

STATISTIQUES INFÉRENTIELLES RÉGRESSION LINÉAIRE À EFFETS MIXTES

Guillaume Pech & Audrey Mazancieux

PLAN

PARTIE I

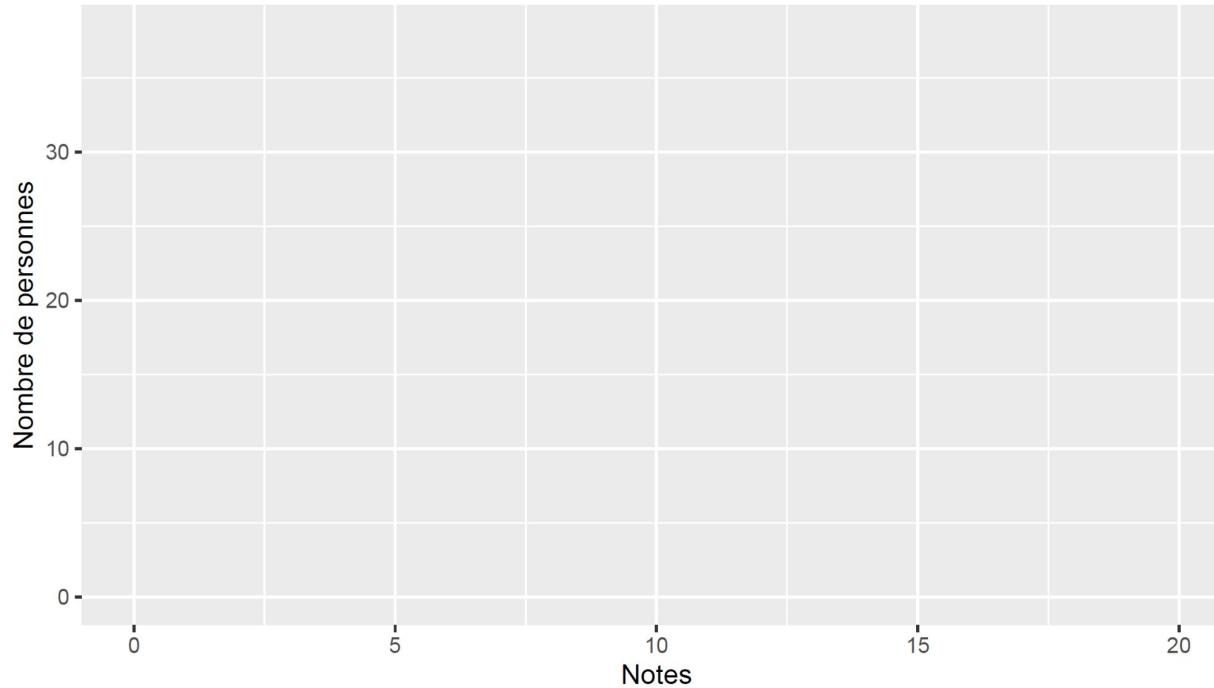
- C'est quoi des données?
- Pourquoi des statistiques inférentielles?
- Régression simple: variable catégorielle
- Régression simple: variable continue
- Conditions d'applications
- Pourquoi utiliser des modèles mixtes

PARTIE II

- Rappel de la partie I
- Modèles mixtes : modéliser les erreurs
 - Pour chaque participant
 - Variabilité de l'intercept
 - Variabilité de la pente
 - Un modèle général : formalisation
- Exemples pratiques

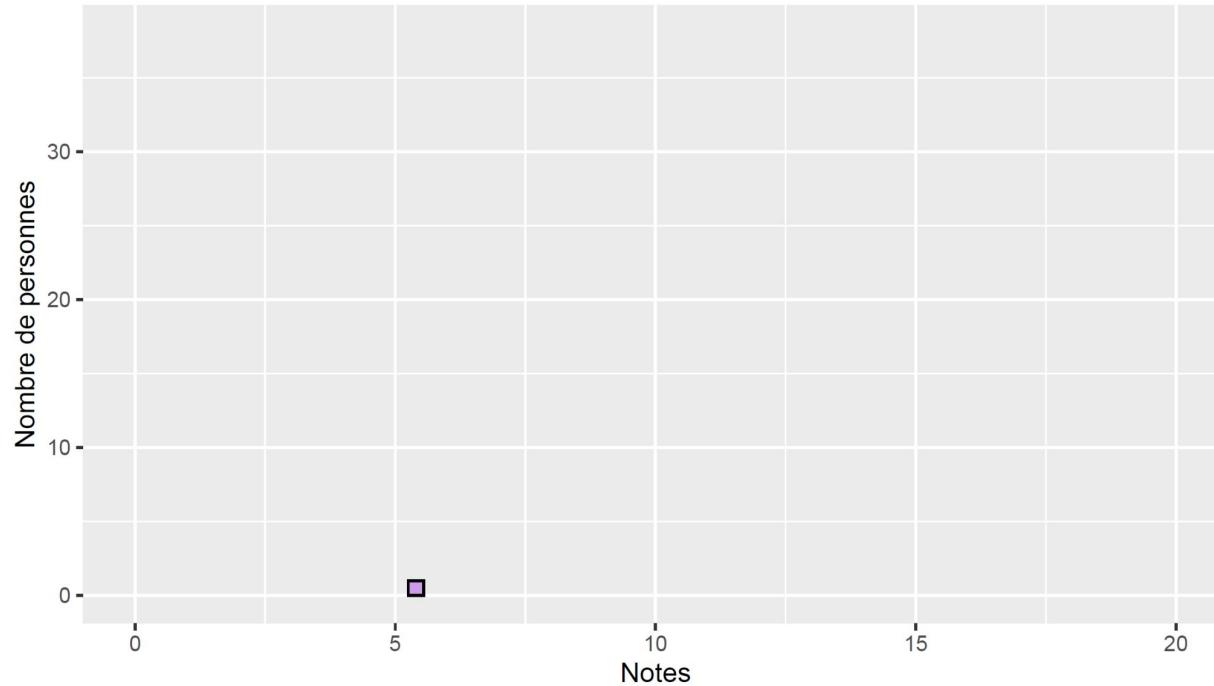
C'EST QUOI DES DONNÉES?

Exemple: des notes à un examen (entre 0 et 20) pour 500 étudiant.es



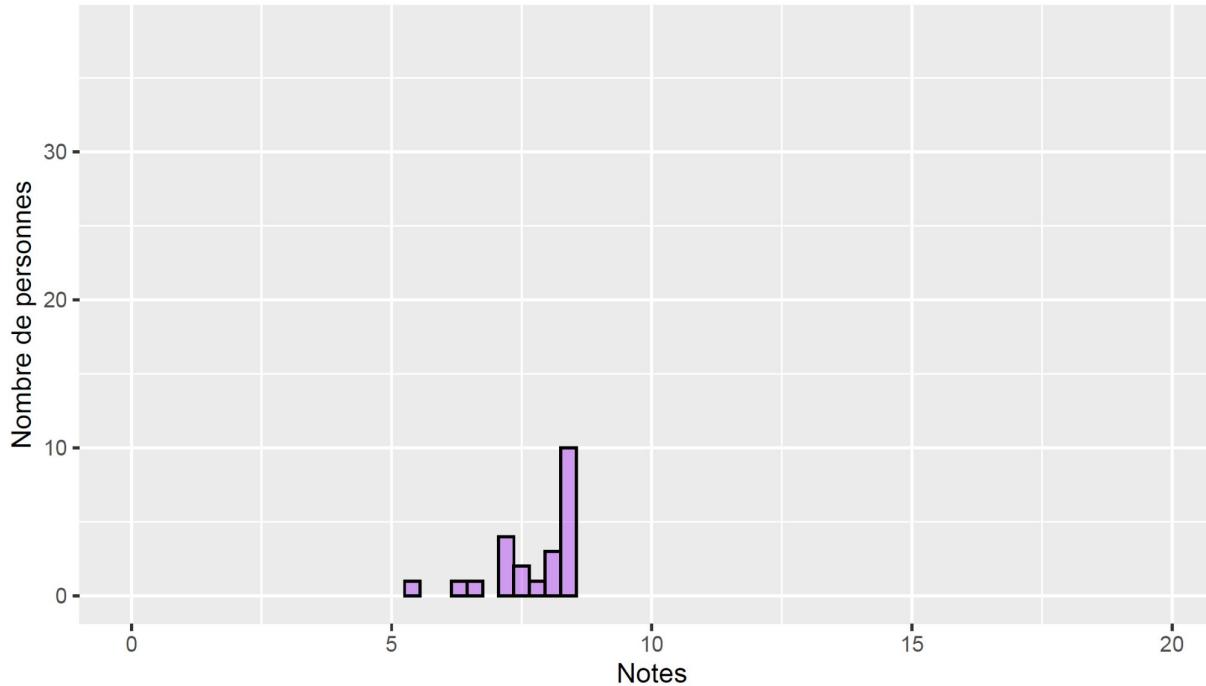
C'EST QUOI DES DONNÉES?

Exemple: des notes à un examen (entre 0 et 20) pour 500 étudiant.es



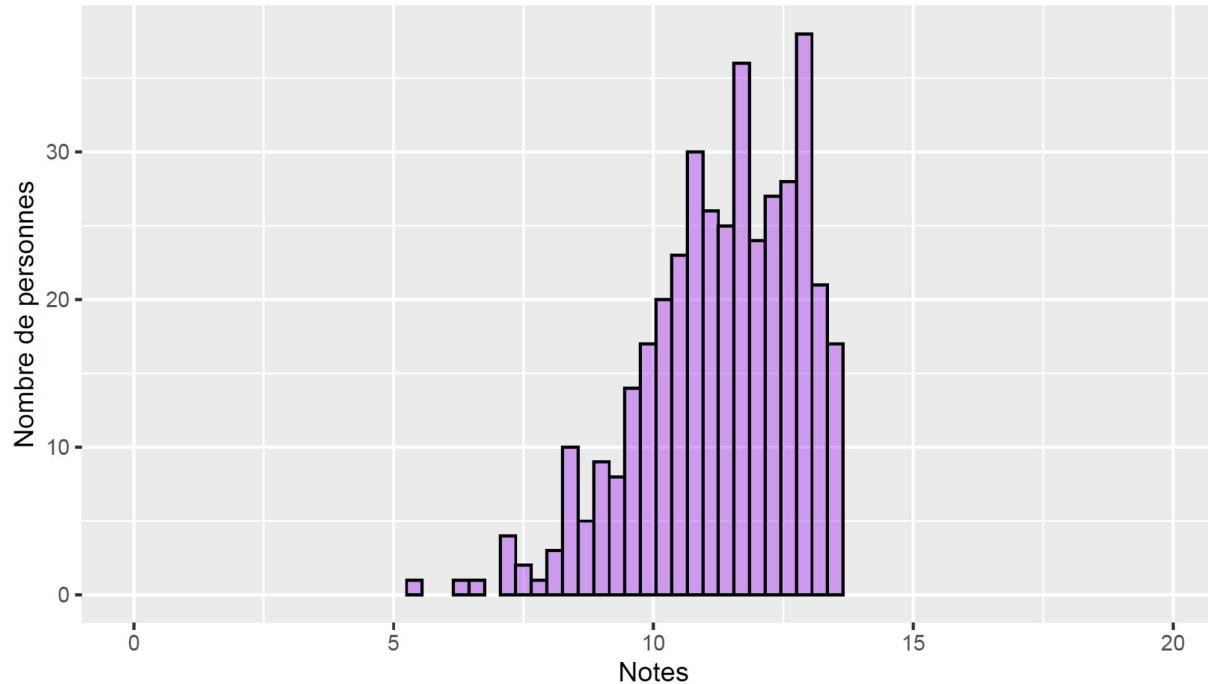
C'EST QUOI DES DONNÉES?

Exemple: des notes à un examen (entre 0 et 20) pour 500 étudiant.es



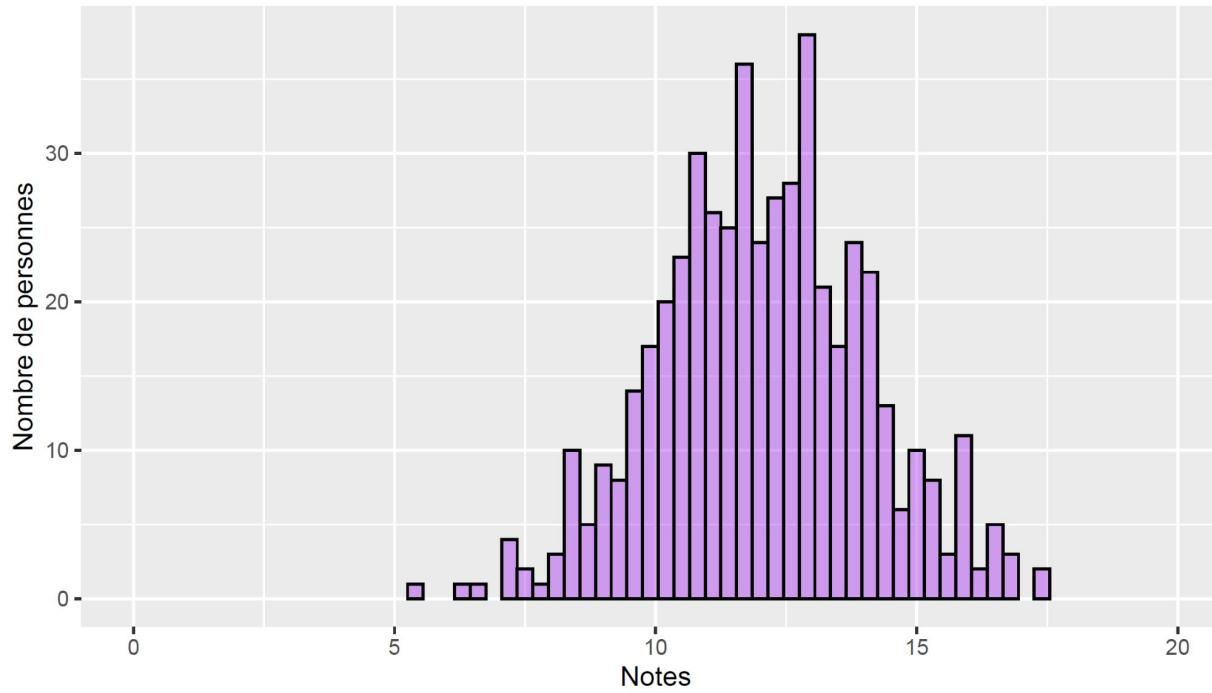
C'EST QUOI DES DONNÉES?

Exemple: des notes à un examen (entre 0 et 20) pour 500 étudiant.es



C'EST QUOI DES DONNÉES?

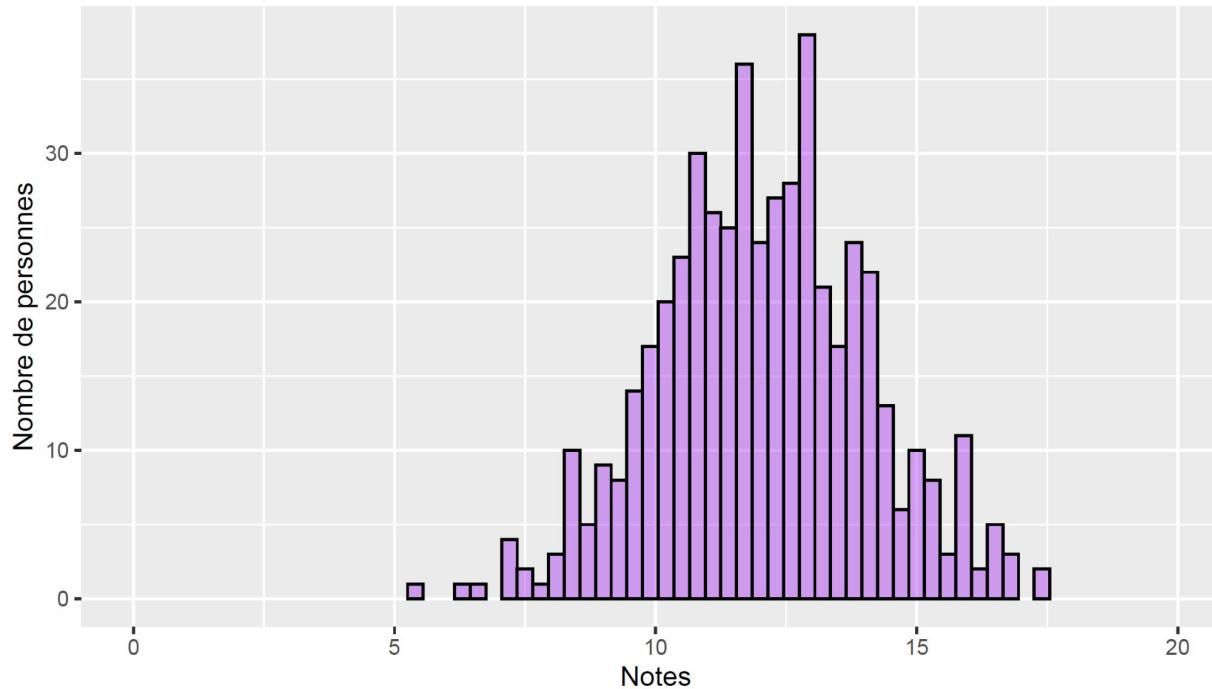
Exemple: des notes à un examen (entre 0 et 20) pour 500 étudiant.es



C'EST QUOI DES DONNÉES?

Exemple: des notes à un examen (entre 0 et 20) pour 500 étudiant.es

C'est une loi
normale !



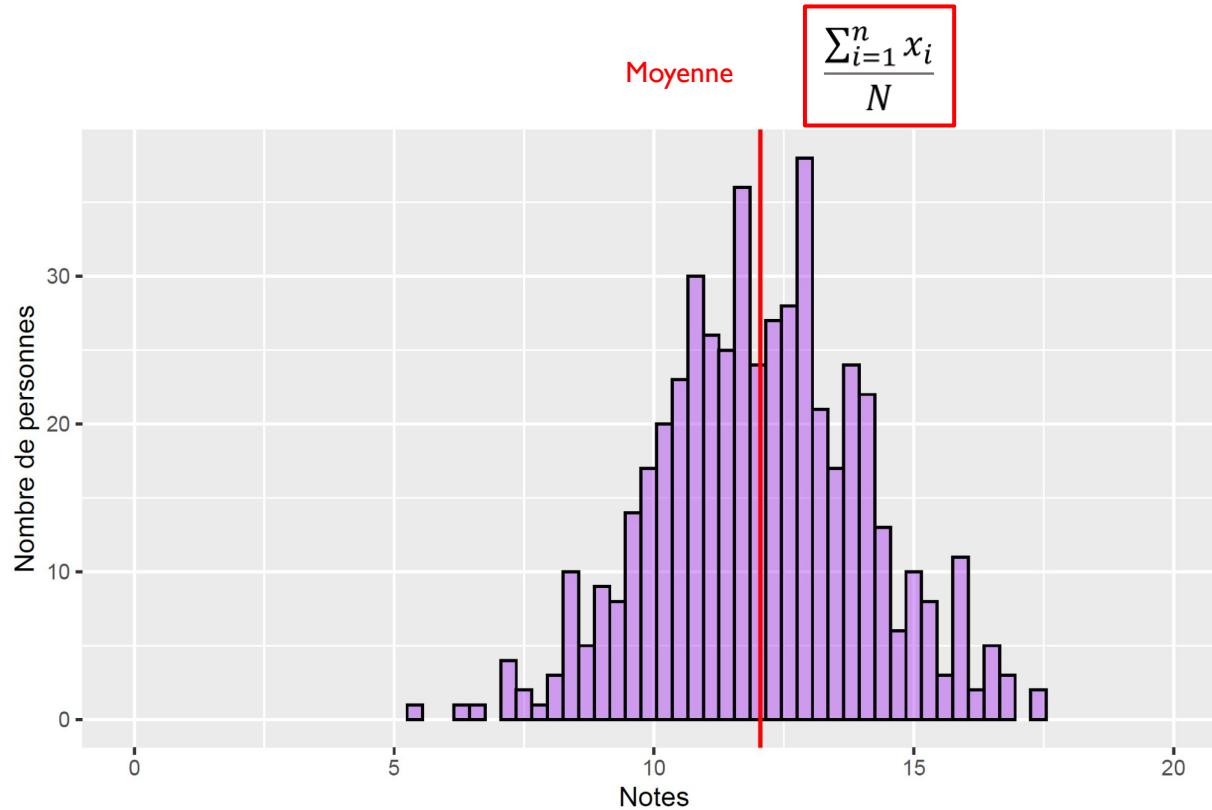
C'EST QUOI UNE LOI NORMALE?

Pourquoi dit-on ‘normale’?

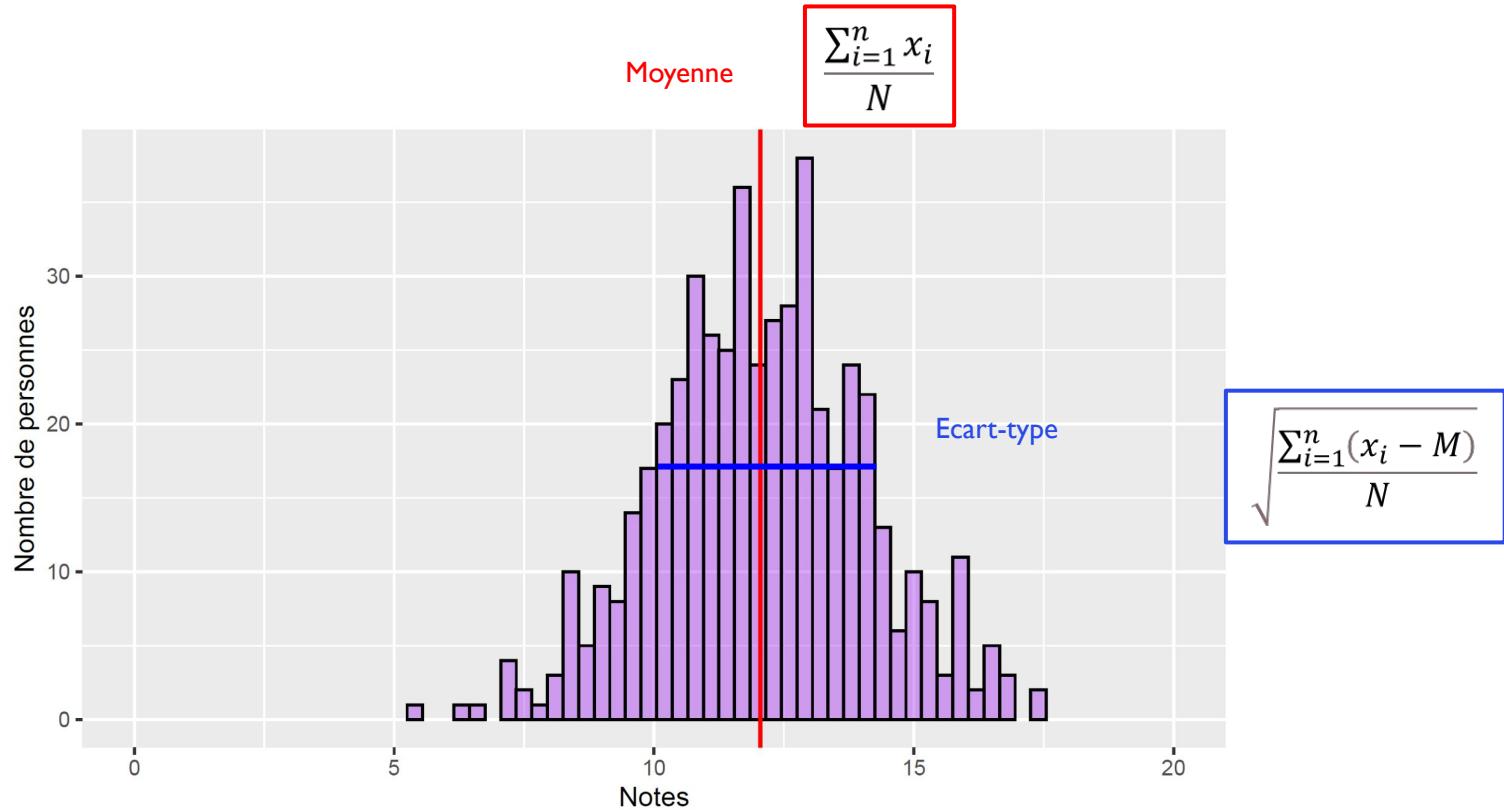
- Pas dans le sens de ‘correct’ dans le sens de ‘référence’. Elle est devenue une référence pour l’analyse des données dans diverses disciplines
- Représente des phénomènes naturels qui tendent à se distribuer selon cette courbe en cloche

La loi normale est facile à résumer avec uniquement 2 indices.

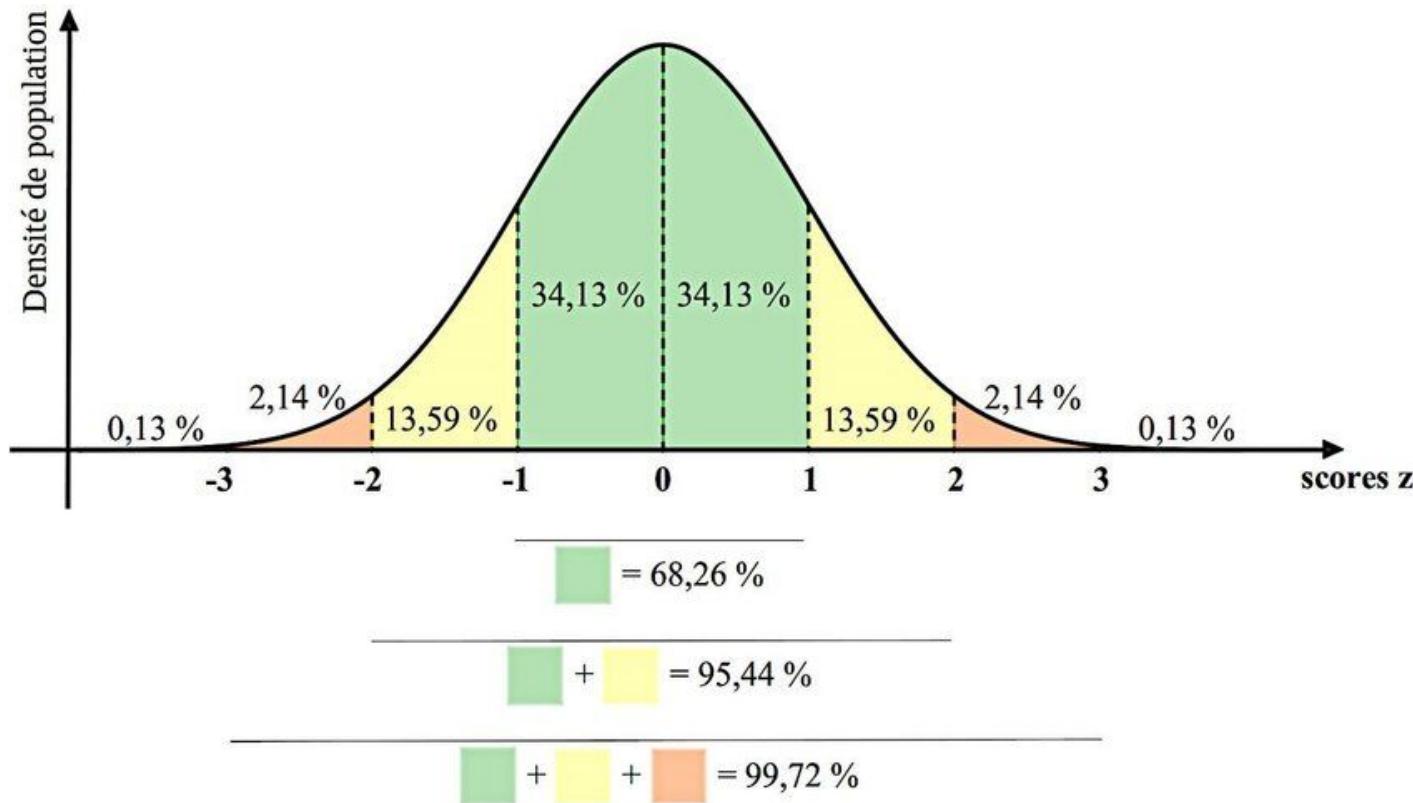
C'EST QUOI UNE LOI NORMALE?



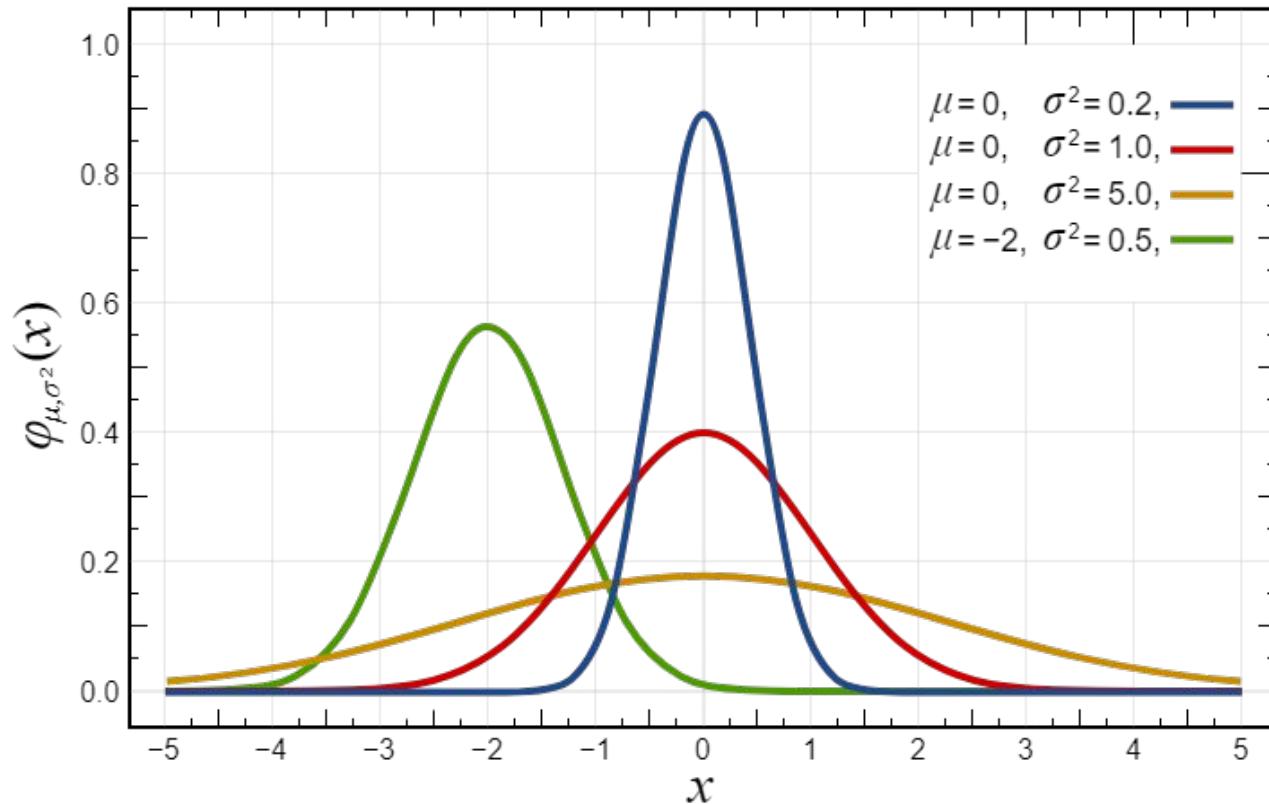
C'EST QUOI UNE LOI NORMALE?



C'EST QUOI UNE LOI NORMALE?



C'EST QUOI UNE LOI NORMALE?



GÉNÉRER DES DONNÉES

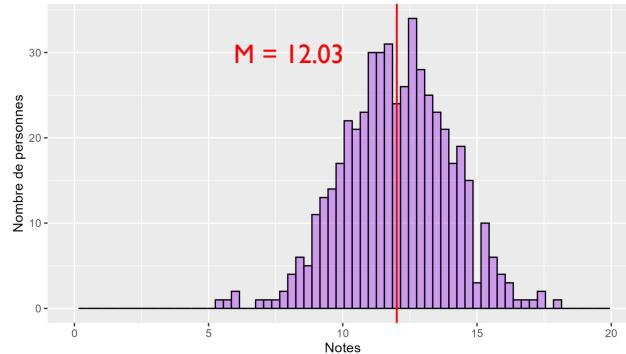
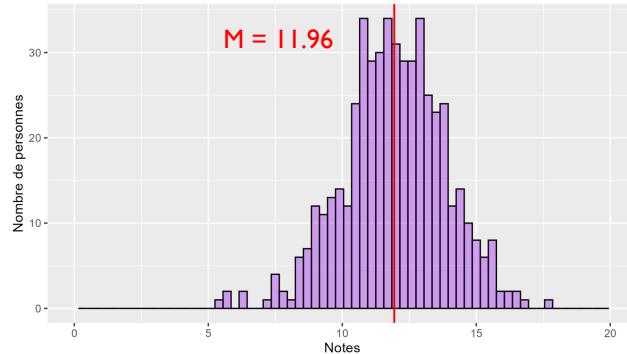
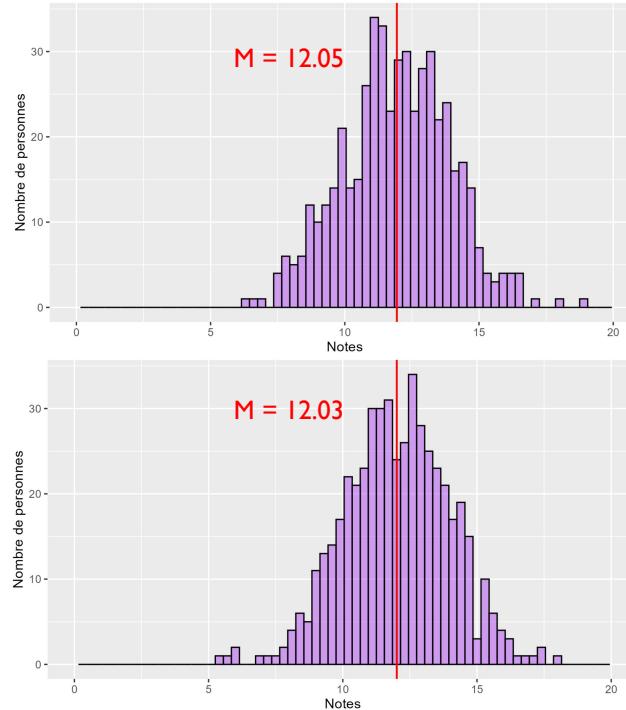
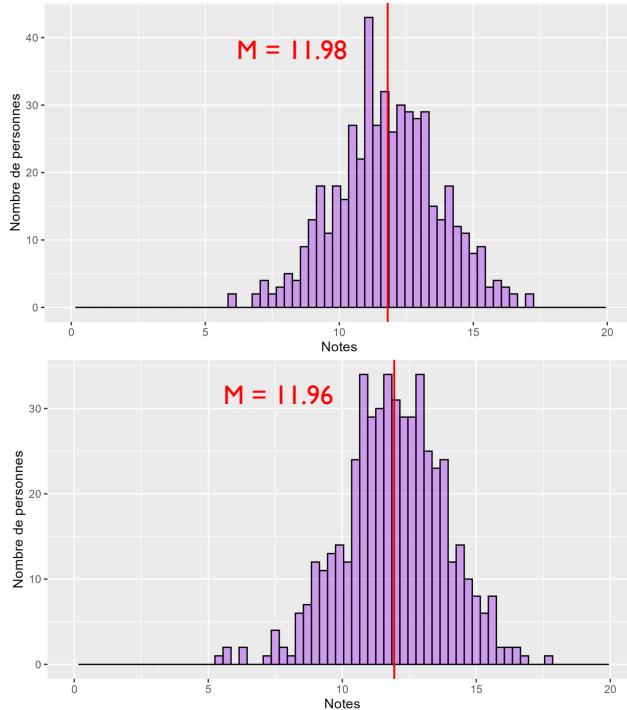
Moyenne = 12 et écart-type = 2

Pour 500 participants

GÉNÉRER DES DONNÉES

Moyenne = 12 et écart-type = 2

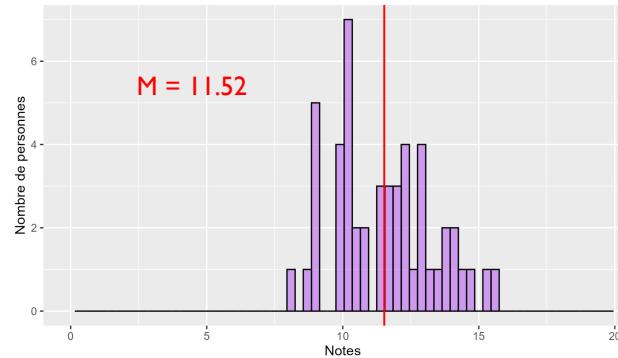
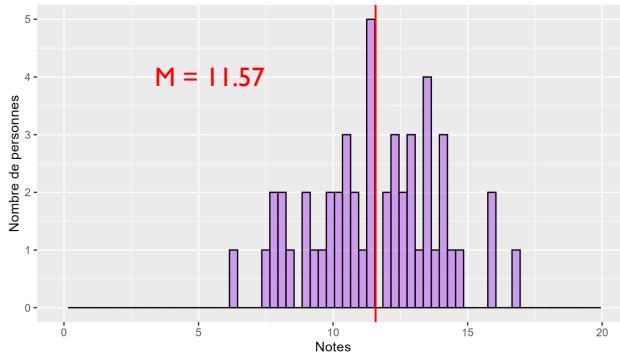
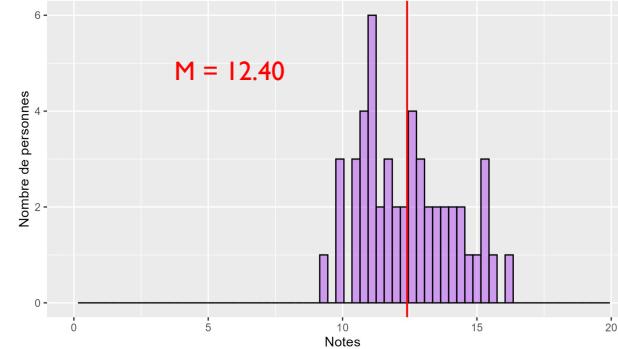
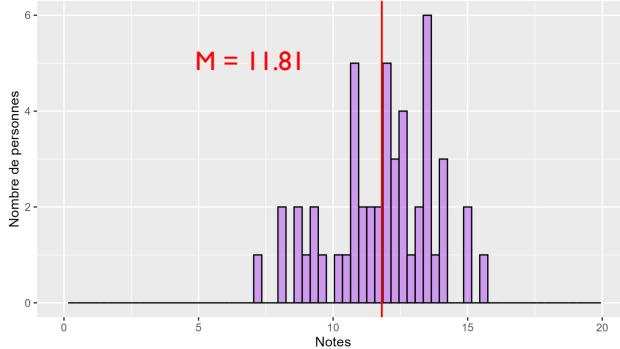
Pour 500 participants



GÉNÉRER DES DONNÉES

Moyenne = 12 et écart-type = 2

Pour 50 participants



DES STATISTIQUES INFÉRENTIELLES?

Objectif: prédire (le comportement)

Faire la démarche inverse, observer des données pour voir ce qui les a générées.

Pour cela nous créons un modèle, c'est à dire que nous simplifions nos données pour essayer de voir quelle(s) règle(s) ont générées nos données.

DES STATISTIQUES INFÉRENTIELLES?

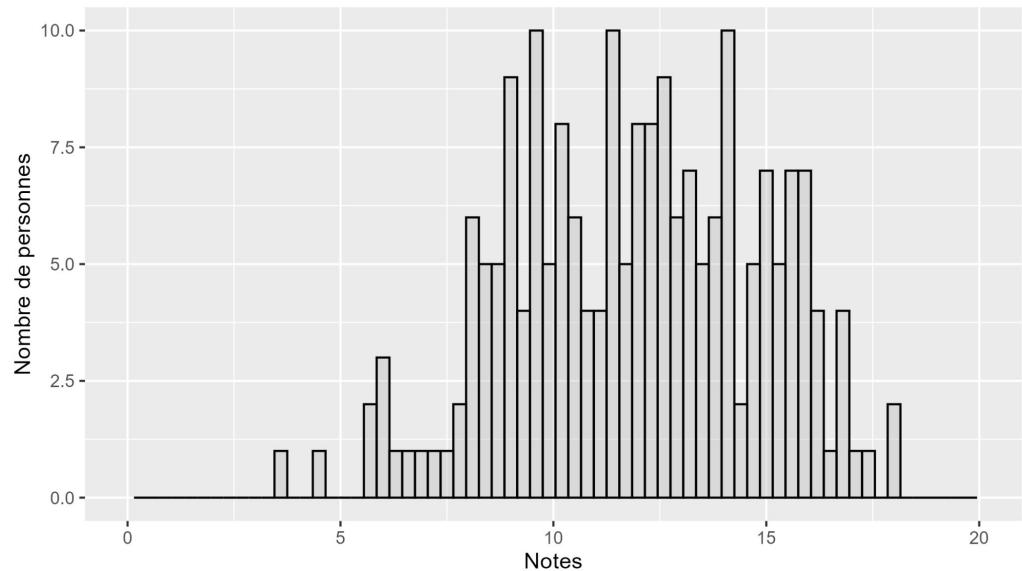
Objectif: prédire (le comportement)

Faire la démarche inverse, observer des données pour voir ce qui les a générées.

Pour cela nous créons un modèle, c'est à dire que nous simplifions nos données pour essayer de voir quelle(s) règle(s) ont générées nos données.

N = 200

Quelles(s) variable(s) peut expliquer ces données?



DES STATISTIQUES INFÉRENTIELLES?

Objectif: prédire (le comportement)

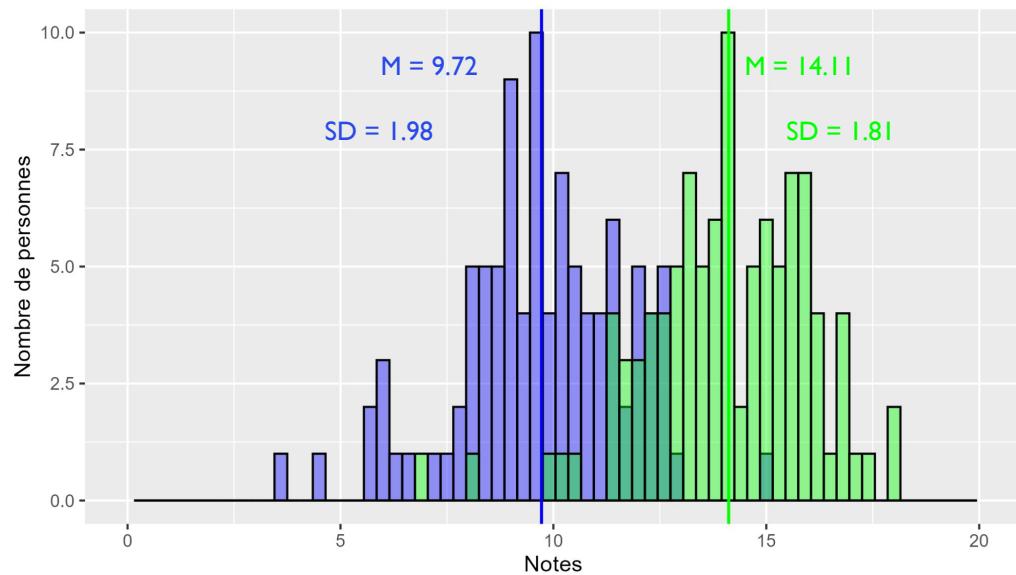
Faire la démarche inverse, observer des données pour voir ce qui les a générées.

Pour cela nous créons un modèle, c'est à dire que nous simplifions nos données pour essayer de voir quelle(s) règle(s) ont générées nos données.

$N = 200$

Dans ce groupe, il y a des gens qui aiment (vert) et des gens qui n'aiment (bleu) pas les statistiques

Notre modèle postule que ces données ont été générées par 2 distributions normales que l'on résume avec leurs moyennes et écarts types.

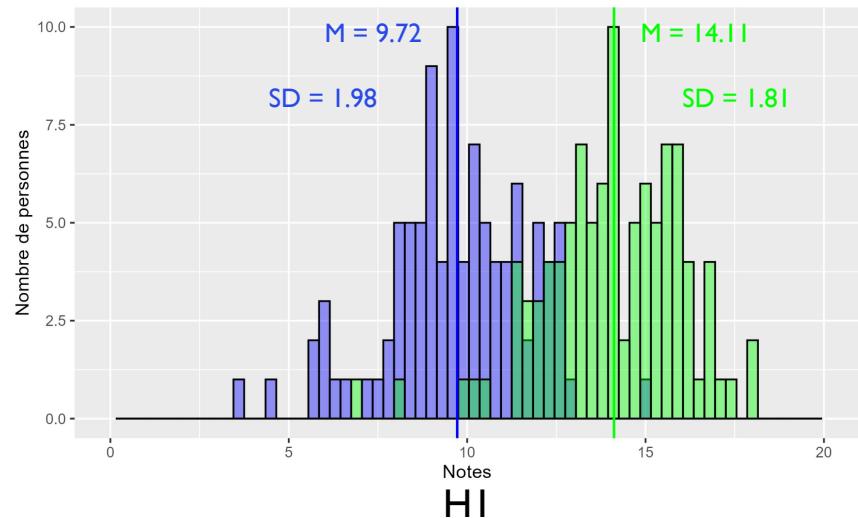
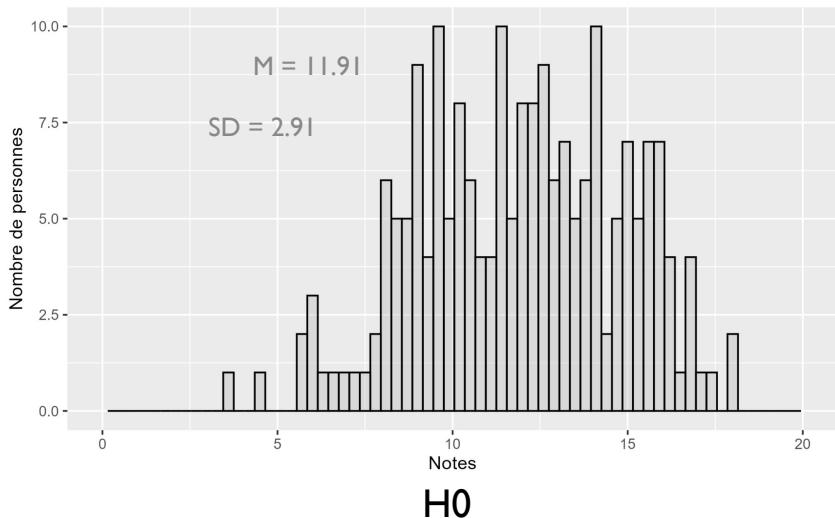


DES STATISTIQUES INFÉRENTIELLES?

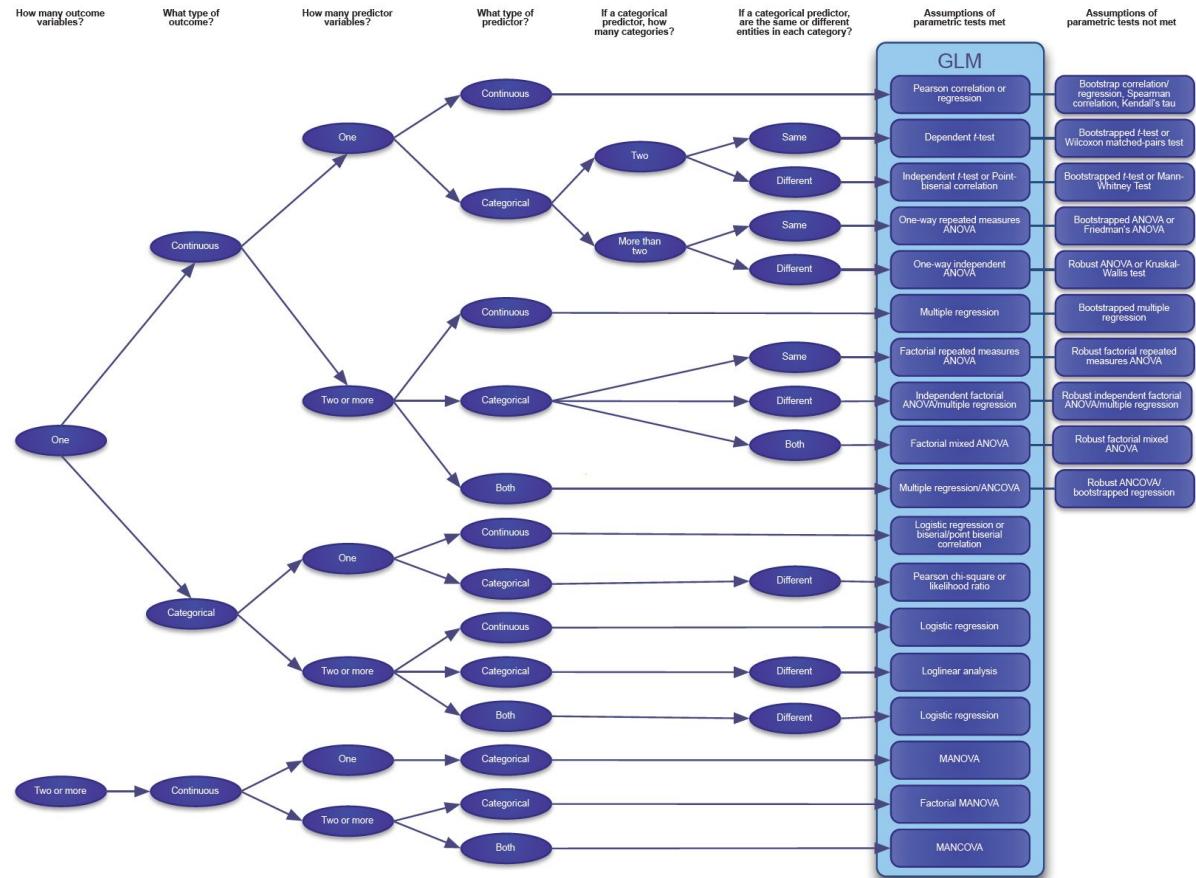
Objectif: prédire (le comportement)

A partir de notre modèle, on peut inférer sur une population à partir des données observées dans notre échantillon.

→ On peut généraliser à tous les gens qui aiment ou n'aiment pas les statistiques.



QUEL TEST CHOISIR?



RÉGRESSION SIMPLE: VARIABLE CATÉGORIELLE

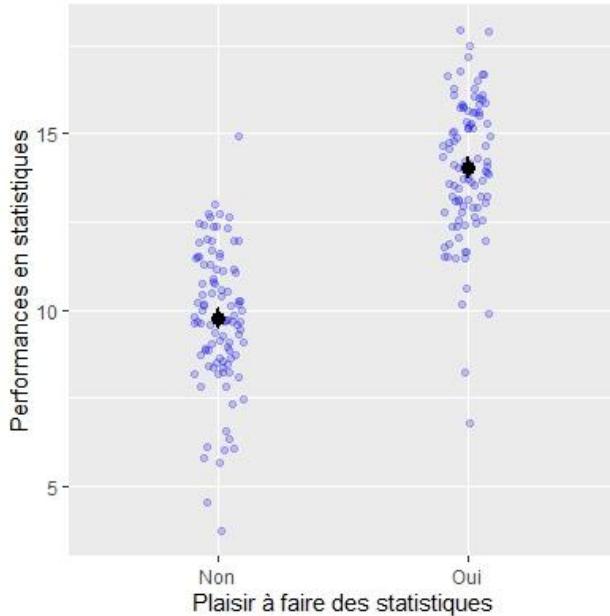
Voir si le plaisir à faire des statistiques (oui / non) prédit les performances en statistiques (note de 0 à 20)

$$\text{performance} \sim \text{plaisir} + \varepsilon$$

ε = erreur de mesure, variance expliquée par des autres variances

Plusieurs tests inférentiels équivalents :

- Test t pour échantillons indépendants
- ANOVA univariée
- Régression linéaire



RÉGRESSION SIMPLE: VARIABLE CATÉGORIELLE

Voir si le plaisir à faire des statistiques (oui / non) prédit les performances en statistiques (note de 0 à 20)

$$\text{performance} \sim \text{plaisir} + \varepsilon$$

MÉTHODE I: Test t pour échantillons indépendants

M_1 = moyenne pour le groupe 1

M_2 = moyenne pour le groupe 2

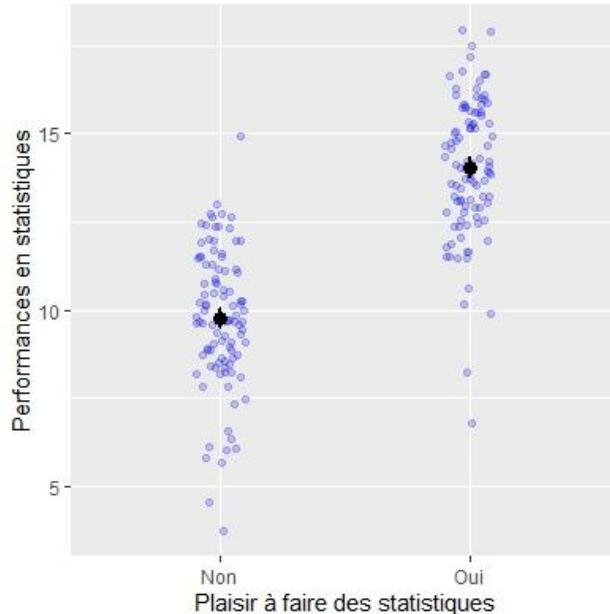
SD_1 = écart type pour le groupe 1

SD_2 = écart type pour le groupe 2

N_1 = nombre de sujet dans le groupe 1

N_2 = nombre de sujet dans le groupe 2

$$t = \frac{M_1 - M_2}{\sqrt{\frac{SD_1^2}{N_1} + \frac{SD_2^2}{N_2}}}$$



RÉGRESSION SIMPLE: VARIABLE CATÉGORIELLE

Voir si le plaisir à faire des statistiques (oui / non) prédit les performances en statistiques (note de 0 à 20)

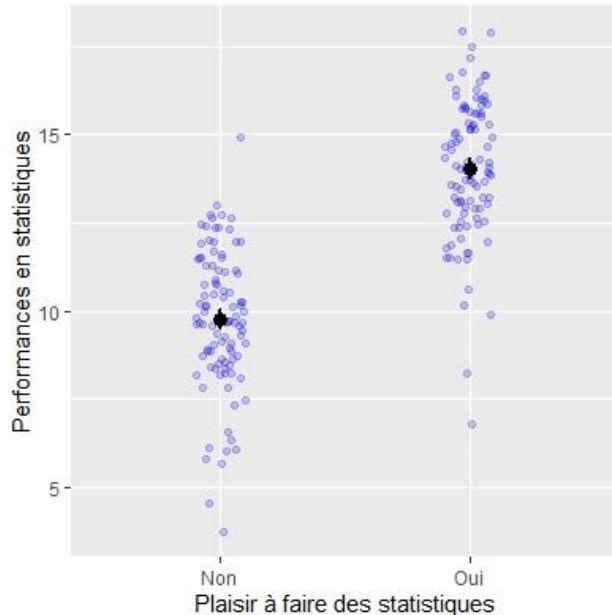
$$\text{performance} \sim \text{plaisir} + \varepsilon$$

MÉTHODE I: Test t pour échantillons indépendants

```
result <- t.test(data$notes ~ data$plaisir_stats, var.equal=TRUE)
result
  Two Sample t-test

  data: data$notes by data$plaisir_stats
  t = -15.478, df = 198, p-value < 2.2e-16
  alternative hypothesis: true difference in means between groups is not equal to 0
  95 percent confidence interval:
  -4.839214 -3.745438
  sample estimates:
  mean in group 'Non'  mean in group 'Oui'
  9.75115      14.04348
```

Différence de moyenne = 4.29



RÉGRESSION SIMPLE: VARIABLE CATÉGORIELLE

Voir si le plaisir à faire des statistiques (oui / non) prédit les performances en statistiques (note de 0 à 20)

$$\text{performance} \sim \text{plaisir} + \varepsilon$$

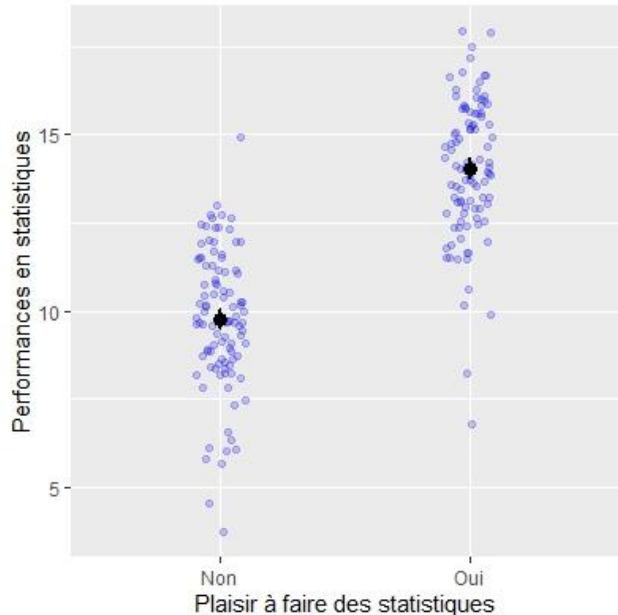
MÉTHODE 2:ANOVA univariée

Somme des carrés des écarts des observations x_i par rapport à la moyenne du groupe M

$$SC_{\text{intra}} = \sum_{i=1}^n (x_i - M)^2$$

Somme des carrés des écarts des moyennes du groupe M par rapport à la moyenne globale

$$SC_{\text{inter}} = \sum_{j=1}^k (M_j - MOY)^2$$



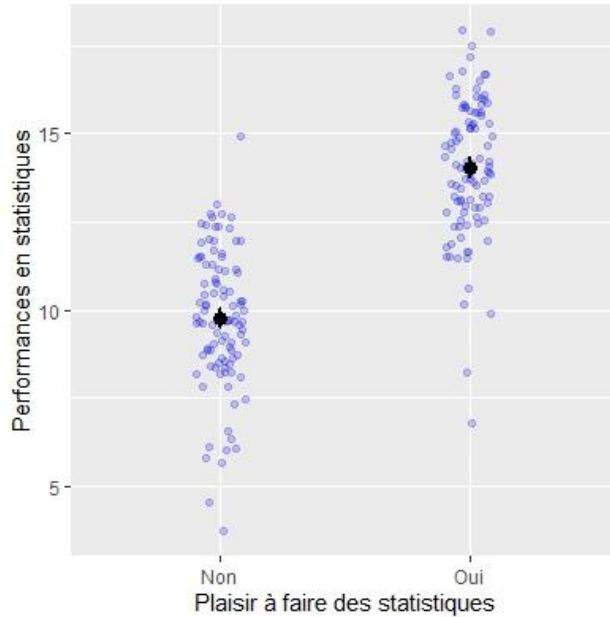
RÉGRESSION SIMPLE: VARIABLE CATÉGORIELLE

Voir si le plaisir à faire des statistiques (oui / non) prédit les performances en statistiques (note de 0 à 20)

$$\text{performance} \sim \text{plaisir} + \varepsilon$$

MÉTHODE 2:ANOVA univariée

$$F = \frac{\frac{SC_{inter}}{ddl_{inter}}}{\frac{SC_{intra}}{ddl_{intra}}}$$



RÉGRESSION SIMPLE: VARIABLE CATÉGORIELLE

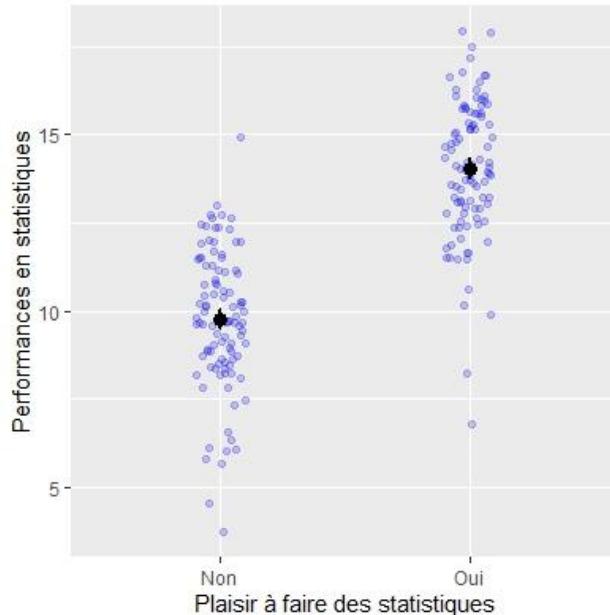
Voir si le plaisir à faire des statistiques (oui / non) prédit les performances en statistiques (note de 0 à 20)

$$\text{performance} \sim \text{plaisir} + \varepsilon$$

MÉTHODE 2:ANOVA univariée

```
result <- aov(data$notes ~ data$plaisir_stats)
summary(result)

      Df Sum Sq Mean Sq F value Pr(>F)
data$plaisir_stats   1  921.1   921.1    239.6 <2e-16 ***
Residuals       198  761.3     3.8
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```



RÉGRESSION SIMPLE: VARIABLE CATÉGORIELLE

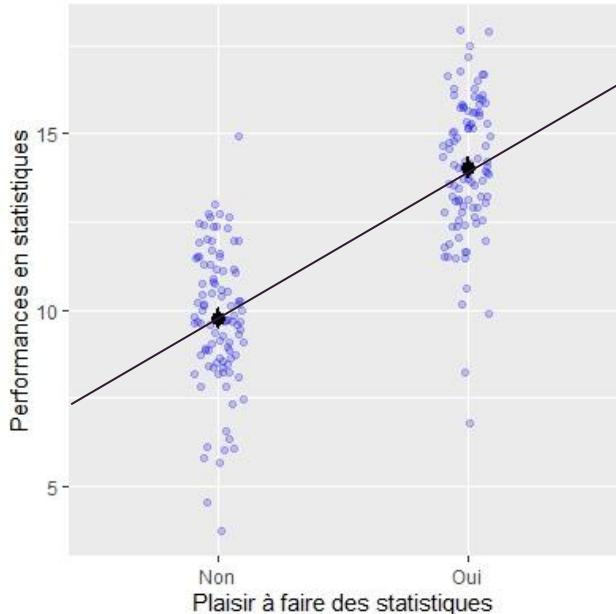
Voir si le plaisir à faire des statistiques (oui / non) prédit les performances en statistiques (note de 0 à 20)

$$\text{performance} \sim \text{plaisir} + \varepsilon$$

MÉTHODE 3: Régression linéaire

Trouver la droite qui minimise le plus la somme des carrées des erreurs

$$y_{ij} = \alpha_0 + \alpha_1 A_{ij} + \varepsilon$$



RÉGRESSION SIMPLE: VARIABLE CATÉGORIELLE

Voir si le plaisir à faire des statistiques (oui / non) prédit les performances en statistiques (note de 0 à 20)

$$\text{performance} \sim \text{plaisir} + \varepsilon$$

MÉTHODE 3: Régression linéaire

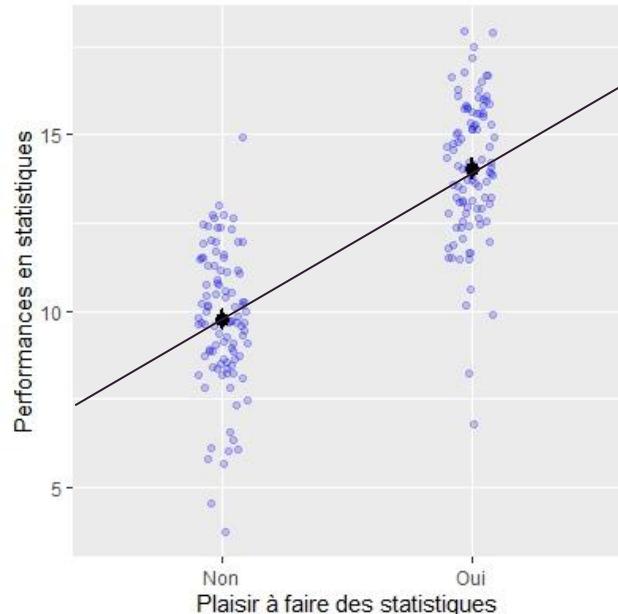
```
reg <- lm(plaisir_stats ~ notes, data=data)
summary(reg)

Call:
lm(formula = notes ~ plaisir_stats, data = data)

Residuals:
    Min      1Q  Median      3Q     Max 
-7.2679 -1.1308 -0.0206  1.4791  5.1633 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 11.8973    0.1387  85.80   <2e-16 ***
plaisir_stats 4.2923    0.2773  15.48   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.961 on 198 degrees of freedom
Multiple R-squared:  0.5475, Adjusted R-squared:  0.5452 
F-statistic: 239.6 on 1 and 198 DF,  p-value: < 2.2e-16
```



RÉGRESSION SIMPLE: VARIABLE CATÉGORIELLE

Voir si le plaisir à faire des statistiques (oui / non) prédit les performances en statistiques (note de 0 à 20)

$$\text{performance} \sim \text{plaisir} + \varepsilon$$

MÉTHODE 3: Régression linéaire

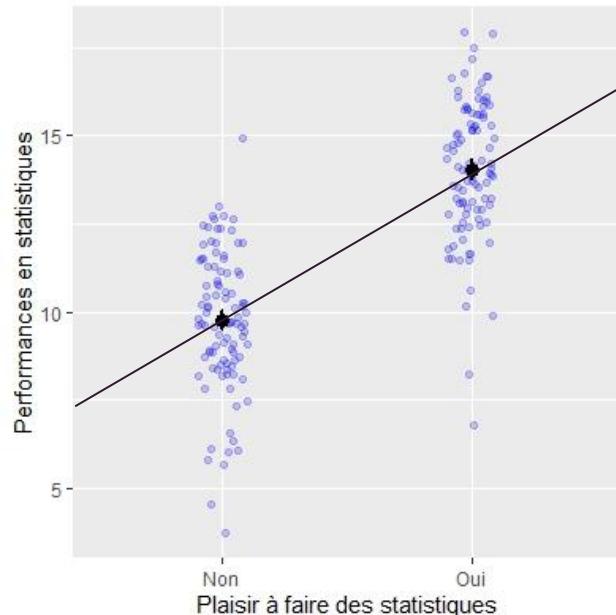
```
reg <- lm(plaisir_stats ~ notes, data=data)
summary(reg)

Call:
lm(formula = notes ~ plaisir_stats, data = data)

Residuals:
    Min      1Q  Median      3Q     Max 
-7.2679 -1.1308 -0.0206  1.4791  5.1633 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 11.8973    0.1387  85.80   <2e-16 ***
plaisir_stats 4.2923    0.2773  15.48   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.961 on 198 degrees of freedom
Multiple R-squared:  0.5475, Adjusted R-squared:  0.5452 
F-statistic: 239.6 on 1 and 198 DF,  p-value: < 2.2e-16
```



RÉGRESSION SIMPLE: VARIABLE CATÉGORIELLE

Voir si le plaisir à faire des statistiques (oui / non) prédit les performances en statistiques (note de 0 à 20)

$$\text{performance} \sim \text{plaisir} + \varepsilon$$

MÉTHODE 3: Régression linéaire

```
reg <- lm(plaisir_stats ~ notes, data=data)
summary(reg)

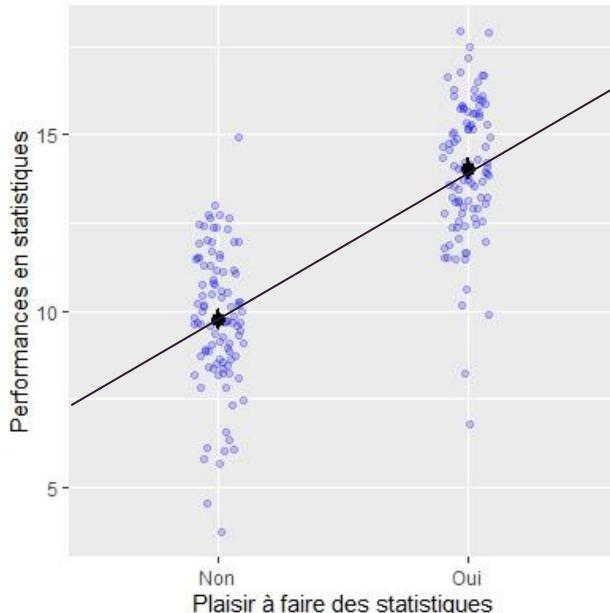
Call:
lm(formula = notes ~ plaisir_stats, data = data)

Residuals:
    Min      1Q  Median      3Q     Max 
-7.2679 -1.1308 -0.0206  1.4791  5.1633 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 11.8973    0.1387  85.80   <2e-16 ***
plaisir_stats 4.2923    0.2773  15.48   <2e-16 ***  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.961 on 198 degrees of freedom
Multiple R-squared:  0.5475, Adjusted R-squared:  0.5452 
F-statistic: 239.6 on 1 and 198 DF,  p-value: < 2.2e-16
```

Différence de moyenne = 4.29



RÉGRESSION SIMPLE: VARIABLE CONTINUE

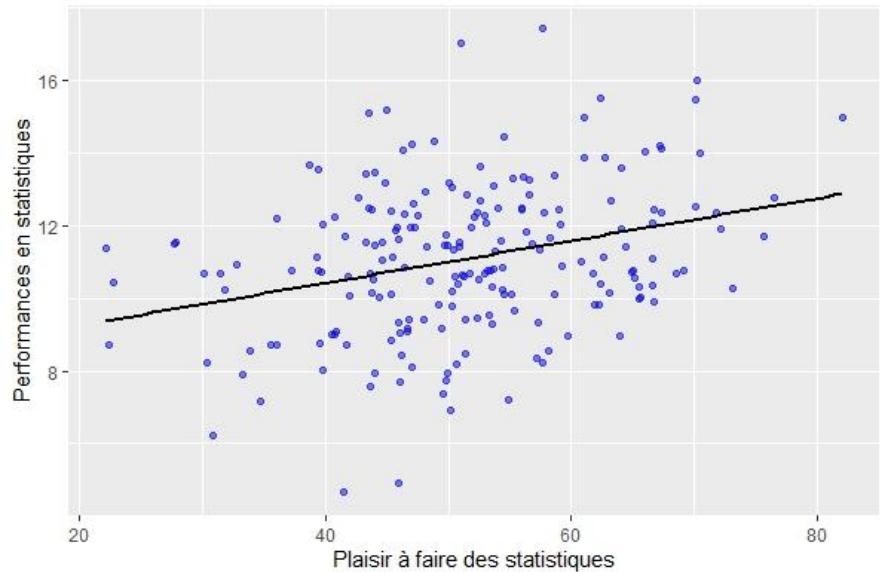
Voir si le plaisir à faire des statistiques (de 0 à 100) prédit les performances en statistiques (note de 0 à 20)

$$\text{performance} \sim \text{plaisir} + \varepsilon$$

Régression linéaire

Trouver la droite qui minimise la plus la somme des carrées des erreurs

$$y_{ij} = \alpha_0 + \alpha_1 A_{ij} + \varepsilon$$



RÉGRESSION SIMPLE: VARIABLE CONTINUE

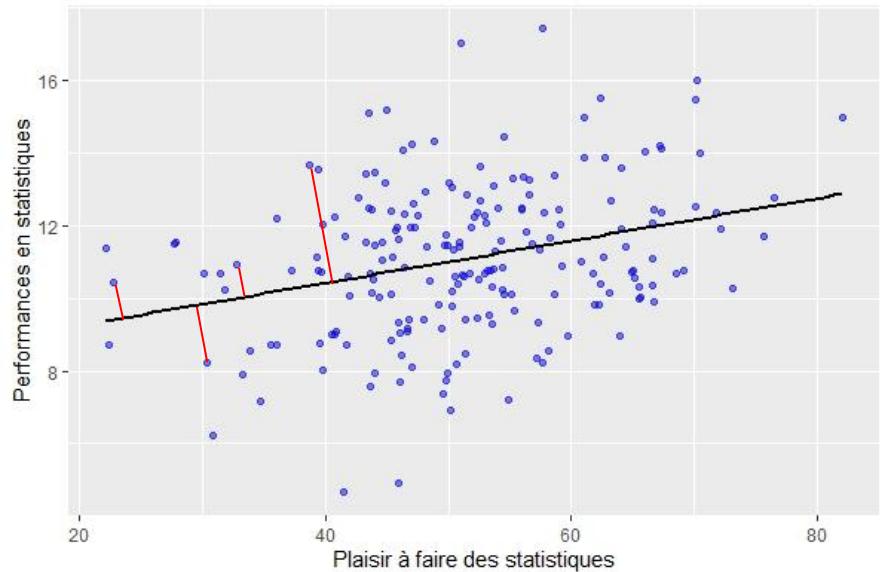
Voir si le plaisir à faire des statistiques (de 0 à 100) prédit les performances en statistiques (note de 0 à 20)

$$\text{performance} \sim \text{plaisir} + \varepsilon$$

Régression linéaire

Trouver la droite qui minimise la plus la somme des carrées des erreurs

$$y_{ij} = \alpha_0 + \alpha_1 A_{ij} + \varepsilon$$



RÉGRESSION SIMPLE: VARIABLE CONTINUE

Voir si le plaisir à faire des statistiques (de 0 à 100) prédit les performances en statistiques (note de 0 à 20)

$$\text{performance} \sim \text{plaisir} + \varepsilon$$

Régression linéaire

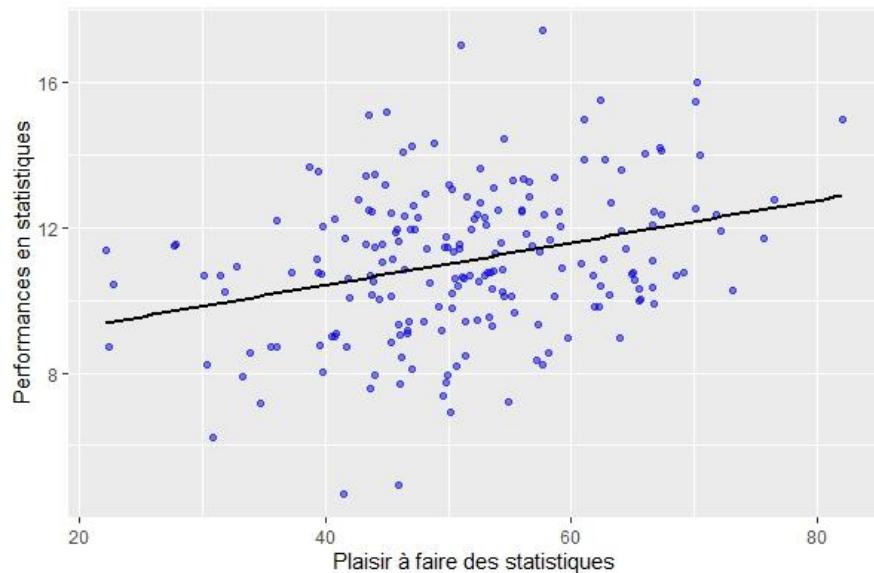
```
reg <- lm(notes ~ plaisir_stats_continue, data=data2)
summary(reg)

Call:
lm(formula = notes ~ plaisir_stats_continue, data = data2)

Residuals:
    Min      1Q  Median      3Q     Max 
-5.8550 -1.4381  0.0792  1.2939  5.9688 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 8.08926   0.68127 11.874 < 2e-16 ***
plaisir_stats_continue 0.05831   0.01298  4.493 1.19e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.992 on 198 degrees of freedom
Multiple R-squared:  0.09251, Adjusted R-squared:  0.08792 
F-statistic: 20.18 on 1 and 198 DF,  p-value: 1.195e-05
```



RÉGRESSION SIMPLE: VARIABLE CONTINUE

Voir si le plaisir à faire des statistiques (de 0 à 100) prédit les performances en statistiques (note de 0 à 20)

$$\text{performance} \sim \text{plaisir} + \varepsilon$$

Régression linéaire

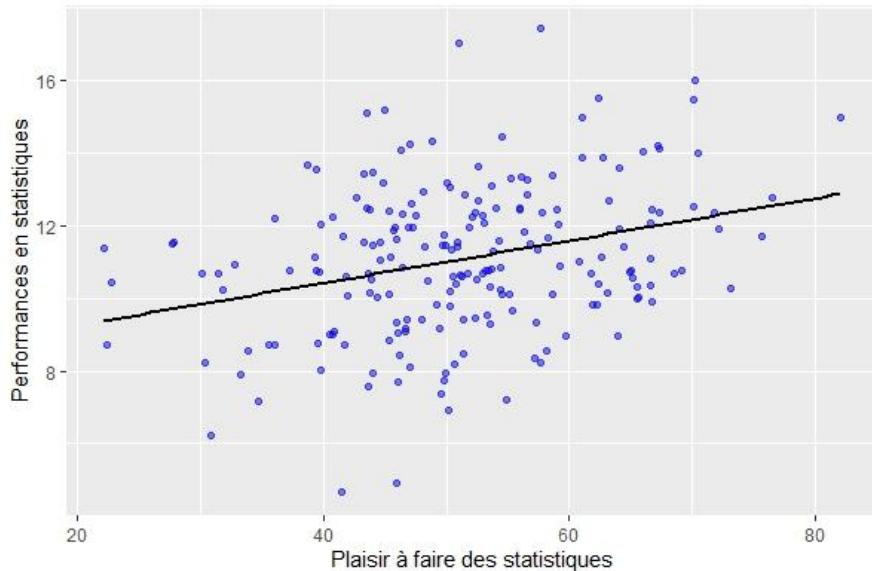
```
reg <- lm(notes ~ plaisir_stats_continue, data=data2)
summary(reg)

Call:
lm(formula = notes ~ plaisir_stats_continue, data = data2)

Residuals:
    Min      1Q  Median      3Q     Max 
-5.8550 -1.4381  0.0792  1.2939  5.9688 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  8.08926   0.68127 11.874 < 2e-16 ***
plaisir_stats_continue 0.05831   0.01298  4.493 1.19e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.992 on 198 degrees of freedom
Multiple R-squared:  0.09251, Adjusted R-squared:  0.08792 
F-statistic: 20.18 on 1 and 198 DF,  p-value: 1.195e-05
```



RÉGRESSION SIMPLE: VARIABLE CONTINUE

Voir si le plaisir à faire des statistiques (de 0 à 100) prédit les performances en statistiques (note de 0 à 20)

$$\text{performance} \sim \text{plaisir} + \varepsilon$$

Régression linéaire

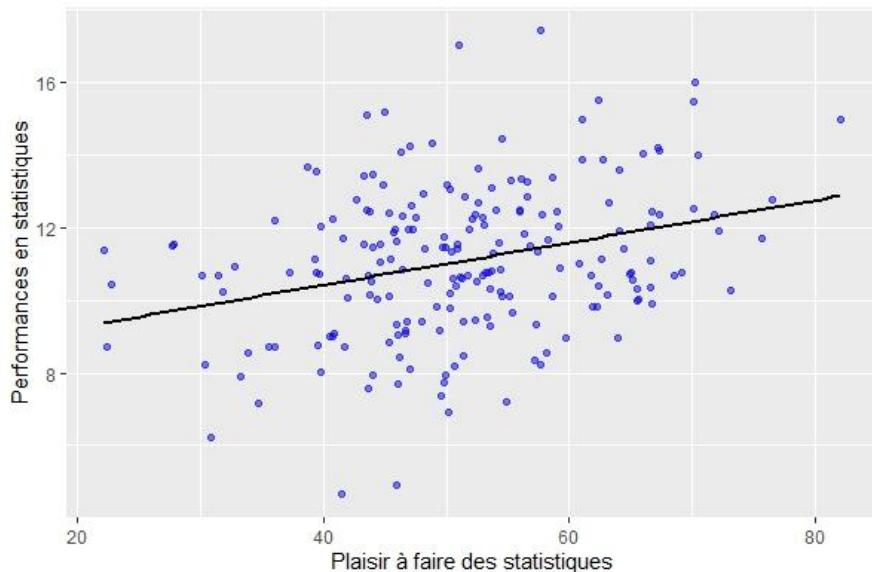
```
reg <- lm(notes ~ plaisir_stats_continue, data=data2)
summary(reg)

Call:
lm(formula = notes ~ plaisir_stats_continue, data = data2)

Residuals:
    Min      1Q  Median      3Q     Max 
-5.8550 -1.4381  0.0792  1.2939  5.9688 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 8.08926   0.68127 11.874 < 2e-16 ***
plaisir_stats_continue 0.05831   0.01298  4.493 1.19e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.992 on 198 degrees of freedom
Multiple R-squared:  0.09251, Adjusted R-squared:  0.08792 
F-statistic: 20.18 on 1 and 198 DF,  p-value: 1.195e-05
```



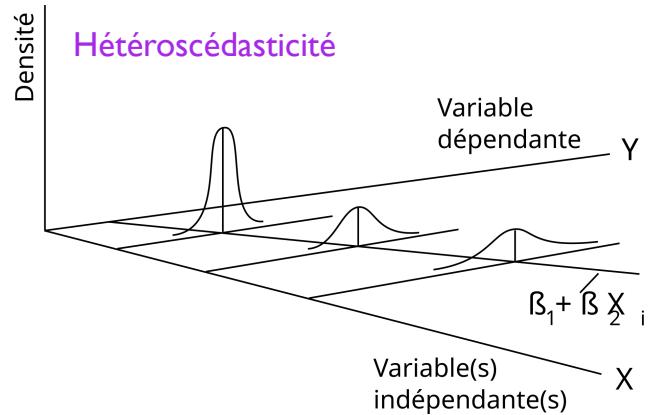
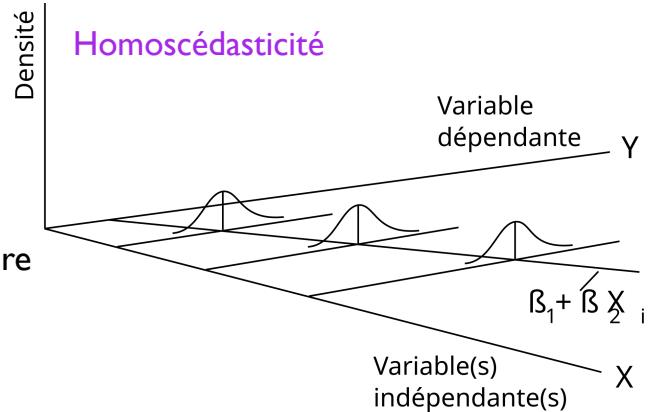
CONDITIONS D'APPLICATIONS

Pour pouvoir utiliser des tests paramétriques, il faut s'assurer que:

CONDITIONS D'APPLICATIONS

Pour pouvoir utiliser des tests paramétriques, il faut s'assurer que:

- 1) Les résidus (erreurs ε) sont distribués de façon normale
- 2) Les résidus (erreurs ε) de nos différents groupes ont une variance similaire
→ On teste l'absence d'hétérocédasticité

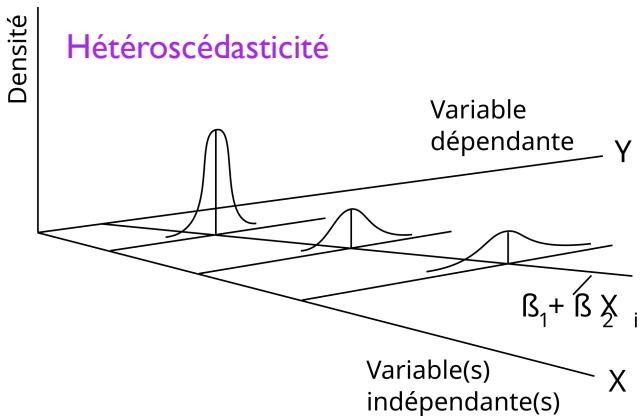
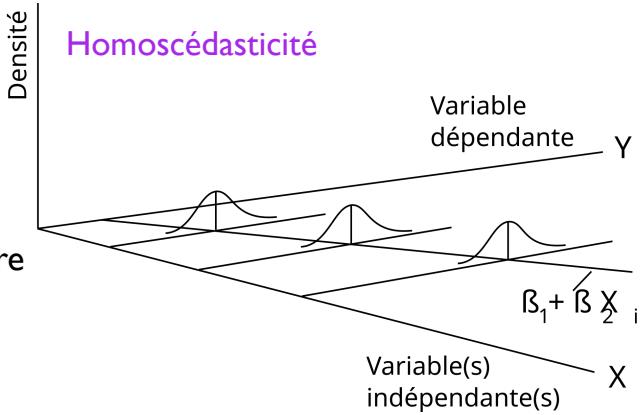


CONDITIONS D'APPLICATIONS

Pour pouvoir utiliser des tests paramétriques, il faut s'assurer que:

- 1) Les résidus (erreurs ε) sont distribué de façon normale
- 2) Les résidus (erreurs ε) de nos différents groupes ont une variance similaire
→ On teste l'absence d'hétérocédasticité
- 3) L'indépendance des résidus (erreurs ε)

Vérifier aussi les données extrêmes / aberrantes qui peuvent biaiser les estimations (car ont un fort poids dans la moyenne).



POURQUOI UTILISER DES MODÈLES MIXTES?

Un exemple simple

La performance en statistique est mesurée 1 fois → risque d'erreur important car beaucoup de facteurs peuvent influencer cette mesure (type d'examen, horaire, humeur, difficulté de l'évaluation, etc.)

→ On décide plus faire plusieurs mesures (ex. 30 examens).

POURQUOI UTILISER DES MODÈLES MIXTES?

Un exemple simple

La performance en statistique est mesurée 1 fois → risque d'erreur important car beaucoup de facteurs peuvent influencer cette mesure (type d'examen, horaire, humeur, difficulté de l'évaluation, etc.)

→ On décide plus faire plusieurs mesures (ex. 30 examens).

Approche classique: Moyenner les notes aux 30 examens (indépendance des résidus) et faire une régression linéaire

MAIS

- Perte d'information
- Plus grand risque d'erreurs de type I

→ Solution: utilisation de modèles à effets mixtes !

POURQUOI UTILISER DES MODÈLES MIXTES?

Une structure hiérarchique

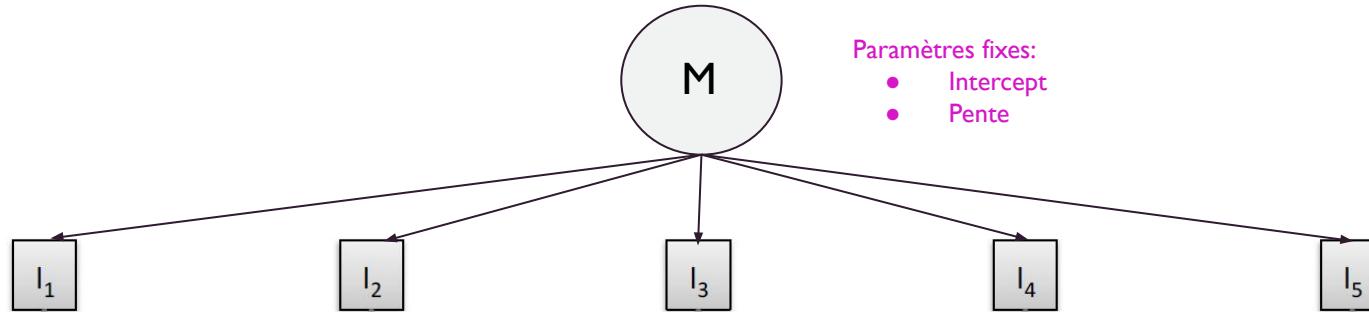
Collecter plusieurs données sur les mêmes participants crée une structure hiérarchique des données.



POURQUOI UTILISER DES MODÈLES MIXTES?

Une structure hiérarchique

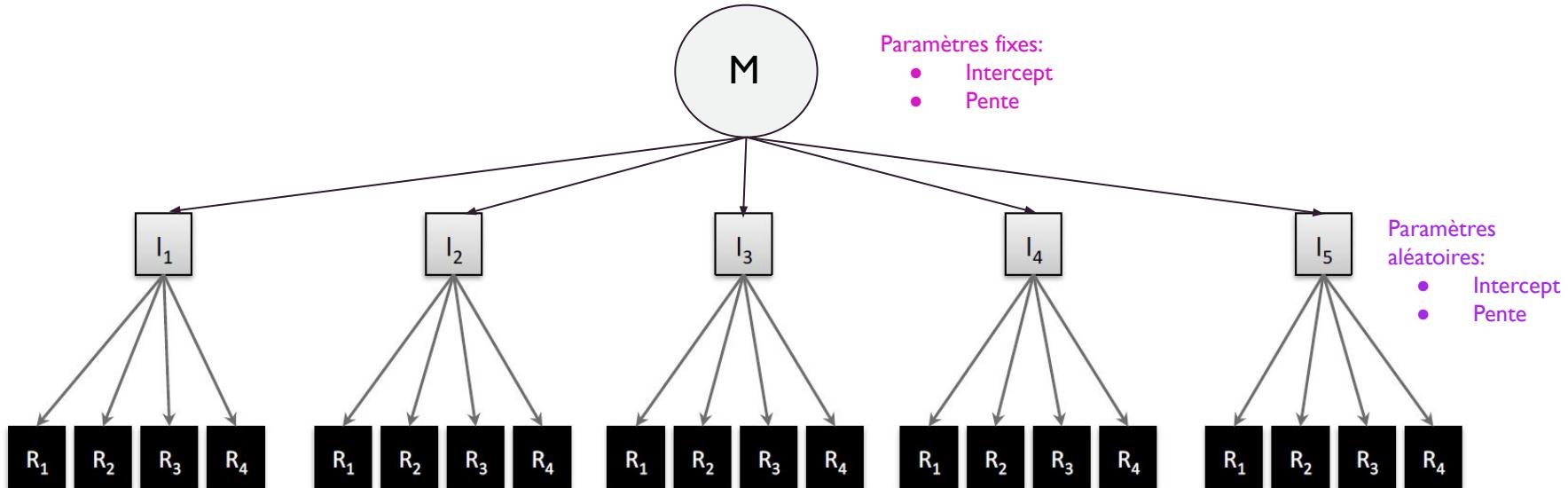
Collecter plusieurs données sur les mêmes participants crée une structure hiérarchique des données.



POURQUOI UTILISER DES MODÈLES MIXTES?

Une structure hiérarchique

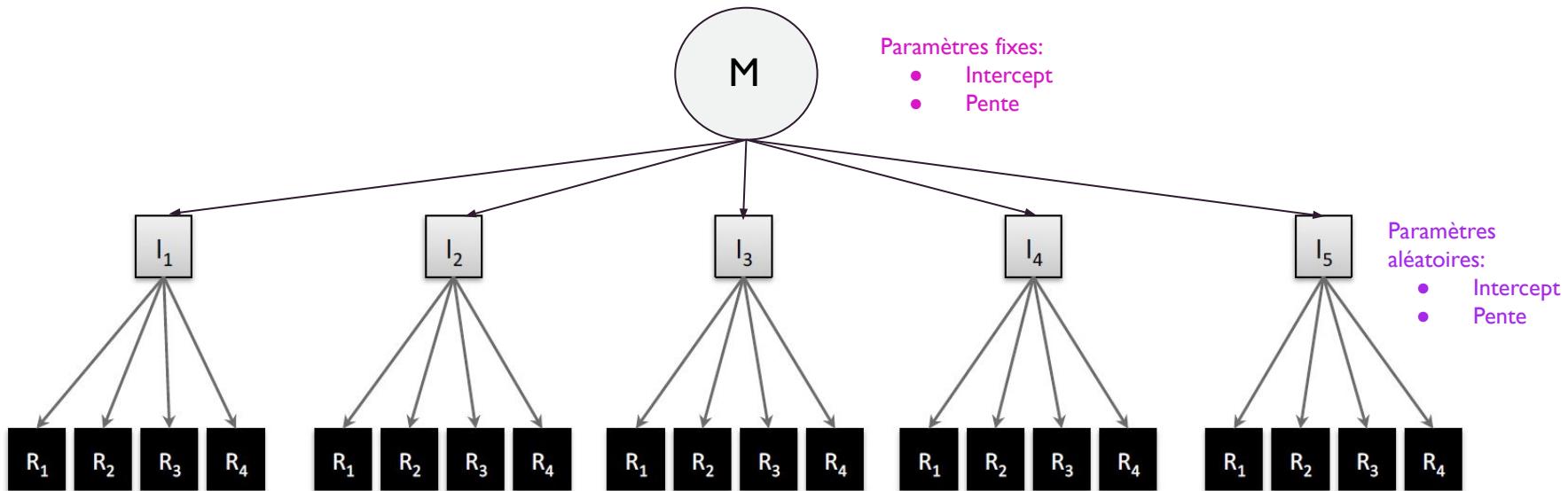
Collecter plusieurs données sur les mêmes participants crée une structure hiérarchique des données.



POURQUOI UTILISER DES MODÈLES MIXTES?

Une structure hiérarchique

Collecter plusieurs données sur les mêmes participants crée une structure hiérarchique des données.



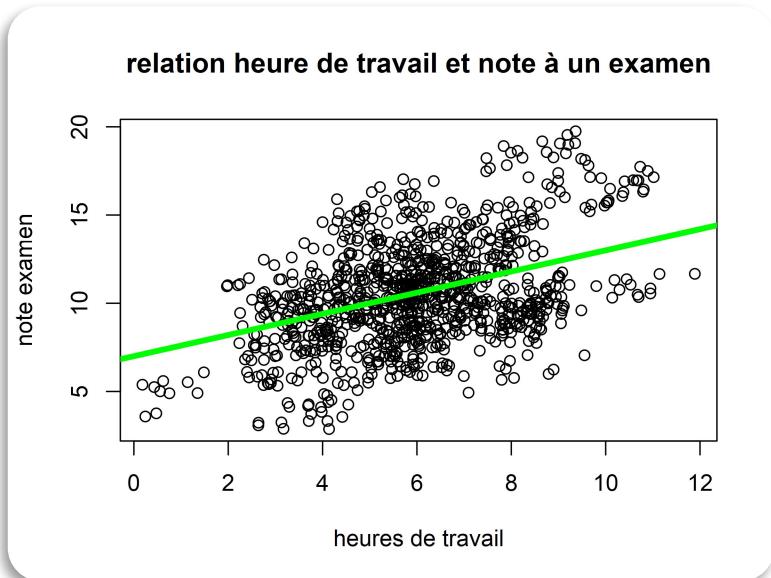
Deux niveaux de variabilités : intra-sujets et inter-sujets.

Les approches classiques ne prennent pas en compte la variabilité intra-sujet.

POURQUOI UTILISER DES MODÈLES MIXTES?

Est-ce que le nombre d'heure passé à travailler une matière prédit la note à l'examen?

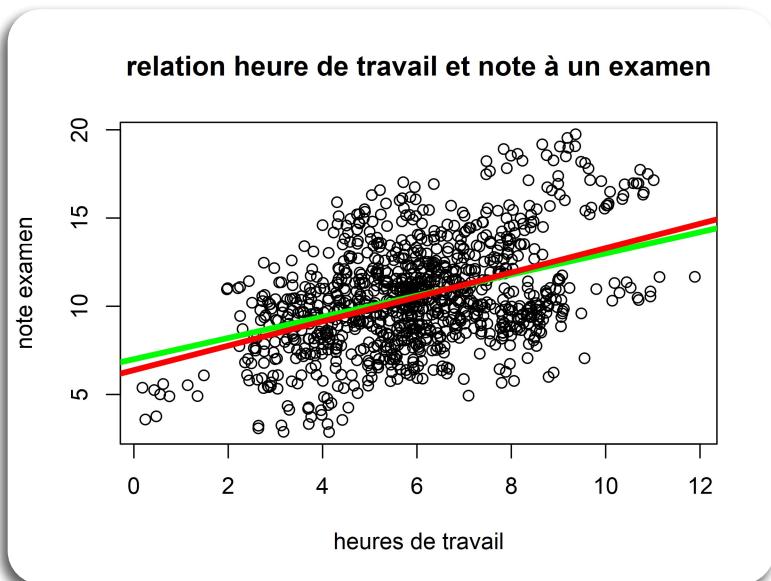
Données simulées : on sait que l'intercept réel (dans la population) est 7 et pente de 0.6



POURQUOI UTILISER DES MODÈLES MIXTES?

Est-ce que le nombre d'heure passé à travailler une matière prédit la note à l'examen?

Données simulées : on sait que l'intercept réel (dans la population) est 7 et pente de 0.6



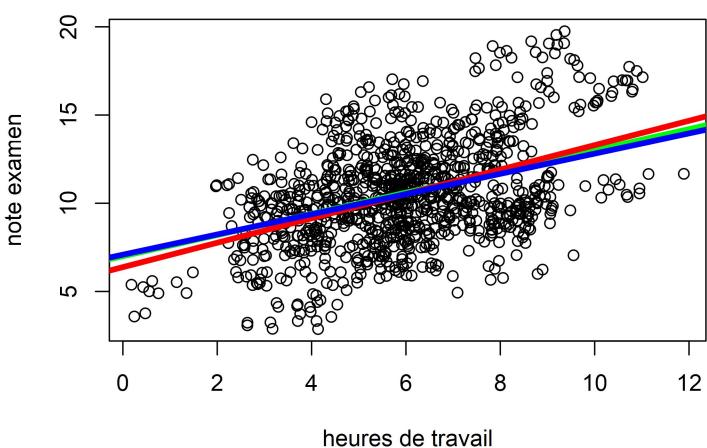
Simple régression linéaire :
intercept = 6.37982, pente = 0.69

POURQUOI UTILISER DES MODÈLES MIXTES?

Est-ce que le nombre d'heure passé à travailler une matière prédit la note à l'examen?

Données simulées : on sait que l'intercept réel (dans la population) est 7 et pente de 0.6

relation heure de travail et note à un examen



Simple régression linéaire :

intercept = 6.37982, pente = 0.69

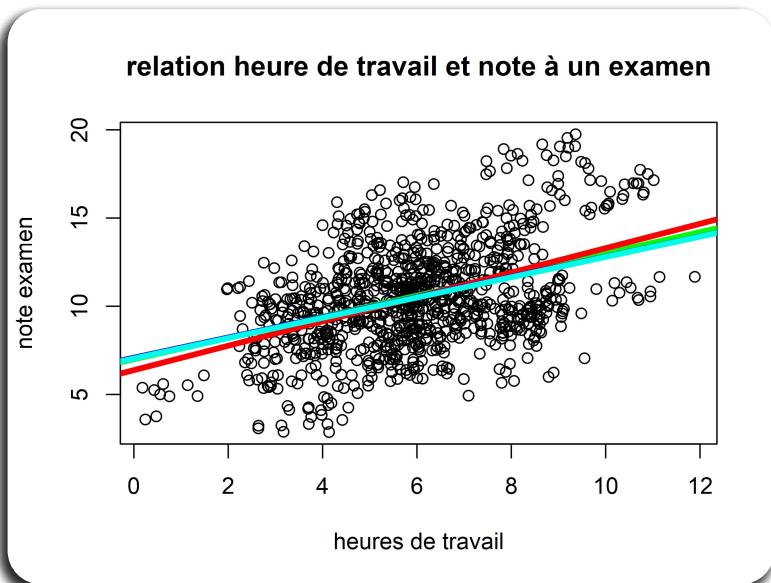
Régression linéaire mixte avec intercept par participant :

intercept = 7.08, pente = 0.57

POURQUOI UTILISER DES MODÈLES MIXTES?

Est-ce que le nombre d'heure passé à travailler une matière prédit la note à l'examen?

Données simulées : on sait que l'intercept réel (dans la population) est 7 et pente de 0.6



Simple régression linéaire :

intercept = 6.37982, pente = 0.69

Régression linéaire mixte avec intercept par participant :

intercept = 7.08, pente = 0.57

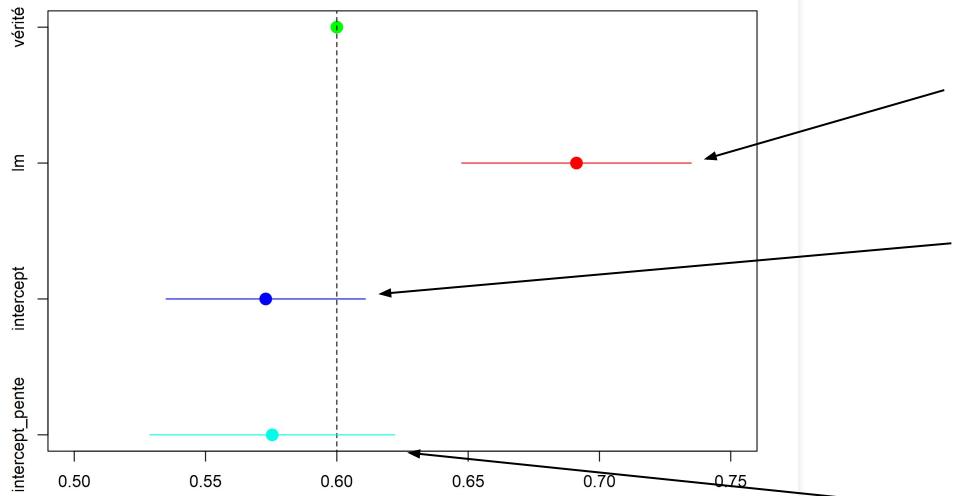
Régression linéaire mixte avec intercept et pente par participant :

intercept = 7.04, pente = 0.58

POURQUOI UTILISER DES MODÈLES MIXTES?

Est-ce que le nombre d'heure passé à travailler une matière prédit la note à l'examen?

Données simulées : on sait que l'intercept réel (dans la population) est 7 et pente de 0.6

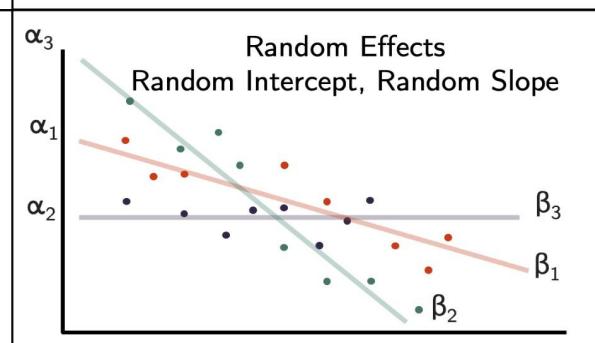
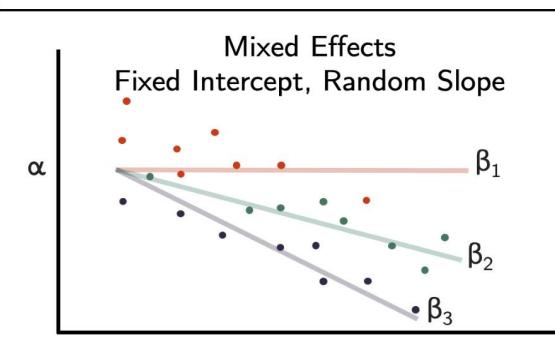
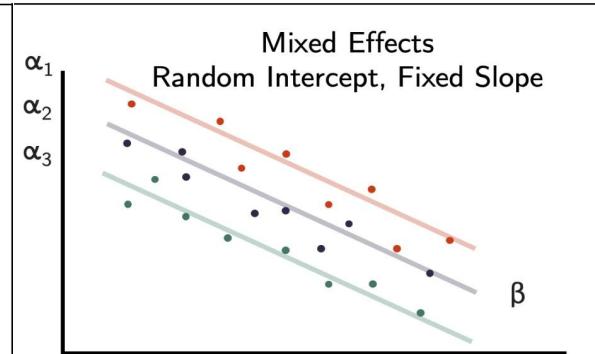
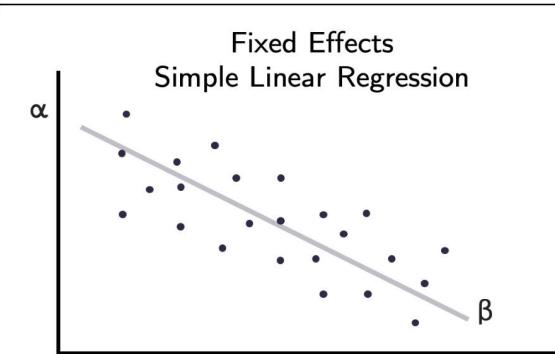


Simple régression linéaire :
intercept = 6.37982, pente = 0.69,
erreur standard = 0.044

Régression linéaire mixte avec intercept par participant :
intercept = 7.08, pente = 0.57,
erreur standard = 0.038

Régression linéaire mixte avec intercept et pente par participant :
intercept = 7.04, pente = 0.58,
erreur standard = 0.047

POURQUOI UTILISER DES MODÈLES MIXTES?



RÉSUMÉ DU JOUR

- L'importance de la loi normale → moyenne et écart type
- Modèle statistique
 - T-test
 - ANOVA
 - Régression simple
- Conditions d'applications des tests paramétrique: importance de l'indépendance des résidus
- Avantages des modèles mixtes: effets fixes et effets aléatoires