

Analyse de variance - Modèles à effets fixes

Mathieu Emily

Laboratoire de mathématiques appliquées
Agrocampus Ouest, Rennes



Plan

1 Introduction de la problématique

- Des questions “types”
- Notion de variabilité

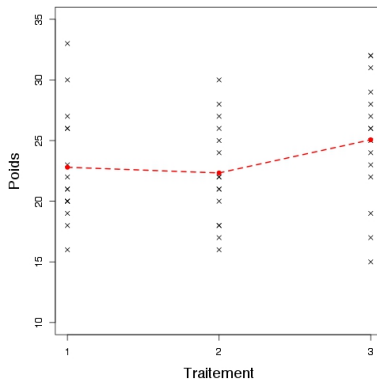
2 Modèles à effets fixes

- Modèles à 1 facteur
- Modèles à 2 facteurs
- Extensions

- Des questions “types”
- Notion de variabilité

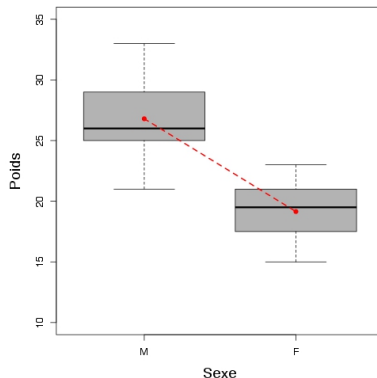
- Modèles à 1 facteur
 - Modèle
 - Estimation
 - Test global
 - Analyse Post-Hoc
 - Hypothèses du modèle et tests non paramétriques
- Modèles à 2 facteurs
 - Modèles
 - Estimation
 - Test global
 - Interaction
- Extensions

Quel est l'effet d'un traitement sur le poids de poussins ?



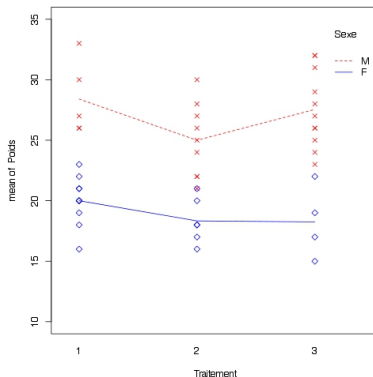
- Poids : variable à expliquer **quantitative**
- Traitement : variable explicative **qualitative** ou **facteur**

Quel est l'effet du sexe sur le poids de poussins ?



- Poids : variable à expliquer **quantitative**
- Sexe : variable explicative **qualitative** ou **facteur**
- Remarque : on utilise une autre représentation de la variabilité par des “boxplot”

Quel est l'effet du traitement et du sexe sur le poids de poussins ?



- Poids : variable à expliquer **quantitative**
- Traitement et Sexe : 2 **facteurs**
- Interprétation des **interactions** : à faire avec précaution !

Objectifs généraux de l'Analyse de variance

- **Expliquer** une variable quantitative par 1 ou plusieurs **facteurs**.
- Expliquer une variable revient à expliquer sa **variabilité** ou variance.
 - ▶ La variabilité mesure la dispersion de la variable
 - ▶ La variabilité se quantifie par la **somme des carrés des écarts par rapport à la moyenne**.

Objectif de l'analyse de variance (Version 1)

Evaluer le fait de considérer les observations d'une même catégorie d'un facteur **égale à la moyenne empirique**, sur la variabilité de la variable à expliquer.

Caractéristiques fondamentales

- **Planification expérimentale**
 - ▶ Orthogonalité des facteurs
 - ▶ Plan **équilibré** (mêmes effectifs pour chaque cellule du plan)
 - ▶ Plan **complet** : toutes les combinaisons de facteurs ont été expérimentées
- **Hypothèses de modélisation**
 - ▶ Modèles **paramétriques**
 - ▶ Modèles non-paramétriques

Plan

- 1 Introduction de la problématique
 - Des questions “types”
 - Notion de variabilité
- 2 Modèles à effets fixes
 - Modèles à 1 facteur
 - Modèle
 - Estimation
 - Test global
 - Analyse Post-Hoc
 - Hypothèses du modèle et tests non paramétriques
 - Modèles à 2 facteurs
 - Modèles
 - Estimation
 - Test global
 - Interaction
 - Extensions

La variabilité

- La variabilité est la notion **de base** de l'Analyse de variance.
 - ▶ Toute la complexité d'un phénomène est quantifiée par sa variabilité.
- Un objectif statistique de l'Analyse de variance est de rendre compte du **maximum de variabilité** d'un phénomène en tenant compte du **minimum d'information**.
 - ▶ L'information utilisée est caractérisée par la dimension des facteurs ou **degrés de liberté**.
 - ▶ Compromis classique en statistique entre **qualité d'ajustement** et **parcimonie**.

Objectif de l'analyse de variance (Version 2)

Evaluer la variabilité expliquée par un facteur en pondérant par rapport à la dimension (ou degré de liberté) de ce facteur.

Exemple jouet

- Considérons le phénomène de la visualisation d'une photo¹.



- Le phénomène est **caractérisé par la variabilité** des pixels de l'image.
- La variabilité est **mesurée par la Somme des Carrés des Ecart (SCE)** pour 256 couleurs vaut :

$$SCE_T = 9232 \quad \text{et} \quad \text{ddl} = 256 - 1 = 255$$

1. La photo "lena" sert d'image de test pour les algorithmes de traitement d'image et est devenue un standard industriel et scientifique.

Exemple jouet : modèle 1

- Un spécialiste connaît les **zones de couleur moyenne**.
 - ▶ Peut-on **conserver une bonne** visualisation du phénomène ?
- Les pixels “de couleur moyenne” sont ramenés à leur **moyenne empirique** (et de même pour les autres pixels).



- Pensez-vous que l'information donnée par le spécialiste est **pertinente** ?

Exemple jouet : modèle 1

- Pour évaluer un modèle on compare la variabilité du **modèle** à celle de la **résiduelle**
- Evaluation numérique :

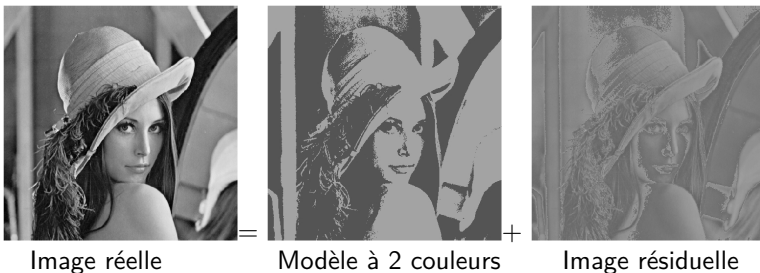
$$SCE_{M1} = \mathbf{160}$$

$$SCE_{R1} = \mathbf{9072}$$

$$F_{M1} = \mathbf{4.48} \quad p.val = \mathbf{0.035}$$

Exemple jouet : modèle 2

- Un autre spécialiste connaît les zones **foncées**.
 - ▶ Peut-on conserver une bonne visualisation du phénomène ?



$$SCE_{M2} = \mathbf{6330} \quad SCE_{R2} = \mathbf{2901}$$

$$F_{M2} = 554.23 \quad p.val < 2.2e - 16$$

Exemple jouet 3

- Un non-spécialiste ne connaît rien et affecte des pixels foncés **au hasard**.
 - Peut-on conserver une bonne visualisation du phénomène ?



Image réelle

=



Modèle à 2 couleurs

+

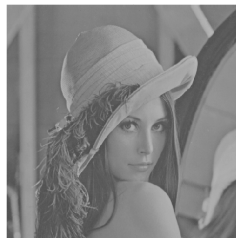


Image résiduelle

$$SCE_{M3} = \mathbf{0.02} \quad SCE_R = \mathbf{9231}$$

$$F_{M3} = \mathbf{0.00076} \quad p.val = 0.97$$

Exemple jouet : bilan

- Dans les exemples précédents, nous avons cherché à **modéliser** (ou simplifier) un phénomène complexe (photo avec 256 couleurs) par un phénomène plus simple (photo à 2 couleurs)
 - ▶ La pertinence du modèle est évaluée par la **part de variabilité conservée par le modèle**.
 - ▶ Cette mesure s'effectue par le calcul de la **somme des carrés des écarts**.
- En détails on remarque que :
 - ▶ Modèle 1 : information pertinent mais non-linéaire.
 - ▶ Modèle 2 : information très pertinente.
 - ▶ Modèle 3 : information pas du tout pertinente.

Plan

1 Introduction de la problématique

- Des questions “types”
- Notion de variabilité

2 Modèles à effets fixes

- Modèles à 1 facteur
- Modèles à 2 facteurs
- Extensions

Exemple illustratif

- Espèce étudiée : **Vache**
- Taille d'échantillon : $n = 30$ **observations**
- Une variable à expliquer : le gain moyen de poids sur 30 jours.
- 2 variables explicatives :
 - ▶ Milieu à 3 modalités (Bon, médiocre et moyen)
 - ▶ Génotype à 2 modalités : Pur et Croisé

```
> charolais
      Genotype Milieu GMQ
1      Croisé   Bon  834
2      Croisé   Bon  780
3      Croisé   Bon  813
4      Croisé   Bon  806
-      -      -      -
```

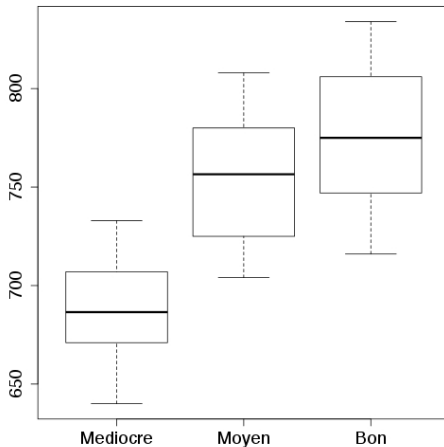
Démarche à suivre

- L'étude de l'impact de facteurs sur une variable quantitative peut se faire à **différents niveaux**.
- Du point de vue statistique, il convient de respecter l'ordre du schéma d'analyse suivant :
 - ① Pertinence du modèle **global**
 - ② Effet de **chaque facteur**
 - ③ Comparaison des **modalités** pour un facteur donné

Plan

- 1 Introduction de la problématique
 - Des questions “types”
 - Notion de variabilité
- 2 **Modèles à effets fixes**
 - **Modèles à 1 facteur**
 - Modèle
 - Estimation
 - Test global
 - Analyse Post-Hoc
 - Hypothèses du modèle et tests non paramétriques
 - Modèles à 2 facteurs
 - Modèles
 - Estimation
 - Test global
 - Interaction
 - Extensions

Illustration : effet du Milieu sur le gain de poids



- Faire l'exemple sous R

Plan

- 1 Introduction de la problématique
 - Des questions “types”
 - Notion de variabilité
- 2 Modèles à effets fixes
 - Modèles à 1 facteur
 - Modèle
 - Estimation
 - Test global
 - Analyse Post-Hoc
 - Hypothèses du modèle et tests non paramétriques
 - Modèles à 2 facteurs
 - Modèles
 - Estimation
 - Test global
 - Interaction
 - Extensions

Ecriture du modèle

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

- i représente le **niveau** du facteur ($i = 1, \dots, I$)
- j représente l'indice de répétition ($j = 1, \dots, n_i$)
 - ▶ Si le plan est **équirépété** : $n_i = J$
- **Hypothèses de modélisation :**
 - ▶ $\forall i, j, \varepsilon_{ij} \sim \mathcal{N}(0, \sigma)$
 - ▶ $\forall i' \neq i \text{ ou } j' \neq j : \text{Cov}(\varepsilon_{ij}, \varepsilon_{i'j'}) = 0$

Interprétation du modèle

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

- La variable Y_{ij} est la somme
 - ▶ d'un terme **constant** μ
 - ▶ d'un effet **spécifique** à chaque niveau du facteur α_i
 - ▶ un **résidu** aléatoire gaussien ε_{ij}
- Le modèle possède $I + 1$ paramètres :
 - ▶ La dimension paramétrique du modèle est I !
 - ▶ Une **contrainte de liaison** doit être explicitée pour estimer les paramètres.

Contraintes sur les paramètres

- Nullité du terme constant : $\mu = 0$

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

- **Nullité de la somme des effets individuels**

$$\sum_{i=1}^I \alpha_i = 0$$

- Nullité de la somme pondérée des effets individuels

$$\sum_{i=1}^I n_i \alpha_i = 0$$

- Une modalité de référence : i^*

$$\alpha_{i^*} = 0$$

Plan

- 1 Introduction de la problématique
 - Des questions “types”
 - Notion de variabilité
- 2 **Modèles à effets fixes**
 - **Modèles à 1 facteur**
 - Modèle
 - **Estimation**
 - Test global
 - Analyse Post-Hoc
 - Hypothèses du modèle et tests non paramétriques
 - Modèles à 2 facteurs
 - Modèles
 - Estimation
 - Test global
 - Interaction
 - Extensions

Estimation des paramètres : principe des Moindres Carrés Ordinaires

- On cherche à minimiser l'**écart entre les observations et le modèle** :

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

- Pour chaque modalité i , minimiser $\sum_{j=1}^{n_i} (Y_{ij} - \widehat{Y}_{ij})^2$ se fait par :

$$\widehat{\mu}_i = Y_{i.} = \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{ij}$$

L'estimation du modèle permet d'utiliser le modèle pour la prédiction

Y_{ij} est prédite par son **espérance** \widehat{Y}_{ij}

- Pour le modèle avec nullité de la constante, l'estimation des paramètres permet donc de prédire chaque modalité du facteur par sa **fréquence observée**.

Autres contraintes

- Somme des effets individuels nulle

$$\hat{\mu} = \frac{1}{I} \sum_{i=1}^I Y_{i.} \quad \text{et} \quad \forall i, \hat{\alpha}_i = Y_{i.} - \hat{\mu}$$

- Somme pondérée des effets individuels nulle

$$\hat{\mu} = Y_{..} = \frac{1}{I} \sum_{i=1}^I n_i Y_{i.} \quad \text{et} \quad \forall i, \hat{\alpha}_i = Y_{i.} - Y_{..}$$

- Une modalité de référence : i^*

$$\hat{\mu} = Y_{i^*}. \quad \text{et} \quad \forall i, \hat{\alpha}_i = Y_{i.} - Y_{i^*}.$$

Estimation des résidus

- Degrés de liberté : **n-l**
 - ▶ n données
 - ▶ l paramètres estimés dans le modèle
- Un estimateur **non biaisé** de σ est donné par

$$\widehat{\sigma^2} = \frac{1}{n-l} \sum_{i=1}^l \sum_{j=1}^{n_i} (Y_{ij} - \widehat{Y}_{ij})^2$$

Sortie R

```
> modele.simple <- lm(GMQ ~ Milieu,data=charolais)
> summary(modele.simple)
```

```
Call:
lm(formula = GMQ ~ Milieu, data = charolais)
```

```
Residuals:
    Min       1Q   Median       3Q      Max
-60.00 -25.25   2.50  24.75  58.00
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  688.00      10.45   65.831 < 2e-16 ***
MilieuMoyen   66.00       14.78   4.466 0.000128 ***
MilieuBon     88.00       14.78   5.954 2.39e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 33.05 on 27 degrees of freedom
Multiple R-squared:  0.5872,    Adjusted R-squared:  0.5566
F-statistic: 19.2 on 2 and 27 DF,  p-value: 6.496e-06
```

```
> options(contrasts=c("contr.sum","contr.sum"))
> modele.simple.contrainte <- lm(GMQ ~ Milieu,data=charolais)
> summary(modele.simple.contrainte)
```

```
Call:
lm(formula = GMQ ~ Milieu, data = charolais)
```

```
Residuals:
    Min       1Q   Median       3Q      Max
-60.00 -25.25   2.50  24.75  58.00
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  739.333      6.034 122.531 < 2e-16 ***
Milieu1     -51.333      8.533  -6.016 2.03e-06 ***
Milieu2      14.667      8.533   1.719  0.0971 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 33.05 on 27 degrees of freedom
Multiple R-squared:  0.5872,    Adjusted R-squared:  0.5566
F-statistic: 19.2 on 2 and 27 DF,  p-value: 6.496e-06
```

Sortie SAS avec la proc glm

```
proc glm data=charolais
  class Milieu;
  model GMQ=Milieu / solution ss1 ss3;
run;
```

Results Viewer - sashtml

Dependent Variable: GMQ

Source	DDL	Somme des carrés	Moyenne quadratique	Valeur F	Pr > F
Model	2	41946.66667	20973.33333	19.20	< .0001
Error	27	29490.00000	1092.22222		
Corrected Total	29	71436.66667			

r.carré	Coef de Var	Racine MSE	GMQ Moyenne
0.587187	4.470079	33.04879	739.3333

Source	DDL	Type I SS	Moyenne quadratique	Valeur F	Pr > F
Milieu	2	41946.66667	20973.33333	19.20	< .0001

Source	DDL	Type III SS	Moyenne quadratique	Valeur F	Pr > F
Milieu	2	41946.66667	20973.33333	19.20	< .0001

Paramètre	Valeur estimée	Erreur type	Valeur du test t	Pr > t
Intercept	754.0000000	10.45094360	72.15	< .0001
Milieu Bon	22.0000000	14.77986618	1.49	0.1482
Milieu Mediocre	-66.0000000	14.77986618	-4.47	0.0001
Milieu Moyen	0.0000000	-	-	-

Utilisation d'un modèle d'ANOVA

- Un modèle d'ANOVA est rarement utilisé pour **prédire un phénomène**.
- Un modèle d'ANOVA est surtout utilisé pour savoir si le facteur considéré a un **effet** ou non sur la variable réponse.

Question statistique

La prise en compte des différents niveaux du facteurs contribue-t-elle **significativement** à expliquer la variabilité de Y ?

- Retour sous R

Plan

- 1 Introduction de la problématique
 - Des questions “types”
 - Notion de variabilité
- 2 Modèles à effets fixes
 - Modèles à 1 facteur
 - Modèle
 - Estimation
 - Test global
 - Analyse Post-Hoc
 - Hypothèses du modèle et tests non paramétriques
 - Modèles à 2 facteurs
 - Modèles
 - Estimation
 - Test global
 - Interaction
 - Extensions

Décomposition de la variabilité

$$\begin{aligned}
 SCE_T &= \sum_{i=1}^I \sum_{j=1}^{n_i} (Y_{ij} - Y_{..})^2 \\
 &\vdots \\
 &= \sum_{i=1}^I n_i (Y_{i.} - Y_{..})^2 + \sum_{i=1}^I \sum_{j=1}^{n_i} (Y_{ij} - Y_{i.})^2 \\
 &= \text{SCE}_M + \text{SCE}_R
 \end{aligned}$$

- SCE_T (resp. SCE_M , SCE_R) : somme des carrés des écarts Totale (resp. Modèle, Résiduelle)
- SCE_M : variabilité inter-classe = variabilité expliquée par le modèle
- SCE_R : variabilité intra-classe = variabilité résiduelle

Hypothèses de test

- Pour tester l'effet d'un facteur on peut :

- ▶ faire une **comparaison de modèles** :

$$\mathcal{H}_0 : Y_{ij} = \mu + \varepsilon_{ij} \quad \text{contre} \quad \mathcal{H}_1 : Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

- ▶ ce qui revient à :

$$\mathcal{H}_0 : \forall i \quad \alpha_i = 0 \quad \text{contre} \quad \mathcal{H}_1 : \exists i \quad \alpha_i \neq 0$$

- Pour ce test, on va comparer la variabilité expliquée par le modèle, SCE_M , à la variabilité résiduelle, SCE_R .

- ▶ Si les 2 variabilités sont du **même ordre** de grandeur il n'y a pas d'effet du facteur.
- ▶ Pour comparer SCE_M et SCE_R , il faut diviser par leurs **degrés de liberté** et comparer CM_M et CM_R :

$$CM_M = \frac{SCE_M}{I - 1} \quad \text{et} \quad CM_R = \frac{SCE_R}{n - I}$$

Test global

- On peut montrer :

$$\mathbb{E}[CM_M] = \sigma^2 + \frac{1}{I-1} \sum_{i=1}^I n_i \alpha_i^2 \quad \text{et} \quad \mathbb{E}[CM_R] = \sigma^2$$

- Sous \mathcal{H}_0 on a :

$$\mathbb{E}[CM_M] = \mathbb{E}[CM_R]$$

$$\frac{SCE_M}{\sigma^2} \sim_{\mathcal{H}_0} \chi_{I-1}^2 \quad \text{et} \quad \frac{SCE_R}{\sigma^2} \sim_{\mathcal{H}_0} \chi_{n-I}^2$$

Ainsi :

$$F_{\text{Obs}} = \frac{CM_M}{CM_R} \sim_{\mathcal{H}_0} \mathcal{F}_{n-I}^{I-1}$$

Tableau d'analyse de variance

Variabilité	SCE	ddl	CM	Statistique
Facteur	$\sum_i n_i (Y_{i.} - Y_{..})^2$	$I - 1$	$SCE_M / (I - 1)$	$F_{Obs} = CM_M / CM_R$
Résiduelle	$\sum_{ij} (Y_{ij} - Y_{i.})^2$	$n - I$	$SCE_R / (n - I)$	
Totale	$\sum_{i,j} (Y_{ij} - Y_{..})^2$	$n - 1$		

- Règle de décision :
 - ▶ Si $F_{Obs} < \mathcal{F}_{n-I}^{I-1}(1 - \alpha)$, on **accepte** \mathcal{H}_0 au seuil α .
 - ▶ Si $F_{Obs} > \mathcal{F}_{n-I}^{I-1}(1 - \alpha)$, on **rejette** \mathcal{H}_0 au seuil α .
- Remarque : le test est **unilatéral**. Pour rejeter \mathcal{H}_0 , on cherche à avoir $CM_M > CM_R$.

Tableau d'analyse de variance avec R

```
> anova(modele.simple)
Analysis of Variance Table

Response: GMQ
          Df Sum Sq Mean Sq F value    Pr(>F)
Milieu      2  41947 20973.3   19.202 6.496e-06 ***
Residuals 27  29490  1092.2
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- On note donc un effet **significatif** du milieu sur le gain moyen de poids.
- Remarque : le tableau d'ANOVA ne dépend pas de la **contrainte**.

Tableau d'analyse de variance avec SAS

```
proc glm data=charolais
  class Milieu;
  model GMQ=Milieu / solution ss1 ss3;
run;
```

Results Viewer - sashtml

Dependent Variable: GMQ

Source	DDL	Somme des carrés	Moyenne quadratique	Valeur F	Pr > F
Model	2	41946.66667	20973.33333	19.20	< .0001
Error	27	29490.00000	1092.22222		
Corrected Total	29	71436.66667			

r.carré	Coef de Var	Racine MSE	GMQ Moyenne
0.587187	4.470079	33.04879	739.3333

Source	DDL	Type I SS	Moyenne quadratique	Valeur F	Pr > F
Milieu	2	41946.66667	20973.33333	19.20	< .0001

Source	DDL	Type III SS	Moyenne quadratique	Valeur F	Pr > F
Milieu	2	41946.66667	20973.33333	19.20	< .0001

Paramètre	Valeur estimée	Erreur type	Valeur du test t	Pr > t
Intercept	754.0000000	10.45094360	72.15	< .0001
Milieu Bon	22.00000000	14.77986618	1.49	0.1482
Milieu Mediocre	-66.00000000	14.77986618	-4.47	0.0001
Milieu Moyen	0.00000000	-	-	-

Test de conformité

- $\mathcal{H}_0 : \alpha_i = c$ vs $\mathcal{H}_0 : \alpha_i \neq c$
- Sous la contrainte de nullité de somme des effets on :

$$\hat{\alpha}_i \sim \mathcal{N}(\alpha_i, \sigma_{\hat{\alpha}_i}^2)$$

- On estime $\sigma_{\hat{\alpha}_i}^2$ par :

$$\widehat{\sigma_{\hat{\alpha}_i}^2} = \frac{I-1}{I} \frac{\widehat{\sigma^2}}{J} = \frac{I-1}{I} \frac{CM_R}{J}$$

- Ainsi :

$$t_{Obs} = \frac{\hat{\alpha}_i - c}{\widehat{\sigma_{\hat{\alpha}_i}^2}} \sim_{\mathcal{H}_0} \mathcal{T}_{n-I}$$

- On peut donc construire un test de Student à partir de $|t_{Obs}|$:
 - ▶ Si $|t_{Obs}| < t_{n-I}(1 - \alpha/2)$ on accepte \mathcal{H}_0 au seuil α
 - ▶ Si $|t_{Obs}| > t_{n-I}(1 - \alpha/2)$ on rejette \mathcal{H}_0 au seuil α

Plan

- 1 Introduction de la problématique
 - Des questions “types”
 - Notion de variabilité
- 2 **Modèles à effets fixes**
 - **Modèles à 1 facteur**
 - Modèle
 - Estimation
 - Test global
 - **Analyse Post-Hoc**
 - Hypothèses du modèle et tests non paramétriques
 - Modèles à 2 facteurs
 - Modèles
 - Estimation
 - Test global
 - Interaction
 - Extensions

Analyse Post-Hoc : comparaison de 2 moyennes

- Si $\mathcal{H}_0 : \forall i \alpha_i = 0$ est rejetée, on peut rechercher plus précisément les **modalités du facteur qui diffèrent**.
- Soit i et i' , 2 modalités du facteur, on s'intéresse à tester l'hypothèse : $\mathcal{H}_0 : \forall i \mu_i = \mu_{i'}$.
- On utilise alors un principe de comparaison de **moyennes** :

$$\hat{\mu}_i - \hat{\mu}_{i'} \sim \mathcal{N} \left(\mu_i - \mu_{i'}, \sqrt{\left(\frac{1}{n_i} + \frac{1}{n_{i'}} \right) \sigma^2} \right)$$

- Il faut estimer σ^2 : on utilise l'hypothèse d'homoscédasticité et l'information de la **résiduelle** (et donc son degré de liberté : $ddl_R = n - I$) :

$$t_{i,i'} = \frac{\hat{\mu}_i - \hat{\mu}_{i'} - (\mu_i - \mu_{i'})}{\hat{\sigma} \sqrt{\frac{1}{n_i} + \frac{1}{n_{i'}}}} \sim \mathcal{T}_{ddl_R}$$

- On pose alors : $LSD = t_{ddl_R}(1 - \alpha/2) \hat{\sigma} \sqrt{\frac{1}{n_i} + \frac{1}{n_{i'}}}$
 - ▶ Si $|\mu_i - \mu_{i'}| > LSD$, on **rejette** \mathcal{H}_0

Analyse Post-Hoc : comparaison de > 2 moyennes

- Attention : si l'on effectue plusieurs fois cette comparaison, **le niveau α du test n'est plus respecté.**
- Une stratégie consiste à **corriger** ce niveau α par une procédure de correction de tests multiples (**Bonferroni** par exemple)
- D'autres procédures s'appuient sur la statistique d'**écart studentisé** :

$$q_{r,ddl} = \frac{M_{r_1} - M_{r_2}}{\sqrt{\frac{1}{2} \left(\frac{\hat{\sigma}^2}{n_r} + \frac{\hat{\sigma}^2}{n_1} \right)}} = t_{r_1, r_2} \sqrt{2}$$

- ▶ r_1 et r_2 sont les moyennes à comparer
 - ▶ ddl est le nombre de degré de liberté de la résiduelle
 - ▶ pour l'ANOVA à 1 facteur : $r = I - 1$ et $ddl = n - I$.
 - ▶ **Tukey**(-Kramer) HSD, Duncan (**MRT**), Newman-Keuls (**SNK**)
- La méthode de **Scheffé** s'appuie sur les contrastes.

Exemple avec sorties R

• Retour sous R



```
> pairwise.t.test(charolais$GMQ,charolais$Milieu,p.adjust="none")
```

Pairwise comparisons using t tests with pooled SD

data: charolais\$GMQ and charolais\$Milieu

```
      Mediocre Moyen
Moyen 0.00013      -
Bon    2.4e-06  0.14820
```

P value adjustment method: none

```
> pairwise.t.test(charolais$GMQ,charolais$Milieu,p.adjust="bonferroni")
```

Pairwise comparisons using t tests with pooled SD

data: charolais\$GMQ and charolais\$Milieu

```
      Mediocre Moyen
Moyen 0.00038      -
Bon    7.2e-06  0.44461
```

P value adjustment method: bonferroni

```
> require(multcomp)
```

```
> modele.lway <- lm(GMQ ~ Milieu,data=charolais)
```

```
> tuk <- glht(modele.lway,linfct=mcp(Milieu="Tukey"))
```

```
> summary(tuk)
```

Simultaneous Tests for General Linear Hypotheses

Multiple Comparisons of Means: Tukey Contrasts

Fit: lm(formula = GMQ ~ Milieu, data = charolais)

Linear Hypotheses:

	Estimate	Std. Error	t value	Pr(> t)
Moyen - Mediocre == 0	66.00	14.78	4.466	<0.001 ***
Bon - Mediocre == 0	88.00	14.78	5.954	<0.001 ***
Bon - Moyen == 0	22.00	14.78	1.489	0.312

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Adjusted p values reported -- single-step method)

Exemple avec sorties R

- Retour sous R
- `pairwise.t.test` (pk stats) et `glht` (pk multcomp)

```
> pairwise.t.test(charolais$GMQ,charolais$Milieu,p.adjust="none")
```

Pairwise comparisons using t tests with pooled SD

data: charolais\$GMQ and charolais\$Milieu

```
      Mediocre Moyen
Moyen 0.00013  -
Bon    2.4e-06  0.14820
```

P value adjustment method: none

```
> pairwise.t.test(charolais$GMQ,charolais$Milieu,p.adjust="bonferroni")
```

Pairwise comparisons using t tests with pooled SD

data: charolais\$GMQ and charolais\$Milieu

```
      Mediocre Moyen
Moyen 0.00038  -
Bon    7.2e-06  0.44461
```

P value adjustment method: bonferroni

```
> require(multcomp)
```

```
> modele.lway <- lm(GMQ ~ Milieu,data=charolais)
```

```
> tuk <- glht(modele.lway,linfct=mcp(Milieu="Tukey"))
```

```
> summary(tuk)
```

Simultaneous Tests for General Linear Hypotheses

Multiple Comparisons of Means: Tukey Contrasts

Fit: `lm(formula = GMQ ~ Milieu, data = charolais)`

Linear Hypotheses:

	Estimate	Std. Error	t value	Pr(> t)
Moyen - Mediocre == 0	66.00	14.78	4.466	<0.001 ***
Bon - Mediocre == 0	88.00	14.78	5.954	<0.001 ***
Bon - Moyen == 0	22.00	14.78	1.489	0.312

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Adjusted p values reported -- single-step method)

- Le package `agricole` permettent de créer des groupes entre les modalités.

Exemple avec sorties SAS : utilisation de l'option lsmeans

```
proc glm data=charolais
  class Milieu;
  model GMQ=Milieu / solution ss1 ss3;
  lsmeans Milieu / adjust = T;
  lsmeans Milieu / adjust = BON;
  lsmeans Milieu / adjust = TUKEY;

run;
```

Le Système SAS

The GLM Procedure
Least Squares Means

Milieu	GMQ LSMEAN	Nombre LSMEAN
Bon	776.000000	1
Mediocre	688.000000	2
Moyen	754.000000	3

Least Squares Means for effect Milieu
Pr > |t| for H0: LSMean(i)-LSMean(j)
Dependent Variable: GMQ

i/j	1	2	3
1		<.0001	0.1482
2	<.0001		0.0001
3	0.1482	0.0001	

Le Système SAS

The GLM Procedure
Least Squares Means

Adjustment for Multiple Comparisons: Bonferroni

Milieu	GMQ LSMEAN	Nombre LSMEAN
Bon	776.000000	1
Mediocre	688.000000	2
Moyen	754.000000	3

Least Squares Means for effect Milieu
Pr > |t| for H0: LSMean(i)-LSMean(j)
Dependent Variable: GMQ

i/j	1	2	3
1		<.0001	0.4446
2	<.0001		0.0004
3	0.4446	0.0004	

Le Système SAS

The GLM Procedure
Least Squares Means
Adjustment for Multiple Comparisons: Tukey

Milieu	GMQ LSMEAN	Nombre LSMEAN
Bon	776.000000	1
Mediocre	688.000000	2
Moyen	754.000000	3

Least Squares Means for effect Milieu
Pr > |t| for H0: LSMean(i)-LSMean(j)
Dependent Variable: GMQ

i/j	1	2	3
1		<.0001	0.3121
2	<.0001		0.0004
3	0.3121	0.0004	

Méthode des contrastes

- Un contraste est une combinaison linéaire des moyennes (par modalité) telle que la somme des coefficients (C_1, \dots, C_I) valent 0 : $\sum_i C_i = 0$.

$$\text{contraste} = \sum_{i=1}^I C_i Y_{i.}$$

- La SCE expliquée par le contraste est donnée par :

$$SCE_{\text{contraste}} = \frac{\left(\sum_{i=1}^I C_i Y_{i.} \right)^2}{\sum_{i=1}^I C_i^2 / n_i} = CM_{\text{contraste}} \quad \text{car 1 ddl}$$

- En comparant à la résiduelle on a :

$$F_{\text{Obs}} = \frac{CM_{\text{contraste}}}{CM_R} \sim_{\mathcal{H}_0} \mathcal{F}_{\text{ddl}_R}^1$$

- Par exemple $C_1 = 1$, $C_2 = 1$ et $C_3 = -2$ permet de savoir les 2 modalités "1" et "2" sont meilleurs que la modalité "3" :

$$\mathcal{H}_0 : Y_{1.} + Y_{2.} - 2Y_{3.} = 0$$

Exemple de contrastes (1)

- Hypothèse nulle : La modalité "Bon" de Milieu a la même GMQ que les modalités combinées "Médiocre" et "Moyen".
- Cela revient à $\mathcal{H}_0 : Y_1. + Y_2. - 2Y_3. = 0$ d'où le contraste : $C_1 = 1$, $C_2 = 1$ et $C_3 = -2$.
- Pour les logiciels, il faut faire attention à l'ordre des modalités interprété pour chaque variable :
 - Pour SAS, on peut utiliser l'option contrast 'a' Milieu 2 -1 1;

```
> levels(charolais$Milieu)
[1] "Bon"      "Médiocre" "Moyen"
> summary(gllht(modele.lway, lmfct=mcp(Milieu=c(-2,1,1))))
```

Simultaneous Tests for General Linear Hypotheses

Multiple Comparisons of Means: User-defined Contrasts

Fit: lm(formula = GMQ ~ Milieu, data = charolais)

Linear Hypotheses:

	Estimate	Std. Error	t value	Pr(> t)
1 = 0	-110.0	25.6	-4.297	0.000201 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Adjusted p values reported -- single-step method)

Contraste	DDL	SC contrastée	Moyenne quadratique	Valeur F	Pr > F
a	1	20166.66667	20166.66667	18.46	0.0002

- Remarque : R propose un test t (Student) tandis que SAS propose un test F (Fisher)
 - $t_{Obs}^2 = F_{Obs}$ ici car $ddl = 1$

Exemple de contrastes (2)

- Hypothèse nulle : La modalité "Médiocre" de Milieu a la même GMQ que la modalité et "Moyen".
- Cela revient à $\mathcal{H}_0 : Y_{1.} - Y_{2.} = 0$ d'où le contraste : $C_1 = 1$, $C_2 = -1$ et $C_3 = 0$.
- Pour les logiciels, il faut faire attention à l'ordre des modalités interprété pour chaque variable :
 - Pour SAS, on peut utiliser l'option contrast 'a' Milieu 0 1 1;

```
> levels(charolais$Milieu)
[1] "Bon"      "Médiocre" "Moyen"
> summary(glmt(modele.lway, lmfct=mcp(Milieu=c(0,1,-1))))
```

Simultaneous Tests for General Linear Hypotheses

Multiple Comparisons of Means: User-defined Contrasts

Fit: lm(formula = GMQ ~ Milieu, data = charolais)

Linear Hypotheses:

	Estimate	Std. Error	t value	Pr(> t)
1 = 0	-66.00	14.78	-4.466	0.000128 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Adjusted p values reported -- single-step method)

Contraste	DDL	SC contrastée	Moyenne quadratique	Valeur F	Pr > F
a	1	21780.00000	21780.00000	19.94	0.0001

Plan

- 1 Introduction de la problématique
 - Des questions “types”
 - Notion de variabilité
- 2 Modèles à effets fixes
 - Modèles à 1 facteur
 - Modèle
 - Estimation
 - Test global
 - Analyse Post-Hoc
 - Hypothèses du modèle et tests non paramétriques
 - Modèles à 2 facteurs
 - Modèles
 - Estimation
 - Test global
 - Interaction
 - Extensions

Principales hypothèses

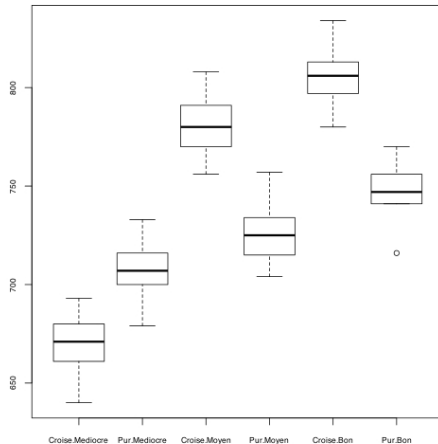


- **Normalité** des résidus :
 - ▶ Test de Shapiro par exemple : `shapiro.test`
- **Homoscédasticité** :
 - ▶ Test de Bartlett par exemple : `bartlett.test`
- **Indépendance** des observations :
 - ▶ Quasiment impossible à tester en pratique
- Variante non-paramétrique :
 - ▶ Test de Mann-Whitney si $I = 2$
 - ▶ Test de Kruskal-Wallis si $I > 2$

Plan

- 1 Introduction de la problématique
 - Des questions “types”
 - Notion de variabilité
- 2 Modèles à effets fixes
 - Modèles à 1 facteur
 - Modèle
 - Estimation
 - Test global
 - Analyse Post-Hoc
 - Hypothèses du modèle et tests non paramétriques
 - Modèles à 2 facteurs
 - Modèles
 - Estimation
 - Test global
 - Interaction
 - Extensions

Illustration : effet du Milieu et du génotype sur le gain de poids



Plan

- 1 Introduction de la problématique
 - Des questions “types”
 - Notion de variabilité
- 2 Modèles à effets fixes
 - Modèles à 1 facteur
 - Modèle
 - Estimation
 - Test global
 - Analyse Post-Hoc
 - Hypothèses du modèle et tests non paramétriques
 - Modèles à 2 facteurs
 - Modèles
 - Estimation
 - Test global
 - Interaction
 - Extensions

Ecriture du modèle

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \varepsilon_{ijk}$$

- i représente le niveau du premier facteur ($i = 1, \dots, I$)
- j représente le niveau du deuxième facteur ($j = 1, \dots, J$)
- k représente l'indice de répétition ($k = 1, \dots, n_{ij}$)
 - ▶ Si le plan est équilibré : $n_{ij} = K$
- Hypothèses de modélisation :
 - ▶ $\forall i, j, k, \varepsilon_{ijk} \sim \mathcal{N}(0, \sigma)$
 - ▶ $\forall i' \neq i$ ou $j' \neq j$ ou $k' \neq k$: $\text{Cov}(\varepsilon_{ikj}, \varepsilon_{i'j'k'}) = 0$

Interprétation du modèle

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \varepsilon_{ijk}$$

- La variable Y_{ijk} est la somme
 - ▶ d'un terme constant μ
 - ▶ d'un effet spécifique à chaque niveau du premier facteur α_i
 - ▶ d'un effet spécifique à chaque niveau du **second** facteur β_j
 - ▶ un résidu aléatoire gaussien ε_{ijk}
- Le modèle possède **$I + J + 1$** paramètres :
 - ▶ La dimension paramétrique du modèle est **$I + J - 1$** !
 - ▶ Une **contrainte de liaison** doit être explicitée pour estimer les paramètres.

Plan

- 1 Introduction de la problématique
 - Des questions “types”
 - Notion de variabilité
- 2 Modèles à effets fixes
 - Modèles à 1 facteur
 - Modèle
 - Estimation
 - Test global
 - Analyse Post-Hoc
 - Hypothèses du modèle et tests non paramétriques
 - Modèles à 2 facteurs
 - Modèles
 - **Estimation**
 - Test global
 - Interaction
 - Extensions

Estimation des paramètres : principe des Moindres Carrés Ordinaires

- On se place dans le cas **équilibré** $n_{ij} = K$.
- On cherche à minimiser l'écart entre les observations et le modèle :

$$\sum_{i=1}^I \sum_{j=1}^J \sum_{k=1}^{n_{ij}} (Y_{ijk} - \widehat{Y}_{ijk})^2$$

- Nullité de la **somme des effets individuels** ($\sum_{i=1}^I \alpha_i = 0$ et $\sum_{j=1}^J \beta_j = 0$)

$$\widehat{\mu} = Y_{...} \quad \text{et} \quad \forall i, \widehat{\alpha}_i = Y_{i..} - Y_{...} \quad \text{et} \quad \forall j, \widehat{\beta}_j = Y_{.j.} - Y_{...}$$

- Pour les résidus :
 - Degrés de liberté : $n - I - J + 1$
 - Estimateur **sans biais** de σ :

$$\widehat{\sigma}^2 = \frac{1}{n - I - J + 1} \sum_{i=1}^I \sum_{j=1}^J \sum_{k=1}^K (Y_{ijk} - \widehat{Y}_{ijk})^2$$

Plan

- 1 Introduction de la problématique
 - Des questions “types”
 - Notion de variabilité
- 2 Modèles à effets fixes
 - Modèles à 1 facteur
 - Modèle
 - Estimation
 - Test global
 - Analyse Post-Hoc
 - Hypothèses du modèle et tests non paramétriques
 - Modèles à 2 facteurs
 - Modèles
 - Estimation
 - Test global
 - Interaction
 - Extensions

La variabilité du modèle

- Dans le cas **complet** (toutes les combinaisons des facteurs sont testés) et **équilibré** (chaque combinaison est testée un même nombre de fois), on a :

$$\sum_{i,j,k} (Y_{ijk} - Y_{...})^2 = \sum_{i,j,k} (Y_{i..} - Y_{...})^2 + \sum_{i,j,k} (Y_{.j.} - Y_{...})^2 + \sum_{i,j,k} (Y_{ijk} - (Y_{i..} + Y_{.j.} - Y_{...}))^2$$

$$\text{SCE}_T = \text{SCE}_{F_1} + \text{SCE}_{F_2} + \text{SCE}_R$$

Variabilité	Somme des carrés	ddl
Facteur 1	$\text{SCE}_{F_1} = KJ \sum_j (Y_{i..} - Y_{...})^2$	$I - 1$
Facteur 2	$\text{SCE}_{F_2} = KI \sum_j (Y_{.j.} - Y_{...})^2$	$J - 1$
Résiduelle	$\text{SCE}_R = \sum_{i,j,k} (Y_{ijk} - (Y_{i..} + Y_{.j.} - Y_{...}))^2$	$IJK - I - J - 1$
Totale	$\text{SCE}_T = \sum_{i,j,k} (Y_{ijk} - Y_{...})^2$	$IJK - 1$

Test global

- On peut montrer que :

$$\mathbb{E}[CM_{F_1}] = \sigma^2 + \frac{JK}{I-1} \sum_{i=1}^I \alpha_i^2, \quad \mathbb{E}[CM_{F_2}] = \sigma^2 + \frac{IK}{J-1} \sum_{j=1}^J \beta_j^2, \quad \mathbb{E}[CM_R] = \sigma^2$$

- Pour tester l'**effet d'un facteur** (Facteur 1 par exemple), on utilise le même principe qu'une ANOVA à 1 facteur :

$$F_{Obs}^{F_1} = \frac{CM_{F_1}}{CM_R} \sim_{\mathcal{H}_0} \mathcal{F}_{n-I-J+1}^{I-1}$$

- ▶ Règle de décision :

- Si $F_{Obs}^{F_1} < \mathcal{F}_{n-I-J+1}^{I-1}(1-\alpha)$, on accepte \mathcal{H}_0 au seuil α .
- Si $F_{Obs}^{F_1} > \mathcal{F}_{n-I-J+1}^{I-1}(1-\alpha)$, on rejette \mathcal{H}_0 au seuil α .

Plan

- 1 Introduction de la problématique
 - Des questions “types”
 - Notion de variabilité
- 2 Modèles à effets fixes
 - Modèles à 1 facteur
 - Modèle
 - Estimation
 - Test global
 - Analyse Post-Hoc
 - Hypothèses du modèle et tests non paramétriques
 - Modèles à 2 facteurs
 - Modèles
 - Estimation
 - Test global
 - Interaction
 - Extensions

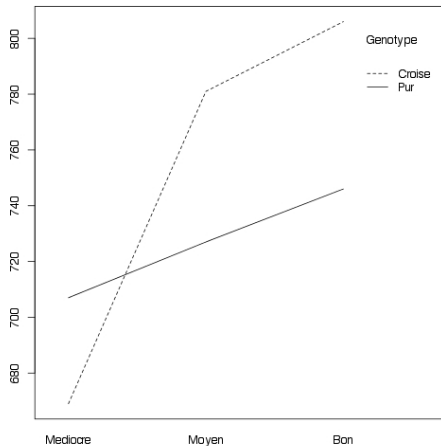
Notion d'interaction

- Dans le modèle simple on suppose que les effets des deux facteurs s'**ajoutent** simplement :
 - ▶ L'effet d'une modalité d'un facteur (α_i par exemple) est la **même** quelle que soit la modalité du second facteur.
- Cette hypothèse est vraie lorsque les effets des 2 facteurs sur la variable réponse sont indépendants.

Interaction

En analyse de variance, il y a **interaction** lorsque les effets de 2 facteurs appliqués simultanément ne peuvent pas être déduits des **moyennes** des réponses des 2 facteurs pris **séparément**.

Illustration : effet du Milieu et du Génotype sur le gain de poids



Ecriture du modèle

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \alpha\beta_{ij} + \varepsilon_{ijk}$$

- i représente le niveau du premier facteur ($i = 1, \dots, I$)
- j représente le niveau du premier facteur ($j = 1, \dots, J$)
- k représente l'indice de répétition ($k = 1, \dots, n_{ij}$)
 - ▶ Si le plan est équirépété : $n_{ij} = K$
- Hypothèses de modélisation :
 - ▶ $\forall i, j, k, \varepsilon_{ijk} \sim \mathcal{N}(0, \sigma)$
 - ▶ $\forall i' \neq i$ ou $j' \neq j$ ou $k' \neq k$: $\text{Cov}(\varepsilon_{ikj}, \varepsilon_{i'j'k'}) = 0$
- $\alpha\beta_{ij}$ correspond à l'**effet spécifique de l'interaction** des facteurs.
- Dans ce modèle le nombre de degré de liberté des résidus est donné par :

$$n - 1 - (I - 1) - (J - 1) - (I - 1)(J - 1) = \mathbf{n - IJ}$$

Estimation

- On utilise le principe des **Moindres Carrés Ordinaires**.
- Avec comme contrainte la **nullité de la somme des effets** :

$$\sum_{i=1}^I \alpha_i = 0, \quad \sum_{j=1}^J \beta_j = 0, \quad \forall i : \sum_{j=1}^J \alpha\beta_{ij} = 0 \quad \forall j : \sum_{i=1}^I \alpha\beta_{ij} = 0$$

et dans le cas **complet et équilibré**, on a

$$\hat{\mu} = Y_{...} , \quad \hat{\alpha}_i = Y_{i..} - Y_{...} , \quad \hat{\beta}_j = Y_{.j.} - Y_{...}$$

$$\hat{\alpha}\hat{\beta}_{ij} = Y_{ij.} - Y_{i..} - Y_{.j.} + Y_{...}$$

- Pour les **résidus**, on a :

$$\widehat{\sigma^2} = \frac{1}{n - IJ} \sum_{i=1}^I \sum_{j=1}^J \sum_{k=1}^K (Y_{ijk} - \widehat{Y}_{ijk})^2 = \frac{1}{n - IJ} \sum_{i=1}^I \sum_{j=1}^J \sum_{k=1}^K \varepsilon_{ijk}^2$$

La variabilité du modèle

- Dans le cas complet (toutes les combinaisons des facteurs sont testés) et équilibré (chaque combinaison est testée un même nombre de fois), on a :

$$\begin{aligned} \sum_{i,j,k} (Y_{ijk} - Y_{...})^2 &= \sum_{i,j,k} (Y_{i..} - Y_{...})^2 + \sum_{i,j,k} (Y_{.j.} - Y_{...})^2 \\ &\quad + \sum_{i,j,k} (Y_{ij.} - (Y_{i..} + Y_{.j.} - Y_{...}))^2 + \sum_{i,j,k} (Y_{ijk} - Y_{ij.})^2 \\ \text{SCE}_T &= \text{SCE}_{F_1} + \text{SCE}_{F_2} + \text{SCE}_{F_{12}} + \text{SCE}_R \end{aligned}$$

Variabilité	Somme des carrés	ddl
Facteur 1	$SCE_{F_1} = KJ \sum_i (Y_{i..} - Y_{...})^2$	$I - 1$
Facteur 2	$SCE_{F_2} = KI \sum_j (Y_{.j.} - Y_{...})^2$	$J - 1$
Interaction	$SCE_{F_{12}} = K \sum_i \sum_j (Y_{ij.} - Y_{...})^2$	$(I - 1)(J - 1)$
Résiduelle	$SCE_R = \sum_{i,j,k} (Y_{ijk} - Y_{ij.})^2$	$IJ(K - 1)$
Totale	$SCE_T = \sum_{i,j,k} (Y_{ijk} - Y_{...})^2$	$IJK - 1$

Test global

- On peut montrer que :

$$\mathbb{E}[CM_{F_1}] = \sigma^2 + \frac{JK}{I-1} \sum_{i=1}^I \alpha_i^2, \quad \mathbb{E}[CM_{F_2}] = \sigma^2 + \frac{IK}{J-1} \sum_{j=1}^J \beta_j^2,$$

$$\mathbb{E}[CM_{F_{12}}] = \sigma^2 + \frac{K}{(I-1)(J-1)} \sum_{i=1}^I \sum_{j=1}^J \alpha_i \beta_j^2, \quad \mathbb{E}[CM_R] = \sigma^2$$

- Les tests sur l'effet d'un facteur sont :

$$F_{Obs}^{F_1} = \frac{CM_{F_1}}{CM_R} \sim_{\mathcal{H}_0} \mathcal{F}_{IJ(K-1)}^{(I-1)}, \quad F_{Obs}^{F_2} = \frac{CM_{F_2}}{CM_R} \sim_{\mathcal{H}_0} \mathcal{F}_{IJ(K-1)}^{(J-1)}$$

- Pour tester l'effet d'**interaction** :

$$F_{Obs}^{F_{12}} = \frac{CM_{F_{12}}}{CM_R} \sim_{\mathcal{H}_0} \mathcal{F}_{IJ(K-1)}^{(I-1)(J-1)}$$

- Les règles de décision des tests s'appuient sur l'unilatéralité du test.

Plan

- 1 Introduction de la problématique
 - Des questions “types”
 - Notion de variabilité
- 2 Modèles à effets fixes
 - Modèles à 1 facteur
 - Modèle
 - Estimation
 - Test global
 - Analyse Post-Hoc
 - Hypothèses du modèle et tests non paramétriques
 - Modèles à 2 facteurs
 - Modèles
 - Estimation
 - Test global
 - Interaction
 - Extensions

ANOVA à $n (> 2)$ facteurs

- Les modèles d'analyse de variance se généralise facilement à un nombre **plus important** de facteurs (ANOVA à n facteurs).
 - ▶ Les principes de test des effets sont identiques
 - ▶ Le nombre de paramètres **augmente fortement** avec des interactions d'ordre élevé
 - ▶ Interprétation **très complexe** d'interaction d'ordre > 2
- On peut alors utiliser une **écriture matricielle** qui donne des résultats généraux qui permette de :
 - ▶ lier l'ANOVA à la **régression linéaire multiple** (Voir TD)
 - ▶ avoir des formules générales des estimateurs
 - ▶ traiter le cas des données **déséquilibrées**

Voir Exemple Code R

Décomposition de la variabilité (cas déséquilibré)

$$\begin{aligned}
 y_{ijk} - y_{...} &= (y_{ijk} - y_{ij.}) + (y_{i..} - y_{...}) + (y_{.j.} - y_{...}) + (y_{ij.} - y_{i..} - y_{.j.} + y_{...}) \\
 (y_{ijk} - y_{...})^2 &= (y_{ijk} - y_{ij.})^2 + (y_{i..} - y_{...})^2 + (y_{.j.} - y_{...})^2 + (y_{ij.} - y_{i..} - y_{.j.} + y_{...})^2 \\
 &\quad + \sum_{\ell}^6 DP_{\ell}(ijk) \\
 SST &= \sum_{i=1}^I \sum_{j=1}^J \sum_{k=1}^{n_{ij}} (y_{ijk} - y_{ij.})^2 + \sum_{i=1}^I n_i (y_{i..} - y_{...})^2 + \sum_{j=1}^J n_j (y_{.j.} - y_{...})^2 \\
 &\quad + \sum_{i=1}^I \sum_{j=1}^J n_{ij} (y_{ij.} - y_{i..} - y_{.j.} + y_{...})^2 \\
 &\quad + \sum_{i=1}^I \sum_{j=1}^J \sum_{k=1}^{n_{ij}} \sum_{\ell=1}^6 DP_{\ell}(ijk)
 \end{aligned}$$

Décomposition de la variabilité (cas déséquilibré) - Les doubles produits

- Etant donné que $\forall x_{ij}$ indépendant de k :

$$\sum_{i=1}^I \sum_{j=1}^J \sum_{k=1}^{n_{ij}} x_{ij} (y_{ijk} - y_{ij.}) = \sum_{i=1}^I \sum_{j=1}^J x_{ij} \left(\sum_{k=1}^{n_{ij}} (y_{ijk} - y_{ij.}) \right) = 0,$$

nous avons : $DP_1 = DP_2 = DP_3 = 0$.

- Après quelques lignes de calculs simples mais (très) fastidieux :

$$\sum_{i=1}^I \sum_{j=1}^J \sum_{k=1}^{n_{ij}} DP_4(ijk) + DP_5(ijk) + DP_6(ijk) = 2 \left(ny_{...}^2 - \sum_{i=1}^I \sum_{j=1}^J n_{ij} y_{i..} y_{.j.} \right)$$

Décomposition de la variabilité (cas déséquilibré) - Réécriture

$$\begin{aligned}
 SST &= \sum_{i=1}^I \sum_{j=1}^J \sum_{k=1}^{n_{ij}} (y_{ijk} - y_{ij.})^2 + \sum_{i=1}^I n_i (y_{i..} - y_{...})^2 + \sum_{j=1}^J n_j (y_{.j.} - y_{...})^2 \\
 &\quad + \sum_{i=1}^I \sum_{j=1}^J n_{ij} (y_{ij.} - y_{i..} - y_{.j.} + y_{...})^2 \\
 &\quad + \sum_{i=1}^I \sum_{j=1}^J \sum_{k=1}^{n_{ij}} \sum_{\ell=1}^6 DP_{\ell}(ijk) \\
 &= SSE + SSF_1 + SSF_2 + SSF_{12} + SDP
 \end{aligned}$$

- Pour un plan équilibré ($n_{ij} = n_0 \forall (ij)$) : $SDP=0$.
- Si $SDP \neq 0$, il n'est pas possible d'affecter SDP à une unique source de variation parmi :
 - ▶ F_1
 - ▶ F_2
 - ▶ F_{12}

Décomposition de la variabilité (cas déséquilibré)

- Dans un plan qui n'est pas complet-équilibré la **somme des SCE** n'est pas nécessairement égale à la **SCE_T** .
- On distingue alors plusieurs types de sortie dont :
 - ▶ Somme des carrés de **Type I** qui correspondent à la variabilité expliquée par un effet sachant que les effets **précédemment** écrits sont déjà dans le modèle.
 - ▶ Somme des carrés de **Type III** qui correspondent à la variabilité expliquée par **l'apport d'un effet sachant que tous les autres effets sont déjà présents dans le modèle**.
- L'**ordre d'inclusion** des facteurs est donc importante dans l'analyse de Type I
- Remarque : Type I et Type III sont **équivalentes** dans un plan complet-équilibré

Calcul des SC pour l'interaction

- Type I : différence des SC résiduels entre
 - ▶ $Y_{ijk} = \mu + \alpha_i + \beta_j + \varepsilon_{ijk}$
 - ▶ $Y_{ijk} = \mu + \alpha_i + \beta_j + \alpha\beta_{ij} + \varepsilon_{ijk}$
- Type III : différence des SC résiduels entre
 - ▶ $Y_{ijk} = \mu + \alpha_i + \beta_j + \varepsilon_{ijk}$
 - ▶ $Y_{ijk} = \mu + \alpha_i + \beta_j + \alpha\beta_{ij} + \varepsilon_{ijk}$

Calcul des SC pour l'interaction

- Type I : différence des SC résiduels entre
 - ▶ $Y_{ijk} = \mu + \alpha_i + \beta_j + \varepsilon_{ijk}$
 - ▶ $Y_{ijk} = \mu + \alpha_i + \beta_j + \alpha\beta_{ij} + \varepsilon_{ijk}$
- Type III : différence des SC résiduels entre
 - ▶ $Y_{ijk} = \mu + \alpha_i + \beta_j + \varepsilon_{ijk}$
 - ▶ $Y_{ijk} = \mu + \alpha_i + \beta_j + \alpha\beta_{ij} + \varepsilon_{ijk}$
- Type I et Type III donnent les mêmes résultats

Calcul des SC pour le facteur 2

- Type I : différence des SC résiduels entre

- ▶ $Y_{ijk} = \mu + \alpha_i + \varepsilon_{ijk}$

- ▶ $Y_{ijk} = \mu + \alpha_i + \beta_j + \varepsilon_{ijk}$

- Type III : différence des SC résiduels entre

- ▶ $Y_{ijk} = \mu + \alpha_i + \alpha\beta_{ij} + \varepsilon_{ijk}$

- ▶ $Y_{ijk} = \mu + \alpha_i + \beta_j + \alpha\beta_{ij} + \varepsilon_{ijk}$

Calcul des SC pour le facteur 1

- Type I : différence des SC résiduels entre

- ▶ $Y_{ijk} = \mu + \varepsilon_{ijk}$

- ▶ $Y_{ijk} = \mu + \alpha_i + \varepsilon_{ijk}$

- Type III : différence des SC résiduels entre

- ▶ $Y_{ijk} = \mu + \beta_j + \alpha\beta_{ij} + \varepsilon_{ijk}$

- ▶ $Y_{ijk} = \mu + \alpha_i + \beta_j + \alpha\beta_{ij} + \varepsilon_{ijk}$

Exemple 2 : modèle sans facteur d'interaction

- On reprend les données de Vache sans les 2 premières observations.

```
> mod1 <- lm(GMQ ~ Milieu,data=charol)
> anova(mod1)
Analysis of Variance Table

Response: GMQ
          Df Sum Sq Mean Sq F value    Pr(>F)    
Milieu      2  34537  17268.7   16.845 2.33e-05 ***
Residuals  25   25629   1025.2                
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> sum((mean(charol$GMQ)-predict(mod1))^2)
[1] 34537.5
```

```
> mod2 <- lm(GMQ ~ Milieu+Genotype,data=charol)
> anova(mod2)
Analysis of Variance Table

Response: GMQ
          Df Sum Sq Mean Sq F value    Pr(>F)    
Milieu      2  34537  17268.7   18.5835 1.331e-05 ***
Genotype    1   3327   3327.5    3.5808  0.07057 .
Residuals  24   22302    929.2                
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> sum((mean(charol$GMQ)-predict(mod2))^2)-sum((mean(charol$GMQ)-predict(mod1))^2)
[1] 3327.5
```

```
> mod1_bis <- lm(GMQ ~ Genotype,data=charol)
> anova(mod1_bis)
Analysis of Variance Table

Response: GMQ
          Df Sum Sq Mean Sq F value    Pr(>F)    
Genotype    1   1982   1982.4    0.8859 0.3553
Residuals  26   58185   2237.9
```

```
> require(car)
> Anova(mod2,type="III")
Anova Table (Type III tests)

Response: GMQ
          Sum Sq Df    F value    Pr(>F)    
(Intercept) 14964183 1 16103.5061 < 2.2e-16 ***
Milieu       35883  2   19.3073 1.006e-05 ***
Genotype      3327  1    3.5808  0.07057 .
Residuals    22302 24
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Anova Type I et Type III - SAS

```
proc glm data=charolais2
  class Milieu Genotype;
  model GMQ=Milieu Genotype/ solution ss1 ss3;
run;
```

Results Viewer - sashtml

Le Système SAS

The GLM Procedure

Dependent Variable: GMQ

Source	DDL	Somme des carrés	Moyenne quadratique	Valeur F	Pr > F
Model	3	37865.00000	12621.66667	13.58	<.0001
Error	24	22302.00000	929.25000		
Corrected Total	27	60167.00000			

r-carré	Coef de Var	Racine MSE	GMQ Moyenne
0.629332	4.150252	30.48360	734.5000

Source	DDL	Type I SS	Moyenne quadratique	Valeur F	Pr > F
Milieu	2	34537.50000	17268.75000	18.58	<.0001
Genotype	1	3327.50000	3327.50000	3.58	0.0706

Source	DDL	Type III SS	Moyenne quadratique	Valeur F	Pr > F
Milieu	2	35882.56410	17941.28205	19.31	<.0001
Genotype	1	3327.50000	3327.50000	3.58	0.0706

Source	DDL	Type III SS	Moyenne quadratique	Valeur F	Pr > F
Error	24	22302.00000	929.25000		

Calcul des SC pour le facteur 2

- Type I : différence des SC résiduels entre
 - ▶ $Y_{ijk} = \mu + \alpha_i + \varepsilon_{ijk}$
 - ▶ $Y_{ijk} = \mu + \alpha_i + \beta_j + \varepsilon_{ijk}$
- Type III : différence des SC résiduels entre
 - ▶ $Y_{ijk} = \mu + \alpha_i + \alpha\beta_{ij} + \varepsilon_{ijk}$
 - ▶ $Y_{ijk} = \mu + \alpha_i + \beta_j + \alpha\beta_{ij} + \varepsilon_{ijk}$

Calcul des SC pour le facteur 2

- Type I : différence des SC résiduels entre
 - ▶ $Y_{ijk} = \mu + \alpha_i + \varepsilon_{ijk}$
 - ▶ $Y_{ijk} = \mu + \alpha_i + \beta_j + \varepsilon_{ijk}$
- Type III : différence des SC résiduels entre
 - ▶ $Y_{ijk} = \mu + \alpha_i + \alpha\beta_{ij} + \varepsilon_{ijk}$
 - ▶ $Y_{ijk} = \mu + \alpha_i + \beta_j + \alpha\beta_{ij} + \varepsilon_{ijk}$
- Type I et III donnent les mêmes résultats

Calcul des SC pour le facteur 1

- Type I : différence des SC résiduels entre
 - ▶ $Y_{ijk} = \mu + \varepsilon_{ijk}$
 - ▶ $Y_{ijk} = \mu + \alpha_i + \varepsilon_{ijk}$
- Type III : différence des SC résiduels entre
 - ▶ $Y_{ijk} = \mu + \beta_j + \varepsilon_{ijk}$
 - ▶ $Y_{ijk} = \mu + \alpha_i + \beta_j + \varepsilon_{ijk}$
 - ▶ $SS_{F1} = 58185 - 22302 = 35883$

Démarche à suivre

- Du point de vue statistique, il convient de respecter l'ordre du schéma d'analyse suivant :
 - ① Pertinence du modèle global
 - Modèle global vs. modèle constant
 - ② Effet de chaque facteur
 - On commence par les effets d'interaction
 - ③ Comparaison des modalités pour un facteur donné
 - Comparaisons multiples
 - Test sur les contrastes