Université de Montpellier
Département de Mathématiques Appliquées

Outstanding 2 : Anova

Atelier Projet — HAX916X

Réalisé par :
DIALLO Ousmane
ATTOUMANI Ibrahim

STATISTIQUE
SCIENCE DES DONNÉES
UNIVERSITÉ DE MONTPELLIER

L'INSTITUT agro Montpellier

Année Universitaire 2025 – 2026

Table des matières

1. Modélisation

L'analyse de la variance (ou ANOVA) est une méthode permettant de modéliser la relation entre une variable quantitative et une ou plusieurs variables qualitatives.

Définition 1 : L'analyse de variance à un facteur

Quand il y a une seule variable explicative qualitative `A` à `I` modalités, on parle d'analyse de variance à 1 facteur.
Le modèle s'écrit:

$$y_{i,k} = \mu + \alpha_i + \epsilon_{i,k}$$

où:

- μ est la moyenne général,
- α_i est l'effet de la modalité i du facteur `A`,
- $\epsilon_{i,k}$ est un résidus avec k l'indice de répétition.

Cette méthode permet la comparaison de `K` moyennes.

Nous allons nous intéresser ici au cas de deux variables explicatives, donc à l'analyse de variance à deux facteurs

Définition 1 : L'analyse de variance à deux facteurs

Quand il y a deux variables explicatives qualitatives `A` et `B`, et `Y` une variable à expliquer quantitative, nous notons `I` le nombre de modalités de la variable `A` et `J` celui de la variable `B`.
Le modèle s'écrit:

$$y_{i,j,k} = \mu + \alpha_i + \beta_j + \gamma_{i,j} + \epsilon_{i,j,k}$$

où:

- μ est la moyenne général,
- α_i est l'effet de la modalité i du facteur `A`,
- β_j est l'effet de la modalité j du facteur `B`,
- $\gamma_{i,j}$ est un terme d'interaction entre les facteurs `A` et `B`,
- $\epsilon_{i,j,k}$ est un résidus avec k l'indice de répétition.

On teste les effets de A, de B et de l'interaction $A \times B$ avec des tests de Fisher qui comparent la variabilité expliquée à la variabilité résiduelle.

On commence toujours par tester l'interaction. Si elle est significative, cela signifie que l'effet de A dépend de B (et inversement), et il n'est plus utile de tester séparément A et B.

Les hypothèses du test d'interaction sont :

$$(H_0)_{AB} : \forall \quad (i,j), \gamma_{ij} = 0, \qquad (H_1)_{AB} : \exists (i,j), \quad \gamma_{ij} \neq 0. \qquad (1)$$

On teste ensuite séparément l'effet de chaque facteur. Par exemple, pour tester l'effet du facteur A, on compare :

$$\begin{cases} H_0 : y_{ijk} = \mu + \beta_j + \varepsilon_{ijk} & \text{(modèle sans le facteur } A\text{)} \\ H_1 : y_{ijk} = \mu + \alpha_i + \beta_j + \varepsilon_{ijk} & \text{(modèle avec } A\text{, sans interaction)} \end{cases}$$

Une fois le modèle final choisi (celui où les facteurs retenus sont significatifs), on estime les paramètres μ, α_i, β_j et γ_{ij}.

Cependant, le nombre total de paramètres $1 + I + J + IJ$ dépasse le nombre de paramètres effectivement estimables : seuls IJ sont indépendants. On doit donc imposer $1 + I + J$ contraintes pour rendre le modèle identifiable.

Les contraintes classiques sont :

- **Analyse par cellule** :

$$\mu = 0, \quad \alpha_i = 0 \ \forall i, \quad \beta_j = 0 \ \forall j.$$

- **Cellule de référence** :

$$\alpha_1 = 0, \qquad \beta_1 = 0, \quad \gamma_{i1} = 0 \ \forall i, \qquad \gamma_{1j} = 0 \ \forall j.$$

- **Contraintes de somme** :

$$\sum_i \alpha_i = 0, \qquad \sum_j \beta_j = 0, \quad \sum_j \gamma_{ij} = 0 \ \forall i, \quad \sum_i \gamma_{ij} = 0 \ \forall j.$$

1.1. Application directe: anova à effet fixe

Dans cette partie, nous allons utiliser le jeu de données **Ozone**.
Nous souhaitons analyser la relation entre le maximum journalier de la concentration en ozone (en $\mu g/m^3$) et deux facteurs :

- la direction du `vent`, classée en quatre secteurs : `Nord`, `Sud`, `Est`, `Ouest`,
- la précipitation, classée en deux modalités : `Sec` et `Pluie`.

Nous disposons de 112 observations relevées durant l'été 2001.

La variable A admet $I = 4$ modalités et la variable B a $J = 2$ modalités.

```
# importation
ozone <- read.table("data/ozone.txt",header=T)
# selectionner les variables d'etude
vars <- c("maxO3", "vent", "pluie" )
```

```r
# dataframe avec les variables d'etude
df <- data.frame(ozone[,vars])
```

Dans la suite, nous utiliserons la fonction AovSum du package **FactoMineR**.

```r
# modele d'interaction
mod_inter <- AovSum(maxO3 ~ vent + pluie + vent:pluie, data=df)

# le test de Fisher
mod_inter$Ftest
```

```
##                SS  df      MS F value    Pr(>F)
## vent         3227   3  1075.6  1.7633    0.1588
## pluie       10996   1 10996.5 18.0271 4.749e-05 ***
## vent:pluie   1006   3   335.5  0.5500    0.6493
## Residuals   63440 104   610.0
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

La fonction renvoie deux types de résultats : ceux des tests de Fisher et ceux des tests de Student.

Nous commençons par examiner les résultats des tests de Fisher, qui évaluent l'effet global de chaque facteur et de leur interaction. La première colonne indique les sommes des carrés, la deuxième les degrés de liberté et la troisième les carrés moyens. La quatrième colonne présente la statistique de test F et la cinquième donne la p-value, c'est-à-dire la probabilité d'obtenir une valeur aussi grande si l'hypothèse nulle est vraie.

Comme la p-value (0.65) est largement supérieure à 5%, nous ne pouvons pas rejeter l'hypothèse nulle. Nous concluons donc que l'interaction n'est pas significative.

Ainsi, le modèle peut être réécris de la manière suivant:

$$y_{i,j,k} = \mu + \alpha_i + \beta_j + \epsilon_{i,j,k} \quad (2)$$

Nous estimons alors le modèle sans interaction (2).

```r
# modele sans interaction
mod_sans_inter <- AovSum(maxO3 ~ vent + pluie, data=df)

# le test de Fischer
mod_sans_inter$Ftest
```

```
##                SS  df      MS F value    Pr(>F)
## vent         3791   3  1263.8  2.0982    0.1048
## pluie       16159   1 16159.4 26.8295 1.052e-06 ***
## Residuals   64446 107   602.3
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

La p-value associée à l'effet **vent** ($Pr(> F) = 0.1048$) est supérieur à 5% donc on conclut à la non significativité du facteur **vent**. C'est à dire il n'y a pas de d'effet de la directition du vent sur le maximux d'ozone journalier. Ainsi, on obtient un modèle d'analyse de variance un seul facteur (**pluie**):

$$y_{i,j,k} = \mu + \beta_j + \epsilon_{i,j,k} \quad (2)$$

Remarque : La significativité su facteur α_i

Si le facteur α_i étais significative, on aurait garder le modèle sous $(H_1)_A$:

$$y_{i,j,k} = \mu + \alpha_i + \beta_j + \epsilon_{i,j,k}$$

Puis, on aurait estimer les coefficients du modèle.

Nous allons à présent ajuster le modèle d'analyse de variance un seul facteur:

```
# modele a un facteur (pluie)
mod_simple <- AovSum(maxO3 ~ pluie, data=ozone)

# le test de Fisher
mod_simple$Ftest
```

```
##              SS  df      MS F value    Pr(>F)
## pluie     19954   1 19954.2  32.166 1.157e-07 ***
## Residuals 68238 110   620.3
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

La probabilité critique associée à l'effet **pluie** ($Pr(> F) = 1.157e - 07$) est très faible :

on en conclut que le facteur **pluie** affecte le maximum d'ozone journalier. Pour interpréter cet effet, il faut estimer les coefficients.

Les estimations des coeeficients dépende aux contraintes utilisées. Tous les résultats sur l'estimation des coefficients sont donnés dans l'objet **Ttest**.

```
# le test de Studebt
mod_simple$Ttest
```

```
##                Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)    87.11796   2.419555 36.005777 1.066479e-62
## pluie - Pluie -13.72262   2.419555 -5.671545 1.156980e-07
## pluie - Sec    13.72262   2.419555  5.671545 1.156980e-07
```

On obtient une matrice comportant pour chaque lignes, 4 colonnes: son estimation (**Estimate**) son écart-type estimé (**Std. Error**), et enfin la valeur de la statistique de test (**t value**) et

de la probabilité critique ($Pr(>|t|)$) associées au test de Student:

$$(H_0) : \forall i, \quad \alpha_i = 0 \quad vs \quad \exists i, \quad \alpha_i \neq 0$$

L'estimation de μ (87.12), notée Intercept, correspond à la moyenne du maximum d'ozone. L'effet du temps humide, α_1, est estimé à une diminution de $-13.72\,\mu g/m^3$ du maximum d'ozone. Comme les coefficients doivent vérifier $\alpha_1 + \alpha_2 = 0$ (contrainte sur la somme des coefficients), l'effet du temps sec est donc de $+13.72\,\mu g/m^3$.