

# Science des données : module 10



ANOVA

Philippe Grosjean & Guyliann Engels

Université de Mons, Belgique  
Laboratoire d'Écologie numérique des Milieux aquatiques

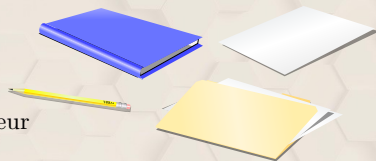


<http://biodatascience-course.sciviews.org>  
[sdd@sciviews.org](mailto:sdd@sciviews.org)

**U**MONS

# Objectifs du cours

- Se rappeler le principe de l'ANOVA à 1 facteur
- Étendre à un modèle à 2 facteurs
- Déterminer les différents types de modèles à 2 facteurs et choisir l'analyse correspondante



# ANOVA à un facteur - Rappel 1

- Les tests de Student sont limités à la comparaison de **deux** variables quantitatives (deux échantillons indépendants).
- **Comment comparer les moyennes de plus de deux échantillons simultanément ?**
- En faisant autant de tests de Student qu'il y a de couples de moyennes à comparer, **le risque de se tromper dans au moins une de ces comparaisons vaut :**

échantillons	2	3	4	6	8	10
risque $\alpha = 5\%$	5%	12%	20%	37%	51%	63%

=> Il faut travailler différemment : **ANOVA (ANalysis Of VAriance)**

# ANOVA à un facteur - Rappel 2

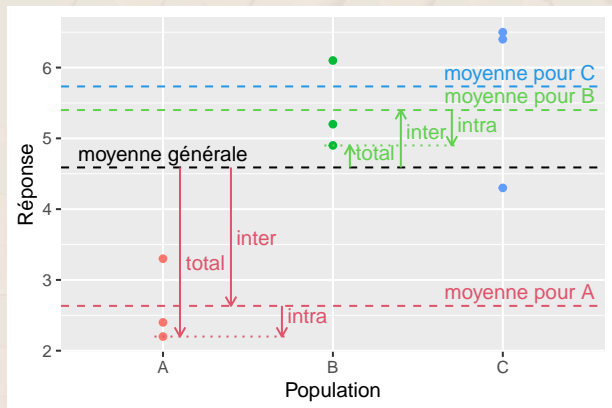
## ■ Hypothèses:

$$H_0 : \mu_1 = \mu_2 = \dots = \mu_p$$

$H_1$  : au moins une moyenne est différente

## ■ Décomposition de la variance

Comme pour la variance, les parts de variance *inter* et *intra* se calculent par la somme des carrés des distances divisées par les ddl.



## ANOVA à un facteur - Rappel 3

- Calcul des sommes des carrés (inter- et intragroupes)

$i$  = indice des observations au sein du jeu de données de 1 à  $n$ ,  $j$  = facteurs (sous-populations de 1 à  $p$ ),  $\bar{y}_j$  = moyenne de la  $j^{\text{ème}}$  population:

$$S_{inter} = \sum_{i=1}^n (\bar{y}_j - \bar{y})^2 \quad S_{intra} = \sum_{i=1}^n (y_{ij} - \bar{y}_j)^2$$

- Association du nombre de degrés de liberté :
  - $p - 1$  pour l'intergroupe
  - $n - p$  pour l'intragroupe
- Construction du **tableau de l'ANOVA** :

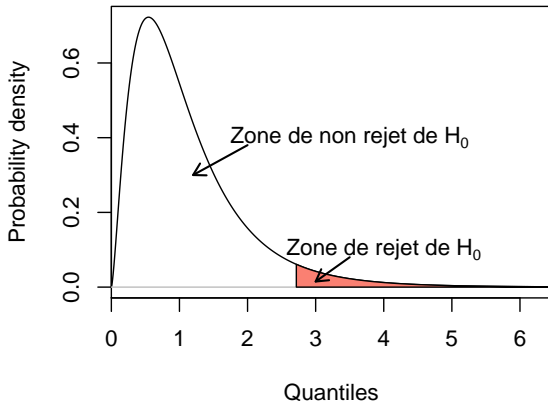
Type	Ddl	Somme carrés	Carré moyen ( $CM$ )	Statistique $F_{obs}$	P (>)
Inter (facteur)	$p - 1$	$S_{inter}$	$S_{inter}/ddl_{inter}$	$CM_{inter}/CM_{intra}$	...
Intra (résidus)	$n - p$	$S_{intra}$	$S_{intra}/ddl_{intra}$		

- Calcul de la **statistique  $F_{obs}$**  comme le **rapport des carrés moyens**.

# La distribution $F$ - Rappel

- Distribution asymétrique, n'admettant que des valeurs nulles ou positives.

La zone de rejet a une aire égale au seuil  $\alpha$  choisi  
Il suffit de positionner la valeur de  $F_{\text{obs}}$  sur le graphe pour savoir où l'on se situe



- **Deux paramètres** : les degrés de liberté au numérateur et au dénominateur (ici =  $ddl_{\text{inter}}$  et  $ddl_{\text{intra}}$ , respectivement).

# ANOVA à un facteur - Conditions d'application

- Modèle correspondant :

$$y_{ij} = \mu + \tau_j + \epsilon_{ij}$$

avec :  $\tau_j = \text{cste}$  et  $\epsilon_{ij} \sim N(0, \sigma) = \text{résidus (residuals)}$

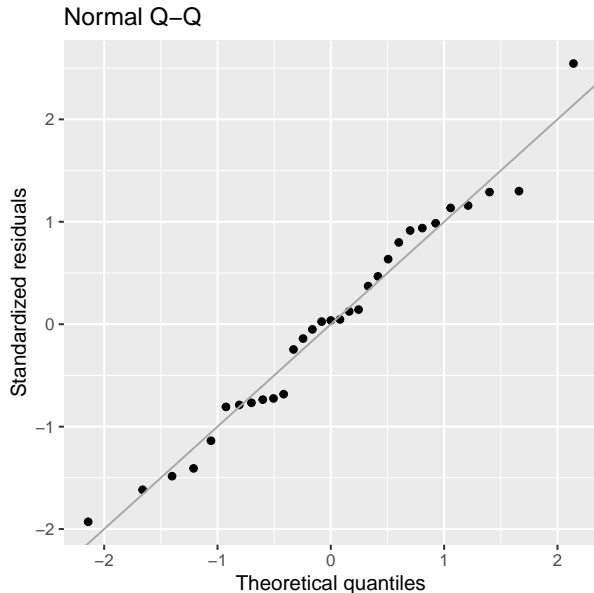
- Conditions d'application du test :

- échantillon aléatoire,
- observations indépendantes,
- variable **réponse** quantitative,
- une variable **explicative** qualitative à 3 niveaux ou plus,
- distribution **normale** des résidus,
- **homoscédasticité** (même variance intragroupes, **homoscedasticity** en anglais opposé à **hétéroscédasticité** = variance différente entre les groupes).

# Vérification de la normalité des résidus

- Résidus =  $y_{ij} - \bar{y}_j$
- Graphe  
quantile-quantile

Si les points s'alignent le long d'une droite, les quantiles respectifs correspondent  
 => la distribution observée se conforme à la distribution théorique





# Test de l'homoscédasticité

Différents tests d'hypothèse existent (Batlett, Levene, etc.).

■ Les hypothèses sont :

**$H_0$ : homoscédasticité  $\Rightarrow \text{var}_1 = \text{var}_2 = \dots = \text{var}_p$**

**$H_1$ : hétéroscédasticité  $\Rightarrow$  au moins une variance différente**

## Rejet de $H_0$ - Tests de comparaisons multiples

- Si  $H_0$  n'est pas rejetée, pas de problèmes, on considère que les moyennes ne sont pas significativement différentes les unes des autres au seuil  $\alpha$
- Si, par contre,  $H_0$  est rejetée, **nous ne savons pas encore quelle(s) moyenne(s) est(sont) différente(s) des autres**. Pour cela, il faut encore réaliser une **comparaison multiple des moyennes** (dit aussi **test *post hoc***, car il doit nécessairement être précédé de l'ANOVA).
  - Le test le plus simple est celui de **Bonferroni**. Il consiste à effectuer des tests de Student 2 à 2, mais en apportant une correction au seuil  $\alpha$  utilisé par chaque test individuel.
  - Le test de **Sheffé** est une alternative, mais très conservatif (et peu utilisé)
  - Le meilleur test est celui de **Tukey** (implémenté dans R).
  - **Pour mémoire** : pour des comparaisons à un témoin, on peut utiliser le test de **Dunnet** (pas dans R).

# ANOVA à un facteur - Démonstration

- **Iris**, étude de la longueur des pétales en fonction de l'espèce.
- Résolution de l'exemple dans R (démonstration...).
- **Pour mémoire : dans R, le modèle de l'ANOVA à un facteur s'écrit :**

Reponse  $\sim$  Facteur

- *Attention aux transformations éventuellement nécessaires pour l'homoscédasticité et la distribution normale des résidus !*

# ANOVA à deux facteurs

- **Extrêmement varié** : de nombreux modèles différents existent (facteurs fixes ou aléatoires, plan équilibré ou non, avec ou sans répétitions, facteurs imbriqués, etc.)
- **Expérience/analyse factorielle (factorial analysis)**. Permet plus que de tester deux facteurs simultanément : permet aussi de déterminer s'il y a des **interactions** entre les deux facteurs (variations différentes du facteur B pour les différents niveaux du facteur A).
- Exemple d'un jeu de données connu : **ToothGrowth** => facteur A = substitut alimentaire, facteur B = dose.
- Distinction entre **plan équilibré** et **non équilibré** : le plan équilibré est à la fois plus puissant et plus facile à calculer. Donc, il vaut mieux essayer d'obtenir le même nombre d'observations par niveau, autant que possible.

## Attention

Il faut des vraies répétitions pas des **pseudo-répétitions** !

# ANOVA à deux facteurs croisés sans réplicas

- Le manque de réplicas ne nous permet pas d'étudier les interactions entre les deux facteurs... nous devons considérer ici qu'elles n'existent pas (mais l'analyse sera incorrecte si cette hypothèse est fausse)!
- $y_{ijk} = \mu + \tau_{1j} + \tau_{2k} + \epsilon_{ijk}$
- $\epsilon_{ijk} \sim N(0, \sigma) = \text{résidus (residuals)}$
- Dans R, le modèle s'écrit :

Reponse  $\sim$  Facteur1 + Facteur2

## ANOVA à deux facteurs croisés sans réplicas - Exemple

- Rendement de blé (tonnes/ha) : 4 variétés testés dans 3 fermes :

	Variété A	Variété B	Variété C	Variété D
Ferme X	0.327	0.500	0.442	0.471
Ferme Y	0.532	0.599	0.516	0.638
Ferme Z	0.269	0.308	0.241	0.305

# ANOVA à deux facteurs croisés avec réplicas

- Cette fois-ci, on peut étudier les interactions
- $y_{ijk} = \mu + \tau_{1j} + \tau_{2k} + \tau_{1j \cdot \tau_{2k}} + \epsilon_{ijk}$
- $\epsilon_{ijk} \sim N(0, \sigma) = \text{résidus (residuals)}$
- Dans R, le modèle s'écrit :

Reponse ~ Facteur1 + Facteur2 + Facteur1 : Facteur2

ou en abrégé :

Reponse ~ Facteur1 \* Facteur2

## ANOVA à deux facteurs croisés avec réplicas - Exemple

- Rendement de blé (tonnes/ha) : 4 variétés testés dans 3 fermes :

	Variété A	Variété B	Variété C	Variété D
Ferme X	0.327	0.500	0.442	0.471
	0.280	0.510	0.463	0.460
Ferme Y	0.532	0.599	0.516	0.638
	0.526	0.637	0.499	0.655
Ferme Z	0.269	0.308	0.241	0.305
	0.277	0.286	0.228	0.314



# ANOVA à deux facteurs hiérarchisés

- Un facteur est **imbriqué dans l'autre** (hierarchically nested factor)
- $y_{ijk} = \mu + \tau_{1j} + \tau_{2k}(\tau_{1j}) + \epsilon_{ijk}$
- $\epsilon_{ijk} \sim N(0, \sigma) =$  **résidus** (residuals)
- Dans **R**, le modèle s'écrit :

Reponse ~ Facteur1 + Facteur2 %in% Facteur1

## ANOVA à deux facteurs hiérarchisés - Exemple

- Mesures de contamination bactérienne de différentes eaux par différents étudiants (un étudiant mesure une seule eau 3 fois) :

Eaux	Mesure 1	Mesure 2
Egout	(ét. A) 2700 / 2800 / 1700	(ét. B) 2600 / 3000 / 3200
Polluée	(ét. C) 52 / 49 / 61	(ét. D) 68 / 75 / 83
Propre	(ét. E) 5.9 / 7.6 / 16.0	(ét. F) 5.6 / 5.9 / 6.3