

GC – Probabilités et Statistique

<http://moodle.epfl.ch/course/view.php?id=14271>

Cours 11b

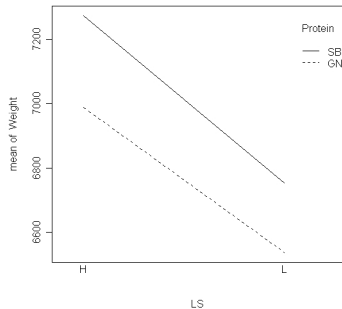
- Expériences factorielles
- 2-way ANOVA (anova à deux voies)
- Modèle général linéaire

Plan d'expérience factorielle et interaction

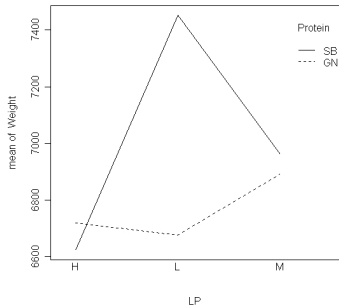
- Exemple : étude de hibernation
 - *Question générale* : Comment les changements dans l'environnement de l'animal provoquent l'animal de commencer à hiberner ?
 - *Question spécifique* : Quel est l'effet du changement de la durée du jour sur la concentration de l'enzyme de la pompe sodium dans deux organes du hamster doré ?
- Comparer deux (ou plus) ensembles de conditions dans la *même expérience* : long/ court ET coeur/cerveau
- Dans cet exemple, il y a 4 combinaisons de conditions :
 - Long/Coeur, Long/Cerveau, Court/Coeur, Court/Cerveau
- Interaction = 'différence des différences'
- Il y a une *interaction* quand l'effet de l'association des traitements n'est pas la somme des effets des traitements
- En cas d'interaction, l'effet d'un traitement *varie suivant qu'il est ou pas associé* à l'autre
- L'interprétation des effets est plus difficile en cas d'interaction

Interaction

pas d'interaction



interaction



Avantages des expériences factorielles

- *Plus efficace* (puissante) qu'une série des expériences étudiant un facteur à la fois
- Permet l'estimation d'*interaction* entre ensembles de conditions qui influence la réponse
- Toutes les données sont utilisées pour l'estimation des effets

ANOVA à deux voies – Introduction

- Étude simultanée d'un facteur A à I modalités et d'un facteur B à J modalités
- Pour chaque couple de modalités (A, B) :
 - on a un échantillon
 - tous les échantillons sont de *mêmes tailles* n (plan équilibré)
- Suppositions :
 - les populations étudiées suivent une distribution normale
 - les variances des populations sont toutes égales (homoscédasticité)
 - les échantillons sont prélevés aléatoirement et indépendamment dans les populations

Exemple 11b.1

- La durée de vie (heures) d'une pile pourrait dépendre sur le type de matière et la température de l'appareil utilisé
- $n = 4$ piles sont testées pour chaque combinaison de type et température
- Les 36 tests sont effectués dans un ordre aléatoire
- L'étude s'adresse les questions :
 - Quels sont les effets du type et température sur la durée de vie
 - Est-ce qu'il y a un type de matière qui prolonge la durée de vie uniformément, quel que soit la température

Exemple 11b.1, cont. : données

Material Type	Temperature (°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

Modèle complet

- Le *modèle complet* : avec interactions
- $y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ijk}$
- $E[\epsilon_{ijk}] = 0$, $Var(\epsilon_{ijk}) = \sigma^2$, $Cov(\epsilon_{ijk}, \epsilon_{i'j'k'}) = 0$ si $(ijk) \neq (i'j'k')$

Tableau d'ANOVA

source	df	SC	CM	F
A	$I - 1$	$nJ \sum_{i=1}^I (\bar{y}_{i..} - \bar{y}_{...})^2$	SC_A / df_A	CM_A / CM_{err}
B	$J - 1$	$nI \sum_{j=1}^J (\bar{y}_{.j.} - \bar{y}_{...})^2$	SC_B / df_B	CM_B / CM_{err}
AB	$(I - 1)(J - 1)$	$n \sum_{j=1}^J \sum_{i=1}^I (y_{ij.} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{y}_{...})^2$	SC_{AB} / df_{AB}	CM_{AB} / CM_{err}
erreur	$IJ(n - 1)$	$\sum_{k=1}^n \sum_{j=1}^J \sum_{i=1}^I (y_{ijk} - \bar{y}_{ij.})^2$	SC_{err} / df_{err}	
total (corr.)	$nIJ - 1$	$\sum_{k=1}^n \sum_{j=1}^J \sum_{i=1}^I (y_{ijk} - \bar{y}_{...})^2$		

* : **n** = nombre PAR cellule (pas la taille de l'échantillon)

Exemple 11b.1, cont. : sorties

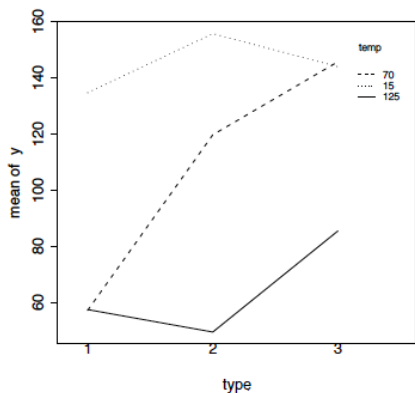
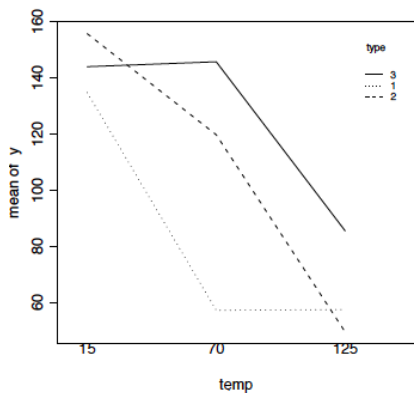
Analysis of Variance Table

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			

- Quelles sont vos conclusions ??

Exemple 11b.1, cont. : interaction plot, 2 vues



Tests d'hypothèses

- Test d'interaction

$$H: \gamma_{ij} = 0, i = 1, \dots, I - 1; j = 1, \dots, J - 1$$

- Statistique de test :

$$F_{AB} = CM_{AB} / CM_{\text{erreur}} \sim F_{(I-1)(J-1), IJ(n-1)} \text{ sous } H$$

- Test d'effet du facteur A

$$H: \alpha_i = 0, i = 1, \dots, I - 1$$

- Statistique de test :

$$F_A = CM_A / CM_{\text{erreur}} \sim F_{I-1, IJ(n-1)} \text{ sous } H$$

- Test d'effet du facteur B

$$H: \beta_j = 0, j = 1, \dots, J - 1$$

- Statistique de test :

$$F_B = CM_B / CM_{\text{erreur}} \sim F_{J-1, IJ(n-1)} \text{ sous } H$$

Modèle additif

- Le *modèle additif* : sans interactions
- $y_{ijk} = \mu + \alpha_i + \beta_j + \epsilon_{ijk}$
- $E[\epsilon_{ijk}] = 0$, $Var(\epsilon_{ijk}) = \sigma^2$, $Cov(\epsilon_{ijk}, \epsilon_{i'j'k'}) = 0$ si $(ijk) \neq (i'j'k')$

Tableau d'ANOVA

source	df	SC	CM	F
A	$I - 1$	$nJ \sum_{i=1}^I (\bar{y}_{i..} - \bar{y}_{...})^2$	SC_A / df_A	CM_A / CM_{err}
B	$J - 1$	$nI \sum_{j=1}^J (\bar{y}_{.j.} - \bar{y}_{...})^2$	SC_B / df_B	CM_B / CM_{err}
erreur	$nIJ - I - J + 1$	$\sum_{k=1}^n \sum_{j=1}^J \sum_{i=1}^I (y_{ijk} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{y}_{...})^2$	SC_{err} / df_{err}	
total (corr.)	$nIJ - 1$	$\sum_{k=1}^n \sum_{j=1}^J \sum_{i=1}^I (y_{ijk} - \bar{y}_{...})^2$		

* : **n** = nombre PAR cellule (pas la taille de l'échantillon)

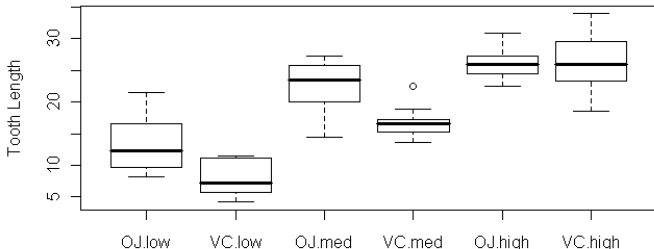
- Quels sont les hypothèses et STs ??



Example 11b.2 : ToothGrowth

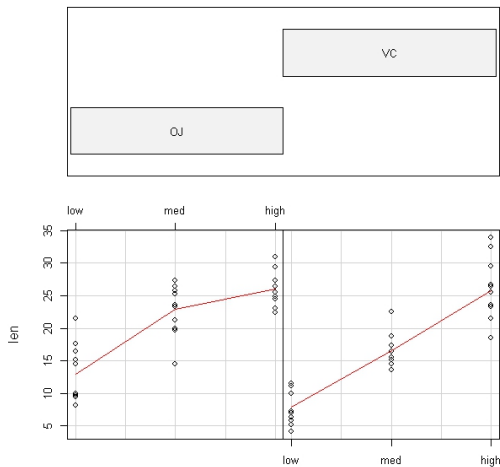
- “The response is the length of odontoblasts (teeth) in each of 10 guinea pigs at each of three dose levels of Vitamin C (0.5, 1, and 2 mg) with each of two delivery methods (orange juice or ascorbic acid).”

Boxplots of Tooth Growth Data



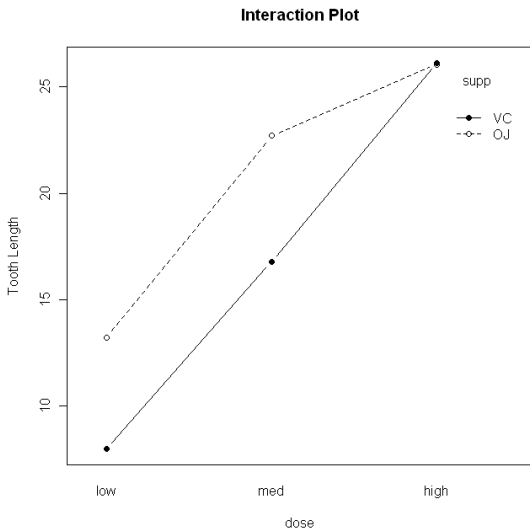
Exemple 11b.2, cont : Graphiques

Given : supp



ToothGrowth data: length vs dose, given type of supplement

Exemple 11b.2, cont. : Interaction plot



Exemple 11b.2, cont : Tableau d'anova sorties

```
> 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

Tailles des échantillons non équilibrées

- Dans le cas équilibré, les effets et les interactions pourraient être estimés *indépendamment*
- C'est grâce à l'orthogonalité des sous-espaces qui correspondent aux différents effets du modèle
- Ce n'est plus le cas si les tailles des échantillons sont différentes (cas non équilibré) :
$$SC_{\text{Modèle}} \neq SCA + SCB + SCAB$$
- Pour un plan non équilibré, les estimations des effets doivent être *ajustées* (pour les autres effets dans le modèle)
- On ne peut plus faire des tests $F = \frac{CM_x}{CM_{\text{Erreur}}}$
- Il faut faire des tests d'un *sous-modèle*

Exemple 11b.2, cont : Sous-ensemble non équilibré

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

Exemple 11b.2, cont. : supp 1er

```
> # 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		

Exemple 11b.2, cont : doselev 1er

```
> # 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		

Modèles linéaires

- Le modèle de régression et le modèle d'ANOVA sont *modèles linéaires*
- Un modèle linéaire est linéaire *dans les paramètres*
- Exemples – linéaire ou non ?
 - $y = \beta_0 + \beta_1 x_1^2 + \beta_2 \log x_2 + \epsilon$
 - $y = \beta_0 + e^{\beta_1 x_1^2} + \beta_2 \log x_2 + \epsilon$
 - $y = e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon}$
- Régression : X quantitative(s) – continue(s), Y continue
- ANOVA : X qualitative(s), Y continue

Modèle linéaire général

- Le modèle : $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$
- $\mathbf{Y}$ un vecteur (ou une matrice) des mesures (multivariées)
- $\mathbf{X}$ une matrice des variables explicatives
- $\boldsymbol{\beta}$ un vecteur (ou une matrice) des paramètres
- $\boldsymbol{\epsilon}$ un vecteur (ou une matrice) des erreurs/bruit
- $\boldsymbol{\epsilon} \sim MVN(0, \Sigma)$
- Normalement, l'estimation des paramètres est fait par la méthode de moindres carrés (autres méthodes possibles)

Exemples

- *Régression* (simple ou multiple) : la variable réponse et les variables prédictrices sont *continues (quantitatives)*
- *ANOVA* (un ou plusieurs facteurs) : la variable réponse est *continue*, les variables explicatives sont *qualitatives*
- *ANCOVA* : la fusion d'ANOVA et de la régression
 - la variable réponse *continue (quantitative)* est modélisée en fonction de deux (ou plus) variables prédictrices dont *l'une au moins est qualitative*
 - ANCOVA teste si certains facteurs ont un effet sur la variable de réponse après avoir enlevé la variance dont les prédicteurs quantitatifs (les *covariables*) sont responsables
- *MANOVA, MANCOVA* : la réponse est *multivariée*

Variables indicatrices pour le modèle

- La forme matricielle pour le modèle linéaire :

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

- Selon la forme de la matrice $\mathbf{X}$, on est dans le cas :
 - de la *régression linéaire* ($\mathbf{X}$ est alors composée de la variable constante 1 et des p variables explicatives), ou
 - du *modèle factoriel* ($\mathbf{X}$ est composée des **variables indicatrices associées aux niveaux** du (ou des) facteur(s))
 - de *ancova* ($\mathbf{X}$ est composée des variables continue(s) et indicatrice(s))
- En général, le modèle pourrait contenir des variables de types différents (modèle linéaire générale)