

S8_B2_Variabilités interindividuelles

Table of contents

1	Contexte : données hiérarchiques	1
2	Variabilité inter-individuelle	2
2.A	Définition intuitive	2
2.B	Interprétation	2
3	Variabilité inter-centre	3
3.A	Idée de base	3
3.B	Coefficient de corrélation intraclasse (ICC)	3
4	Modèle linéaire mixte : séparer les deux variabilités	4
4.A	Modèle avec intercept aléatoire de centre	4
4.B	Implications pratiques	4
5	Pourquoi on ne retrouve pas ces variances avec des modèles séparés	4
5.A	Modèles “par centre”	4
5.B	Modèles “par patient” (données longitudinales)	5
6	Illustration rapide en R	6
6.A	Simulation avec centres	6
6.B	Modèle mixte	7
6.C	Modèles séparés par centre	8
6.D	Résumé	10

1 Contexte : données hiérarchiques

On considère un schéma très courant :

- des patients indexés par $i = 1, \dots, n_j$,
- regroupés dans des centres (hôpitaux, services) indexés par $j = 1, \dots, J$,
- pour chaque patient on observe une réponse Y_{ij} (par exemple : score, biomarqueur, indicateur binaire),
- et des covariables X_{ij} (âge, sexe, gravité, traitement, etc.).

Si on ignore la structure en centres, on pourrait écrire un modèle linéaire simple :

$$Y_{ij} = \beta_0 + \beta_1 X_{ij} + \varepsilon_{ij}.$$

Mais en pratique :

- les patients diffèrent entre eux (même à X_{ij} égal),
- les centres diffèrent entre eux (même profil de patients, pratiques différentes, etc.).

Les modèles linéaires mixtes servent à séparer ces sources de variabilité.

Plan :

1. Définir la variabilité inter-individuelle.
2. Définir les variations inter-centre.
3. Voir comment un modèle mixte les représente.
4. Expliquer pourquoi on ne peut pas estimer ces variances avec des modèles séparés par centre / par patient.
5. Illustrer avec un petit exemple en R. □

2 Variabilité inter-individuelle

2.A Définition intuitive

Même dans un centre donné et pour les mêmes covariables X_{ij} , deux patients n'auront pas exactement la même valeur de Y_{ij} .

Cette dispersion entre patients, une fois qu'on a pris en compte les covariables, est la variabilité inter-individuelle.

Dans un modèle linéaire simple :

$$Y_{ij} = \beta_0 + \beta_1 X_{ij} + \varepsilon_{ij},$$

on suppose souvent :

$$\varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2),$$

où :

- ε_{ij} est l'erreur individuelle,
- σ^2 est la variance inter-individuelle résiduelle : à covariables égales (et dans un même centre si on l'a inclus), elle mesure à quel point les patients diffèrent entre eux.

2.B Interprétation

On peut lire σ^2 comme :

- la “dispersion” des résultats individuels autour de ce que le modèle prédit,
- ce qui reste comme variabilité entre patients après avoir ajusté sur les covariables et, le cas échéant, les effets de centre.

□

3 Variabilité inter-centre

3.A Idée de base

Même si l'on ajuste sur les mêmes covariables X_{ij} , deux centres peuvent :

- prendre en charge les patients différemment,
- avoir des équipes ou des ressources différentes,
- avoir des populations de patients légèrement différentes.

Donc, à covariables égales, la moyenne de Y_{ij} peut différer d'un centre à l'autre.

On introduit alors un effet aléatoire de centre u_j :

$$Y_{ij} = \beta_0 + u_j + \beta_1 X_{ij} + \varepsilon_{ij}.$$

Avec un modèle à effets aléatoires, on suppose typiquement :

$$u_j \sim \mathcal{N}(0, \tau^2), \quad \varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2),$$

et u_j indépendant de ε_{ij} .

- τ^2 : variance inter-centre (variabilité entre les centres),
- σ^2 : variance inter-individuelle résiduelle (variabilité entre patients à l'intérieur d'un centre).

Au total :

$$\text{Var}(Y_{ij} \mid X_{ij}) = \tau^2 + \sigma^2.$$

3.B Coefficient de corrélation intraclasse (ICC)

Une quantité clé est l'ICC (intra-class correlation) :

$$\rho = \frac{\tau^2}{\tau^2 + \sigma^2}.$$

- Si ρ est proche de 0 : la variabilité est surtout inter-individuelle, les centres se ressemblent.
- Si ρ est élevé : les différences entre centres expliquent une proportion importante de la variance totale.

□

4 Modèle linéaire mixte : séparer les deux variabilités

4.A Modèle avec intercept aléatoire de centre

On résume :

$$Y_{ij} = \beta_0 + u_j + \beta_1 X_{ij} + \varepsilon_{ij},$$

avec :

- $u_j \sim \mathcal{N}(0, \tau^2)$: effet aléatoire de centre,
- $\varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$: erreur résiduelle individuelle.

Le modèle donne :

- une estimation de β_1 (effet moyen de la covariable),
- une estimation de τ^2 (variabilité inter-centre),
- une estimation de σ^2 (variabilité inter-individuelle).

On obtient donc deux niveaux de variabilité clairement séparés.

4.B Implications pratiques

- Meilleure estimation des erreurs standards de β_1 (on ne fait pas “comme si” tout était indépendant).
- Possibilité de quantifier :
 - à quel point les centres diffèrent entre eux (via τ^2 et ρ),
 - à quel point les patients diffèrent au sein d’un centre (via σ^2).
 - On peut généraliser à d’autres centres (on modélise une distribution des centres, pas seulement ceux observés).

□

5 Pourquoi on ne retrouve pas ces variances avec des modèles séparés

5.A Modèles “par centre”

Supposons qu’au lieu d’un modèle mixte unique, on ajuste un modèle linéaire séparé dans chaque centre :

$$Y_{ij} = \beta_{0j} + \beta_{1j} X_{ij} + \varepsilon_{ij}^{(j)}, \quad j = 1, \dots, J.$$

On obtient alors une collection de paramètres :

- $\hat{\beta}_{0j}$,

- $\hat{\beta}_{1j}$,

mais il n'y a pas de paramètre τ^2 dans un modèle global.

On pourrait regarder la variance empirique des $\hat{\beta}_{0j}$ entre centres, mais cette variance mélange :

- la vraie variabilité inter-centre,
- la variabilité d'estimation (certains centres ont peu de patients, donc $\hat{\beta}_{0j}$ est très instable).

Sans modèle commun qui impose :

$$\beta_{0j} = \beta_0 + u_j, \quad u_j \sim \mathcal{N}(0, \tau^2),$$

il est impossible de séparer proprement :

- “ce qui vient vraiment de différences entre centres” (τ^2),
- de “ce qui vient du bruit d'échantillonnage” sur chaque $\hat{\beta}_{0j}$.

Donc :

- dans des modèles séparés par centre, on n'a pas de paramètre explicite pour la variance inter-centre,
- on ne peut pas estimer τ^2 ni un ICC propre,
- on perd la structure hiérarchique globale.

5.B Modèles “par patient” (données longitudinales)

Même idée si l'unité d'analyse est la visite et que l'on a plusieurs visites par patient.

Un modèle mixte pour des données longitudinales pourrait être :

$$Y_{ij} = \beta_0 + b_i + \beta_1 t_{ij} + \varepsilon_{ij},$$

avec :

- $b_i \sim \mathcal{N}(0, \omega^2)$: effet aléatoire patient (niveau moyen propre à chaque patient),
- $\varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$: bruit intra-patient.

Alors :

- ω^2 = variabilité inter-individuelle (entre patients),
- σ^2 = variabilité intra-individuelle (au sein d'un même patient).

Si, au lieu de ça, on ajuste un modèle séparé pour chaque patient :

$$Y_{ij} = \beta_{0i} + \beta_{1i} t_{ij} + \varepsilon_{ij}^{(i)},$$

on obtient une collection de pentes et d'intercepts :

- $\hat{\beta}_{0i}, \hat{\beta}_{1i}$,

mais :

- pas de paramètre ω^2 pour décrire la distribution des effets patients,
- la dispersion des $\hat{\beta}_{0i}$ et $\hat{\beta}_{1i}$ est un mélange de :
 - vraie variabilité entre patients,
 - bruit d'estimation (surtout quand chaque patient a peu de mesures).

On ne peut donc pas :

- décomposer proprement la variance entre et au sein des patients,
- prédire pour un nouveau patient à partir d'une distribution $b_i \sim \mathcal{N}(0, \omega^2)$,
- bénéficier du shrinkage (effets des petits échantillons ramenés vers la moyenne).

□

6 Illustration rapide en R

6.A Simulation avec centres

On simule des données avec :

- $J = 10$ centres,
- $n_j = 50$ patients par centre,
- un effet centre u_j de variance τ^2 ,
- un bruit individuel ε_{ij} de variance σ^2 .

```
J <- 10
n_per_center <- 50

center <- rep(1:J, each = n_per_center)

# Covariable individuelle
x <- rnorm(J * n_per_center, mean = 0, sd = 1)

# Paramètres "vrais"
beta0 <- 2
beta1 <- 1
tau <- 1 # écart-type inter-centre
sigma <- 2 # écart-type inter-individuel

# Effets aléatoires de centre
u <- rnorm(J, mean = 0, sd = tau)
u_center <- u[center]
```

```
# Bruits individuels
eps <- rnorm(J * n_per_center, mean = 0, sd = sigma)

# Réponse
y <- beta0 + u_center + beta1 * x + eps

dat <- data.frame(
  y = y,
  x = x,
  center = factor(center)
)

head(dat)
```

	y	x	center
1	0.6428930	-0.56047565	1
2	1.6002347	-0.23017749	1
3	4.7217458	1.55870831	1
4	1.8798106	0.07050839	1
5	0.2945232	0.12928774	1
6	1.6435736	1.71506499	1

6.B Modèle mixte

On ajuste un modèle mixte avec intercept aléatoire de centre : (c'est à dire un effet aléatoire sur l'intercept, qui correspond aux différences de niveau moyen entre centres)

```
mod_mixed <- lmer(y ~ x + (1 | center), data = dat)
summary(mod_mixed)
```

```
Linear mixed model fit by REML ['lmerMod']
Formula: y ~ x + (1 | center)
Data: dat
```

REML criterion at convergence: 2148.5

Scaled residuals:

Min	1Q	Median	3Q	Max
-2.64638	-0.63464	0.02181	0.67665	2.72271

Random effects:

Groups	Name	Variance	Std.Dev.
center	(Intercept)	0.5698	0.7548
Residual		4.1129	2.0280

Number of obs: 500, groups: center, 10

Fixed effects:

	Estimate	Std. Error	t value
(Intercept)	1.57420	0.25537	6.164
x	1.10000	0.09432	11.663

Correlation of Fixed Effects:

(Intr)
x -0.013

Dans la partie Random effects, on retrouve :

- une estimation de τ^2 (variance inter-centre),
- une estimation de σ^2 (variance résiduelle).

On peut donc en déduire un ICC estimé :

```
var_comp <- as.data.frame(VarCorr(mod_mixed))
tau2_hat <- var_comp$vcov[var_comp$grp == "center"]
sigma2_hat <- var_comp$vcov[var_comp$grp == "Residual"]

icc_hat <- tau2_hat / (tau2_hat + sigma2_hat)
icc_hat
```

```
[1] 0.1216758
```

6.C Modèles séparés par centre

Si on fait un modèle séparé par centre :

```
coefs_by_center <- lapply(split(dat, dat$center), function(dd) {
  coef(lm(y ~ x, data = dd))
})

coefs_by_center
```

```
$`1`
(Intercept)          x
  1.626168    1.200121
```

```
$`2`
(Intercept)          x
  0.6134143    1.2447135
```

```
$`3`
(Intercept)          x
  2.811325    1.029217
```

```
$`4`
```

```
(Intercept)      x
      2.360334    1.266255
```

\$`5`

```
(Intercept)      x
      1.064226    1.149654
```

\$`6`

```
(Intercept)      x
      2.0856707   0.8961968
```

\$`7`

```
(Intercept)      x
      0.9985427   0.9285005
```

\$`8`

```
(Intercept)      x
      0.3389266   1.3435537
```

\$`9`

```
(Intercept)      x
      2.1766631   0.7207045
```

\$`10`

```
(Intercept)      x
      1.598613    1.330416
```

On obtient :

- une liste d'intercepts et de pentes par centre,
- mais aucun paramètre global pour la variance inter-centre.

La variance empirique des intercepts estimés :

```
betas0 <- sapply(coefs_by_center, function(b) b[1])
var(betas0)
```

```
[1] 0.6462605
```

mélange :

- la vraie dispersion entre centres,
- et la variabilité d'estimation due au fait que chaque centre a une taille finie (n_j).

Il n'y a pas, ici, d'équivalent direct de τ^2 et de l'ICC estimés proprement par le modèle mixte.

□

6.D Résumé

- La variabilité inter-individuelle correspond à la dispersion entre patients après ajustement sur les covariables (variance σ^2).
- Les variations inter-centre correspondent aux différences de niveau moyen (ou de pente) entre centres (variance τ^2).
- Un modèle linéaire mixte permet de :
 - séparer ces deux composantes τ^2 et σ^2 ,
 - calculer un ICC $\rho = \tau^2 / (\tau^2 + \sigma^2)$,
 - ajuster correctement les erreurs standards des effets fixes.
- Si l'on ajuste des modèles séparés par centre ou par patient :
 - on n'a plus de paramètre global τ^2 (ou ω^2),
 - la variabilité entre coefficients estimés mélange vraie variabilité et bruit d'estimation,
 - on ne peut pas quantifier proprement la variabilité inter-centre / inter-individuelle ni la réinjecter dans un cadre prédictif.

Les modèles mixtes sont justement construits pour modéliser explicitement ces variabilités, au lieu de les laisser “perdues” dans des modèles séparés.