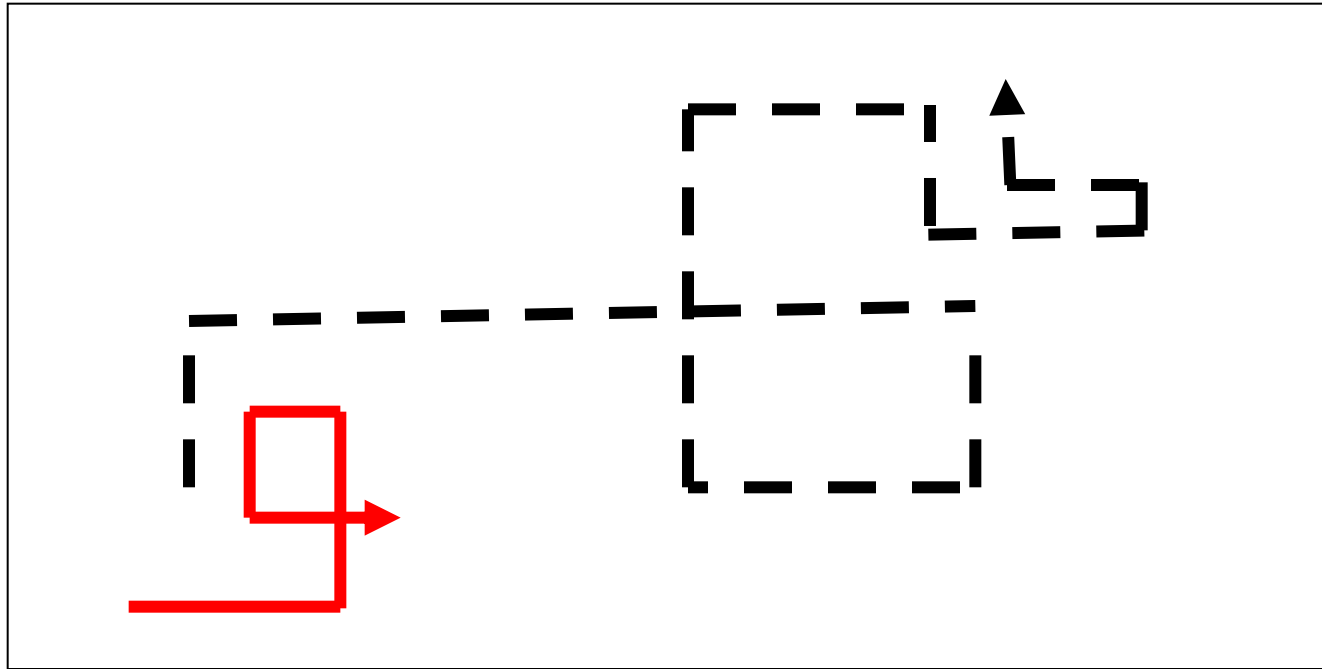
$Y_i$  is the distance traveled by the rat during the experiment.



# Description of the Experiment

Passive rat 

Active rat 



It is assumed that a successful drug increase the distance traveled by the rat.



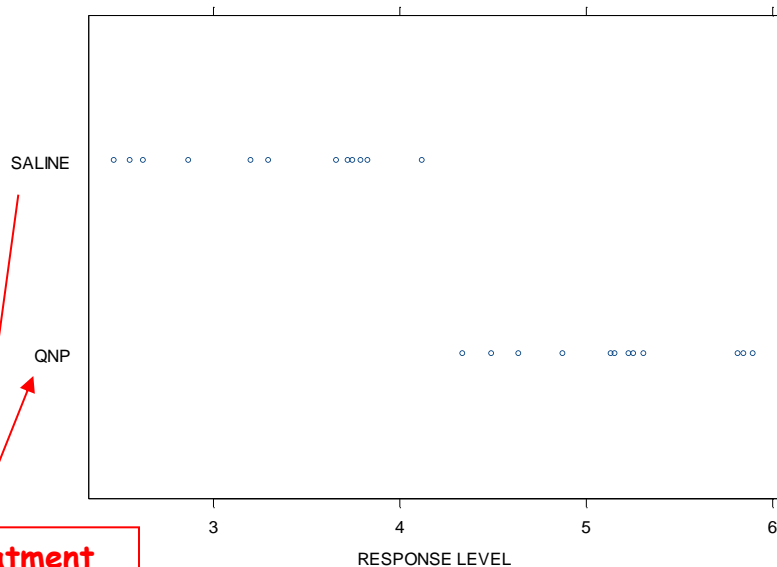
# The Scientific Question

- Does the drug increase the distance traveled by the rat ?

A good drug is expected to improve the rats' performance,  
i.e. to increase the distance travel by the rat

# Graphical display of the data (1)

A strip plot



Treatment groups

The response level  
(on log scale)

The data

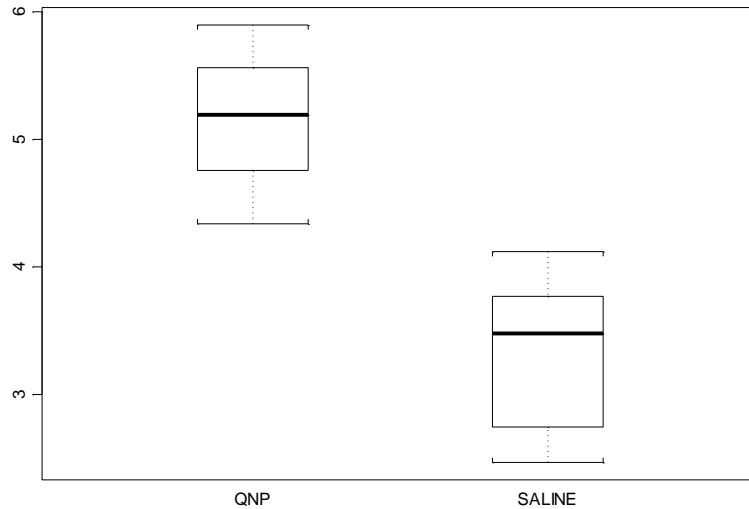
22	QNP	186.6145
11	QNP	103.3529
4	QNP	191.3850
16	QNP	334.9845
7	QNP	89.2831
13	QNP	345.5070
2	QNP	169.5161
20	QNP	173.1491
19	QNP	130.9634
8	QNP	363.4392
10	QNP	76.5340
24	QNP	202.1145
1	SALINE	12.8458
17	SALINE	44.3092
15	SALINE	41.3581
6	SALINE	24.5560
23	SALINE	61.5525
18	SALINE	38.8464
5	SALINE	27.0107
12	SALINE	45.9960
21	SALINE	13.7927
14	SALINE	42.4009
3	SALINE	17.5861
9	SALINE	11.7937

Response

Treatment group

# Graphical display of the data (2)

A boxplot plot



The response level  
(on log scale)

Treatment  
groups

The data

22	QNP	186.6145
11	QNP	103.3529
4	QNP	191.3850
16	QNP	334.9845
7	QNP	89.2831
13	QNP	345.5070
2	QNP	169.5161
20	QNP	173.1491
19	QNP	130.9634
8	QNP	363.4392
10	QNP	76.5340
24	QNP	202.1145
1	SALINE	12.8458
17	SALINE	44.3092
15	SALINE	41.3581
6	SALINE	24.5560
23	SALINE	61.5525
18	SALINE	38.8464
5	SALINE	27.0107
12	SALINE	45.9960
21	SALINE	13.7927
14	SALINE	42.4009
3	SALINE	17.5861
9	SALINE	11.7937

Response

Treatment group

# Descriptive Statistics: Overall mean

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
2.468	3.569	4.229	4.244	5.173	5.896

The function **summary( )** was used in order to calculate the overall mean. The **response** is **ldist** (the distance traveled on log scale)

```
> summary(Biophar$ldist)
```

# Groups' means

```
> tapply(ldist,Biophar$Treat,mean)
```

QNP	SALINE
5.164716	3.323143

```
> tapply(ldist,Biophar$Treat,sd)
```

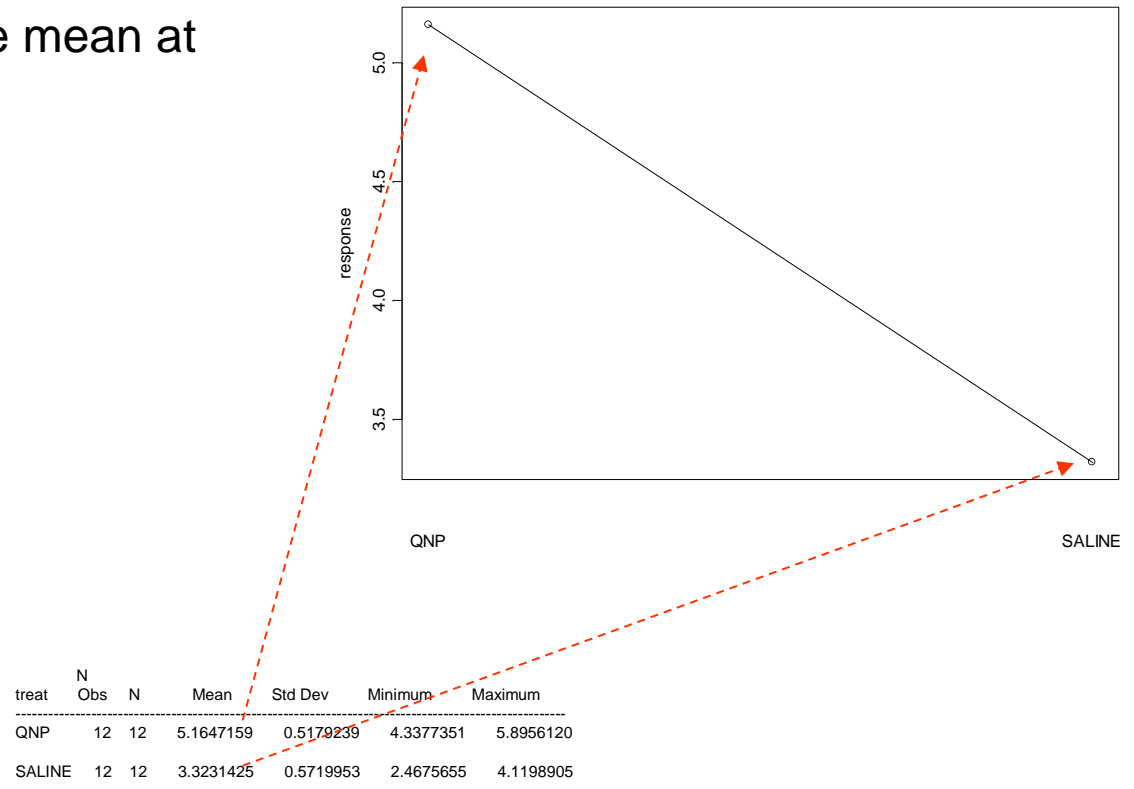
QNP	SALINE
0.5179239	0.5719953

```
tapply(ldist, Biophar$Treat, mean)  
tapply(ldist, Biophar$Treat, sd)
```

Using the function `tapply ( )`, the means and standard deviations are calculated for each level of the factor (i.e., for each treatment group)

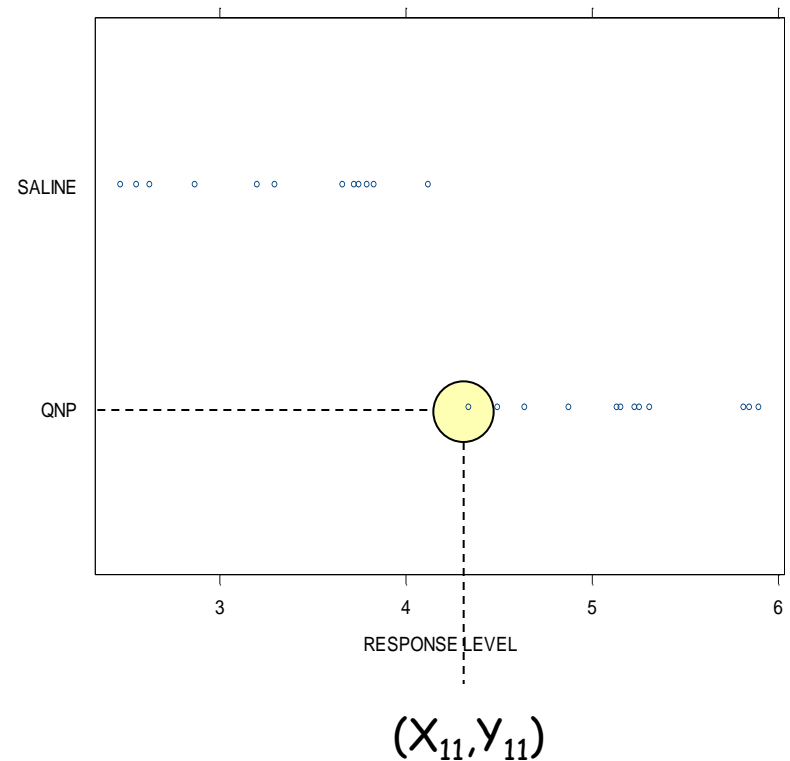
# Graphical Display (3)

The mean plot, shows the mean at each factor level.



# ANOVA Terminology

- The distance traveled in is the dependent variable. This is the response.
- The treatment group is the independent variable and it called the factor. In this example the factor has two levels.



# Data Structure

- We have two variables, the factor ( $x$ ) and the response ( $Y$ ).
- The value of  $X$  is equal for all subjects from the same treatment group. This value is the factor level.

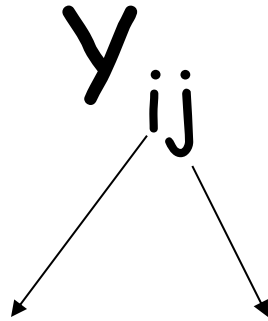
22	QNP	186.6145
11	QNP	103.3529
4	QNP	191.3850
16	QNP	334.9845
7	QNP	89.2831
13	QNP	345.5070
2	QNP	169.5161
20	QNP	173.1491
19	QNP	130.9634
8	QNP	363.4392
10	QNP	76.5340
24	QNP	202.1145
1	SALINE	12.8458
17	SALINE	44.3092
15	SALINE	41.3581
6	SALINE	24.5560
23	SALINE	61.5525
18	SALINE	38.8464
5	SALINE	27.0107
12	SALINE	45.9960
21	SALINE	13.7927
14	SALINE	42.4009
3	SALINE	17.5861
9	SALINE	11.7937

The factor:  
the  
treatment  
group

The response:  
the distance  
traveled ( $y_i$ )



# Data Structure: notation (1)



Group  
number:  $i=1,2$

Observation  
number:  
 $j=1,2,3,\dots,12$

The  
sample  
size at  
the group.

22	QNP	186.6145
11	QNP	103.3529
4	QNP	191.3850
16	QNP	334.9845
7	QNP	89.2831
13	QNP	345.5070
2	QNP	169.5161
20	QNP	173.1491
19	QNP	130.9634
8	QNP	363.4392
10	QNP	76.5340
24	QNP	202.1145
1	SALINE	12.8458
17	SALINE	44.3092
15	SALINE	41.3581
6	SALINE	24.5560
23	SALINE	61.5525
18	SALINE	38.8464
5	SALINE	27.0107
12	SALINE	45.9960
21	SALINE	13.7927
14	SALINE	42.4009
3	SALINE	17.5861
9	SALINE	11.7937

Group 1:

$i=1$

$n_1=12$

Group 2:

$i=2$

$n_2=12$

$y_{212}$ :  
Observation  
number 12 in  
group 2

# Data Structure: notation (2)

Number of Group: I

Sample size: n

$n = n_1 + n_2 + \dots + n_k$

Overall mean:  $\bar{Y}_{..}$

Mean of group j:  $\bar{Y}_{j.}$

Sample size in group j:  $n_j$

22	QNP	186.6145
11	QNP	103.3529
4	QNP	191.3850
16	QNP	334.9845
7	QNP	89.2831
13	QNP	345.5070
2	QNP	169.5161
20	QNP	173.1491
19	QNP	130.9634
8	QNP	363.4392
10	QNP	76.5340
24	QNP	202.1145

1	SALINE	12.8458
17	SALINE	44.3092
15	SALINE	41.3581
6	SALINE	24.5560
23	SALINE	61.5525
18	SALINE	38.8464
5	SALINE	27.0107
12	SALINE	45.9960
21	SALINE	13.7927
14	SALINE	42.4009
3	SALINE	17.5861
9	SALINE	11.7937

Group 1: The group mean

$\bar{Y}_{1.}$

Group 2: The group mean

$\bar{Y}_{2.}$

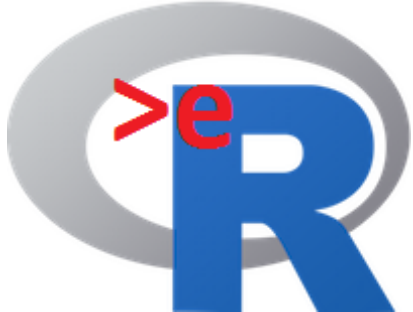
# Descriptive statistics

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
2.468	3.569	4.229	4.244	5.173	5.896

Overall mean:

Mean of group i:

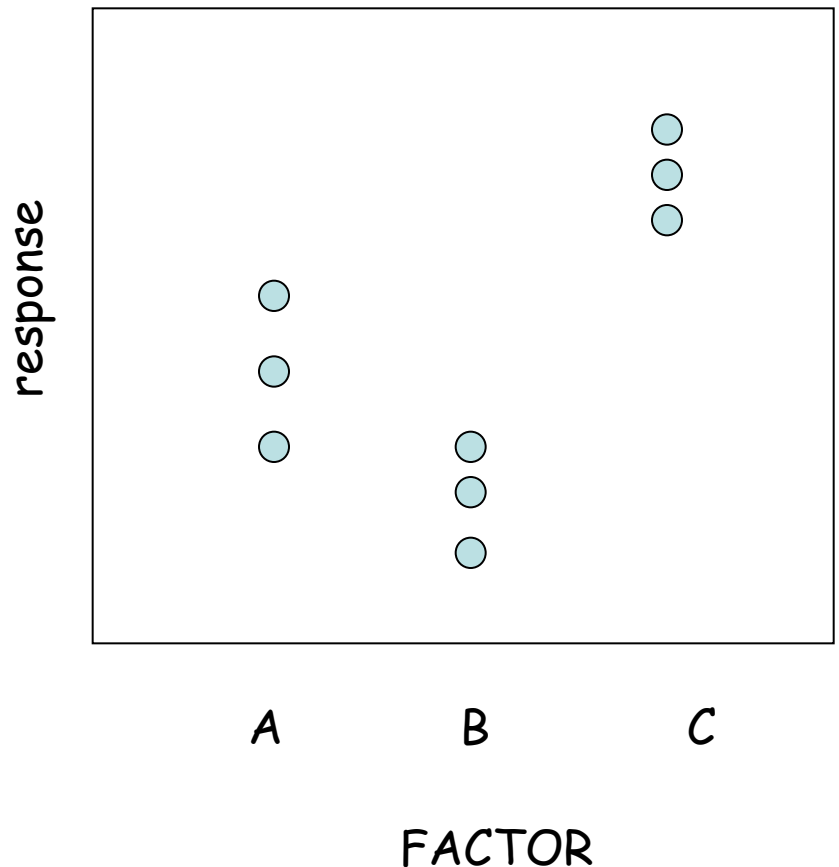
```
> tapply(ldest,Biophar$Treat,mean)
      QNP      SALINE
5.164716 3.323143
```



## Sources of Variability

# What is a One-Way ANOVA Model?

- An One-Way ANOVA model is a statistical model which aims to explain the **variability** of the response variable.
- The question of primary interest is **IF THE MEAN RESPONSE IS DIFFERENT** across the factor levels.



# Two Sources of Variability

- The main concept in ANOVA models, and in particular One-way ANOVA is to decompose the total variability of the response into two parts.

**total** variability = variability **within** the groups + variability **between** the groups

- An ANOVA model is a model in which we explain the total variability with these two sources.

# A very simple example

- One factor experiment.
- The factor has three levels (1,2,3).
- Three observations at each level.

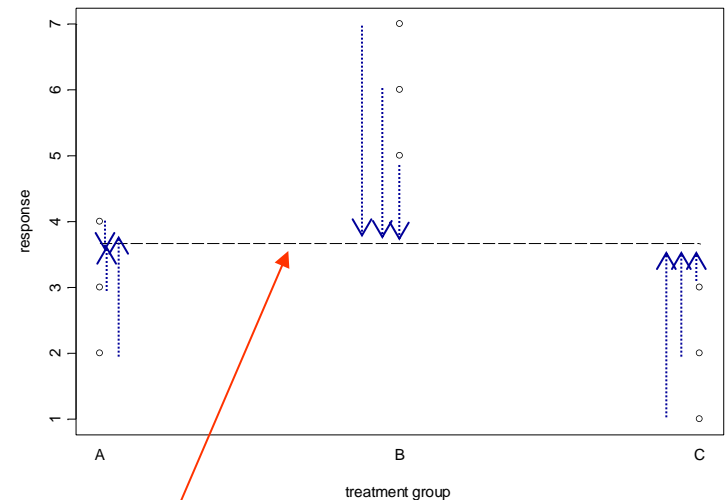
group	$y_{ij}$	Group mean
1 1 1	2 3 4	3
2 2 2	5 6 7	6
3 3 3	1 2 3	2

Overall mean: 3.6666

# Two Sources of Variability: the total variability

The total sum of squares (SST) is the sum of squared distance between the observation and the overall mean.

$$(2 - 3.666)^2 + (3 - 3.666)^2 + (4 - 3.666)^2 + \dots, (2 - 3.666)^2 + (3 - 3.666)^2 = 32$$



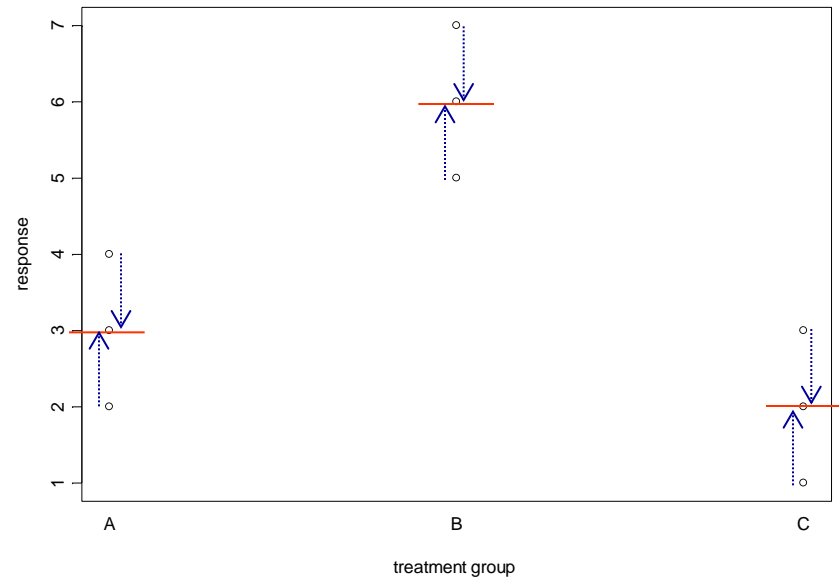
The overall  
mean=3.66667



# Two Sources of Variability: the variability within the groups

The sum of squares with the groups in the sum of squared difference between the observations at each group to the group mean.

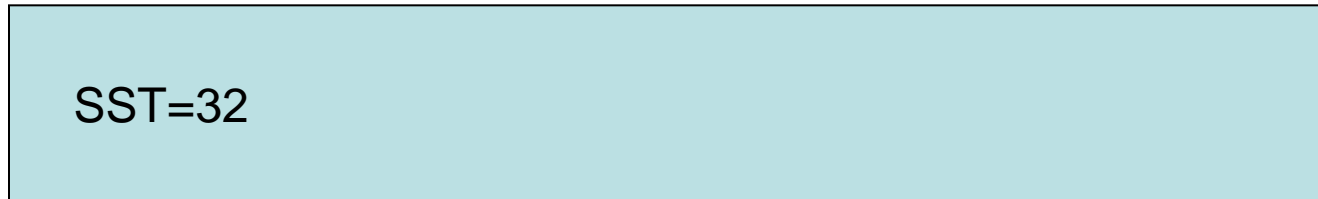
$$\begin{array}{lcl} A & (2-3)^2 + (3-3)^2 + (4-3)^2 = 2 & \\ B & (5-6)^2 + (6-6)^2 + (7-6)^2 = 2 & + \\ C & (1-2)^2 + (2-2)^2 + (3-2)^2 = 2 & + \\ & \hline & 6 & \end{array}$$



Groups means: 3 (group A), 6 (group B) and 2 (group C)

# Two Sources of Variability

Total variability



SSW=6

SSB=26

Variability within  
the groups

Variability  
between the  
groups

$$SST = SSW + SSB$$

In the slides for the class we use the notations:  
 $SST = SSE + SSTR$

# Two Sources of Variability: the variability between the groups

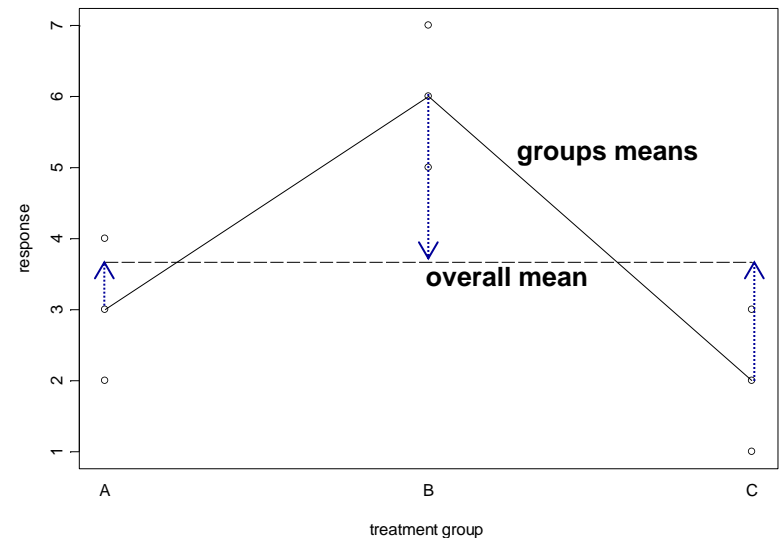
The sum of squares between the groups in the sum of squared difference between the group means to the overall mean.

$$3 \times (3 - 3.666)^2 + 3 \times (6 - 3.666)^2 + 3 \times (2 - 3.666)^2 = 26$$

groups means

overall mean

The sample size at each group



# Two Sources of Variability

- Recall that the aim of the analysis is to test if the mean response across the factor levels are equal.
- In the next few slides we focus on two datasets. In the first there is no difference in the mean response across the factor levels and the second is an example of a dataset in which the means are not equal.
- Mind that: when we say "the means" we mean the parameters (or the population mean) and NOT the sample means !!!!!

# Hypothetical Experiment

- We consider a one factor experiment in which the factor has three levels: A, B and C.
- There are 3 observations at each factor level.
- Sample size:  $N=9$ .
- Sample size per group:  $n_i=3$ ,  $i=1,2,3$ .
- Number of groups:  $I=3$ .

# Example 1: data1

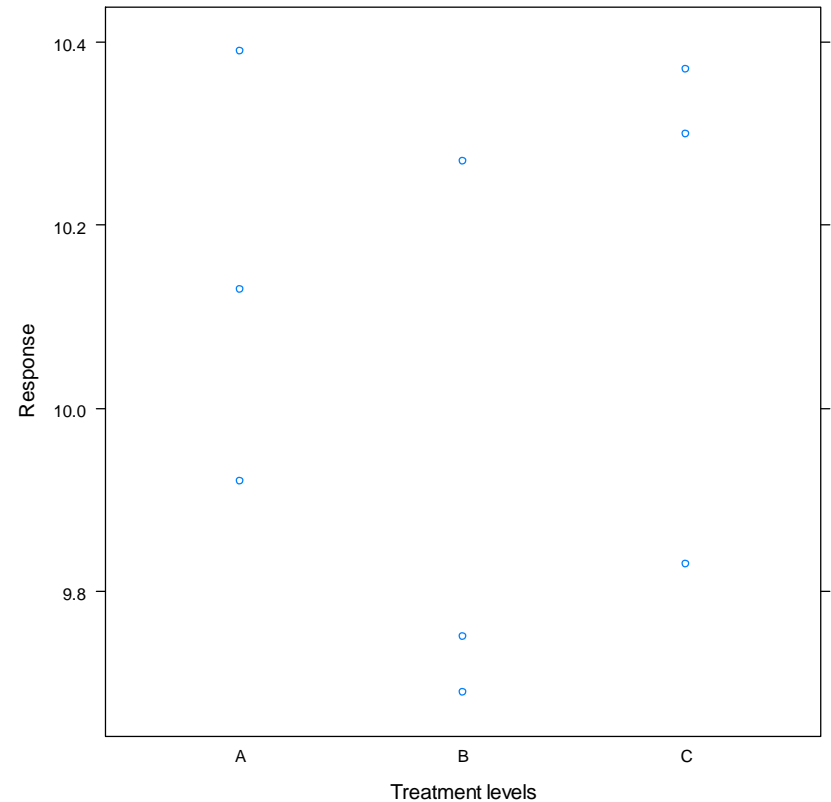
```
> tapply(data1$response, list(data1$Treat), mean)
```

A	B	C
10.146667	9.903333	10.166667

```
> tapply(data1$response, list(data1$Treat), sd)
```

A	B	C
0.2354428	0.3189566	0.2936551

Means' groups and dot plot



**EXAMPLE 1**

# Example 2: data2

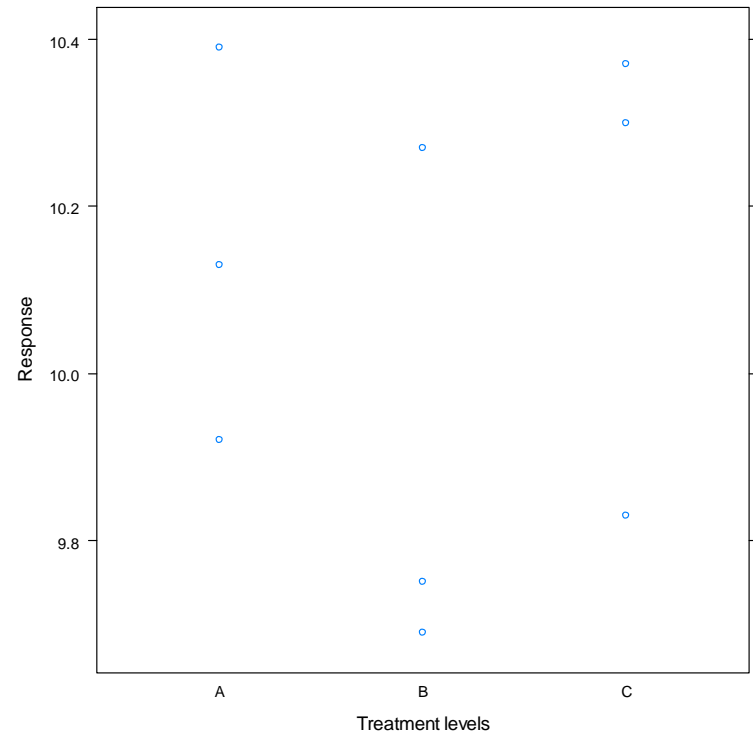
```
> tapply(data2$response, list(data2$Treat), mean)
```

A	B	C
9.886667	12.010000	14.073333

```
> tapply(data2$response, list(data2$Treat), sd)
```

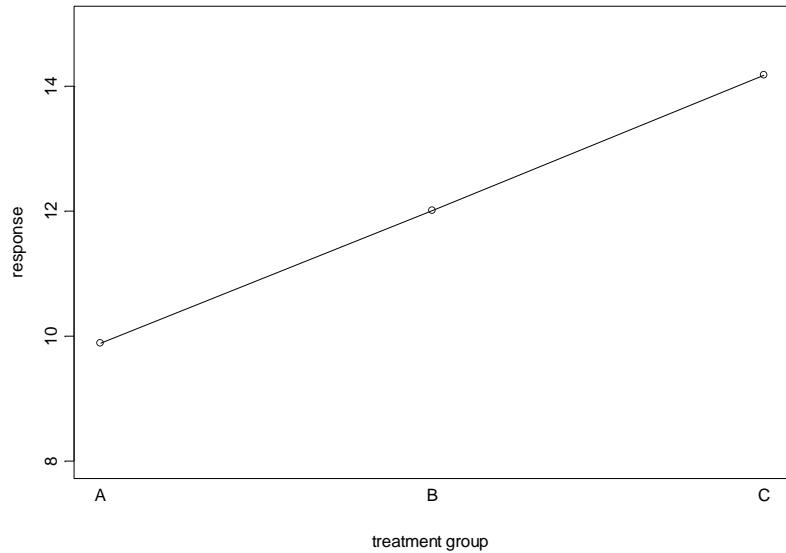
A	B	C
0.59743898	0.59270566	0.07371115

Means' groups and dot plot

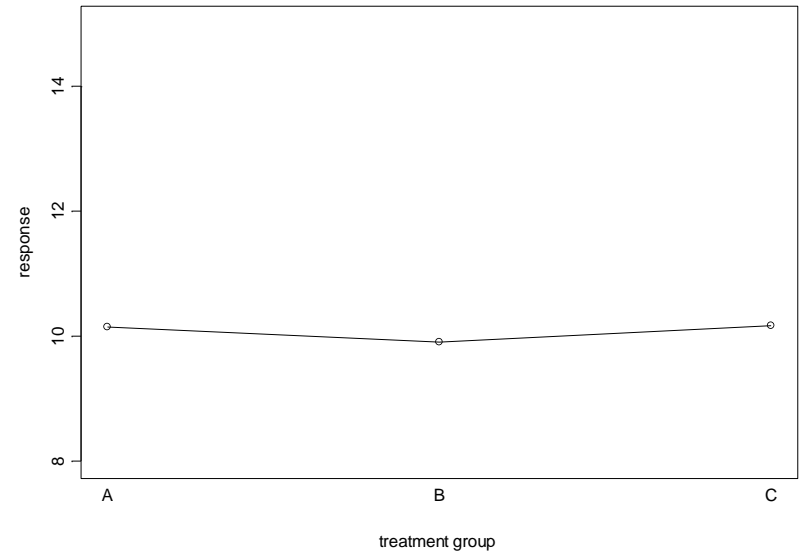


EXAMPLE 2

# Means plot



EXAMPLE 2



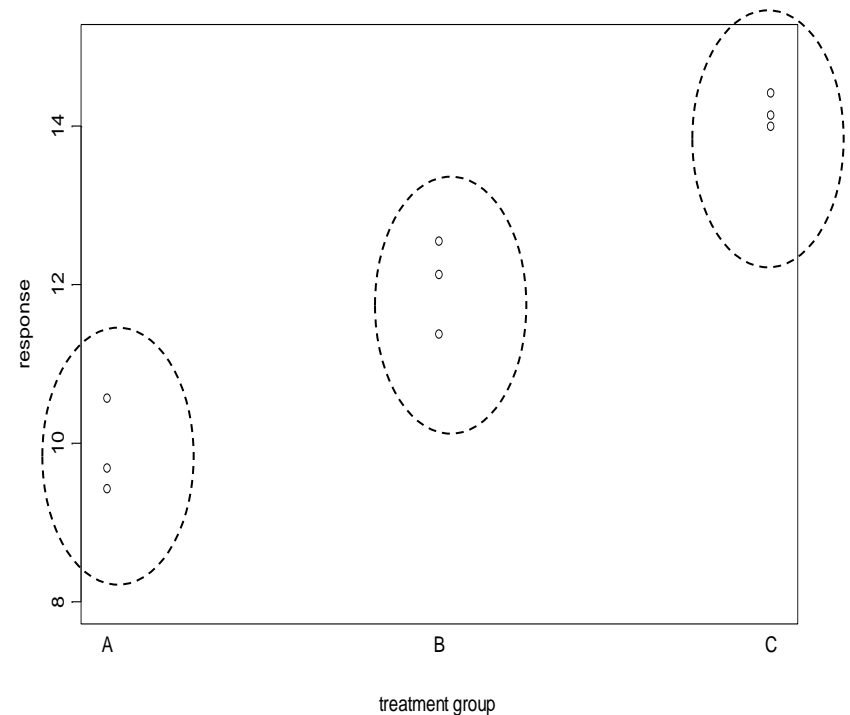
EXAMPLE 1



# Example 2: The Variability within the group

- The variability within a group is the sum of squares between the response and the group mean.
- Within each group, the variability seems to be more or less the same.
- We can see that the value of the response depends on the factor level.

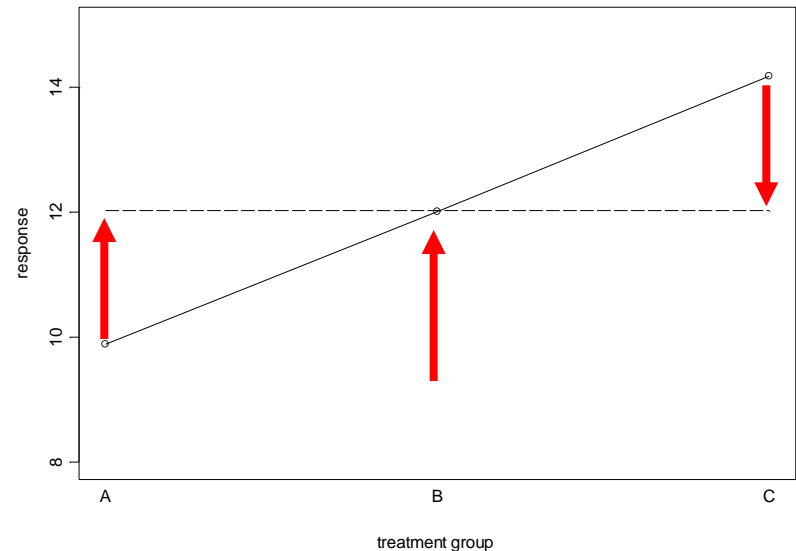
EXAMPLE 2



# Example 2: The variability between the groups

- The difference between the group mean to the overall mean.

## EXAMPLE 2





One-way ANOVA using R:  
the `aov()` function

# The function `aov( )` in R

Fitting one-way ANOVA model  
with function `aov( )`

```
> fit.data2 <- aov(response ~ Treat, data = data2)
```

factor

```
> aov(fit.data2)
```

Call:

```
aov(formula = fit.data2)
```

Terms:

	Treat	Residuals
Sum of Squares	26.294067	1.427333
Deg. of Freedom	2	6

Residual standard error: 0.4877385

Estimated effects may be unbalanced

The `aov(fit.data2)` gives  
information about the data  
structure and about the  
observations used for the  
analysis

# Sources of variability

- The second part of the output is the ANOVA table.
- The column **Sum of Squares** presents the between, the within and the total sum of squares

The SAS System  
The ANOVA Procedure

Dependent Variable: response

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	27.60708889	13.80354444	54.92	0.0001
Error	6	1.50793333	0.25132222		
Corrected Total	8	29.11502222			

The variability Within the groups

The variability between the groups

EXAMPLE 2

# Sources of variability

- The column **Sum of Squares** presents the between and the within (residual) sum of squares

```
> summary(fit.data2)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Treat	2	26.294	13.147	55.27	0.000136 ***
Residuals	6	1.427	0.238		

---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

The variability  
Within the groups

The variability between the groups

EXAMPLE 2

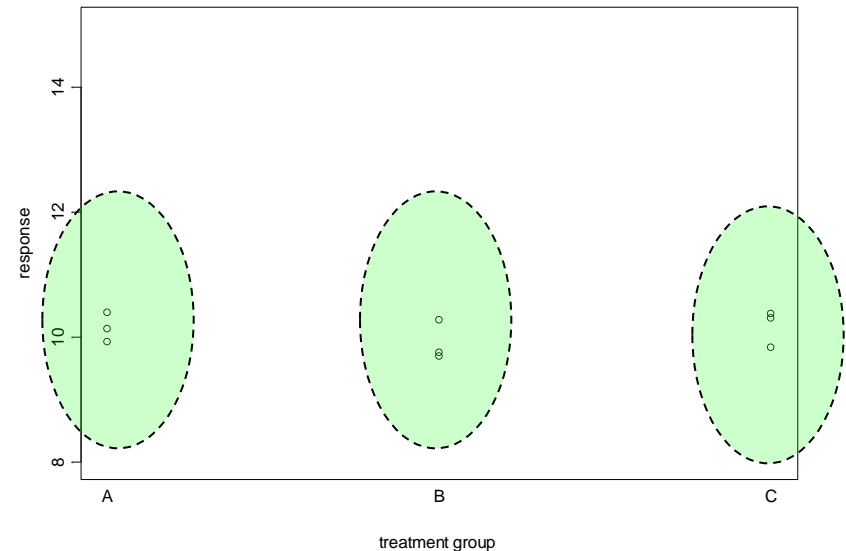
# The variability within the groups (SSW)

```
> summary(fit.data1)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Treat	2	0.1290	0.06448	0.795	0.494
Residuals	6	0.4868	0.08113		

SSW: The error sum of squares is the within group sum of squares. This is the within group source of variability.

## EXAMPLE 1



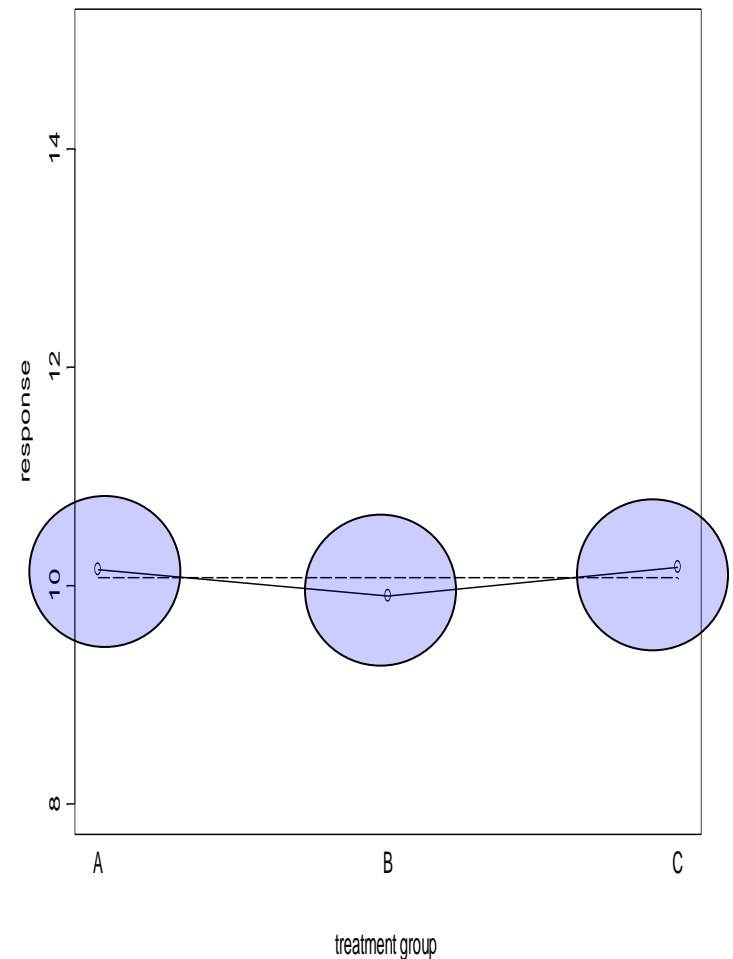
# The variability between the groups (SSB)

## EXAMPLE 1

```
> summary(fit.data1)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Treat	2	0.1290	0.06448	0.795	0.494
Residuals	6	0.4868	0.08113		

SSB: The model sum of squares is the between group sum of squares. This is the between group source of variability.





# Dgrees of freedom, mean squares and the F-value

```
> summary(fit.data1)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Treat	2	0.1290	0.06448	0.795	0.494
Residuals	6	0.4868	0.08113		

I-1

n-I

$$\frac{0.06448}{0.08113} = 0.79$$

n=9, I=3:

n-1=8

n-I=6

I-1=2

Where:

I - is the factor level

n - is the sample size

$$0.12895556 / 2 = 0.06447778$$

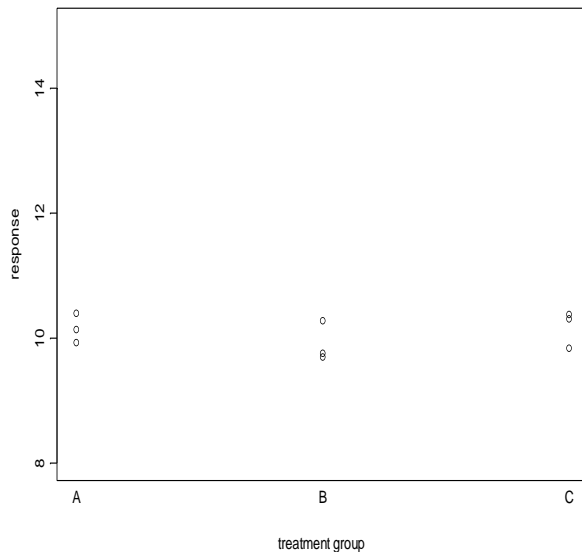
$$0.48680000 / 6 = 0.08113333$$

**EXAMPLE 1**

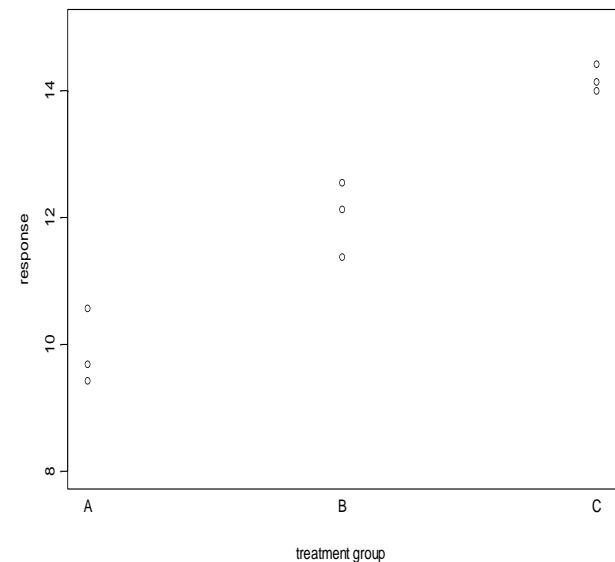
...and now to the **big** question...

HOW CAN WE TEST IF THE MEAN RESPONSE IS EQUAL ACROSS THE GROUPS?

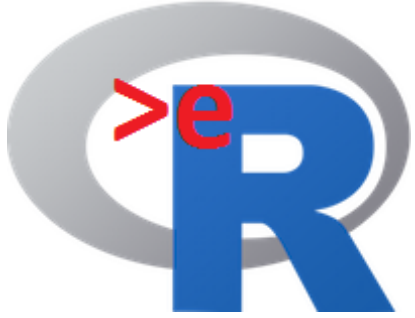
EXAMPLE 1



EXAMPLE 2



What is the main difference between the two examples?



## Model formulation and hypotheses testing

# One-Way ANOVA model

- The one way ANOVA model is a statistical model which we use in order to test the null hypothesis that the mean response across the factor level equal.
- It does not tell us which one is different.

# One-Way ANOVA model: model formulation

$$Y_{ij} = \mu_i + \varepsilon_{ij}$$

Parameters: fixed but unknown and needed to be estimated

Random error, assumed to follow normal distribution with constant variance.

$$\varepsilon_{ij} \sim N(0, \sigma^2)$$

Model assumptions are:

1. The random error is normal distributed.
2. The variance is constant across the factor levels.

# The Null Hypothesis: No treatment effect

- For a model in which the factor has three levels we wish to test the null hypothesis:

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

- This means that we want to test if the means across all factor levels are equal.
- Mind that: we test if the parameters ( $\mu_j$ ) are equal, not is the sample means ( $\bar{Y}_j$ ).

# Test Statistic

Within group sum of squares

$$SSW = \sum_{i=1}^I \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{i.})^2$$

Between group sum of squares

$$SSB = \sum_{i=1}^I n_i (\bar{Y}_{i.} - \bar{Y}_{..})^2$$

$$F = \frac{SSB / (I - 1)}{SSW / (N - I)} = \frac{MSB}{MSW}$$

The test statistic,  $F$ , is the ratio between the mean of the between sum of squares ( $SSB$ ) and the mean of the within sum of squares.

# Test Statistic in R

Within group sum of squares/dgree of fredom  
Between group sum of squares/dgree of fredom

$$\frac{SSB / (I - 1)}{SSW / (N - I)} = \frac{MSB}{MSW} = F$$

```
> summary(fit.data1)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Treat	2	0.1290	0.06448	0.795	0.494
Residuals	6	0.4868	0.08113		

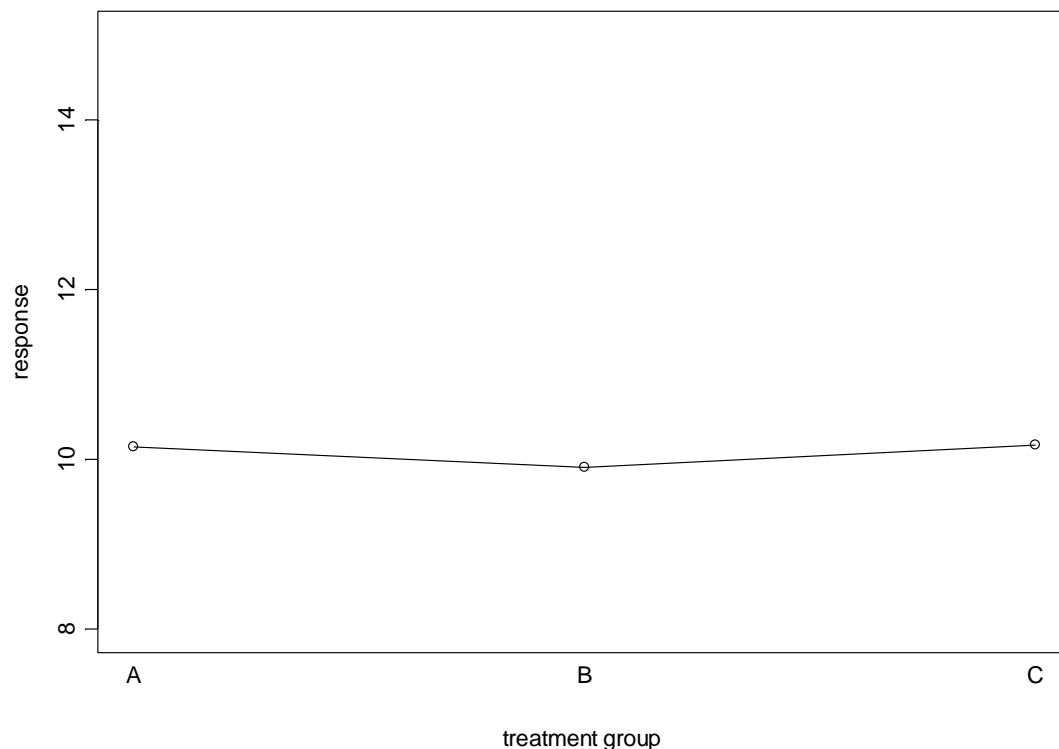


# Intuition behind the F-test

If there is no difference between the mean response across the treatment groups, then we expect that  $SSB$  will be relatively small (since the group means are closed to the overall mean).

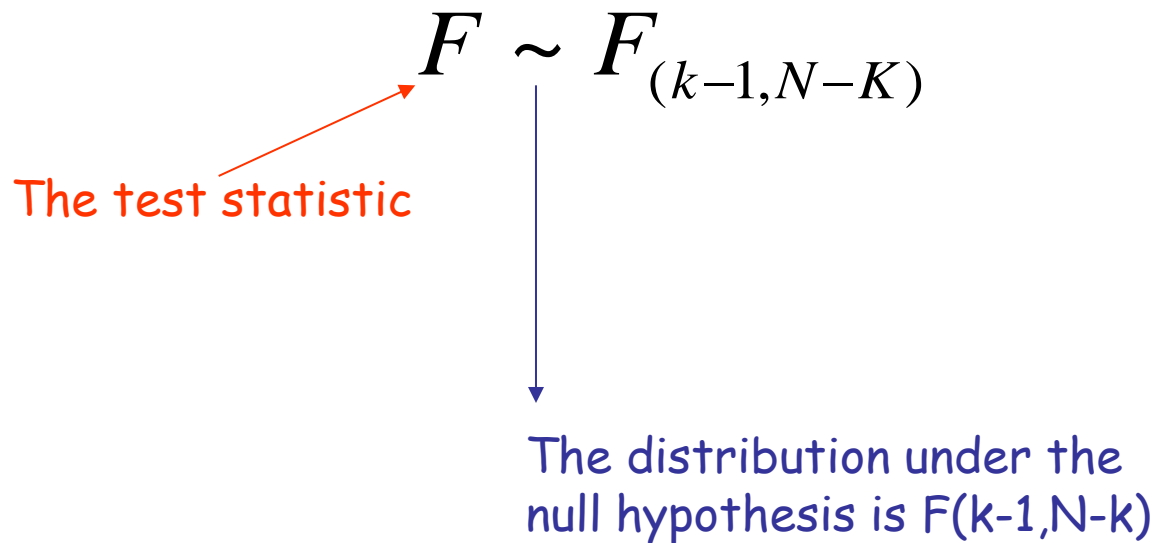
This means that we will reject the null hypothesis for a "large" value of  $SSB$  or a "large" value of  $F$ .

What is a large value? In example 1  $F=0.79$ , is it large?



# Distribution of the Test Statistic Under the null Hypothesis

- Under the null hypothesis, the distribution of the test statistic  $F = \frac{SSB/(K-1)}{SSW/N-K} = \frac{MSB}{MSW}$  is known.



# Inference

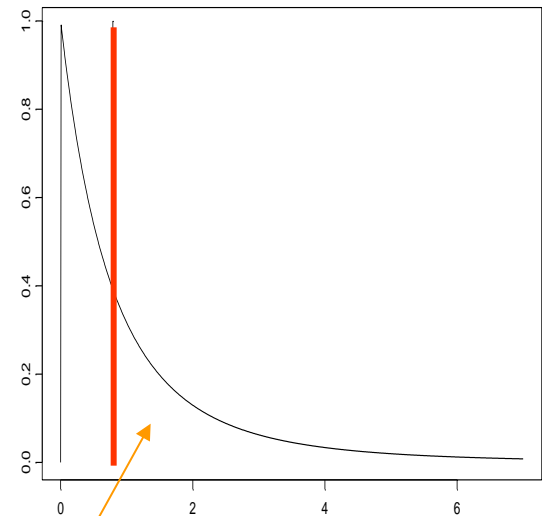
## The decision rule:

If the P-value  $> 0.05$  we do not reject the null hypothesis.

If the P-value  $< 0.05$  we reject the null hypothesis.

Conclusion: we do not reject the null hypothesis since  $P\text{-value}=0.49 > 0.05$ .

$F(2,6)$



	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Treat	2	0.1290	0.06448	0.795	0.494
Residuals	6	0.4868	0.08113		

$K-1=2, N-K=6$

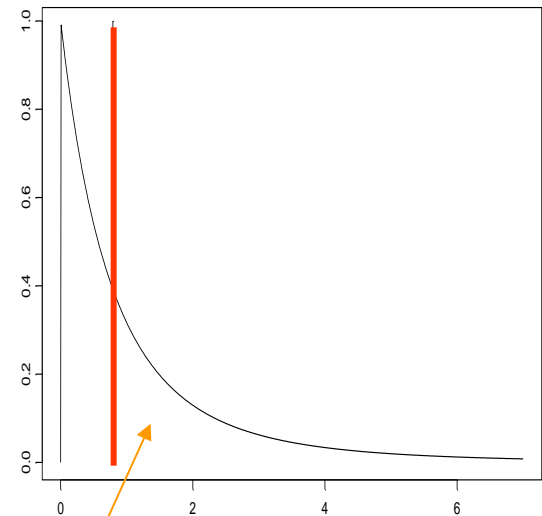
# Inference

The plot shows the distribution of the test statistic (the density function) under the null. The red line is the observed value of the test statistic, i.e. 0.79.

The P-value ( $\Pr > F$  in the SAS output) is the probability to observe an extreme value, i.e. Values that greater or equal to 0.79.

The P-value is the area under the curve in the right side of the red line.

$F(2,6)$



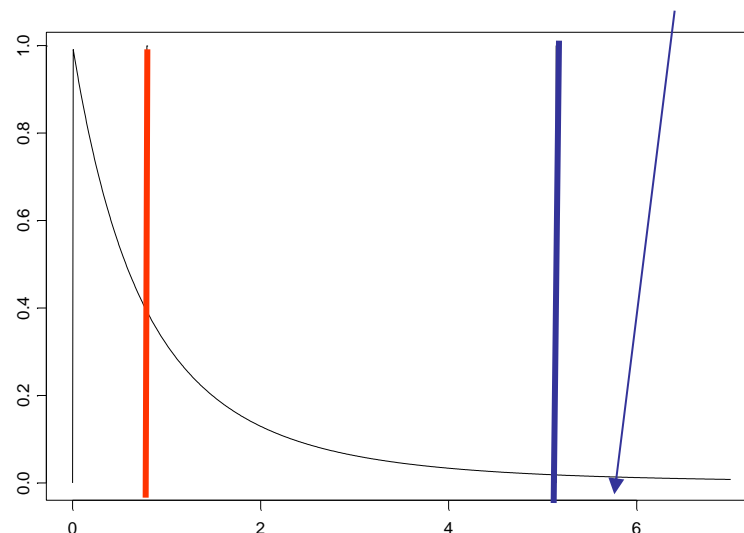
$I-1 = 2, n - I = 6$

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Treat	2	0.1290	0.06448	0.795	0.494
Residuals	6	0.4868	0.08113		

# Inference

- The P-value is the probability to observed the value of the test statistics (0.79) under the null hypotheses.
- If  $P\text{-value} > 0.05$  we do not reject the null hypothesis.

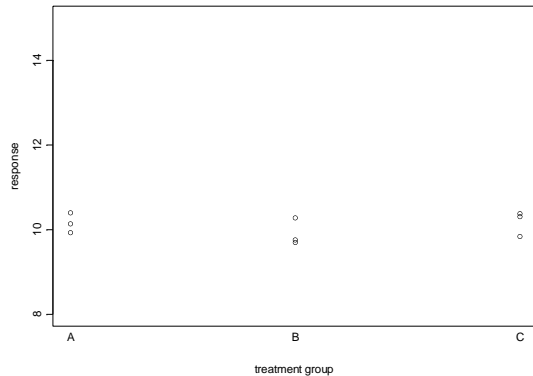
Critical value: the probability to observed a vlaue higher than the critical value if 0.05



	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Treat	2	0.1290	0.06448	0.795	0.494
Residuals	6	0.4868	0.08113		

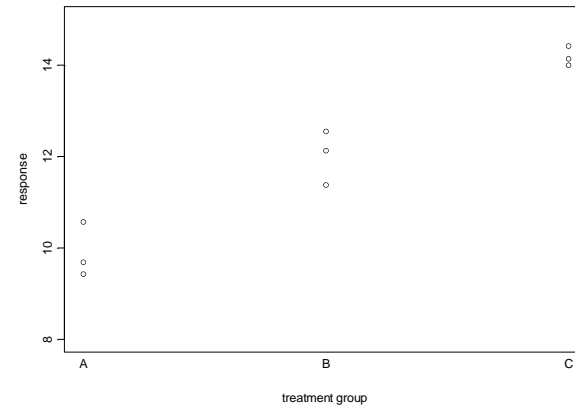
# Using the ANOVA table to test the null hypothesis

EXAMPLE 1



	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Treat	2	0.1290	0.06448	0.795	0.494
Residuals	6	0.4868	0.08113		

EXAMPLE 2



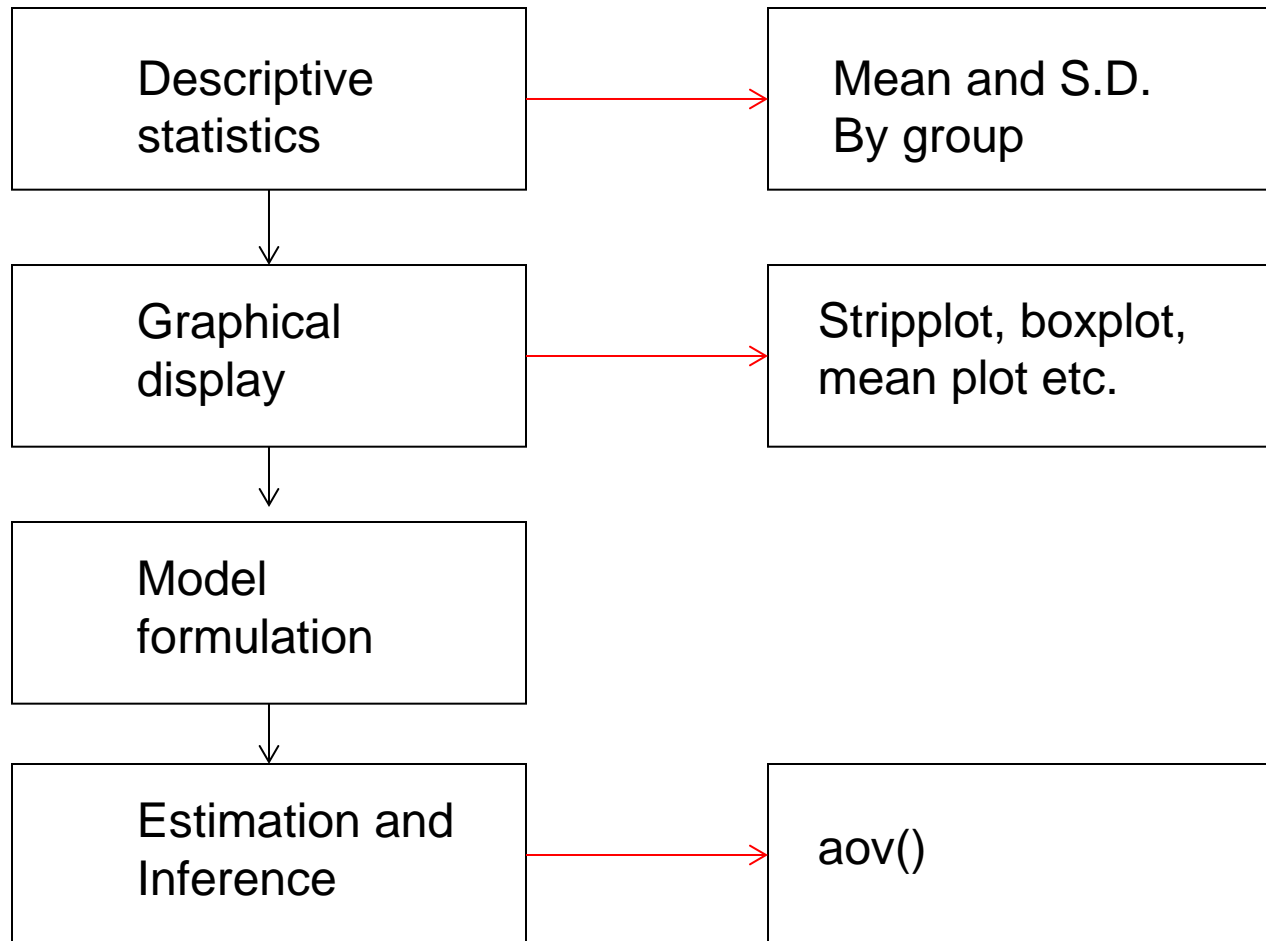
	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Treat	2	26.294	13.147	55.27	0.000136 ***
Residuals	6	1.427	0.238		

---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05  
'.' 0.1 ' ' 1

Example 1:  $P=0.49 > 0.05$ , we do not reject  $H_0$

Example 2:  $P=0.0001 < 0.05$ , we reject  $H_0$

# Fitting ANOVA model: the step of the analysis





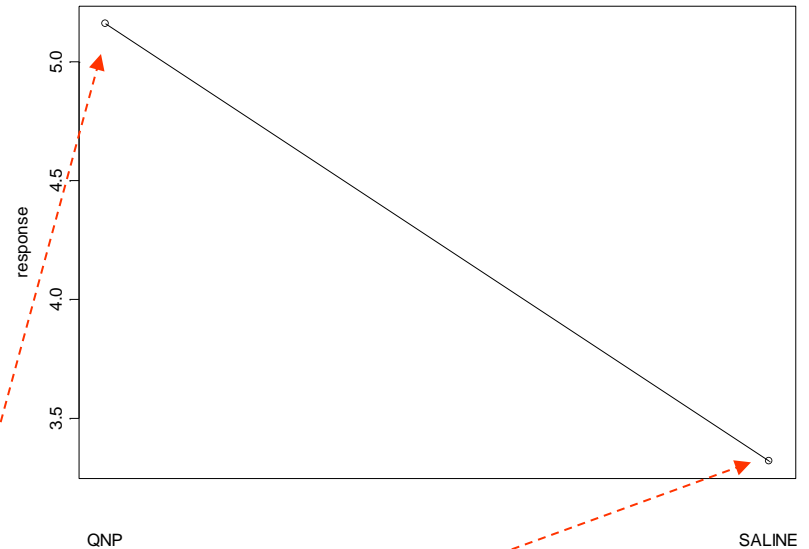
Analysis of the pharmaceutical experiment



# Descriptive Analysis

The sample mean in the active drug group is equal to 5.16 higher than the sample mean in the control group (3.23).

The variability seems to be equal in the two groups (S.D=0.52 in the active drug group compared with S.D=0.57 in the control group).

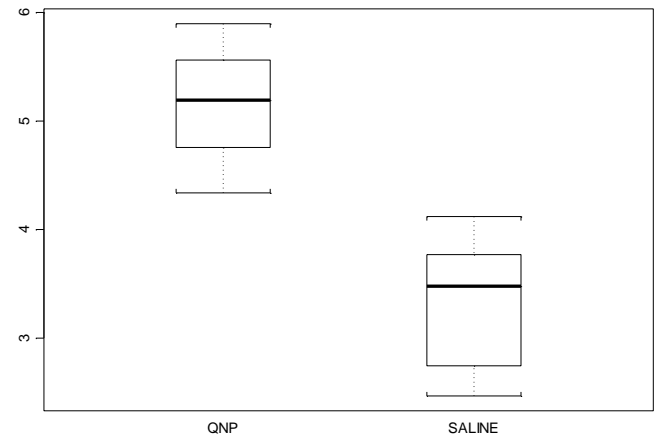
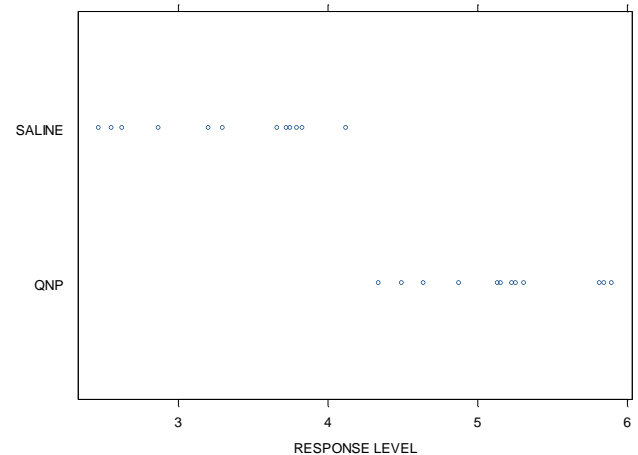


treat	N	Obs	Mean	Std Dev	Minimum	Maximum
QNP	12	12	5.1647159	0.5179239	4.3377351	5.8956120
SALINE	12	12	3.3231425	0.5719953	2.4675655	4.1198905

# Graphical displays

Both the stripplot and the boxplot indicate that the response is higher in the active drug group.

The stripplot shows clearly that the within variability in the two groups is almost the same.



# Model formulation

- We consider the following one-way ANOVA model

$$Y_{ij} = \mu_i + \varepsilon_{ij}$$

- Here,  $\mu_j$  is the mean response in treatment group  $j$ ,  $j = 1, 2$  and  $Y_{ij}$  is the distance traveled (on log scale) by the  $i$ 'th rat in  $j$ 'th treatment group .
- Sample sizes were equal in both treatment group,  $n_1 = n_2 = 12$  and  $N = 24$ .
- It is further assumed that  $\varepsilon_{ij} \sim N(0, \sigma^2)$ .

# Inference

We wish to get the hypothesis that the drug has no effect on the response. Formally we test the following hypotheses:

$$H_0 : \mu_1 = \mu_2$$

$$H_1 : \mu_1 \neq \mu_2$$

We use the F statistic in order to test the hypotheses

$$F = \frac{SSB / (K - 1)}{SSW / N - K} = \frac{MSB}{MSW}$$

Under the null hypotheses:

$$F \sim F_{(1,22)}$$

```
> summary(fit.biophar)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Treat	1	20.35	20.348	68.35	3.41e-08 ***
Residuals	22	6.55	0.298		

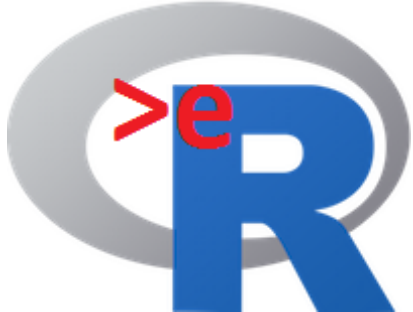
---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

The model was estimated using the function `aov()` in R.

F = 68.35 with P < 0.001.

Conclusion: The null hypothesis is rejected (p < 0.05) and we conclude that distance traveled by rats which received the active drug is higher than the distance traveled by the rats from the control group.

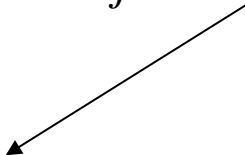


## Model diagnostic

# The one-way ANOVA model

The one-way ANOVA models has two components: the unknown parameters and the stochastic part.

$$Y_{ij} = \mu_i + \varepsilon_{ij}$$



Parameter- unknown and  
fixed



Random error

Our aim is to estimate and to make inference about the parameters. However, the validity of the inference depend is the model assumptions hold

# Parameters and parameter estimates

- Parameters

$$\mu_j.$$

- Parameter estimates

$$\hat{\mu}_j.$$

The parameters in the model represent the populations mean. The parameters estimates are for the populations mean are the same means at each group.

$$\hat{\mu}_1. = \bar{Y}_1.$$

$$\hat{\mu}_2. = \bar{Y}_2.$$

$$\hat{\mu}_3. = \bar{Y}_3.$$

# Model assumptions (1)

$$\varepsilon_{ij} \sim N(0, \sigma^2)$$



Normal distribution of the  
random error

Constant variance

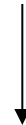
The random component of the model is assumed to follow normal distribution with mean zero and constant variance. However, this is only an assumption that needs to be validated.



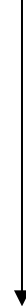
# The residuals

The random error is not observed and, similar to the linear regression model, we estimate it with the residual  $e_i$

$$e_{ij} = Y_{ij} - \bar{Y}_i.$$



The observation



Parameter estimate for the group  
mean

# Model assumptions (2)

- Normality

qq normal plot for the residuals. If the normality assumption holds, the points qnormal plot should be a straight line.

- Constant variance

We use two plots to check this assumption: (1) scatterplot for the response and (2) boxplot for the residuals.



Model diagnostic in R

Illustration with three examples

# Example 1: The Data

```
> tapply(Data1$Response,list(Data1$Treat),mean)
```

```
      1      2      3  
9.786667 10.538333  9.906667
```

```
> tapply(Data1$Response,list(Data1$Treat),sd)
```

```
      1      2      3  
0.4274420 0.4956780 0.3274548
```

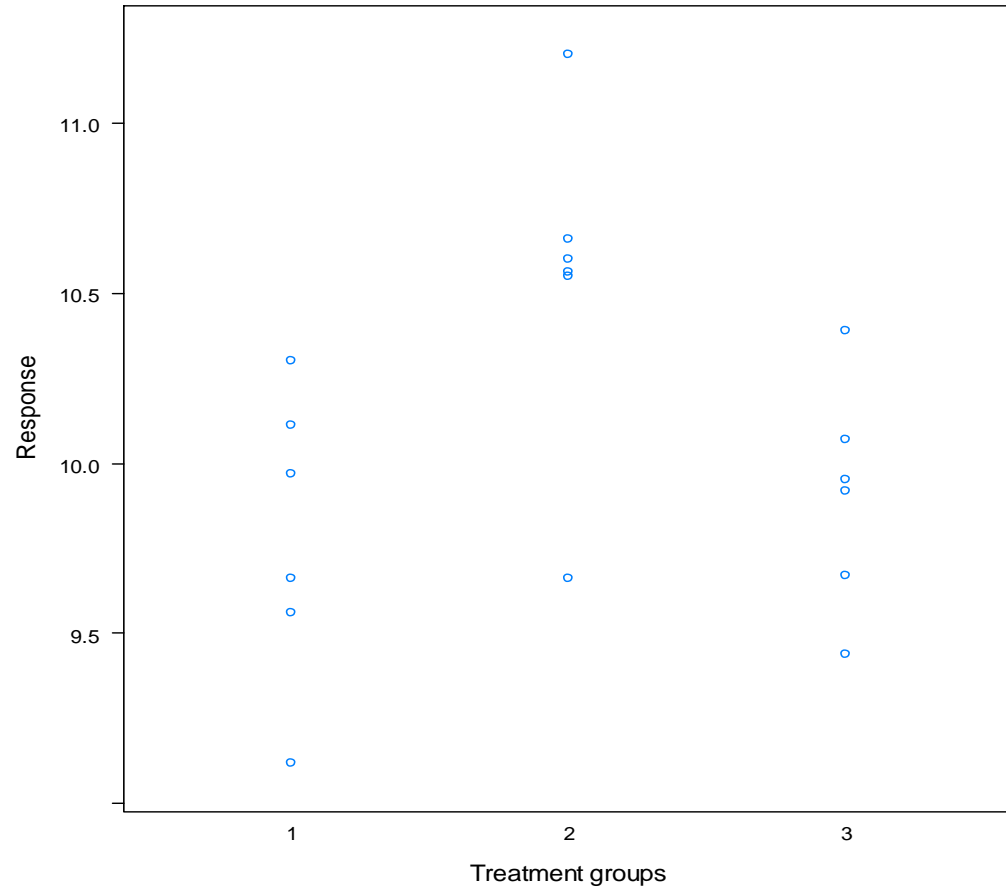
3 groups ( $l=3$ ).

6 observations per group  
( $n_i=6$ ).

The Data

1	1	10.30
2	1	9.12
3	1	9.97
4	1	9.56
5	1	9.66
6	1	10.11
7	2	10.56
8	2	10.60
9	2	9.66
10	2	10.55
11	2	10.20
12	2	10.66
13	3	9.67
14	3	9.44
15	3	10.39
16	3	9.95
17	3	9.92
18	3	10.07

# Graphical display: dotplot by treatment group



# Output

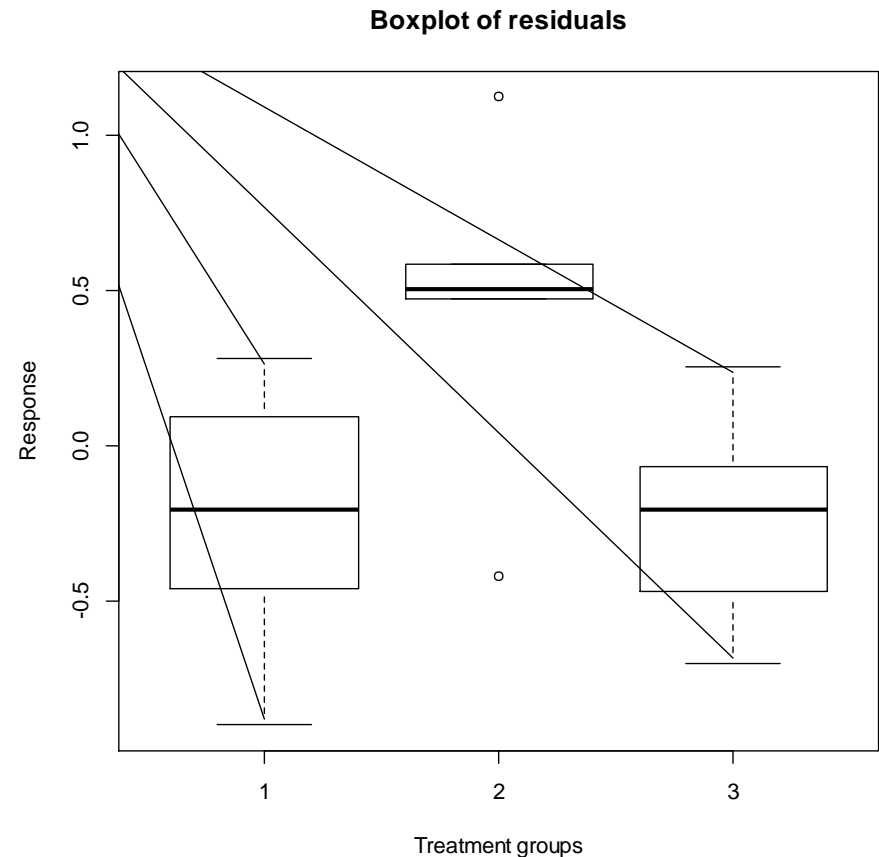
```
> Data1$predict <- fit.Data1$fit  
> Data1$resid <- fit.Data1$resid  
> print(Data1)
```

- Note that the predicted values are simply the group means.

	Treat	Response	predict	resid
1	1	10.30	10.01722	0.28277778
2	1	9.12	10.01722	-0.89722222
3	1	9.97	10.01722	-0.04722222
4	1	9.56	10.01722	-0.45722222
5	1	9.66	10.01722	-0.35722222
6	1	10.11	10.01722	0.09277778
7	2	10.56	10.07722	0.48277778
8	2	10.60	10.07722	0.52277778
9	2	9.66	10.07722	-0.41722222
10	2	10.55	10.07722	0.47277778
11	2	11.20	10.07722	1.12277778
12	2	10.66	10.07722	0.58277778
13	3	9.67	10.13722	-0.46722222
14	3	9.44	10.13722	-0.69722222
15	3	10.39	10.13722	0.25277778
16	3	9.95	10.13722	-0.18722222
17	3	9.92	10.13722	-0.21722222
18	3	10.07	10.13722	-0.06722222

# Graphical display: boxplot for the residuals

```
> boxplot(split(fit.Data1$resid,Treat),  
+         xlab = "Treatment groups",  
+         ylab = "Response",  
+         main = "Boxplot of residuals")
```



If the variance is constant we expect to see the same boxplots for all treatment groups.

# Levene's test

- The Levene's test is a formal test for constant variance.

Levene's Test for Homogeneity of Variance (center = median)

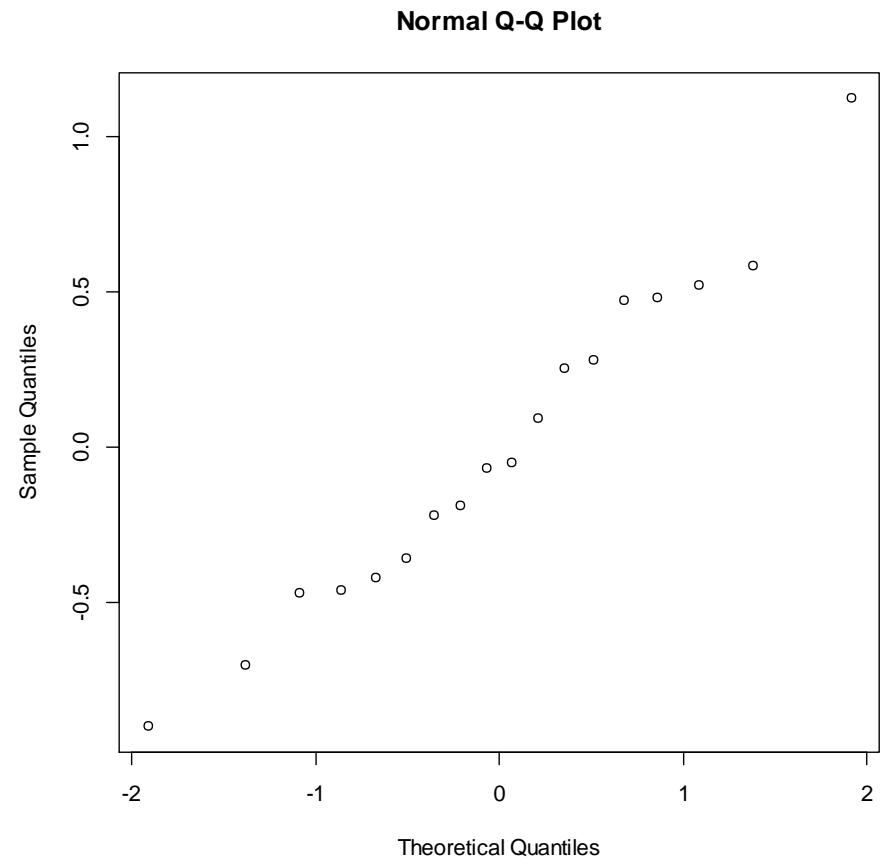
	Df	F	value	Pr(>F)
group	2	0.2251	0.8011	
	15			

```
> library(car)
> leveneTest(Response~factor(Treat), data = Data1)
```



# Normal probability plot for the residuals

- If the normality assumption holds, the points qqnormal plot should be a straight line.



# Example 2: The data

```
> tapply(Data2$Response,  
         list(Data2$Treat),mean)
```

Treatment means

	1	2	3
	10.026667	10.305000	7.766667

```
> tapply(Data2$Response,  
         list(Data2$Treat),sd)
```

Treatment SDs

	1	2	3
	0.2677063	0.9178181	2.6375266

3 groups ( $l=3$ ).

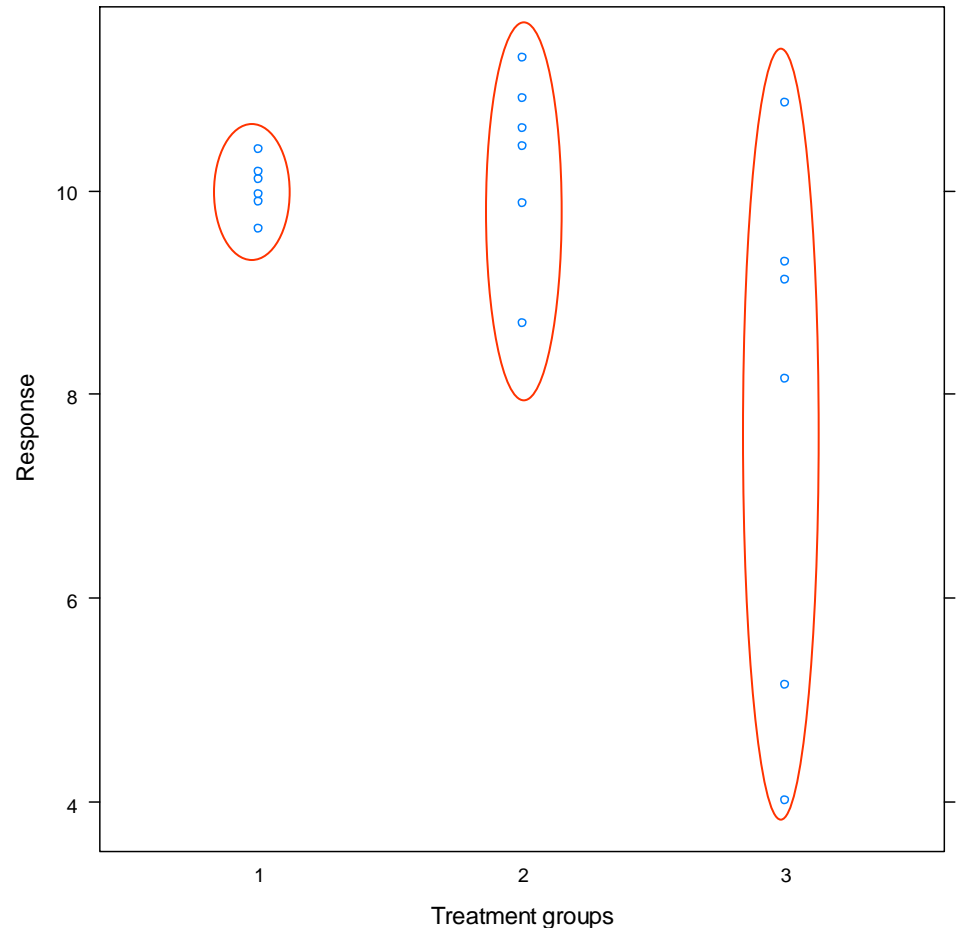
6 observations per group  
( $n_i=6$ ).

## The Data

Treat Response		
1	1	10.18
2	1	10.11
3	1	9.96
4	1	9.89
5	1	10.40
6	1	9.62
7	2	10.44
8	2	8.70
9	2	9.88
10	2	10.90
11	2	11.30
12	2	10.61
13	3	4.02
14	3	9.12
15	3	8.15
16	3	10.86
17	3	5.15
18	3	9.30

# Graphical display: dotplot by treatment group

The variability within each treatment group is not constant.



# Inference and levene's test

Levene's Test for Homogeneity of Variance  
(center = median)

	Df	F value	Pr(>F)
group	2	4.4458	0.03047 *
	15		

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Treat	1	15.32	15.323	5.185	0.0369 *
Residuals	16	47.29	2.955		

---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

```
> fit.Data2 <- aov(Response ~ Treat, data = Data2)
```

```
> summary(fit.Data2)
```

```
> library(car)
```

```
> leveneTest(Response~factor(Treat), data = Data2)
```

# Predicted values and residuals

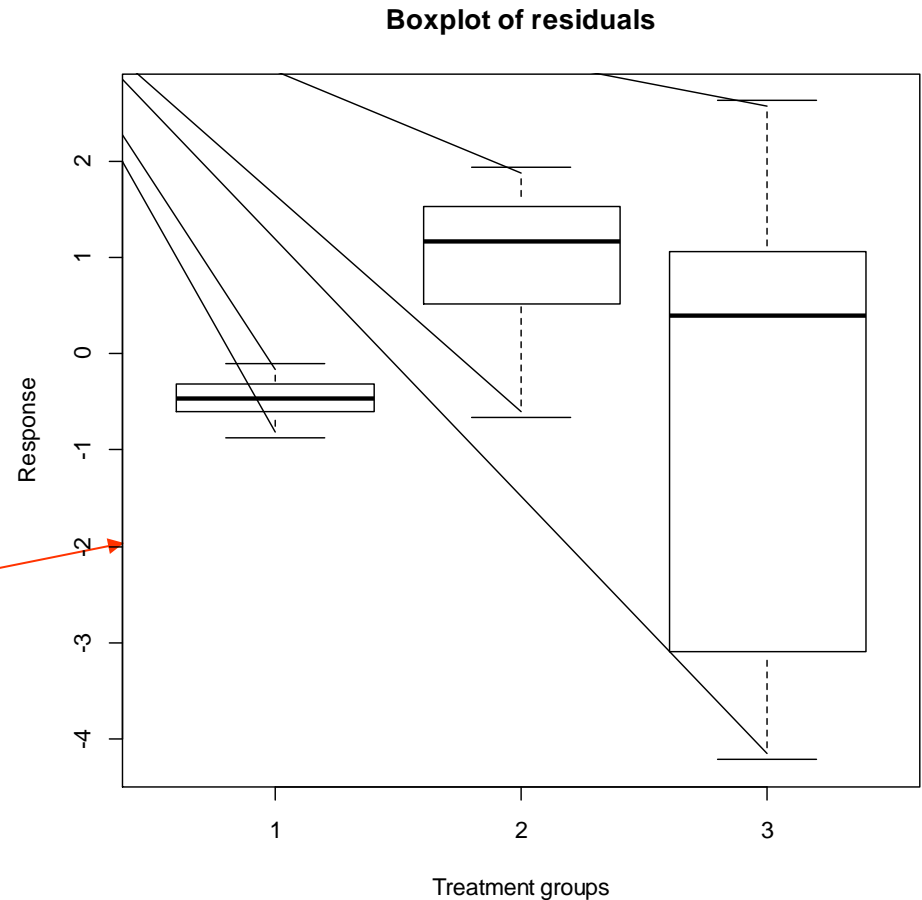
```
> Data2$predict <- fit.Data2$fit  
> Data2$resid <- fit.Data2$resid  
> print(Data2)
```

	Treat	Response	predict	resid
1	1	10.18	10.496111	-0.31611111
2	1	10.11	10.496111	-0.38611111
3	1	9.96	10.496111	-0.53611111
4	1	9.89	10.496111	-0.60611111
5	1	10.40	10.496111	-0.09611111
6	1	9.62	10.496111	-0.87611111
7	2	10.44	9.366111	1.07388889
8	2	8.70	9.366111	-0.66611111
9	2	9.88	9.366111	0.51388889
10	2	10.90	9.366111	1.53388889
11	2	11.30	9.366111	1.93388889
12	2	10.61	9.366111	1.24388889
13	3	4.02	8.236111	-4.21611111
14	3	9.12	8.236111	0.88388889
15	3	8.15	8.236111	-0.08611111
16	3	10.86	8.236111	2.62388889
17	3	5.15	8.236111	-3.08611111
18	3	9.30	8.236111	1.06388889

# Graphical display: Boxplot for the residuals

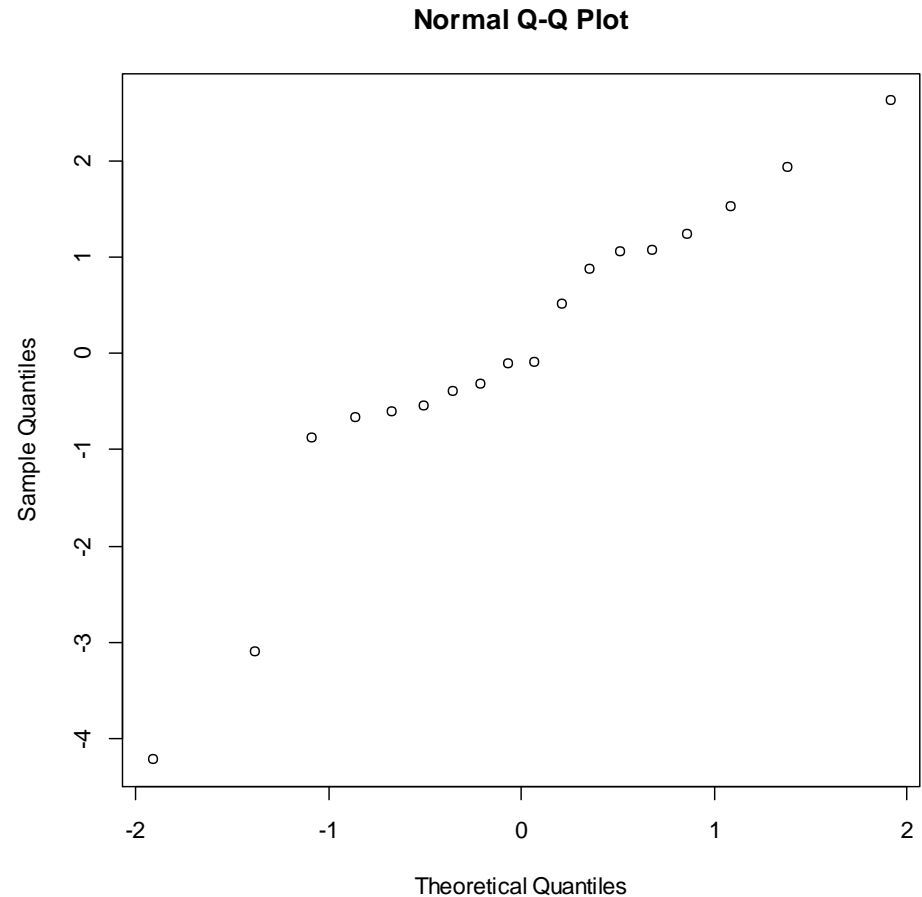
```
> boxplot(split(fit.Data2$resid,Treat),  
+         xlab = "Treatment groups",  
+         ylab = "Response",  
+         main = "Boxplot of residuals")
```

The boxplot indicates that the variability is not constant.



# Normal probability plot for the residuals

- Taking into account that  $n=20$ , the qqnormal plot does not indicate a problem with the normality assumption.





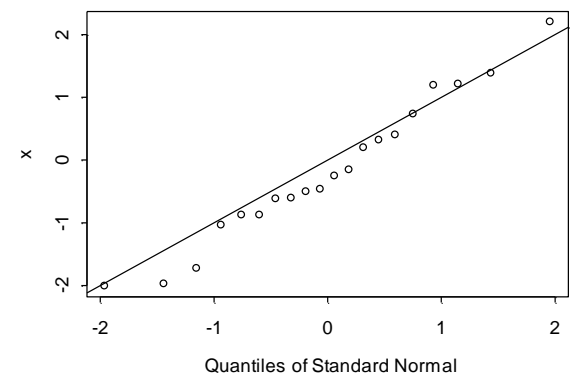
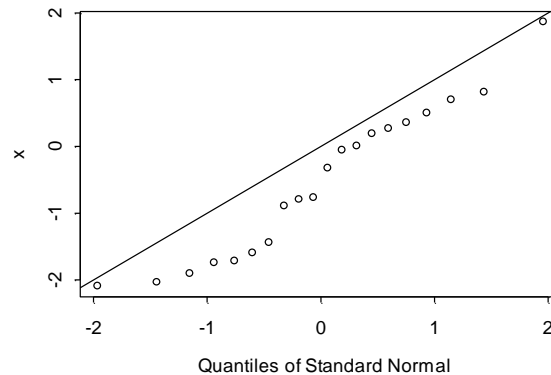
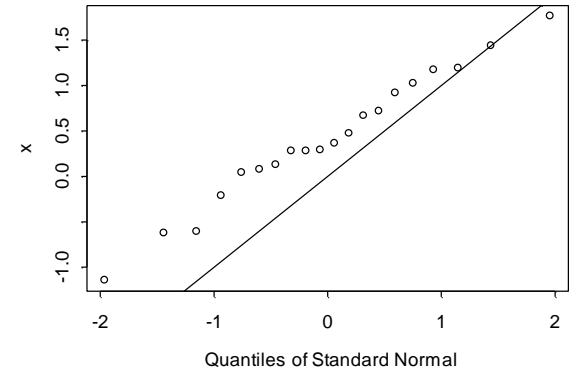
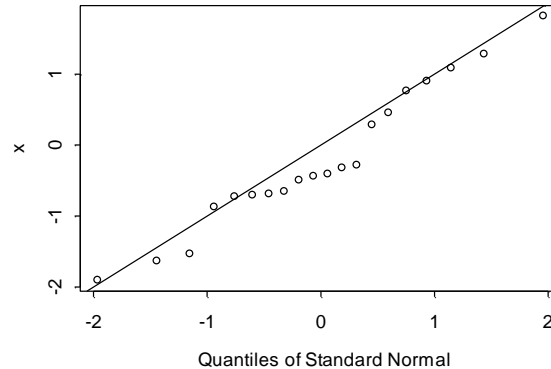
Model diagnostic in R: normal probability plot



# 4 samples of size 20

Example of 4  
qqnormal plot for  
4 random  
samples from  
normal  
distribution with  
 $n=20$ .

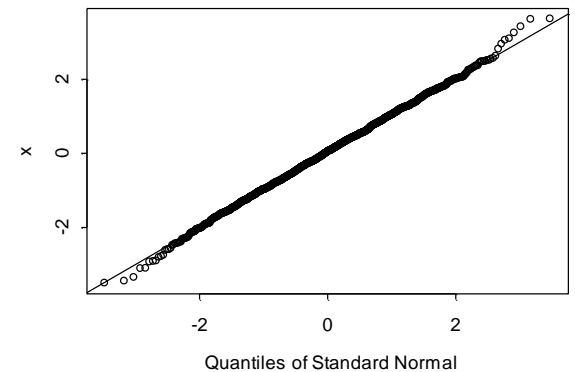
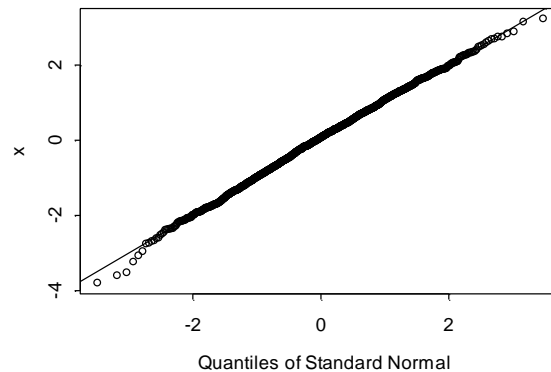
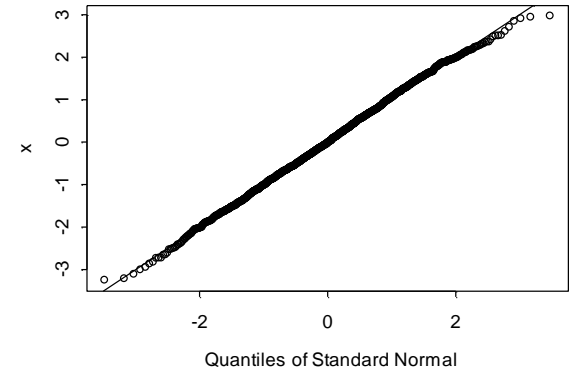
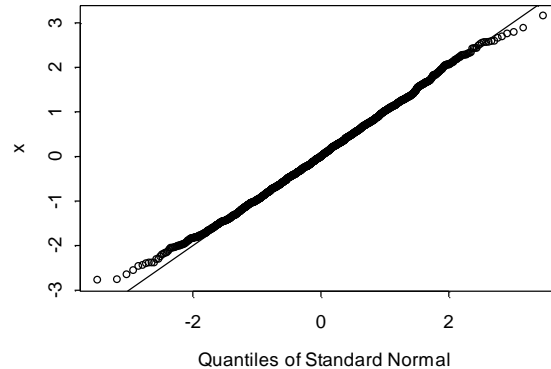
$$U \sim N(0,1)$$



# 4 samples of size 2000

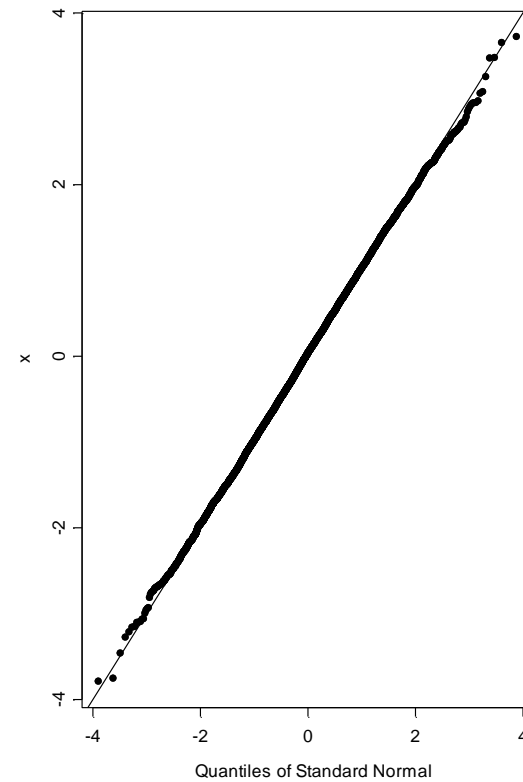
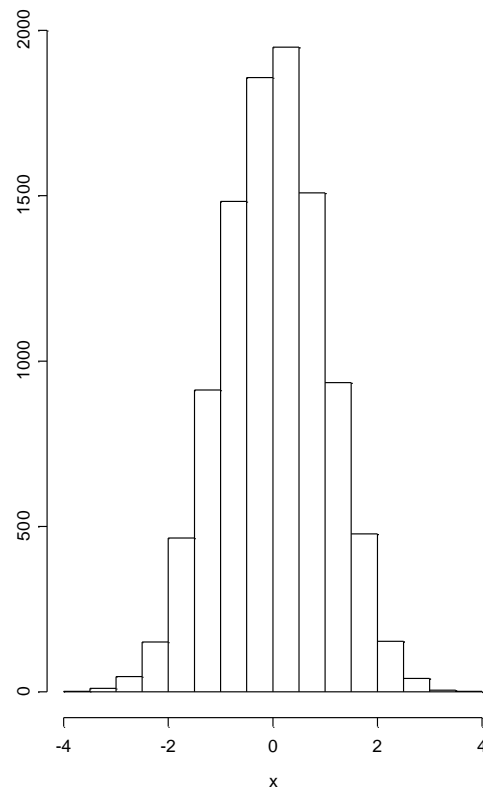
Example of 4  
qqnormal plot for  
4 random  
samples from  
normal  
distribution with  
 $n=2000$ .

$$U \sim N(0,1)$$



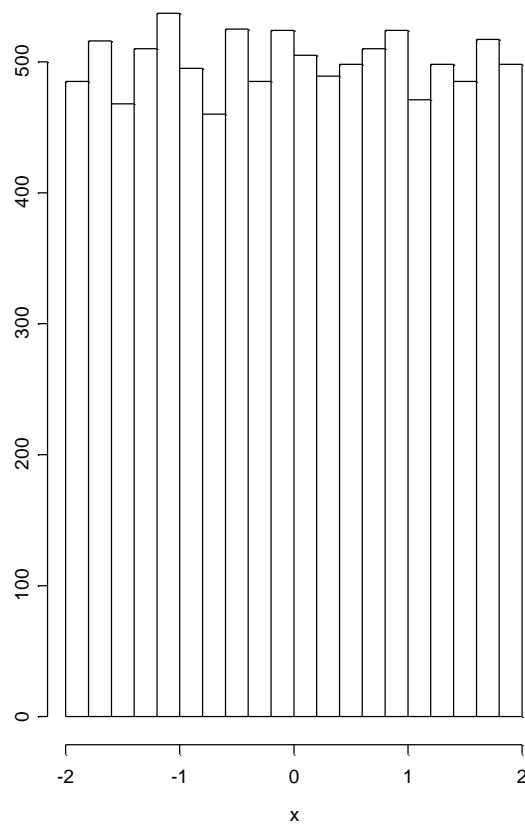
3 Examples of normal probability plots  
for samples of size 10000

$N(0,1)$

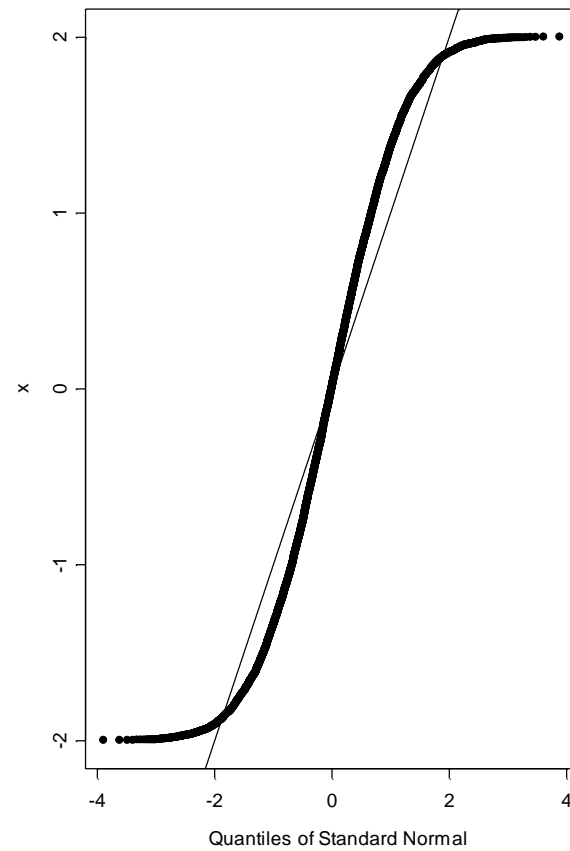


$N=10000$

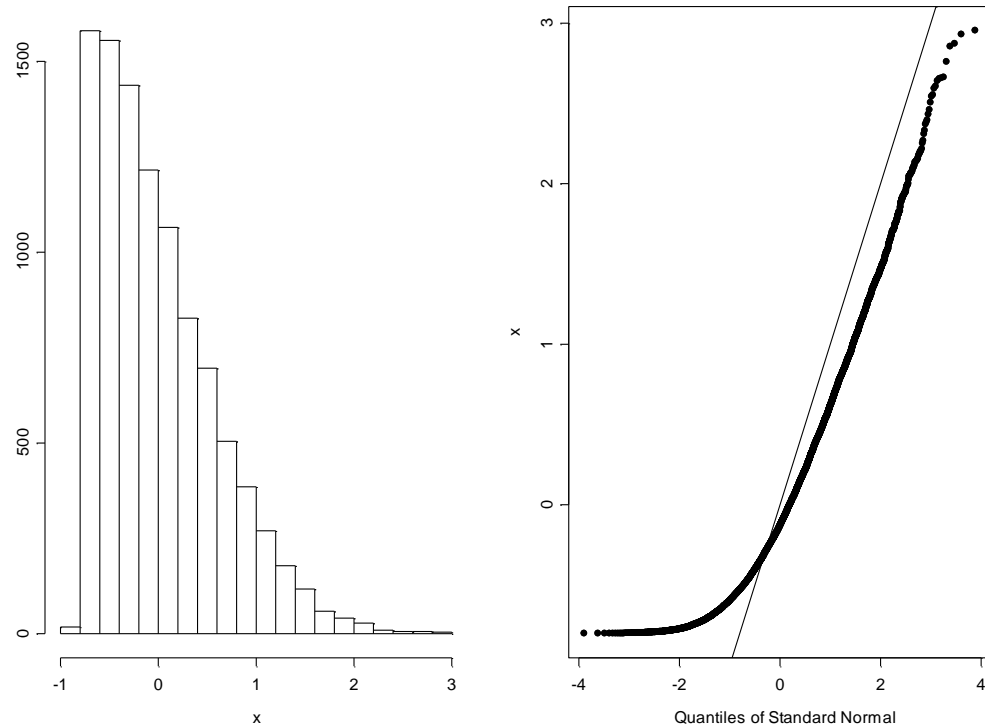
$U(-2,2)$



$N=10000$



# Skewed distribution



N=10000

# Bottom line for model diagnostic

- We use three graphical displays for model diagnostic: dotplot for the response, boxplot for the residuals and qqnormal plot for the residuals.
- In this course, the model diagnostic is only descriptive. Formal test can be used in order to test if the variance is constant or the residuals follow a normal distribution.



## One-Way ANOVA Model: Multiple Testing



# The problem of Multiplicity (1)

- Testing one hypothesis using significant level of 0.05 means that the Type I error is equal to 0.05.
- Type I error is the probability to reject the null hypothesis when the null hypothesis is correct.
- In other words, we make a decision to reject the null hypothesis and the probability that we are wrong is 0.05.

# The problem of Multiplicity (2)

- If we need to test  $k$  null hypotheses simultaneously and we use for each null hypothesis significant level of 0.05 the overall significant level will be much higher.
- This means that when we have a problem with multiple testing (when we need to test more than one null hypothesis) we need to adjust the significant level of each test that we do.

# The problem of Multiplicity (3): bonferroni correction

- There are many ways to adjust for multiple testing
- In this course we use the bonferroni correction.
- Suppose that we have  $K$  tests and we want to keep the overall significant level at 0.05, bonferroni proposed to test each test with significant level of:

The diagram shows the Bonferroni correction formula:  $\frac{\alpha}{K} = \frac{0.05}{K}$ . The entire fraction is enclosed in a solid blue rectangular box. Inside this box, the Greek letter  $\alpha$  is enclosed in a dashed red rectangular box. A red dashed arrow points from the text 'The overall significant level' to the  $\alpha$  in the dashed box. A solid blue arrow points from the text 'The per-comparison significant level' to the  $K$  in the denominator of the left fraction. A solid black arrow points from the '0.05' in the right fraction to the text 'For  $\alpha=0.05$ '.

The overall significant level

$$\frac{\alpha}{K} = \frac{0.05}{K}$$

The per-comparison significant level

For  $\alpha=0.05$

# The problem of Multiplicity (4): bonferroni correction

- For example, if we have 3 tests to do and we want to keep the overall significant level at 0.05

The diagram illustrates the Bonferroni correction formula:  $\frac{\alpha}{3} = \frac{0.05}{3}$ . A red dashed box highlights the  $\alpha$  in the numerator, with a red arrow pointing to the text "The overall significant level". A blue solid box highlights the entire fraction  $\frac{\alpha}{3}$ , with a blue arrow pointing to the text "The per-comparison significant level". A black arrow points from the equals sign to the text "For  $\alpha=0.05$ ".

The overall significant level

The per-comparison significant level

For  $\alpha=0.05$

- This means that for each test we use significant level of  $0.05/3=0.01666667$

# The null hypothesis for the F-test

- For one-way ANOVA model with three levels' factor the null hypothesis is

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

- The null hypothesis states that the three populations' mean are equal.

- If we reject the null hypothesis we conclude that the populations' means are not equal BUT we do not know which population is different from the other.
- For example, it could be that populations' mean 1 and 2 are not different and they are both different from the third population mean.

# The null hypothesis for the F-test

- If we reject the null hypothesis we might want to test each pair of means.
- This means that we have three tests to perform and if we want to keep the overall significant level at 0.05 we need to test each hypothesis at significant level of  $0.05/3$ .

$$H_{0,1} : \mu_1 = \mu_2$$

$$H_{0,2} : \mu_1 = \mu_3$$

$$H_{0,3} : \mu_2 = \mu_3$$

# Example 3: the Data

```
> tapply(Data3$Response, list(Data3$Treat), mean)
```

```
      1      2      3  
10.20000 10.18000 12.47833
```

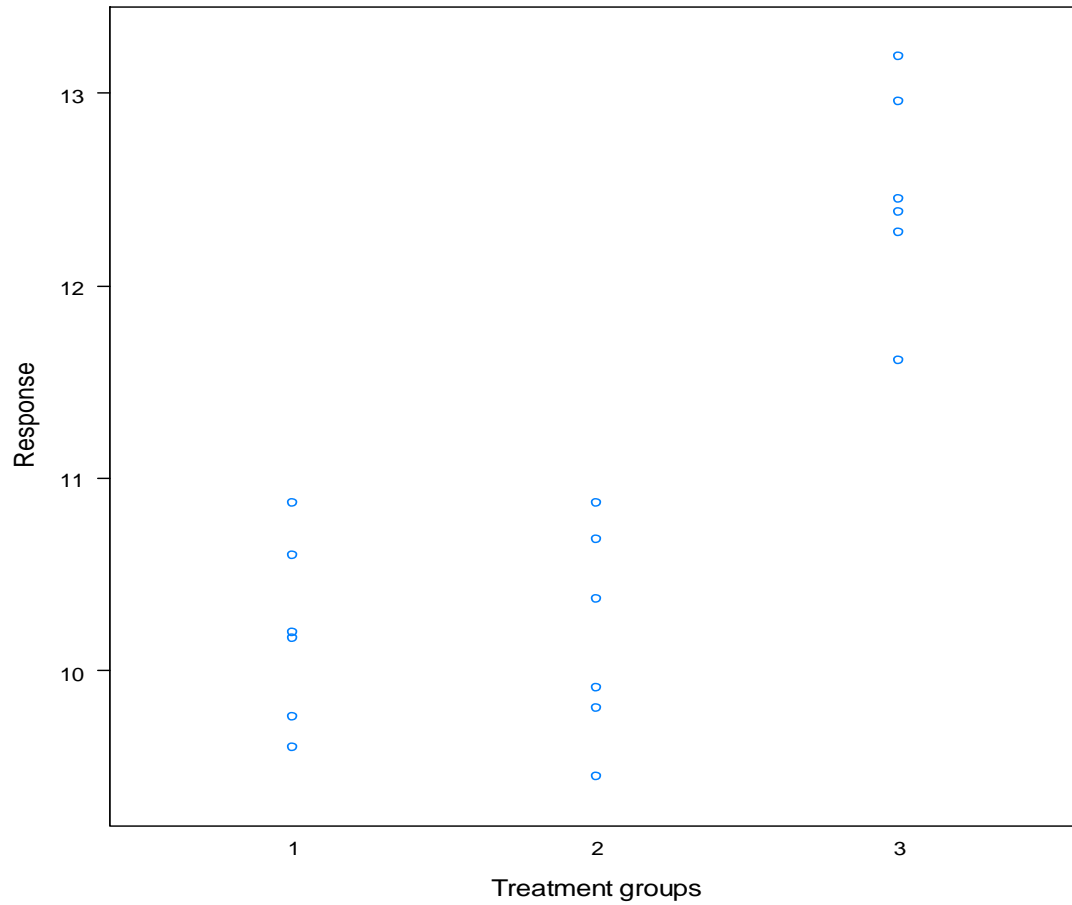
We can see that the sample means in group 1 and 2 are very closed to each other (10.20 and 10.18).

These patterns reveal also in the following two plots.

The Data

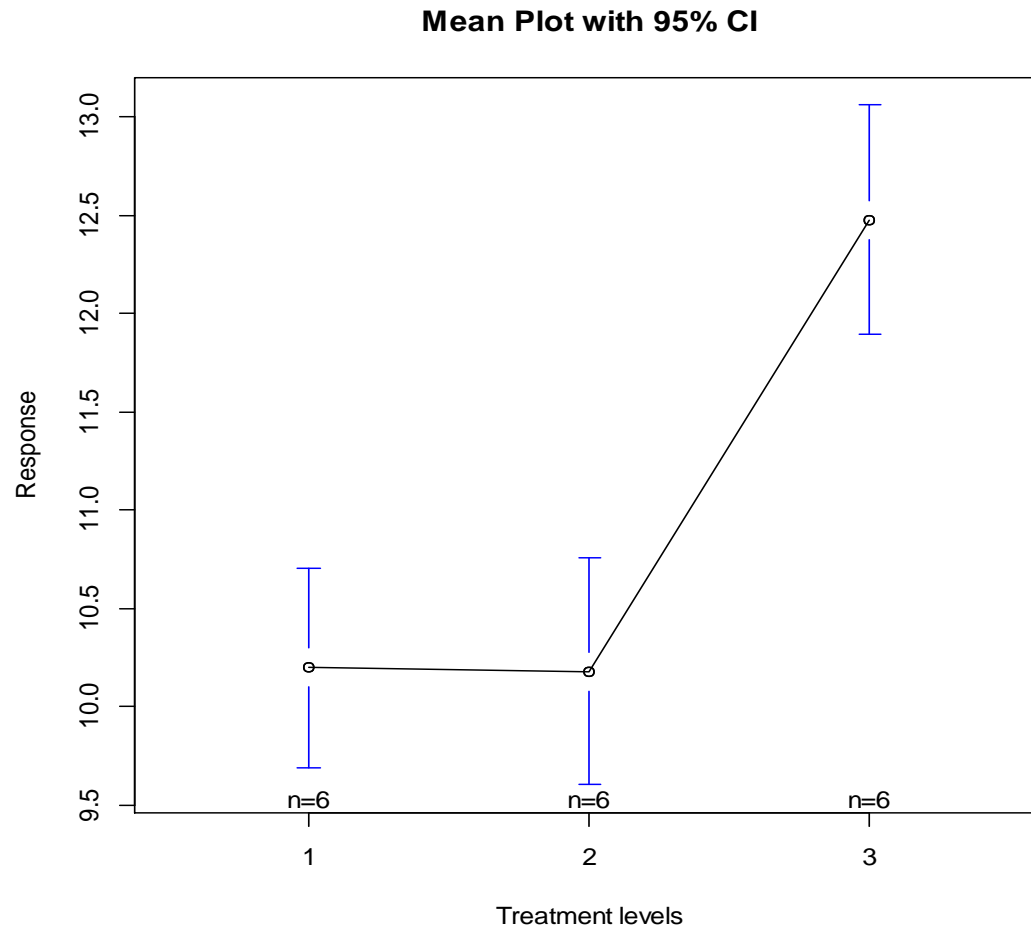
Treat		Response
1	1	9.76
2	1	10.60
3	1	10.20
4	1	10.87
5	1	10.17
6	1	9.60
7	2	10.68
8	2	10.87
9	2	9.91
10	2	9.80
11	2	9.45
12	2	10.37
13	3	12.28
14	3	12.38
15	3	12.45
16	3	12.96
17	3	11.61
18	3	13.19

# Dotplot by treatment group





# Sample means by treatment group



# Inference

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

The null hypothesis is **rejected**. We can conclude that populations' means are not equal.

The patterns reveal in the plots indicates that it could be that not all populations' means are different.

In the next step we would like to test each pair separately.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Treat	1	15.572	15.57	25.97	0.000108 ***
Residuals	16	9.593	0.60		

---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

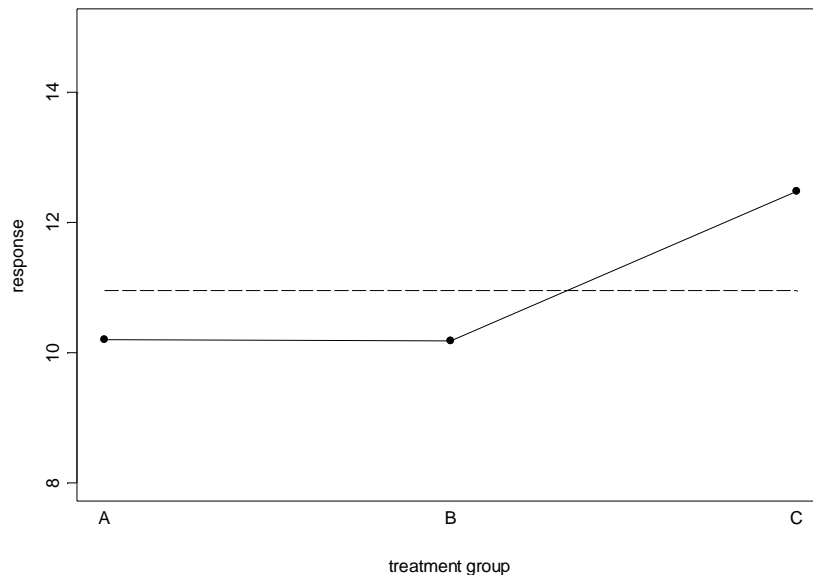
```
> fit.Data3 <- aov(Response ~ Treat, data = Data3)
> summary(fit.Data3)
```

The null hypothesis was rejected, what does it mean ?

- It mean that there are differences among the populations' mean.
- It does not mean that all the populations' means are equal.

# Multiple comparisons

- The next step is to test each pair.
- We have 3 tests to perform.
- Each test will be performed in significant level of  $0.05/3$ .



$$H_{0,1} : \mu_1 = \mu_2$$

$$H_{0,2} : \mu_1 = \mu_3$$

$$H_{0,3} : \mu_2 = \mu_3$$

# Using function `pairwise.t.test()` for multiple comparisons

Pairwise comparisons using t tests with pooled SD

data: Response and Treat

	1	2
2	1	-
3	6.2e-06	5.6e-06

P value adjustment method: bonferroni

Treatment 3 is different from treatments 1 and 2

Treatment 1 is not different from treatment 2

```
> pairwise.t.test(Response, Treat,  
+   p.adj = "bonferroni", data = Data3)
```

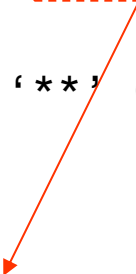
The bonferroni adjustment

The null hypothesis that the means in treatment 1 and 2 are equal is not rejected.

Each P-value should be compared with  $0.05/3$  !!!

# aov ( ) output: global null hypothesis

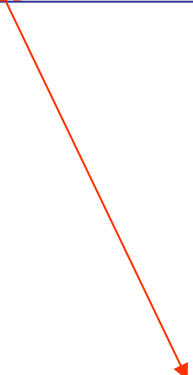
```
      Df Sum Sq Mean Sq F value    Pr(>F)
Treat   1 15.572    15.57    25.97 0.000108 ***
Residuals 16  9.593     0.60
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.'
                 0.1 ' ' 1
```



Global F test: the null hypothesis  $H_0 : \mu_1 = \mu_2 = \mu_3$  is rejected.

# Multiple comparisons with option `p.adj = "none"`

```
> pairwise.t.test(Response, Treat,  
+   p.adj = "none", data = Data3)
```



T tests for all pairwise comparisons, however it does not make correction for the Type I error rate across the pairwise tests.

Pairwise comparisons using t tests with pooled SD

data: Response and Treat

	1	2
2	0.95	-
3	2.1e-06	1.9e-06

P value adjustment method: none

$H_{0,1} : \mu_1 = \mu_3$

$H_{0,2} : \mu_2 = \mu_3$

$H_{0,3} : \mu_1 = \mu_2$

The diagram illustrates the mapping between the p-values in the table and the corresponding null hypotheses. A red line starts from the p-value 0.95 (comparison between groups 1 and 2) and points to the hypothesis  $H_{0,3} : \mu_1 = \mu_2$ . A blue line starts from the p-value 2.1e-06 (comparison between groups 1 and 3) and points to the hypothesis  $H_{0,1} : \mu_1 = \mu_3$ . A green line starts from the p-value 1.9e-06 (comparison between groups 2 and 3) and points to the hypothesis  $H_{0,2} : \mu_2 = \mu_3$ .



# More about Multiplicity

- A link for an online course about multiple testing (at more advanced level):

<https://github.com/eR-Biostat/Courses/tree/master/Under%20graduate%20courses/Basic%20concept%20in%20statistical%20inference%20using%20R%20-II>