# MODÈLES LINÉAIRES MIXTES ET GÉNÉRALISÉS POUR L'ANALYSE DES DONNÉES

Félix L’Heureux Bilodeau

28 novembre 2024

RQRAD

Réseau québécois de recherche en agriculture durable

# Plan de la présentation

Objectifs :
Boite d’outils pour les dispositifs fréquents en agriculture

Besoin d’une base en R et en statistiques

La présentation et le fichier de code vous sera transmis.

## RÉGRESSION LINÉAIRE

Introduction/révision

## MODÈLES MIXTES

Quand l’utiliser et déterminer les facteurs aléatoires

## MODÈLES GÉNÉRALISÉS

Introduction aux différentes distributions

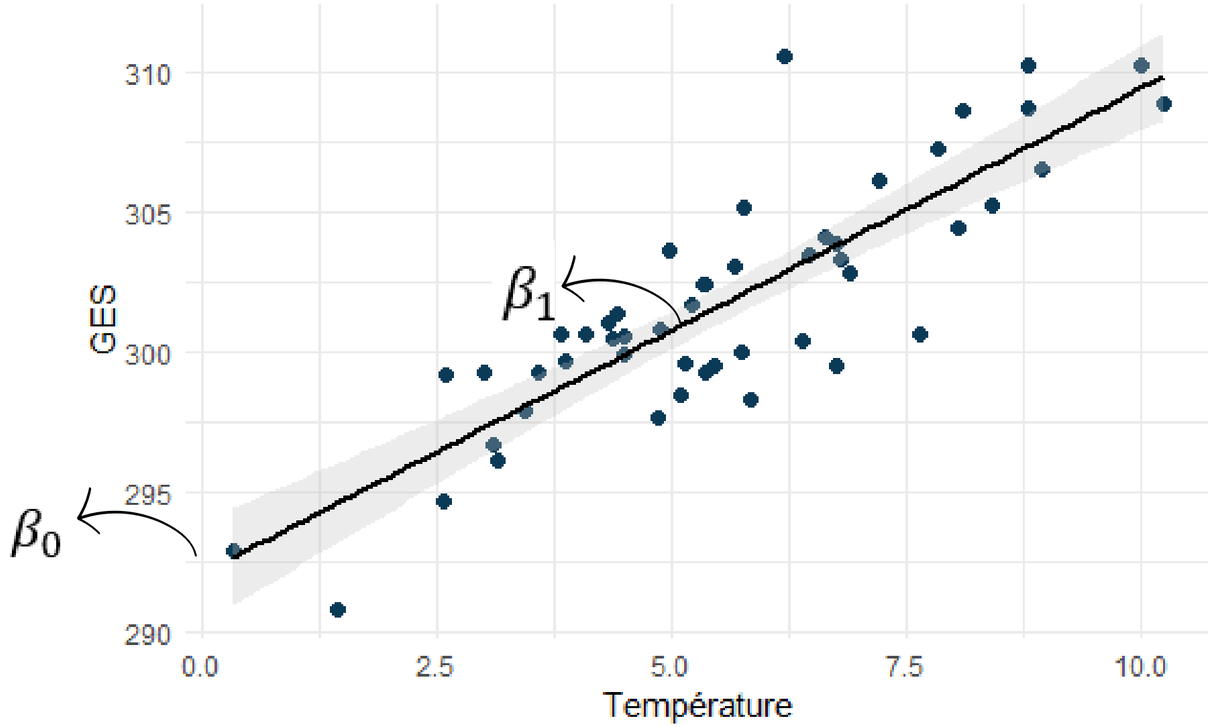
## STATISTIQUES BAYÉSIENNES

Introduction

# LA RÉGRESSION LINÉAIRE

Régression simple:

$$y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

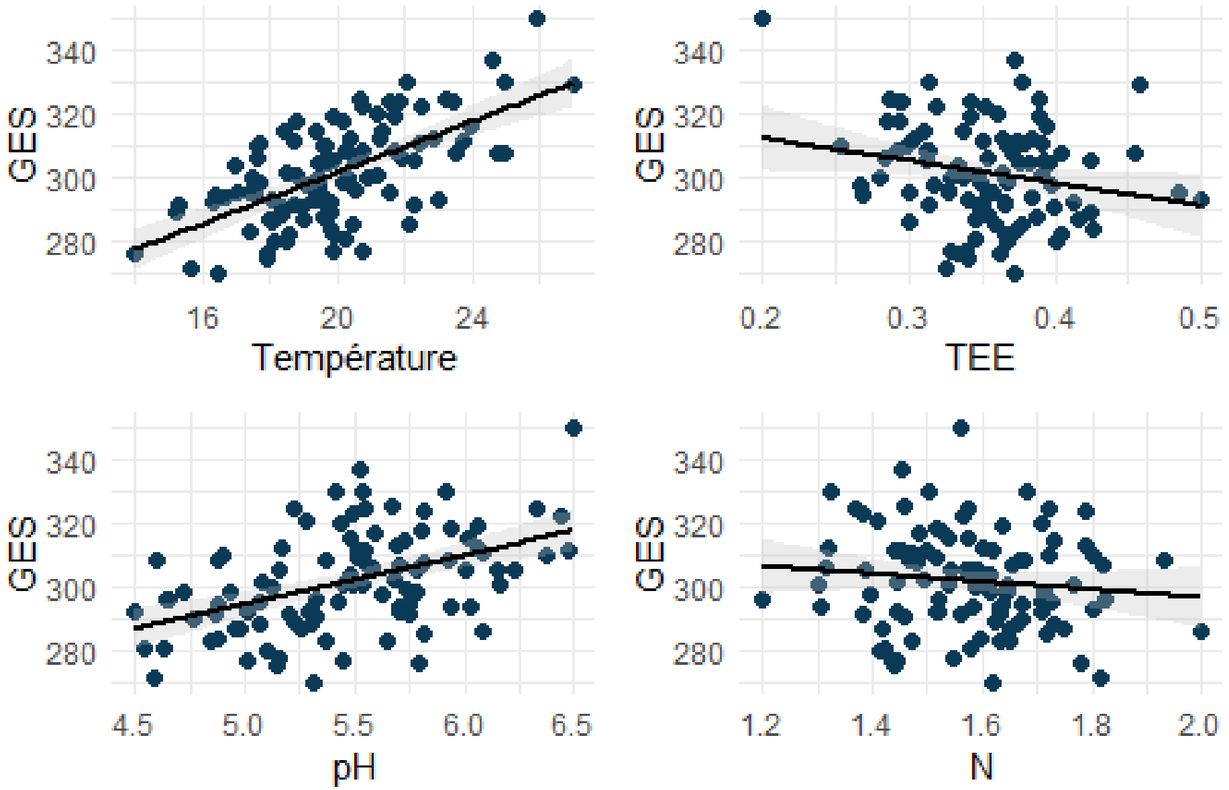


# LA RÉGRESSION LINÉAIRE

Pour une relation de cause à effet

Régression multiple:

$$y_i = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \beta_3 X_{3,i} + \beta_4 X_{4,i} + \epsilon_i$$



# LA RÉGRESSION LINÉAIRE

Pour une relation de cause à effet

Régression multiple:

$$y_i = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \beta_3 X_{3,i} + \beta_4 X_{4,i} + \epsilon_i$$

```
mod <- lm(GES ~ Temperature + TEE + pH + N, data = df)
summary(mod)
```

```
Call:
lm(formula = GES ~ Temperature + TEE + pH + N, data = df)

Residuals:
     Min       1Q   Median       3Q      Max
-26.2940  -6.0784   0.9272   7.5761  19.9238

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  228.4750    19.0155  12.015  < 2e-16 ***
Temperature    3.7620     0.4742   7.934 4.15e-12 ***
TEE         -104.7863    22.4969  -4.658 1.04e-05 ***
pH             8.7169     2.3770   3.667 0.000404 ***
N             -8.0200     7.2517  -1.106 0.271544
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.45 on 95 degrees of freedom
Multiple R-squared:  0.5693,    Adjusted R-squared:  0.5512
F-statistic: 31.39 on 4 and 95 DF,  p-value: < 2.2e-16
```

# LA RÉGRESSION LINÉAIRE

Les suppositions de la régression

1. Indépendance
    a. les Y sont indépendant entre eux
2. Normalité
    a. les données et erreurs suivent une distribution N
3. Existance
    a. à chaque X, il existe une distribution de Y dans la population
4. Homoscédasticité
    a. La variance de Y est la même pour chaque X
5. Linéarité
    a. La relation Y ~ X est linéaire
6. Mesure de X exact

# LA RÉGRESSION LINÉAIRE

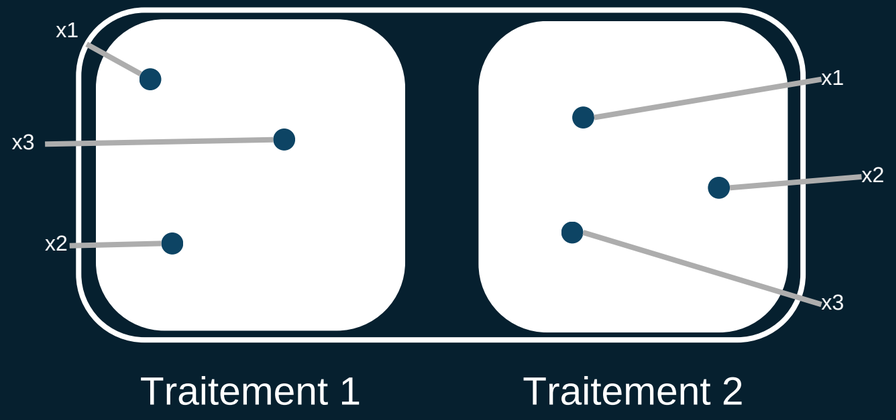
Les suppositions de la régression

1. Indépendance
    a. les Y sont indépendant entre eux
2. Normalité
    a. les données et erreurs suivent une distribution N
3. Existance
    a. à chaque x, il existe une distribution de Y dans la population
4. Homoscédasticité
    a. La variance de Y est la même pour chaque X
5. Linéarité
    a. La relation Y ~ X est linéaire
6. Mesure de X exact

# INDÉPENDANCE DES DONNÉES

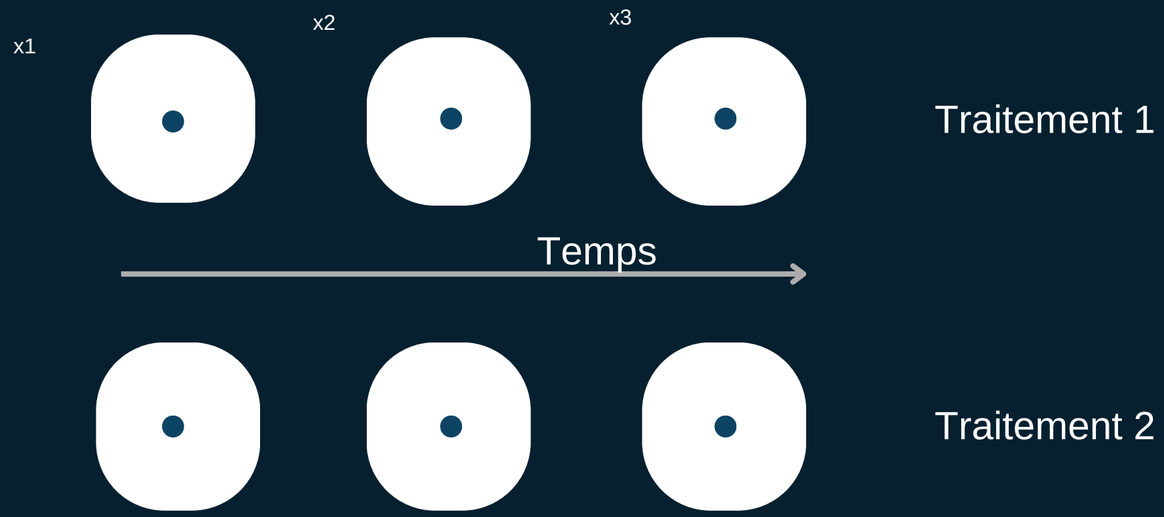
La pseudo-répétition

1. simple : plusieurs observation mais 1 seule UE par traitement




La pseudo-répétition

2. Temporel : mesure successive sur les même UE




🚩 ANOVA ?

# LE MODÈLE LINÉAIRE MIXTE

Permet de prendre en compte la structure des données (dispositif)
Pour mesure répétés, blocs, tiroir ...

Facteur aléatoire : «Il s'agit généralement de facteurs de regroupement dont vous souhaitez contrôler l'effet dans votre modèle, mais dont l'effet spécifique sur la variable de réponse ne vous intéresse pas. Ils nous permettent donc de structurer le processus d'erreur.» CSBQ

C'est un effet imprévisible

$$y_i = \beta X_i + Z_i b_i + \epsilon_i$$

où

$$\beta X_i = \beta_0 + \beta_1 x_i + \cdots + \beta_k x_k$$

$$\beta = effets\ fixes$$

$$Z_i b_i = b_1 x_1 + \cdots + b_k x_k$$

$$b_i = effet\ aléatoire \sim N(0, \sigma_b^2)$$

# DISPOSITIF EN BLOC

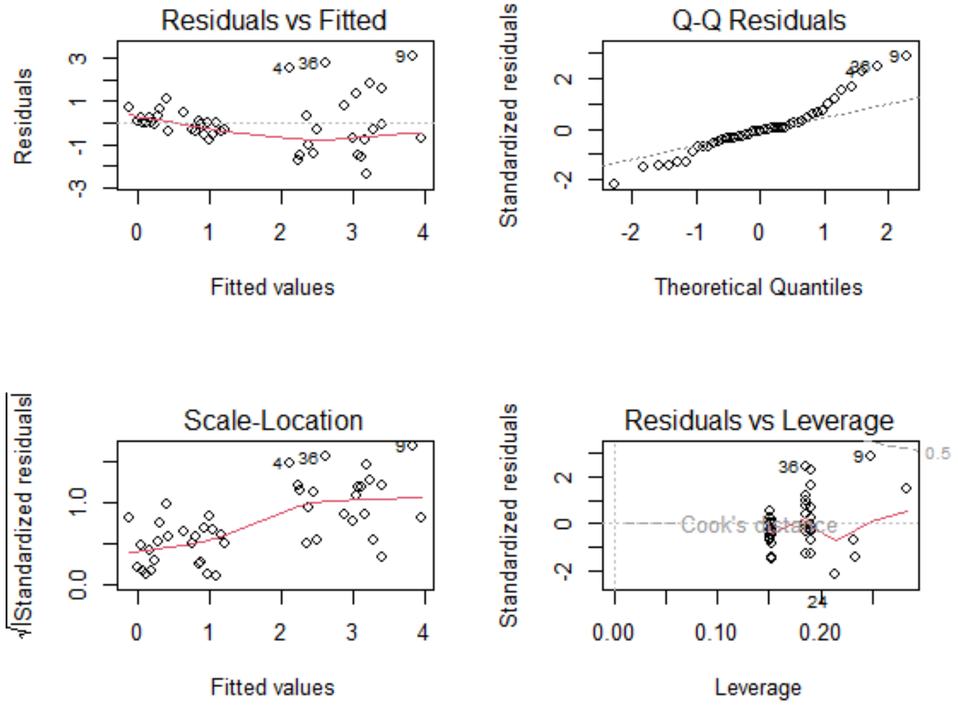
```
> head(Sophie,10)
# A tibble: 10 × 4
   Bloc  Trt           Saule    Rdt
   <fct> <fct>         <fct>  <dbl>
 1 1     AB            Pr     0.919
 2 1     AB            Sm     0.0424
 3 1     AB-MRF1-BRF   Pr     3.70
 4 1     AB-MRF1-BRF   Sm     4.63
 5 1     AB-MRF1-CI    Pr     4.38
 6 1     AB-MRF1-CI    Sm     0.796
 7 1     Témoin        Pr     1.13
 8 1     Témoin        Sm     0.615
 9 1     AB-MRF2-BRF   Pr     6.95
10 1     AB-MRF2-BRF   Sm     1.51
```

# DISPOSITIF EN BLOC

```
> head(Sophie,10)
# A tibble: 10 × 4
   Bloc  Trt           Saule  Rdt
   <fct> <fct>         <fct>  <dbl>
 1 1     AB            Pr     0.919
 2 1     AB            Sm     0.0424
 3 1     AB-MRF1-BRF   Pr     3.70
 4 1     AB-MRF1-BRF   Sm     4.63
 5 1     AB-MRF1-CI    Pr     4.38
 6 1     AB-MRF1-CI    Sm     0.796
 7 1     Témoin        Pr     1.13
 8 1     Témoin        Sm     0.615
 9 1     AB-MRF2-BRF   Pr     6.95
10 1     AB-MRF2-BRF   Sm     1.51
```

```
mod1 <- lm(Rdt ~ Trt + Saule + Bloc, data = Sophie)
```

```
plot(mod1)
```



# DISPOSITIF EN BLOC

```
> head(Sophie,10)
# A tibble: 10 × 4
   Bloc  Trt           Saule  Rdt
   <fct> <fct>         <fct>  <dbl>
 1 1     AB            Pr     0.919
 2 1     AB            Sm     0.0424
 3 1     AB-MRF1-BRF   Pr     3.70
 4 1     AB-MRF1-BRF   Sm     4.63
 5 1     AB-MRF1-CI    Pr     4.38
 6 1     AB-MRF1-CI    Sm     0.796
 7 1     Témoin        Pr     1.13
 8 1     Témoin        Sm     0.615
 9 1     AB-MRF2-BRF   Pr     6.95
10 1     AB-MRF2-BRF   Sm     1.51
```

```
mod1 <- lm(Rdt ~ Trt + Saule + Bloc, data = Sophie)
```

```
summary(mod1)
```

```
Call:
lm(formula = Rdt ~ Trt + Saule + Bloc, data = Sophie)

Residuals:
     Min       1Q   Median       3Q      Max
-2.03122 -0.51724 -0.02872  0.64128  2.61516

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)      0.3638     0.5404   0.673 0.505263
TrtAB            0.1959     0.5843   0.335 0.739420
TrtAB-BRF        0.1694     0.5843   0.290 0.773572
TrtAB-MRF1-BRF   2.2254     0.5843   3.809 0.000541 ***
TrtAB-MRF1-CI    2.3842     0.5843   4.081 0.000247 ***
TrtAB-MRF2-BRF   3.1777     0.6781   4.686 4.12e-05 ***
SaulePr          0.7891     0.3495   2.258 0.030289 *
Bloc2           -1.0406     0.4770  -2.181 0.035956 *
Bloc3           -0.4174     0.5092  -0.820 0.417932
Bloc4            0.2054     0.4900   0.419 0.677600
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.169 on 35 degrees of freedom
Multiple R-squared:  0.6308,    Adjusted R-squared:  0.5359
F-statistic: 6.644 on 9 and 35 DF,  p-value: 1.748e-05
```

# DISPOSITIF EN BLOC

## En modèle mixte

```
library(nlme)
```

```
modM <- lme(Rdt ~ Trt + Saule,~1|Bloc, data = Sophie)
```

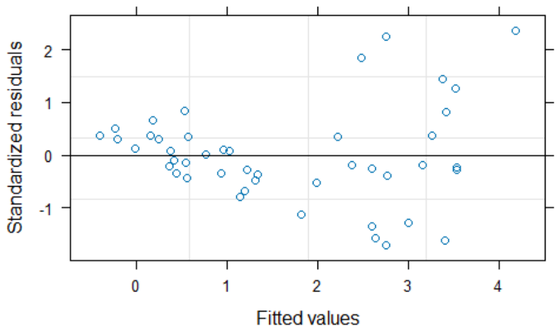
# DISPOSITIF EN BLOC

## En modèle mixte

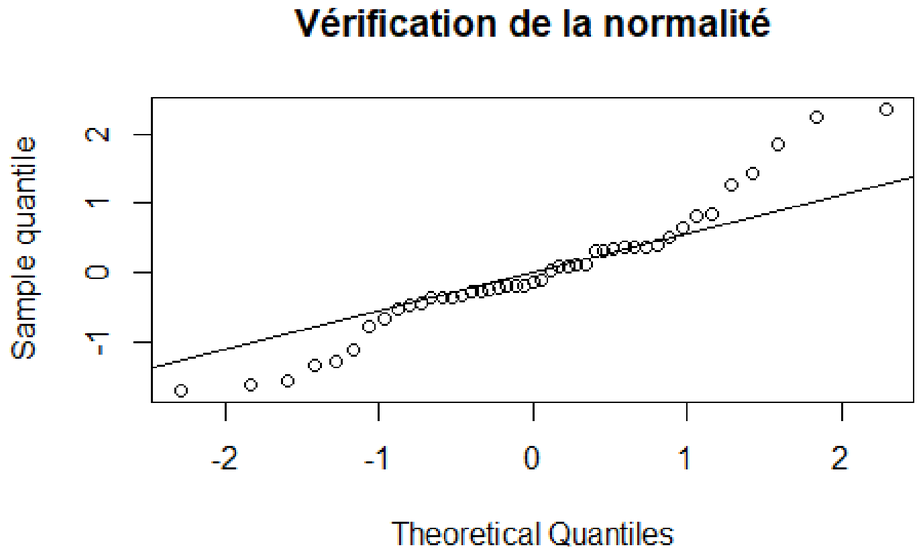
```
library(nlme)
```

```
modM <- lme(Rdt ~ Trt + Saule,~1|Bloc, data = Sophie)
```

```
plot(modM)
```

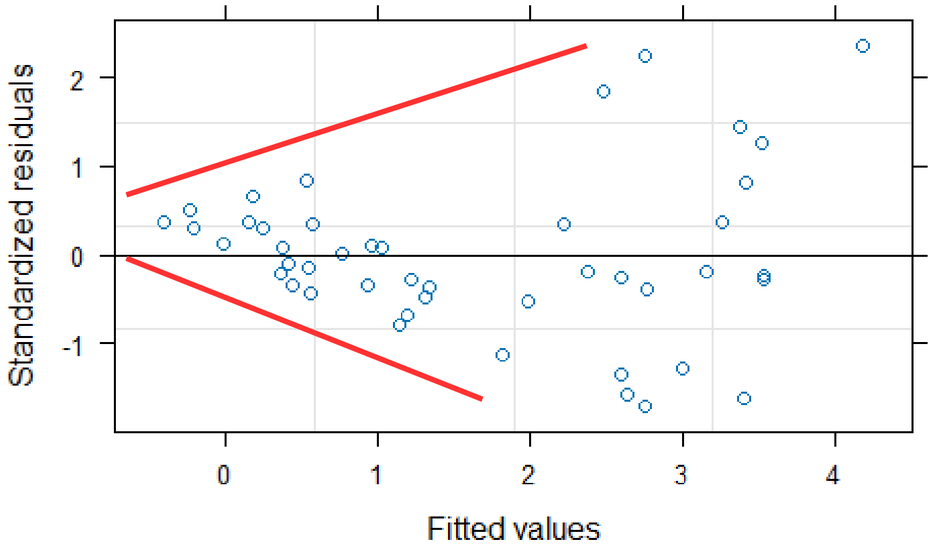


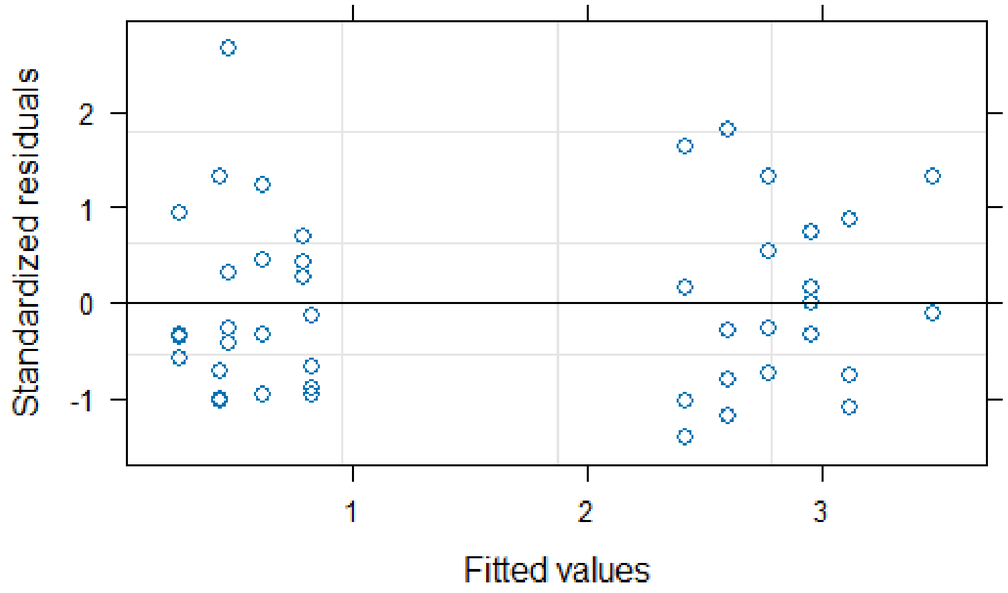

```
residus<-residuals(modM,type = 'pearson')
plot(qqnorm(residus), main = 'Vérification de la normalité',
qqline(residus)
```



# DISPOSITIF EN BLOC

```
modM2 <- lme(Rdt ~ Trt + Saule,~1|Bloc,weights=varConstPower(), data = Sophie)
plot(modM2)
```



# DISPOSITIF EN BLOC

## En modèle mixte

```
modM2 <- lme(Rdt ~ Trt + Saule,~1|Bloc,weights=varConstPower(),
             data = Sophie)
```

```
> anova(modM2)
            numDF denDF  F-value p-value
(Intercept)     1    35 63.99426  <.0001
Trt             5    35  6.84377  0.0002
Saule           1    35  4.74826  0.0361
```

```
> summary(modM2)
Linear mixed-effects model fit by REML
  Data: Sophie
       AIC      BIC    logLik
  128.0837 146.0971 -53.04183

Random effects:
 Formula: ~1 | Bloc
        (Intercept)  Residual
StdDev: 3.036901e-05 0.1313601

Variance function:
 Structure: Constant plus power of variance covariate
 Formula: ~fitted(.)
 Parameter estimates:
   const    power
2.782203 2.282310
Fixed effects:  Rdt ~ Trt + Saule
                    Value Std.Error DF  t-value p-value
(Intercept)     0.2641202 0.1561109 35 1.691876  0.0996
TrtAB           0.1769548 0.2003417 35 0.883265  0.3831
TrtAB-BRF       0.2111179 0.2017497 35 1.046435  0.3025
TrtAB-MRF1-BRF  2.1598475 0.5501184 35 3.926151  0.0004
TrtAB-MRF1-CI   2.3372110 0.6145295 35 3.803253  0.0005
TrtAB-MRF2-BRF  2.8586265 1.0353230 35 2.761096  0.0091
SaulePr         0.3553771 0.1630880 35 2.179051  0.0361
```

Variabilité entre les blocs

Beta_0

Beta_i

# DISPOSITIF EN BLOC

## Résultat

```r
# Obtenir les moyennes marginales estimées pour chaque niveau de Trt
emmeans_modM2 <- emmeans(modM2, ~ Trt)

# Comparaisons multiples avec ajustement de p-valeurs
comparisons <- contrast(emmeans_modM2, method = "pairwise", adjust = "tukey")

# Afficher les résultats
summary(comparisons)
```

```
> summary(comparisons)
 contrast                          estimate    SE df t.ratio p.value
 Témoin - AB                        -0.1770 0.200 35  -0.883  0.9480
 Témoin - (AB-BRF)                  -0.2111 0.202 35  -1.046  0.8987
 Témoin - (AB-MRF1-BRF)             -2.1598 0.550 35  -3.926  0.0048
 Témoin - (AB-MRF1-CI)              -2.3372 0.615 35  -3.803  0.0067
 Témoin - (AB-MRF2-BRF)             -2.8586 1.040 35  -2.761  0.0883
 AB - (AB-BRF)                      -0.0342 0.207 35  -0.165  1.0000
 AB - (AB-MRF1-BRF)                 -1.9829 0.552 35  -3.592  0.0118
 AB - (AB-MRF1-CI)                  -2.1603 0.616 35  -3.505  0.0148
 AB - (AB-MRF2-BRF)                 -2.6817 1.040 35  -2.588  0.1273
 (AB-BRF) - (AB-MRF1-BRF)           -1.9487 0.553 35  -3.527  0.0140
 (AB-BRF) - (AB-MRF1-CI)            -2.1261 0.617 35  -3.447  0.0172
 (AB-BRF) - (AB-MRF2-BRF)           -2.6475 1.040 35  -2.554  0.1363
 (AB-MRF1-BRF) - (AB-MRF1-CI)       -0.1774 0.801 35  -0.221  0.9999
 (AB-MRF1-BRF) - (AB-MRF2-BRF)      -0.6988 1.160 35  -0.605  0.9900
 (AB-MRF1-CI) - (AB-MRF2-BRF)       -0.5214 1.190 35  -0.439  0.9978
```

# DISPOSITIF EN BLOC

## Résultat

```
library(AICcmodavg)
```

```
######## Illustration des résultats
nv <- expand.grid(Bloc = factor(seq(from = 1, to = 4, by = 1)),
                  Trt = factor(c('Témoin', 'AB', 'AB-BRF','AB-MRF1-BRF'
                                 ,'AB-MRF1-CI','AB-MRF2-BRF')),
                  Saule = factor(c('Pr','Sm')))

pred1 <- predictSE.lme(modM2, newdata = nv, se.fit = TRUE)

nv$Pv <- pred1$fit
nv$se <- pred1$se.fit
nv$Rt <- (nv$Pv)


nv$min <- (nv$Pv - 1.96*nv$se)
nv$max <- (nv$Pv + 1.96*nv$se)
```

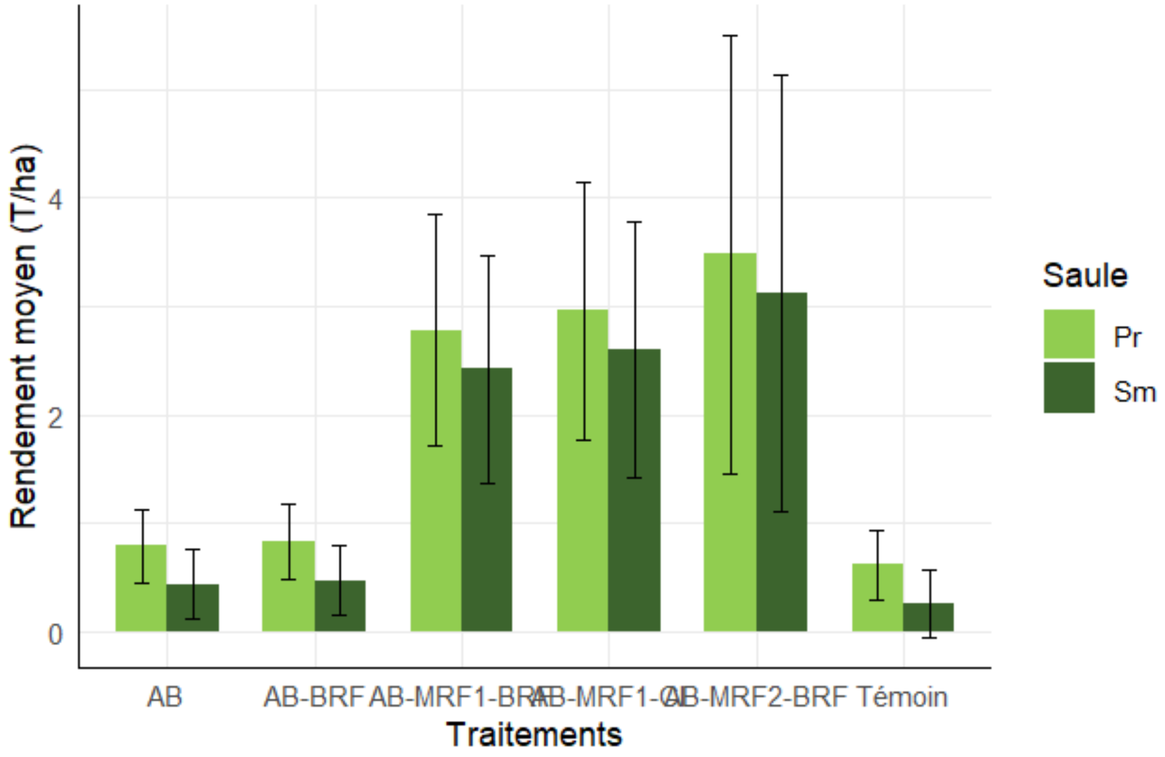
| | Bloc | Trt | Saule | Pv | se | min | max |
|---|---|---|---|---|---|---|---|
| 1 | 1 | Témoin | Pr | 0.6194972 | 0.1641178 | 0.29782628 | 0.9411682 |
| 2 | 2 | Témoin | Pr | 0.6194972 | 0.1641178 | 0.29782628 | 0.9411682 |
| 3 | 3 | Témoin | Pr | 0.6194972 | 0.1641178 | 0.29782628 | 0.9411682 |
| 4 | 4 | Témoin | Pr | 0.6194972 | 0.1641178 | 0.29782628 | 0.9411682 |
| 5 | 1 | AB | Pr | 0.7964521 | 0.1726486 | 0.45806091 | 1.1348433 |
| 6 | 2 | AB | Pr | 0.7964521 | 0.1726486 | 0.45806091 | 1.1348433 |
| 7 | 3 | AB | Pr | 0.7964521 | 0.1726486 | 0.45806091 | 1.1348433 |
| 8 | 4 | AB | Pr | 0.7964521 | 0.1726486 | 0.45806091 | 1.1348433 |
| 9 | 1 | AB-BRF | Pr | 0.8306152 | 0.1746020 | 0.48839534 | 1.1728350 |
| 10 | 2 | AB-BRF | Pr | 0.8306152 | 0.1746020 | 0.48839534 | 1.1728350 |
| 11 | 3 | AB-BRF | Pr | 0.8306152 | 0.1746020 | 0.48839534 | 1.1728350 |
| 12 | 4 | AB-BRF | Pr | 0.8306152 | 0.1746020 | 0.48839534 | 1.1728350 |

# DISPOSITIF EN BLOC

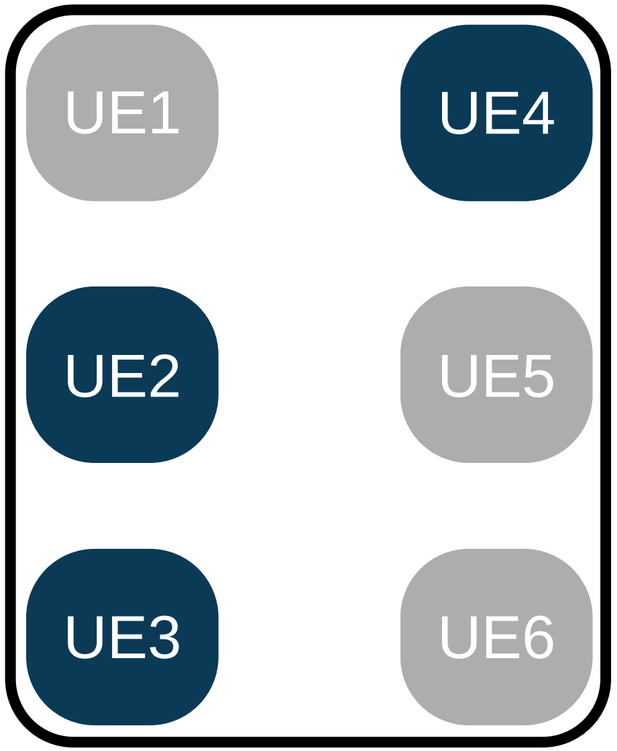
## Résultat

```
library(AICcmodavg)
```

```
######## Illustration des résultats
nv <- expand.grid(Bloc = factor(seq(from = 1, to = 4, by = 1)),
                  Trt = factor(c('Témoin', 'AB', 'AB-BRF','AB-MRF1-BRF'
                                 ,'AB-MRF1-CI','AB-MRF2-BRF')),
                  Saule = factor(c('Pr','Sm')))

pred1 <- predictSE.lme(modM2, newdata = nv, se.fit = TRUE)

nv$Pv <- pred1$fit
nv$se <- pred1$se.fit
nv$Rt <- (nv$Pv)


nv$min <- (nv$Pv - 1.96*nv$se)
nv$max <- (nv$Pv + 1.96*nv$se)
```



# DISPOSITIF EN MESURES RÉPÉTÉES



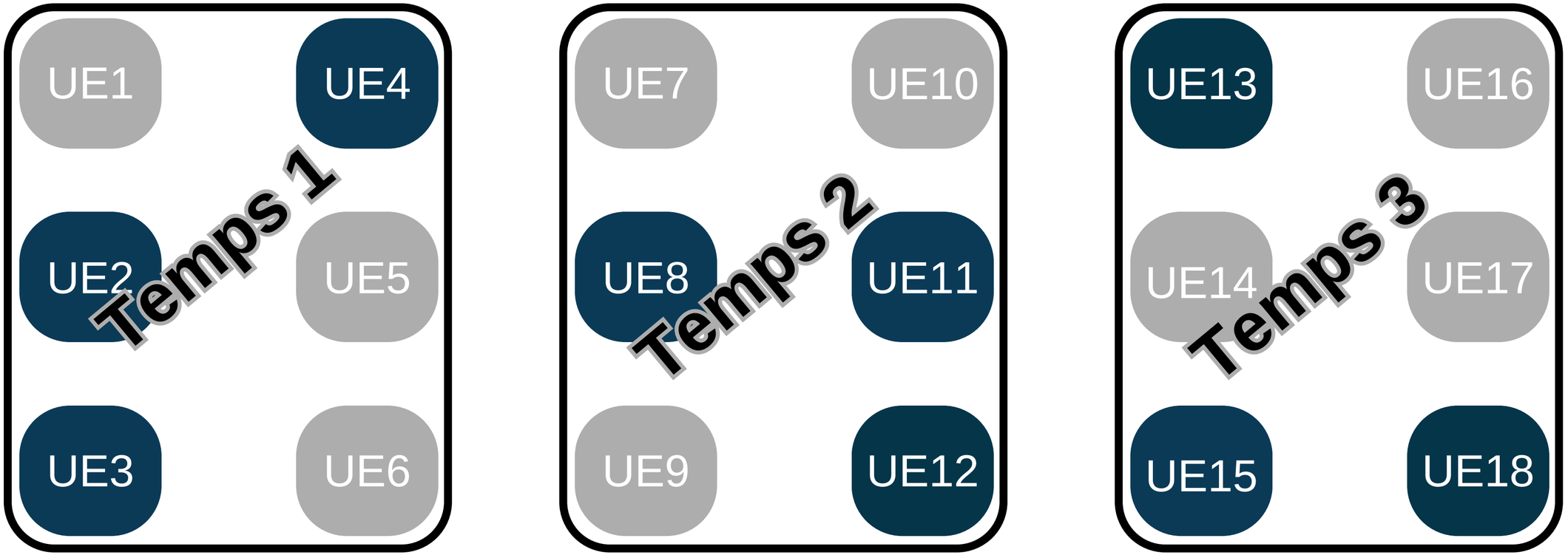

Mesures de l'N
Temps 1
Temps 2
Temps 3
...

Problème si on utilise une régression linéaire ?

Est-ce possible de régler ce problème en changeant la conception du dispositif ?

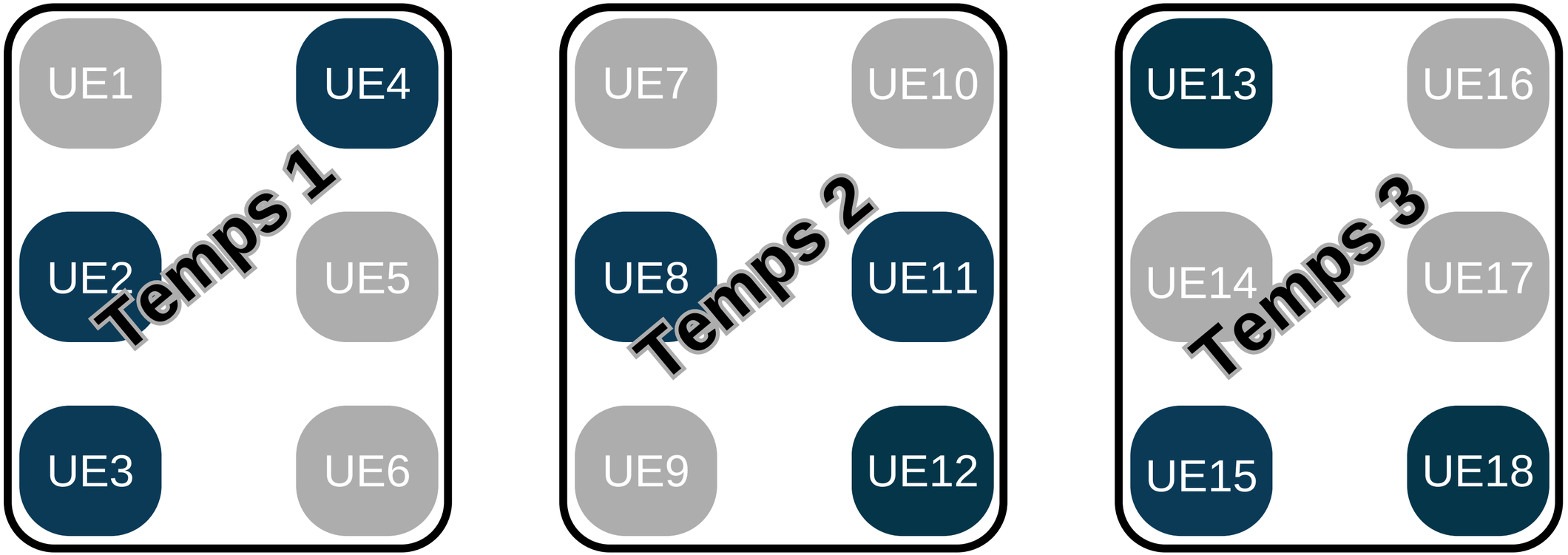
# DISPOSITIF EN MESURES RÉPÉTÉES

Méthode alternative

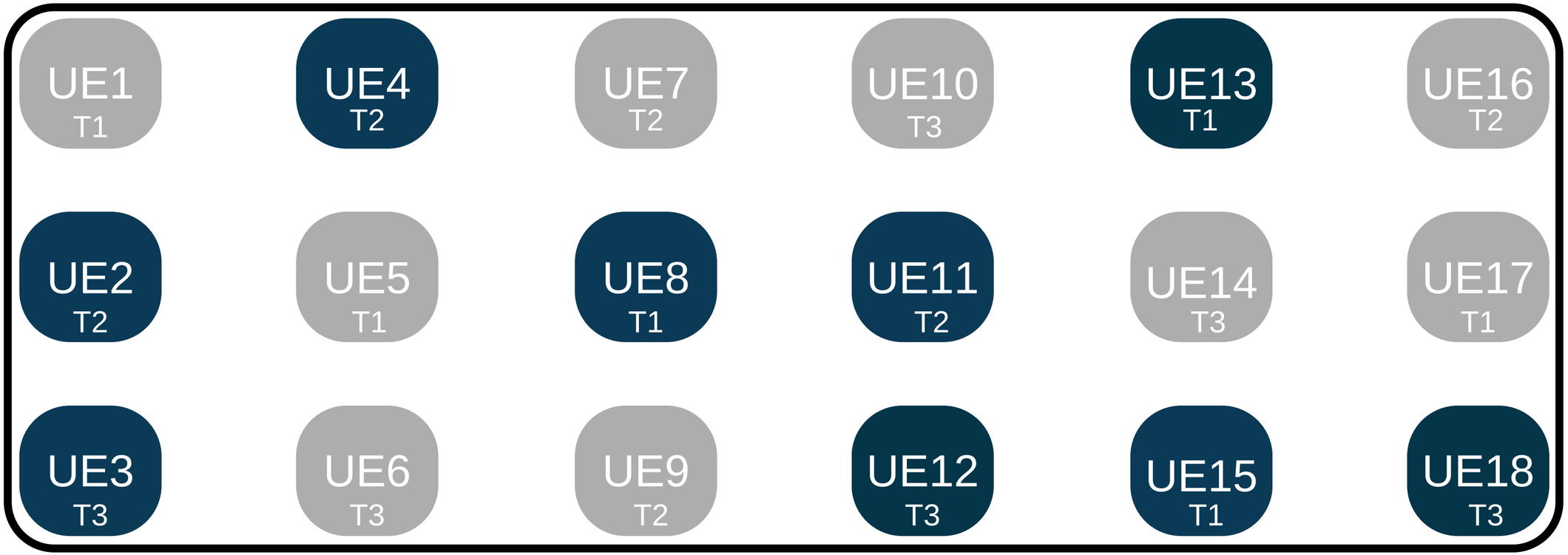


# DISPOSITIF EN MESURES RÉPÉTÉES

Problème ?



# DISPOSITIF EN MESURES RÉPÉTÉES



# DISPOSITIF EN MESURES RÉPÉTÉES

```
> head(Flux,10)
# A tibble: 10 × 9
   Semaine Colonne Sol    BRF   Plante  Tair  Tsol   TEE     Rt
     <dbl> <fct>   <fct>  <fct> <fct>  <dbl> <dbl> <dbl>  <dbl>
 1       1 2       Argile Non   Non     23.3  23.5  24.5 116.
 2       1 3       Loam   Oui   Oui     23.3  23.5  33.3 546.
 3       1 4       Sable  Non   Non     23.3  23.3  21.6   8.14
 4       1 5       Loam   Oui   Non     23    23.3  25.3 493.
 5       1 6       Argile Oui   Non     23    23.3  21.3 452.
 6       1 7       Loam   Non   Oui     23    23.9  31.1 170.
 7       1 8       Loam   Oui   Oui     23.3  24.1  24.4 728.
 8       1 9       Argile Non   Oui     23    24.1  19.4 212.
 9       1 10      Sable  Oui   Non     23.3  24    26.1 329.
10       1 11      Loam   Non   Non     23.3  23.2  29.3 131.
```

# DISPOSITIF EN MESURES RÉPÉTÉES

```
> head(Flux,10)
# A tibble: 10 × 9
   Semaine Colonne Sol    BRF   Plante  Tair  Tsol   TEE     Rt
     <dbl> <fct>   <fct>  <fct> <fct>  <dbl> <dbl> <dbl>  <dbl>
 1       1 2       Argile Non   Non     23.3  23.5  24.5 116.
 2       1 3       Loam   Oui   Oui     23.3  23.5  33.3 546.
 3       1 4       Sable  Non   Non     23.3  23.3  21.6   8.14
 4       1 5       Loam   Oui   Non     23    23.3  25.3 493.
 5       1 6       Argile Oui   Non     23    23.3  21.3 452.
 6       1 7       Loam   Non   Oui     23    23.9  31.1 170.
 7       1 8       Loam   Oui   Oui     23.3  24.1  24.4 728.
 8       1 9       Argile Non   Oui     23    24.1  19.4 212.
 9       1 10      Sable  Oui   Non     23.3  24    26.1 329.
10       1 11      Loam   Non   Non     23.3  23.2  29.3 131.
```

**Effets fixes ?**

**Effets aléatoires ?**

# DISPOSITIF EN MESURES RÉPÉTÉES

```
> head(Flux,10)
# A tibble: 10 × 9
   Semaine Colonne Sol    BRF   Plante  Tair  Tsol   TEE     Rt
     <dbl> <fct>   <fct>  <fct> <fct>  <dbl> <dbl> <dbl>  <dbl>
 1       1 2       Argile Non   Non     23.3  23.5  24.5 116.
 2       1 3       Loam   Oui   Oui     23.3  23.5  33.3 546.
 3       1 4       Sable  Non   Non     23.3  23.3  21.6   8.14
 4       1 5       Loam   Oui   Non     23    23.3  25.3 493.
 5       1 6       Argile Oui   Non     23    23.3  21.3 452.
 6       1 7       Loam   Non   Oui     23    23.9  31.1 170.
 7       1 8       Loam   Oui   Oui     23.3  24.1  24.4 728.
 8       1 9       Argile Non   Oui     23    24.1  19.4 212.
 9       1 10      Sable  Oui   Non     23.3  24    26.1 329.
10       1 11      Loam   Non   Non     23.3  23.2  29.3 131.
```

## Effets fixes ?

**Semaine + Sol + BRF +Plante +Tair +TEE**

## Effets aléatoires ?

**Colonne**

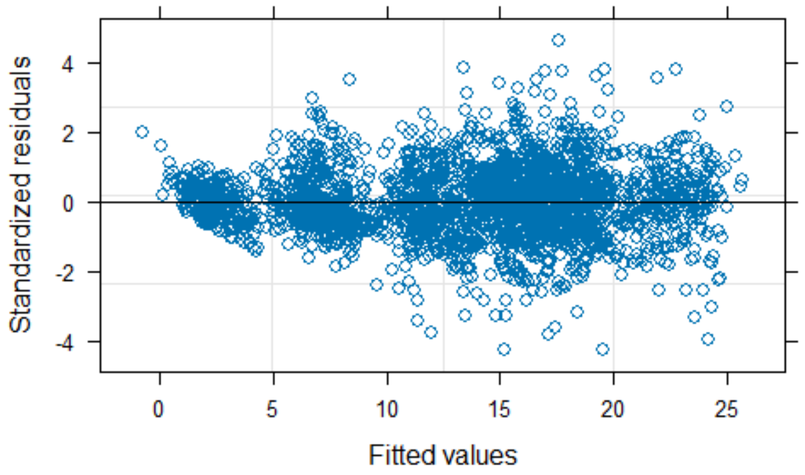
# DISPOSITIF EN MESURES RÉPÉTÉES

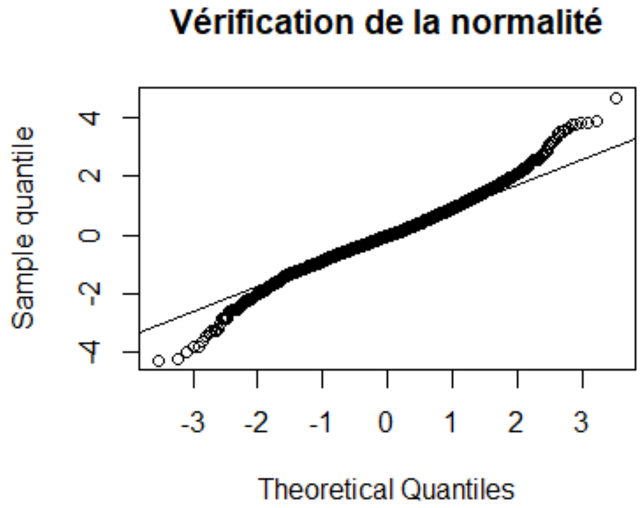
```
modR <- lme(Rt~Semaine+Sol+Plante+BRF+Tair+TEE,
            random =~1|Colonne,
            data=Flux,na.action=na.exclude)
```

# DISPOSITIF EN MESURES RÉPÉTÉES

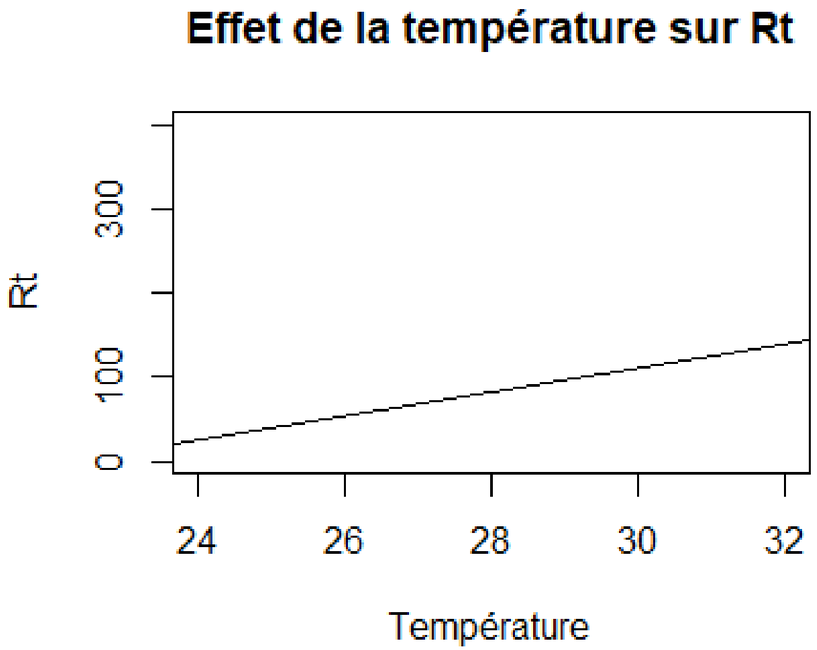
## Vérification des postulats

```
plot(modR)
residusR<-residuals(modR,type = 'pearson')
plot(qqnorm(residusR), main = 'Vérification de la normalité',
qqline(residusR)
```



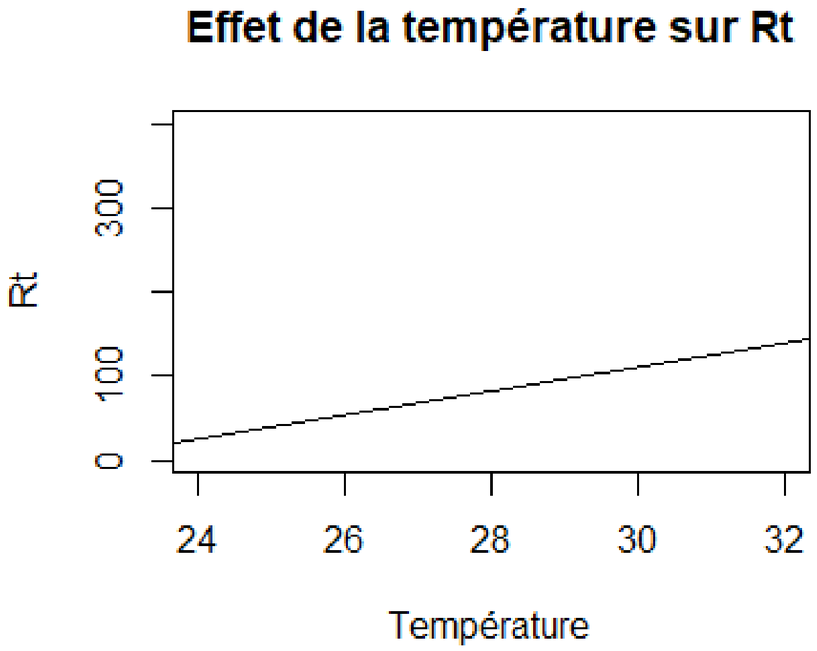



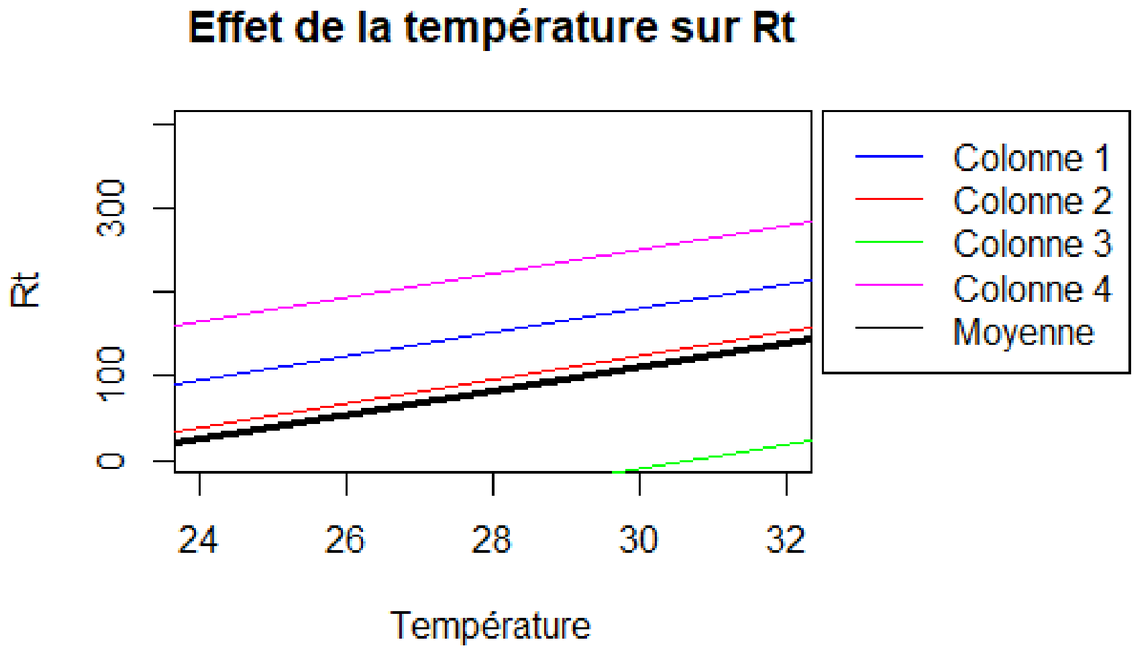
# DISPOSITIF EN MESURES RÉPÉTÉES



# DISPOSITIF EN MESURES RÉPÉTÉES

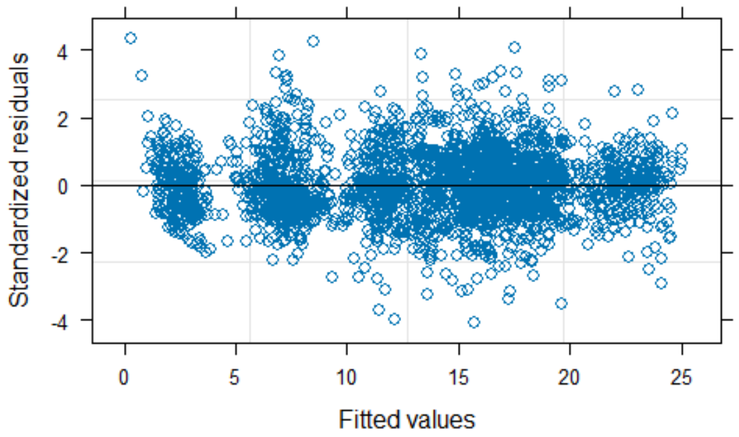
```
random =~1|Colonne,
```



# DISPOSITIF EN MESURES RÉPÉTÉES

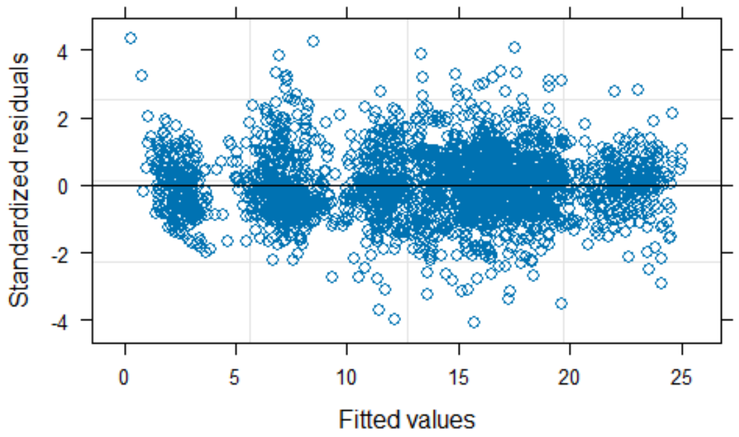
```
modR <- lme(Rtt~Semaine+Sol+Plante+BRF+Tair+TEE,
            random =~1|Colonne,weights=varConstPower(),
            data=Flux,na.action=na.exclude)
```




```
Linear mixed-effects model fit by REML
  Data: Flux
       AIC      BIC    logLik
  12215.42 12285.19 -6095.712

Random effects:
 Formula: ~1 | Colonne
        (Intercept)  Residual
StdDev:    1.874998 0.3160978

Variance function:
 Structure: Constant plus power of variance covariate
 Formula: ~fitted(.)
 Parameter estimates:
    const     power
2.8910385 0.6952419
Fixed effects:  Rtt ~ Semaine + Sol + Plante + BRF + Tair + TEE
                Value Std.Error   DF   t-value p-value
(Intercept) -1.300805 0.8593330 2407 -1.513739  0.1302
Semaine     -0.016794 0.0044837 2407 -3.745619  0.0002
SolLoam      1.789852 0.5665312   67  3.159318  0.0024
SolSable    -3.201163 0.5628075   67 -5.687847  0.0000
PlanteOui    6.512809 0.4623359   67 14.086747  0.0000
BRFOui       8.171785 0.4628007   67 17.657247  0.0000
Tair         0.337734 0.0288167 2407 11.720074  0.0000
TEE         -0.000179 0.0084515 2407 -0.021142  0.9831
```

# DISPOSITIF EN MESURES RÉPÉTÉES

```
modR <- lme(Rtt~Semaine+Sol+Plante+BRF+Tair+TEE,
            random =~1|Colonne,weights=varConstPower(),
            data=Flux,na.action=na.exclude)
```




```
Linear mixed-effects model fit by REML
  Data: Flux
       AIC      BIC    logLik
  12215.42 12285.19 -6095.712

Random effects:
 Formula: ~1 | Colonne
        (Intercept)  Residual
StdDev:    1.874998 0.3160978

Variance function:
 Structure: Constant plus power of variance covariate
 Formula: ~fitted(.)
 Parameter estimates:
    const     power
2.8910385 0.6952419
Fixed effects:  Rtt ~ Semaine + Sol + Plante + BRF + Tair + TEE
                Value Std.Error   DF   t-value p-value
(Intercept) -1.300805 0.8593330 2407 -1.513739  0.1302
Semaine     -0.016794 0.0044837 2407 -3.745619  0.0002
SolLoam      1.789852 0.5665312   67  3.159318  0.0024
SolSable    -3.201163 0.5628075   67 -5.687847  0.0000
PlanteOui    6.512809 0.4623359   67 14.086747  0.0000
BRFOui       8.171785 0.4628007   67 17.657247  0.0000
Tair         0.337734 0.0288167 2407 11.720074  0.0000
TEE         -0.000179 0.0084515 2407 -0.021142  0.9831
```

$$\frac{1.8749^2}{(1.8749^2 + 0.3160^2)} = 97\ \%$$

ne doit pas être top petit (10E-3) car problème d'estimation de la variance

# STRUCTURES HIÉRARCHIQUES

## Avec effets nichés (imbriqués)

### Dispositif en tiroirs



Semances

S1

S2

```
modT <- lme(y ~ Semance + Fertilisant,
            random =~1|Bloc/Fertilisant,
            data=Ferti,na.action=na.exclude)
```

# STRUCTURES CROISÉS

**Dispositif en carré latin**

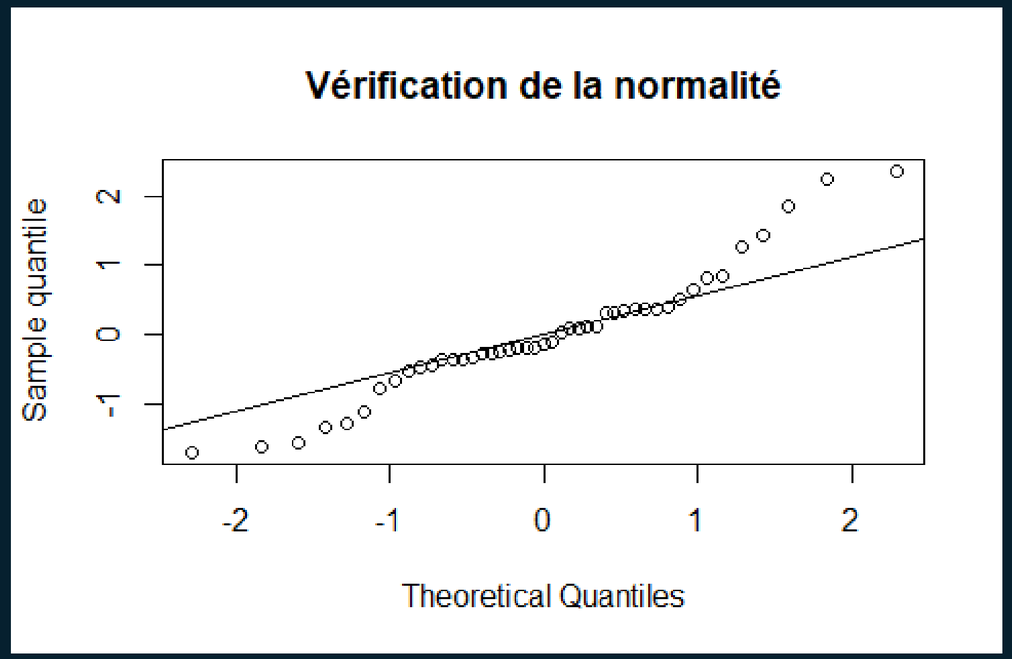
| | | | | |
|---|---|---|---|---|
| B | E | D | C | A |
| A | D | E | B | C |
| D | A | C | E | B |
| C | B | A | D | E |
| E | C | B | A | D |

```
library(lme4)
```

```
modele_mixe <- lmer(y_i ~ Traitement + (1 | Ligne) + (1 | Colonne),
                    data = carre_latin_df)
```

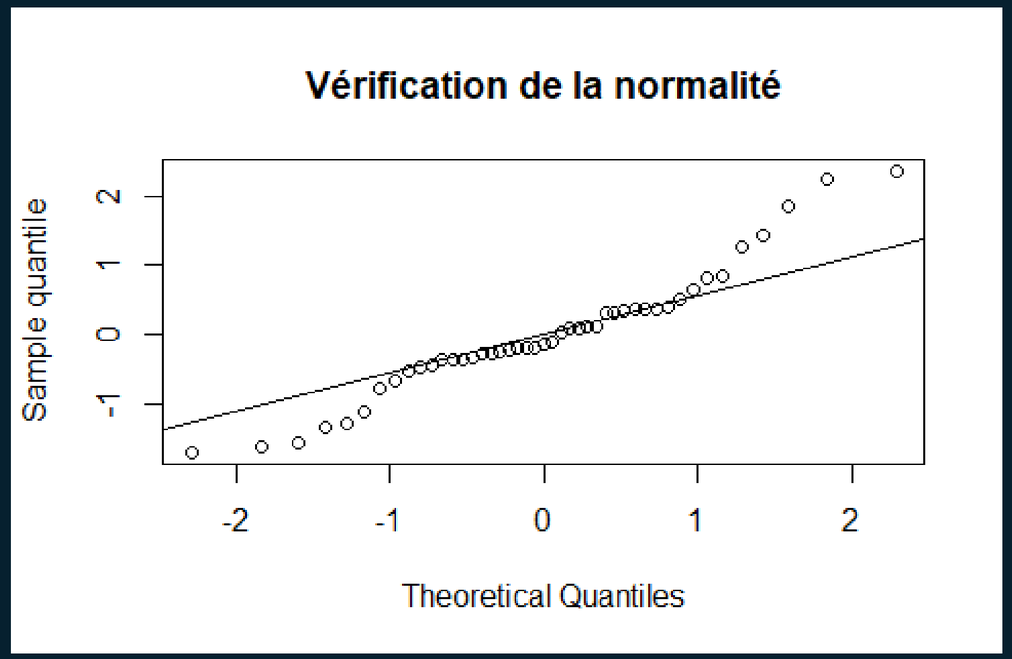
```
> head(carre_latin_df,10)
   Ligne Colonne Traitement      y_i
1      1       1          A 11.37096
2      1       2          B 16.51152
3      1       3          C 18.61114
4      1       4          D 22.55953
5      1       5          E 31.89519
6      2       1          B 14.89388
7      2       2          C 22.28665
8      2       3          D 22.34354
9      2       4          E 31.21467
10     2       5          A 10.40427
```

# NORMALITÉ DES DONNÉES

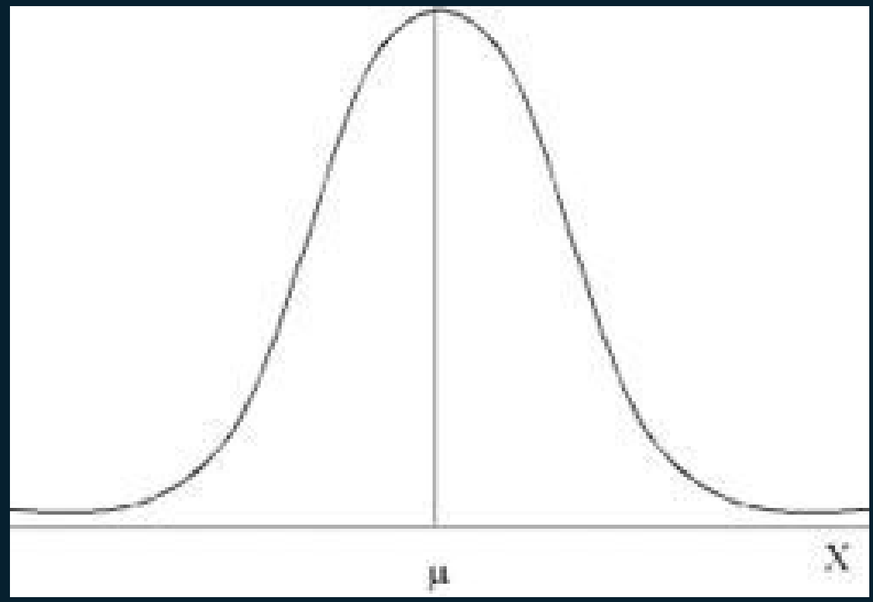



On fait quoi avec une non normalité ?

# NORMALITÉ DES DONNÉES



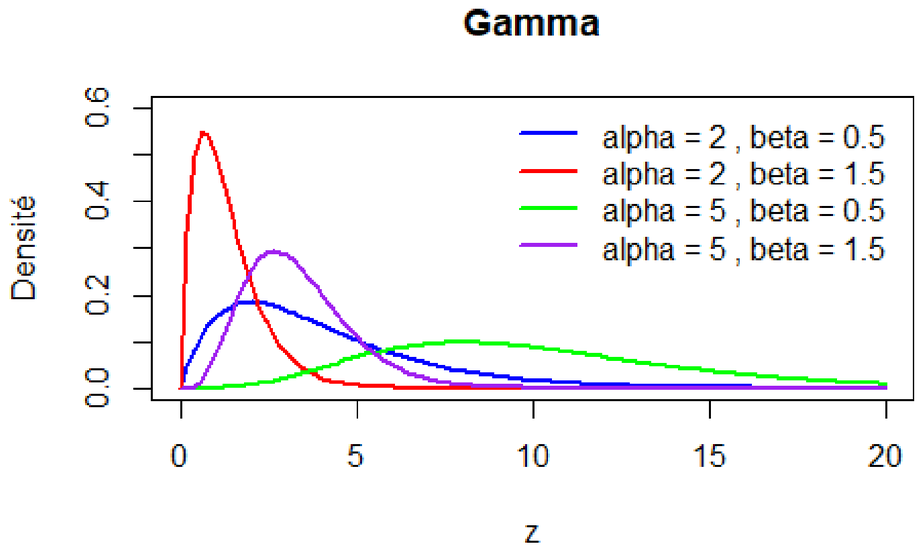

Est-ce que mes données font du sens avec la définition d’une loi normale ?




On fait quoi avec une non normalité ?

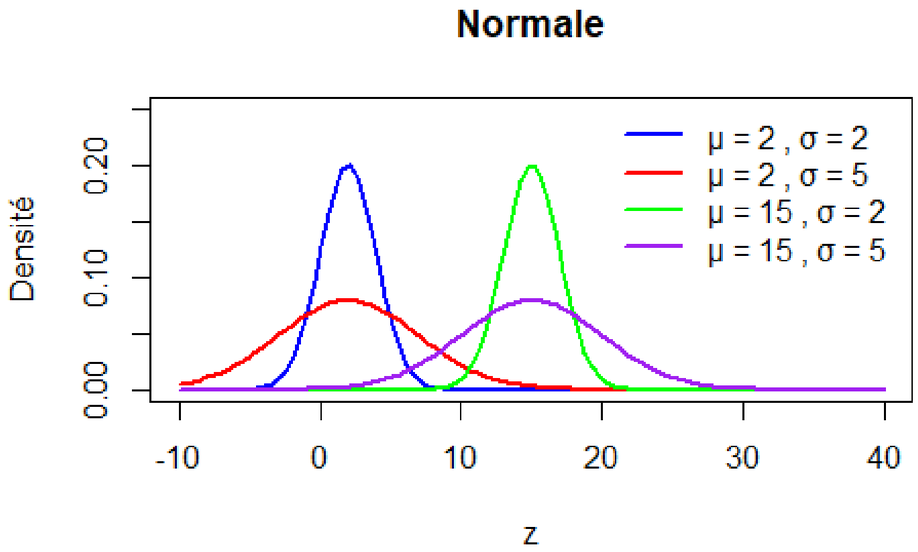
# UTILISER UNE AUTRE DISTRIBUTION : LES MODÈLES GÉNÉRALISÉS

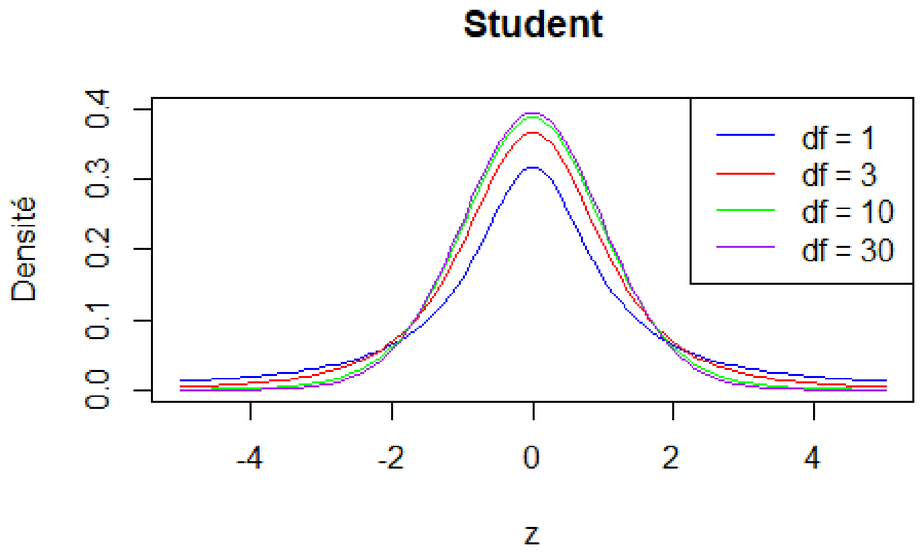
# LES DISTRIBUTION POUR VALEURS CONTINUES



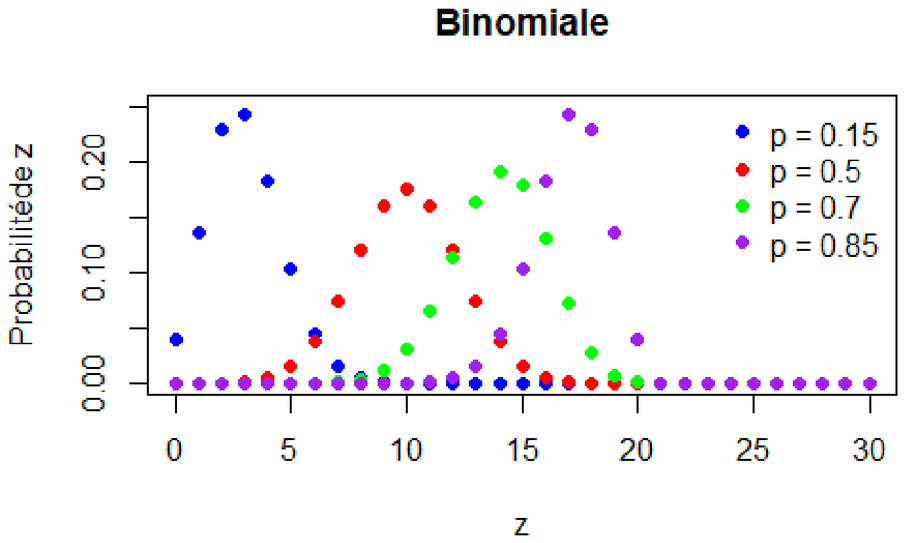

Exemples:

- Rendement
- Distance
- Poid



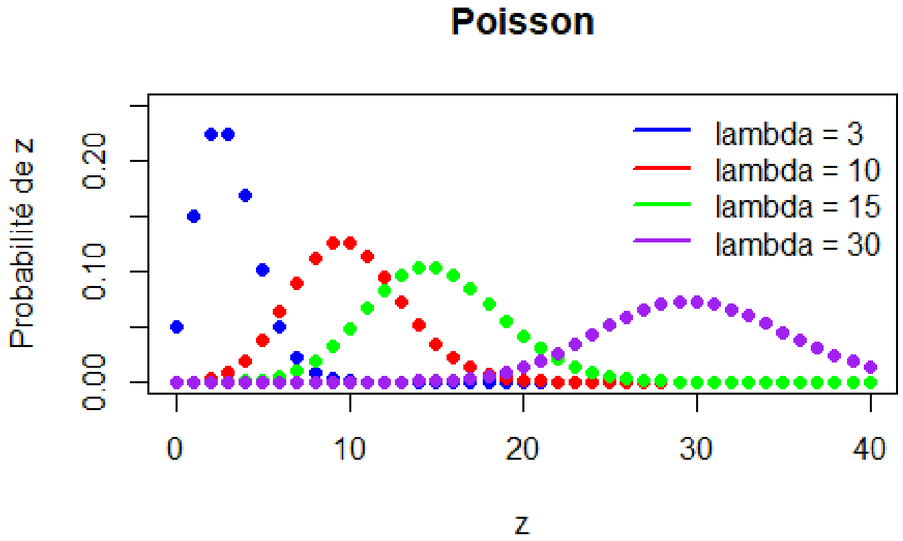



# LES DISTRIBUTION POUR VALEURS DISCRÈTES



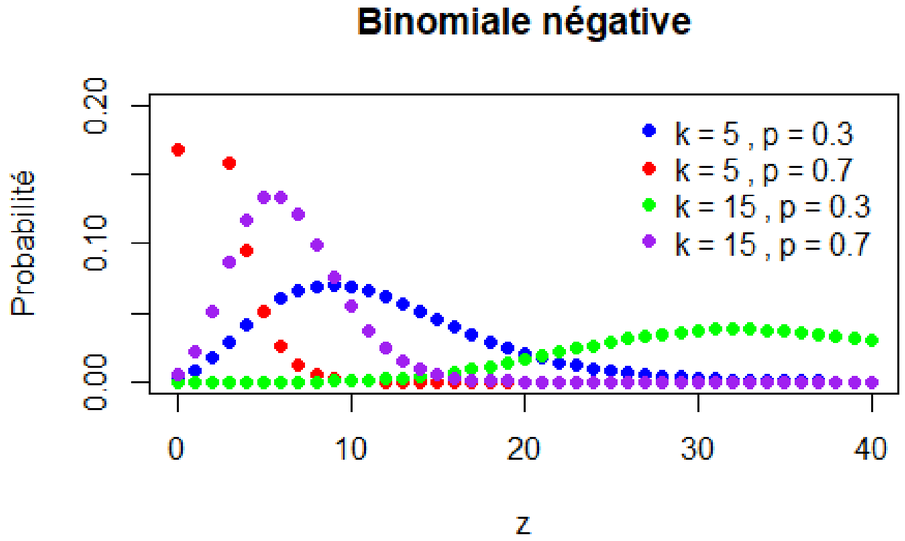

Exemples:

- Vivant/mort
- Bon/mauvais
- Santé/malade
- Germé/non




Exemples:

- Insecte
- Nombre d'espèce
- Nombre de mauvais herbe



# EXEMPLE AVEC GAMMA

## En bloc complet

```
modG <- glmer(Rdt ~ Trt + Saule + (1 | Bloc),
              data = Sophie,
              family = Gamma(link = "log"))
```

```
> head(Sophie,10)
# A tibble: 10 × 4
   Bloc  Trt          Saule  Rdt
   <fct> <fct>        <fct>  <dbl>
 1 1     AB           Pr     0.919
 2 1     AB           Sm     0.0424
 3 1     AB-MRF1-BRF  Pr     3.70
 4 1     AB-MRF1-BRF  Sm     4.63
 5 1     AB-MRF1-CI   Pr     4.38
 6 1     AB-MRF1-CI   Sm     0.796
 7 1     Témoin       Pr     1.13
 8 1     Témoin       Sm     0.615
 9 1     AB-MRF2-BRF  Pr     6.95
10 1     AB-MRF2-BRF  Sm     1.51
```

```
> summary(modG)
Generalized linear mixed model fit by maximum likelihood (Laplace Approximation) ['glmerMod']
 Family: Gamma  ( log )
Formula: Rdt ~ Trt + Saule + (1 | Bloc)
   Data: Sophie

     AIC      BIC   logLik deviance df.resid
   111.0    127.3    -46.5     93.0       36

Scaled residuals:
    Min      1Q  Median      3Q     Max
-1.3894 -0.7906 -0.1699  0.5041  2.7903

Random effects:
 Groups   Name        Variance Std.Dev.
 Bloc     (Intercept) 0.04431  0.2105
 Residual             0.42719  0.6536
Number of obs: 45, groups:  Bloc, 4

Fixed effects:
             Estimate Std. Error t value Pr(>|z|)
(Intercept)   -1.2526     0.3060  -4.094 4.25e-05 ***
TrtAB          0.3913     0.3365   1.163  0.24491
TrtAB-BRF      0.5037     0.3445   1.462  0.14367
TrtAB-MRF1-BRF 1.8728     0.3318   5.645 1.66e-08 ***
TrtAB-MRF1-CI  1.9571     0.3382   5.786 7.20e-09 ***
TrtAB-MRF2-BRF 2.1729     0.3820   5.688 1.28e-08 ***
SaulePr        0.6017     0.2072   2.903  0.00369 **
```

# EXEMPLE AVEC STUDENT

```
mod <- lm(y ~ TEE + Traitement, data =data)
```

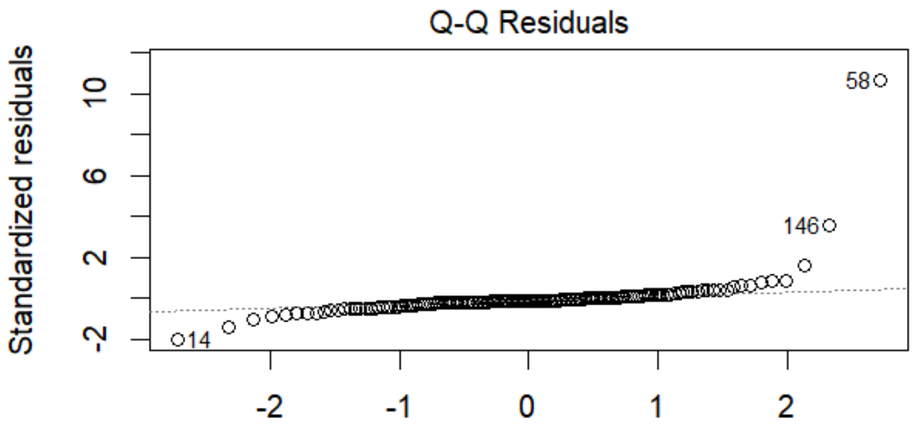
```
> head(data,10)
           y       TEE Traitement
1  11.741277 36.67207          A
2  14.993483 47.54224          A
3  10.480904 32.99427          A
4  14.289005 37.66846          A
5   2.138507 13.59034          A
6  10.875119 44.65837          A
7   8.626283 30.04558          A
8   8.614412 18.93025          A
9   8.854446 26.49677          A
10 14.329603 45.45225          A
```

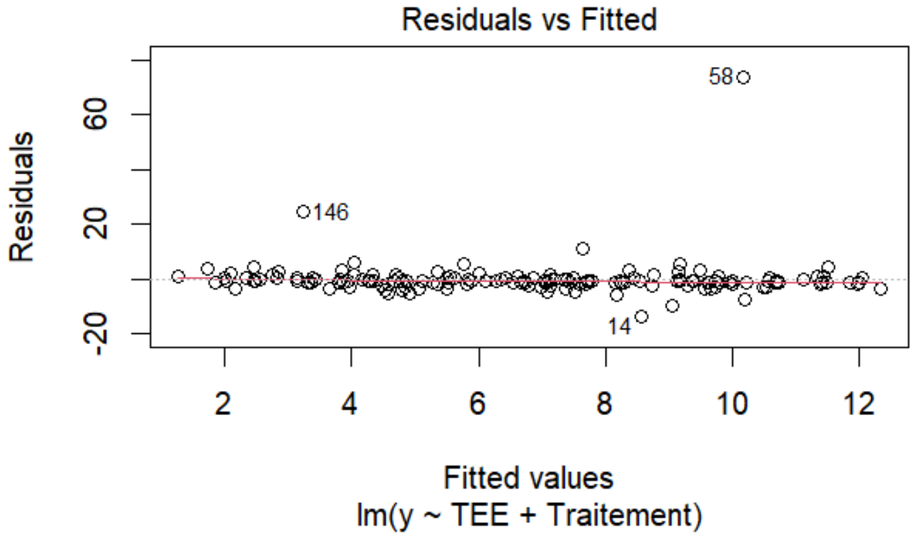
```
> summary(mod)

Call:
lm(formula = y ~ TEE + Traitement, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-13.791  -1.429  -0.759   0.456  73.879

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.19104    1.72159   0.692    0.490
TEE          0.22376    0.04986   4.488 1.44e-05 ***
TraitementB -2.24665    1.14160  -1.968    0.051 .
```

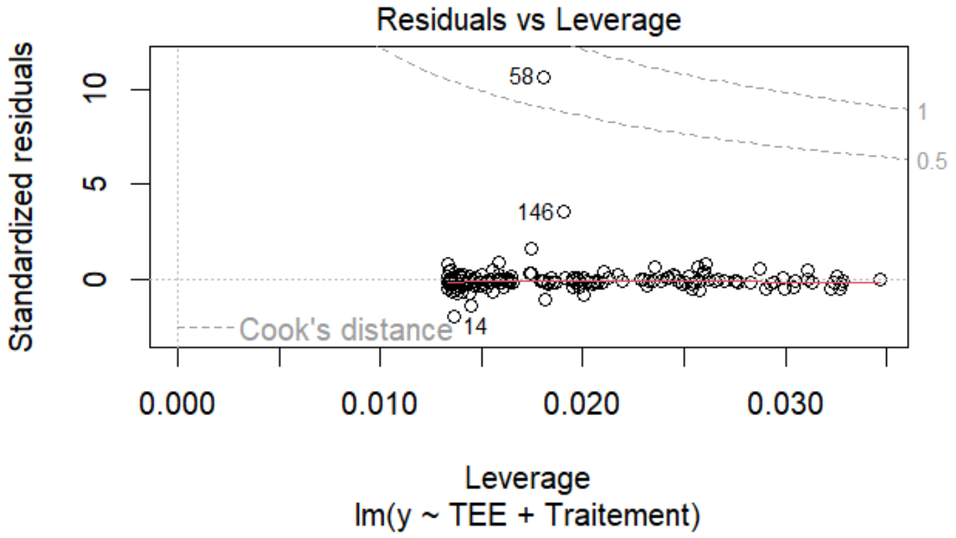
# EXEMPLE AVEC STUDENT

## Vérifications des suppositions

```
plot(mod)
```



# EXEMPLE AVEC STUDENT

```
library(hett)
modS <- tlm(y ~ TEE + Traitement, data = data,
            estDof = TRUE)
```

```
> summary(modS)
Location model :

Call:
tlm(lform = y ~ TEE + Traitement, data = data, estDof = TRUE)

Residuals:
      Min         1Q     Median         3Q        Max
-12.87298   -0.78143   -0.03788    1.07149   74.95380

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.987463   0.342239   2.885   0.0045 **
TEE          0.202064   0.009912  20.386  < 2e-16 ***
TraitementB -1.952782   0.226941  -8.605 1.07e-14 ***
```

# EXEMPLE AVEC BETA

```
> head(data,10)
   Ferme     Travail          MO
1      1      Labour 0.05399735
2      1 Non-Labour 0.10551937
3      1      Labour 0.09901891
4      1 Non-Labour 0.10583416
5      1      Labour 0.06182087
6      1 Non-Labour 0.10517067
7      1      Labour 0.05430205
8      1 Non-Labour 0.09816151
9      1      Labour 0.08850356
10     1 Non-Labour 0.07683616
```

```
# Modèle mixte bêta
modB <- glmmTMB(
  MO ~ Travail + (1 | Ferme),
  data = data,
  family = beta_family(link = "logit")
```

# EXEMPLE AVEC BETA

```
> head(data,10)
   Ferme     Travail          MO
1      1      Labour 0.05399735
2      1 Non-Labour 0.10551937
3      1      Labour 0.09901891
4      1 Non-Labour 0.10583416
5      1      Labour 0.06182087
6      1 Non-Labour 0.10517067
7      1      Labour 0.05430205
8      1 Non-Labour 0.09816151
9      1      Labour 0.08850356
10     1 Non-Labour 0.07683616
```

```
> summary(modB)
 Family: beta  ( logit )
Formula:          MO ~ Travail + (1 | Ferme)
Data: data

     AIC      BIC   logLik deviance df.resid
  -788.5   -776.4    398.2   -796.5      146

Random effects:

Conditional model:
 Groups Name        Variance Std.Dev.
 Ferme  (Intercept) 0.002981 0.0546
Number of obs: 150, groups:  Ferme, 5

Dispersion parameter for beta family ():  255

Conditional model:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)       -2.60581    0.03742  -69.63   <2e-16 ***
TravailNon-Labour  0.33038    0.03744    8.82   <2e-16 ***
```

# EXEMPLE AVEC BETA

$$\beta_0 = -2.6058$$

$$\beta_1 = 0.3304$$

$$MO_{labour} = \frac{\exp(\beta_0)}{1 + \exp(\beta_0)} = 0.0687 = 6.87\%$$

$$MO_{Non-labour} = \frac{\exp(\beta_0 + \beta_1)}{1 + \exp(\beta_0 + \beta_1)} = 0.09318 = 9.32\%$$

# EXEMPLE AVEC POISSON

```
modP <- glmer(Insecte ~ Trt + Temps + (1|UE),
                        family = poisson(),
                        data = data)
```

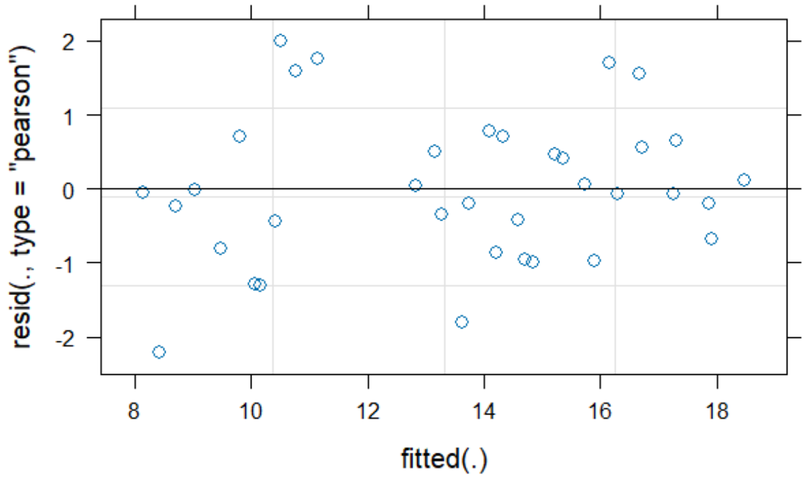
```
> head(data,10)
      Trt UE Temps Insecte
1  Trt 1  1     1       8
2  Trt 1  1     2       2
3  Trt 1  1     3       8
4  Trt 1  1     4       9
5  Trt 1  2     1       6
6  Trt 1  2     2       9
7  Trt 1  2     3      16
8  Trt 1  2     4      17
9  Trt 1  3     1       7
10 Trt 1  3     2      12
```

```
> summary(modP)
Generalized linear mixed model fit by maximum likelihood (Laplace Approximation) ['glmerMod']
 Family: poisson  ( log )
Formula: Insecte ~ Trt + Temps + (1 | UE)
   Data: data

     AIC      BIC   logLik deviance df.resid
   212.3    220.2   -101.1    202.3       31

Scaled residuals:
    Min      1Q  Median      3Q     Max
-1.9154 -0.4847 -0.2332  0.3914  1.8758

Random effects:
 Groups Name        Variance Std.Dev.
 UE     (Intercept) 0.05777  0.2403
Number of obs: 36, groups:  UE, 9

Fixed effects:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.87792    0.20349   9.228  < 2e-16 ***
TrtTrt 2     0.70301    0.23239   3.025  0.00249 **
TrtTrt 3     0.51724    0.23458   2.205  0.02746 *
Temps        0.08595    0.04124   2.084  0.03712 *
```

# EXEMPLE AVEC POISSON

## Vérifications des suppositions

```
plot(modP)
```




## surdispersion

```
#Calcul de la statistique chi-carré (somme des résidus de Pearson au carré)
chi2 <- sum(residuals(modP, type = "pearson")^2)

# Calcul de c_hat (surdispersion) en divisant par les degrés de liberté résiduels
ddl_residuels <- df.residual(modP)
c_hat <- chi2 / ddl_residuels

# Affichage de c_hat
c_hat
```

```
> c_hat
[1] 1.11965
```

**Ok**

**entre 1 et 4 - ajuster la surdispersion du modèle**

**> 4 - utiliser une autre distribution (binomiale-négative)**

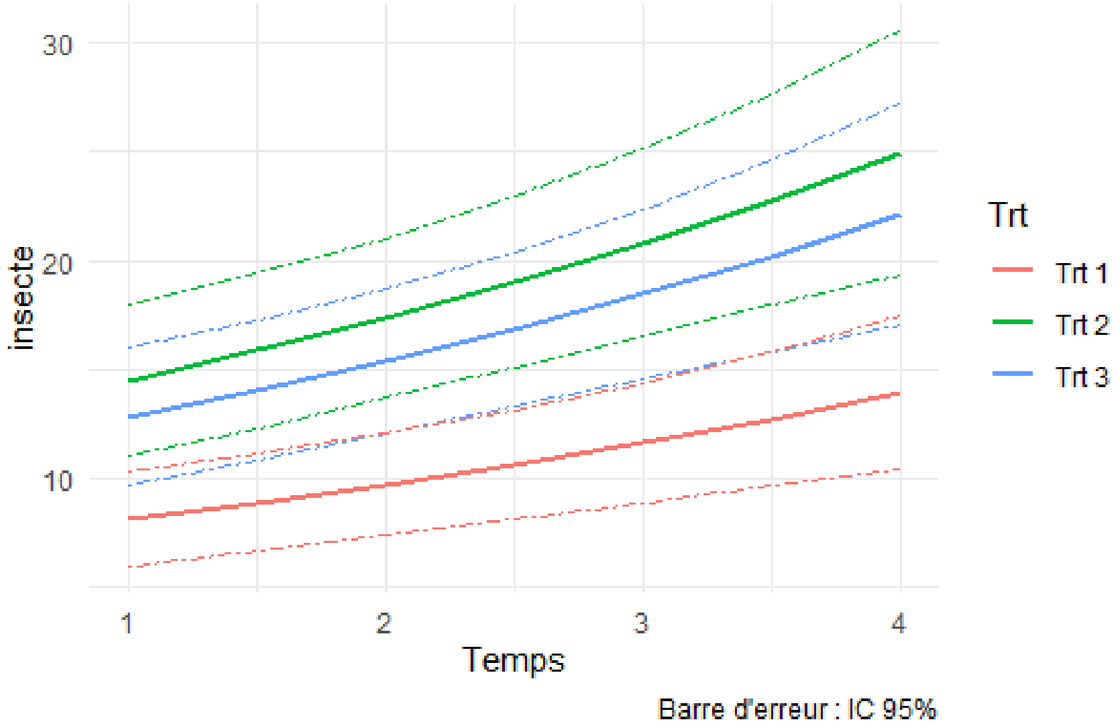
# EXEMPLE AVEC POISSON

## Résultats

```
library(multcomp) #comparaison multiple
```

```
# Comparaison des niveaux de Trt dans un modèle glm
glht_res <- glht(modP, linfct = mcp(Trt = "Tukey"))

# Résultats des comparaisons multiples
summary(glht_res)
```

```
Linear Hypotheses:
                   Estimate Std. Error z value Pr(>|z|)
Trt 2 - Trt 1 == 0   0.5835     0.1594   3.660   <0.001 ***
Trt 3 - Trt 1 == 0   0.4631     0.1610   2.876   0.0113 *
Trt 3 - Trt 2 == 0  -0.1204     0.1499  -0.803   0.7007
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
(Adjusted p values reported -- single-step method)
```



# EXEMPLE AVEC BINOMIALE

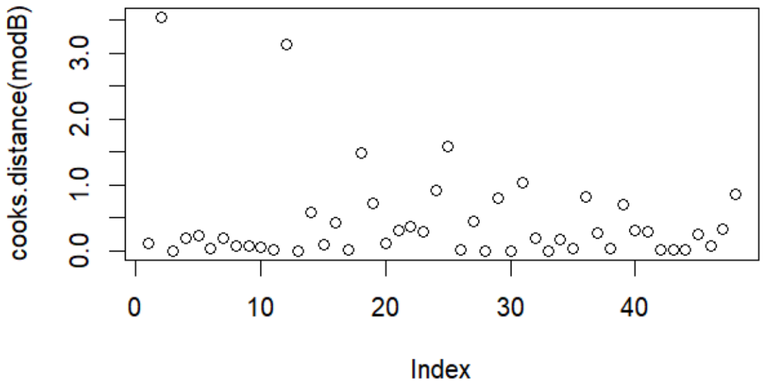
## (régression logistique)

```
# A tibble: 10 × 5
   Bloc  Trt           Saule Survie    Nb
   <fct> <fct>         <fct>   <dbl> <dbl>
 1 1     AB            Pr      0.577   248
 2 1     AB            Sm      0.113   248
 3 1     AB-MRF1-BRF   Pr      0.847   248
 4 1     AB-MRF1-BRF   Sm      0.734   248
 5 1     AB-MRF1-CI    Pr      0.75    248
 6 1     AB-MRF1-CI    Sm      0.673   248
 7 1     Témoin        Pr      0.722   248
 8 1     Témoin        Sm      0.383   248
 9 1     AB-MRF2-BRF   Pr      0.847   248
10 1     AB-MRF2-BRF   Sm      0.657   248
```

```
modB <- glm(Survie ~ Trt+Bloc+Saule,
            family = binomial(link = logit),
            weights = Nb, data = Sophie)
```

```
> c_hat(modB) #library(AICcmodavg)
'c-hat' 17.11 (method: pearson estimator)
```

```
plot(cooks.distance(modB)) #Valeurs extrêmes
```



# EXEMPLE AVEC BINOMIALE NÉGATIVE

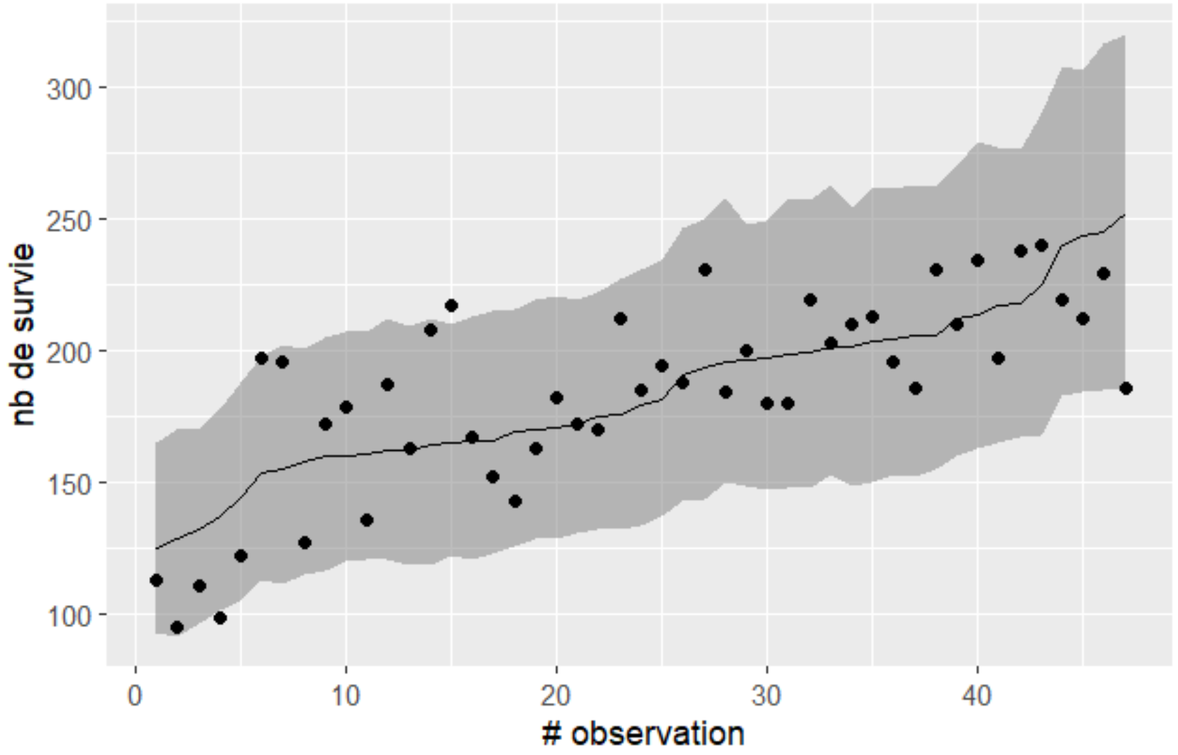
```
> head(Sophie,10)
# A tibble: 10 × 6
   Bloc  Trt          Saule Survie    Nb Vivant
   <fct> <fct>        <fct>   <dbl> <dbl>  <dbl>
 1 1     AB           Pr      0.577   248    143
 2 1     AB-MRF1-BRF  Pr      0.847   248    210
 3 1     AB-MRF1-BRF  Sm      0.734   248    182
 4 1     AB-MRF1-CI   Pr      0.75    248    186
 5 1     AB-MRF1-CI   Sm      0.673   248    167
 6 1     Témoin       Pr      0.722   248    179
 7 1     Témoin       Sm      0.383   248     95
 8 1     AB-MRF2-BRF  Pr      0.847   248    210
 9 1     AB-MRF2-BRF  Sm      0.657   248    163
10 1     AB-BRF       Pr      0.790   248    196
```

```
library(MASS)
modBN <- glm.nb(Vivant ~ Trt+Bloc+Saule, data = Sophie,
                control = glm.control(maxit = 5000))
```

```
> chi2 <- sum(residuals(modBN, type = "pearson")^2)
> chi2 / df.residual(modBN)
[1] 1.298335
```

```
> summary(modBN, dispersion = 1.29)

Call:
glm.nb(formula = Vivant ~ Trt + Bloc + Saule, data = Sophie,
    control = glm.control(maxit = 5000), init.theta = 73.62515766,
    link = log)

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)    4.86060    0.07483  64.958  < 2e-16 ***
TrtAB          0.05509    0.08328   0.662 0.508282
TrtAB-BRF      0.24415    0.07929   3.079 0.002077 **
TrtAB-MRF1-BRF 0.28087    0.07910   3.551 0.000384 ***
TrtAB-MRF1-CI  0.24987    0.07926   3.152 0.001620 **
TrtAB-MRF2-BRF 0.22994    0.07937   2.897 0.003766 **
Bloc2         -0.02967    0.06633  -0.447 0.654616
Bloc3          0.17406    0.06543   2.660 0.007812 **
Bloc4          0.05956    0.06592   0.904 0.366252
SaulePr        0.21523    0.04598   4.681 2.85e-06 ***
```

# EXEMPLE AVEC BINOMIALE NÉGATIVE




```r
#valeurs prédite du modèle
sim_nb <- simulate(modBN, nsim = 1000, re.form = NULL, newdata = Sophie)
sim_pred <- mutate(Sophie, pred = predict(modBN, type = "response"),
                           q025 = apply(sim_nb, 1, quantile, probs = 0.025),
                           q975 = apply(sim_nb, 1, quantile, probs = 0.975)) %>%
  arrange(pred)

#graphique
ggplot(sim_pred, aes(x = 1:nrow(sim_pred), y = pred, ymin = q025, ymax = q975)) +
  geom_ribbon(alpha = 0.3) +
  geom_line() +
  geom_point(aes(y = Vivant))+
  xlab(" # observation") + ylab("nb de survie")
```

# Statistiques Bayésiennes

# CRITIQUE DE L'UTILISATION DE LA VALEUR P

Yaddanapudi, L. N. (2016). The American Statistical Association statement on P-values explained. Journal of Anaesthesiology Clinical Pharmacology, 32(4), 421-423

THE AMERICAN STATISTICIAN
2016, VOL. 70, NO. 2, 129–133
http://dx.doi.org/10.1080/00031305.2016.1154108

Taylor & Francis
Taylor & Francis Group

EDITORIAL

## The ASA's Statement on *p*-Values: Context, Process, and Purpose

In February 2014, George Cobb, Professor Emeritus of Mathematics and Statistics at Mount Holyoke College, posed these questions to an ASA discussion forum:

> Q: Why do so many colleges and grad schools teach $p = 0.05$?
> A: Because that's still what the scientific community and journal editors use.
> Q: Why do so many people still use $p = 0.05$?
> A: Because that's what they were taught in college or grad school.

Cobb's concern was a long-worrisome circularity in the sociology of science based on the use of bright lines such as $p < 0.05$: "We teach it because it's what we do; we do it because it's what we teach." This concern was brought to the attention of the ASA Board.

The ASA Board was also stimulated by highly visible discussions over the last few years. For example, ScienceNews (Siegfried 2010) wrote: "It's science's dirtiest secret: The 'scientific method' of testing hypotheses by statistical analysis stands on a flimsy foundation." A November 2013, article in Phys.org Science News Wire (2013) cited "numerous deep flaws" in null hypothesis significance testing. A ScienceNews article (Siegfried 2014) on February 7, 2014, said "statistical techniques for testing hypotheses …have more flaws than Facebook's privacy policies." A week later, statistician and "Simply Statistics" blogger Jeff Leek responded. "The problem is not that people use P-values poorly," Leek wrote, "it is that the vast majority of data analysis is not performed by people properly trained to perform data analy-

2014) and a statement on risk-limiting post-election audits (American Statistical Association 2010). However, these were truly policy-related statements. The VAM statement addressed a key educational policy issue, acknowledging the complexity of the issues involved, citing limitations of VAMs as effective performance models, and urging that they be developed and interpreted with the involvement of statisticians. The statement on election auditing was also in response to a major but specific policy issue (close elections in 2008), and said that statistically based election audits should become a routine part of election processes.

By contrast, the Board envisioned that the ASA statement on *p*-values and statistical significance would shed light on an aspect of our field that is too often misunderstood and misused in the broader research community, and, in the process, provides the community a service. The intended audience would be researchers, practitioners, and science writers who are not primarily statisticians. Thus, this statement would be quite different from anything previously attempted.

The Board tasked Wasserstein with assembling a group of experts representing a wide variety of points of view. On behalf of the Board, he reached out to more than two dozen such people, all of whom said they would be happy to be involved. Several expressed doubt about whether agreement could be reached, but those who did said, in effect, that if there was going to be a discussion, they wanted to be involved.

# UTILITÉ DES MODÈLE STATISTIQUE

On veut inférer sur la population à partir d’un échantillon.

- Estimer des grandeurs
- Tester des hypothèses
- Tirer des conclusion générales

# STATISTIQUES FRÉQUENTISTES

**Exemple**

On réalise une étude comprenant deux conditions :
contrôle & traitement

On veux connaitre s'il y a une différence entre les 2

Une méta-analyse nous indique que la différence moyenne entre ces deux conditions est de 3 unités → Quelle implication pour notre modèle ?

# STATISTIQUES FRÉQUENTISTES

**Exemple**

On réalise une étude comprenant deux conditions : contrôle & traitement

On veux connaitre s'il y a une différence entre les 2

Une méta-analyse nous indique que la différence moyenne entre ces deux conditions est de 3 unités

$H0 : \mu1 = \mu2$

$H1 : \mu1 \neq \mu2$

# STATISTIQUES FRÉQUENTISTES

**Exemple**

On réalise une étude comprenant deux conditions :
contrôle & traitement

On veux connaitre s'il y a une différence entre les 2

Une méta-analyse nous indique que la différence moyenne entre ces deux conditions est de 3 unités

Le résultat d'un test t :
$t(14) = 1.23, p = 0.24$

Comment interpréter la valeur p ?

# STATISTIQUES FRÉQUENTISTES

**Exemple**

Le résultat d'un test t :
t(14) = 1.23, p = 0.24

Si H0 est vraie, nous avons 24% de chances d'avoir observer nos données [P(D|T)]

Il y a 24% de probabilité que les données soient compatibles avec H0.

# STATISTIQUES FRÉQUENTISTES

**Exemple**

On peut construire une intervalle de confiance à 95 %

X ± 1,96 SE/ racine(n)

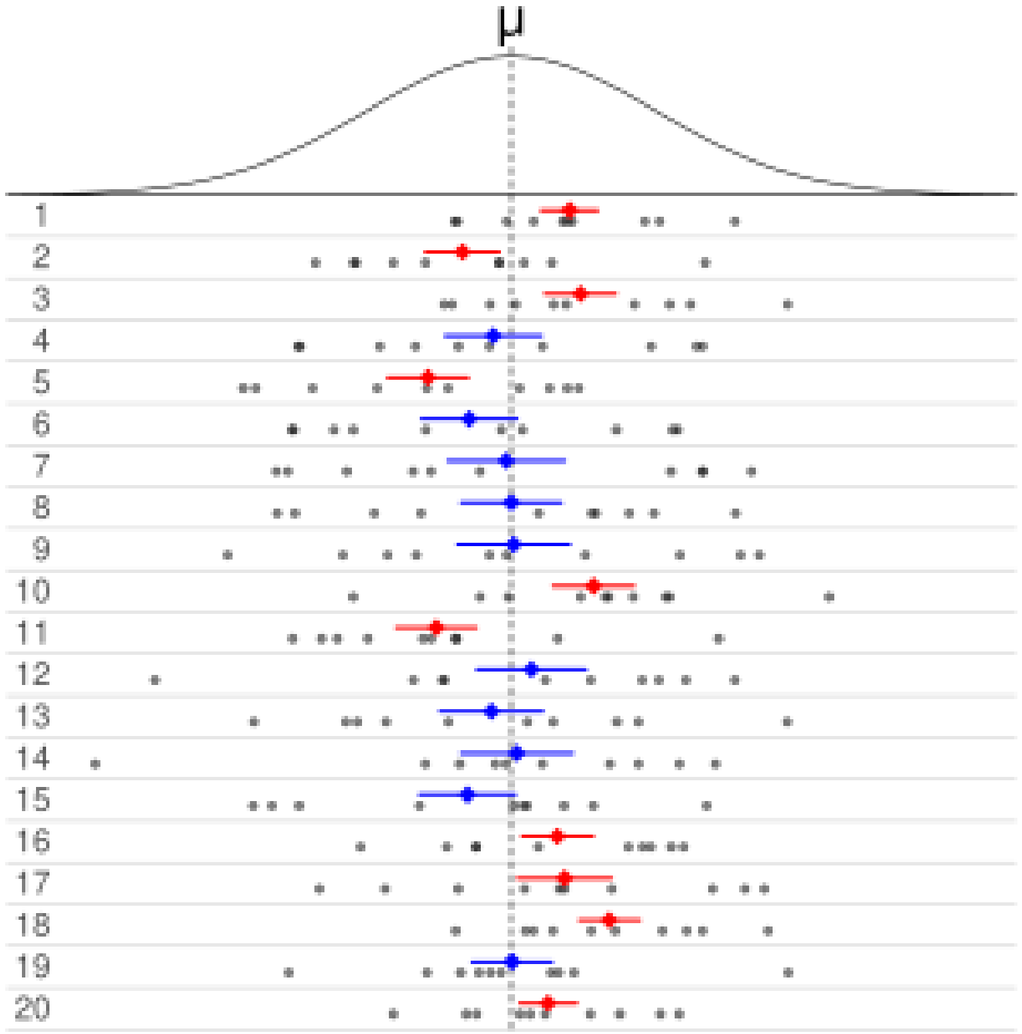
IC μ1 - μ2 : [- 0,02 ; 0,04]

# STATISTIQUES FRÉQUENTISTES

## Exemple

Interprétation :
Si on échantillonne 100 fois et on fait 100 test t, la vrai valeur sera comprise dans l'IC 95 fois.

(contre-intuitif)



# STATISTIQUES BAYÉSIENNES

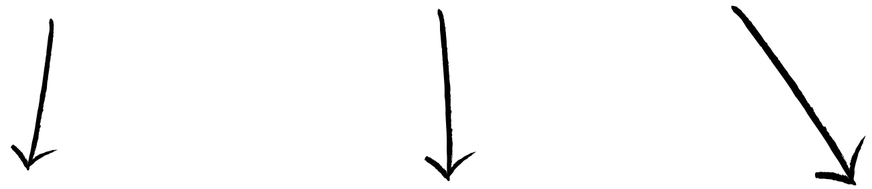
Dans une approche idéale d’inférence, on devrait :

- obtenir un maximum d’information sur le paramètre et non sur les données
- avoir le choix d’intégrer de l’information antérieure, pour éviter de repartir à zéro pour chaque test

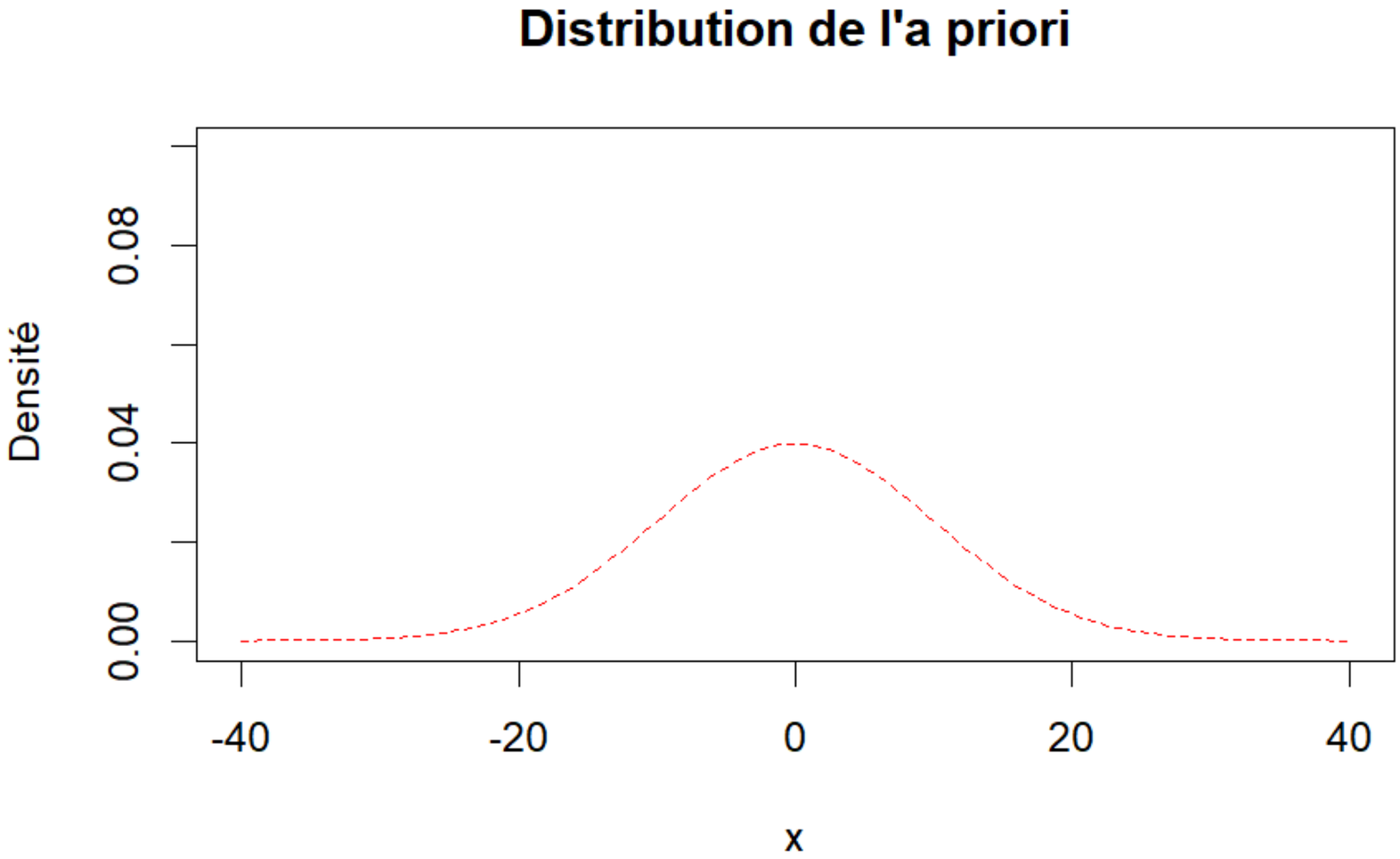
# STATISTIQUES BAYÉSIENNES

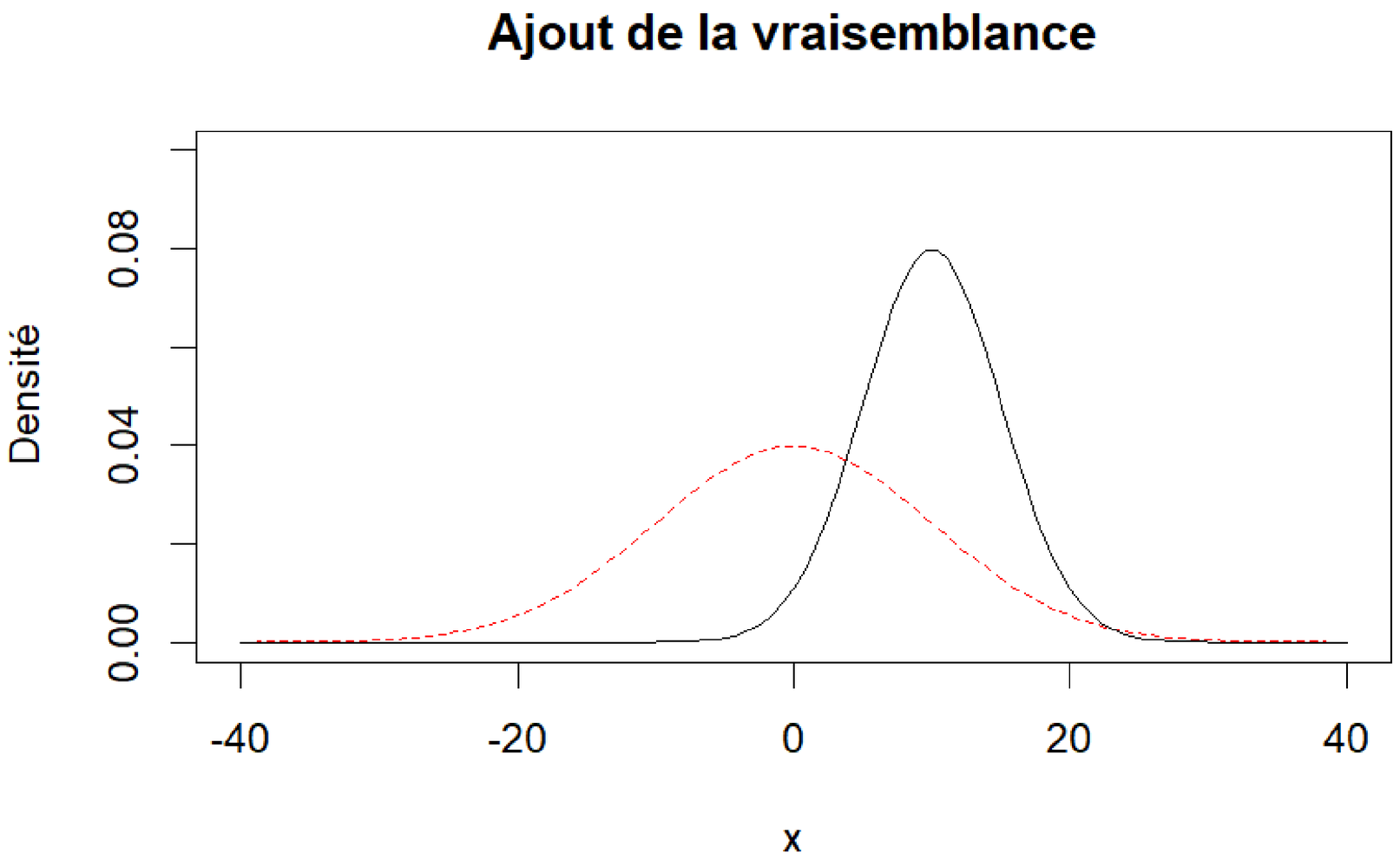
Équation de Bayes (1760), étayé par Laplace (1774) :

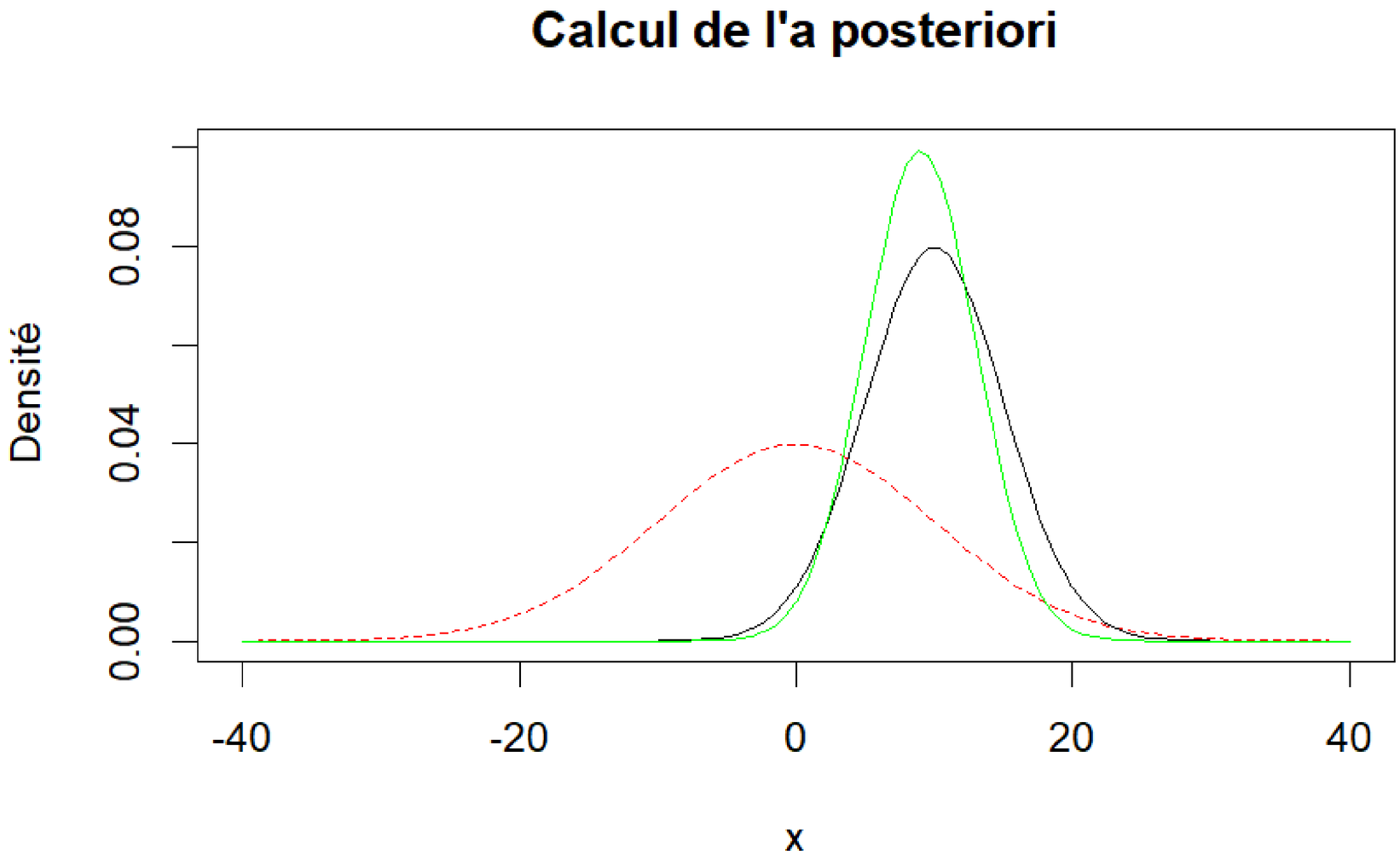
$$P(T|D) = P(D|T) \times P(T) / P(D)$$



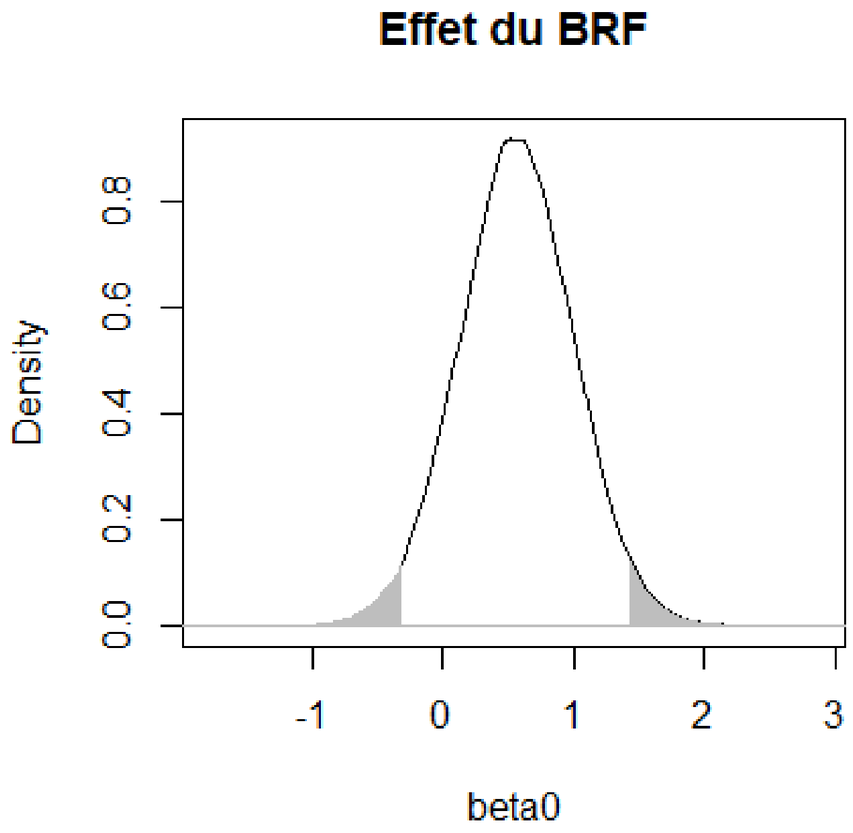
a postériori  vraisemblance  a priori


Distribution de l'a priori
Densité
0.00
0.04
0.08
-40
-20
0
20
40
x


Ajout de la vraisemblance
Densité
0.00
0.04
0.08
-40
-20
0
20
40
x

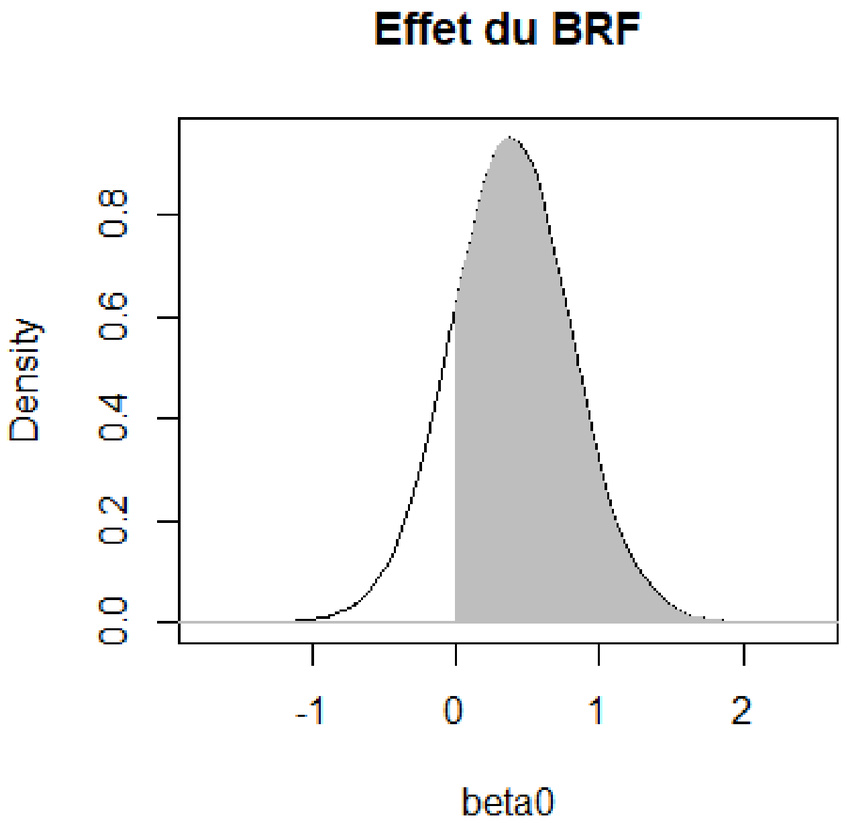

Calcul de l'a posteriori
Densité
0.00
0.04
0.08
-40
-20
0
20
40
x

# EXEMPLE APPLIQUÉ




Le 0 est compris dans l'intervalle de crédibilité (95%) : pas d'effet du traitement.

# EXEMPLE APPLIQUÉ




Le 0 est compris dans l'intervalle de crédibilité (95%) : pas d'effet de mon traitement.

On peut calculer la P(Trt > Témoin):
P = 82 %

# EXEMPLE APPLIQUÉ

| Valeur FB | Interprétation de l'évidence de l'effet |
|---|---|
| < 1 | Nul |
| 1 à 4 | Négligeable |
| 4 à 10 | Substantiel |
| 10 à 30 | Fort |
| 30 à 100 | Très fort |
| > 100 | Décisif |

On peut calculer le facteur de bayes (FB) :

$$FB = P(H1|D)\ P(H0)/P(H0|D)\ P(H1)$$

$$FB = 0.82/(1-0.82) = 4.62$$

# MERCI

*“Tout les modèles sont faux mais certains sont utiles”*

felix.lheureux-bilodeau.1@ulaval.ca

# Ressources utilisées:

*Hobbs and Hooten, Bayesian model, A statistical primer for ecologists. Princeton university press*

*Mazerolle, M. Note de cours : FOR-7046 : Modèles hiérarchiques et inférence bayésienne pour les sciences naturelles. Université Laval*

*Centre de la Science de la Biodiversité du Québec. Série d'ateliers R du CSBQ.https://r.qcbs.ca/fr/workshops/*

*Ivers, H. Conférence à l'AGAA. Au-delà d'un test significatif (ou non) : comment en apprendre davantage sur sa question de recherche à l'aide de l'approche bayésienne. 2022*

felix.lheureux-bilodeau.1@ulaval.ca