GC – Probabilités et Statistique

http://moodle.epfl.ch/course/view.php?id=18431

Lecture 11

- Revision : 1-way ANOVA (anova à une voie)
- Model evaluation (*video only, NOT EXAMINED*)
- Multiple comparisons
- Factorial experiments
- 2-way ANOVA (anova à deux voies)
- General Linear Model

Révision : ANOVA

- Abréviation de *AN*alysis *O*f *VA*riance (analyse de variance)
- Mais c'est un test de différences des *moyennes*
- L'idée :

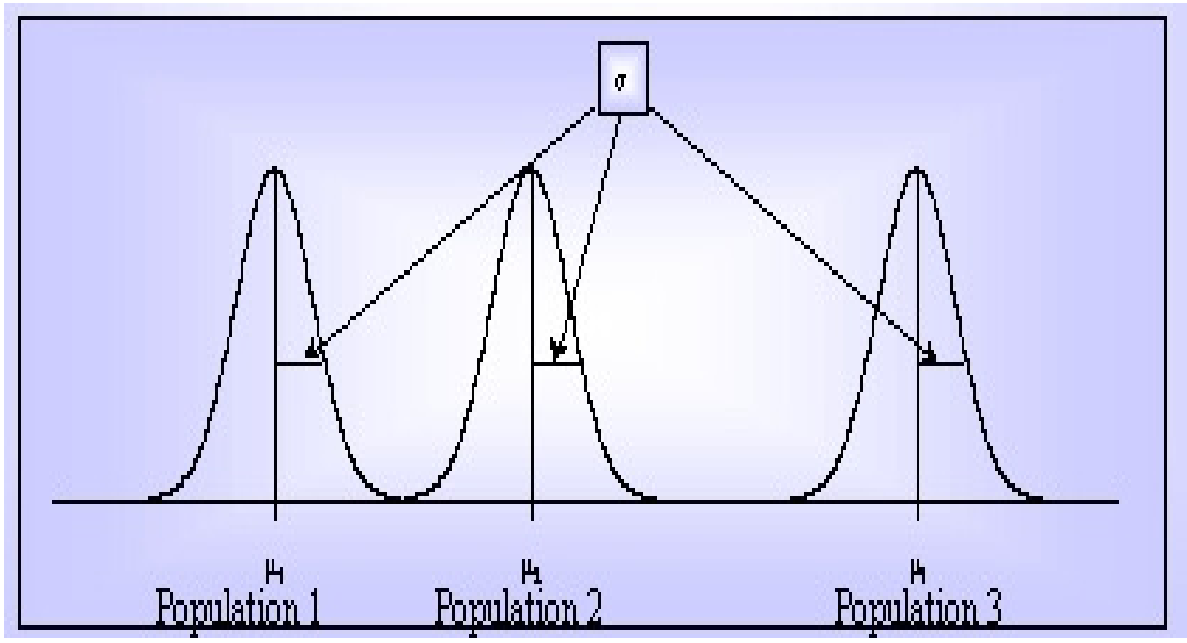


Tableau d'ANOVA

source	df	SC (SS)	CM (MS) ($=SC/\text{df}$)	F	p-valeur
traitements	$k-1$	SCE_{trts}	$SCE_{trts}/(k-1)$	CM_{trts}/CM_{erreur}	$P(F_{obs} > F_{k-1,n-k})$
erreur	$n-k$	SCE_{erreur}	$SCE_{erreur}/(n-k) (= \hat{\sigma}^2)$		
total (corr.)	$n-1$	SCE_{totale}			

*** Assumptions ***

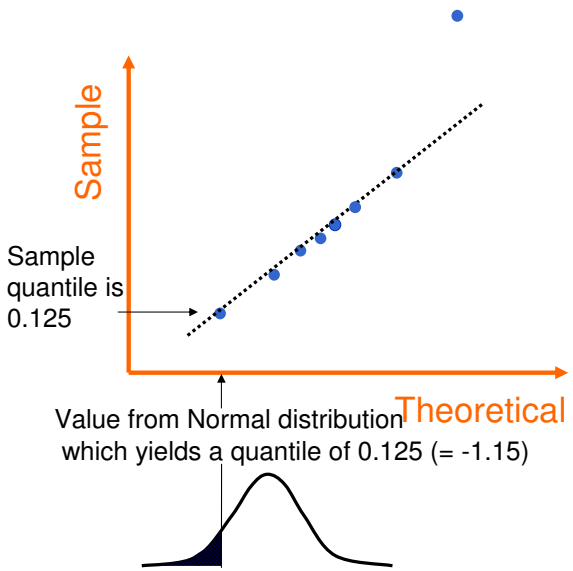
- *Independence* : The k groups (samples) are independent, as well as the individuals within groups ; the ensemble of the n individuals are placed *at random* (*randomization*) between the k modalities for the controled factor A, with n_i individuals receiving treatment i.
- *Homoscedasticity* : The k populations have the same variance ; the factor A acts only on the *mean* of the variable X and does not change its variance
- *Normality* : The variable studied follows a Normal distribution in the k populations compared (or the CLT applied to the means if the n_i are 'sufficiently large')
- Evaluation of model assumptions **WILL NOT BE EXAMINED**

Model evaluation : Normality

- Boxplots of observations (or residuals) should be rather symmetric
- A graph of the sample mean vs. variaces should not display any pattern
- QQ-plot (normal) plot of the observations (or residuals) should form a straight line
- Check whether there are any unusual or influential values

QQ-plot

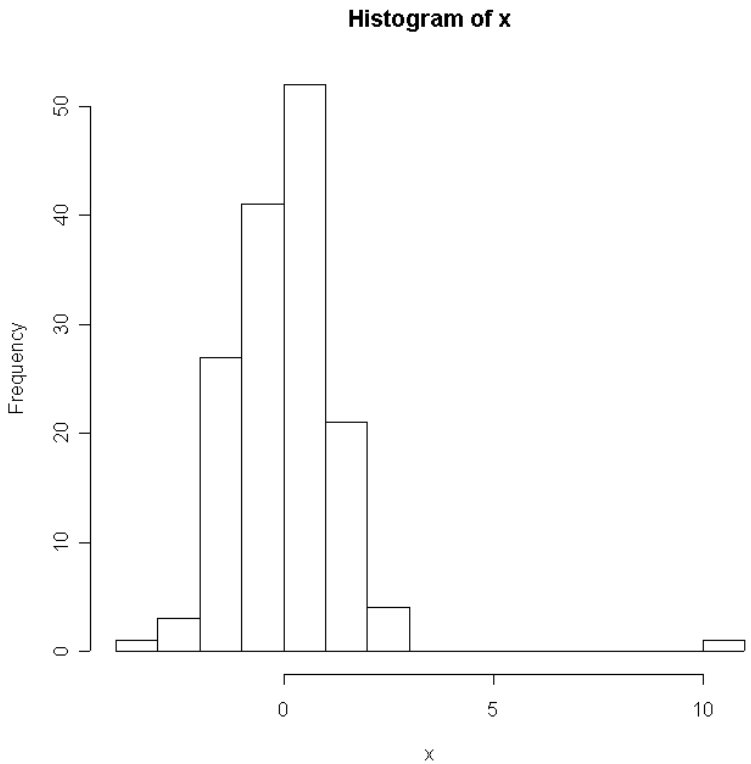
- Quantile-quantile plot (graph)
- Used to determine if a sample follows a particular distribution (for example the Normal distribution) (or to compare two samples)
- A method for identification of outliers when the data are approximatelv Normallv distributed

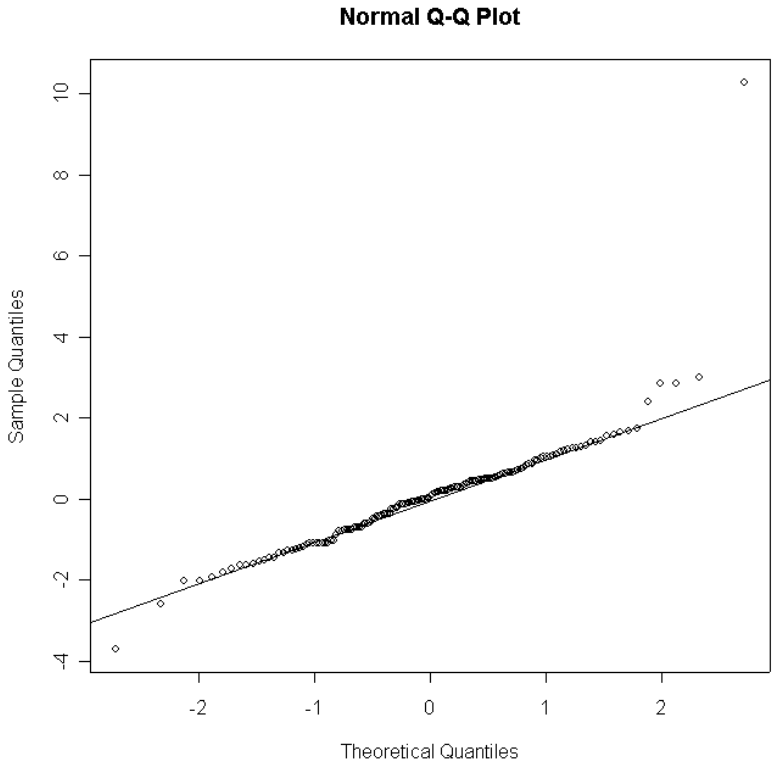


Characteristic deviations from a straight line

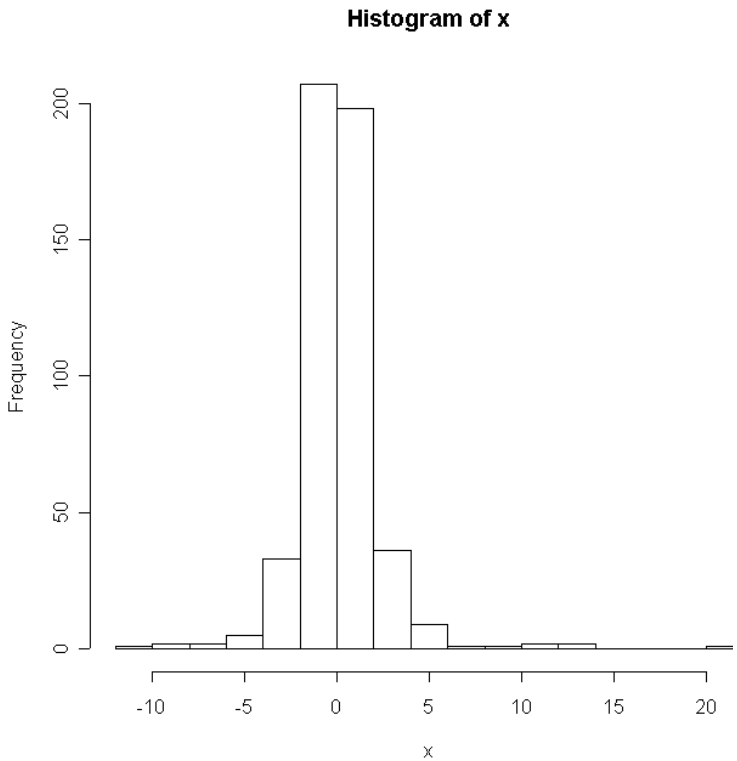
- Outliers
- Curvature at both extremities (long or short tails)
- Convex or concave curvature (asymmetry)
- Horizontal segments, plateaus (discretization)

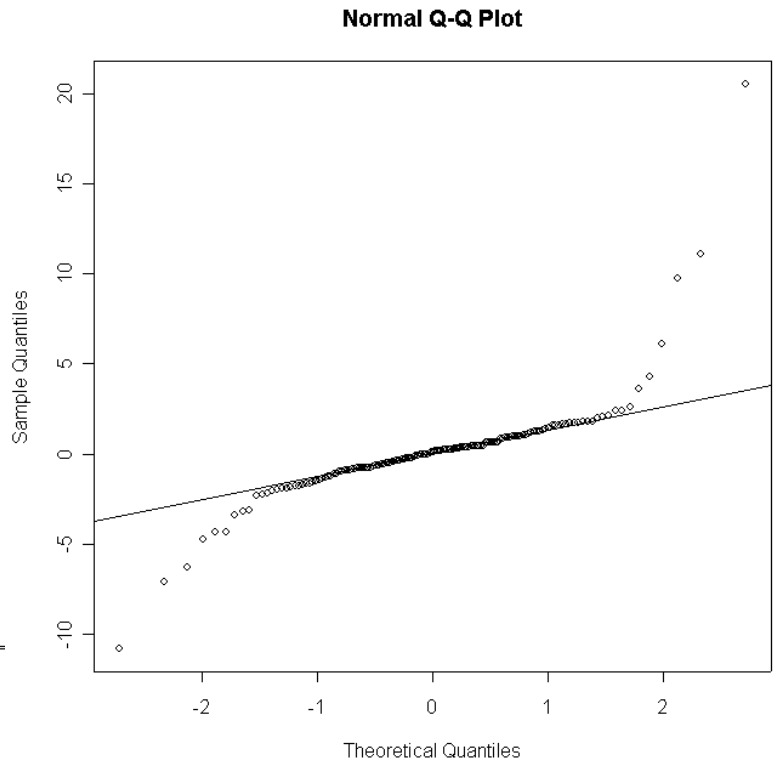
Outliers


Histogram of x
Frequency
0
10
20
30
40
50
0
5
10
x

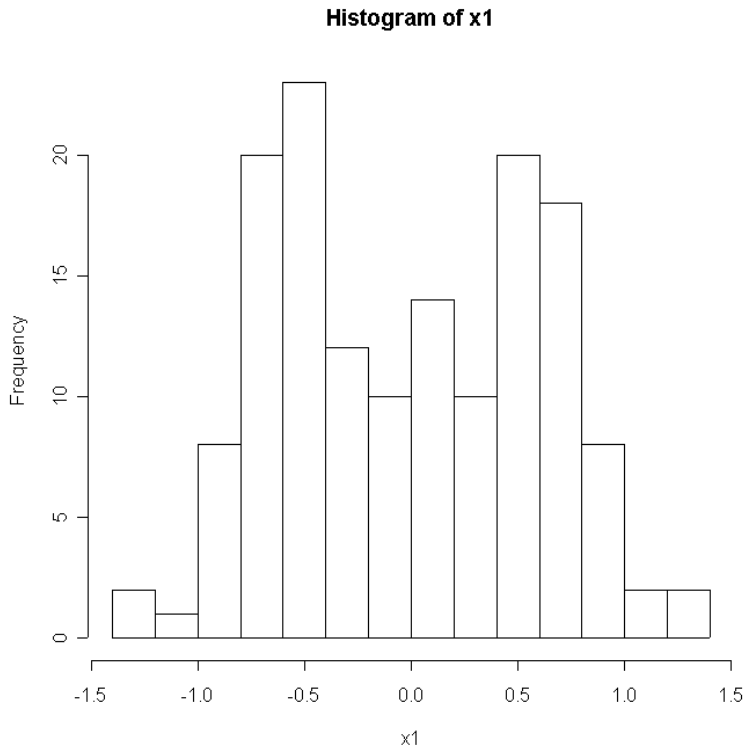

Normal Q-Q Plot
Sample Quantiles
-4
-2
0
2
4
6
8
10
-2
-1
0
1
2
Theoretical Quantiles

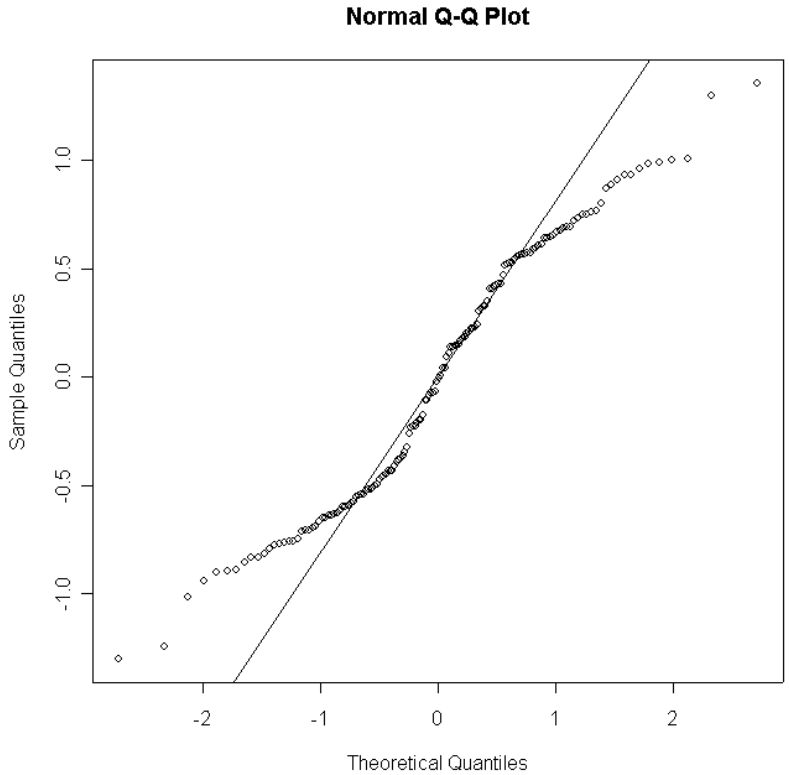
Long tails


Histogram of x
Frequency
0
50
100
150
200
-10
-5
0
5
10
15
20
x

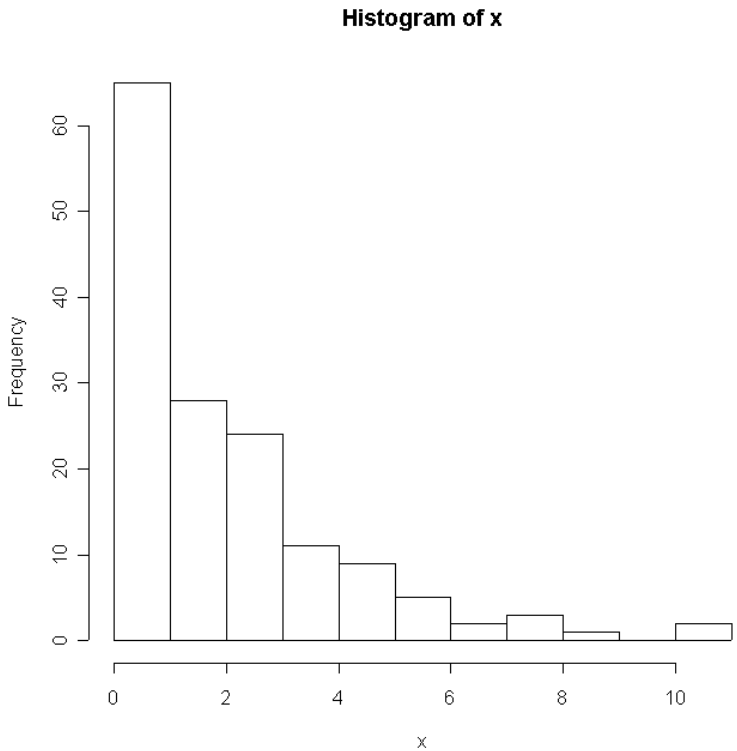

Normal Q-Q Plot
Sample Quantiles
-10
-5
0
5
10
15
20
-2
-1
0
1
2
Theoretical Quantiles

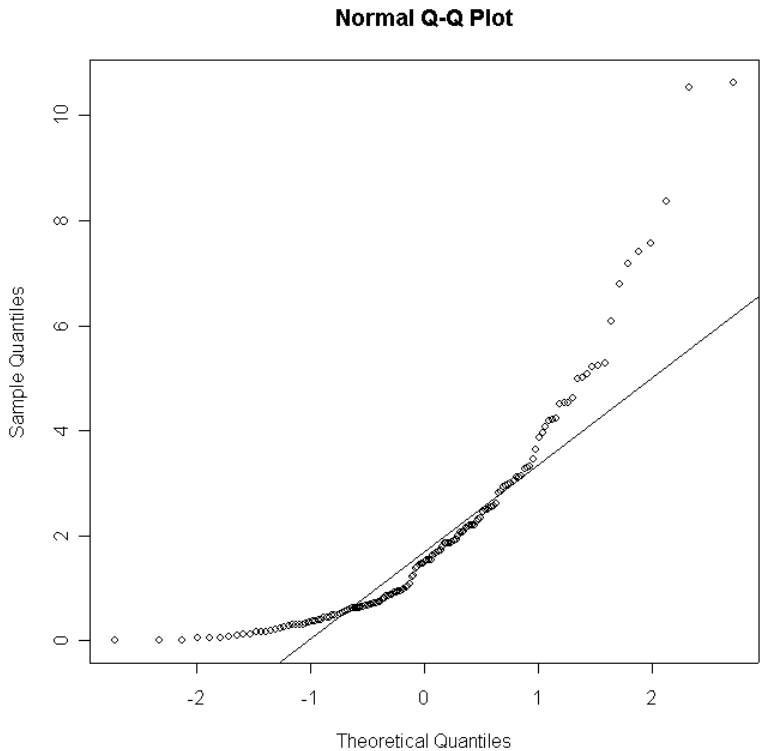
Short tails


Histogram of x1
Frequency
0
5
10
15
20
-1.5
-1.0
-0.5
0.0
0.5
1.0
1.5
x1

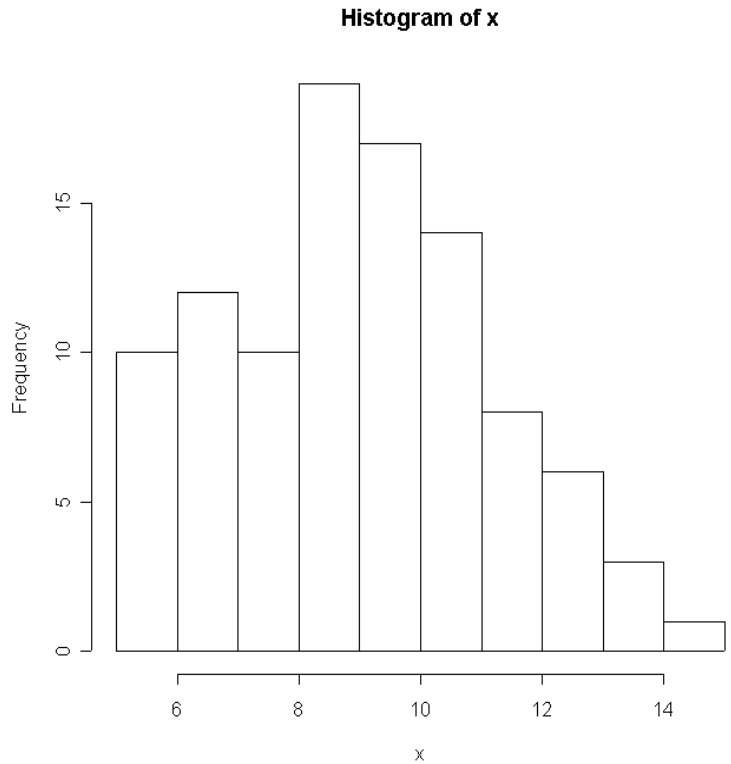

Normal Q-Q Plot
Sample Quantiles
-1.0
-0.5
0.0
0.5
1.0
-2
-1
0
1
2
Theoretical Quantiles

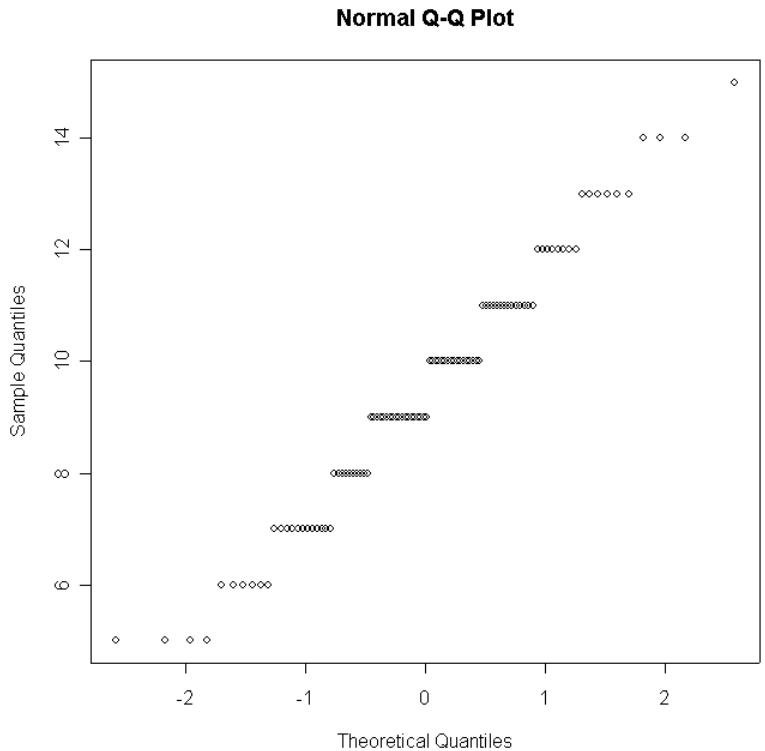
Asymmetry


Histogram of x
Frequency
0
10
20
30
40
50
60
0
2
4
6
8
10
x


Normal Q-Q Plot
Sample Quantiles
0
2
4
6
8
10
-2
-1
0
1
2
Theoretical Quantiles

Discretization

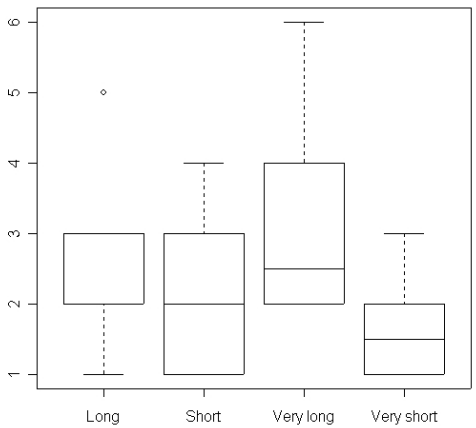


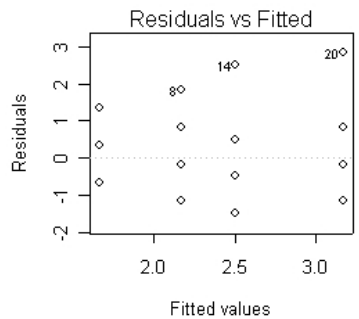


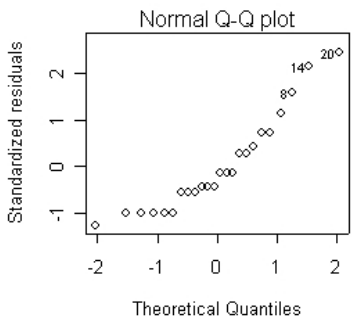
Model evaluation : Homogeneity of variance

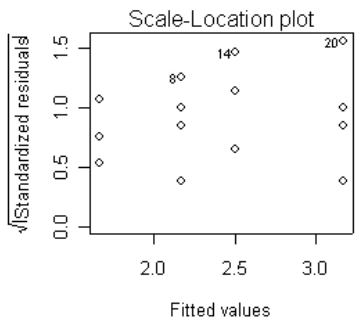
- Boxplots of the observations should show similar variability
- Variability of the residuals should be similar in the graph of residuals versus fitted values
- It is also possible to carry out formal hypothesis tests (e.g. Bartlett, Levene), but these are not useful for diagnosing problems

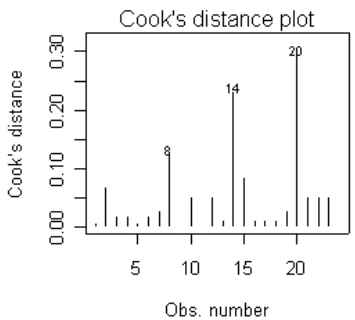
Diagnostic plots











Evaluation of the model : Independence

- Graphics : residuals vs. group mean, might indicated autocorrelation for example
- Normally, treat the question of independence during the conception of the expermient, for example using randomization or perhaps other methods

What does it mean when we reject H ?

- The null hypothesis H is a joint (global) one : that *all* the population means are equal
- When we reject the null hypothesis, that does not mean that all the means are different ! !
- It means that *at least one* is different
- To know which is different, we can carry out '*post hoc*'/*a posteriori* tests (pairs of tests, for example)

ANOVA : after the test

- Once all the conditions for an ANOVA have been verified and the analysis carried out, two conclusions are possible :
  - we reject H
  - we do not have enough evidence to reject H
- If H is not rejected, we conclude that there are not significant differences between group means
- If we DO reject H, typically we are interested in *identifying the modalities/factor levels* that are responsible for the significant result

Problem : control of the Type I error rate α

- For a number k of comparisons, the probability of not rejecting a true H is $(1-\alpha)^k$
- P. ex. : for 4 treatments, there are 6 pairs of means, thus $(1-\alpha)^k = (0.95)^6 = 0.735$
- So α for the *ensemble* (or *family*) of tests is $\alpha_e = 0.265$
- Thus, we are expecting to reject a true H for at least 1 pair $\approx$ 27% of the time, even if the treatments *are identical*
- $\Longrightarrow$ the error α must be controled *while ajusting for the number of comparisons*

Bonferroni method

- To maintain the global level α_e at level α, we must *ajust α for each comparison by the total number of comparisons*
- In this way, α_e is independent of the number of comparisons
- Simples method : method od Bonferroni

$$\alpha' = \alpha/k,$$

  where $k =$ number os comparisons (tests)

- $p_{ajusté} = \min(kp, 1)$
- Bonferroni's method assures that the global level is *smaller than the desired level*
- (That property makes this method *conservative*, and thus less powerful than other methods, but it is applicable for any situation)

Multiple comparisons

- Comparing means of pairs of treatments
- Carried out *after* a significant ANOVA
- Types of comparisons
  - planned (*a priori*) : indpendent of the ANOVA results ; the theory predicts which treatments should be different
  - unplanned (*a posteriori*) : the comparisons are decided *based on the ANOVA results*
- $H : \mu_i = \mu_j$ vs. $A : \mu_i \neq \mu_j$
- Test statistic

$$t = \frac{\overline{y}_i - \overline{y}_j}{\sqrt{\hat{\sigma}^2 \left(1/n_i + 1/n_j\right)}}$$

- $(\hat{\sigma}^2 = MS_{error})$; $\mathrm{df} = df_{error}$

Example 11.1

- Study on different types of brain dominance (left brain, right brain, integration (= both))
- Three groups (eight subjects each) : left brain (active, verbal, logical ; group1), right brain (receptive, spatial, intuitive ; group 2), or integration (group 3)

ANOVA Table

source	df	S	MX	F	p-valeur
group	2	1362.33	681.17	44.614	2.749×10^{-8}
error	21	320.63	15.27		
total (corr.)	23	1682.96			

- We REJECT H, but which group(s) is/are different *??*
- $\overline{y}_1 = 33.625$, $\overline{y}_2 = 15.375$, $\overline{y}_3 = 26.875$

Example 11.1, cont.

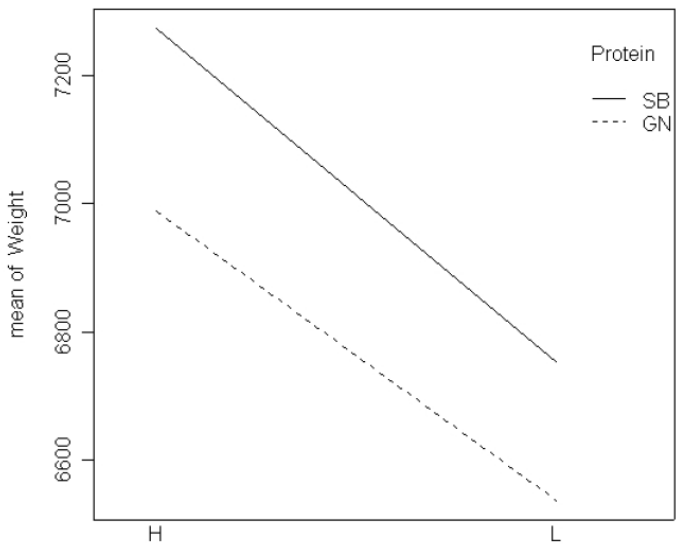
Factorial experimental design and interaction

- Example 11.2 : The lifetime (hours) of a battery could depend on the type of material and the temperature of the apparatus used
- $n = 4$ batteries are tested for **each** combination of type and temperature
- The study addresses the questions :
  - What are the effects of type and temperature on lifetime
  - Is there a type that uniformly affects lifetime, whatever the temperature ?
- We compare *2 sets* of conditions in the *same experiment*

Interaction plot

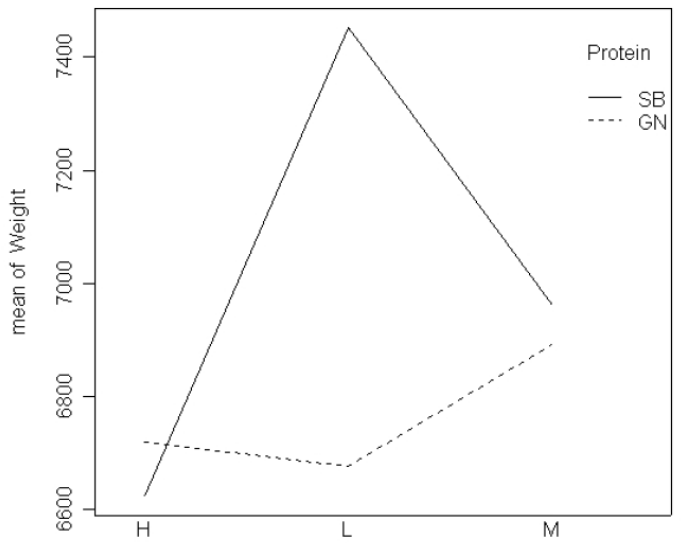
- Interaction = 'difference of differences'
- There is an *interaction* when the effect of the association of combined treatments is not the sum of treatment effects
- In the case of interaction, the effect of a treatment *varies according to whether it is associated with the other treatment*
- The interpretation of individual effects is therefore more difficult when interation is present

pas d'interaction



LS

interaction



LP

Advantages of factorial experiments

- *More efficient* (powerful) than a series of experiments studying one factor at a time
- Permits estimation of *interaction* between sets of conditions that may affect the response
- All data are used for effect estimation

BREAK

2-way ANOVA – Introduction

- Simulataneous study of a factor A with I levels and a factor B with J levels
- For each pair of levels (A, B) :
  - we have a sample
  - all samples are of the *same size* n (balanced design)
- Suppositions :
  - the populations studies are Normally distributed
  - the population variances are all equal (homoscedasticity)
  - the samples are taken randomly and independently in the populations

Example 7.1, cont. : data

Material Type	Temperature (^{o}F) 15		70		125	
1	130	155	34	40	20	70
	74	180	80	75	82	58
2	150	188	136	122	25	70
	159	126	106	115	58	45
3	138	110	174	120	96	104
	168	160	150	139	82	60

Complete model

- The *complete model* : with interaction
- $y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ijk}$
- $E[\epsilon_{ijk}] = 0$, $Var(\epsilon_{ijk}) = \sigma^2$, $Cov(\epsilon_{ijk}, \epsilon_{i'j'k'}) = 0$ if $(ijk) \neq (i'j'k')$

ANOVA table

source	df	SS	MS	F
A	$I-1$	$nJ\sum_{i=1}^{I}(\overline{y}_{i..}-\overline{y}_{...})^2$	SS_A/df_A	MS_A/MS_{err}
B	$J-1$	$nI\sum_{j=1}^{J}(\overline{y}_{.j.}-\overline{y}_{...})^2$	SS_B/df_B	MS_B/MS_{err}
AB	$(I-1)(J-1)$	$n\sum_{j=1}^{J}\sum_{i=1}^{I}(y_{ij.}-\overline{y}_{i..}-\overline{y}_{.j.}+\overline{y}_{...})^2$	SS_{AB}/df_{AB}	MS_{AB}/MS_{err}
error	$IJ(n-1)$	$\sum_{k=1}^{n}\sum_{j=1}^{J}\sum_{i=1}^{I}(y_{ijk}-\overline{y}_{ij.})^2$	SS_{err}/df_{err}	
total (corr.)	$nIJ-1$	$\sum_{k=1}^{n}\sum_{j=1}^{J}\sum_{i=1}^{I}(y_{ijk}-\overline{y}_{...})^2$		

*** : n** = number **PER cell** (not total sample size)

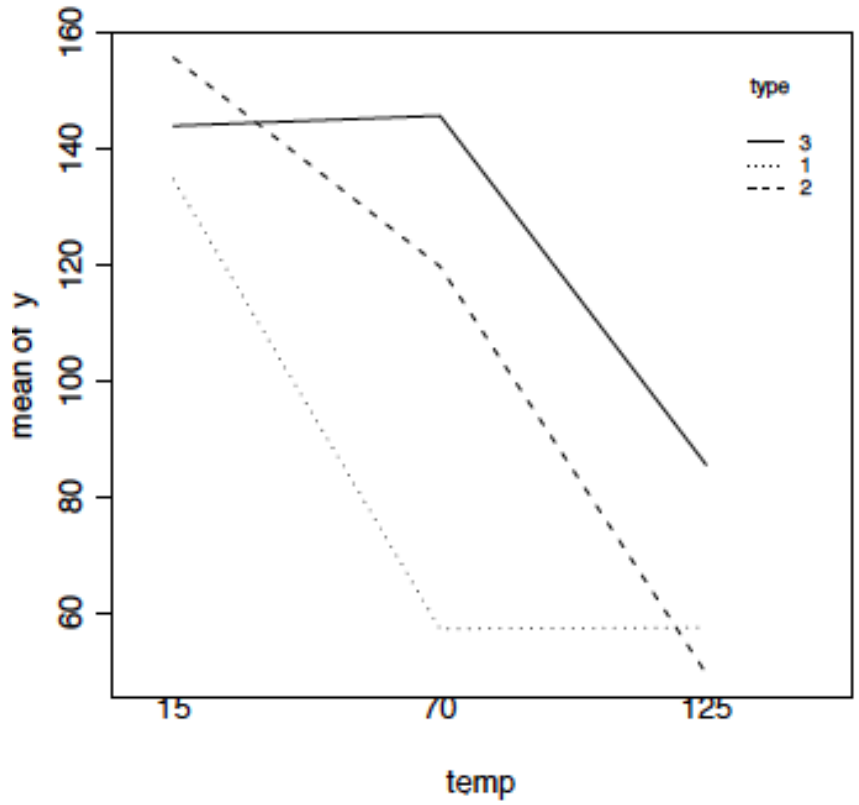
Example 7.1, cont. : output

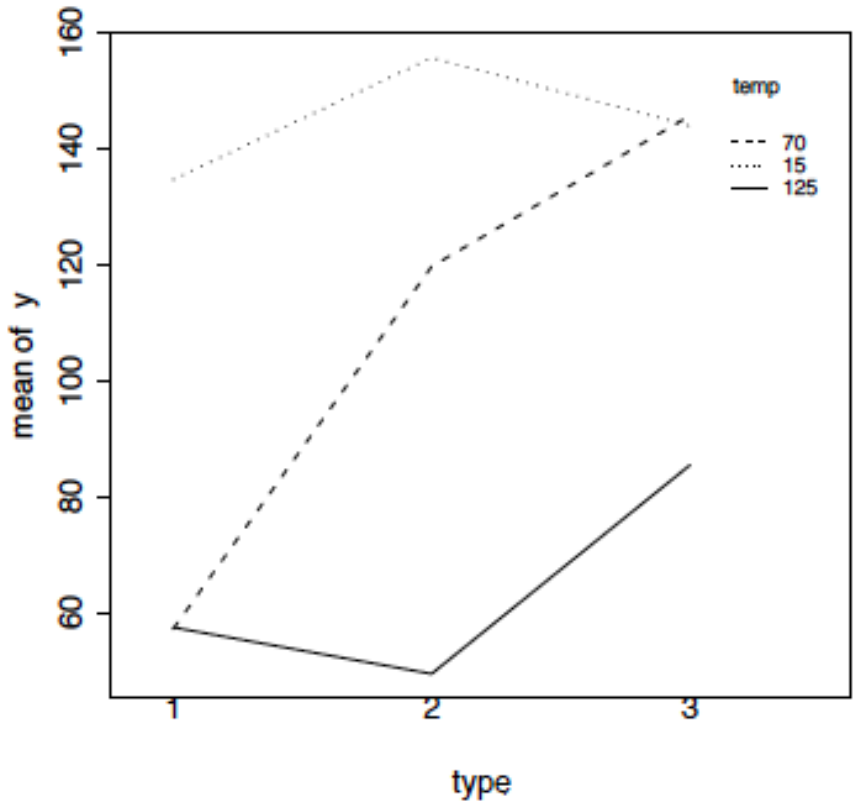
```
Analysis of Variance Table

Response: y
           Df Sum Sq Mean Sq F value    Pr(>F)
type        2  10684    5342  7.9114  0.001976 **
temp        2  39119   19559 28.9677 1.909e-07 ***
type:temp   4   9614    2403  3.5595  0.018611 *
Residuals  27  18231     675
```

- What are your conclusions *??*

Example 11.2, cont. : interaction plot, 2 views

- These 2 graphs show *THE SAME INTERACTIONS*





Hypothesis tests

- Test for interaction
  $H : \gamma_{ij} = 0,\ i = 1, \ldots, I\ ;\ j = 1, \ldots, J$
- Test statistic :
  $F_{AB} = CM_{AB}/CM_{erreur} \sim F_{\underline{(I-1)(J-1),IJ(n-1)}}$ under H
- Test for effect of factor A
  $H : \alpha_i = 0,\ i = 1, \ldots, I$
- Test statistic :
  $F_A = CM_A/CM_{erreur} \sim F_{\underline{I-1,IJ(n-1)}}$ under H
- Test for effect of factor B
  $H : \beta_j = 0,\ j = 1, \ldots, J$
- Test statistic :
  $F_B = CM_B/CM_{erreur} \sim F_{\underline{J-1,IJ(n-1)}}$ under H

Additive model

- The *additive model* : without interactions
- $y_{ijk} = \mu + \alpha_i + \beta_j + \epsilon_{ijk}$
- $E[\epsilon_{ijk}] = 0,\ Var(\epsilon_{ijk}) = \sigma^2,\ Cov(\epsilon_{ijk}, \epsilon_{i'j'k'}) = 0$ id $(ijk) \neq (i'j'k')$

ANOVA Table

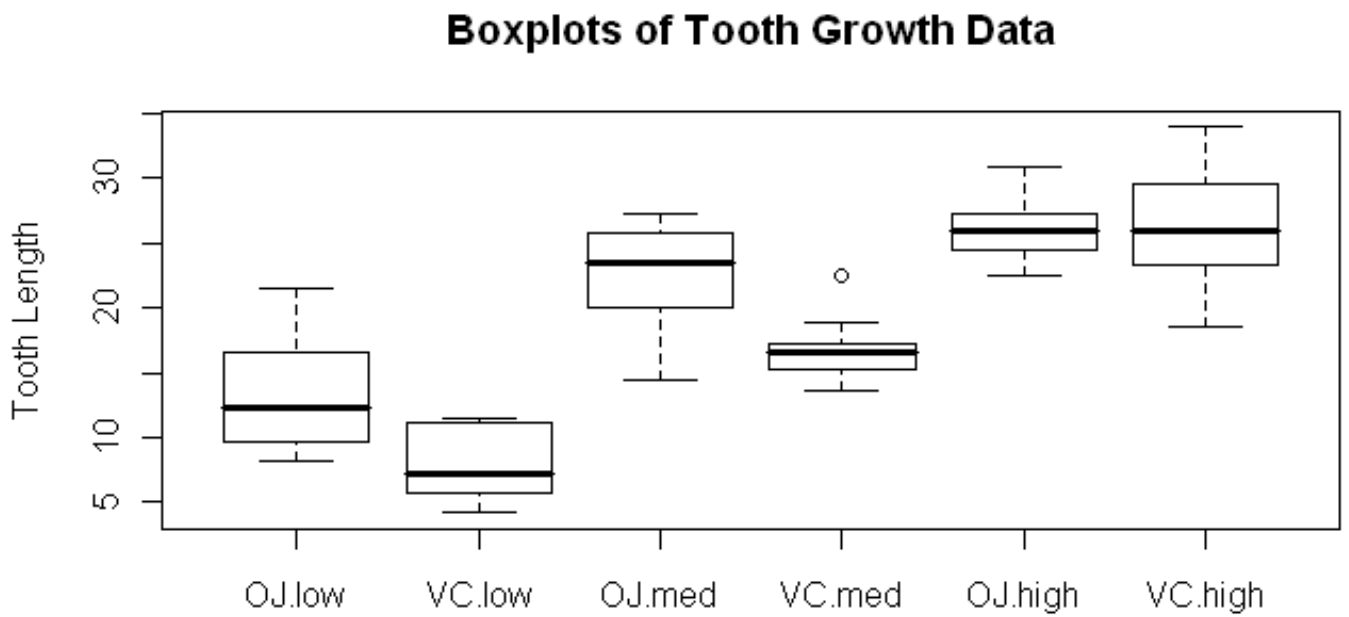
source	df	SS	MS	F
A	$I-1$	$nJ\sum_{i=1}^{I}(\overline{y}_{i\cdot\cdot} - \overline{y}_{\cdots})^2$	SS_A/df_A	MS_A/MS_{err}
B	$J-1$	$nI\sum_{j=1}^{J}(\overline{y}_{\cdot j\cdot} - \overline{y}_{\cdots})^2$	SS_B/df_B	MS_B/MS_{err}
error	$nIJ-I-J+1$	$\sum_{k=1}^{n}\sum_{j=1}^{J}\sum_{i=1}^{I}(y_{ijk} - \overline{y}_{i\cdot\cdot} - \overline{y}_{\cdot j\cdot} + \overline{y}_{\cdots})^2$	SS_{err}/df_{err}	
total (corr.)	$nIJ-1$	$\sum_{k=1}^{n}\sum_{j=1}^{J}\sum_{i=1}^{I}(y_{ijk} - \overline{y}_{\cdots})^2$		

* : **n** = number **PER cell** (not total sample size)

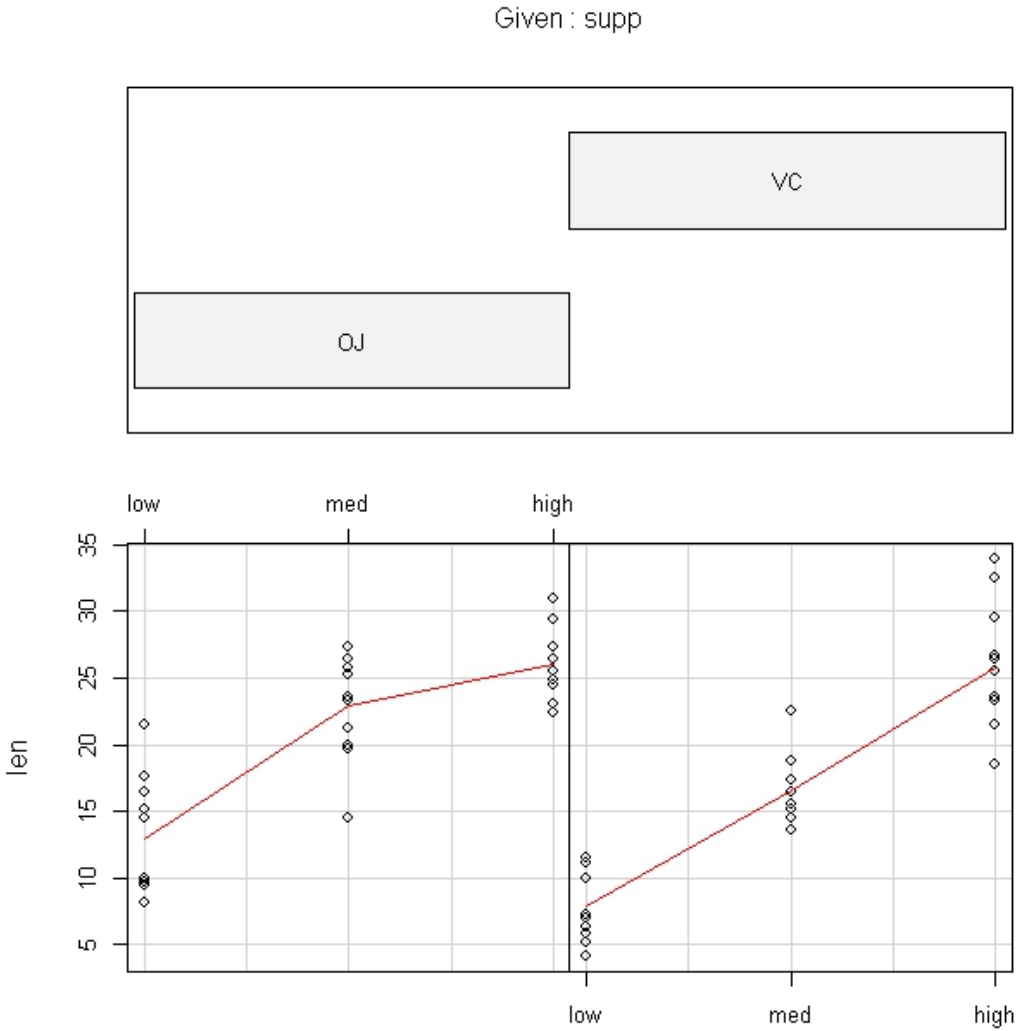
- What are the hypotheses and test statistics *? ?*
  -
  -

Example 11.3 : ToothGrowth

- "The response is the length of odontoblasts (teeth) in each of 10 guinea pigs at each of three dose levels of Vitamin C (0.5, 1, and 2 mg) with each of two delivery methods (orange juice or ascorbic acid)."

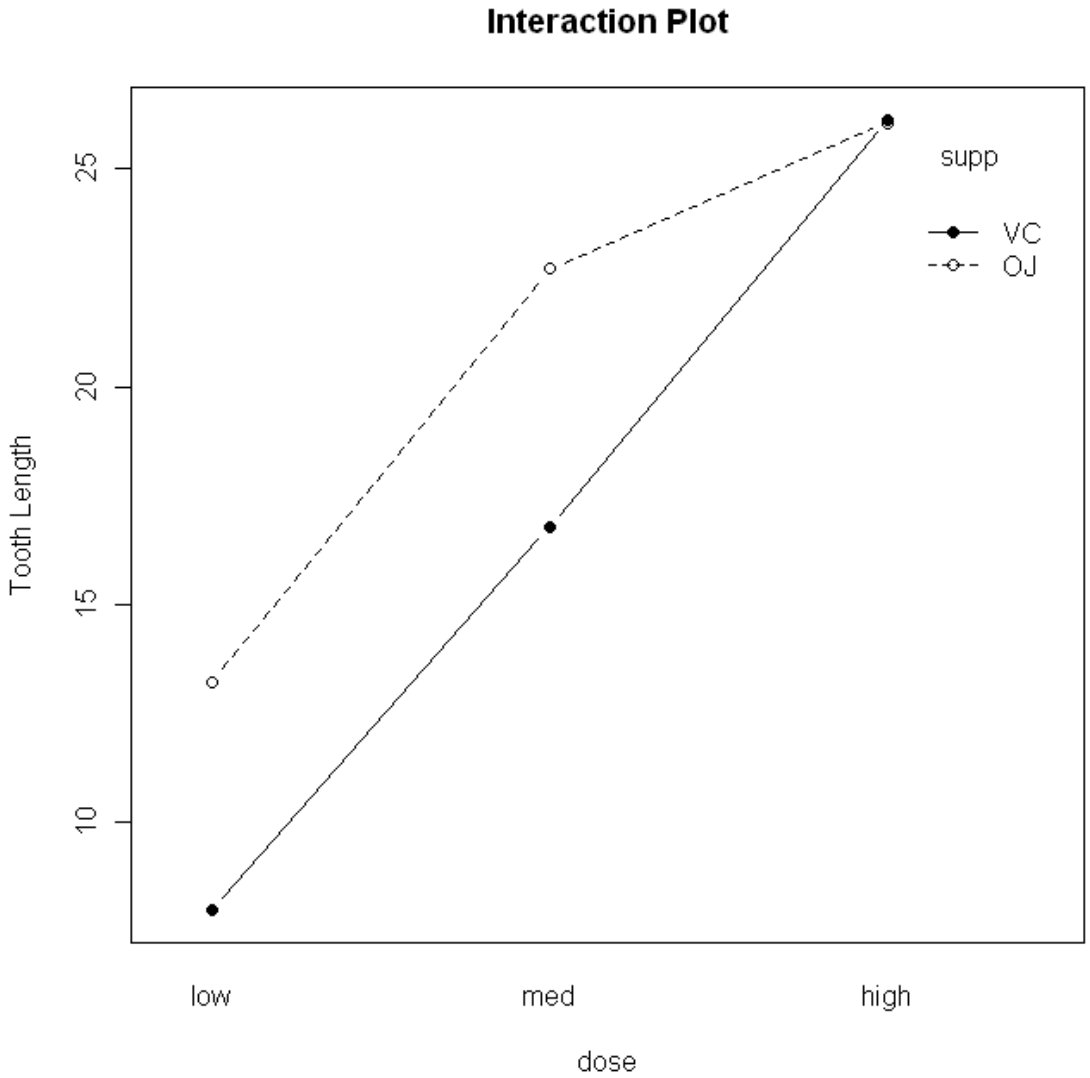


Example 11.3, cont : Graphics



ToothGrowth data: length vs dose, given type of supplement

Example 11.3, cont. : Interaction plot



Example 11.3, cont : ANOVA table output

```
> aov.out = aov(len ~ supp * dose, data=ToothGrowth)
> summary(aov.out)

             Df Sum Sq Mean Sq F value   Pr(>F)
supp          1  205.3   205.3  15.572 0.000231 ***
dose          2 2426.4  1213.2  92.000  < 2e-16 ***
supp:dose     2  108.3    54.2   4.107 0.021860 *
Residuals    54  712.1    13.2
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘
’ 1
```

Unequal sample sizes

- When all sample sizes are equal, the main effects and interactions can be estimated independently *independently*
- That's because of the orthogonality of the sub-spaces that correspond to the different model effects
- This is no longer the case when the sample sizes are different (unbalanced case) :
  $SSModel \neq SSA + SSB + SSAB$
- For an unbalanced design, effect estimation must be *adjusted* (for the other effects in the model) : the estimated values depend on the other terms in the model and their order of entry
- We can no longer carry out tests $F = \frac{MSx}{MSerror}$
- We must carry out sub-model tests

Example 11.3, cont : Unbalanced subset

	L	M	H
VC	4.2 11.5 7.3	16.5 16.5 15.2 17.3	23.6 18.5
OJ	15.2 21.5 17.6 9.7	19.7 23.3	25.5 26.4 22.4 24.5

Example 11.3, cont. : supp 1st

```
> # full interaction model with
> # supp entering first
>
> fit1 <-
    lm(len ~ supp + doselev + supp:doselev,
      data=toothun)
> anova(fit1)
Analysis of Variance Table

Response: len
              Df Sum Sq Mean Sq F value    Pr(>F)
supp           1 174.46  174.46 17.3664 0.0011049
doselev        2 375.75  187.87 18.7012 0.0001495
supp:doselev   2  17.70    8.85  0.8808 0.4377931
Residuals     13 130.60   10.05
```

Example 11.3, cont : doselev 1st

```
> # full interaction model with doselev
> # entering first
>
> fit2 <-
    lm(len ~ doselev + supp + supp:doselev,
      data=toothun)
> anova(fit2)
Analysis of Variance Table

Response: len
              Df Sum Sq Mean Sq F value    Pr(>F)
doselev        2 396.08  198.04 19.7131 0.0001158
supp           1 154.13  154.13 15.3428 0.0017685
doselev:supp   2  17.70    8.85  0.8808 0.4377931
Residuals     13 130.60   10.05
```

Linear models

- The regression model and the ANOVA model are *linear models*
- A linear model is linear *in the parameters*
- Examples – linear or not ?
  - $y = \beta_0 + \beta_1 x_1^2 + \beta_2 \log x_2 + \epsilon$
  - $y = \beta_0 + e^{\beta_1 x_1^2} + \beta_2 \log x_2 + \epsilon$
  - $y = e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon}$
- Regression : X quantitative(s) – continuous, Y continuous
- ANOVA : X qualitative(s), Y continuous

General linear model

- The model : $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$
- $\boldsymbol{Y}$ a vector (or matrix) of measures (multivariate)
- $\boldsymbol{X}$ a matrix of explanatory variables
- $\boldsymbol{\beta}$ a vector (or matrix) of parameters
- $\boldsymbol{\epsilon}$ a vector (or matrix) of errors/noise
- $\boldsymbol{\epsilon} \sim MVN(0, \Sigma)$
- Normally parameter estimations is carried out by the method of least squares (other methods are also possible)

Examples

- *Regression* (simple or multiple) : the reponse variable and the predictor variables are *continues (quantitatives)*
- *ANOVA* (one or multiple factors) : the reponse variable is *continuous*, the explanatory variables are *qualitatives*
- *ANCOVA* : a fusion of ANOVA and regression
  - the reponse variable *continuous (quantitative)* is modeled ad a function of two (or more) predictor variables, of which *at least one is qualitative*
  - ANCOVA tests whether the factors have an effect on the response variabe after having removed the variance for which the quantitative predictors (the *covariates*) are responsible
- *MANOVA*, *MANCOVA* : the response is *multivariate*

Indicator variables for the model

- The matrix form for the linear model :

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

- According to the form of the matrix $\mathbf{X}$, we are in the case of :
  - *linear regression* ($\mathbf{X}$ is then comprised of the constant 1 and p explanatory variables), or
  - *factorial model* ($\mathbf{X}$ is comprised of **indicator variables associated with the levels** of the factor(s))
  - *ancova* ($\mathbf{X}$ is comprised of both qualitative and quantitative variables)
- In general, the model can contain variables of different types (General linear model)