ANOVA à un facteur

S. Jaubert

13 janvier 2020

Contexte : on considère $k = 5$ opérateurs effectuant chacun $n = 20$ mesures.

Dans ce qui va suivre, il faut avoir à l'esprit que nos distributions sont normales et homoscédastiques (même variance)

On va chercher à savoir si la part de dispersion imputable au facteur "Opérateur" est significativement supérieure à la variabilité résiduelle (ou variance de répétabilité notée σ_r^2)

1ère Etape Mesurer les dispersions

On note :

-X_{ij} la $j^{ème}$ mesure ($j = 1...n$) de l'opérateur i ($i = 1...k$)

-$\bar{X} = \frac{1}{N}\Sigma_i\Sigma_j X_{ij}$ $(avec\ N = nk)$ la moyenne totale

-$\bar{X_i} = \frac{1}{n}\Sigma_j X_{ij}$ la moyenne du facteur i

Avant de rentrer dans les détails, donnons le principe général :

La dispersion totale notée **SCT** (somme des carrés totaux) se décompose en deux parties :

- Celle imputable aux facteurs (ici les opérateurs) que l'on notera **SCF** dite aussi **somme des carrés inter-classe**.
- Celle imputable aux résidus **SCR**, **somme des carrés intra-classes**.

$$\begin{array}{ccccc} \sum_{i,j}(X_{ij} - \bar{X})^2 & = & \sum_{i,j}(X_{ij} - \bar{X_i})^2 & + & \sum_{i,j}(\bar{X_i} - \bar{X})^2 \\ SCT & = & SCR & + & SCF \\ dispersion\ totale & = & somme\ des\ carrés\ intra-classes & + & somme\ des\ carrés\ inter-classes \end{array}$$

2ème Etape Calcul des variances factorielles et résiduelles

-Variance résiduelle : $\sigma_r^2 = \frac{SCR}{N-k}$

-Variance Factorielle : $CMF = \frac{SCF}{k-1}$

3ème Etape Test statistique

On va déterminer si la Variance Factorielle est significativement supérieure à la Variance résiduelle

On pose : $F_{k-1,N-k} = \frac{\frac{SCF}{k-1}}{\frac{SCR}{N-k}}$ (nous reconnaissons la statistique de test de Fisher). Si F est supérieure à la valeur seuil théorique selon la distribution de Fisher, avec un risque de 5%, alors le test sera significatif, donc la variabilité factorielle est significativement supérieure à la variabilité résiduelle et on conclut que les facteurs sont différents.

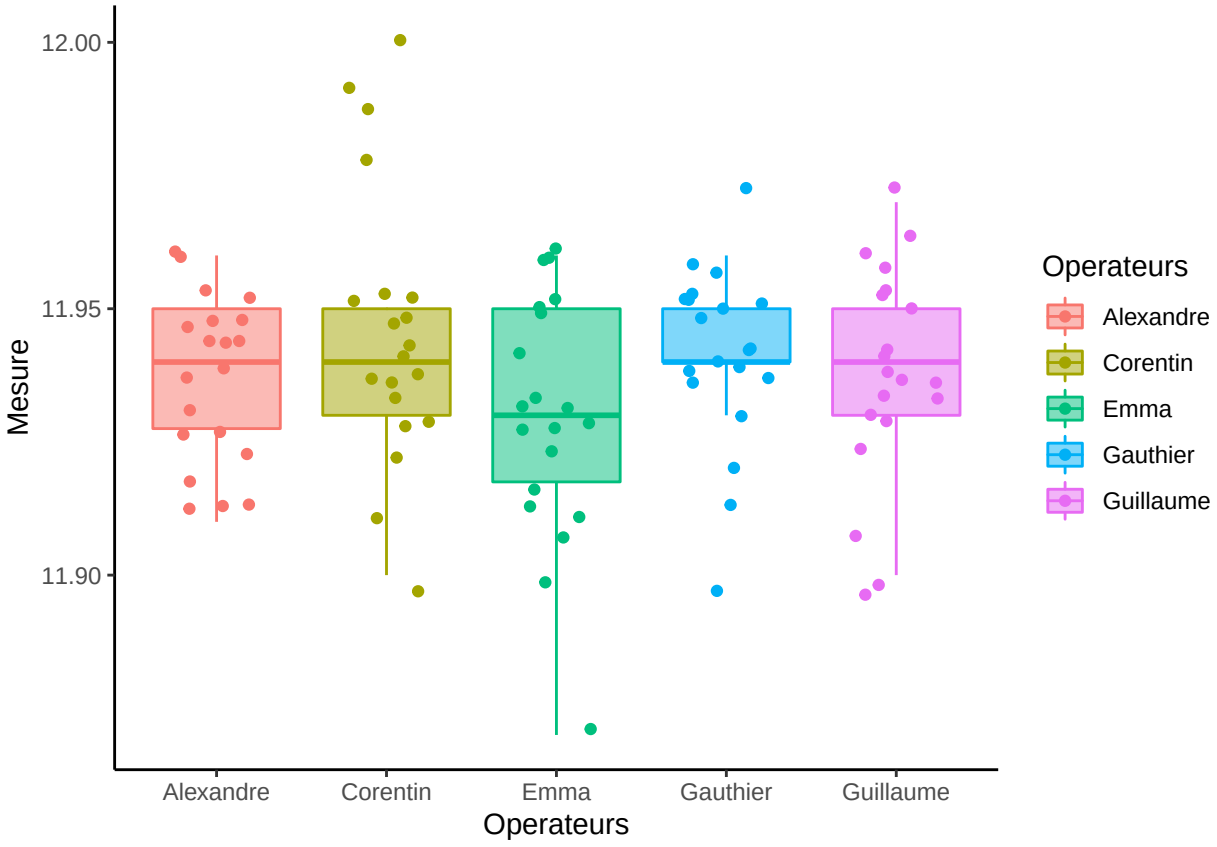
Prenons un exemple et Chargeons les données :

```r
donnees<-read.csv("https://sjaubert.github.io/SPCR/ANOVA_TP_R.csv",sep = ";",dec = ",",header = T)
donnees<-transform(donnees,Operateurs=as.factor(Operateurs))

Mesure=donnees$Mesure
Operateurs=donnees$Operateurs
```

En premier lieu, il est toujours utile de représenter les données pour se faire une première idée.

```r
library(ggplot2)
ggplot(donnees, aes(y=Mesure, x=Operateurs,colour=Operateurs ,fill=Operateurs))+
geom_boxplot(alpha=0.5, outlier.alpha=0)+
geom_jitter(width=0.25)+
theme_classic()
```



A commenter...

Recherche d'une éventuelle variabilité excessive :

```r
library("outliers")
cochran.test(Mesure~Operateurs)
```

```
##
##  Cochran test for outlying variance
##
## data:  Mesure ~ Operateurs
## C = 0.32447, df = 20, k = 5, p-value = 0.1212
```

```
## alternative hypothesis: Group Corentin has outlying variance
## sample estimates:
##    Alexandre     Corentin         Emma     Gauthier    Guillaume
## 0.0002555263 0.0006871053 0.0005207895 0.0002765789 0.0003776316
```

Au regard de la p-value nous pouvons considérer qu'il n'y a pas de variance aberrante

Recherche d'une éventuelle moyenne aberrante :

```
(Moyennes<-tapply(Mesure,Operateurs,mean))
```

```
## Alexandre  Corentin      Emma  Gauthier Guillaume
##   11.9365   11.9465   11.9295   11.9415   11.9375
```

```
grubbs.test(Moyennes)
```

```
##
##  Grubbs test for one outlier
##
## data:  Moyennes
## G.Emma = 1.39665, U = 0.39043, p-value = 0.2978
## alternative hypothesis: lowest value 11.9295 is an outlier
```

Au regard de la p-value nous pouvons considérer qu'il n'y a pas de moyenne aberrante

On peut donc conserver toutes les données

Mettons ces résultats en parallèle avec une ANOVA :

```
res<-aov(Mesure ~ Operateurs)
res
```

```
## Call:
##    aov(formula = Mesure ~ Operateurs)
##
## Terms:
##                 Operateurs Residuals
## Sum of Squares    0.003176  0.040235
## Deg. of Freedom          4        95
##
## Residual standard error: 0.02057975
## Estimated effects may be unbalanced
```

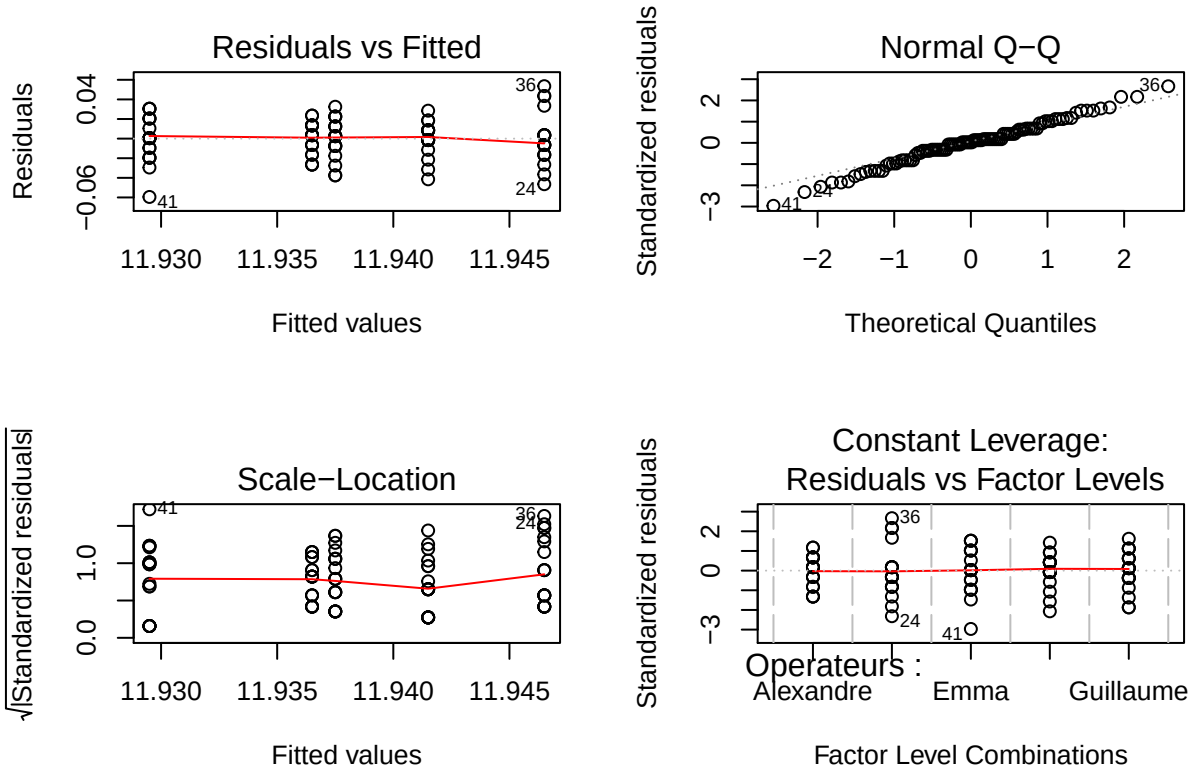
```
summary(res)
```

```
##             Df  Sum Sq   Mean Sq F value Pr(>F)
## Operateurs   4 0.00318 0.0007940   1.875  0.121
## Residuals   95 0.04024 0.0004235
```

La p-value nous indique qu'il n'y a pas de différence significative entre les opérateurs

(retrouvez les calculs précédents "à la main")

```r
par(mfrow=c(2,2)) # permet de séparer la fenêtre graphique en 4 parties (2 lignes, et 2 colonnes)
plot(res)
```



- Le premier graphe nous montre que la valeur des résidus ne semble pas dépendre de l'opérateur
- Le deuxième graphe (quantile-quantile) nous montre que les résidus suivent bien une loi normale
- Le troisième graphe nous montre que les variances des différents groupes sont globalement identiques
- Dans le quatrième graphe nous ne voyons aucune preuve de valeurs aberrantes.

Pour approfondir, quelques compléments théoriques

Nous gardons les même notations que précédemment.

Cherchons un estimateur de la variance de répétabilité

$$X_{ij} = \mu + \underset{\substack{\uparrow \\ \text{Effet de l'opérateur } i}}{\alpha_i} + \underset{\substack{\uparrow \\ \text{Résidu de l'opérateur } i \text{ sur la } j\text{ème mesure}}}{\epsilon_{ij}}$$

où $\epsilon_{ij} \hookrightarrow \mathcal{N}(0;\sigma_r)$ avec σ_r *l'écart type de répétabilité* et $\alpha_i \hookrightarrow \mathcal{N}(0;\sigma_o)$ avec σ_o *l'écart type du facteur opérateur*

On obtient :

$V(X_{ij}) = \sigma_o^2 + \sigma_r^2 = \sigma_R^2$

$SCR = \sum_{i=1}^{k}\sum_{j=1}^{n}\epsilon_{ij}^2$ soit :

$\frac{SCR}{\sigma_r^2} = \sum_{i=1}^{k}\sum_{j=1}^{n}(\frac{\epsilon_{ij}}{\sigma_r})^2$ mais on sait que

$\sum_{j=1}^{n}(\frac{\epsilon_{ij}}{\sigma_r})^2$ suit la loi du χ^2 à $n-1$ ddl alors $\frac{SCR}{\sigma_r^2} \hookrightarrow \chi^2(k(n-1))$ d'où

$\mathbb{E}(\frac{SCR}{\sigma_r^2}) = nk - k = N - k \Leftrightarrow \mathbb{E}(\frac{SCR}{N-k}) = \sigma_r^2$

$\frac{SCR}{N-k}$ est un estimateur sans biais de σ_r^2

Cherchons à présent un estimateur de la variance de reproductibilité

On a vu que :

$SCF = \sum_{i,j}(\bar{X}_i - \bar{X})^2 = n\sum_{i=1}^{k}(\bar{X}_i - \bar{X})^2$ et $\bar{X}_i = \frac{1}{n}\Sigma_j X_{ij} = \frac{1}{n}\Sigma_j(\mu + \alpha_i + \epsilon_{ij})$

$\bar{X}_i = \mu + \alpha_i + \frac{1}{n}\Sigma_j\epsilon_{ij}$

$V(\bar{X}_i) = V(\alpha_i) + \frac{1}{n^2}\Sigma_j V(\epsilon_{ij})$

$V(\bar{X}_i) = \sigma_o^2 + \frac{1}{n}\sigma_r^2$ d'où

$\frac{SCF}{V(\bar{X}_i)} = n\sum_{i=1}^{k}\left(\frac{\bar{X}_i - \bar{X}}{\sqrt{V(\bar{X}_i)}}\right)^2$

$\mathbb{E}(\frac{SCF}{V(\bar{X}_i)}) = n(k-1) \Leftrightarrow \mathbb{E}(SCF) = n(k-1)V(\bar{X}_i)$

Soit : $\mathbb{E}(\frac{SCF}{k-1}) = nV(\bar{X}_i) = n(\sigma_o^2 + \frac{1}{n}\sigma_r^2) = n\sigma_o^2 + \sigma_r^2$

Si on pose $CMF = \frac{SCF}{k-1}$ et $CMR = \frac{SCR}{N-k}$ on a :

$\mathbb{E}(\frac{CMF-CMR}{n}) = \sigma_o^2$ et comme $\sigma_R^2 = \sigma_o^2 + \sigma_r^2$

$\mathbb{E}(\frac{CMF-CMR}{n}) + CMR = \sigma_R^2$

Enfin on posera :

$$\sigma_{R\&R}^2 = \sigma_R^2 + \sigma_r^2$$