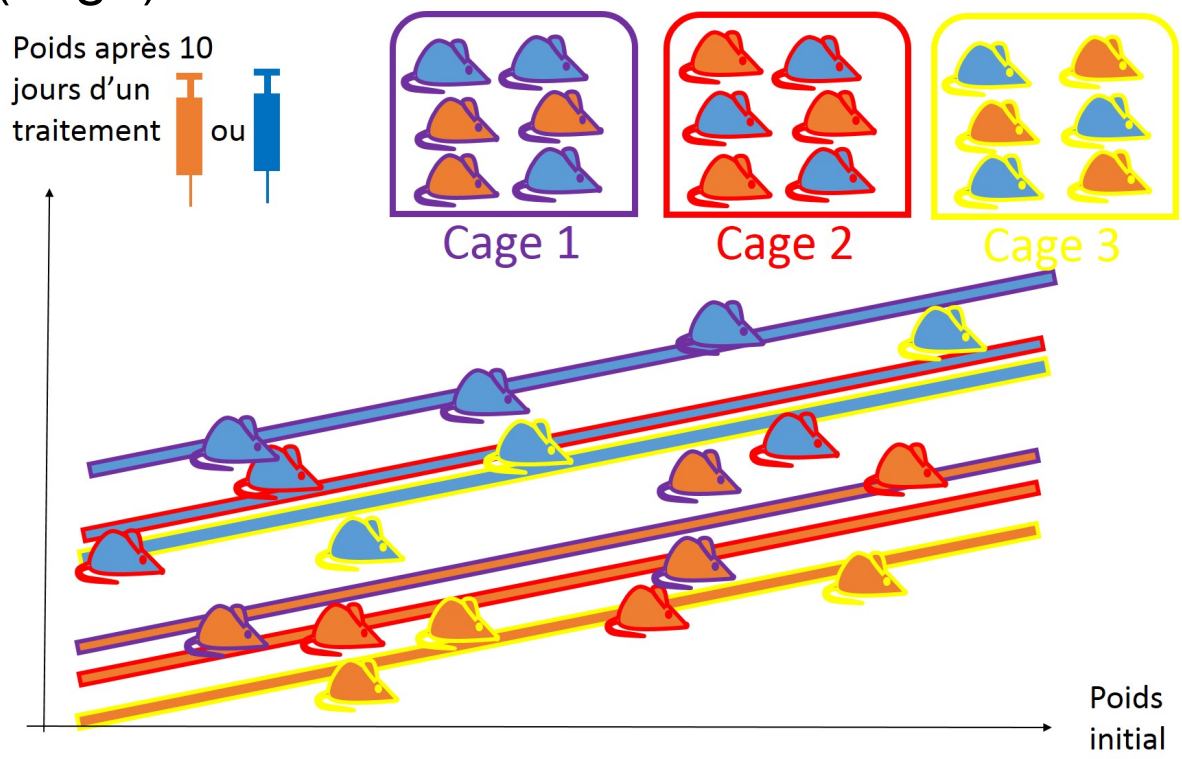
Introduction aux modèles linéaires mixtes et à leur implémentation à l'aide de la fonction `lmer` du package `lme4`

M. L. Delignette-Muller - VetAgro Sup

18 janvier 2024

Illustration

Ex. : modélisation d'une variable quantitative poids 10 jrs ap. traitement) en fonction de deux variables explicatives (poids initial et traitement) en prenant en compte un facteur aléatoire (cage).



Modèle fixe et modèle mixte

- **Facteur fixe** : toutes les modalités étudiées du facteur sont testées dans l'expérience (ex. : facteur "traitement", "sexe", ...)
  L'effet de ce facteur est supposé **prévisible** d'une expérimentation à l'autre (ex. différence entre mâles et femelles).
  Une variable explicative quantitative (ex. : "âge", "temps", ...) est considérée comme fixe (effet = coefficient de régression).
- **Facteur aléatoire** : seul un échantillon aléatoire des modalités du facteur est testé dans l'expérience (ex. : facteur "animal", "cage", "souche bactérienne", ...)
  L'effet de ce facteur est supposé **imprévisible** d'une expérimentation à l'autre (ex. différences entre les cages).

Modèle fixe et modèle mixte en ANOVA1

facteur A fixe (ex. : comparaison de plusieurs traitements)

$X_{ij} = \mu + \alpha_i + \epsilon_{ij}$ avec $\epsilon_{ij} \sim N(0, \sigma)$
Hypothèse nulle testée : $\sum \alpha_i^2 = 0$

Ce qui nous intéresse est l'estimation des effets des différents traitements (caractérisés par μ et les α_i).

facteur A aléatoire (ex. : quantification de la variabilité liée à un facteur "cage")

$X_{ij} = \mu + A_i + \epsilon_{ij}$ avec $\epsilon_{ij} \sim N(0, \sigma)$ et $A_i \sim N(0, \sigma_A)$

Ce qui nous intéresse est l'estimation de la variabilité entre les cages (caractérisé par l'écart type σ_A).

Modèle croisé et modèle hiérarchisé

Exemple d'une analyse de variance à 2 facteurs A et B

Modèle croisé

toutes les modalités testées de B sont croisées avec toutes les modalités testées de A
(ex. : 3 antibiotiques testés sur 5 souches bactériennes, soit 15 groupes en tout)

Modèle hiérarchisé

les modalités possibles de B sont dépendantes de la modalité de A et chaque modalité de B ne peut être associée qu'à une seule modalité de A
(ex. : facteur père et facteur mère sur leur descendance, dans une expérience où chaque père est croisé avec plusieurs

Plan d'expérience avec ou sans répétitions

Exemple d'une analyse de variance à 2 facteurs A et B

Plan d'expérience (ou dispositif) avec répétitions

chaque échantillon (groupe) correspondant à un couple de modalités de A et B contient plusieurs observations (ex. : 3 antibiotiques testés sur 5 souches bactériennes chacun 4 fois)

Plan d'expérience (ou dispositif) sans répétition

chaque échantillon (groupe) correspondant à un couple de modalités de A et B contient une seule observation (ex. : 3 antibiotiques testés sur 5 souches bactériennes chacun une seule fois)

Plan d'expérience équilibré

Un plan d'expérience (ou dispositif) est dit **équilibré** si

- le nombre de répétitions est le même pour tous les groupes,
- le nombre de modalités d'un facteur est le même pour toutes les modalités des autres facteurs.

L'**analyse de variance** classique, calculant des degrés de signification (ou p-value) sur la base d'une décomposition de la somme des carrés des écarts, **requiert généralement un plan d'expérience équilibré**.
L'ajustement de **modèles mixtes permet d'analyser des données issues de dispositifs équilibrés ou non**.

Les deux modèles d'ANOVA 2 dits "classiques"

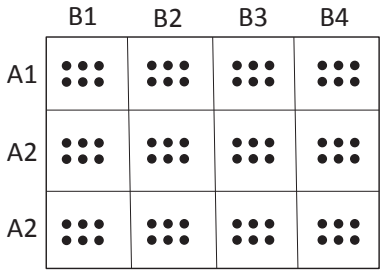
De nombreux cours ou ouvrages abordent l'ANOVA 2 en présentant deux cas parmi les cas possibles :

- l'ANOVA 2 dite "avec répétitions"
  Cette appellation recouvre généralement l'utilisation d'un **modèle croisé fixe avec répétitions**. Un tel modèle ne comporte pas de facteur aléatoire et peut donc être traité avec un modèle linéaire classique.
- l'ANOVA 2 dite "sans répétition"
  Cette appellation recouvre généralement l'utilisation d'un **modèle croisé mixte sans répétition**. On appelle généralement le plan d'expérience associé un dispositif en blocs aléatoires complets.

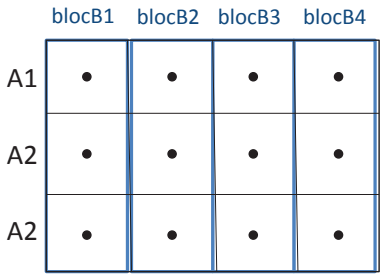
Il serait tout à fait impropre d'appliquer l'une des méthodes correspondant aux deux cas dits "classiques" aux autres cas.

Plans d'expérience classiques - illustration

- ANOVA 2 avec répétitions - modèle fixe



- ANOVA 2 sans répétition - modèle mixte : blocs aléatoires complets



Objectifs pédagogiques

- Savoir caractériser le modèle correspondant à un plan d'expérience donné (identification et nombre de facteurs fixes et aléatoires, hiérarchie éventuelle entre facteurs, ...),
- comprendre le principe des modèles mixtes à partir de l'étude de cas simples à 2 facteurs,
- savoir réaliser et interpréter l'analyse dans le cas de 2 facteurs,
- savoir discuter avec un statisticien dans le cadre de la mise en place d'un plan d'expérience à plus de 2 facteurs et de l'analyse de ses données.

Plan

Exemple de modèle croisé mixte avec répétitions

Une société commercialisant de nombreuses races de souris mutantes à l'usage des laboratoires décide de changer le type de nourriture de ses animaux. Une expérience est entreprise afin de comparer les performances de deux aliments (A : p=2 modalités). Cinq souches (B : q=5 modalités) de souris sont tirées au sort sur le catalogue de la société. Pour chaque souche, huit animaux sont choisis, quatre (n = 4) recevant un des deux aliments. La variable dépendante est le gain de poids.

L'objectif de ce protocole est de comparer les 2 aliments en prenant en compte la variabilité liée à la souche de souris.

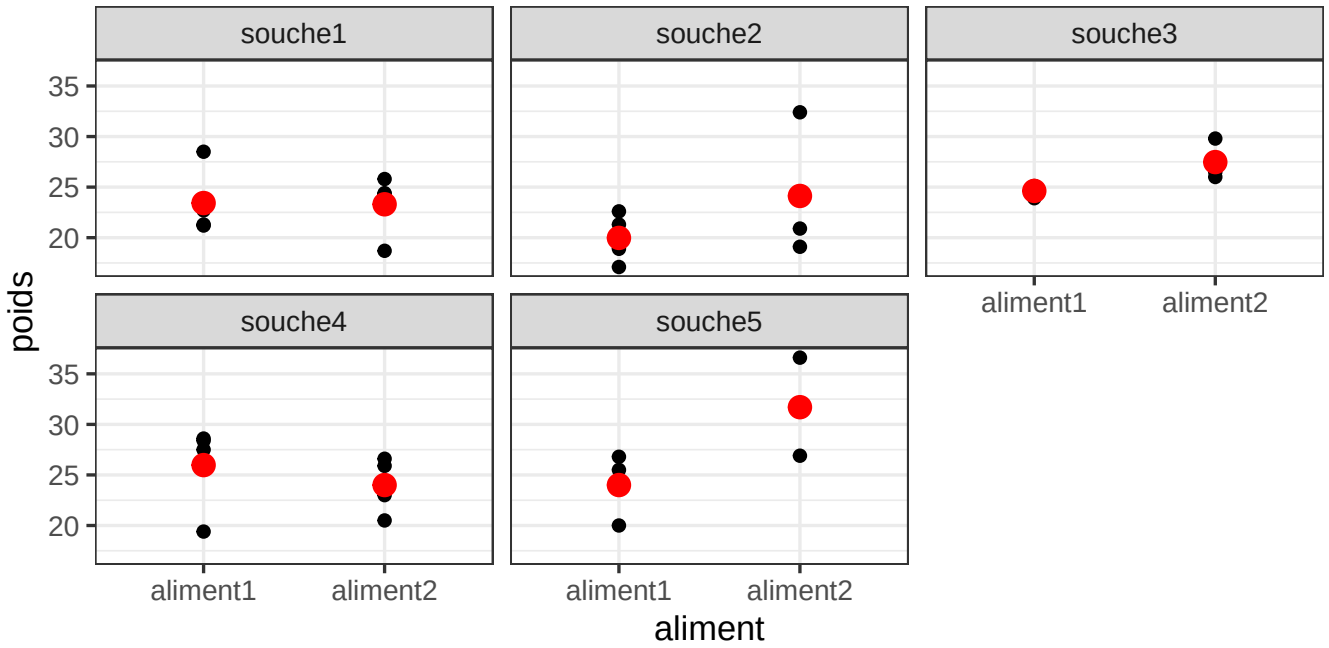
Visualisation des données avec R

Il convient de vérifier le type de chaque variable (`Factor` ou `num`) et le nombre de modalités de chaque variable qualitative et d'examiner le plan d'expérience.

```
> d <- read.table("DATA/NOURRITU.txt", header = TRUE,
+                 stringsAsFactors = TRUE)
> str(d)

'data.frame':        40 obs. of  3 variables:
 $ aliment: Factor w/ 2 levels "aliment1","aliment2": 1 1 1 1
 $ souche : Factor w/ 5 levels "souche1","souche2",..: 1 1 1 1
 $ poids  : num  28.5 21.2 21.3 22.7 25.8 24.3 18.7 24.4 21.3

> xtabs(~ aliment + souche, data = d)

          souche
aliment    souche1 souche2 souche3 souche4 souche5
  aliment1       4       4       4       4       4
  aliment2       4       4       4       4       4
```

Une représentation possible des données avec `ggplot2`

```
> ggplot(data = d, aes(x = aliment, y = poids)) + geom_point()
+   facet_wrap(~ souche)  + stat_summary(fun = mean,
+   geom = "point", colour = "red", size = 3)
```



Modèles théoriques envisageables

Deux modèles mixtes peuvent en théorie être utilisés dans un tel cas, un modèle complet et un modèle simplifié négligeant une éventuelle interaction entre le facteur fixe et le facteur aléatoire (c'est-à-dire une variabilité de l'effet fixe due facteur aléatoire).

Modèle complet

$X_{ijk} = \mu + \alpha_i + B_j + AB_{ij} + \epsilon_{ijk}$ avec $\epsilon_{ijk} \sim N(0, \sigma)$,
$B_j \sim N(0, \sigma_B)$ et $AB_{ij} \sim N(0, \sigma_{AB})$

Modèle simplifié sans interaction

$X_{ijk} = \mu + \alpha_i + B_j + \epsilon_{ijk}$ avec $\epsilon_{ijk} \sim N(0, \sigma)$ et
$B_j \sim N(0, \sigma_B)$

Ajustement du modèle complet sur l'exemple

Il semble plus raisonnable ici d'ajuster le modèle complet sur notre exemple, prenant en compte la variabilité de l'effet "aliment" qui apparaît sur le graphe.

```
> (mm <- lmer(poids ~ aliment + (aliment|souche), data = d))
Linear mixed model fit by REML ['lmerMod']
Formula: poids ~ aliment + (aliment | souche)
   Data: d
REML criterion at convergence: 215
Random effects:
 Groups   Name             Std.Dev. Corr
 souche   (Intercept)      1.44
          alimentaliment2  2.89     -0.12
 Residual                  3.43
Number of obs: 40, groups:  souche, 5
Fixed Effects:
    (Intercept)  alimentaliment2
          23.60             2.52
```

Interprétation des sorties de la fonction `lmer`

Estimation des effets aléatoires (composantes de la variance)

$\hat{\sigma}_{souche} = 1.44$, $\hat{\sigma}_{interaction} = 2.89$ et $\hat{\sigma}_{residuel} = 3.43$

`Corr` correspond à un coefficient de corrélation estimé entre les 2 effets aléatoires (B et AB), le modèle réellement ajusté étant en fait un petit peu plus compliqué que celui présenté dans la diapositive précédente.

Estimation des effets fixes

moyenne de référence $= 23.6$,
Effet aliment$= 2.52$,

Calcul des intervalles de confiance associés aux effets estimés

```
> confint(mm)
                   2.5 % 97.5 %
.sig01              0.00   4.01
.sig02             -1.00   1.00
.sig03              0.00   6.94
.sigma              2.71   4.48
(Intercept)        21.45  25.75
alimentaliment2    -1.10   6.14
```

Au vu de l'intervalle de confiance sur l'effet fixe ([-1.10 ; 6.14]), on ne peut pas conclure à une différence significative entre les performances des deux aliments.

Ajustement du modèle simplifié sur l'exemple

Le modèle simplifié donnerait des résultats quelque peu différents, non pas sur l'estimation ponctuelle des effets fixes, mais sur leurs intervalles de confiance

```
> (mmsimpl <- lmer(poids ~ aliment + (1|souche), data = d))
> fixef(mmsimpl)
    (Intercept) alimentaliment2
          23.60            2.52
> confint(mmsimpl)
                   2.5 % 97.5 %
.sig01             0.000   4.22
.sigma             2.936   4.70
(Intercept)       21.216  25.98
alimentaliment2    0.195   4.84
```

Petite conclusion pour cet exemple

- On ne met pas en évidence d'effet significatif du facteur "aliment" sur le gain de poids des souris dans cette expérience.
- Le nombre de modalités testées du facteur aléatoire "souche" dans cette expérience est faible et on ne peut exclure la présence d'une interaction entre les 2 facteurs c'est-à-dire d'une variabilité de l'effet du facteur "aliment" liée à la souche.
- Il serait sans doute intéressant de recueillir des données sur un plus grand nombre de souches, quitte à prendre moins de souris par souche.

Exemple de modèle croisé aléatoire avec répétitions

Pour quantifier l'effet inhibiteur de la croissance d'une souche d'*Escherichia coli* producteur de shiga-toxines (STEC) par une souche lactique (lactique), on dépose une goutte d'une suspension de la souche lactique sur une gélose préalablement ensemencée avec la souche STEC et on mesure, après incubation, un diamètre d'inhibition de la croissance autour de la goutte. Cette expérience a été réalisée avec 12 souches de STEC (A : 12 modalités) et 9 souches de lactique (B : 9 modalités), en trois répétitions (n = 3).

On souhaite étudier l'impact des 2 facteurs aléatoires "STEC" et "lactique" sur le diamètre d'inhibition.

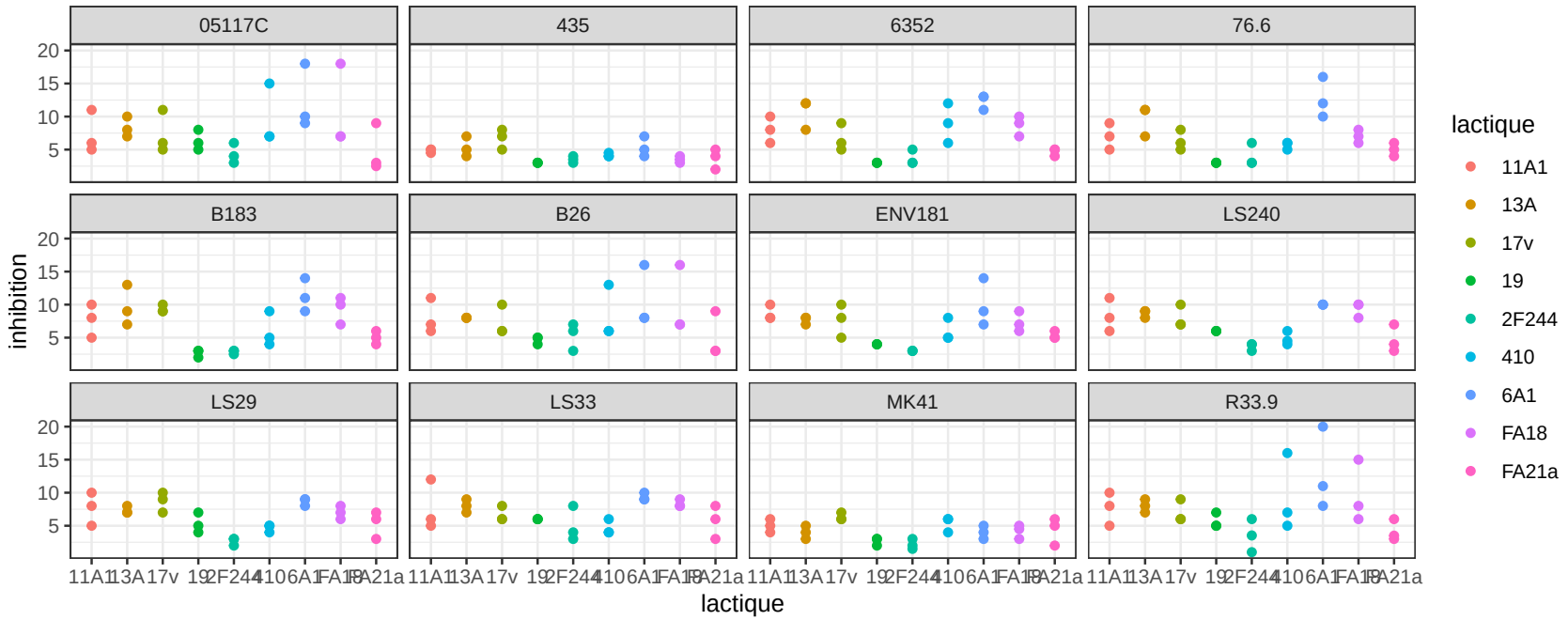
Visualisation des données avec R

Examen des variables et du plan d'expérience

```
> str(d)
'data.frame':        324 obs. of  3 variables:
 $ lactique  : Factor w/ 9 levels "11A1","13A","17v",..: 7 7 7 2 2 2 8
 $ STEC      : Factor w/ 12 levels "05117C","435",..: 7 7 7 7 7 7 7 7
 $ inhibition: num  7 14 9 7 8 8 9 7 6 8 ...
> xtabs(~ STEC + lactique, d)
        lactique
STEC     11A1 13A 17v 19 2F244 410 6A1 FA18 FA21a
  05117C    3   3   3  3     3   3   3    3     3
  435       3   3   3  3     3   3   3    3     3
  6352      3   3   3  3     3   3   3    3     3
  76.6      3   3   3  3     3   3   3    3     3
  B183      3   3   3  3     3   3   3    3     3
  B26       3   3   3  3     3   3   3    3     3
  ENV181    3   3   3  3     3   3   3    3     3
  LS240     3   3   3  3     3   3   3    3     3
  LS29      3   3   3  3     3   3   3    3     3
  LS33      3   3   3  3     3   3   3    3     3
  MK41      3   3   3  3     3   3   3    3     3
```

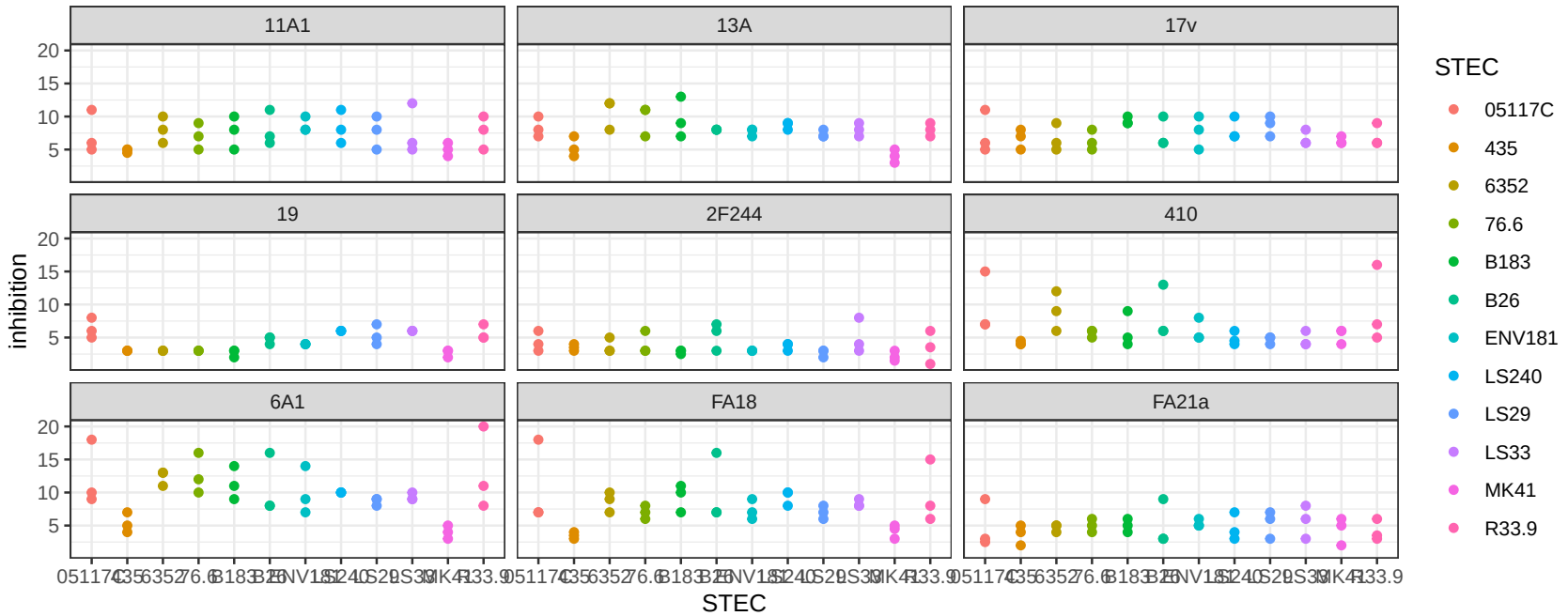
Première représentation graphique avec `ggplot2` - un graphe par souche STEC

```
> ggplot(data = d, aes(x = lactique, y = inhibition, colour = lactique
+   facet_wrap(~ STEC) + geom_point() + theme_bw()
```



Deuxième représentation graphique avec `ggplot2` - un graphe par souche lactique

```
> ggplot(data = d, aes(x = STEC, y = inhibition, colour = STEC)) +
+   facet_wrap(~ lactique) + geom_point()
```



Modèles théoriques envisageables

Deux modèles peuvent en théorie être utilisés dans un tel cas, un modèle complet rarement utilisé et un modèle simplifié plus couramment utilisé dans ce cas, négligeant une éventuelle interaction entre les 2 facteurs aléatoires.

Modèle complet

$X_{ijk} = \mu + A_i + B_j + AB_{ij} + \epsilon_{ijk}$ avec $\epsilon_{ijk} \sim N(0, \sigma)$,
$A_i \sim N(0, \sigma_A)$, $B_j \sim N(0, \sigma_B)$ et $AB_{ij} \sim N(0, \sigma_{AB})$

Modèle simplifié sans interaction

$X_{ijk} = \mu + A_i + B_j + \epsilon_{ijk}$ avec $\epsilon_{ijk} \sim N(0, \sigma)$,
$A_i \sim N(0, \sigma_A)$ et $B_j \sim N(0, \sigma_B)$

Ajustement du modèle simplifié sur l'exemple

```
> (mm <- lmer(inhibition ~  (1|STEC)+ (1|lactique), data = d))
Linear mixed model fit by REML ['lmerMod']
Formula: inhibition ~ (1 | STEC) + (1 | lactique)
   Data: d
REML criterion at convergence: 1496
Random effects:
 Groups   Name        Std.Dev.
 STEC     (Intercept) 1.08
 lactique (Intercept) 1.99
 Residual             2.25
Number of obs: 324, groups:  STEC, 12; lactique, 9
Fixed Effects:
(Intercept)
       6.64
```

Interprétation des sorties de la fonction `lmer`

Estimation des effets aléatoires (composantes de la variance)

$\hat{\sigma}_{STEC} = 1.08$
$\hat{\sigma}_{lactique} = 1.99$
$\hat{\sigma}_{r} = 2.25$

Estimation des effets fixes

moyenne de référence (ici moyenne globale donc) $= 6.64$

Calcul des intervalles de confiance associés aux effets estimés

```
> confint(mm)
                2.5 % 97.5 %
.sig01          0.674   1.80
.sig02          1.238   3.32
.sigma          2.086   2.45
(Intercept)     5.121   8.16
```

Modèle croisé aléatoire ou mixte sans répétition

Pour un modèle mixte ou aléatoire croisé, dans le cas d'un dispositif croisé sans répétition, seul le modèle simplifié (sans interaction) peut être ajusté aux données.
Dans le cas d'un modèle mixte, il conviendra de bien considérer cette hypothèse en amont, notamment si le facteur aléatoire cache un ou plusieurs facteurs fixes qui pourraient entrer en interaction avec le facteur fixe étudié.
Ex : cas d'un facteur "cage", certaines cages contenant des mâles, d'autres des femelles. Il peut être nécessaire de prendre en compte le facteur sexe dans l'analyse des données et son interaction potentielle avec le facteur fixe étudié.

Exemple de modèle croisé mixte sans répétition

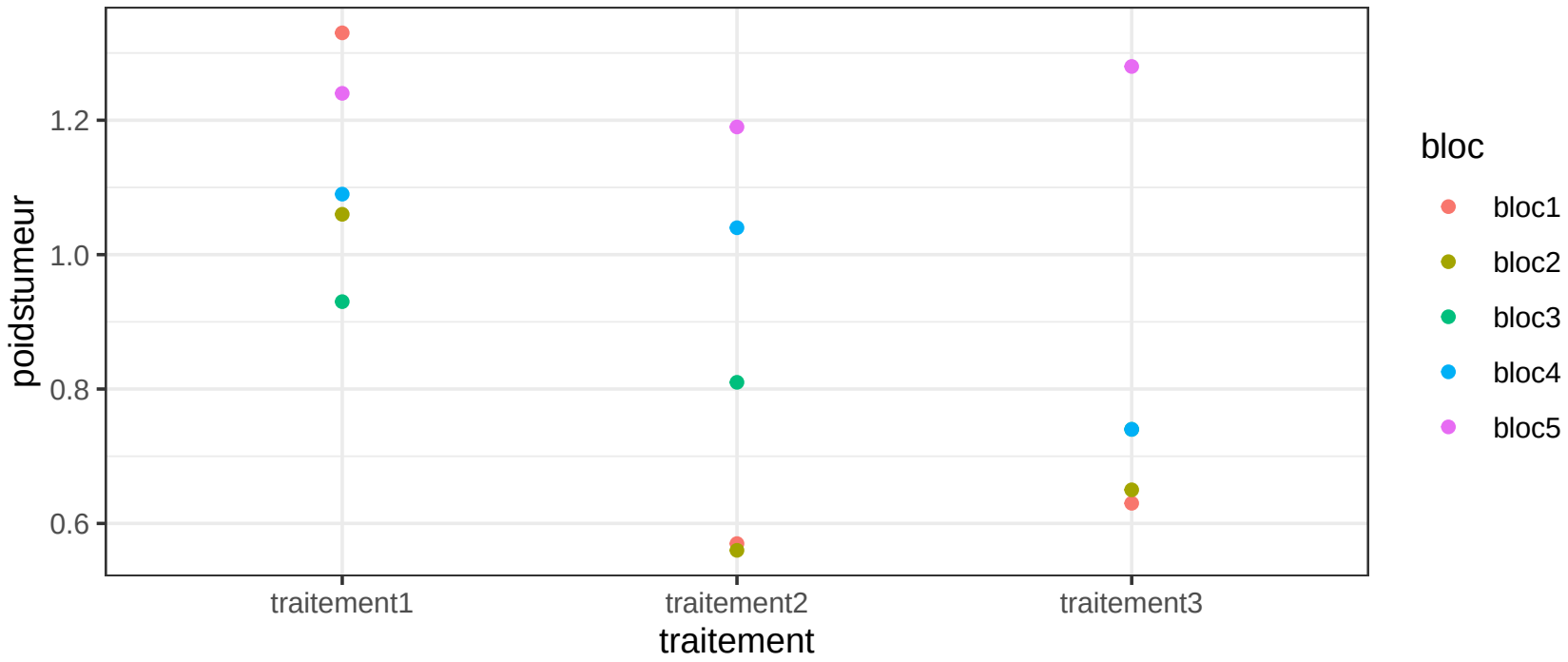
On désire étudier comparativement l'influence de 3 traitements sur le développement d'une tumeur greffée et de ses métastases. On dispose pour cela de 15 animaux. Le critère de jugement choisi est le poids de la tumeur et des métastases après 2 mois de traitement. Cette détermination étant longue et minutieuse, on ne peut sacrifier et examiner que 3 animaux par jour en fin d'expérience, et on décide de ce fait de constituer 5 blocs de 3 animaux recevant les 3 traitements différents.

Visualisation des données avec R

```
> d <- read.table("DATA/tumeur.txt", header = TRUE,
+                 stringsAsFactors = TRUE)
> str(d)
'data.frame':        15 obs. of  3 variables:
 $ poidstumeur: num  1.33 1.06 0.93 1.09 1.24 0.57 0.56 0.81 1
 $ bloc       : Factor w/ 5 levels "bloc1","bloc2",..: 1 2 3 4
 $ traitement : Factor w/ 3 levels "traitement1",..: 1 1 1 1 1
> xtabs(~ traitement + bloc, data = d)
             bloc
traitement    bloc1 bloc2 bloc3 bloc4 bloc5
  traitement1     1     1     1     1     1
  traitement2     1     1     1     1     1
  traitement3     1     1     1     1     1
```

Une représentation possible des données avec `ggplot2`

```
> ggplot(data = d, aes(x = traitement, y = poidstumeur,
+                      colour = bloc)) + geom_point()
```



Modèle théorique

Seul le modèle simplifié supposant l'absence d'interaction entre le facteur fixe et le facteur aléatoire est envisageable du fait de l'absence de répétition.

Modèle sans interaction

$X_{ij} = \mu + \alpha_i + B_j + \epsilon_{ij}$
avec $\epsilon_{ij} \sim N(0, \sigma)$ et
$B_j \sim N(0, \sigma_B)$

Ajustement du modèle sur l'exemple

```
> (mm <- lmer(poidstumeur ~ traitement + (1|bloc), data = d))
Linear mixed model fit by REML ['lmerMod']
Formula: poidstumeur ~ traitement + (1 | bloc)
   Data: d
REML criterion at convergence: 2.94
Random effects:
 Groups   Name        Std.Dev.
 bloc     (Intercept) 0.156
 Residual             0.185
Number of obs: 15, groups:  bloc, 5
Fixed Effects:
          (Intercept)  traitementtraitement2
                1.130                -0.296
traitementtraitement3
               -0.322
```

Interprétation des sorties de la fonction `lmer`

Estimation des effets aléatoires (composantes de la variance)

$\hat{\sigma}_{bloc} = 0.156$
$\hat{\sigma}_{residuel} = 0.185$

Estimation des effets fixes

moyenne de référence (trait. 1) $= 1.130$
différence trait. 2 - trait 1 $= -0.296$
différence trait. 3 - trait 1 $= -0.322$

Calcul des intervalles de confiance associés aux effets estimés

```
> confint(mm)
                              2.5 %  97.5 %
.sig01                        0.000  0.3558
.sigma                        0.113  0.2762
(Intercept)                   0.920  1.3398
traitementtraitement2        -0.522 -0.0697
traitementtraitement3        -0.548 -0.0957
```

Au vu des intervalles de confiance sur les effets fixes (différences entre trait.2 et trait. 1 et entre trait. 3 et trait. 1), on peut conclure à un effet significatif du facteur traitement.

Plan

Exemple de modèle hiérarchisé aléatoire

On mesure le poids à 6 semaines de souris femelles de différentes portées issues de 18 femelles qui ont été croisées avec 6 mâles différents. Deux souris femelles ont été tirées au sort par portée (n=2).

On voudrait quantifier la variabilité des poids des femelles à 6 mois due respectivement au facteur “père” (A : 6 modalités) et au facteur “mère” (B : $18 = 6 \times 3$ modalités, **B emboîté ou niché dans A**).

Visualisation des données avec R - codage

Examen des variables

Bien vérifier le codage explicite de la hiérarchisation !
On ne doit pas retrouver un même nom de mère pour 2 pères différents.

```
> str(d)

'data.frame':         36 obs. of  3 variables:
 $ pere : Factor w/ 6 levels "A","B","C","D",..: 1 1 1 1 1 1 2
 $ mere : Factor w/ 18 levels "m1","m10","m11",..: 1 1 11 11 1
 $ poids: num  19.3 21.9 22.7 24.6 21 19.1 19.3 21.9 20.6 17.9
```

Visualisation des données avec R - plan d'expérience

Examen du plan d'expérience

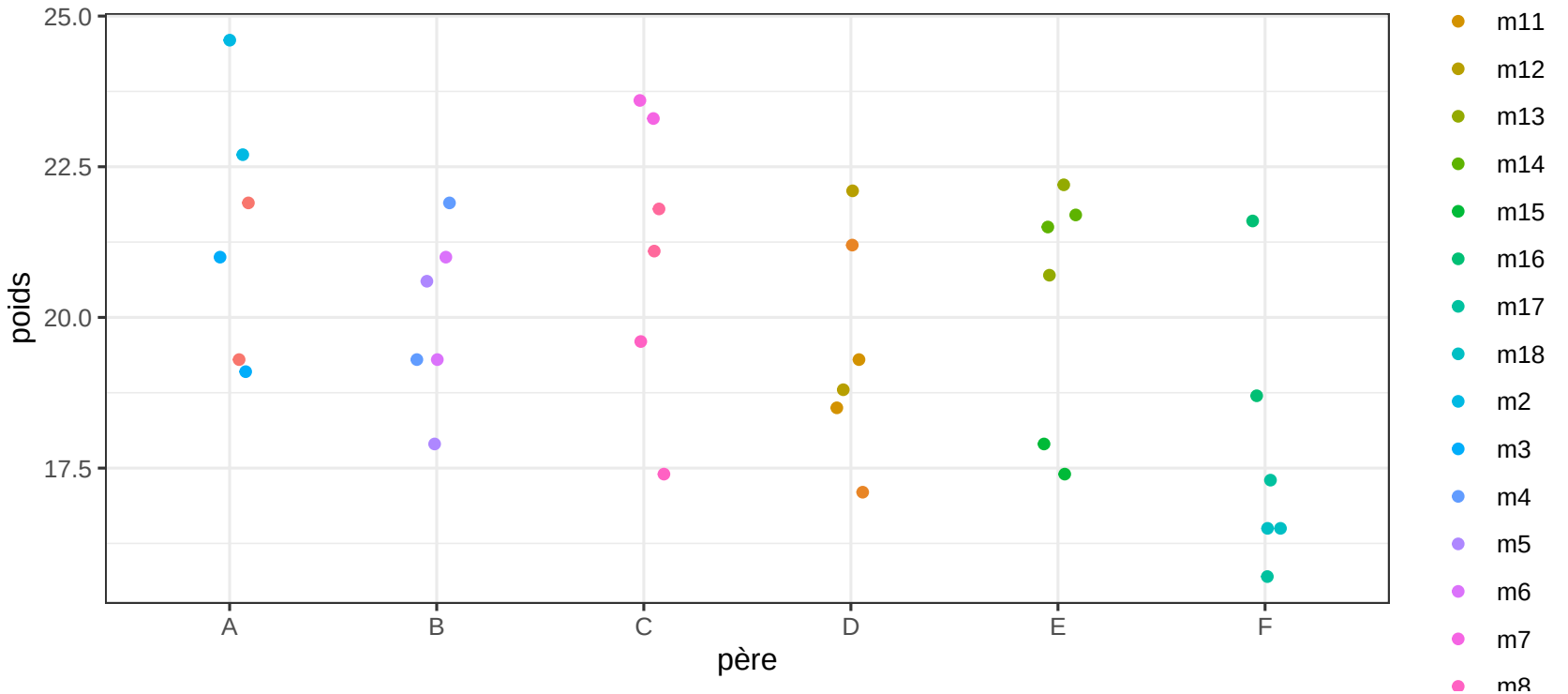
```
> xtabs(~ pere + mere,d)
```

pere	m1	m10	m11	m12	m13	m14	m15	m16	m17	m18	m2	m3	m4	m5	m6	m7	m8
A	2	0	0	0	0	0	0	0	0	0	2	2	0	0	0	0	0
B	0	0	0	0	0	0	0	0	0	0	0	0	2	2	2	0	0
C	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	2
D	0	2	2	2	0	0	0	0	0	0	0	0	0	0	0	0	0
E	0	0	0	0	2	2	2	0	0	0	0	0	0	0	0	0	0
F	0	0	0	0	0	0	0	2	2	2	0	0	0	0	0	0	0

pere	m9
A	0
B	0
C	2
D	0
E	0
F	0

Essai de représentation des données

```
> ggplot(data = d, aes(x = pere, y = poids, colour = mere)) +
+   geom_jitter(height = 0, width = 0.1) + xlab("père")
```



Modèle théorique

Modèle hiérarchisé à 2 facteurs aléatoires A et B

$X_{ijk} = \mu + A_i + B(A)_{ij} + \epsilon_{ijk}$
avec
$\epsilon_{ijk} \sim N(0, \sigma)$,
$A_i \sim N(0, \sigma_A)$ et
$B(A)_{ij} \sim N(0, \sigma_{B(A)})$

Ajustement du modèle sur l'exemple

```
> (mm <- lmer(poids ~ (1|pere/mere), data = d))
Linear mixed model fit by REML ['lmerMod']
Formula: poids ~ (1 | pere/mere)
   Data: d
REML criterion at convergence: 152
Random effects:
 Groups     Name        Std.Dev.
 mere:pere (Intercept) 1.531
 pere      (Intercept) 0.793
 Residual              1.474
Number of obs: 36, groups:  mere:pere, 18; pere, 6
Fixed Effects:
(Intercept)
         20
```

Interprétation des sorties de la fonction `lmer`

Estimation des effets aléatoires (composantes de la variance)

$\hat{\sigma}_{mere(pere)} = 1.53$
$\hat{\sigma}_{pere} = 0.79$
$\hat{\sigma}_{r} = 1.47$

Estimation des effets fixes

moyenne de référence $= 20$

Calcul des intervalles de confiance associés aux effets estimés

```
> confint(mm)
                2.5 % 97.5 %
.sig01          0.504   2.58
.sig02          0.000   2.20
.sigma          1.098   2.13
(Intercept) 18.852  21.15
```

Dans un tel dispositif, la variance associé au facteur du niveau supérieur est plus difficile à estimer que celle associé au facteur du niveau inférieur.

Exemple de modèle hiérarchisé mixte

A partir de 6 souris A, B, C, D, E, F, dites "donneuses" (de cellules cancéreuses), on injecte 200 000 cellules à 24 souris "receveuses". Au bout d'un certain temps on compte le nombre de nodules pulmonaires qui apparaissent chez chacune des 24 souris. Les souris "donneuses" A, B, C ont subi une irradiation aux rayons X, les souris "donneuses" D, E, F, ont subi une irradiation au rayon γ.

On voudrait savoir si le traitement reçu par les donneuses modifie l'avenir des receveuses. On utilisera un modèle hiérarchisé mixte avec un facteur fixe (l'irradiation) et un facteur aléatoire (la souris donneuse) emboîté (ou niché) dans le facteur fixe.

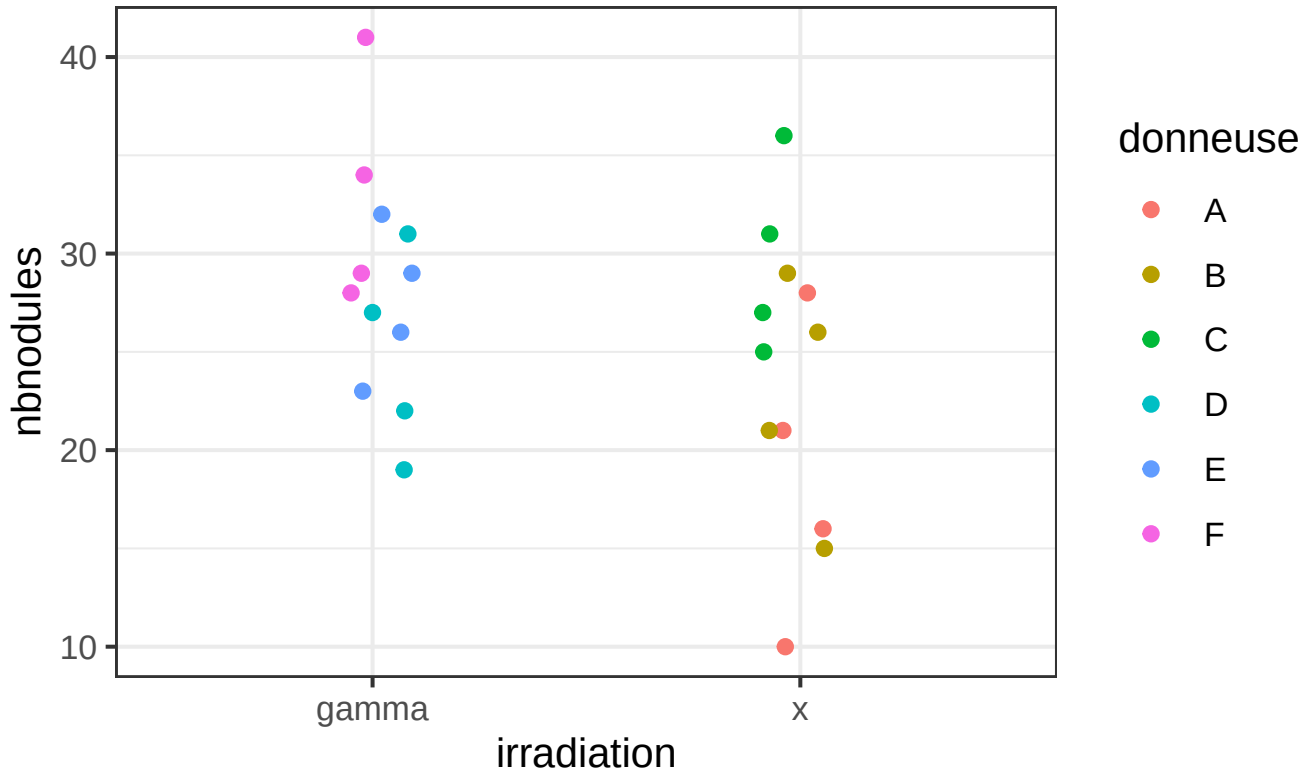
Visualisation des données et du plan d'expérience avec R

Bien vérifier le codage explicite de la hiérarchisation !
On ne doit pas retrouver un un même nom de souris pour 2 traitements différents.

```
> str(d)

'data.frame':        24 obs. of  3 variables:
 $ irradiation: Factor w/ 2 levels "gamma","x": 2 2 2 2 2 2 2 2
 $ donneuse   : Factor w/ 6 levels "A","B","C","D",..: 1 1 1 1
 $ nbnodules  : int  10 16 21 28 15 21 26 29 25 27 ...

> xtabs(~ irradiation + donneuse, data = d)

           donneuse
irradiation A B C D E F
      gamma 0 0 0 4 4 4
      x     4 4 4 0 0 0
```

Essai de représentation des données

```
> ggplot( data = d, aes(x = irradiation, y = nbnodules,
+   colour = donneuse)) + geom_jitter(height = 0, width = 0.1)
```



Modèle théorique

Modèle hiérarchisé à un facteur fixe A et un facteur aléatoire B(A)

$X_{ijk} = \mu + \alpha_i + B(A)_{ij} + \epsilon_{ijk}$
avec
$\epsilon_{ijk} \sim N(0, \sigma)$ et
$B(A)_{ij} \sim N(0, \sigma_{B(A)})$

Ajustement du modèle sur l'exemple

```
> (mm <- lmer(nbnodules ~ irradiation + (1|donneuse),
+             data = d))
Linear mixed model fit by REML ['lmerMod']
Formula: nbnodules ~ irradiation + (1 | donneuse)
   Data: d
REML criterion at convergence: 149
Random effects:
 Groups   Name        Std.Dev.
 donneuse (Intercept) 4.01
 Residual             5.74
Number of obs: 24, groups:  donneuse, 6
Fixed Effects:
 (Intercept)  irradiationx
       28.42         -4.67
```

Interprétation des sorties de la fonction `lmer`

Estimation des effets aléatoires (composantes de la variance)

$\hat{\sigma}_{donneuse} = 4.01$
$\hat{\sigma}_{r} = 5.74$

Estimation des effets fixes

moyenne de référence (rayons γ) $= 28.42$
Effet de l'irradiation (différence $x - \gamma = -4.67$)

Calcul des intervalles de confiance associés aux effets estimés

```
> confint(mm)
                 2.5 % 97.5 %
.sig01            0.00   7.55
.sigma            4.28   8.28
(Intercept)      23.02  33.81
irradiationx    -12.29   2.96
```

On ne met pas en évidence de différence significative entre l'effet des deux types d'irradiation sur le nombre de nodules pulmonaires obtenus chez les receveuses.

Modèle hiérarchisé sans répétition ?

- Peut-on envisager le cas d'un modèle hiérarchisé sans répétition ?
- A quel cas plus simple est-on ramené si l'on n'a pas de répétition pour chaque modalité du facteur de niveau inférieur ?
  Ex. : une seule souris receveuse pour chaque souris donneuse dans l'exemple précédent.
- L'unité d'observation correspond alors au facteur de niveau inférieur.
  - à la souris donneuse dans l'exemple -
  et l'**on est ramené à un modèle linéaire à un seul facteur**.

Modèle hiérarchisé sans répétition ?

- Peut-on envisager le cas d'un modèle hiérarchisé sans répétition ?
- A quel cas plus simple est-on ramené si l'on n'a pas de répétition pour chaque modalité du facteur de niveau inférieur ?
  Ex. : une seule souris receveuse pour chaque souris donneuse dans l'exemple précédent.
- L'unité d'observation correspond alors au facteur de niveau inférieur.
  - à la souris donneuse dans l'exemple -
  et l'**on est ramené à un modèle linéaire à un seul facteur**.

Modèle hiérarchisé sans répétition ?

- Peut-on envisager le cas d'un modèle hiérarchisé sans répétition ?
- A quel cas plus simple est-on ramené si l'on n'a pas de répétition pour chaque modalité du facteur de niveau inférieur ?
  Ex. : une seule souris receveuse pour chaque souris donneuse dans l'exemple précédent.
- L'unité d'observation correspond alors au facteur de niveau inférieur.
  - à la souris donneuse dans l'exemple -
  et l'**on est ramené à un modèle linéaire à un seul facteur**.

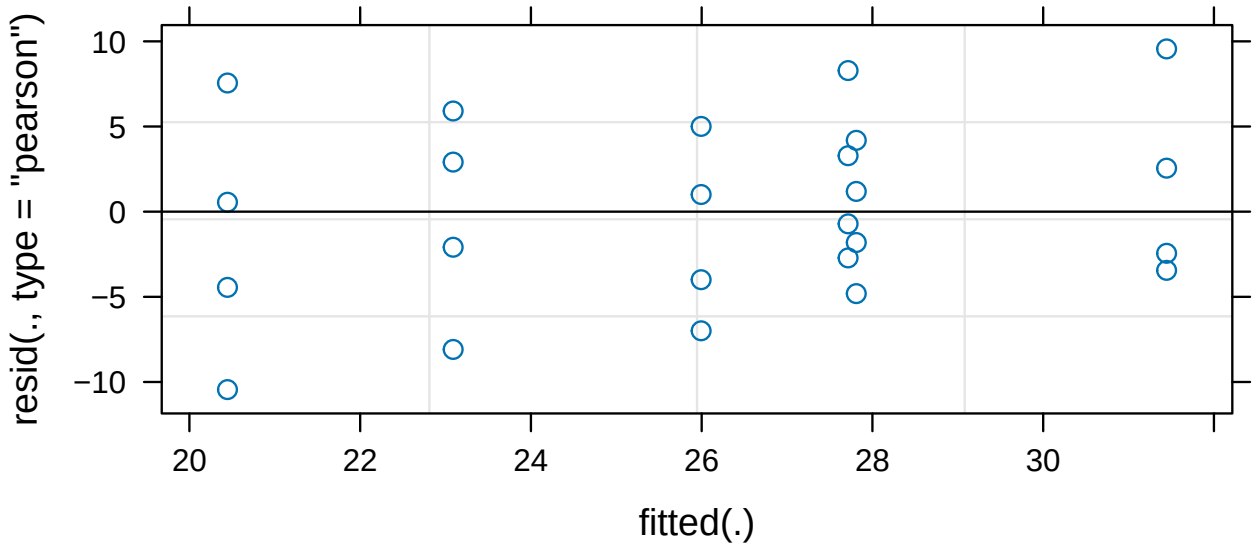
Plan

Graphe des résidus en fonction des valeurs prédites

Quel que soit le modèle, les résidus doivent être distribués aléatoirement suivant une loi normale de variance constante.

Graphe des résidus sur dernier exemple :

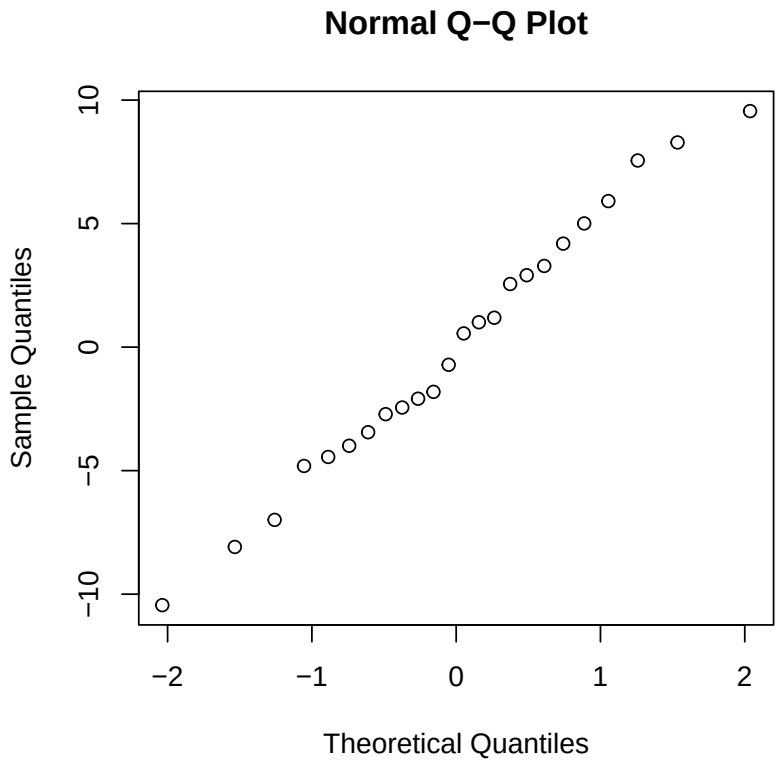
```
> plot(mm)
```



Graphe Quantile-Quantile des résidus

Sur le dernier exemple :

```
> qqnorm(residuals(mm))
```



Graphe Quantile-Quantile des effets aléatoires

Les effets aléatoires sont supposés suivre une loi normale. Sur le dernier exemple, graphe Quantile-Quantile des effets "donneuse" :

```
> qqnorm(ranef(mm)$donneuse[,1])
```

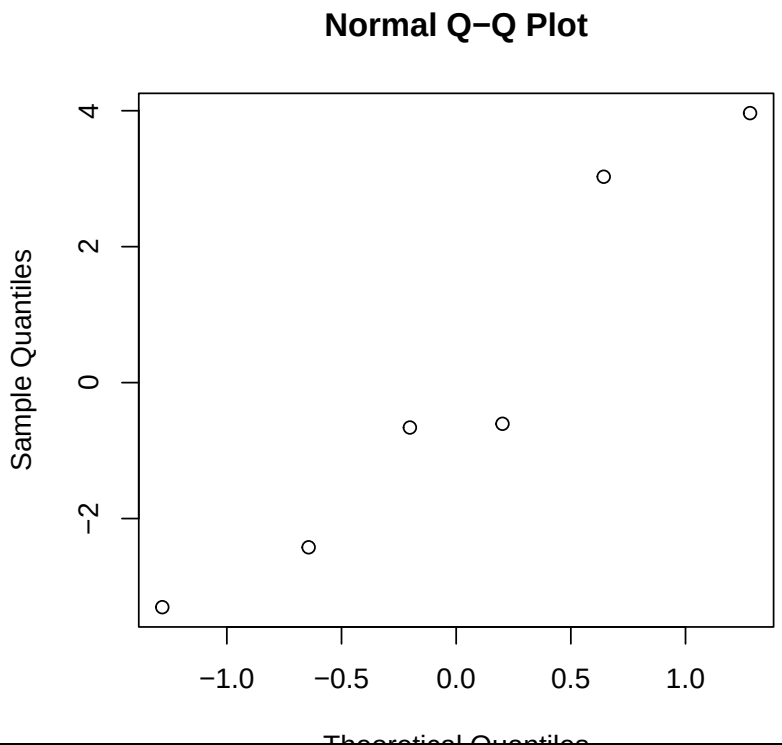


Tableau d'ANOVA sur dispositif déséquilibré ?

Que se passe-t-il si le dispositif est déséquilibré ?
Modifions le jeu de données de l'exemple du modèle croisé mixte sans répétition :

```
> d <- read.table("DATA/tumeur.txt", header = TRUE,
+                 stringsAsFactors = TRUE)
> dincomplet <- d[-1,]
> xtabs(~ traitement + bloc, data = dincomplet)
             bloc
traitement    bloc1 bloc2 bloc3 bloc4 bloc5
  traitement1     0     1     1     1     1
  traitement2     1     1     1     1     1
  traitement3     1     1     1     1     1
```

Tableau d'ANOVA sur dispositif déséquilibré - exemple

Tentons de réaliser un tableau d'analyse de variance classique en changeant l'ordre des 2 facteurs dans le modèle.

```
> summary(aov(poidstumeur ~ traitement + bloc, data = dincomplet))
            Df Sum Sq Mean Sq F value Pr(>F)
traitement   2  0.193  0.0967    5.38  0.038
bloc         4  0.526  0.1315    7.33  0.012
Residuals    7  0.126  0.0180
> summary(aov(poidstumeur ~  bloc + traitement, data = dincomplet))
            Df Sum Sq Mean Sq F value Pr(>F)
bloc         4  0.607  0.1518    8.45 0.0081
traitement   2  0.112  0.0562    3.13 0.1069
Residuals    7  0.126  0.0180
```

Les conclusions tirées de ces 2 tableaux d'ANOVA seraient différente !

Evitez de faire des tableaux d'ANOVA

- Les tableaux d'ANOVA ne s'appliquent pas aux cas déséquilibrés, ni aux cas où l'on a des variables explicatives quantitatives, donc au final s'appliquent rarement !
- Par ailleurs ces tableaux sont peu informatifs quant aux effets estimés, et de ce fait peuvent conduire à une interprétation exclusive, voire abusive, des p-value.

Considérez les tableaux d'ANOVA comme des habitudes anciennes, à abandonner. On sait faire bien mieux avec les modèles.

Absence de valeurs de p dans les sorties de `lmer`

Exemple :

```
> summary(mm)$coefficients
             Estimate Std. Error t value
(Intercept)     28.42       2.85    9.98
irradiationx    -4.67       4.03   -1.16
```

Absence de valeurs de p dans les sorties de `lmer`

`lmer` ne fournit pas de valeurs de p associées à chacun des paramètres du modèle (ddl de la loi de Student associée non connu - le package `lmerTest` peut être utilisé pour cela si vous y tenez),
mais il fournit des **intervalles de confiance** sur les paramètres du modèle (effet fixe ou aléatoire) **valables même en cas de dispositif déséquilibré et ne dépendant pas de l'ordre d'introduction des facteurs dans le modèle**.

Test de la significativité d'un effet aléatoire

L'estimation des composantes de la variance est généralement plus intéressante que le test des effets aléatoires.
Néanmoins `lmer` permet de tester un effet aléatoire par le biais de la comparaison de 2 modèles emboîtés à l'aide de la **statistique du rapport de vraisemblance** (la **partie fixe des 2 modèles** doit être **identique** et l'argument `REML` doit être fixé à `FALSE`)
Mais il serait maladroit de se baser sur le test d'un effet aléatoire pour savoir si on introduit cet effet dans le modèle ou non. Un effet aléatoire qui a un sens biologique, même s'il est non significatif, doit être pris en compte dans l'estimation des effets fixes si c'est techniquement possible.

Récapitulatif des modèles à 2 facteurs

Formules des différents modèles

- A et B aléatoires croisés sans interaction : `(1|A) + (1|B)`
- A fixe et B aléatoire croisés avec interaction : `A + (A|B)`
- A fixe et B aléatoire croisés sans interaction : `A + (1|B)`
- A aléatoire et B aléatoire emboîté dans A : `(1|A/B)`
- A fixe et B aléatoire emboîté dans A : `A + (1|B)`

Les formules sont les mêmes pour une variable explicative quantitative dont l'effet modélisé est forcément un effet fixe (donc codé dans la formule comme un facteur fixe), mais **on suppose alors une relation linéaire et l'effet estimé est le coefficient de régression** (pente) de cette relation linéaire.

Quid des modèles à plus de 2 facteurs

- Les concepts vus précédemment permettent en théorie de traiter tous les modèles à plus de 2 facteurs.
- Le nombre de modèles possibles avec plus de 2 facteurs devient vite grand et le choix d'un modèle plus compliqué (quelles interactions prendre en compte ?).
- l'interprétation des résultats se complique aussi avec l'augmentation du nombre de facteurs et/ou l'augmentation du nombre de modalités par facteur.

Il est généralement recommandé de limiter autant que possible le nombre de facteurs étudiés dans une expérience biologique et de demander un avis statistique dès la planification expérimentale lorsque le nombre de facteurs augmente.

Un exemple de modèle mixte à trois facteurs

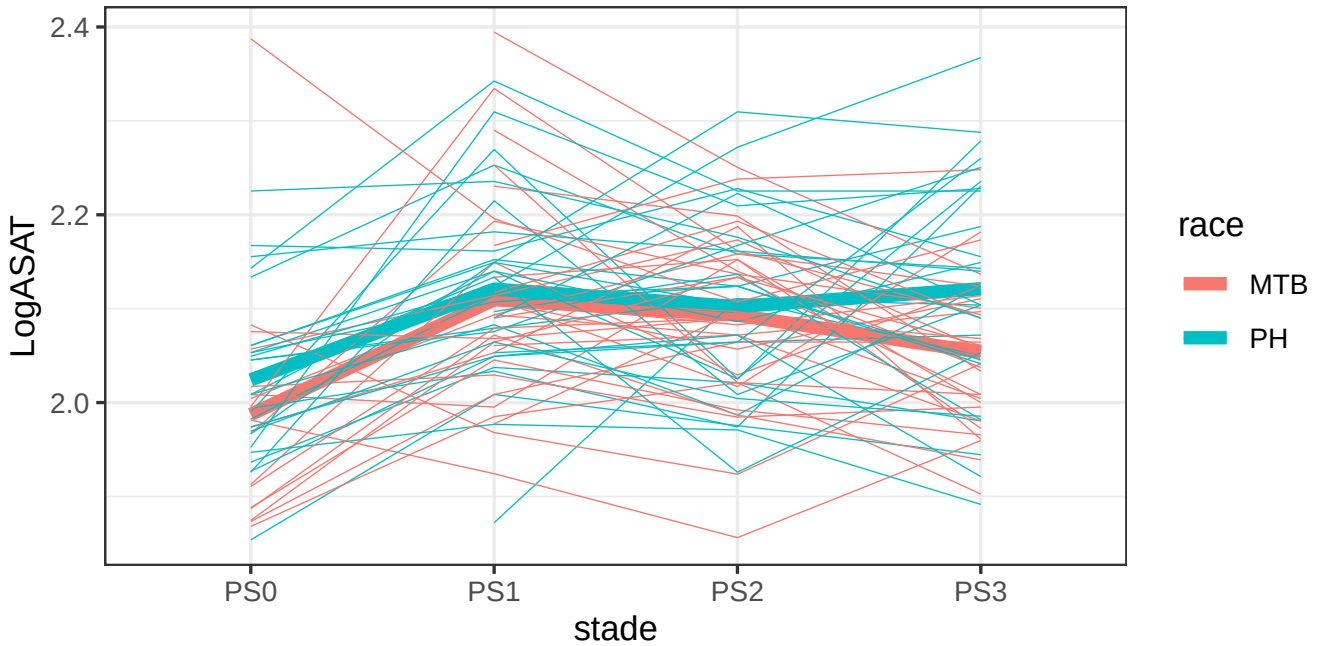
On mesure la teneur en ASAT dans le sang (aspartate aminotransférase) chez 64 vaches de deux races différentes, à quatre stades de lactation différents, PS0 (tarissement), PS1 (vêlage), PS2 (pic de lactation) et PS3 (milieu de lactation). On souhaite étudier l'effet race et l'effet stade de lactation en prenant en compte bien entendu l'effet vache (chaque vache faisant l'objet de plusieurs mesures à différents stades de lactation).
Modèle mixte avec **deux facteurs fixes croisés (stade et race)** et **un facteur aléatoire (vache)** qui est **imbriqué dans le facteur race** et **croisé avec le facteur stade**, pour lequel il joue le rôle d'un facteur bloc.

Codage des données

```
> d <- read.table("DATA/ASAT.txt", header = TRUE,
+                 stringsAsFactors = TRUE)
> str(d)
'data.frame':        238 obs. of  4 variables:
 $ vache  : Factor w/ 64 levels "V1030","V1053",..:
 $ stade  : Factor w/ 4 levels "PS0","PS1","PS2",..
 $ race   : Factor w/ 2 levels "MTB","PH": 1 1 1 1
 $ LogASAT: num  1.97 2.08 2.08 1.99 2.39 ...
```

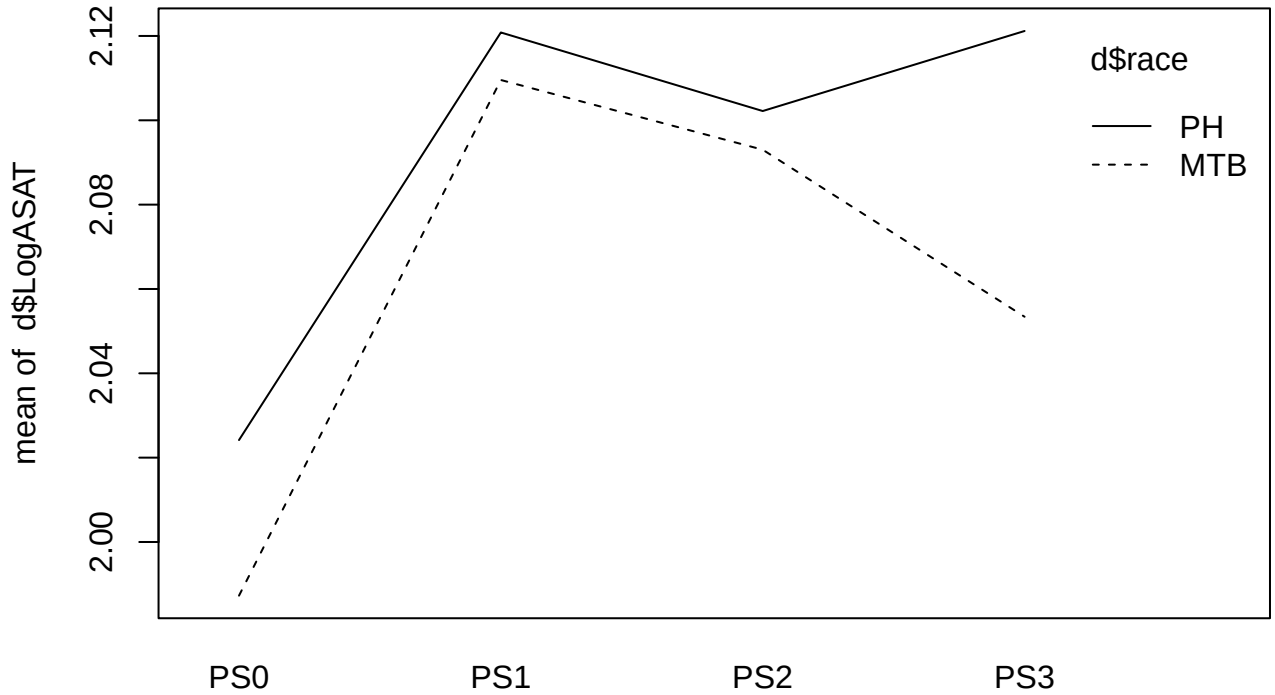
Visualisation des données avec R - tracés individuels

```
> ggplot(data = d, aes(x = stade, y = LogASAT, colour = race,
+    group = vache)) + stat_summary(aes(group = race),
+    geom = "line", lwd = 2, fun = mean) + geom_line(lwd = 0.2)
```



Visualisation des données avec R - graphe d'interaction

```
> interaction.plot(d$stade, d$race, d$LogASAT)
```



Modèle avec interaction entre la race et le stade

```
> m1 <- lmer(LogASAT ~ race + stade + race:stade + (1|vache),
> summary(m1)$coefficients
```

	Estimate	Std. Error	t value
(Intercept)	2.00094	0.0212	94.432
racePH	0.02049	0.0282	0.728
stadePS1	0.10862	0.0226	4.817
stadePS2	0.09209	0.0226	4.083
stadePS3	0.05247	0.0226	2.327
racePH:stadePS1	-0.00921	0.0303	-0.304
racePH:stadePS2	-0.01131	0.0303	-0.374
racePH:stadePS3	0.04729	0.0303	1.563

```
> summary(m1)$varcor
```

Groups	Name	Std.Dev.
vache	(Intercept)	0.0618
Residual		0.0761

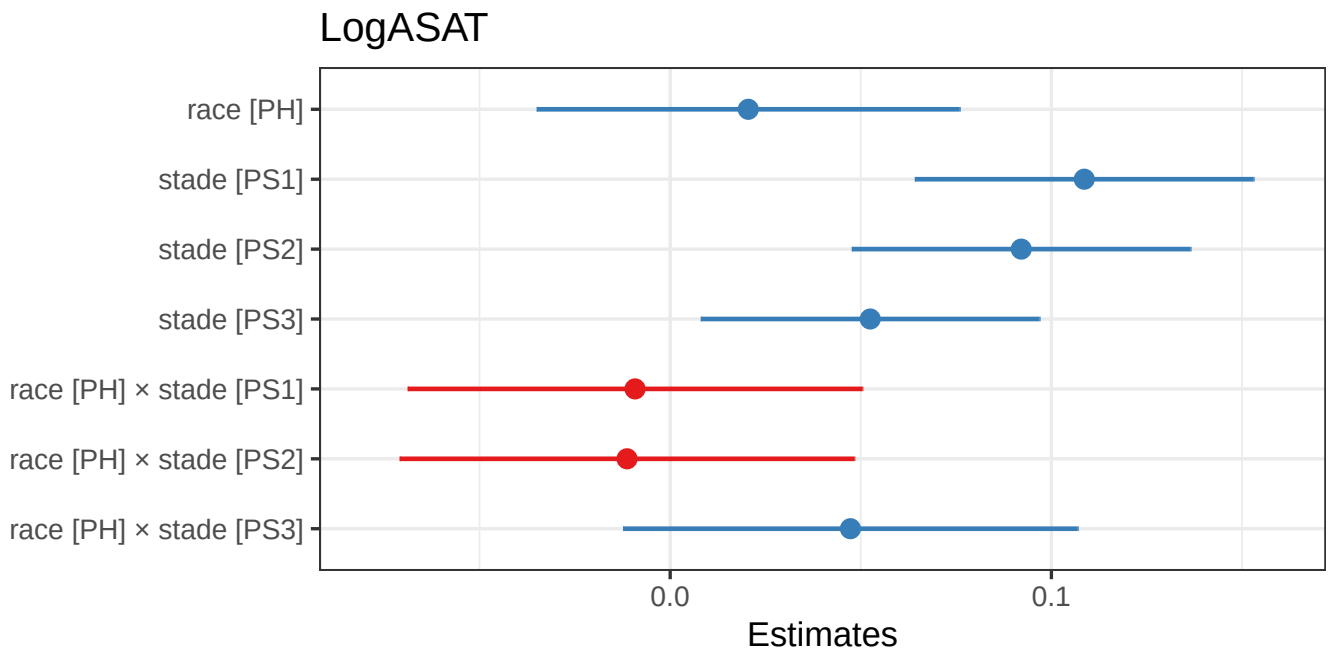
Intervalles de confiance sur les effets estimés

```
> confint(m1)
                    2.5 % 97.5 %
.sig01            0.04696 0.0777
.sigma            0.06755 0.0833
(Intercept)       1.95992 2.0421
racePH           -0.03425 0.0750
stadePS1          0.06499 0.1525
stadePS2          0.04845 0.1359
stadePS3          0.00884 0.0963
racePH:stadePS1 -0.06804 0.0493
racePH:stadePS2 -0.07014 0.0472
racePH:stadePS3 -0.01154 0.1058
```

Intervalles de confiance sur les effets fixes visualisés à l'aide du package `sjPlot`

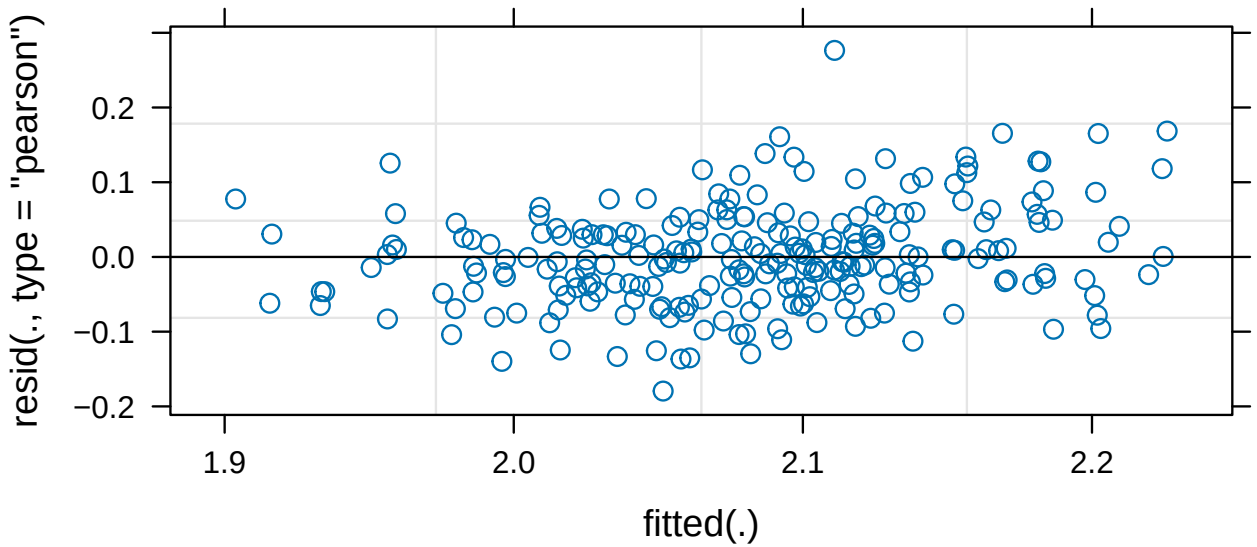
ATTENTION, les IC représentés ne sont pas calculés avec la méthode par défaut proposée par `confint()` (`"Wald"` au lieu de `"profile"`).

```
> plot_model(m1) + ylim(-0.08, 0.16)
```



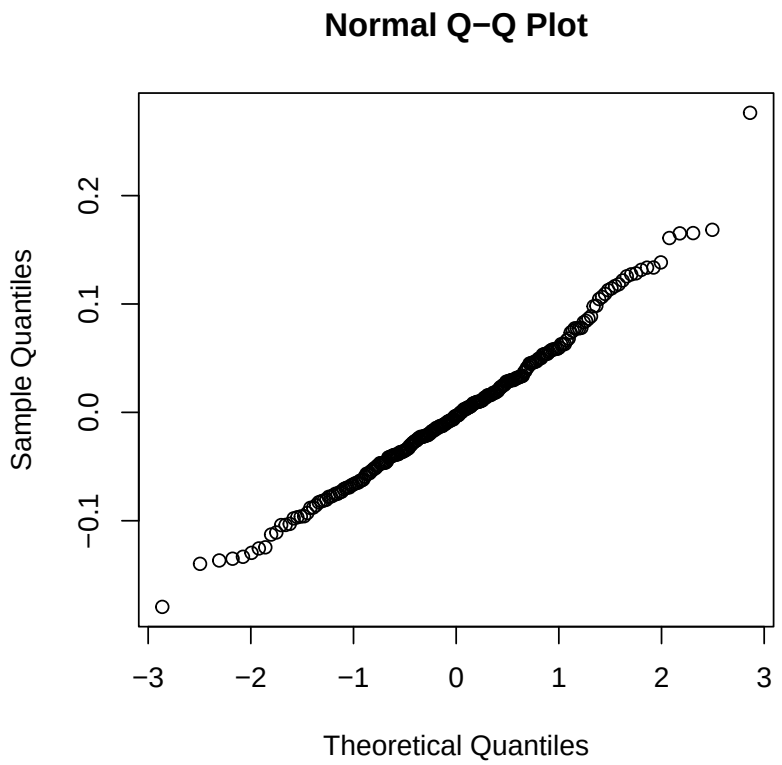
Graphe des résidus en fonction des valeurs prédites

```
> plot(m1)
```



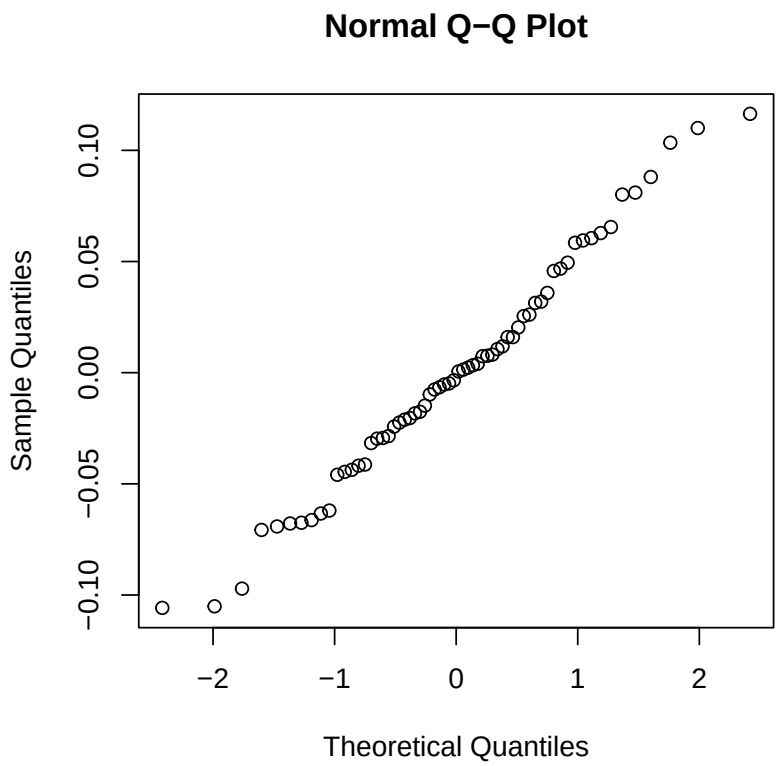
Graphe Quantile-Quantile des résidus

```
> qqnorm(residuals(m1))
```



Graphe Quantile-Quantile des effets aléatoires

```
> qqnorm(ranef(m1)$vache[,1])
```



Conclusion à partir du modèle

Même si l'on n'exclut pas un effet de la race (ASAT des Prim'holstein légèrement supérieur à l'ASAT des Montbéliardes), et une interaction entre la race et le stade (l'ASAT au stade 3 semble être plus élevé qu'au stade 2 pour les Prim'holstein et plus faible pour les Montbéliardes), le seul effet qui apparaît significatif dans cette étude est celui du stade.
Comme il ne semble pas raisonnable d'enlever l'interaction du modèle, les effets fixes sont difficiles à interpréter.
Une possibilité de simplification serait de faire un modèle par race.
Les écarts types intra-individuel et inter-individuel sont respectivement estimés à 0.076 et 0.062.

Exemple de modèle sur la race Prim'holstein

```
> mPH <- lmer(LogASAT ~ stade + (1|vache),
+             data = d[d$race == "PH", ])
> summary(mPH)$coefficients
```

	Estimate	Std. Error	t value
(Intercept)	2.0213	0.0189	106.75
stadePS1	0.0996	0.0200	4.98
stadePS2	0.0809	0.0200	4.05
stadePS3	0.0999	0.0200	5.00

```
> summary(mPH)$varcor
```

Groups	Name	Std.Dev.
vache	(Intercept)	0.0663
Residual		0.0754

Les incontournables à mettre dans un article

- **Statistical analysis**
  - use of a **linear mixed-effect model** (with the lme4 R package)
  - **fixed effects** taken into account (and wether each is considered continuous or categorical if it is not obvious)
  - **interactions** included in the model
  - **random effects** taken into account

  Mieux vaut éviter de mettre une pseudo formule de code R dans la première partie (pas très explicite) !
- **Results**

  A table with **estimated effects** (with confidence intervals) and/or a plot of those estimations.

  Mieux vaut éviter les tables d'ANOVA à l'ancienne avec p-value par effet fixe ! (souvent fausses et pas très informatives)

Conclusion

L'exemple précédent n'est qu'un exemple parmi les multiples possibilités de modèles mixtes à trois facteurs.
Vous avez toutes les briques de légo pour traiter divers exemples.
A vous de jouer avec ces briques de légo pour construire des modèles mixtes qui tiennent la route !

La référence associée au package lme4 :
Douglas Bates, Martin Maechler, Ben Bolker, Steve Walker (2015). Fitting Linear Mixed-Effects Models Using lme4. Journal of Statistical Software, 67(1), 1-48.
`https://www.jstatsoft.org/article/view/v067i01`

Une référence pour aller plus loin

Harrison, X. A., Donaldson, L., Correa-Cano, M. E., Evans, J., Fisher, D. N., Goodwin, C. E., Inger, R. (2018). A brief introduction to mixed effects modelling and multi-model inference in ecology. PeerJ, 6, e4794.
https://peerj.com/articles/4794/