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

- Statistics: 2011-2016, 2017.
- Statistics: 2017.
- Statistics for development : 2018-2022.

The >eR-Biostat initative

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)

https://erbiostat.wixsite.com/erbiostat

UPDATED: 2022

Visit us on Facebook ER-BioStat

GitHub https://github.com/eR-Biostat

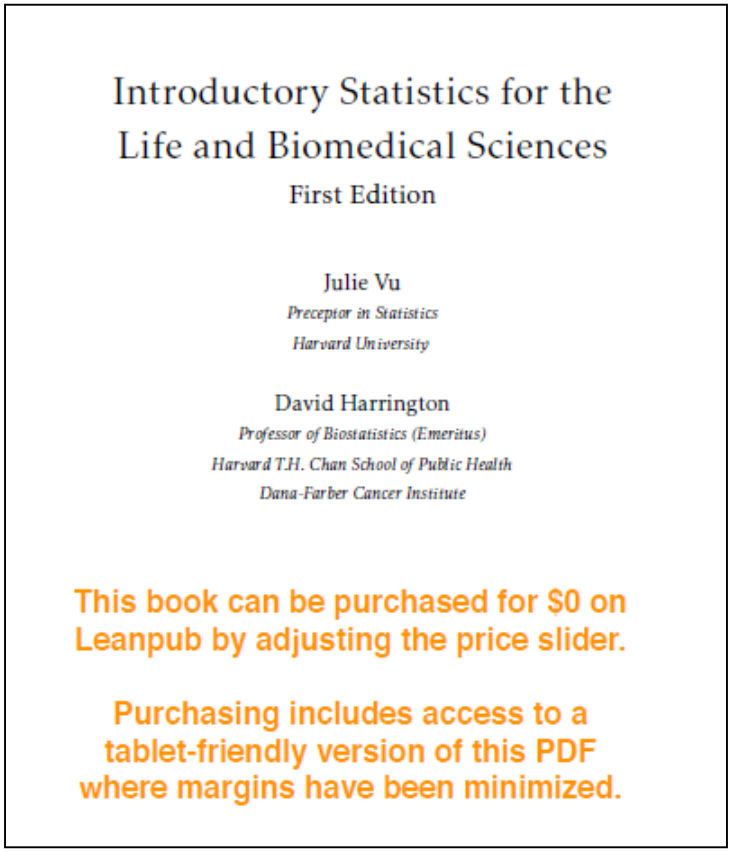
Email: erbiostat@gmail.com

twitter @erbiostat

contents

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

Recommended reading


Introductory Statistics for the Life and Biomedical Sciences

First Edition

Julie Vu
Preceptor in Statistics
Harvard University

David Harrington
Professor of Biostatistics (Emeritus)
Harvard T.H. Chan School of Public Health
Dana-Farber Cancer Institute

This book can be purchased for $0 on Leanpub by adjusting the price slider.

Purchasing includes access to a tablet-friendly version of this PDF where margins have been minimized.

- In this part of the course, we cover mainly Section 5.5 in the book.
- The examples that are used for illustration are not the same as the examples in the book.
- The book is available for free online:

  https://www.openintro.org/book/biostat/

Section 5.5: Comparing means with ANOVA

Graphical displays

- Graphical displays in the course were produced using the old R functions for graphical displays.
- Better figures can be produced using te ggplot2 package.

Introduction:
The one-way ANOVA model

Example: 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

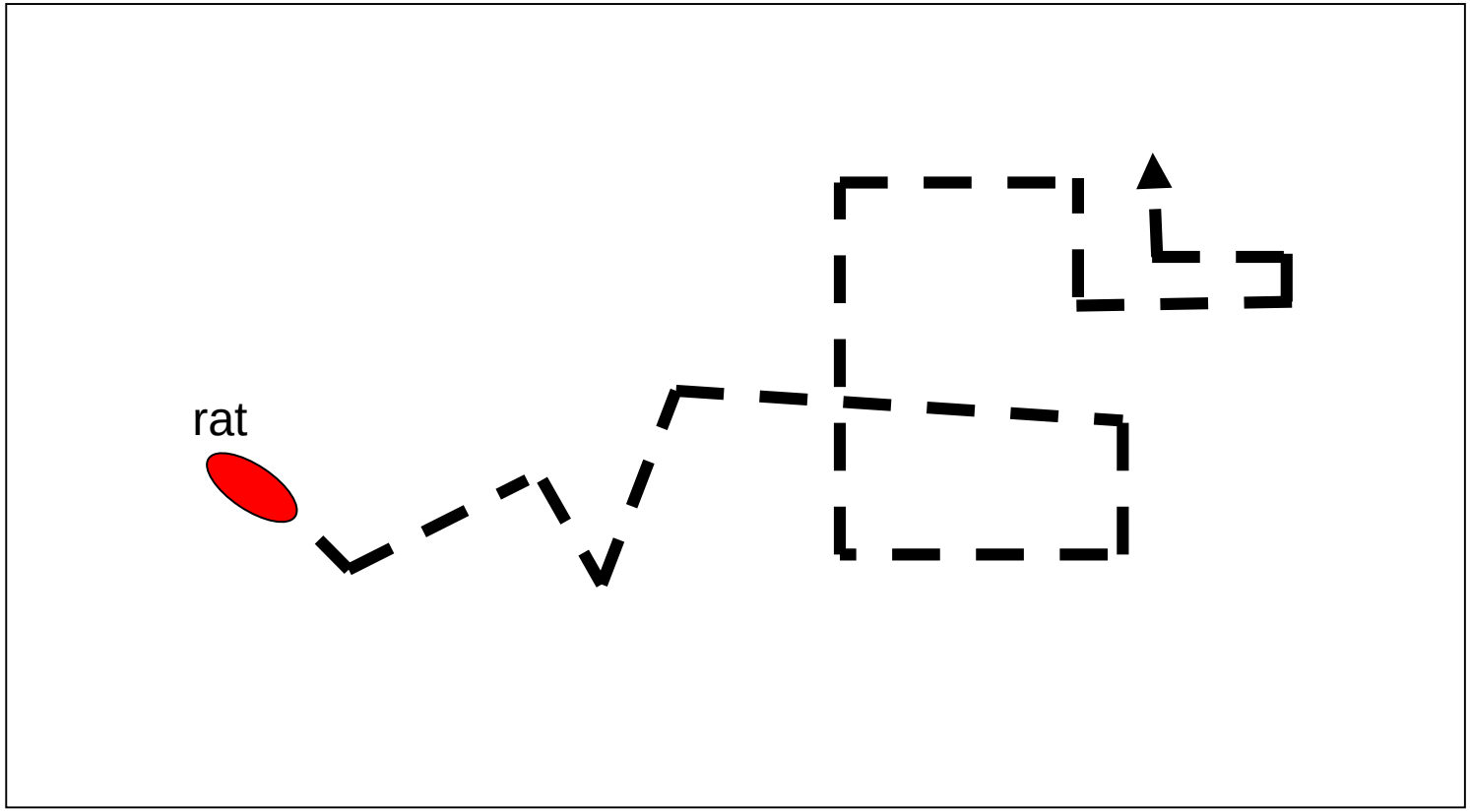
```
> print(Biophar)
Obs  Treat     Dist
22    QNP 186.6145
11    QNP 103.3529
 4    QNP 191.3850
16    QNP 334.9845
 7    QNP  89.2831
13    QNP 345.5070
 2    QNP 169.5161
20    QNP 173.1491
19    QNP 130.9634
 8    QNP 363.4392
10    QNP  76.5340
24    QNP 202.1145
 1 SALINE  12.8458
17 SALINE  44.3092
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

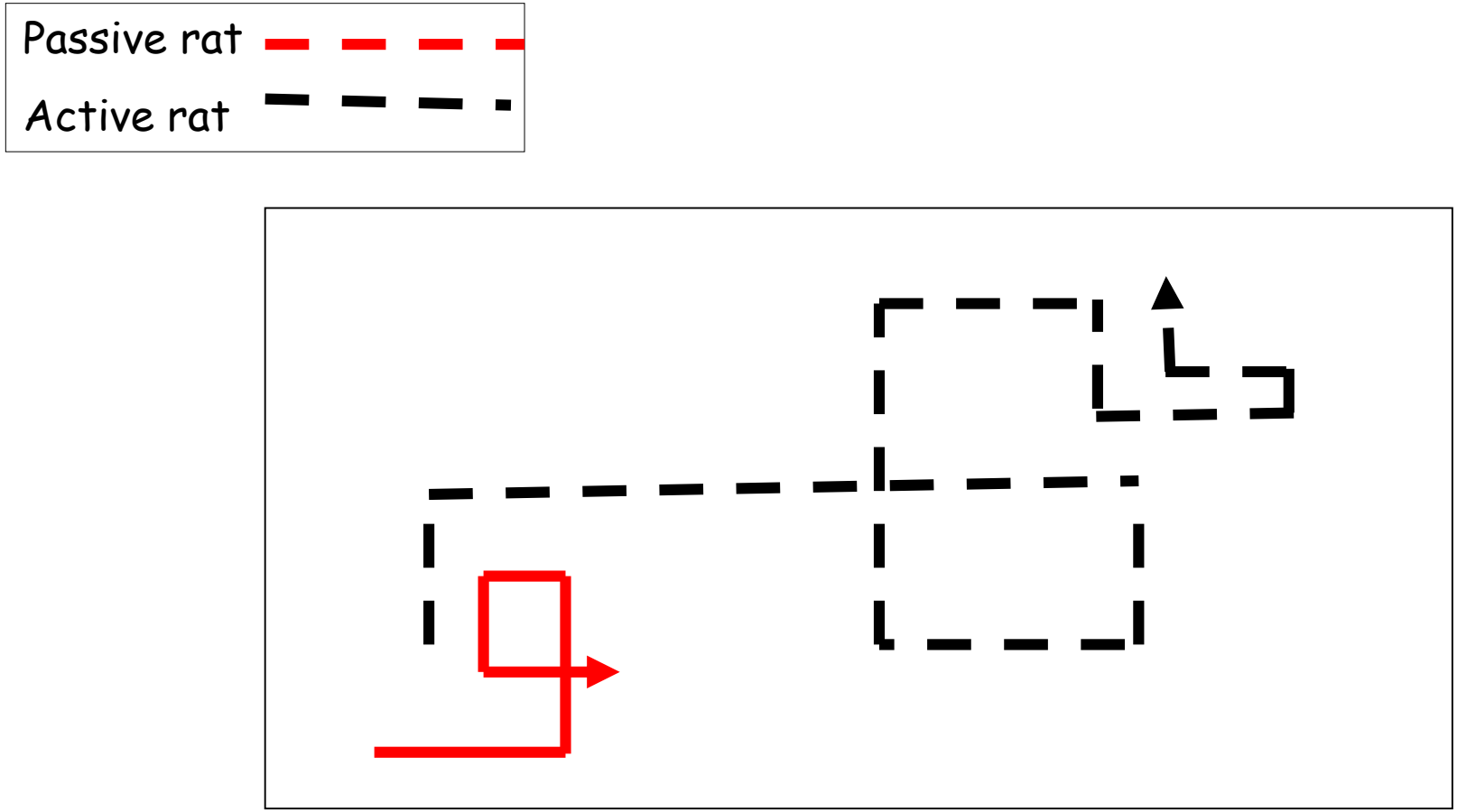
Treatment group

The evaluation of the rat

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



Description of the experiment



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

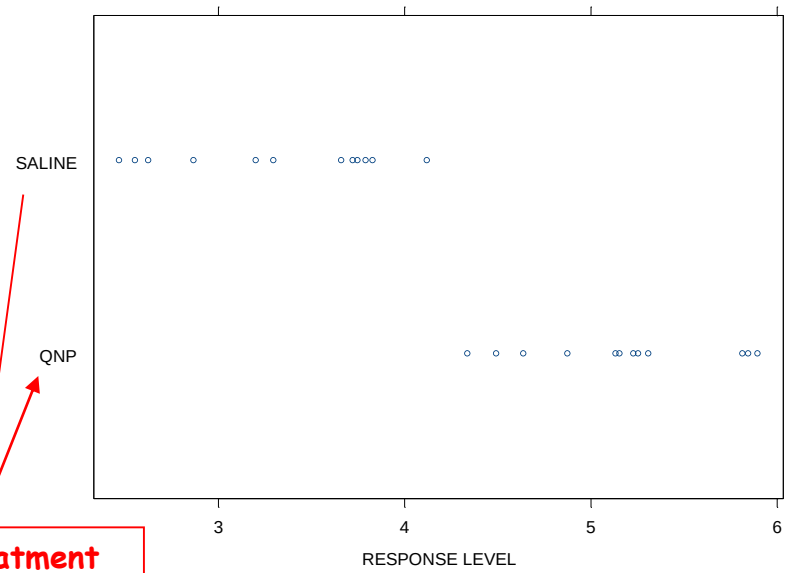
The scientific question

- 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

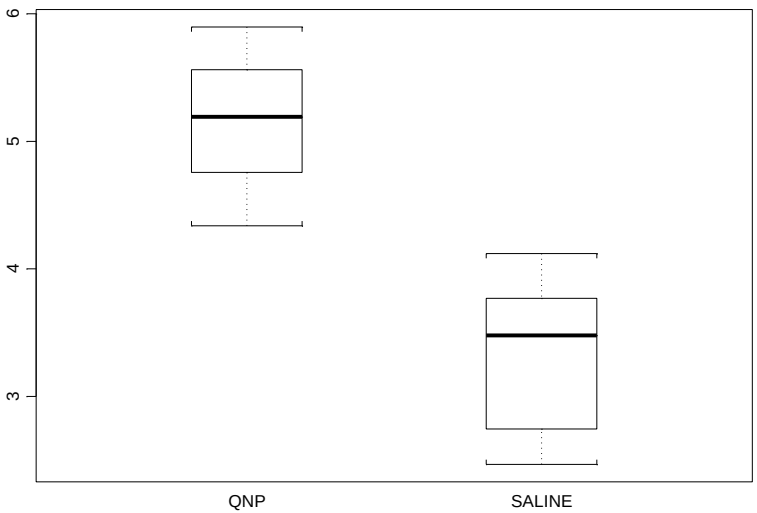
Response

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 distanc traveled on log scale)

> summary(Biophar$ldist)

Groups' means

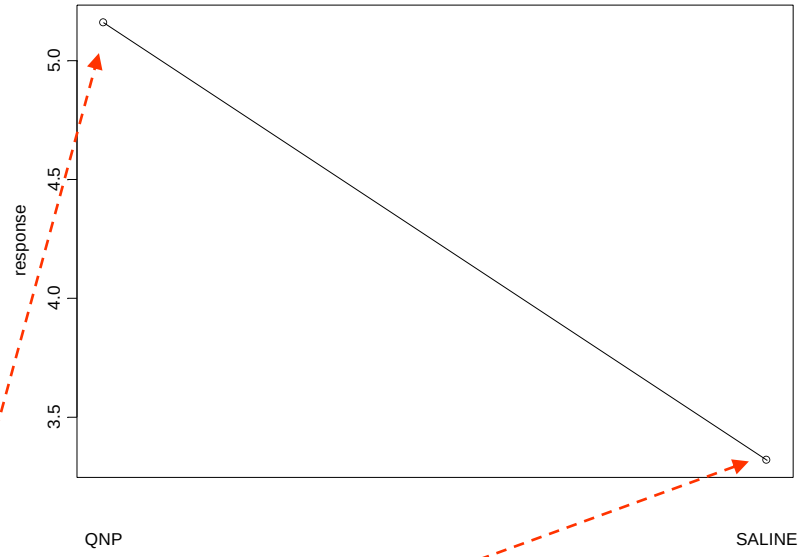
```
> tapply(ldist,Biophar$Treat,mean)
    QNP   SALINE
5.164716 3.323143

> tapply(ldist,Biophar$Treat,sd)
     QNP    SALINE
0.5179239 0.5719953
```

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

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

Graphical dispaly (3)

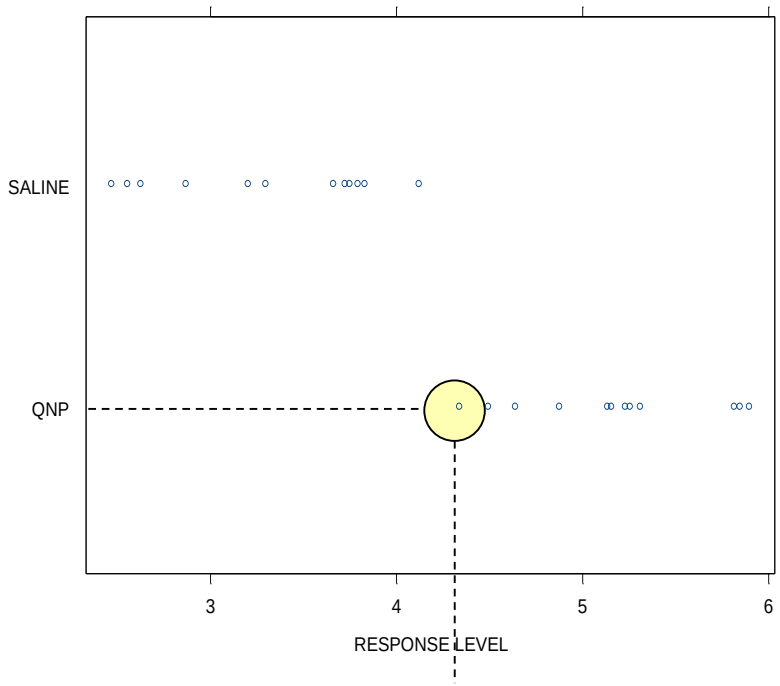
The mean plot, shows the mean at each factor level.



treat	N Obs	N	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

ANOVA terminology

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



(X_{11}, Y_{11})

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)

$$y_{ij}$$

Group number: i=1,2

Observation number:

j=1,2,3,…,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

$Y2_{212}$: Observation number 12 in group 2

Data Structure: notation (2)

Number of Group: I

Sample size: n

$n=n_1+n_2+,\ldots,n_k$

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

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

Sample size in group i: n_i

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

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

Discriptive statistics

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

Overall mean:

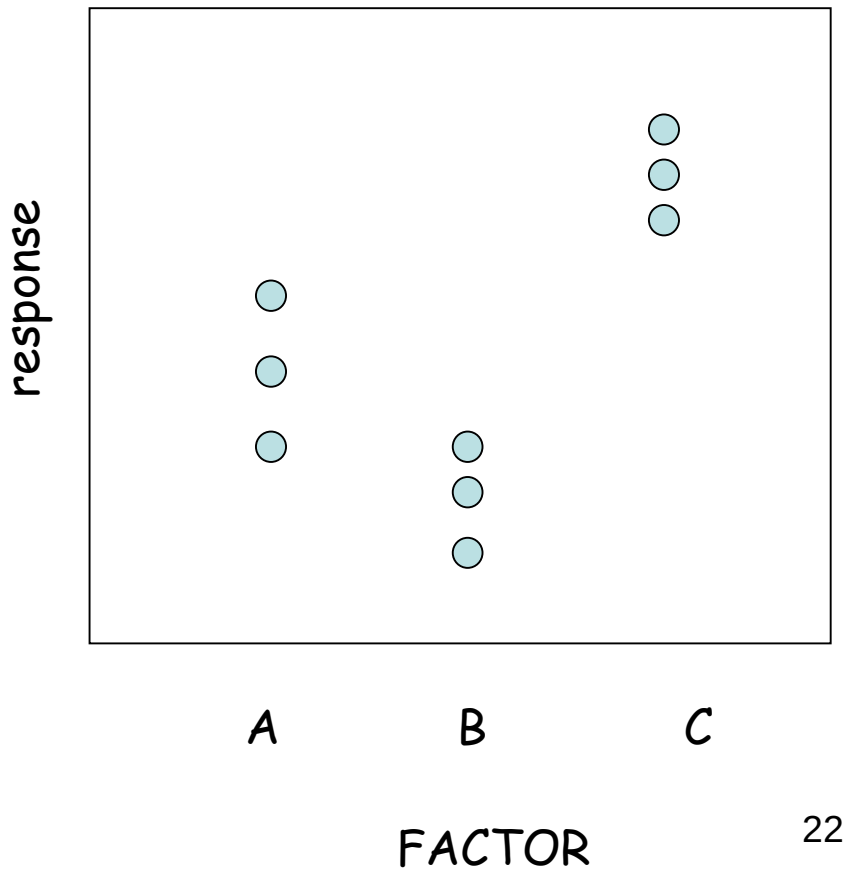
Mean of group i:

```
>
tapply(ldist,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 variabilty=variability within the groups + variability between the groups

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

A very simple example

- A one factor experimant.
- The factor has three levels (1,2,3).
- Three observations at each level.

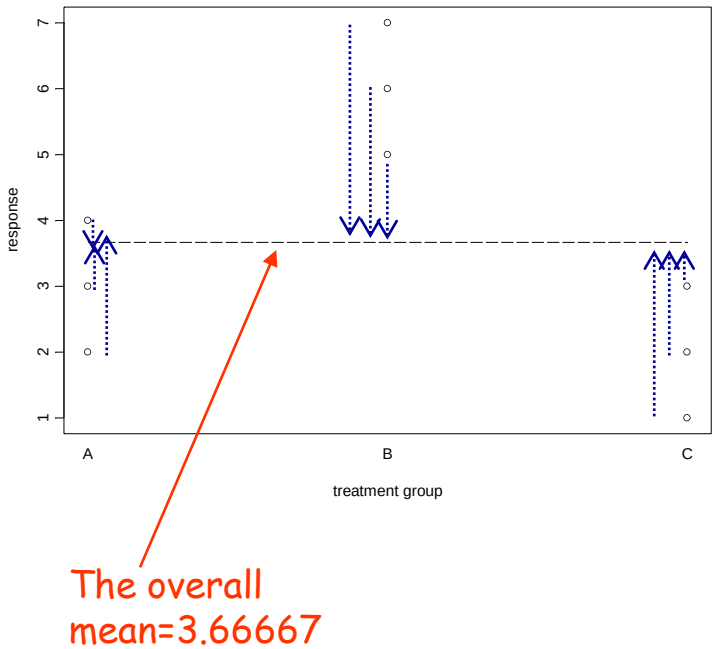
group	Y_{ij}	Group mean
1 1 1	2 3 4	3
2 2 2	5 6 7	6
3 3 3	1 2 3	2

Overall mean: 3.6666

Two Sources of Variability: the total variability

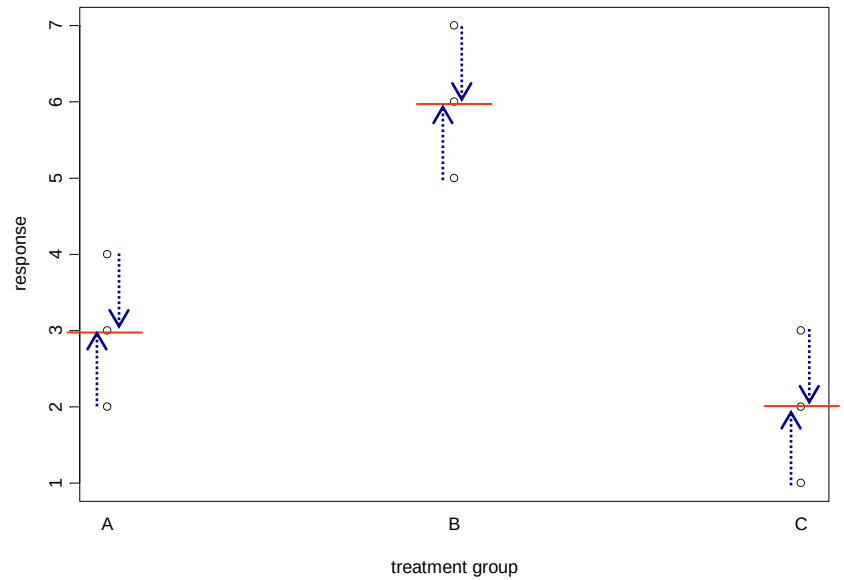
The total sum of squares (SST) is the sum of squared distance between the observations of the overall mean.

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



Two Sources of Variability: the variability within the groups

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



A $(2-3)^2+(3-3)^2+(4-3)^2=2$

B $(5-6)^2+(6-6)^2+(7-6)^2=2$ +

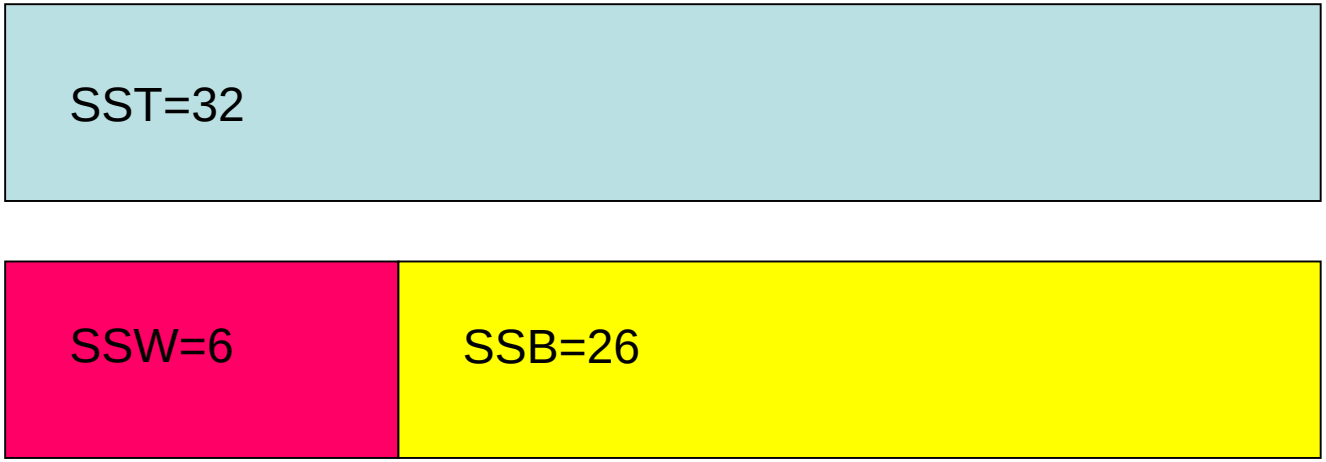
C $(1-2)^2+(2-2)^2+(3-2)^2=2$ +

6

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

Two Sources of Variability

Total variability



Variability within the groups

Variability between the groups

$$SST = SSW + SSB$$

Another notaion: SST=SSE+SSTR

Two Sources of Variability: the variability between the groups

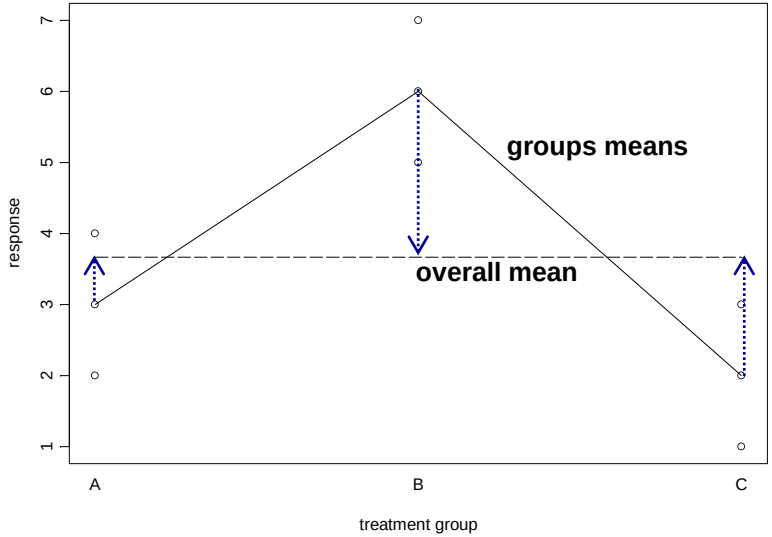
The sum of squares between the groups in the sum of squared diffrence between the group means to the overall mean.

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

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 different in the mean response across the factor levels.
  - The second is an example of a dataset in which the means are not equal.
- Mind that: when we say “the means” we means the parematers (or the population mean) and NOT the sample means !!!!!

Hypothetical experiments

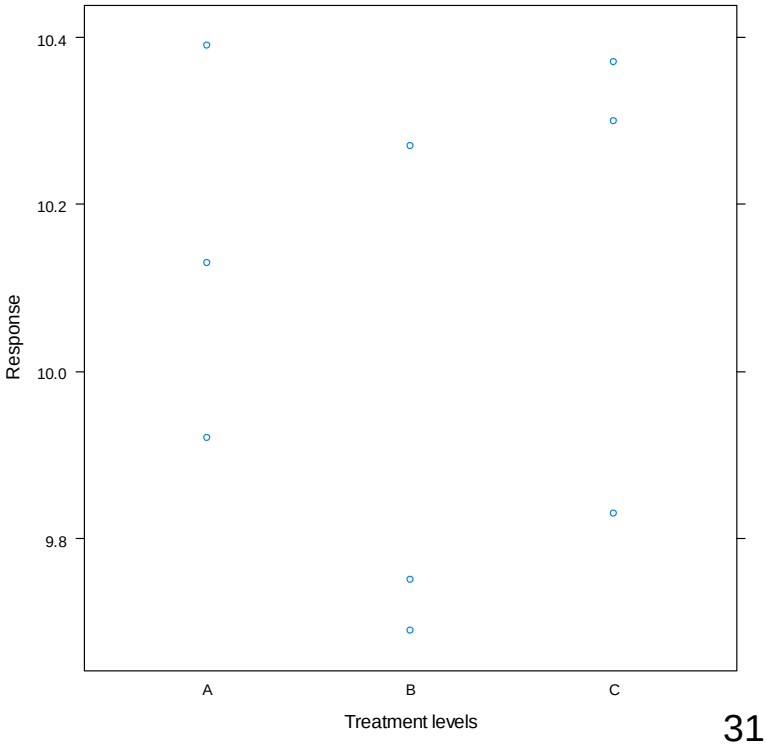
- We consider a one factor experiment in which the factor has thee levels: A, B and C.
- There are 3 observations at each factor level.

- Sample size: N=9.
- Sample size per group: n_i=3, i=1,2,3.
- Number of groups: I=3.

Example 1: data1

Groups' means and dot plot

```
> 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
```

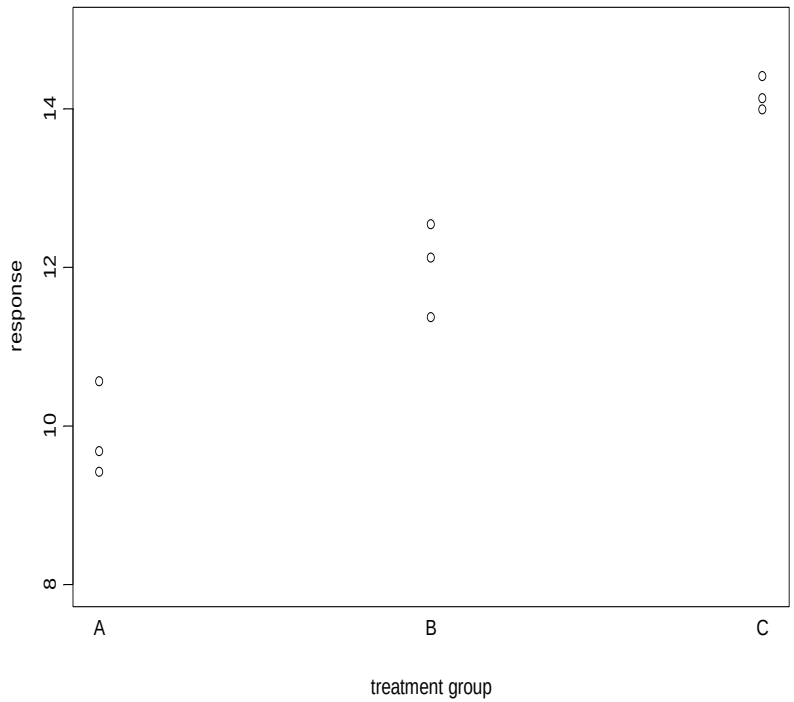


EXAMPLE 1

Example 2: data2

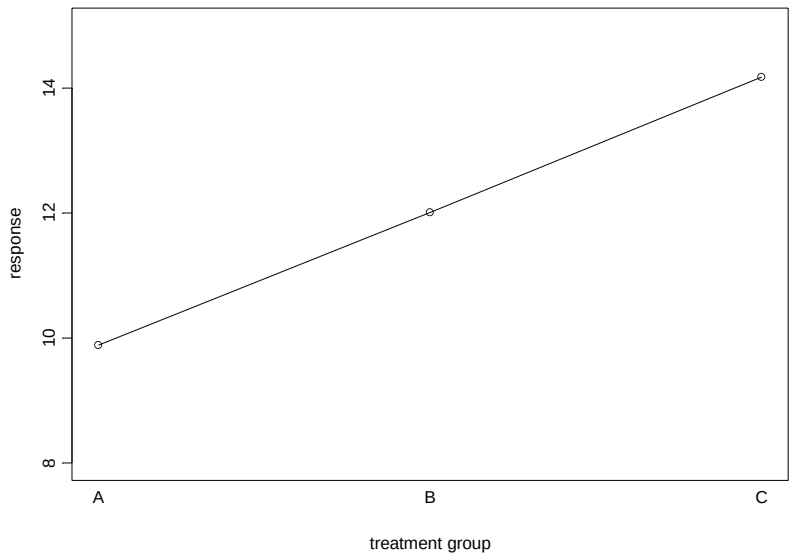
Groups' means and dot plot

```
> 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
```

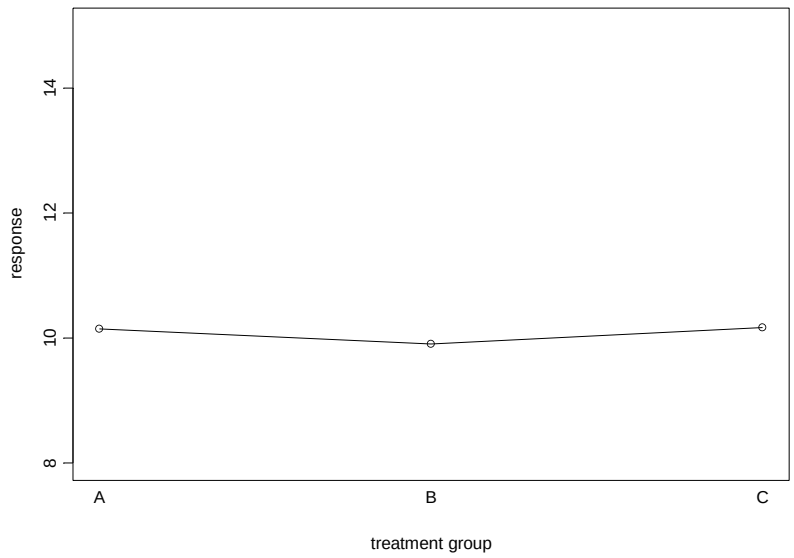


EXAMPLE 2

Means plot



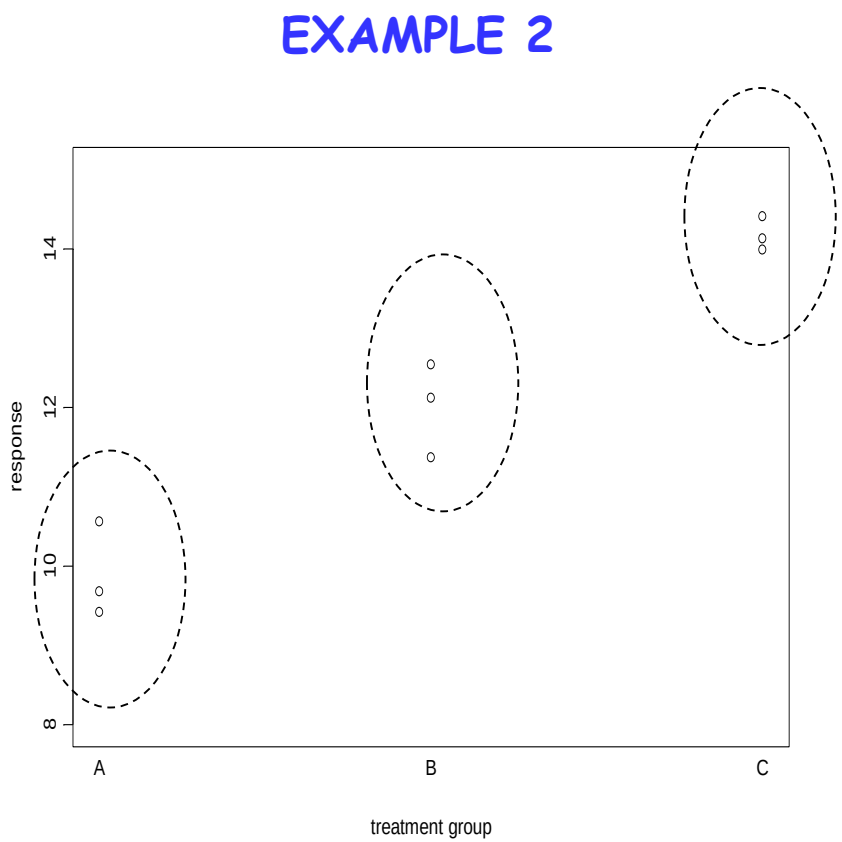
EXAMPLE 2



EXAMPLE 1

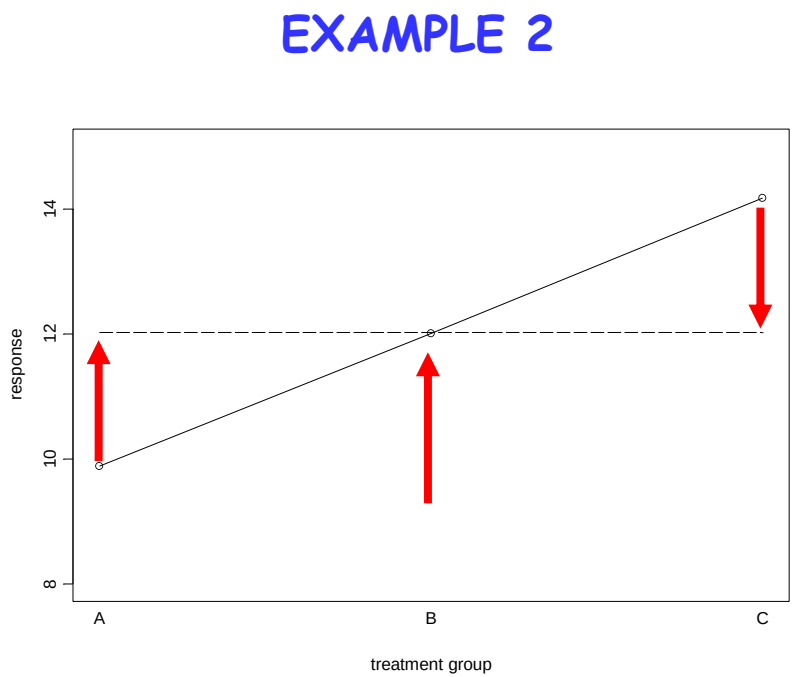
Example 2: the variability within the group

- The variability with a group is the sum of sqaures between the response the group's 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: the variability between the groups

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



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

Output

```
> 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.

EXAMPLE 2

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 Tota	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

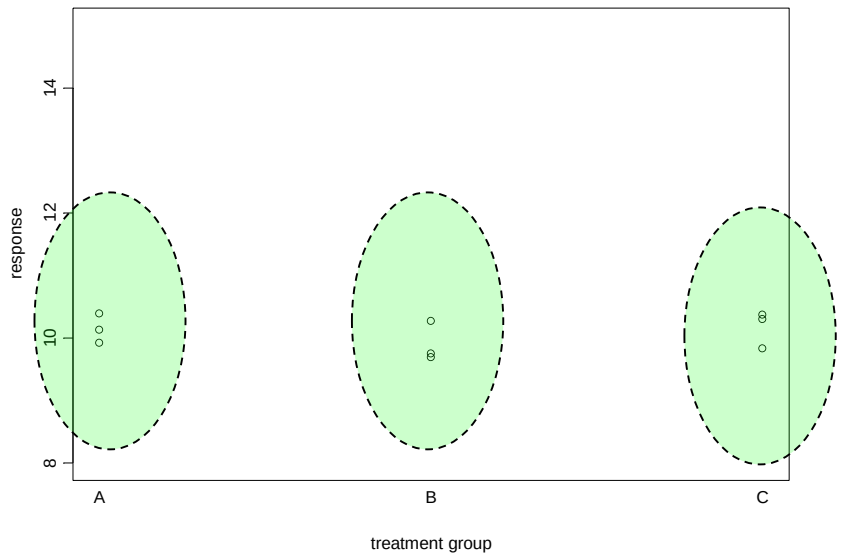
The variability between the groups

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
```

EXAMPLE 1



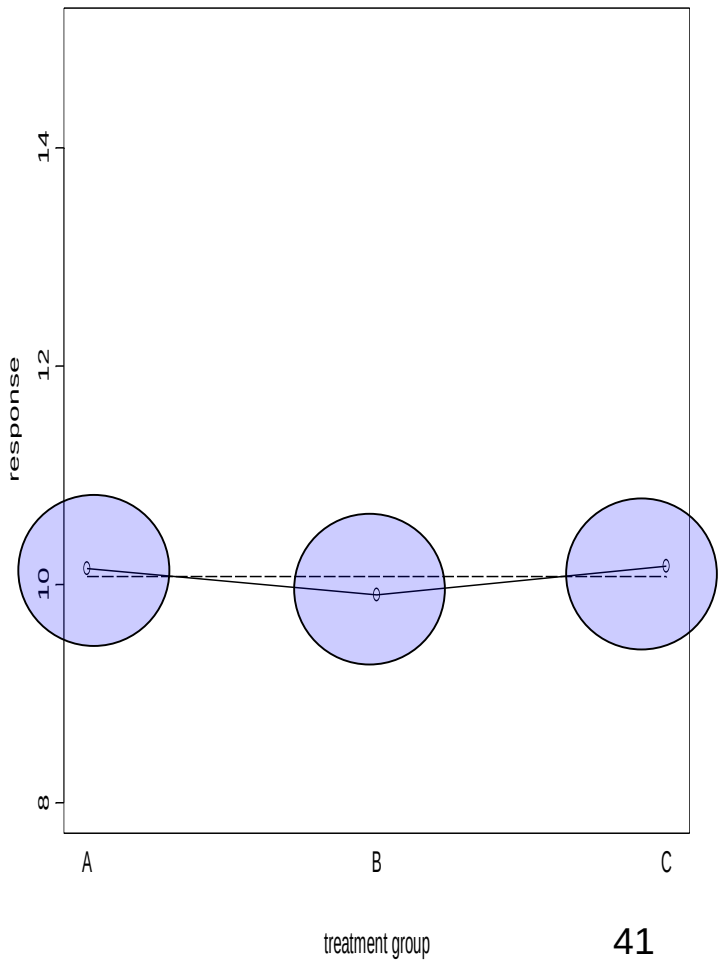
SSW: The error sum of squares is the within group sum of sqaures. This is the within group source of variability.

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 sqaures. This is the 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

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

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

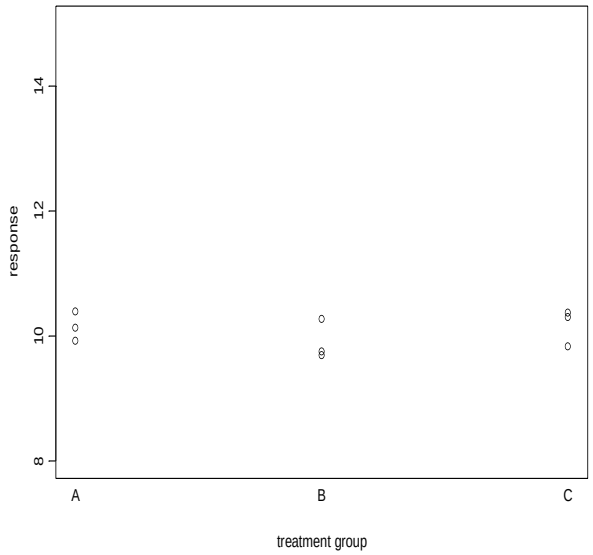
0.48680000 / 6 = 0.08113333

EXAMPLE 1

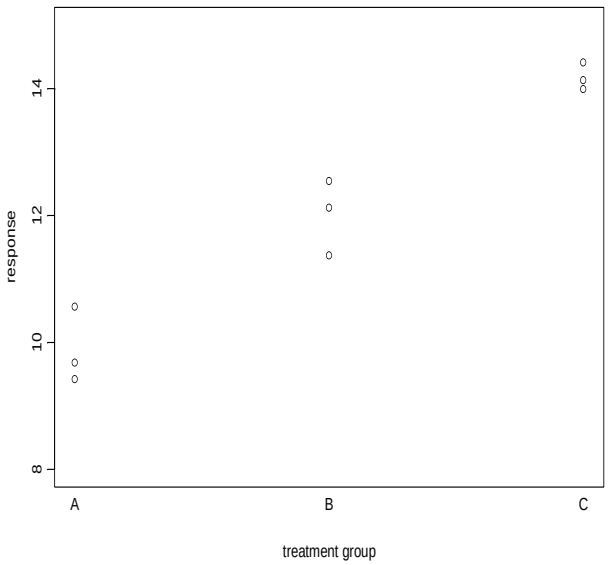
...and now to the big question...

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

EXAMPLE 1



EXAMPLE 2



What is the main diference 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 varaince.

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

Model assumptions are:

1. The random error is nornal distributed.
2. The varaince 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 ($\bar{\mu}_i$) are equal, not if the sample means ($\bar{Y}_i$) are equal.

Test Statistic

Within group sum of squares

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

Between group sum of squares

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

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

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

$$\frac{\text{Within group sum of squares/dgree of fredom}}{\text{Between group sum of squares/dgree of fredom}} = \frac{SSB/(I-1)}{SSW/(N-I)} = \frac{\boxed{MSB}}{\boxed{MSW}} = F$$

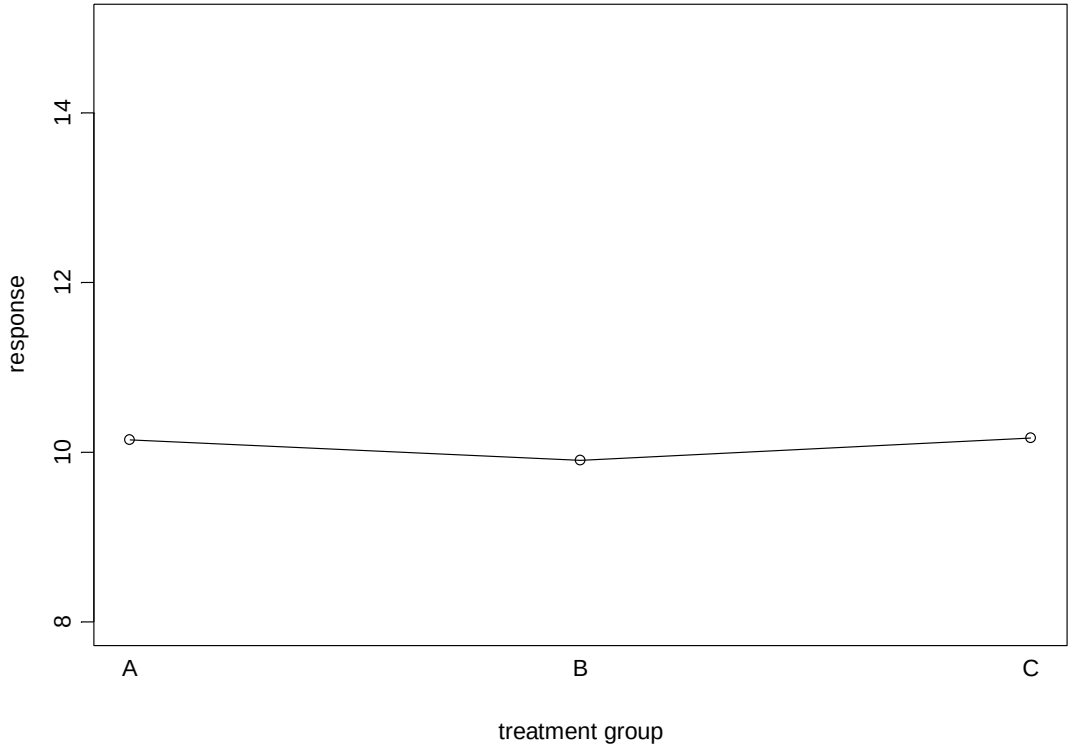
```
> summary(fit.data1)

           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 factor levels, than we expect that SSB will be relativly 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 ?



EXAMPLE 1

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.

$$F \sim F_{(k-1,N-K)}$$

The test statistic

The distribution under the null hypothesis is F(k-1,N-k)

Inference

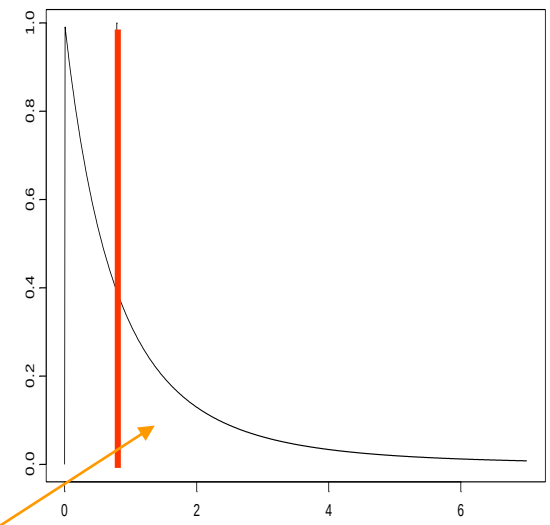
The decision rule:

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-value=0.49 > 0.05.

F(2,6)



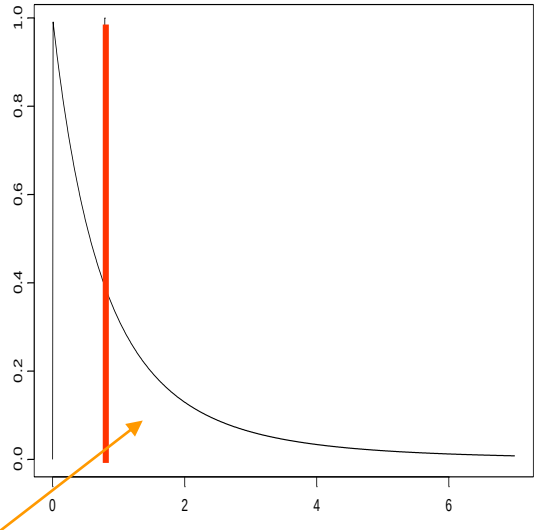
K-1=2, N-K=6

	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 plot shows the distribution of the test statistic (the desity function) under the null. The reed line is the observed value of the test statistic, i.e. 0.79.
- The P-value is the probability to obsrved an exterem 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)



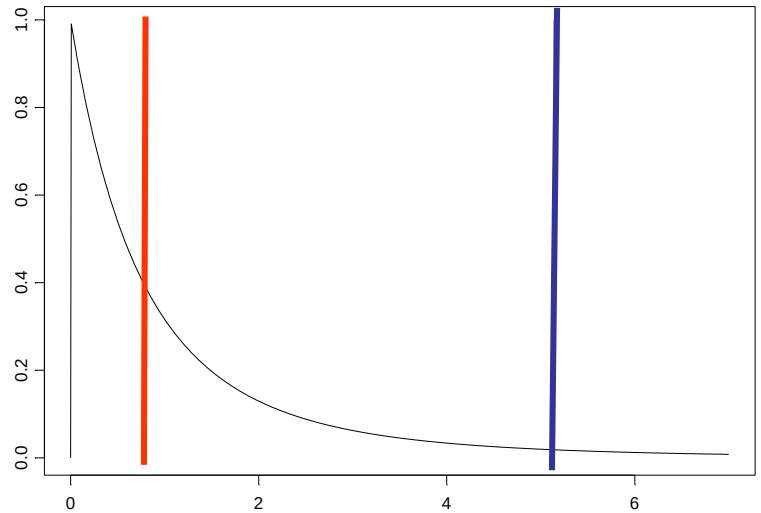
I-1 = 2, n - I = 6

	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 pobability to observed the value of the test statistics (0.79) under the null hypotheses.
- If P-value > 0.05 we do not reject the null hypothesis.

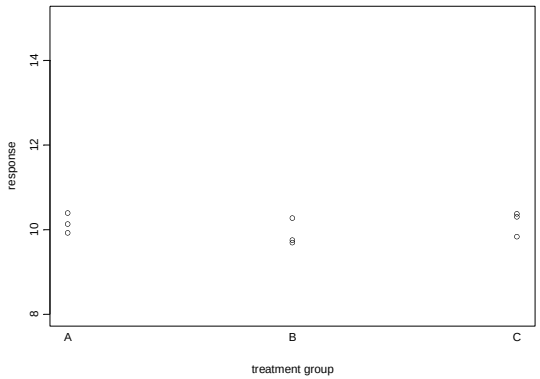
Critical value:



	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		

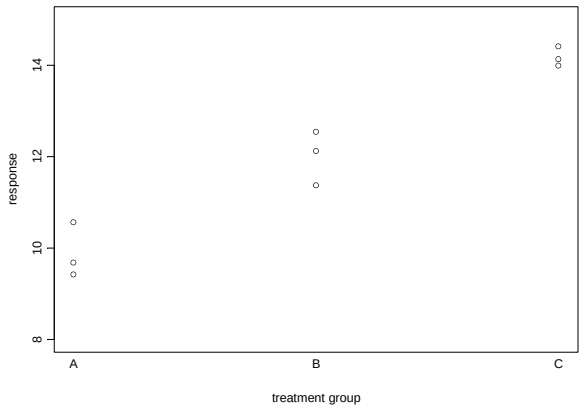
Using the ANOVA table to test the null hypothesis

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

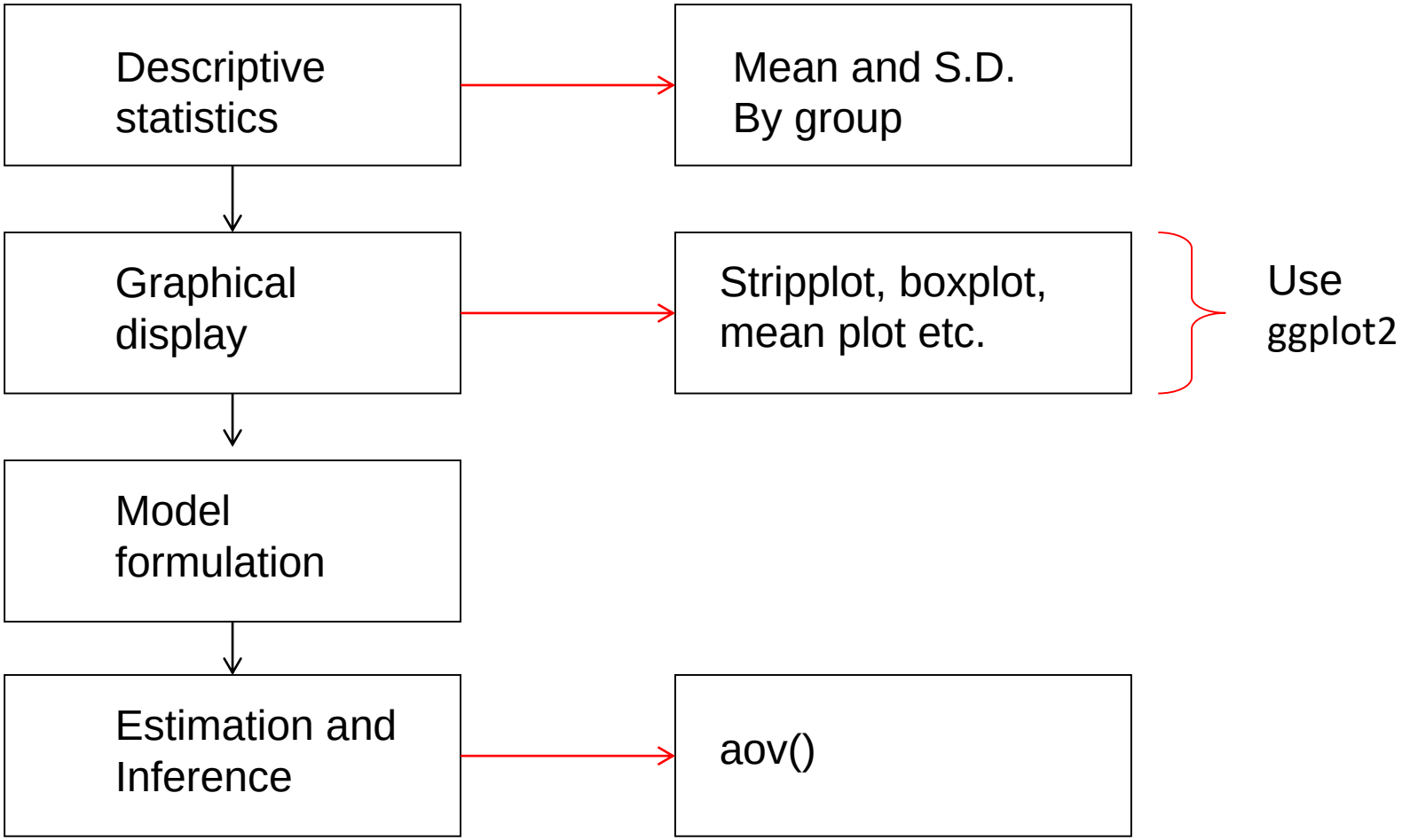


	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
Treat	2	26.294	13.147	55.27	0.000136	***
Residuals	6	1.427	0.238			

Example 1: P=0.49 > 0.05, we do not reject H0

Example 2: P=0.0001 < 0.05, we reject H0

Fitting ANOVA model: the steps of the analysis



Summary

Model

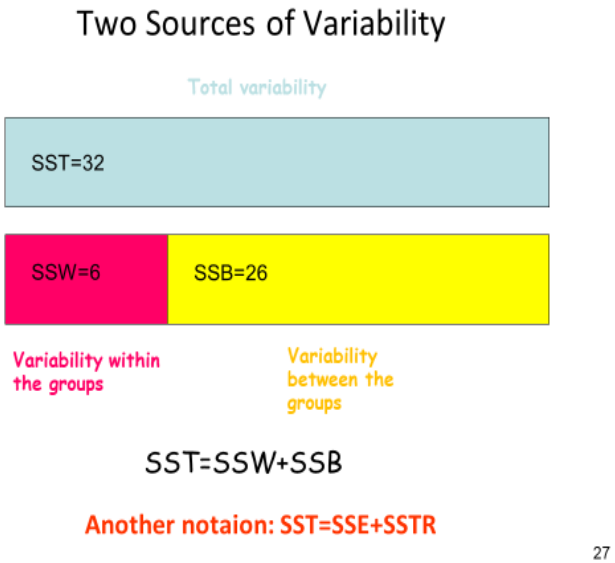
```
> fit.data <- aov(response ~ Treat, data)
```

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

response

Factor mean

Output



- ANOVA table.
- Inference for the null hypothesis:

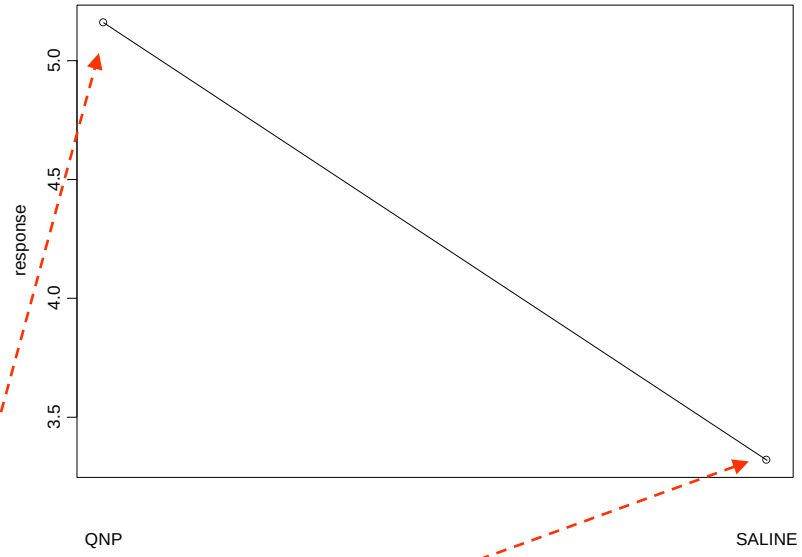
$$H_0 : \mu_1 = \mu_2 = \mu_3$$

Analysis of the pharmacutical experimant

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 agroup compared with S.D=0.57 in teh control group).

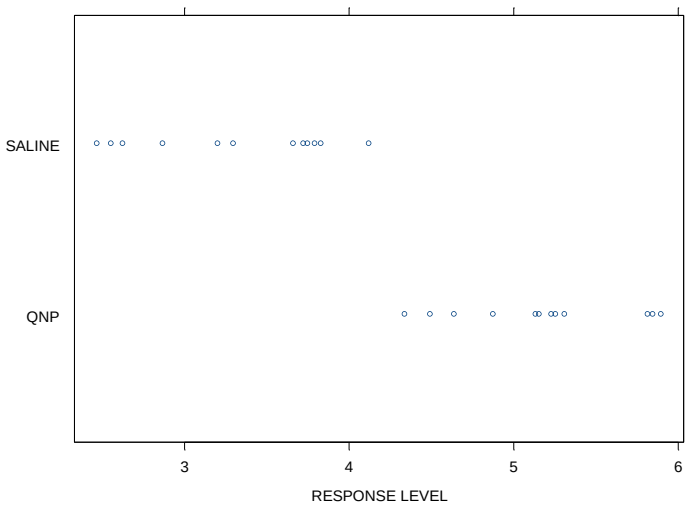


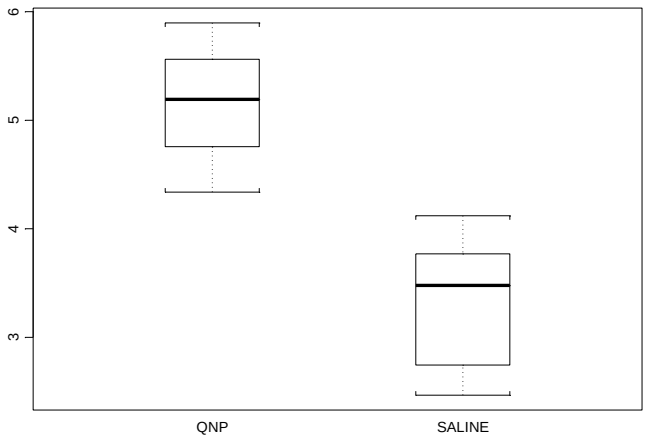
treat	N Obs	N	Mean	Std Dev	Minimum	Maximum
QNP	12	12	5.1647159	0.5179239	4.3377351	5.8956120
SALINE	12	12	3.3231425	0.5719953	2.4675655	4.1198905

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, μ_i is the mean response in treatment group i, i =1,2 and Y_{ij} is the distance travled (on log scale) by the j'th rat in i'th treatment group .

- Sample sizes were equal in both treatment group, $n_1=n_2=12$ and N=24.

- It is further assumed that $\varepsilon_{ij} \sim N\left(0, \sigma^2\right)$.

Inference

We wish to tst the hypothesis that the durg has no effect on the response. Formaly we test the following hypotheses:

$$H_0 : \mu_1 = \mu_2$$
$$H_1 : \mu_1 \neq \mu_2$$

We use the F statistic in ordr 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.

<u>Conclusion</u>: The null hypothesis is rejected (p < 0.05) and we conclude that distanc traveled by rats which recived 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 paramter estimates

- Parameters:

$$\mu_{i.}$$

- Parameter estimates:

$$\hat{\mu}_{i.} = \bar{Y}_{i.}$$

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

$$\begin{aligned} \hat{\mu}_{1.} &= \bar{Y}_{1.} \\ \hat{\mu}_{2.} &= \bar{Y}_{2.} \\ \hat{\mu}_{3.} &= \bar{Y}_{3.} \end{aligned}$$

Model assumptions (1)

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

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.
- This assumption should 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

$$e_{ij} = Y_{ij} - \overline{Y}_{i\cdot}$$

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 qqnormal plot should be a straight line.

- Constant variance

We use two plots to check this assumption: (1) scaterplot 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
```

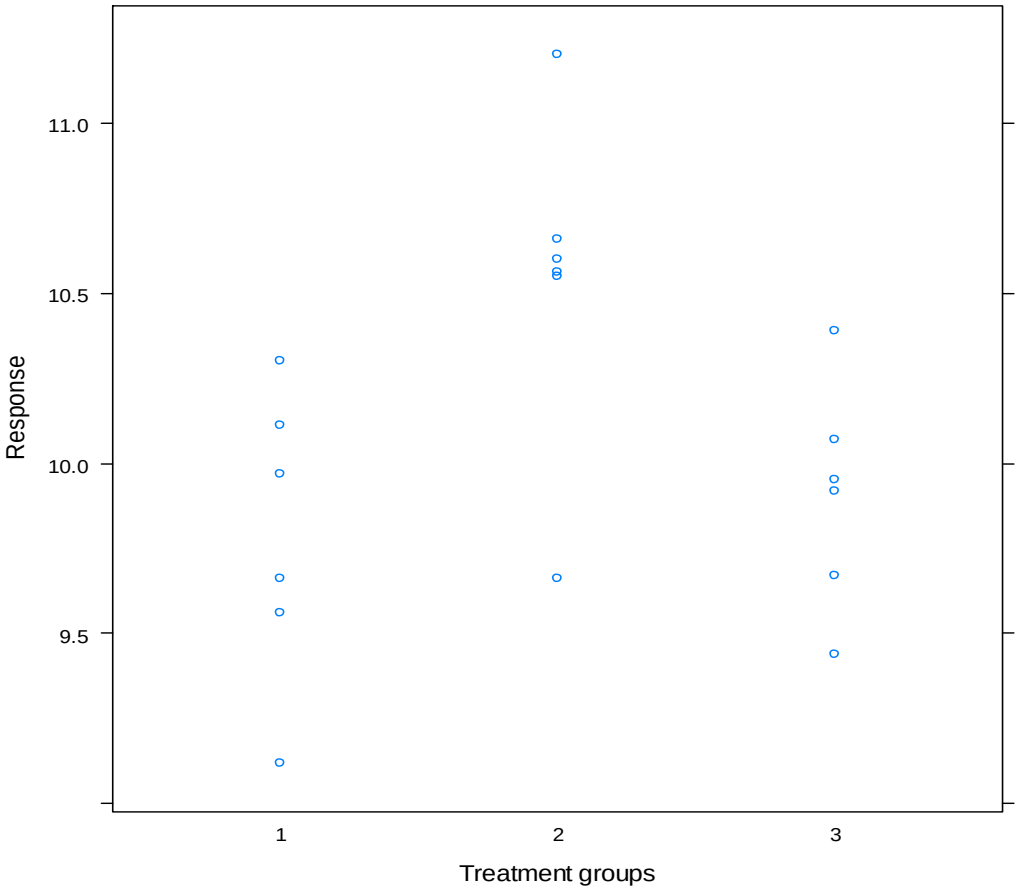
3 groups (I=3).

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

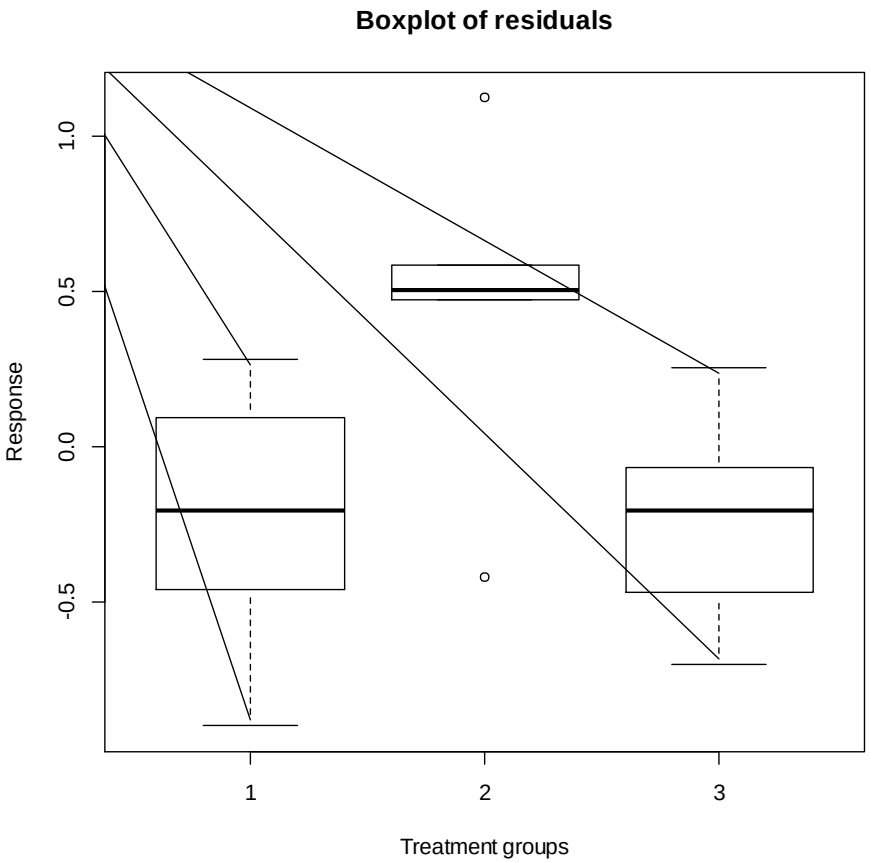
```
> Data1$predict <- fit.Data1$fit
> Data1$resid <- fit.Data1$resid
> print(Data1)
```

	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

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

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.

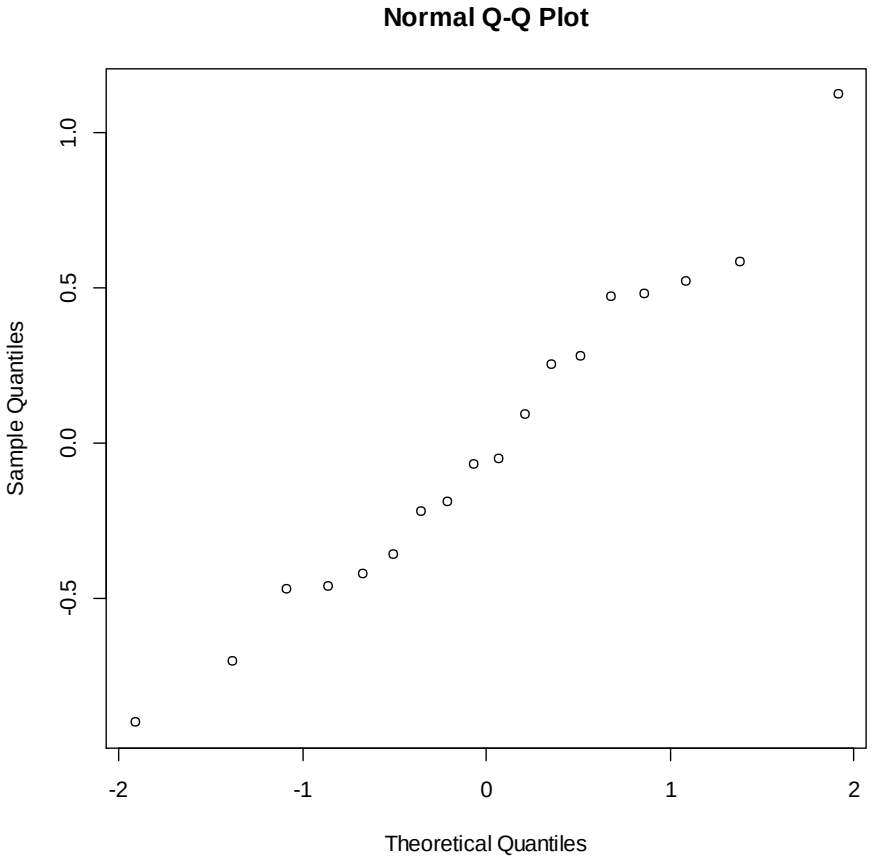
```
> library(car)
> leveneTest(Response~factor(Treat), data = Data1)
```

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

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

We do not reject the null hypothesis.

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 (I=3).

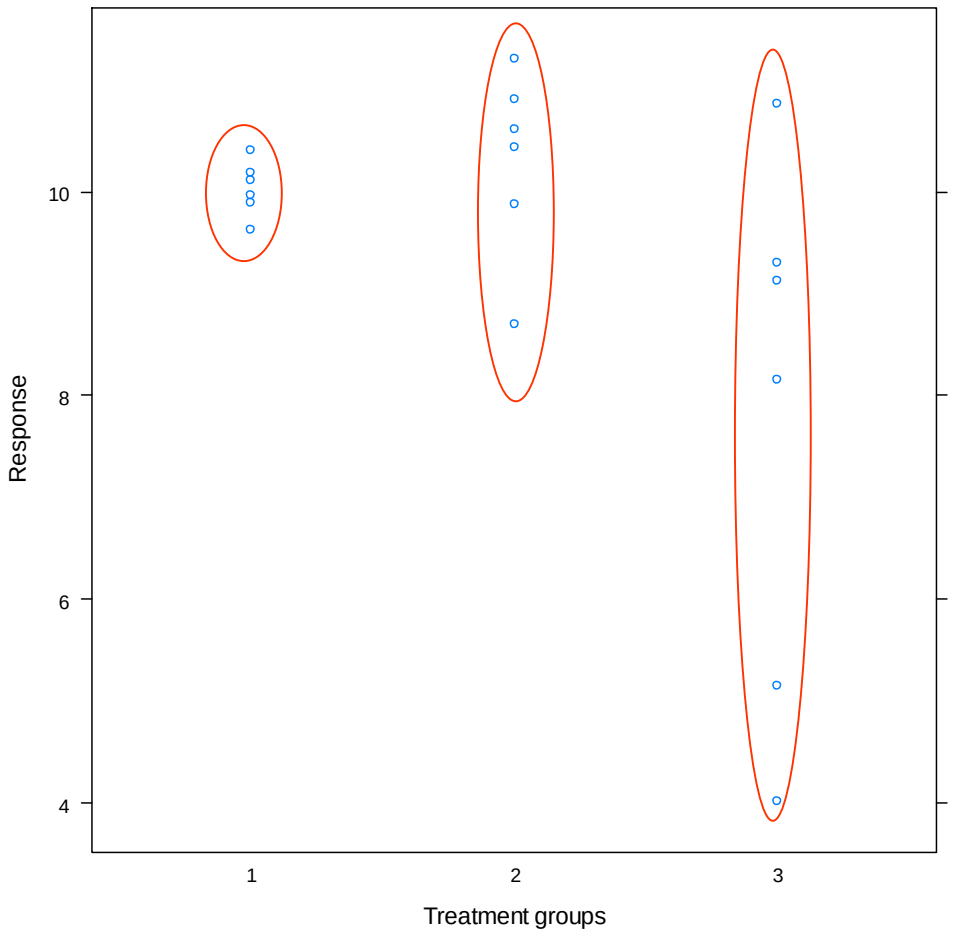
6 observations per group (n_i=6).

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

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

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

ANOVA table

```
          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)
```

One way ANOVA model for testing factor effects.

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

Leveve's test for variance equality across the factor levels.

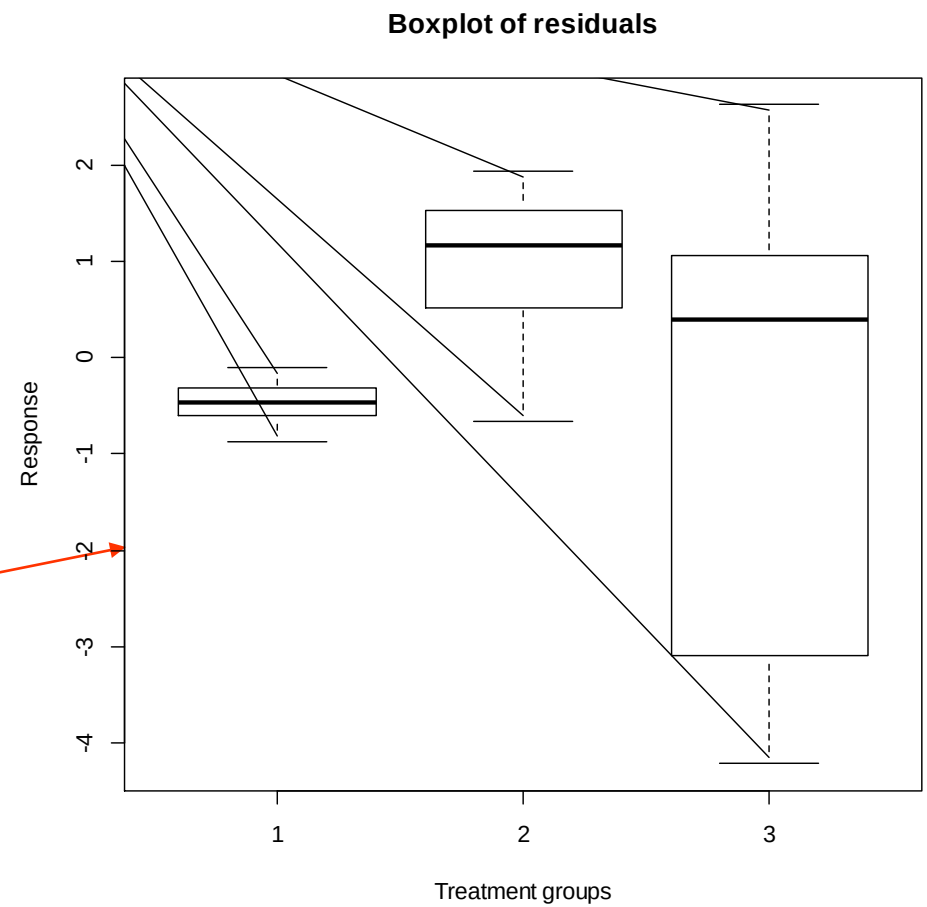
Predicted values and residuals

```
> 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 dispaly: Boxplot for the residuals

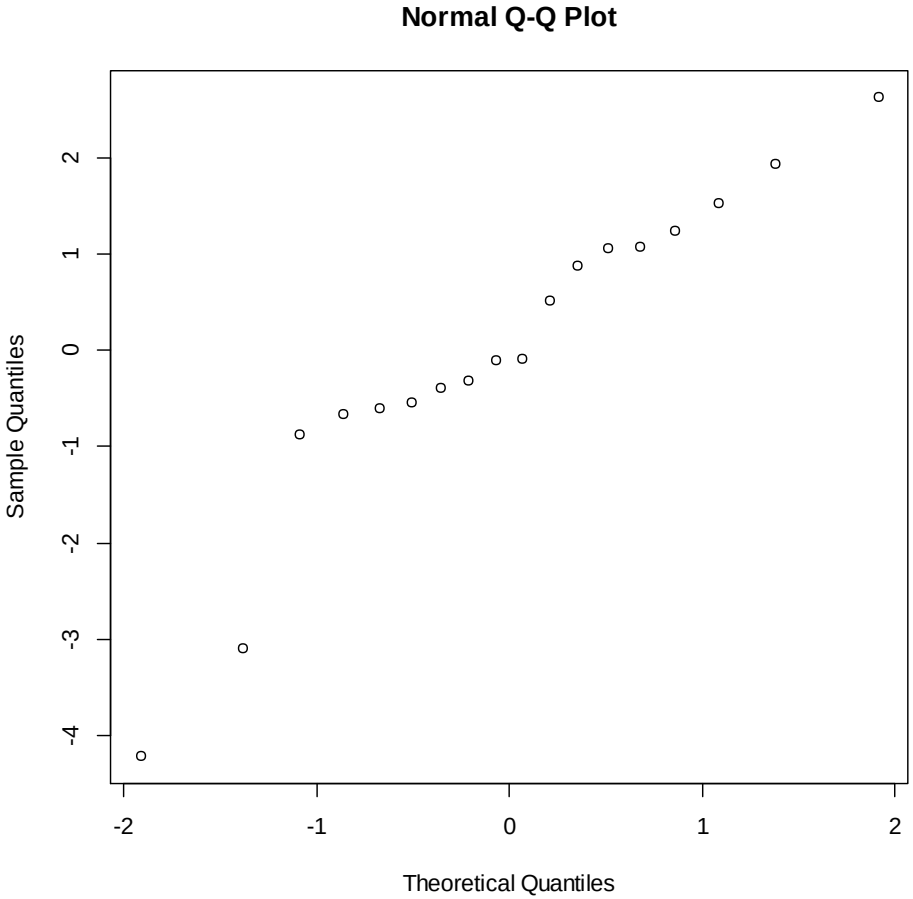
```
> boxplot(split(fit.Data2$resid,Treat),
+     xlab = "Treatment groups",
+     ylab = "Response",
+     main = "Boxplot of residuals")
```



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 indicates on a problem with the normality assumption.

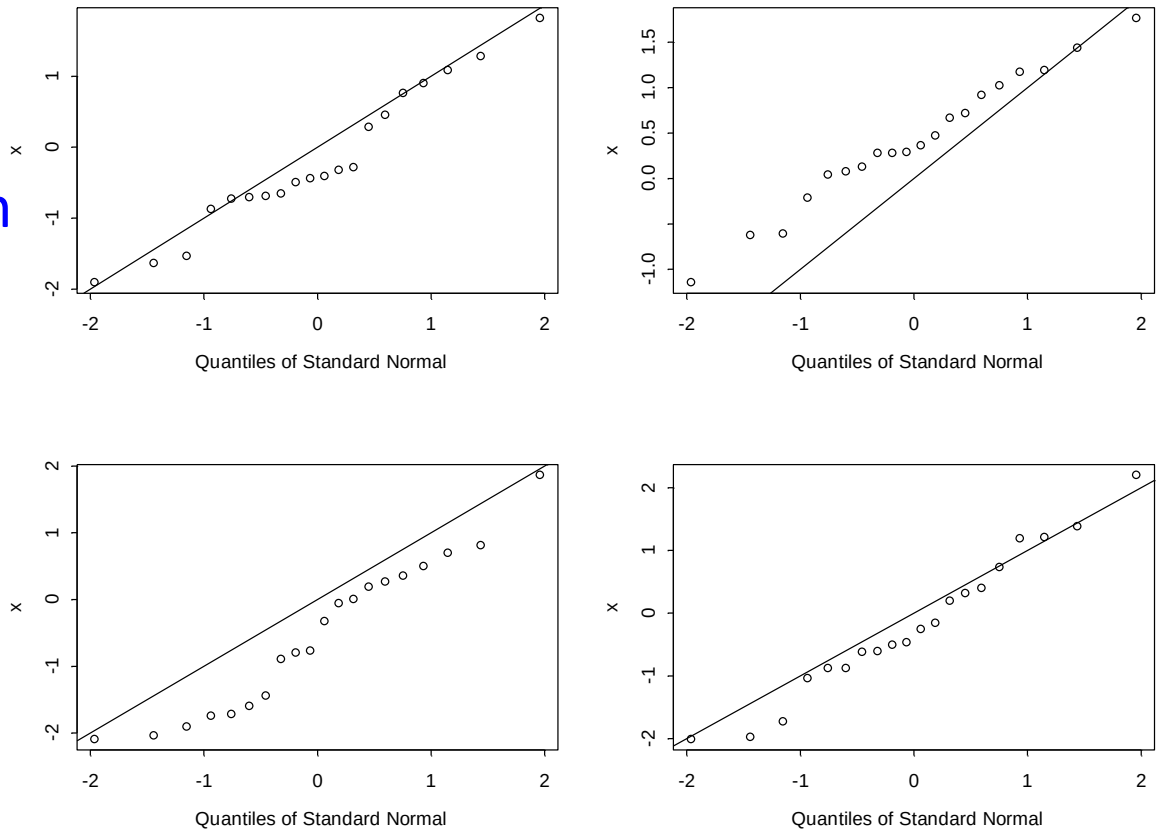


Model diagnostic in R: normal probability plot

4 samples of size 20

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

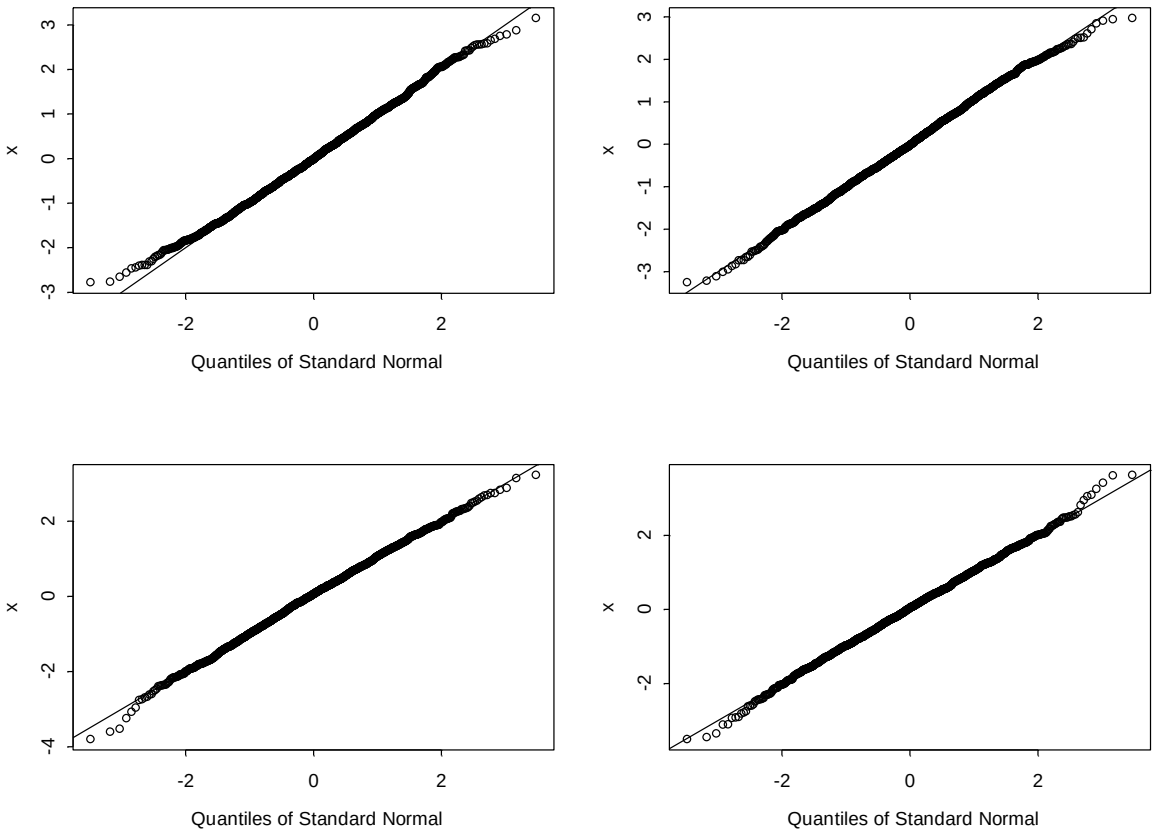
U~N(0,1)



4 samples of size 2000

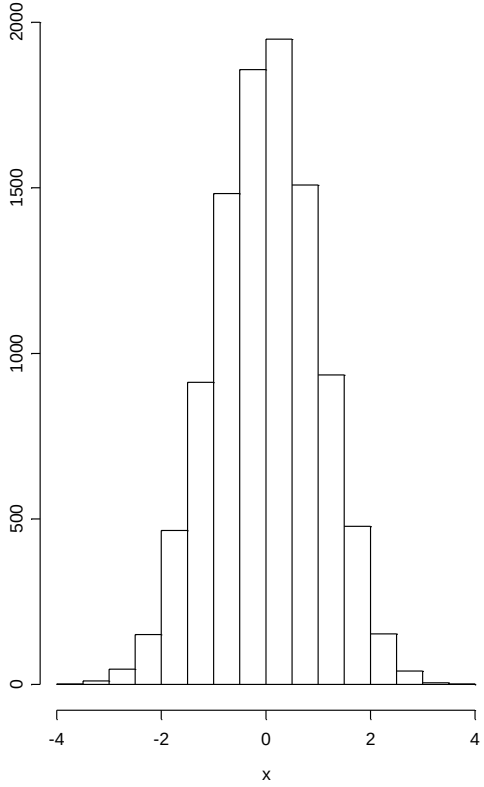
Example of 4 qqnormal plot for 4 random samples from normal distrinution with n=2000.

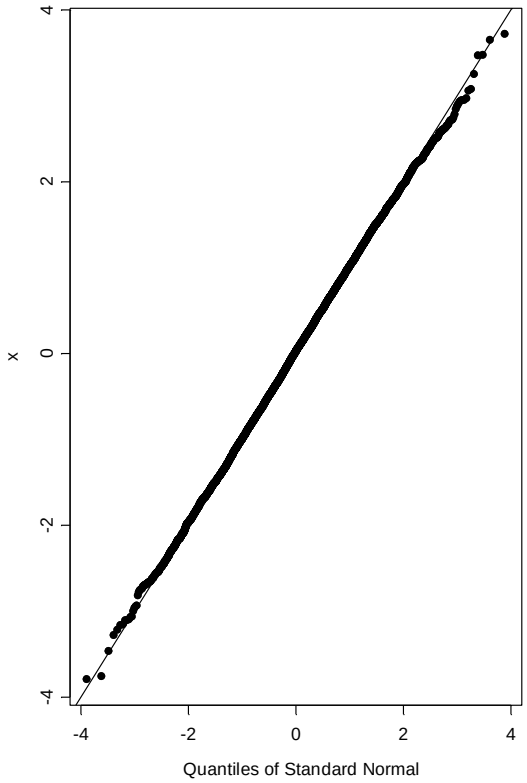
U~N(0,1)



3 Examples of normal probability plots for samples of size 10000

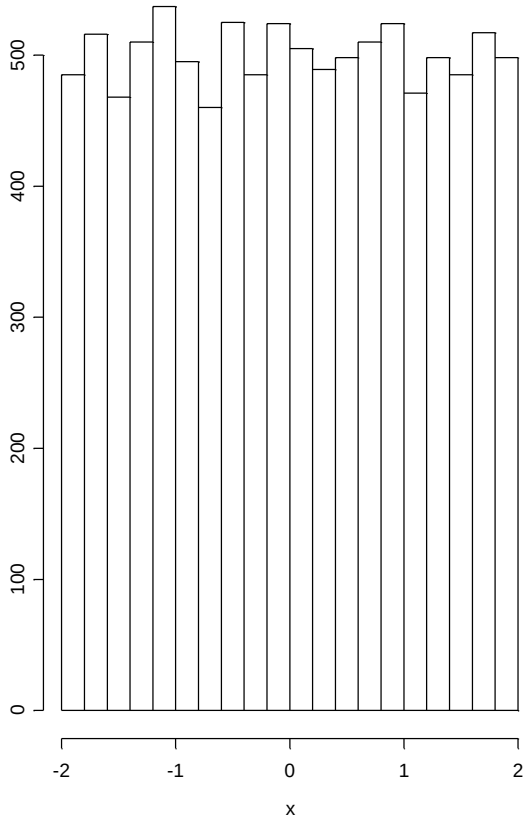
N(0,1)

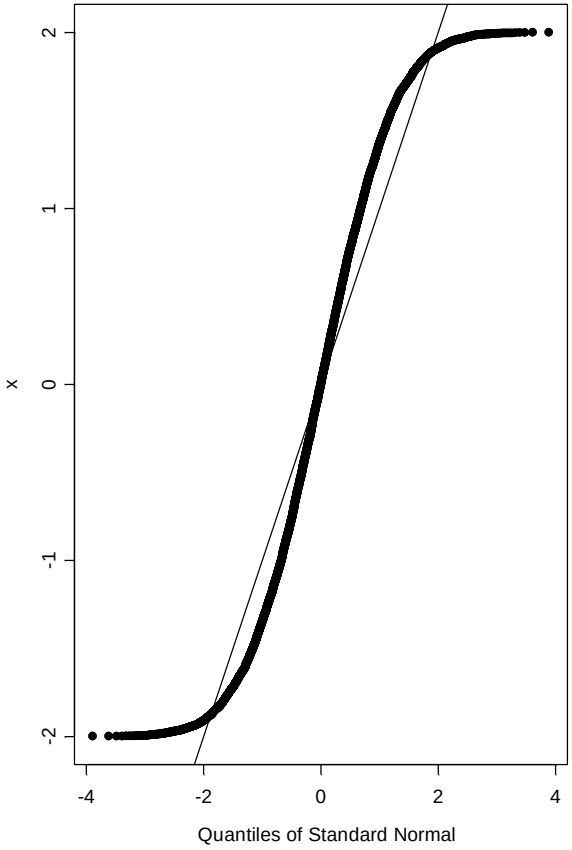

2000
1500
1000
500
0
-4
-2
0
2
4
x


4
2
0
-2
-4
x
-4
-2
0
2
4
Quantiles of Standard Normal

N=10000

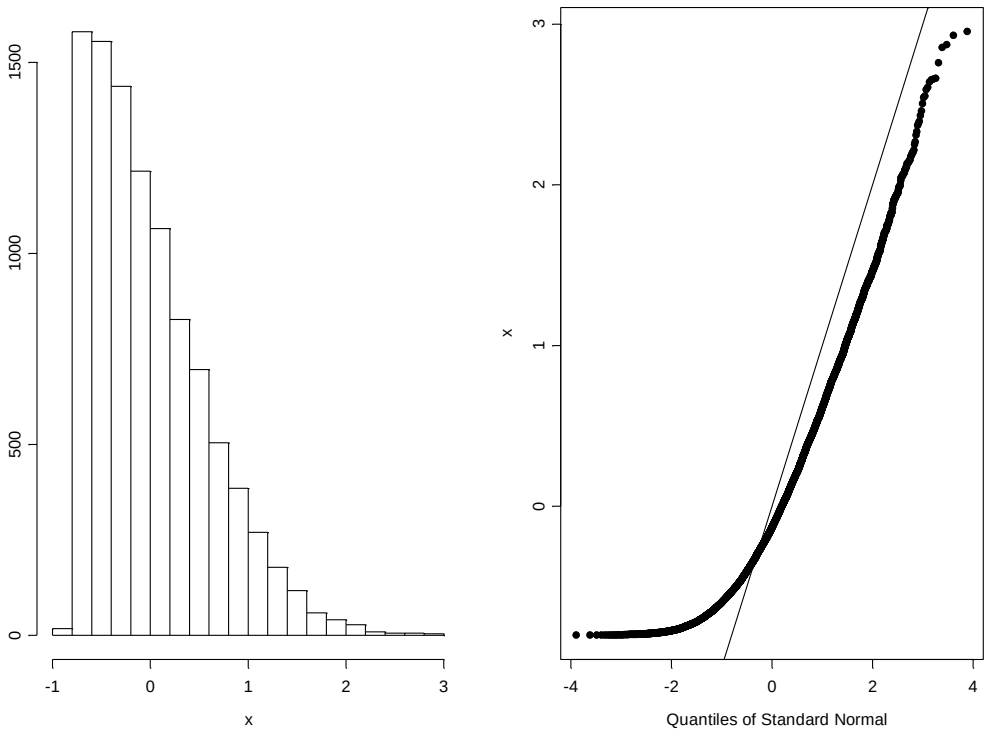
U(-2,2)


0
100
200
300
400
500
-2
-1
0
1
2
x


-2
-1
0
1
2
x
-4
-2
0
2
4
Quantiles of Standard Normal

N=10000

Skewed distribution



N=10000

Summary: R functions

- stripplot(), boxplot().
- aov().
- lm(),glm().
- qqnorm().

Summary: model diagnostic

- We use three graphical displays for model diagnostic:
  - dotplot & boxplot for the response.
  - boxplot for the residuals.
  - 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 (levene's test) or the residuals follow a normal distribution (KS test).

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 simultaneity and we use for each null hypothesis significant level of 0.05 the over 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 hypothesis using significant level of:

The overall significant level

The per-comparison significant level

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

For α=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 overall significant level

$$\frac{\alpha}{3} = \frac{0.05}{3}$$

For α=0.05

The per-comparison significant level

- This means that for each test we use a 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

$$\boxed{H_0 : \mu_1 = \mu_2 = \mu_3}$$

- 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.

$$\boxed{\begin{aligned} H_{0,1}: &\quad \mu_1 = \mu_2 \\ H_{0,2}: &\quad \mu_1 = \mu_3 \\ H_{0,3}: &\quad \mu_2 = \mu_3 \end{aligned}}$$

- 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 a significant level of 0.05/3.

Example 3: the Data

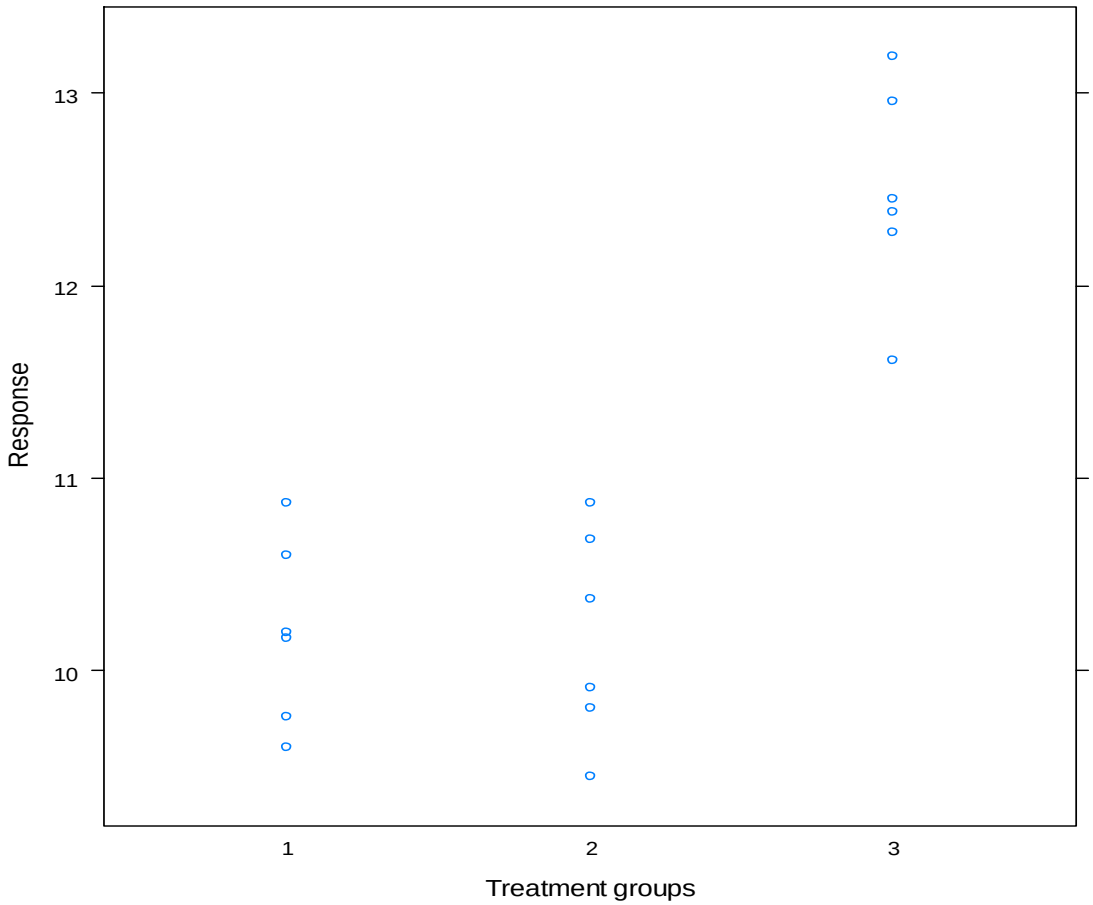
```
> tapply(Data3$Response, list(Data3$Treat),mean)

    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.

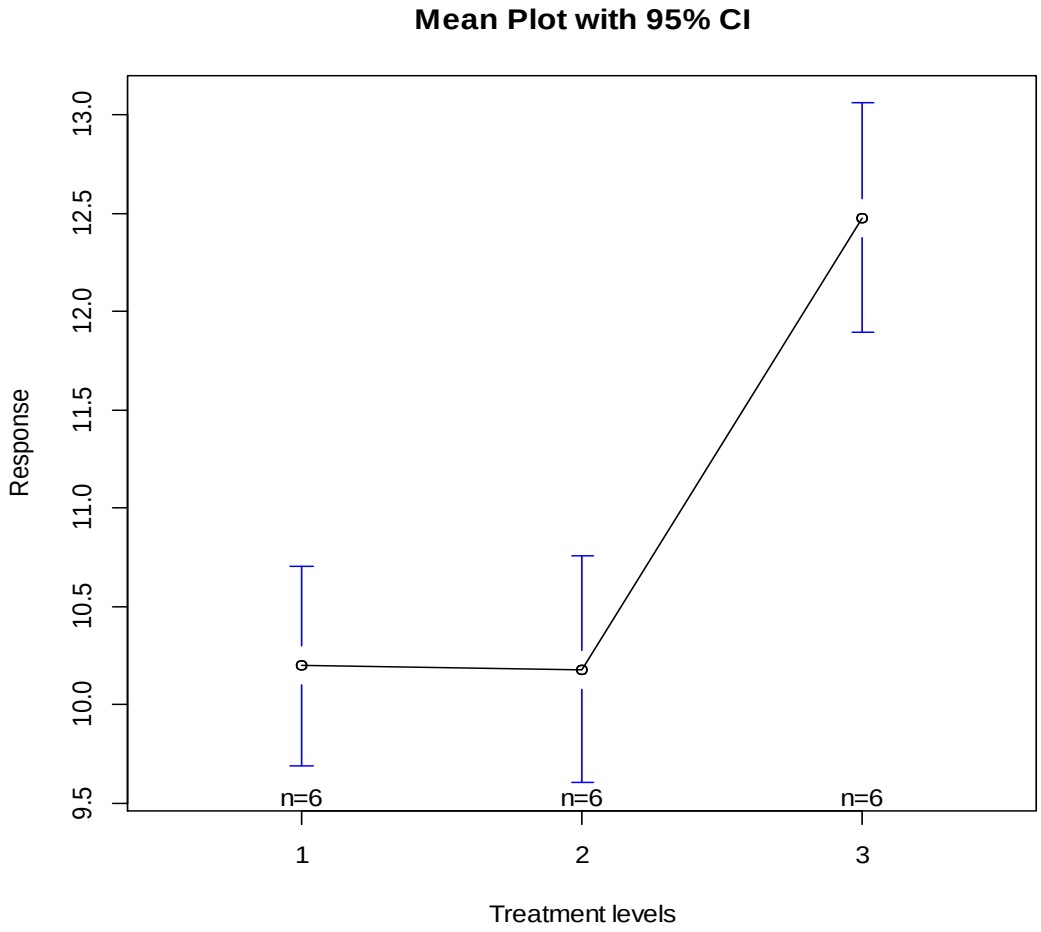
The Data

	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.

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

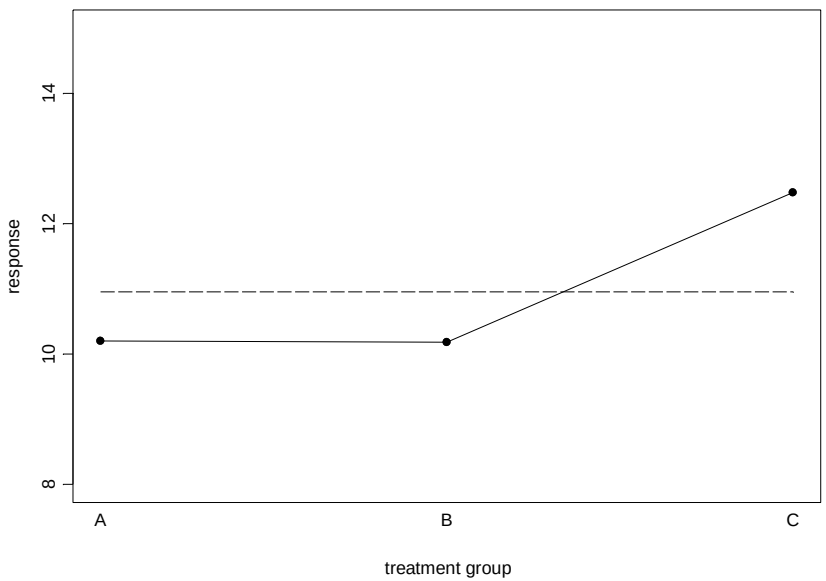
```
            Df Sum Sq Mean Sq F value   Pr(>F)
Treat        1 15.572   15.57   25.97 0.000108 ***
Residuals   16  9.593    0.60
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

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 different from each other.

Muptiple 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.



$$H_{0,1}: \quad \mu_1 = \quad \mu_2$$
$$H_{0,2}: \quad \mu_1 = \quad \mu_3$$
$$H_{0,3}: \quad \mu_2 = \quad \mu_3$$

Using function pairwise.t.test ( ) for multiple comparisoons

```
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.

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 pairwize comparisons, however it does not make correction for the Type I error rate across the pairwise tests.
- P values should be compared with alpha/K.

R output

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

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

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

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

More about Multiplicity

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

https://erbiostat.wixsite.com/inf1/topics