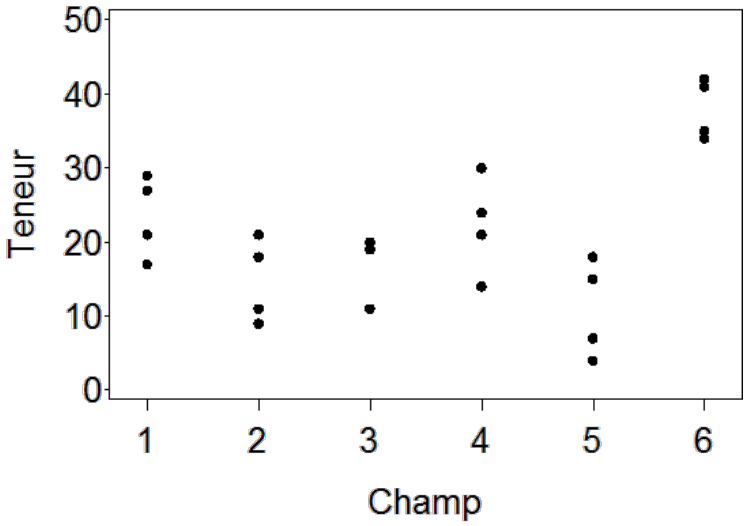
CH 01. Modèles à composantes de variance

Analyse de la variance
à un critère de classification aléatoire



Objectif et plan

Objectif pédagogique

Introduire le concept de **facteur aléatoire** et ce que ce type de facteur implique au niveau de l'inférence statistique

Objectif du modèle d'ANOVA1 aléatoire

Expliquer une variable quantitative Y à partir d'un facteur **catégoriel** dont les niveaux ont été tirés au hasard dans une population de niveaux possibles

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

Plan du chapitre

- Facteurs fixes et aléatoires
- Les hypothèses du modèle d'ANOVA 1 aléatoire
- Le modèle d'ANOVA 1 sous la forme d'une modèle linéaire
- Estimation des paramètres par la méthode des espérances des carrés moyens
- Réécriture du modèle
- Estimation par les méthodes du maximum de vraisemblance et REML
- Inférence et tests d'hypothèse

Facteurs fixes <> facteurs aléatoires

Les facteurs considérés dans le cours LBIRA2110 étaient des **facteurs fixes**.

- Les niveaux du facteur sont connus et peuvent être reproduits (ex: la variété Tripoli, une dose de 120 unités d'azote…)
- Quelques exemples:

  Un effet variété (génétique), un effet dose (fertilisation), un effet labour (phytotechnie)…
- On considère que l'effet d'un niveau particulier d'un facteur est une constante ($\mu_i = \mu + \alpha_i$), et notre objectif est de l'estimer.
- Les questions posées concernent l'effet des différents niveaux des facteurs.

Facteurs fixes <> facteurs aléatoires

On parle de **facteur aléatoire** lorsque les niveaux d'un facteur sont un échantillon aléatoire d'une population de niveaux possibles

- En génétique, le génotype peut être considéré comme un tirage dans un large ensemble de génotypes (cf cours de Génétique des populations)
- Dans un essai agronomique: un facteur « bloc », donc les niveaux sont des sous-parcelles dans un champ, ou des champs dans une région
- Dans un laboratoire, un facteur « date » dont chaque niveau est un jour au cours duquel différents échantillons ont été traités (reproductibilité)

Quelques particularités des facteurs aléatoires:

- Les niveaux du facteur sont **a priori** inconnus et ne peuvent pas être reproduits
  Ce n'est pas toujours le cas: ex. de plantes autogames ou de lignées pures
- L'effet du facteur est **un terme aléatoire**: le $i^{ème}$ niveau est en effet la $i^{ème}$ réalisation d'un tirage aléatoire et ne peut être considéré comme une constante.
- Les questions posées concernent **la population des niveaux possibles**
  - La moyenne μ de l'ensemble des niveaux possibles du facteur
  - L'amplitude de variation entre les niveaux du facteur

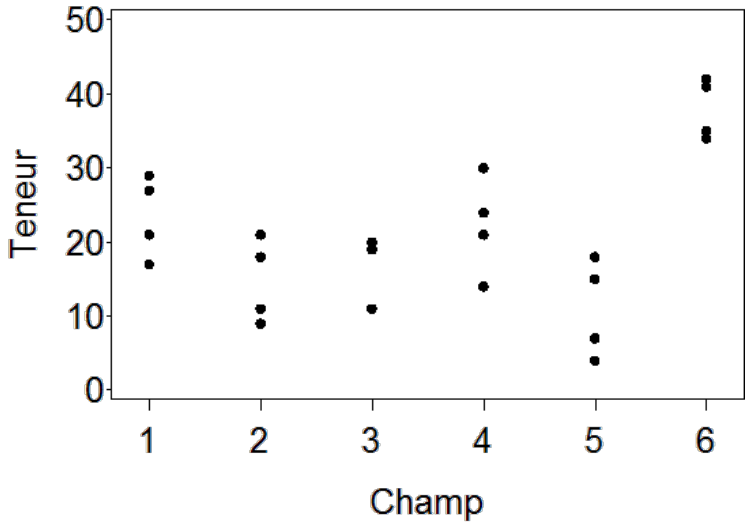
Exemple de motivation

Contexte

On étudie dans une région les risques de contamination de l'aquifère par la culture du maïs en analysant les résidus d'azote dans le sol après la récolte. *On soupçonne l'existence d'une certaine variabilité entre les champs de maïs de la région, dues à des conditions pédologiques ou phytotechniques.* Pour des raisons pratiques, on tire au hasard 6 champs et, dans chacun, on réalise quatre prélèvements.

Données

	Champ	Teneur
1	1	21
2	1	27
3	1	29
4	1	17
5	2	21
6	2	11
7	2	18
8	2	9
9	3	20
10	3	19
11	3	20
12	3	11
13	4	14
14	4	24
15	4	30
16	4	21



Questions

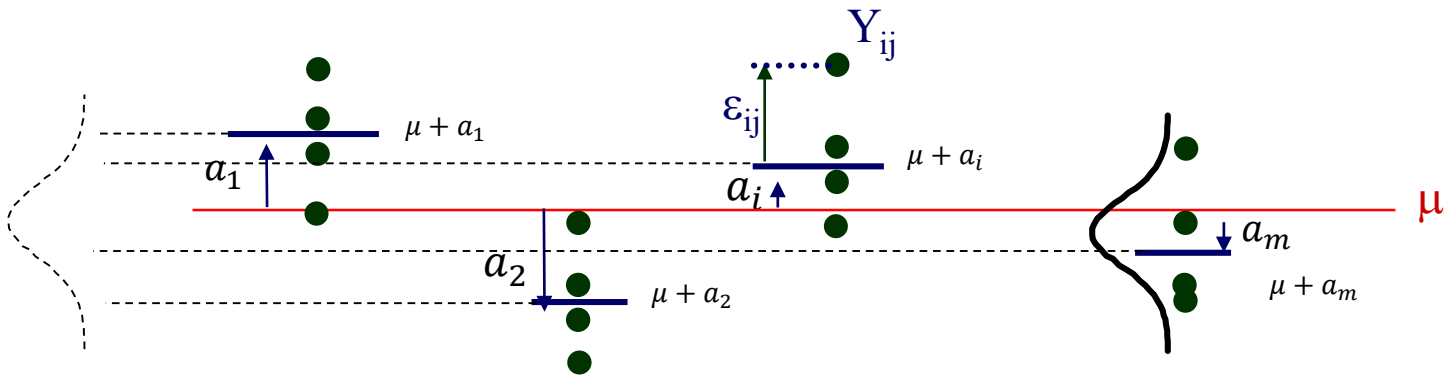
- Estimer la moyenne (+ IC) de la teneur résiduelle en azote **dans la région**
- Quelle est l'importance de la variabilité entre champs et dans chaque champ?

Le modèle de l'ANOVA à un facteur aléatoire

On considère que les observations suivent le modèle probabiliste suivant :

$$Y_{ij} = \mu + a_i + e_{ij} \quad i = 1,\ldots,m \quad j = 1,\ldots,n$$

$$a_i \sim i\, N(0, \sigma_a^2)$$
$$\varepsilon_{ij} \sim i\, N(0, \sigma^2)$$



μ : Résidus azotés moyens attendus (tous champs confondus)

a_i : Effet (déviation **aléatoire**) du $i^{ème}$ champ par rapport à μ.

ε_{ij} : Fluctuation aléatoire due aux différences entre les mesures au sein du champ i

Le modèle de l'ANOVA à un facteur aléatoire

$$Y_{ij} = \mu + a_i + \varepsilon_{ij} \qquad i = 1, \ldots, m \qquad j = 1, \ldots, n$$

$$a_i \sim i\, N(0, \sigma_a^2)$$
$$\varepsilon_{ij} \sim i\, N(0, \sigma^2)$$

Partie fixe — Partie aléatoire

Truc: imaginer que l'on répète l'expérience – les termes qui ne devraient pas changer sont fixes, les autres sont aléatoires.

Le modèle possède deux paramètres de variance à estimer :

- σ_a^2 mesure la variabilité de champ à champ
- σ^2 mesure la variabilité intra-champ

Les paramètres de ce modèle sont $\mu,\ \sigma_a^2,\ \sigma^2$

Les a_i ne sont pas des paramètres du modèle, au contraire de l'ANOVA 1 fixe. Ils sont des **réalisations** de la partie aléatoire du modèle. Songez aux ε_{ij} en ANOVA. La somme des carrés des a_i est une variance – c'est elle le paramètre.

Ecriture du modèle sous forme d'un modèle GLM

$$Y_{ij} = \mu + a_i + \varepsilon_{ij} \quad i = 1, \ldots, m \quad j = 1, \ldots, n$$

$$a_i \sim i\, N(0, \sigma_a^2)$$
$$\varepsilon_{ij} \sim i\, N(0, \sigma^2)$$

$$\boldsymbol{Y} = \begin{bmatrix} Y_{11} \\ Y_{12} \\ Y_{13} \\ Y_{14} \\ Y_{21} \\ Y_{22} \\ Y_{23} \\ Y_{24} \\ \vdots \\ Y_{61} \\ Y_{62} \\ Y_{63} \\ Y_{64} \end{bmatrix} = \begin{bmatrix} \mu + a_1 \\ \mu + a_1 \\ \mu + a_1 \\ \mu + a_1 \\ \mu + a_2 \\ \mu + a_2 \\ \mu + a_2 \\ \mu + a_2 \\ \vdots \\ \mu + a_6 \\ \mu + a_6 \\ \mu + a_6 \\ \mu + a_6 \end{bmatrix} + \begin{bmatrix} \varepsilon_{11} \\ \varepsilon_{12} \\ \varepsilon_{13} \\ \varepsilon_{14} \\ \varepsilon_{21} \\ \varepsilon_{22} \\ \varepsilon_{23} \\ \varepsilon_{24} \\ \vdots \\ \varepsilon_{11} \\ \varepsilon_{62} \\ \varepsilon_{63} \\ \varepsilon_{64} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \mu \\ a_1 \\ a_2 \\ a_3 \\ a_4 \\ a_5 \\ a_6 \end{bmatrix} + \begin{bmatrix} \varepsilon_{11} \\ \varepsilon_{12} \\ \varepsilon_{13} \\ \varepsilon_{14} \\ \varepsilon_{21} \\ \varepsilon_{22} \\ \varepsilon_{23} \\ \varepsilon_{24} \\ \vdots \\ \varepsilon_{11} \\ \varepsilon_{62} \\ \varepsilon_{63} \\ \varepsilon_{64} \end{bmatrix} = \boldsymbol{X_{GLM}\beta_{GLM}} + \boldsymbol{\varepsilon}$$

Ecriture du modèle sous forme d'un modèle GLM

Pour rappel, R réalise le codage de la matrice X d'une manière différente et l'interprétation des paramètres doit être adaptée.

$$\boldsymbol{Y} = \begin{bmatrix} Y_{11} \\ Y_{12} \\ Y_{13} \\ Y_{14} \\ Y_{21} \\ Y_{22} \\ Y_{23} \\ Y_{24} \\ \vdots \\ Y_{61} \\ Y_{62} \\ Y_{63} \\ Y_{64} \end{bmatrix} = \begin{bmatrix} \mu + a_1 \\ \mu + a_1 \\ \mu + a_1 \\ \mu + a_1 \\ \mu + a_2 \\ \mu + a_2 \\ \mu + a_2 \\ \mu + a_2 \\ \vdots \\ \mu + a_6 \\ \mu + a_6 \\ \mu + a_6 \\ \mu + a_6 \end{bmatrix} + \begin{bmatrix} \varepsilon_{11} \\ \varepsilon_{12} \\ \varepsilon_{13} \\ \varepsilon_{14} \\ \varepsilon_{21} \\ \varepsilon_{22} \\ \varepsilon_{23} \\ \varepsilon_{24} \\ \vdots \\ \varepsilon_{11} \\ \varepsilon_{62} \\ \varepsilon_{63} \\ \varepsilon_{64} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \mu \\ a_2 \\ a_3 \\ a_4 \\ a_5 \\ a_6 \end{bmatrix} + \begin{bmatrix} \varepsilon_{11} \\ \varepsilon_{12} \\ \varepsilon_{13} \\ \varepsilon_{14} \\ \varepsilon_{21} \\ \varepsilon_{22} \\ \varepsilon_{23} \\ \varepsilon_{24} \\ \vdots \\ \varepsilon_{11} \\ \varepsilon_{62} \\ \varepsilon_{63} \\ \varepsilon_{64} \end{bmatrix} = \boldsymbol{X_R \beta_R} + \boldsymbol{\varepsilon}$$

Estimation du modèle et espérance des carrés moyens

Décomposition de la variance

$$SST = SSA + SSE$$

$$\sum_{i=1}^{m}\sum_{j=1}^{n}(Y_{ij} - \bar{Y})^2 = \sum_{i=1}^{m}\sum_{j=1}^{n}(\bar{Y}_i - \bar{Y})^2 + \sum_{i=1}^{m}\sum_{j=1}^{n}(Y_{ij} - \bar{Y}_i)^2$$

a_i et ε_{ij}

Degrés de liberté :

SST	*N-1*
SSA	*m-1*
SSE	*N-m*

Espérance des carrés moyens

$$E(MSE) = \sigma^2 \qquad E(MSA) = \sigma^2 + n\sigma_a^2$$

Estimateurs des moindres carrés (+ basés sur les E(CM) pour les variances)

$$\hat{\mu} = \bar{Y} = \frac{1}{N}\sum_{i=1}^{m}\sum_{j=1}^{n} Y_{ij} \qquad \hat{\sigma}^2 = S^2 = MSE = \frac{1}{N-m}\sum_{i=1}^{m}\sum_{j=1}^{n}(Y_{ij} - \bar{Y}_i)^2$$

$$\hat{\sigma}_a^2 = \frac{MSA - MSE}{n}$$

Deux particularités du modèle d'ANOVA1 aléatoire

La variance de Y :

$$\sigma_Y^2 = V(\mu + a_i + \varepsilon_{ij}) = V(a_i + \varepsilon_{ij}) = V(a_i) + V(\varepsilon_{ij}) + 2cov(a_i, \varepsilon_{ij}) = \sigma_a^2 + \sigma^2$$

Les résidus ne dépendent pas du champ

Les Y_{ij} ne sont pas tous indépendants.

Deux observations dans le même champ ($j \neq j'$)

Les résidus sont indépendants

$$cov(Y_{ij}, Y_{ij'}) = cov(\mu + a_i + \varepsilon_{ij}, \mu + a_i + \varepsilon_{ij'}) = cov(a_i, a_i) = \sigma_a^2$$

Deux observations dans des champs distincts ($i \neq i'$, $j \neq j'$)

Les résidus sont indépendants

$$cov(Y_{ij}, Y_{i'j'}) = cov(\mu + a_i + \varepsilon_{ij}, \mu + a_{i'} + \varepsilon_{ij'}) = 0$$

Les champs sont indépendants

La matrice variance-covariance de Y en ANOVA1 aléatoire

$\text{cov}(Y_{ij}, Y_{ij}) : \sigma_Y^2$ | $\text{cov}(Y_{ij}, Y_{ij'})$: obs. d'un même champ | $\text{cov}(Y_{ij}, Y_{i'j'})$: obs. de 2 champs ≠

$$V(Y) = \begin{array}{c|cccccccccccc} ij \rightarrow & 11 & 12 & 13 & 14 & 21 & 22 & 23 & 24 & 31 & 32 & 33 & 34 \\ \hline 11 & \sigma_a^2+\sigma^2 & \sigma_a^2 & \sigma_a^2 & \sigma_a^2 & 0 & \cdots & \cdots & 0 & 0 & \cdots & \cdots & 0 \\ 12 & \sigma_a^2 & \sigma_a^2+\sigma^2 & \sigma_a^2 & \sigma_a^2 & \vdots & & & \vdots & \vdots & & & \vdots \\ 13 & \sigma_a^2 & \sigma_a^2 & \sigma_a^2+\sigma^2 & \sigma_a^2 & \vdots & & & \vdots & \vdots & & & \vdots \\ 14 & \sigma_a^2 & \sigma_a^2 & \sigma_a^2 & \sigma_a^2+\sigma^2 & 0 & \cdots & \cdots & 0 & 0 & \cdots & \cdots & 0 \\ 21 & 0 & \cdots & \cdots & 0 & \sigma_a^2+\sigma^2 & \sigma_a^2 & \sigma_a^2 & \sigma_a^2 & 0 & \cdots & \cdots & 0 \\ 22 & \vdots & & & \vdots & \sigma_a^2 & \sigma_a^2+\sigma^2 & \sigma_a^2 & \sigma_a^2 & \vdots & & & \vdots \\ 23 & \vdots & & & \vdots & \sigma_a^2 & \sigma_a^2 & \sigma_a^2+\sigma^2 & \sigma_a^2 & \vdots & & & \vdots \\ 24 & 0 & \cdots & \cdots & 0 & \sigma_a^2 & \sigma_a^2 & \sigma_a^2 & \sigma_a^2+\sigma^2 & 0 & \cdots & \cdots & 0 \\ 31 & 0 & \cdots & \cdots & 0 & 0 & \cdots & \cdots & 0 & \sigma_a^2+\sigma^2 & \sigma_a^2 & \sigma_a^2 & \sigma_a^2 \\ 32 & \vdots & & & \vdots & \vdots & & & \vdots & \sigma_a^2 & \sigma_a^2+\sigma^2 & \sigma_a^2 & \sigma_a^2 \\ 33 & \vdots & & & \vdots & \vdots & & & \vdots & \sigma_a^2 & \sigma_a^2 & \sigma_a^2+\sigma^2 & \sigma_a^2 \\ 34 & 0 & \cdots & \cdots & 0 & 0 & \cdots & \cdots & 0 & \sigma_a^2 & \sigma_a^2 & \sigma_a^2 & \sigma_a^2+\sigma^2 \end{array} \quad ij \downarrow$$

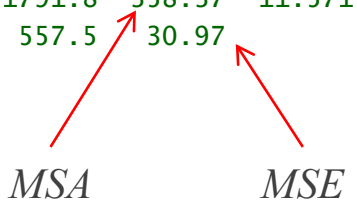
Programme R - package lm

```
dataset <- data.frame("Field" = c( 1,  1,  1,  1,  2,  2,  2,  2,
                                   3,  3,  3,  3,  4,  4,  4,  4,
                                   5,  5,  5,  5,  6,  6,  6,  6),
                      "Nitrogen" = c(21, 27, 29, 17, 21, 11, 18,  9,
                                    20, 19, 20, 11, 14, 24, 30, 21,
                                    7, 15,  18,  4, 41, 42, 35, 34))
dataset$Field <- as.factor(dataset$Field)

mod_lm <- stats::lm(Nitrogen ~ Field, dataset)
summary(mod_lm)
```

```
Call:
lm(formula = Nitrogen ~ Field, data = dataset)

Residuals:
   Min     1Q Median     3Q    Max
-8.250 -4.000  1.625  3.625  7.750

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   23.500      2.783   8.445 1.12e-07 ***
Field2        -8.750      3.935  -2.223  0.03922 *
Field3        -6.000      3.935  -1.525  0.14472
Field4        -1.250      3.935  -0.318  0.75441
Field5       -12.500      3.935  -3.176  0.00523 **
Field6        14.500      3.935   3.685  0.00170 **
```

Programme R - package lm

```
stats::anova(mod_lm)

Analysis of Variance Table

Response: Nitrogen
          Df Sum Sq Mean Sq F value    Pr(>F)
Field      5 1791.8  358.37  11.571 4.106e-05 ***
Residuals 18  557.5   30.97
```



On peut alors retrouver les estimations des paramètres de variance en utilisant les estimateurs des moindres carrés (cf dia 11).

$$\hat{\sigma}^2 = MSE = 30.97$$

$$\hat{\sigma}_a^2 = \frac{MSA - MSE}{n} = \frac{358.37 - 30.97}{4} = 81.85$$

Limites de l'approche GLM

L'approche GLM basée sur la méthode d'estimation des moindres carrés ordinaires

1. Suppose que la matrice variance-covariance des Y_{ij} est $\sigma^2 \mathbf{I}$
2. Ne fournit pas les estimations correctes si les données sont non balancées
3. N'est pas nécessairement la meilleure méthode d'estimation
4. N'estime pas la variance σ_a^2 (on doit la calculer nous-même)
5. Obtient dans certains cas une composante de variance négative
6. Ne fournit pas directement l'estimateur de μ

Nouvelle écriture du modèle linéaire aléatoire

$$Y_{ij} = \mu + a_i + \varepsilon_{ij} \quad i = 1,\ldots,m \quad j = 1,\ldots,n_i \qquad a_i \sim i\, N(0,\sigma_a^2) \qquad \varepsilon_{ij} \sim i\, N(0,\sigma^2)$$

$$\mathbf{Y} = \begin{bmatrix} Y_{11} \\ Y_{12} \\ Y_{13} \\ Y_{14} \\ Y_{21} \\ Y_{22} \\ Y_{23} \\ Y_{24} \\ \vdots \\ Y_{61} \\ Y_{62} \\ Y_{63} \\ Y_{64} \end{bmatrix} = \begin{bmatrix} \mu \\ \mu \\ \mu \\ \mu \\ \mu \\ \mu \\ \mu \\ \mu \\ \vdots \\ \mu \\ \mu \\ \mu \\ \mu \end{bmatrix} + \begin{bmatrix} a_1 \\ a_1 \\ a_1 \\ a_1 \\ a_2 \\ a_2 \\ a_2 \\ a_2 \\ \vdots \\ a_6 \\ a_6 \\ a_6 \\ a_6 \end{bmatrix} + \begin{bmatrix} \varepsilon_{11} \\ \varepsilon_{12} \\ \varepsilon_{13} \\ \varepsilon_{14} \\ \varepsilon_{21} \\ \varepsilon_{22} \\ \varepsilon_{23} \\ \varepsilon_{24} \\ \vdots \\ \varepsilon_{11} \\ \varepsilon_{62} \\ \varepsilon_{63} \\ \varepsilon_{64} \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ \vdots \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} \mu + \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ a_4 \\ a_5 \\ a_6 \end{bmatrix} + \begin{bmatrix} \varepsilon_{11} \\ \varepsilon_{12} \\ \varepsilon_{13} \\ \varepsilon_{14} \\ \varepsilon_{21} \\ \varepsilon_{22} \\ \varepsilon_{23} \\ \varepsilon_{24} \\ \vdots \\ \varepsilon_{11} \\ \varepsilon_{62} \\ \varepsilon_{63} \\ \varepsilon_{64} \end{bmatrix} = \mathbf{X\beta} + \mathbf{Zu} + \boldsymbol{\varepsilon}$$

Partie fixe — Partie aléatoire

$$\boldsymbol{\varepsilon} \sim \mathbf{N}(\mathbf{0},\mathbf{R}) \quad \mathbf{R} = \sigma^2\mathbf{I}_N$$
$$\mathbf{u} \sim \mathbf{N}(\mathbf{0},\mathbf{G}) \quad \mathbf{G} = \sigma_a^2\mathbf{I}_m$$

Structure de la matrice Variance Covariance des Y

$$\mathbf{Y} = \mathbf{X\beta} + \mathbf{Zu} + \boldsymbol{\varepsilon}$$

$\boldsymbol{\varepsilon} \sim \mathbf{N}(\mathbf{0}, \mathbf{R})$ $\quad \mathbf{R} = \sigma^2 \mathbf{I}_N$ Les résidus sont indépendants

$\mathbf{u} \sim \mathbf{N}(\mathbf{0}, \mathbf{G})$ $\quad \mathbf{G} = \sigma_a^2 \mathbf{I}_m$ Les niveaux du facteur aléatoire sont indépendants

$$\mathbf{G} = \begin{bmatrix} \sigma_a^2 & 0 & 0 \\ 0 & \sigma_a^2 & 0 \\ 0 & 0 & \sigma_a^2 \end{bmatrix} = \sigma^2 \boldsymbol{\Lambda}$$

$\mathbf{Y} \sim \mathbf{N}(\mathbf{X\beta}, \mathbf{V})$

$$\mathbf{V} = \mathbf{V}(\mathbf{X\beta} + \mathbf{Zu} + \boldsymbol{\varepsilon}) = \mathbf{V}(\mathbf{Zu}) + \mathbf{V}(\boldsymbol{\varepsilon})$$
$$= \mathbf{ZV}(\mathbf{u})\mathbf{Z}' + \mathbf{R} = \mathbf{ZGZ}' + \mathbf{R}$$

Structure de la matrice Variance Covariance des Y

$$\mathbf{V} = \mathbf{ZGZ}' + \mathbf{R}$$

$$\begin{bmatrix}
\sigma_a^2 & \sigma_a^2 & \sigma_a^2 & \sigma_a^2 & 0 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 & 0 \\
\sigma_a^2 & \sigma_a^2 & \sigma_a^2 & \sigma_a^2 & 0 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 & 0 \\
\sigma_a^2 & \sigma_a^2 & \sigma_a^2 & \sigma_a^2 & 0 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 & 0 \\
\sigma_a^2 & \sigma_a^2 & \sigma_a^2 & \sigma_a^2 & 0 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & \sigma_a^2 & \sigma_a^2 & \sigma_a^2 & \sigma_a^2 & \cdots & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & \sigma_a^2 & \sigma_a^2 & \sigma_a^2 & \sigma_a^2 & \cdots & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & \sigma_a^2 & \sigma_a^2 & \sigma_a^2 & \sigma_a^2 & \cdots & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & \sigma_a^2 & \sigma_a^2 & \sigma_a^2 & \sigma_a^2 & \cdots & 0 & 0 & 0 & 0 \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \cdots & \sigma_a^2 & \sigma_a^2 & \sigma_a^2 & \sigma_a^2 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \cdots & \sigma_a^2 & \sigma_a^2 & \sigma_a^2 & \sigma_a^2 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \cdots & \sigma_a^2 & \sigma_a^2 & \sigma_a^2 & \sigma_a^2 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \cdots & \sigma_a^2 & \sigma_a^2 & \sigma_a^2 & \sigma_a^2
\end{bmatrix} + \sigma^2 \mathbf{I}_{24}$$

Structure de la matrice Variance Covariance des Y

$$\mathbf{V} = \mathbf{ZGZ}' + \mathbf{R}$$

$$\begin{bmatrix}
\sigma_a^2+\sigma^2 & \sigma_a^2 & \sigma_a^2 & \sigma_a^2 & 0 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 & 0 \\
\sigma_a^2 & \sigma_a^2+\sigma^2 & \sigma_a^2 & \sigma_a^2 & 0 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 & 0 \\
\sigma_a^2 & \sigma_a^2 & \sigma_a^2+\sigma^2 & \sigma_a^2 & 0 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 & 0 \\
\sigma_a^2 & \sigma_a^2 & \sigma_a^2 & \sigma_a^2+\sigma^2 & 0 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & \sigma_a^{2+\sigma^2} & \sigma_a^2 & \sigma_a^2 & \sigma_a^2 & \cdots & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & \sigma_a^2 & \sigma_a^2+\sigma^2 & \sigma_a^2 & \sigma_a^2 & \cdots & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & \sigma_a^2 & \sigma_a^2 & \sigma_a^2+\sigma^2 & \sigma_a^2 & \cdots & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & \sigma_a^2 & \sigma_a^2 & \sigma_a^2 & \sigma_a^2+\sigma^2 & \cdots & 0 & 0 & 0 & 0 \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \cdots & \sigma_a^2+\sigma^2 & \sigma_a^2 & \sigma_a^2 & \sigma_a^2 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \cdots & \sigma_a^2 & \sigma_a^2+\sigma^2 & \sigma_a^2 & \sigma_a^2 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \cdots & \sigma_a^2 & \sigma_a^2 & \sigma_a^2+\sigma^2 & \sigma_a^2 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \cdots & \sigma_a^2 & \sigma_a^2 & \sigma_a^2 & \sigma_a^2+\sigma^2
\end{bmatrix}$$

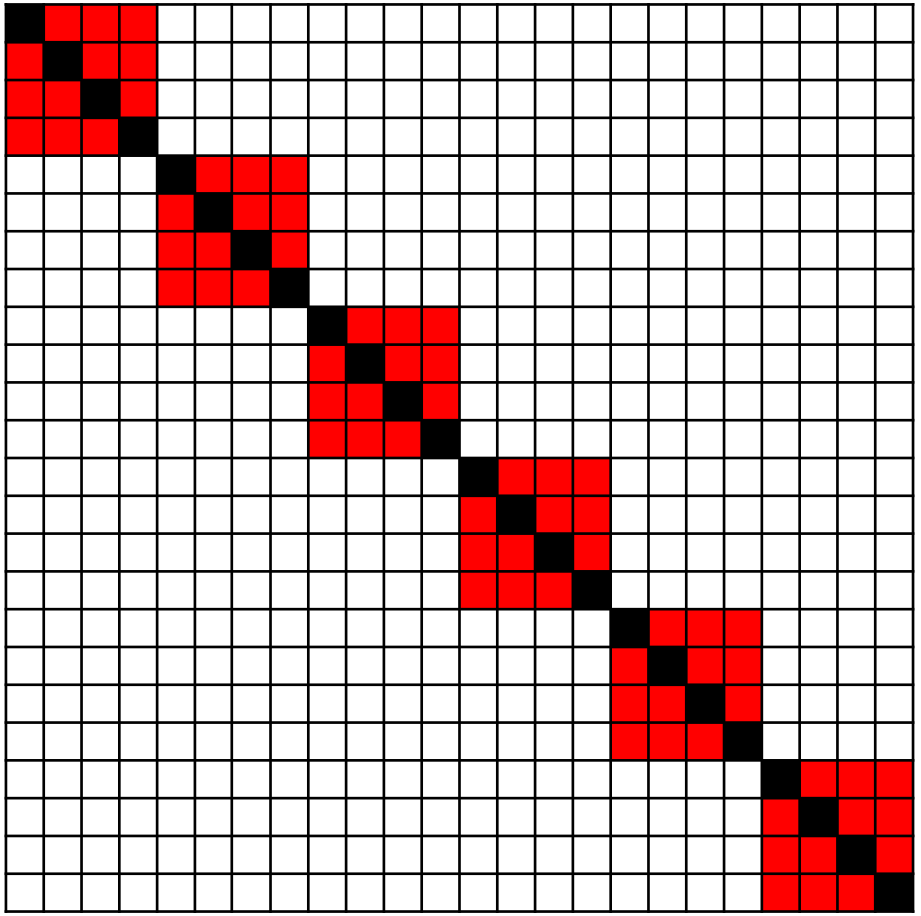
Cette structure « bloc-diagonale » correspond à la forme attendue (cf dia 12), contrairement à celle supposée par le méthode GLM (simple diagonale).

Structure de la matrice Variance Covariance des Y

$$\mathbf{V} = \begin{bmatrix} \mathbf{V}_1 & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \ddots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{V}_6 \end{bmatrix}$$

$$\mathbf{V}_i = \begin{bmatrix} \sigma_a^2 + \sigma^2 & \sigma_a^2 & \sigma_a^2 & \sigma_a^2 \\ \sigma_a^2 & \sigma_a^2 + \sigma^2 & \sigma_a^2 & \sigma_a^2 \\ \sigma_a^2 & \sigma_a^2 & \sigma_a^2 + \sigma^2 & \sigma_a^2 \\ \sigma_a^2 & \sigma_a^2 & \sigma_a^2 & \sigma_a^2 + \sigma^2 \end{bmatrix}$$

$$\mathbf{V} = I_6 \otimes J_4 \sigma_a^2 + I_{24} \sigma^2$$



Estimation des paramètres

Dans le modèle d'ANOVA 1 aléatoire, la présence de deux termes de variance (σ^2 et σ_a^2) fait que la méthode des moindres carrés ne fonctionne pas.

Solution: Utiliser la méthode du maximum de vraisemblance

- Soit Y une variable aléatoire de distribution de probabilité g(Y;θ)
- Soit $\theta=(\theta_1, \theta_2, \ldots \theta_p)$ le vecteur des p paramètres inconnus à estimer
- Soit $Y_1, Y_2 \ldots Y_N$ un échantillon de données indépendantes

La fonction de vraisemblance associée est définie par $L(\theta) = \prod_{i=1}^{N} g(y_i ; \theta)$

L'estimateur de maximum de vraisemblance de θ est la valeur de θ qui maximise L(θ)

En pratique, on maximise le log de la vraisemblance qui, pour le modèle aléatoire, s'écrit:

$$l = \log L = -\frac{N}{2}\log 2\,\pi - \frac{1}{2}\log|\mathbf{V}| - \frac{1}{2}(\mathbf{y} - \mathbf{X\beta})'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X\beta})$$

Estimation par la méthode du maximum de vraisemblance

Dérivation de l par rapport aux paramètres de la partie fixe :

$$\frac{\partial l}{\partial \boldsymbol{\beta}} = -\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\boldsymbol{\beta} + \mathbf{X}'\mathbf{V}^{-1}\mathbf{y} = 0 \quad \Rightarrow \quad \boldsymbol{\beta} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}$$

Cet estimateur coïncide avec l'estimateur des moindes carrés généralisés (GLS). On note que pour $\mathbf{V}=\sigma^2\mathbf{I}$, on obtient l'estimateur des moindres carrés ordinaires vu dans le cas des modèles fixes: $\boldsymbol{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$.

!! Il existe autant de valeurs de $\boldsymbol{\beta}$ que d'inverses généralisées de $(\mathbf{X'V}^{-1}\mathbf{X})$. Cependant, ont peut montrer que $\mathbf{X}\boldsymbol{\beta}$ est unique et ne dépend pas de $(\mathbf{X'V}^{-1}\mathbf{X})^{-}$.

$$\Rightarrow \quad \mathbf{X}\boldsymbol{\beta} = \mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}$$

Dérivation de l par rapport aux paramètres de la partie aléatoire :

$$\frac{\partial l}{\partial \sigma_i^2} = -\frac{1}{2}\mathrm{tr}(\mathbf{V}^{-1}\mathbf{Z}_j\mathbf{Z}_i') - \frac{1}{2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'\mathbf{V}^{-1}\mathbf{Z}_i\mathbf{Z}_j'\mathbf{V}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) = 0$$

On résoud d'abord la seconde équation en substituant $\mathbf{X}\boldsymbol{\beta} = \mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}$.

La résolution de ces équations fournit les estimateurs $\widehat{\boldsymbol{\beta}}$ et $\widehat{\mathbf{V}}$.

Estimation par la méthode du maximum de vraisemblance

Dans le cas de l'ANOVA1 aléatoire, on obtient :

$$\hat{\mu} = \bar{Y} \qquad \hat{\sigma}^2 = MSE \qquad \hat{\sigma}_a^2 = \frac{1}{n}\left[\left(1 - \frac{1}{m}\right) MSA - MSE\right]$$

On note que l'estimateur ML de $\sigma_a{}^2$ est biaisé. Nous avions vu plus haut que l'estimateur non biaisé valait :

$$\hat{\sigma}_a^2 = \frac{MSA - MSE}{n}$$

Pour contourner ce problème, on a recours à la méthode du maximum de vraisemblance restreint (REML).

Estimation par la méthode du REML

La méthode du maximum de vraisemblance restreint (REML) fournit une procédure alternative d'estimation des composantes de variance qui permet de séparer l'estimation des paramètres de la partie fixe et aléatoire du modèle.

REML construit une matrice $\mathbf{K}$ telle que $\mathbf{K'X} = \mathbf{0}$. Sous ces conditions :

$$\mathbf{K}'\mathbf{y} \sim \mathbf{N}(\mathbf{0}, \mathbf{K}'\mathbf{V}\mathbf{K})$$

La vraisemblance de $\mathbf{K'y}$ ne contient plus les termes fixes du modèle (β) :

$$l = \log L = -\frac{1}{2} r_{\mathbf{K}} \log 2\pi - \frac{1}{2} \log|\mathbf{K}'\mathbf{V}\mathbf{K}| - \frac{1}{2} \mathbf{y}'\mathbf{K}(\mathbf{K}'\mathbf{V}\mathbf{K})^{-1}\mathbf{K}'\mathbf{y}$$

La maximisation de cette vraisemblance mène donc directement à l'estimation des termes de $\widehat{\mathbf{V}}$ (σ_a^2 et σ^2).

Les estimateurs REML des composantes de variance ne dépendent pas de β !

Estimation par la méthode du REML

On obtient ensuite les estimateurs GLS des termes fixes :

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\widehat{\mathbf{V}}^{-1}\mathbf{X})^{-}\mathbf{X}'\widehat{\mathbf{V}}^{-1}\mathbf{y}$$

Dans le cas de l'ANOVA1 aléatoire, on obtient :

$$\hat{\mu} = \bar{Y} \qquad \hat{\sigma}^2 = MSE \qquad \hat{\sigma}_a^2 = \frac{MSA - MSE}{n}$$

On note que l'estimateur REML de σ_a^2 est non biaisé

On utilisera dès lors la méthode REML pour ajuster les modèles mixtes.

Modélisation R, package lme4

```
require(lme4)

mod_lmer <- lme4::lmer(Nitrogen ~ 1 + (1 | Field), dataset)
summary(mod_lmer)
```

fixe

aléatoire

```
Linear mixed model fit by REML ['lmerMod']
Formula: Nitrogen ~ 1 + (1 | Field)
   Data: dataset

REML criterion at convergence: 159.7

Scaled residuals:
    Min      1Q  Median      3Q     Max
-1.4656 -0.7993  0.2719  0.6990  1.4094

Random effects:
 Groups   Name        Variance Std.Dev.
 Field    (Intercept) 81.85    9.047
 Residual             30.97    5.565
Number of obs: 24, groups:  Field, 6

Fixed effects:
            Estimate Std. Error t value
(Intercept)   21.167      3.864   5.478
```

L'objectif du modèle d'ANOVA1 aléatoire

L'objectif de l'étude est d'estimer la moyenne et la variabilité des reliquats d'azote à l'échelle de la région. Il ne s'agit donc pas de prédire les reliquats de tel ou tel champ en particulier.

Parenthèse: on peut caractériser la variabilité présente à l'aide du rapport :

$$\frac{\sigma_a^2}{\sigma_Y^2} = \frac{\sigma_a^2}{\sigma_a^2 + \sigma^2}$$

Ce rapport mesure la proportion de la variabilité totale qui est prise en compte par σ_a^2 . Il varie entre 0 et 1:

- 0 : tous les champs ont les mêmes reliquats, la variabilité de Y est due exclusivement aux différences entre échantillons dans un champ $\Rightarrow \sigma_a^2 = 0$ et $\sigma_Y^2 = \sigma^2$
- 1 : les échantillons d'un même champ sont tous identiques, la variabilité des Y est due exclusivement aux différences entre champs $\Rightarrow \sigma_a^2 = \sigma_Y^2$ et $\sigma^2 = 0$).

Ce rapport est nommé coefficient de corrélation intraclasse parce qu'il est égal à la corrélation entre deux échantillons du même champ (cf matrice $\mathbf{V}$)

L'hypothèse principale du modèle d'ANOVA1 aléatoire

L'hypothèse principale du modèle ANOVA1 aléatoire est que les reliquats des différents champs sont égaux :

$$H_0 : \sigma_a^2 = 0$$

$$H_1 : \sigma_a^2 > 0$$

Cette hypothèse peut être testée en partant des espérances des carrés moyens:

$$E(MSA) = \sigma^2 + n\sigma_a^2$$
$$E(MSE) = \sigma^2$$

Le rapport suivant sera proche de 1 si $\sigma_a^2 = 0$ et supérieur à 1 si $\sigma_a^2 > 0$

$$F_{obs} = \frac{MSA}{MSE} \sim F\,(m-1, N-m) \text{ sous } H_0$$

Cette méthode est utilisée dans une approche GLM – cf le résultat de la table `anova(mod_lm)` à la dia 27.

Elle est toutefois délicate à généraliser pour des modèles plus complexes ou lorsque les données sont non balancées.

L'hypothèse principale du modèle d'ANOVA1 aléatoire

Une manière alternative de tester cette hypothèse est d'utiliser le test du rapport de vraisemblance de modèles emboîtés.

On désire comparer les deux modèles suivants :

H_0: Modèle restreint : $Y_{ij} = \mu + \varepsilon_{ij}$ $\quad \varepsilon_{ij} \sim iN(0, \sigma^2)$

H_1: Modèle complet : $Y_{ij} = \mu + a_i + \varepsilon_{ij}$ $\quad a_i \sim iN(0, \sigma_a^2)$ $\quad \varepsilon_{ij} \sim iN(0, \sigma^2)$

avec des vraisemblances $L(R)$ et $L(F)$.

La statistique de test du rapport de vraisemblance est la suivante :

$$LRT = -2\log_e\left[\frac{L(R)}{L(F)}\right] = -2[\log_e L(R) - \log_e L(F)] \sim \chi^2_{(df)} \text{ sous } H_0$$

avec df le nombre de paramètres soustraits du modèle complet (ici $df = 1$ puisqu'on enlève un paramètre: σ_a^2).

On rejette l'hypothèse H_0 lorsque $LRT > \chi^2_{(0.95;df)}$

Résultats de la modélisation R – package lmerTest

```
require(lmerTest)

lmerTest::ranova(mod_lmer)


ANOVA-like table for random-effects: Single term deletions

Model:
Nitrogen ~ (1 | Field)
            npar  logLik    AIC    LRT Df Pr(>Chisq)
<none>         3 -79.826 165.65
(1 | Field)    2 -87.428 178.86 15.204  1  9.652e-05 ***
```

Inférence sur la moyenne de la population

L'estimateur non biaisé de la moyenne de la population est : $\hat{\mu} = \bar{Y}$

La variance de cet estimateur est $\text{var}_\infty(\widehat{\boldsymbol{\beta}}) = (\mathbf{X}'\widehat{\mathbf{V}}^{-1}\mathbf{X})^-$

d'où l'on obtient $\sigma^2_{\bar{Y}} = \frac{n\sigma_a^2+\sigma^2}{N}$

La présence du terme $n\sigma_a^2$ vient de la covariance entre les différentes valeurs de Y_{ij} au sein des champs (pour rappel: la moyenne $\bar{Y}$ est une somme de v.a. Y_{ij} – ici ces Y_{ij} ne sont pas indépendants, donc on retrouve un terme de covariance).

Comme $E[MSA] = n\sigma_a^2 + \sigma^2$, on peut montrer que :

$$\frac{\bar{Y} - \mu}{\sqrt{MSA/N}} \sim t\,(m-1)$$

Et on obtient l'intervalle de confiance pour μ :

$$\bar{Y} \pm t(1-\alpha/2\,;\ m-1)\sqrt{\frac{MSA}{N}}$$

Résultats de la modélisation R - package emmeans

```
require(emmeans)

emmeans::emmeans(mod_lmer)
```

```
1       emmean   SE df lower.CL upper.CL
 overall   21.2 3.86  5     11.2     31.1

Degrees-of-freedom method: kenward-roger
Confidence level used: 0.95
```

$\bar{Y}$

$$\hat{\sigma}_{\bar{Y}} = \sqrt{\frac{MSA}{N}} = \sqrt{\frac{358.37}{24}}$$

Prédiction des effets aléatoires

Dans certains cas, on peut être amené à vouloir estimer les a_i (c'est notamment le cas en génétique quantitative où on souhaite estimer la valeur des différents génotypes au sein de la population).

L'exercice est différent de celui de l'estimation de la valeur moyenne d'un effet fixe. Les différents niveaux du facteur aléatoire étant tirés d'une population, il s'en suit que $a_i \sim N(0, \sigma_a^2)$. Le modèle aléatoire pose donc que leur espérance est nulle.

La valeur « réalisée » d'un niveau d'un facteur aléatoire est plutôt son espérance conditionnelle aux valeurs observées. On peut montrer que:

$$\mathrm{E}[\,\mathbf{u}|\mathbf{y}] = \mathbf{G}\mathbf{Z}'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$$

et on trouve:

$$\hat{\mathbf{u}} = \hat{\mathbf{G}}\mathbf{Z}'\hat{\mathbf{V}}^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})$$

Pour marquer cette distinction, on parlera de prédiction des a_i tandis qu'on réserve le terme d'estimation pour les paramètres du modèle (μ, σ_a^2 et σ^2).

Cette prédiction est dénommée Best Linear Unbiased Predictor (BLUP).

Diverses sorties de R – packages lme4, lmerTest

```
lmerTest::ranef(mod_lmer)

$Field
  (Intercept)
1   2.1316725
2  -5.8620994
3  -3.3497711
4   0.9897051
5  -9.2880016
6  15.3784945

with conditional variances for "Field"
```

valeurs des $\hat{a}_i$

prédiction ≠ estimation

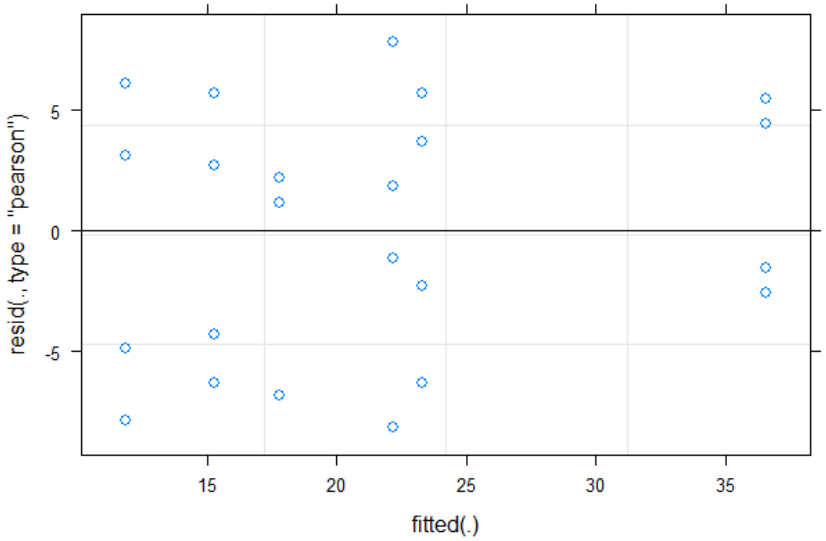
```
lme4::coef(mod_lmer)

$Field
  (Intercept)
1    23.29834
2    15.30457
3    17.81690
4    22.15637
5    11.87867
6    36.54516
```

$\hat{\mu} + \hat{a}_i$ (BLUP)

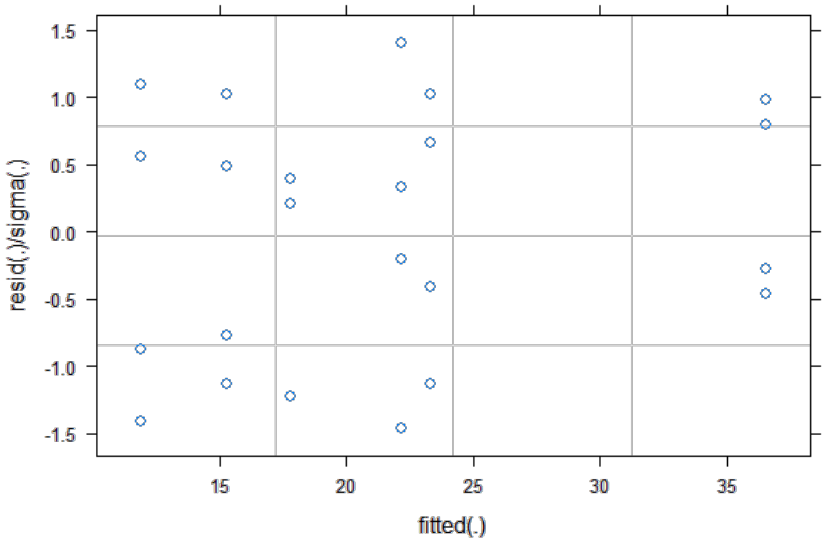
```
emmeans::emmeans(mod_lm, ~Field)

 Field emmean   SE df lower.CL upper.CL
 1       23.5 2.78 18    17.65     29.3
 2       14.8 2.78 18     8.90     20.6
 3       17.5 2.78 18    11.65     23.3
 4       22.2 2.78 18    16.40     28.1
 5       11.0 2.78 18     5.15     16.8
 6       38.0 2.78 18    32.15     43.8
```

Graphiques diagnostiques - package lme4

```
plot(mod_lmer)
```

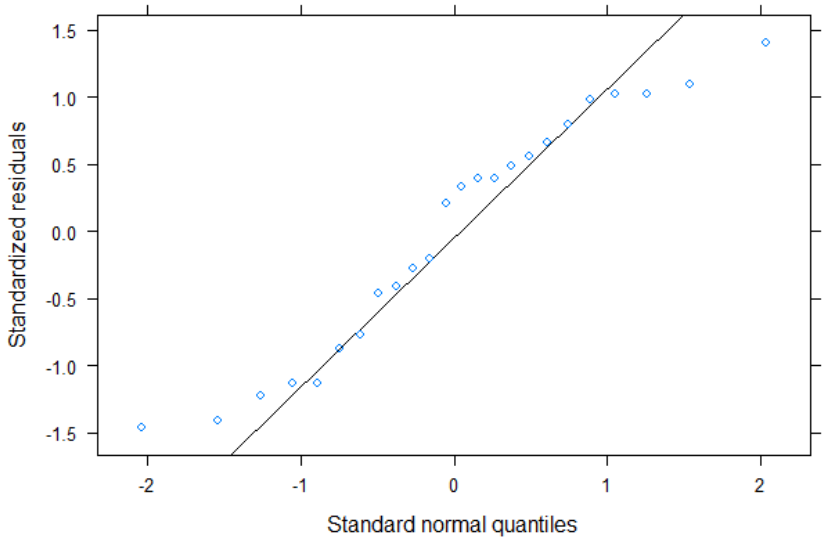



```
plot(mod1, (resid(.) / sigma(.)) ~ fitted(.))
```

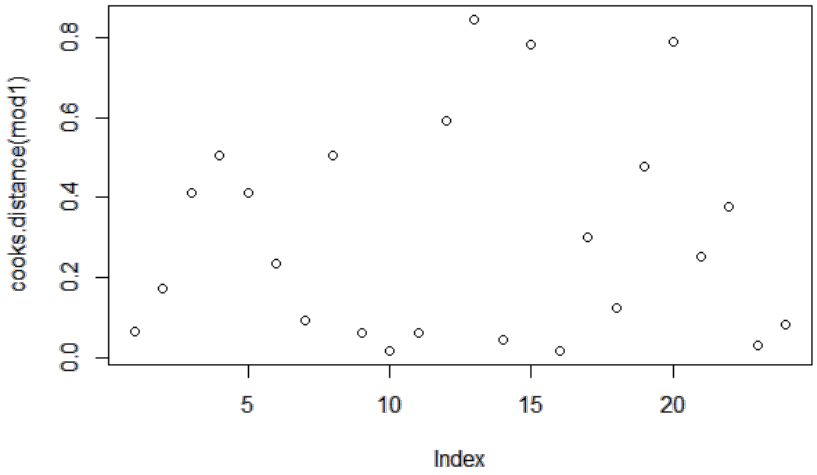


Graphiques diagnostiques - package lme4

```
lattice::qqmath(mod_lmer, id = 0.05)
```




```
plot(cooks.distance(mod_lmer))
```



Facteurs fixes <> facteurs aléatoires

Voir par exemple
https://web.ma.utexas.edu/users/mks/statmistakes/fixedvsrandom.html

La décision de traiter un facteur comme fixe ou aléatoire n'est pas toujours évidente, même pour le statisticien chevronné. Les conséquences au niveau de l'inférence peuvent cependant être importantes. Il est donc important de pouvoir justifier pourquoi on choisit de traiter tel ou tel facteur comme fixe ou aléatoire.

Exemples de situations moins évidentes:

- Une étude sur les effets du non-labour est réalisée dans 5 fermes :
  - Les 5 fermes étudiées sont, a priori, un échantillon des fermes de la région.
  - Comme on imagine qu'il y aura des différences entre fermes, on mettra un facteur « ferme » dans notre modèle d'ANOVA.
  - Si l'on s'intéresse aux 5 fermes de l'échantillon, on traitera le facteur « ferme » comme fixe.
  - Si on souhaite que nos conclusions s'appliquent à l'ensemble des fermes de la région, on traitera le facteur « ferme » comme aléatoire.