# ANALYSE DE VARIANCE

## Plan du Chapitre 4





FPSTAT INRA 1997 IV-Analyse de Variance-1

Analyse de variance

test d'une structuration des données

à l'aide de variables explicatives

de type qualitatif = facteurs



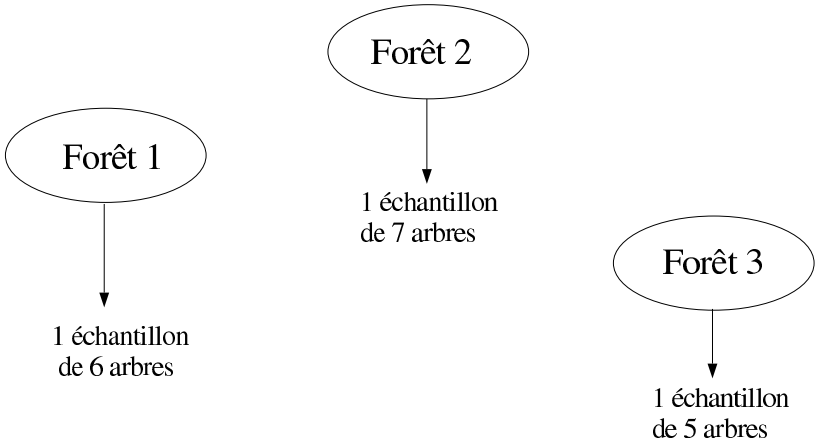
Lors d'une analyse de variance, on se pose la question de la structuration des données.

On teste un type de structuration par rapport à une structuration plus simple ; les structurations envisagées sont établies à l'aide de variables de type qualitatif appelées des facteurs. Tester une structuration contre une structuration plus simple, revient à tester un modèle contre un sous modèle emboîté. Deux possiblités existent pour ce test :

— le test est non significatif : on s'arrête, la structuration envisagée n'apporte rien par rapport à la structuration plus simple.

— le test est significatif : il faut comparer les niveaux des facteurs de structuration par une comparaison multiple de moyennes.

# Analyse de variance à un facteur




| Forêt 1 | Forêt 2 | Forêt 3 |
|---|---|---|
| 23.4 | 18.9 | 22.5 |
| 24.4 | 21.1 | 22.9 |
| 24.6 | 21.1 | 23.7 |
| 24.9 | 22.1 | 24.0 |
| 25.0 | 22.5 | 24.0 |
| 26.2 | 23.5 | |
| | 24.5 | |



**Exemple : Étude des forêts à travers la hauteur d'arbres**

Présenter la notion de structuration : au lieu d'avoir 18 arbres pris au hasard, on a 6+7+5 arbres pris dans 3 forêts différentes.

Le facteur, la variable qualitative qui structure les données (permet de faire des lots différents dans les données) est la forêt. Il présente 3 niveaux, encore appelés modalités.

La structuration apparaît dans le tableau des données sous forme de 3 colonnes, chacune correspondant à 1 niveau du facteur.

La variable étudiée dans chaque forêt est la hauteur des arbres.

## Le vecteur des données

$$Y_n = \begin{pmatrix} Y_1 = 23.4 \\ Y_2 = 24.4 \\ Y_3 = 24.6 \\ Y_4 = 24.9 \\ Y_5 = 25.0 \\ Y_6 = 26.2 \\ Y_7 = 18.9 \\ Y_8 = 21.1 \\ Y_9 = 21.1 \\ Y_{10} = 22.1 \\ Y_{11} = 22.5 \\ Y_{12} = 23.5 \\ Y_{13} = 24.5 \\ Y_{14} = 22.5 \\ Y_{15} = 22.9 \\ Y_{16} = 23.7 \\ Y_{17} = 24.0 \\ Y_{18} = 24.0 \end{pmatrix} \rightarrow Y_{ir} = \begin{pmatrix} Y_{11} = 23.4 \\ Y_{12} = 24.4 \\ Y_{13} = 24.6 \\ Y_{14} = 24.9 \\ Y_{15} = 25.0 \\ Y_{16} = 26.2 \\ Y_{21} = 18.9 \\ Y_{22} = 21.1 \\ Y_{23} = 21.1 \\ Y_{24} = 22.1 \\ Y_{25} = 22.5 \\ Y_{26} = 23.5 \\ Y_{27} = 24.5 \\ Y_{31} = 22.5 \\ Y_{32} = 22.9 \\ Y_{33} = 23.7 \\ Y_{34} = 24.0 \\ Y_{35} = 24.0 \end{pmatrix}$$



Dans les chapitres précédents, les données étaient présentées sous forme d'un vecteur.

En maintenant un vecteur, on fait apparaître la structuration à l'aide d'indices.

— Un premier indice correspondant au niveau considéré du facteur i = 1, 2, 3.

— Un deuxième indice correspondant à l'individu échantillonné dans chaque forêt.

A titre de petit exercice on peut demander aux stagiaires la valeur de $Y_{14}$, $Y_{32}$ $\cdots$

## Question posée

La structuration en forêts explique t-elle
la variabilité des données ?

Les forêts sont elles différentes ?

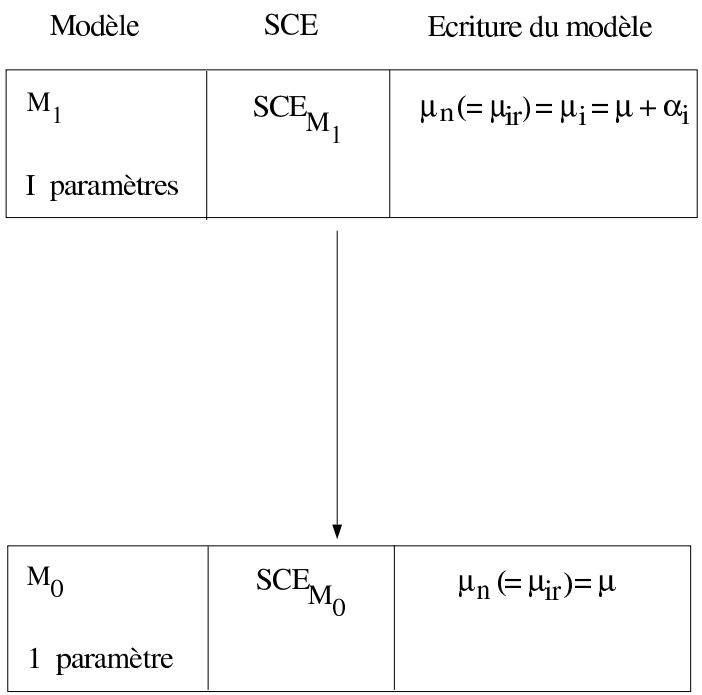
Si oui : quelles forêts sont différentes ?



Le transparent reprend les deux notions présentées en 1 :

— le test de 2 modèles

— la comparaison multiple de moyennes.






Le modèle $M_1$ est testé par l'analyse de variance contre $M_0$, qui est un sous−modèle emboîté dans $M_1$ (plus simple que $M_1$).

La formulation utilisée présente des modèles de l'espérance de $\mu_{ir}$.

Dans le cas d'une analyse à un facteur, deux modèles seulement peuvent être comparés :

— $M_1$, où $\alpha_i$ traduit l'effet du facteur (forêt, pour l'exemple). L'existence de différences entre forêts est testée

— $M_0$, il n'y a pas de différences entre les forêts ; en espérance tous les individus sont identiques.

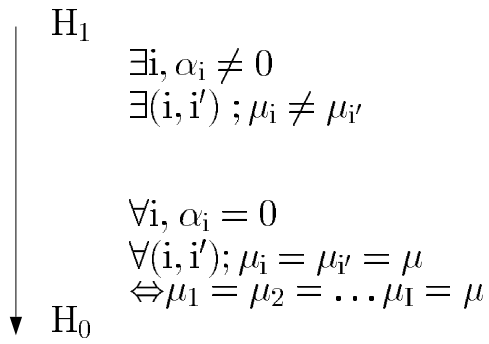
Chaque modèle se caractérise par son nombre de paramètres (I/1) et sa somme de carrés d'écarts résiduelle $(SCE_{M_1}/SCE_{M_0})$.

Le modèle $M_1$ est écrit sous deux formes équivalentes

$\mu_n = \mu_i$ : forme irréductible

$\mu_n = \mu + \alpha_i$ : forme réductible

## Hypothèses testées






*Ce transparent peut être superposé au précédent*

Détail sur les hypothèses. Traduire les symboles dans le cas de l'exemple forêt :

$H_1$ : correspondant au modèle $M_1$ ; il existe au moins une forêt différente des autres ; il y a au moins un $\alpha_i$ non nul.

$H_0$ : toutes les forêts sont identiques ; il n'y a pas d'effet forêt, les $\alpha_i$ sont nuls.

FPSTAT INRA 1997 IV-Analyse de Variance-7

## Commandes et fichier sous SAS

```
data trv ;
infile 'foret3' ;
input foret hauteur ;
run ;
proc glm ;
class foret ;
model hauteur = foret ;
run ;
```

Fichier des données

```
1  23.4
1  24.4
1  24.6
1  24.9
1  25.0
1  26.2
2  18.9
2  21.1
2  21.1
2  22.1
2  22.5
2  23.5
2  24.5
3  22.5
3  22.9
3  23.7
3  24.0
3  24.0
```


FPSTAT INRA 1997 IV-Analyse de Variance-8

## Sorties SAS

General Linear Models Procedure

Class Level Information

| Class | Levels | Values |
|---|---|---|
| Foret | 3 | 1 2 3 |

Number of observations in data set = 18

General Linear Models Procedure

Dependent Variable : HAUTEUR

| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| Model | 2 | 25.30930159 | 12.65465079 | 7.30 | 0.0061 |
| Error | 15 | 26.00014286 | 1.73334286 | | |
| Corrected Total | 17 | 51.30944444 | | | |

| R-Square | C.V. | Root MSE | HAUTEUR Mean |
|---|---|---|---|
| 0.493268 | 5.651840 | 1.316565 | 23.29444 |

General Linear Models Procedure

Dependent Variable : HAUTEUR

| Source | DF | Type I SS | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| FORET | 2 | 25.30930159 | 12.65465079 | 7.30 | 0.0061 |

| Source | DF | Type III SS | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| FORET | 2 | 25.30930159 | 12.65465079 | 7.30 | 0.0061 |



La sortie SAS se compose de deux tableaux correspondant aux deux étapes de l'analyse:

— un tableau "modèle" qui teste le modèle le plus complexe contre le modèle le plus simple. C'est la première étape, on s'intéresse au F du modèle, on recherche la présence d'une structuration expliquant la variabilité des données. Si ce test de F est significatif, on s'intéresse aux facteurs qui structurent significativement les données.

— on lit alors le tableau "facteur"

## Tableau "Modèle"

General Linear Models Procedure

Class Level Information

| Class | Levels | Values |
|---|---|---|
| Foret | 3 | 1 2 3 |

Number of observations in data set = 18

General Linear Models Procedure

Dependent Variable : HAUTEUR

| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| Model | 2 | 25.30930159 | 12.65465079 | 7.30 | 0.0061 |
| Error | 15 | 26.00014286 | 1.73334286 | | |
| Corrected Total | 17 | 51.30944444 | | | |

| R-Square | C.V. | Root MSE | HAUTEUR Mean |
|---|---|---|---|
| 0.493268 | 5.651840 | 1.316565 | 23.29444 |

| | | | | |
|---|---|---|---|---|
| Model | I-1 | $SCE_{M_0} - SCE_{M_1}$ | $\frac{SCE_{M_0} - SCE_{M_1}}{I-1}$ | $\frac{\left(SCE_{M_0} - SCE_{M_1}\right) / I-1}{SCE_{M_1} / N-I}$ |
| Error | N-I | $SCE_{M_1}$ | $\frac{SCE_{M_1}}{N-I}$ | |
| Corrected Total | N-1 | $SCE_{M_0}$ | | |

| R-Square | C.V. | Root MSE | HAUTEUR Mean |
|---|---|---|---|
| $\frac{SCE_{M_0} - SCE_{M_1}}{SCE_{M_0}}$ | 100x(Root MSE)/Hauteur Mean | $\sqrt{\frac{SCE_{M_1}}{N-I}}$ | Y.. |



Les différents éléments du tableau "modèle" sont exploités en se référant aux écritures utilisées dans les chapitres précédents.

FPSTAT INRA 1997 IV-Analyse de Variance-10

## Tableau "Facteurs"

General Linear Models Procedure

Dependent Variable : HAUTEUR

| Source | DF | Type I SS | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| FORET | 2 | 25.30930159 | 12.65465079 | 7.30 | 0.0061 |

| Source | DF | Type III SS | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| FORET | 2 | 25.30930159 | 12.65465079 | 7.30 | 0.0061 |

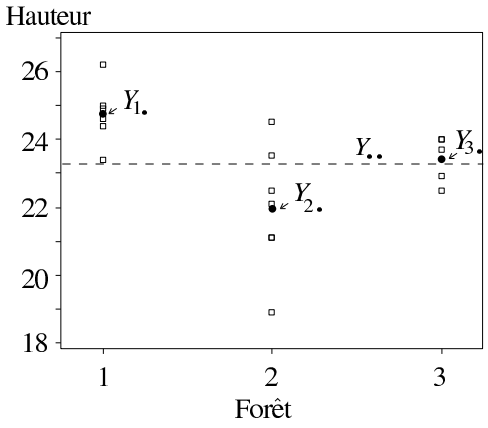
| | | | |
|---|---|---|---|
| FORET | I-1 | $SCE_{M_0} - SCE_{M_1}$ | $\left(\frac{SCE_{M_0} - SCE_{M_1}}{SCE_{M_1} / N\text{-}I}\right) / I\text{-}1$ |



Dans le cas d'une analyse de variance à un facteur, deux modèles seulement peuvent être testés. Aussi les éléments du tableau "facteurs" correspondent-ils à ceux du test du modèle complexe $M_1$ dans le tableau "modèle", puisque ce modèle $M_1$ correspond à la prise en compte du facteur étudié. Tester l'effet du facteur correspond dans ce cas à tester le modèle le plus complet.

Dans ce tableau apparaissent des sommes de carrés de type I et de type III qui sont égales. Au cours du chapitre "analyse de variance à deux facteurs" nous expliquerons à quoi correspondent ces types I et III ; par défaut ces deux types de sommes de carrés sont calculées par la procédure GLM de SAS.

FPSTAT INRA 1997 IV-Analyse de Variance-11




$$SCE_{totale} = SCE_{facteur} + SCE_{résiduelle}$$

$$SCE_{totale} = \sum_{n} (Y_n - Y..)^2$$

$$SCE_{facteur} = \sum_{i} \sum_{r} (Y_{i.} - Y..)^2 = \text{variabilité inter} - \text{classes}$$

$$SCE_{résiduelle} = \sum_{i} \sum_{r} (Y_{ir} - Y_{i.})^2 = \text{variabilité intra} - \text{classe}$$



**Objectif : relier l'approche classique (effet facteur ; variabilité inter− classe/intra− classe) par rapport à l'approche modèle/sous modèle.**

On utilise le graphique pour visualiser la variabilité considérée par les différentes SCE :

— $SCE_{totale}$ = somme des carrés des écarts de chaque carré à la ligne pointillée

— $SCE_{facteur}$ = somme des carrés des écarts des ronds (moyenne de chaque facteur) à la ligne = variabilité inter−classes

— $SCE_{résiduelle}$ = somme des carrés des écarts des carrés au rond / classe = variabilité intra−classe

Quelle que soit la somme de carrés $SCE_{totale}, SCE_{facteur}, SCE_{résiduelle,}$ la somme porte sur tous les individus observés.

$$SCE_{totale} = SCE_{M_0} : \text{variabilité totale}$$

$$SCE_{facteur} = SCE_{M_0} - SCE_{M_1} : \text{variabilité inter} - \text{classe}$$

$$SCE_{résiduelle} = SCE_{M_1} : \text{variabilité intra} - \text{classe}$$

$$\text{d}'\text{où}$$

$$F = \frac{SCE_{facteur}/\ I-1}{SCE_{résiduelle}/\ N-I} = \frac{(SCE_{M_0} - SCE_{M_1})/\ I-1}{SCE_{M_1}/\ N-I}$$



La $SCE_{totale}$ = le calcul de l'écart des données à une valeur à laquelle on les résume, la moyenne générale. C'est la définition de la somme des carrés résiduelle du modèle $M_0$, le modèle le plus simple.

La $SCE_{facteur}$ = le calcul de l'écart entre un résumé pour certaines valeurs, la moyenne de la classe à laquelle appartiennent ces valeurs, à un autre résumé pour ces valeurs, la moyenne générale. Chaque résumé correspond à un modèle différent : la moyenne des classes correspond au modèle prenant en compte le facteur, modèle $M_1$, la moyenne générale correspond au modèle $M_0$. En calculant l'écart entre les deux résumés on s'intéresse à l'apport du modèle $M_1$ par rapport au modèle $M_0$.

La $SCE_{résiduelle}$ = calcul de l'écart entre la valeur observée et un résumé par le modèle. Par définition c'est la somme des carrés des écarts résiduelle de $M_1$

FPSTAT INRA 1997 IV-Analyse de Variance-13

Test du modèle $M_1$/ modèle $M_0$

Estimation

Démonstration (facultative) à utiliser si les stagiaires veulent aller plus loin

$$SCE_{M_0} = SCE_{M_0} - SCE_{M_1} + SCE_{M_1}$$

• $M_0 \leftrightarrow \mu_n = \mu$ avec $Y_{..} = \hat{\mu} = \hat{\mu}_{n_{M_0}} = \hat{Y}_{n_{M_0}}$

$$SCE_{totale} = \sum_{i=1}^{N} (Y_n - Y_{..})^2 = \sum_{i=1}^{N} \left(Y_n - \hat{Y}_{n_{M_0}}\right)^2$$

donc $SCE_{totale} = SCE_{M_0}$

• $M_1 \leftrightarrow \mu_n = \mu_i$ avec $Y_{i.} = \hat{\mu}_i = \hat{\mu}_{n_{M_1}} = \hat{Y}_{n_{M_1}}$

$$\begin{aligned} SCE_{facteur} &= \sum_i \sum_r (Y_{i.} - Y_{..})^2 \\ &= \sum_i \sum_r [(Y_n - Y_{..}) - (Y_n - Y_{i.})]^2 \\ &= \sum_i \sum_r \left[\left(Y_n - \hat{Y}_{n_{M_0}}\right) - \left(Y_n - \hat{Y}_{n_{M_1}}\right)\right]^2 \\ &= SCE_{M_0} - SCE_{M_1} \end{aligned}$$

$$\begin{aligned} SCE_{résiduelle} &= \sum_i \sum_r (Y_{ir} - Y_{i.})^2 \\ &= \sum_i \sum_r \left(Y_n - \hat{Y}_{n_{M_1}}\right)^2 \\ &= SCE_{M_1} \end{aligned}$$

Complément de démonstration fourni



Démonstration que $SCE_{facteur} = SCE_{M_0} - SCE_{M_1}$

$$\begin{aligned} SCE_{facteur} &= \sum_i \sum_r (Y_{i.} - Y_{..})^2 \\ &= \sum_i \sum_r [(Y_n - Y_{..}) - (Y_n - Y_{i.})]^2 \\ &= \sum_i \sum_r \left[\left(Y_n - \hat{Y}_{n_{M_0}}\right) - \left(Y_n - \hat{Y}_{n_{M_1}}\right)\right]^2 \end{aligned}$$

$n_i$ : effectif de la classe i

on pose $a = Y_n - \hat{Y}_{n_{M_0}}$ $b = Y_n - \hat{Y}_{n_{M_1}}$ pour simplifier

$$\begin{aligned} &= \sum_i \sum_r (a - b)^2 \\ &= \sum_i \sum_r \left(a^2 - 2ab + b^2\right) \\ &= \sum_i \sum_r a^2 + \sum_i \sum_r b^2 - 2 \sum_i \sum_r ba \end{aligned}$$

on pose $a = Y_n - \hat{Y}_{n_{M_0}} = \left(Y_n - \hat{Y}_{n_{M_1}}\right) + \left(\hat{Y}_{n_{M_1}} - \hat{Y}_{n_{M_0}}\right) = b + c$

avec $c = \hat{Y}_{n_{M_1}} - \hat{Y}_{n_{M_0}}$

donc l'égalité devient

$$\begin{aligned} &= \sum_i \sum_r a^2 + \sum_i \sum_r b^2 - 2 \sum_i \sum_r b^2 - 2 \sum_i \sum_r bc \\ &= \sum_i \sum_r a^2 - \sum_i \sum_r b^2 - 2 \sum_i \sum_r bc \end{aligned}$$

Calcul de : $\sum_i \sum_r bc$

$$= \sum_i \sum_r \left[ \left( Y_n - \widehat{Y}_{n_{M_1}} \right) \left( \widehat{Y}_{n_{M_1}} - \widehat{Y}_{n_{M_0}} \right) \right]$$

$$= \sum_i \sum_r \left( Y_n \widehat{Y}_{n_{M_1}} - \widehat{Y}^2_{n_{M_1}} - Y_n \widehat{Y}_{n_{M_0}} + \widehat{Y}_{n_{M_1}} \widehat{Y}_{n_{M_0}} \right)$$

$$= \underbrace{\sum_i \sum_r Y_n \widehat{Y}_{n_{M_1}}}_{(1)} - \underbrace{\sum_i \sum_r \widehat{Y}^2_{n_{M_1}}}_{(2)} - \underbrace{\sum_i \sum_r Y_n \widehat{Y}_{n_{M_0}}}_{(3)} + \underbrace{\sum_i \sum_r \widehat{Y}_{n_{M_1}} \widehat{Y}_{n_{M_0}}}_{(4)}$$

Calcul de (1), (2), (3), (4)

$$(1) \sum_i \sum_r Y_n \widehat{Y}_{n_{M_1}} = \sum_{ir} Y_{ir} Y_{i.} = \sum_i Y_{i.} \sum_{r=1}^{n_i} Y_{ir} = \sum_i n_i Y^2_{i.} \text{ car} \begin{cases} \widehat{Y}_{n_{M_1}} = Y_{i.} = \frac{\sum_r Y_{ir}}{n_i} \end{cases}$$

$$(2) \sum_i \sum_r \widehat{Y}^2_{n_{M_1}} = \sum_i n_i Y^2_{i.} \qquad (2) = (1)$$

$$(3) \sum_i \sum_r Y_n \widehat{Y}_{n_{M_0}} = \widehat{Y}_{n_{M_0}} \sum_i \sum_r Y_{ir} = N \widehat{Y}^2_{n_{M_0}} \text{ car} \begin{cases} \widehat{Y}_{n_{M_0}} = Y_{..} = \frac{\sum_r \sum_i Y_{ir}}{N} \\ \sum_i n_i = N \end{cases}$$

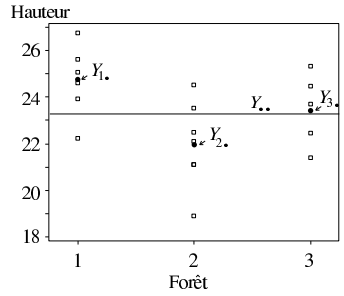
$$(4) \sum_i \sum_r \widehat{Y}_{n_{M_1}} \widehat{Y}_{n_{M_0}} = \widehat{Y}_{n_{M_0}} \sum_i \sum_r \widehat{Y}_{n_{M_1}} = \widehat{Y}_{n_{M_0}} \sum_i n_i Y_{i.} = \widehat{Y}_{n_{M_0}} \sum_i \sum_r Y_{ir} = N \widehat{Y}^2_{n_{M_0}}$$

donc $\sum_i \sum_r bc = (1) - (1) - (3) + (3) = 0$

l'égalité devient $\sum_i \sum_r a^2 - \sum_i \sum_r b^2$, en revenant à l'expression de a et b

(cqfd !) $SCE_{facteur} = \sum_i \sum_r \left( Y_n - \widehat{Y}_{n_{M_0}} \right)^2 - \sum_i \sum_r \left( Y_n - \widehat{Y}_{n_{M_1}} \right)^2 = SCE_{M_0} - SCE_{M_1}$


FPSTAT INRA 1997 IV-Analyse de Variance-13



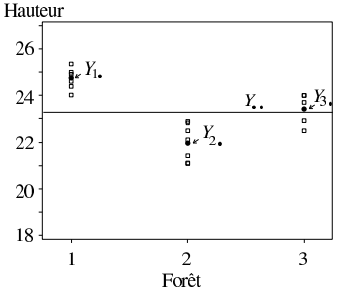

$Y_{1.} = 24.75$

$Y_{2.} = 21.96$

$Y_{3.} = 23.42$

$(Y_{1.} + Y_{2.} + Y_{3.})/3 = 23.38$

$H_0$ ou $H_1$ ?




$Y_{1.} = 24.75$

$Y_{2.} = 21.96$

$Y_{3.} = 23.42$

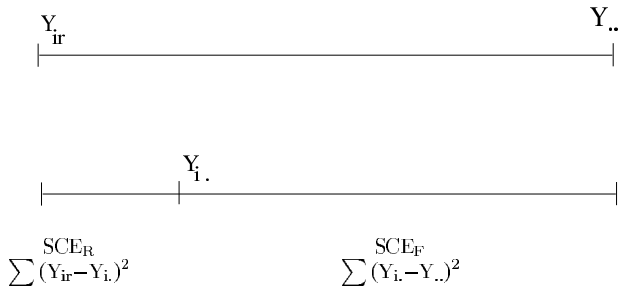
$(Y_{1.} + Y_{2.} + Y_{3.})/3 = 23.38$

$H_0$ ou $H_1$ ?



La valeur d'un test F et sa probabilité dépendent de la dispersion des valeurs à l'intérieur de chaque classe, puisque le test F est le rapport de la variabilité inter−classes à la variabilité intra−classe.

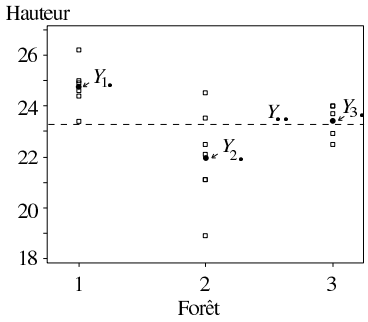
$$SCE_T : \sum (Y_{ir} - Y_{..})^2$$




$$SCE_T = SCE_F + SCE_R$$

si $SCE_F >> SCE_R \Rightarrow H_1$

si $SCE_F \simeq SCE_R \Rightarrow H_0$






$SCE_T = SCE_{totale}$
$SCE_F = SCE_{facteur}$
$SCE_R = SCE_{résiduelle}$

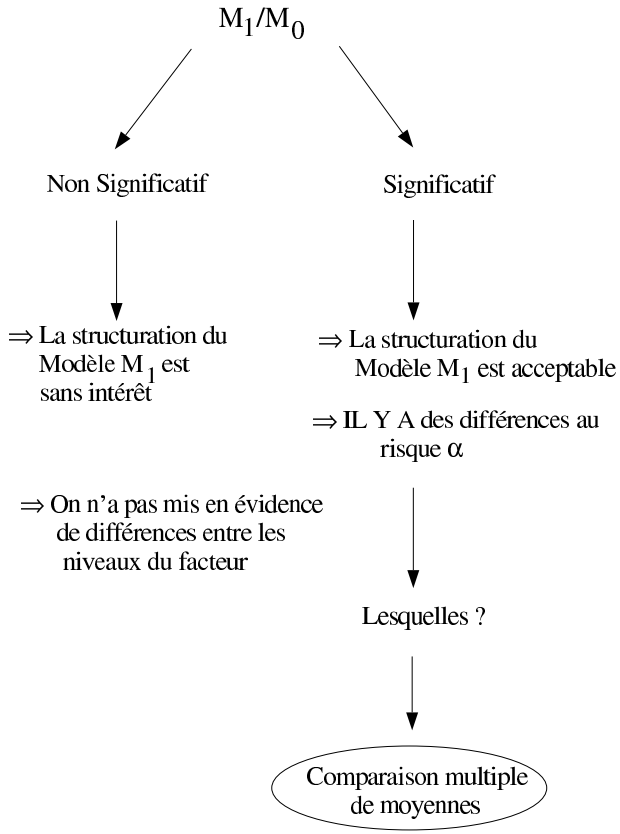
— Une variabilité intra−classe trop grande ne permet pas la mise en évidence de différences (d'une structuration) $\Rightarrow H_0$

— Une variabilité intra−classe faible facilite la mise en évidence de différences (d'une structuration) $\Rightarrow H_1$

FPSTAT INRA 1997 IV-Analyse de Variance-15

## Analyse de Variance à un facteur
## Synthèse

TEST F modèle / ss modèle





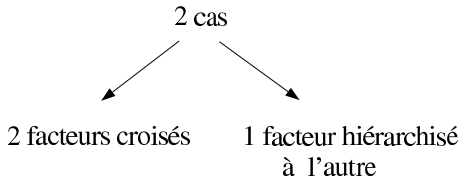

# Attention :

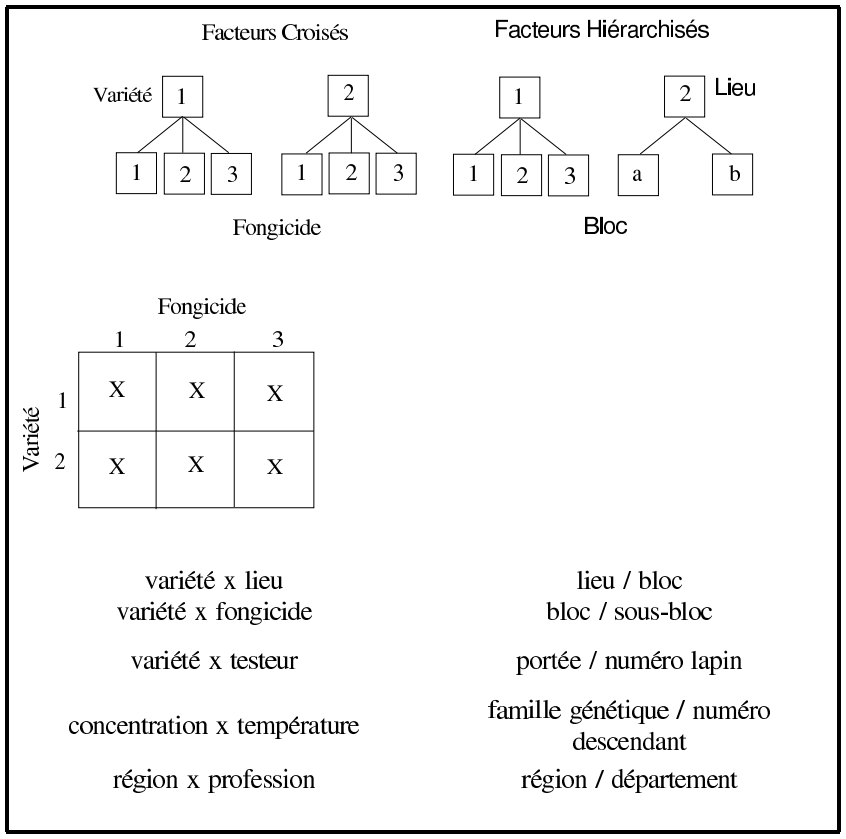
l'acceptation de $H_0$ signifie que l'on n'a pas mis en évidence de différence.

Ne connaissant pas $\beta$ (risque de deuxième espèce Prob($H_0/H_1$)), il faut rester prudent dans la formulation.

Par contre pour $H_1$, on connaît $\alpha$, on peut être plus affirmatif.

# Analyse de Variance à deux facteurs






Lorsque l'on considère 2 facteurs croisés, tous les niveaux du facteur 2 sont combinés avec les niveaux du facteur 1. L'étude de ces combinaisons est possible expérimentalement.

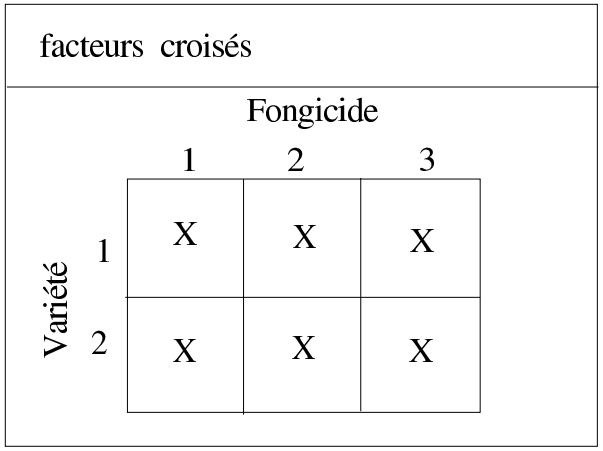
Lorsque l'on considère 1 facteur hiérarchisé à l'autre, cela signifie que les niveaux du deuxième facteur sont différents selon le niveau considéré du premier facteur. De tels dispositifs sont généralement mis en place parce qu'expérimentalement il est impossible de combiner les niveaux des 2 facteurs.

FPSTAT INRA 1997 IV-Analyse de Variance-17

# Analyse de Variance à deux facteurs croisés

Facteur A, I niveaux

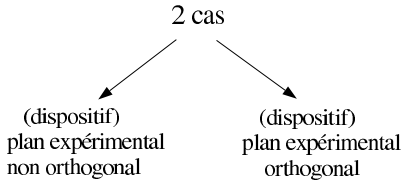
Facteur B, J niveaux




les IJ combinaisons des niveaux des facteurs A et B peuvent être étudiées



FPSTAT INRA 1997 IV-Analyse de Variance-18




Étude du nombre de jours avant germination pour des variétés de carottes

| $Y_{ijk}$ | Variété 1 | Variété 2 | Variété 3 |
|---|---|---|---|
| Sol 1 | 6 10 11 | 13 15 | 14 22 |
| Sol 2 | 12 19 15 18 | 31 | 18 9 12 |

ex : dispositif déséquilibré

Étude de la teneur en huile de populations de tournesol

| Origine | AFRIQUE | HONGRIE | MAROC |
|---|---|---|---|
| testeur 1 | 43.54 45.30 | 44.25 42.55 | 47.28 49.40 |
| testeur 2 | 47.21 47.73 | 44.34 46.49 | 47.75 49.47 |

ex : dispositif équirépété



Deux cas peuvent se présenter.

L'exemple de dispositif expérimental orthogonal présenté (équirépété) est le plus couramment rencontré. Cependant des dispositifs orthogonaux peuvent être non équirépétés ; ils satisfont, alors, à la condition suivante :

nb répétitions proportionnel : $\begin{cases} \text{d'une ligne à l'autre} \\ \text{d'une colonne à l'autre} \end{cases}$

c'est à dire $n_{ij} = \frac{n_{i+} n_{+j}}{N}$

$$n_{i+} = \sum_j n_{ij}$$
$$n_{+j} = \sum_i n_{ij}$$
$$N = \sum_{i,j} n_{ij}$$

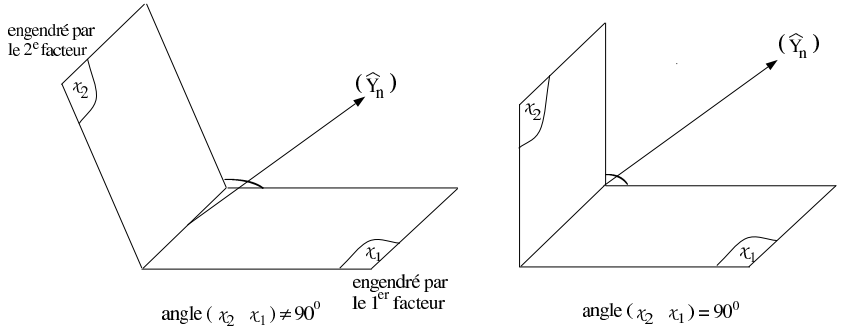
Exemple de dispositifs orthogonaux déséquilibrés

| | | |
|---|---|---|
| $n_{11} = 2$ | $n_{12} = 2$ | $n_{13} = 2$ |
| $n_{21} = 5$ | $n_{22} = 5$ | $n_{33} = 5$ |

| | | |
|---|---|---|
| $n_{11} = 2$ | $n_{12} = 3$ | $n_{13} = 7$ |
| $n_{21} = 2$ | $n_{22} = 3$ | $n_{23} = 7$ |

Les plans orthogonaux qui vont être présentés sont équirépétés

## Illustration géométrique






*Ce transparent est facultatif.*

Il présente une représentation dans l'espace de la notion d'orthogonalité.

FPSTAT INRA 1997 IV-Analyse de Variance-20

# Analyse de Variance à deux facteurs croisés cas orthogonal

**Exemple d'Applications de l'Analyse de la Variance dans le cas équirépété**

Etude de la teneur en huile de populations de tournesol

Données expérimentales

| Origine | AFRIQUE | HONGRIE | MAROC |
|---|---|---|---|
| testeur 1 | 43.54<br>45.30 | 44.25<br>42.55 | 47.28<br>49.40 |
| testeur 2 | 47.21<br>47.73 | 44.34<br>46.49 | 47.75<br>49.47 |

Facteur A = testeur, I = 2 niveaux

Facteur B = origine, J = 3 niveaux

répétitions par combinaison AB, r = 2

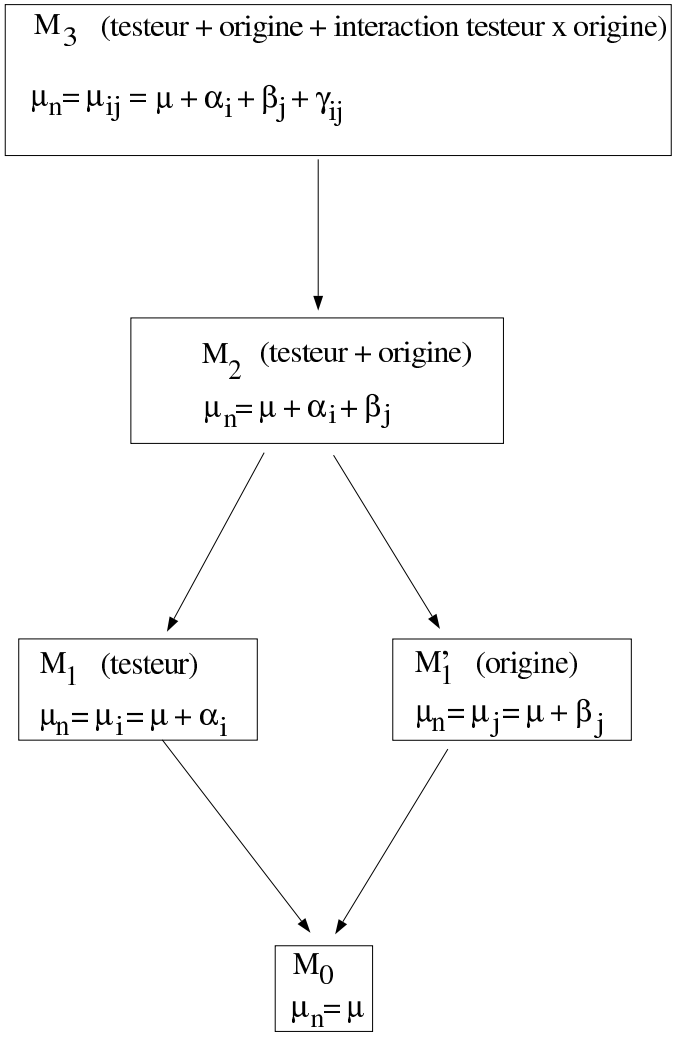


On dispose de 3 origines géographiques.

On croise des plantes de chaque population avec 2 variétés appelées testeur 1 et testeur 2, pour connaître la valeur en croisement de chaque population.

Le caractère étudié, la teneur en huile, est mesuré sur les hybrides issus de chaque croisement.

FPSTAT INRA 1997 IV-Analyse de Variance-21






Le dispositif étant présenté, il convient de s'intéresser aux différents modèles correspondant aux structurations possibles :

— modèle $M_3$ : la structuration est due à la combinaison des 2 facteurs, c'est à dire leurs effets additifs et leur interaction.

— modèle $M_2$ : c'est le modèle additif ; les 2 facteurs participent à la structuration mais n'interagissent pas.

— $M_1, M_1'$ : un seul facteur structure les données.

On obtient 5 modèles, 2 modèles ne sont pas emboîtés l'un dans l'autre.

On peut demander aux stagiaires quels sont les modèles emboîtés, ceux qui ne le sont pas.

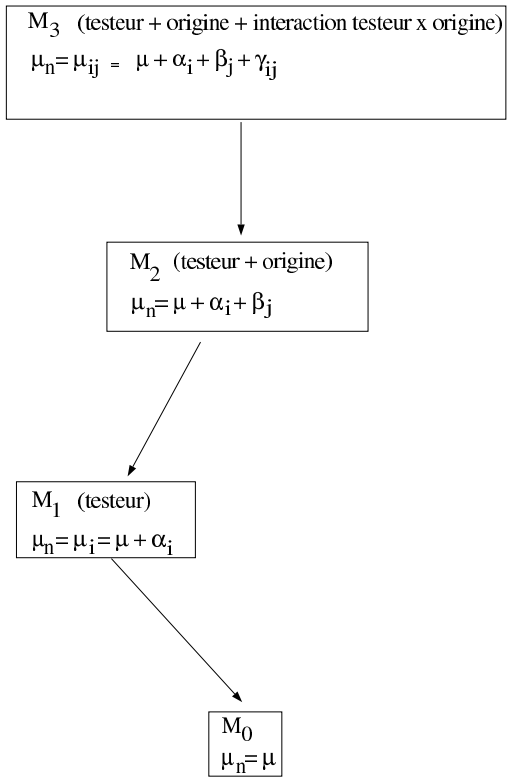
Pour chaque modèle, les formes irréductibles et réductibles sont données (sauf pour le modèle additif, pour lequel la forme irréductible n'existe pas de manière simple).

**Attention** : Si le nombre de répétitions est égale à 1, l'interaction n'est pas estimable. Le modèle $M_3$ ne peut pas être étudié.

## Fichier des données et Commandes SAS

```
1 1 43.54
1 1 45.30
1 2 44.25
1 2 42.55
1 3 47.28
1 3 49.40
2 1 47.21
2 1 47.73
2 2 44.34
2 2 46.49
2 3 47.75
2 3 49.47
```

```
data trv ;
infile 'tourn' ;
input testeur origine huile ;
run ;
proc glm ;
class testeur origine ;
model huile = testeur origine testeur*origine ;
run ;
```






Le modèle écrit selon cet ordre correspond à la succession des modèles $M_0 \rightarrow M_1 \rightarrow M_2 \rightarrow M_3$

FPSTAT INRA 1997 IV-Analyse de Variance-23

# Sortie SAS

General Linear Models Procedure

Class Level Information

| Class | Levels | Values |
|---|---|---|
| Testeur | 2 | 1 2 |
| Origine | 3 | 1 2 3 |

Number of observations in data set = 12

General Linear Models Procedure

Dependent Variable : HUILE

| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| Model | 5 | 47.18144167 | 9.43628833 | 6.18 | 0.0233 |
| Error | 6 | 9.16665000 | 1.52777500 | | |
| Corrected Total | 11 | 56.34809167 | | | |

| R-Square | C.V. | Root MSE | HUILE Mean |
|---|---|---|---|
| 0.837321 | 2.671010 | 1.236032 | 46.27583 |

General Linear Models Procedure

Dependent Variable : HUILE

| Source | DF | Type I SS | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| TESTEUR | 1 | 9.48740833 | 9.48740833 | 6.21 | 0.0470 |
| ORIGINE | 2 | 33.74581667 | 16.87290833 | 11.04 | 0.0097 |
| TESTEUR*ORIGINE | 2 | 3.94821667 | 1.97410833 | 1.29 | 0.3415 |

| Source | DF | Type III SS | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| TESTEUR | 1 | 9.48740833 | 9.48740833 | 6.21 | 0.0470 |
| ORIGINE | 2 | 33.74581667 | 16.87290833 | 11.04 | 0.0097 |
| TESTEUR*ORIGINE | 2 | 3.94821667 | 1.97410833 | 1.29 | 0.3415 |



FPSTAT INRA 1997 IV-Analyse de Variance-24