


Plan du cours

Partie 4. Les tests sur les différences de moyennes

- 1. Test de conformité : moyenne observée vs théorique
- 2. Test d'homogénéité : Comparer les moyennes de plusieurs échantillons
 - 2.1. Comparaison de deux échantillons indépendants
 - 2.1.1. Test paramétrique : t de Student (Welch)
(Comparaison de variances : test F de Fisher)
 - 2.1.2. Test non paramétrique : Wilcoxon-Mann-Whitney
 - 2.3. Comparaison de deux échantillons appariés
 - 2.3.1. Test paramétrique : t de Student apparié
 - 2.3.2. Test non paramétrique : Wilcoxon apparié
 - 2.4. Comparaison de trois échantillons ou plus**
 -  **2.4.1. Test non paramétrique : Kruskal-Wallis**
 - 2.4.2. Test paramétrique : ANOVA 1 facteur

Comparaison >2 échantillons indépendants – Non paramétrique

Les données :

On mesure 1 VA quantitative ou qualitative ordinale X (VD) sur k échantillons indépendants (VI à $k > 2$ modalités) de n_i individus.

Principe :

Test global de comparaison de k échantillons **basé sur les rangs** des observations.

→ Généralisation du test de Wilcoxon-Mann-Whitney à la comparaison de > 2 échantillons.

→ S'applique quand les conditions de ANOVA 1 facteur ne sont pas respectées.

Comparaison >2 échantillons indépendants – Non paramétrique

Les hypothèses :

H_0 : les sommes des rangs des k échantillons sont proches

→ Ils proviennent vraisemblablement d'une même population statistique.

H_1 : Il y a au moins 1 échantillon dont la somme des rangs \neq significativement des autres

→ Ils proviennent d'au moins 2 populations statistiques différentes.

Conditions d'application :

- Les échantillons sont indépendants
- Eviter effectifs d'échantillons trop petits $\ll 5$
- Pas trop d'ex aequo

Comparaison >2 échantillons indépendants – Non paramétrique

La statistique de test :

$$H = \frac{12}{N(N+1)} \times \sum_{k=1}^k \left[\frac{r_i^2}{n_i} \right] - 3(N+1)$$

k : nb total d'échantillons

N : nb total d'observations (sur k échantillons)

n_i : nb d'observations dans l'échantillon i

r_i : somme des rangs dans l'échantillon i

En cas d'ex-aequo, calculer le H_{obs*} corrigé : $H_{obs*} = H_{obs} / \text{correction}$

Avec :

$$\text{correction} = 1 - \frac{\sum_{k=1}^k (e_r^3 - e_r)}{N^3 - N}$$

e_r : nombre d'ex-aequo pour un rang "r" donné

Exemple : si 4 ex-aequo pour 1 rang donné, e_r=4

Comparaison >2 échantillons indépendants – Non paramétrique

La méthode:

1. Fusionner les valeurs de l'ensemble des échantillons en 1 seule série.
2. Classer par ordre croissant les N observations en conservant une trace de leur échantillon d'origine + en repérant les ex-aequos.
3. Assigner un rang à chaque observation (rang moyen en cas d'ex-aequo).
4. Calculer la somme des rangs affectés à chaque échantillon ($= r_i$).
5. Calculer la St_{obs} et la comparer avec sa distribution théorique sous H_0 .

NB : si les rangs des valeurs des éch. sont « bien mélangés », les r_i seront proches les 1 des autres

Comparaison >2 échantillons indépendants – Non paramétrique

2.4.1.1. Pour $k \leq 5$:

Le $H_{\text{théo}}$ se lit dans la table de Kruskal-Wallis pour un nombre k d'échantillon et « n_i » observations au seuil α fixé.

2.4.1.2. Pour $k > 5$:

Le $H_{\text{théo}}$ se lit dans la table de la loi du Chi^2 avec ddl = $k-1$ au seuil α fixé.

Règle de décision :

On compare H_{obs} (*: corrigé des ex-aequo si besoin) au $H_{\text{théo}}$.

Si $H_{\text{obs}(*)} \geq H_{\text{théo}}$ alors on rejette H_0 et accepte H_1 .

Comparaison >2 échantillons indépendants – Non paramétrique

Application :

Au sein de 3 sites du bassin d'Arcachon, le nombre de couples de canards nichant dans 5 quadrats aléatoires a été comptabilisé.



		Q1	Q2	Q3	Q4	Q5
C :	Cap ferret	18	22	26	18	na
A :	Audenge	51	37	25	43	39
D :	Delta de l'Eure	37	23	17	33	33

Ces 3 sites sont ils exploités de la même façon par les canards ou existe-t-il un ou plusieurs lieux privilégiés ?

→ Voir la réalisation du test « à la main » en vidéo sur Moodle ←

Importer un jeu de données sous R

Importer sous R le data frame «canard» (Moodle, **format csv –anglosaxon- : sep="," , dec="."**).

(rappeler **chemin d'accès** de votre répertoire de travail si besoin)

```
canard_df<-read.table("canard.csv", sep="," , dec = ".",header = T)
```

```
canard_df$site<-as.factor(canard_df$site)  
str(canard_df)
```

```
'data.frame':  14 obs. of  2 variables:  
 $ nb  : int  18 22 26 18 51 37 25 43 39 37 ...  
 $ site: Factor w/ 3 levels "A","C","D": 2 2 2 2 1 1 1 1 1 3 ...
```

	nb	site
1	18	C
2	22	C
3	26	C
4	18	C
5	51	A
6	37	A
7	25	A
8	43	A
9	39	A
10	37	D
11	23	D
12	17	D
13	33	D
14	33	D

Effectuer le test sous R

Code :

```
boxplot(nb~site, data=canard_df)
```

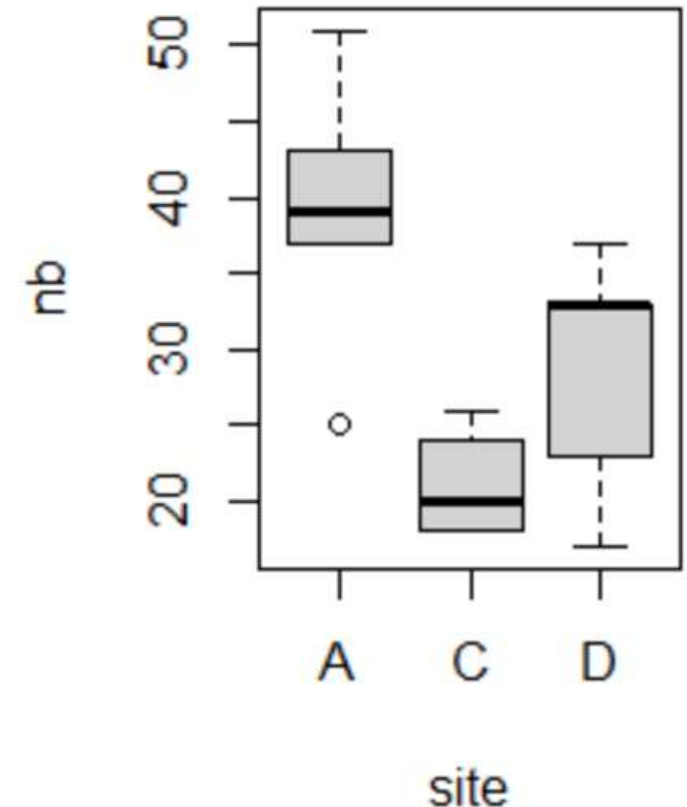
```
kruskal.test(nb~site, data=canard_df)
```

Sortie :

```
Kruskal-Wallis rank sum test
```

```
data: nb by site
```

```
Kruskal-Wallis chi-squared = 6.7301, df = 2, p-value = 0.03456
```



→ H1: Différence significative entre au moins 2 échantillons ...

Test de Kruskal-Wallis – tests « post-hoc »

Si on conclue à H_0 : « les distributions sont les même dans les k échantillons »

→ On **s'arrête là**...

→ On conclue qu'aucune différence significative n'a pu être révélée avec le plan utilisé.

Quand **la conclusion du test est H_1** : « au moins 2 des k échantillons différent »

→ on peut effectuer des **tests** de comparaisons multiples « **post-hoc** » (ex : 2 à 2) pour déterminer quels échantillons sont statistiquement différents des autres.

→ On cherche à savoir où se situent ces différences?

Il existe plusieurs tests (non détaillés ici) qui peuvent être réalisé facilement avec R...

Effectuer le test sous R – test « post-hoc »

Code : `install.packages("PMCMRplus")`
`library(PMCMRplus)`

Dunn test: valide pour plans équilibrés ou déséquilibrés

```
res2<-kwAllPairsDunnTest(nb~site, canard_df)  
summary(res2)
```

Effectuer le test sous R – test « post-hoc »

Code :

```
install.packages("PMCMRplus")  
library(PMCMRplus)
```

Dunn test: valide pour plans équilibrés ou déséquilibrés

```
res2<-kwAllPairsDunnTest(nb~site, canard_df)  
summary(res2)
```

Sortie : Pairwise comparisons using Dunn's all-pairs test

```
data: nb by site  
alternative hypothesis: two.sided  
P value adjustment method: holm  
H0
```

	z	value	Pr(> z)
C - A == 0	2.538	0.033403	*
D - A == 0	1.669	0.190412	
D - C == 0	0.965	0.334380	

Les échantillons A et C sont significativement différents

Etoiles = niveaux significativité p-value

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

Plan du cours

Partie 4. Les tests sur les différences de moyennes

1. Test de conformité : moyenne observée vs théorique
2. Test d'homogénéité : Comparer les moyennes de plusieurs échantillons
 - 2.1. Comparaison de deux échantillons indépendants
 - 2.1.1. Test paramétrique : t de Student (Welch)
(Comparaison de variances : test F de Fisher)
 - 2.1.2. Test non paramétrique : Wilcoxon-Mann-Whitney
 - 2.3. Comparaison de deux échantillons appariés
 - 2.3.1. Test paramétrique : t de Student apparié
 - 2.3.2. Test non paramétrique : Wilcoxon apparié
 - 2.4. Comparaison de trois échantillons ou plus**
 - 2.4.1. Test non paramétrique : Kruskal-Wallis
 - 2.4.2. Test paramétrique : ANOVA 1 facteur**

Comparaison >2 échantillons indépendants – Paramétrique

Les données :

- 1 variable VD quantitative Y.
- 1 VI qualitative X à k modalités définissant les k échantillons indépendants à comparer.



Principe :

On utilise l'**ANOVA à 1 facteur** quand on veut :

- **Comparer les moyennes des k (>2) échantillons/groupes :**

$$A \neq B \neq C ?$$

*“ Est ce que les **moyennes** des groupes sont **globalement différentes** ou pas ? ”*

- **Tester l'effet de la VI (modalités fixées par expérimentateur) sur la VD :**

$$A = B < C ?$$

*“ Quelle est l'**influence des différentes modalités** de la VI sur les variations de la VD? ”*

→ Généralisation du test de Student à la comparaison de k moyennes.

→ Equivalent paramétrique du test de Kruskal-Wallis (test global, $\alpha_{\text{global}}=5\%$).

Comparaison >2 échantillons indépendants – Paramétrique

Les hypothèses :

H_0 : $\bar{Y}_1 = \bar{Y}_2 = \dots = \bar{Y}_k = \bar{\bar{Y}}$ (les moyennes sont similaires (\rightarrow même population statistique))

H_1 : au moins une des moyennes est différente des autres (\rightarrow au moins 2 populations)

- **Si on ne rejette pas H_0** : le plan d'échantillonnage n'a pas pu révéler de variation significative globale entre les échantillons (hasard d'échantillonnage)

\rightarrow On s'arrête là!

- **Si on rejette H_0** : les conditions dans lesquelles ont été réalisés les k échantillons (modalités du facteur) sont de nature à modifier significativement les moyennes d'échantillons \bar{Y}_j

\rightarrow On utilisera un test « post-hoc » (ex: Tukey HSD) pour identifier les échantillons qui diffèrent significativement des autres ...

Comparaison >2 échantillons indépendants – Paramétrique

Les conditions d'application :

- **Indépendances des résidus** = les échantillons sont aléatoires et indépendants

Dépend du plan d'expérience : pas d'autocorrélation temporelle (mesures répétées) ou spatiale

→ *Autocorrélogramme (lag plot), test de Durbin-Watson*

- **Les résidus doivent être normalement distribués**

→ *Histogramme des résidus, Qqplot, test Shapiro*

- **Homoscédasticité des résidus (égalité des variances)**

→ *Test de Bartlett (pour plans équilibrés) ou de Levene*

➔ On doit **d'abord effectuer le test** pour obtenir les **résidus** et vérifier ces conditions.

Comparaison >2 échantillons indépendants – Paramétrique

Ces conditions sont toutefois assez flexible...

- L'ANOVA est un test robuste, plutôt peu sensible au non respect de la normalité (>30).
- On peut s'affranchir de l'homoscédasticité si les effectifs sont ~ égaux (surtout si grands).

... mais faut pas pousser !

En revanche, manque d'homoscédasticité cumulé à des effectifs inégaux = grave problème!

Que faire quand les conditions ne sont pas remplies ?

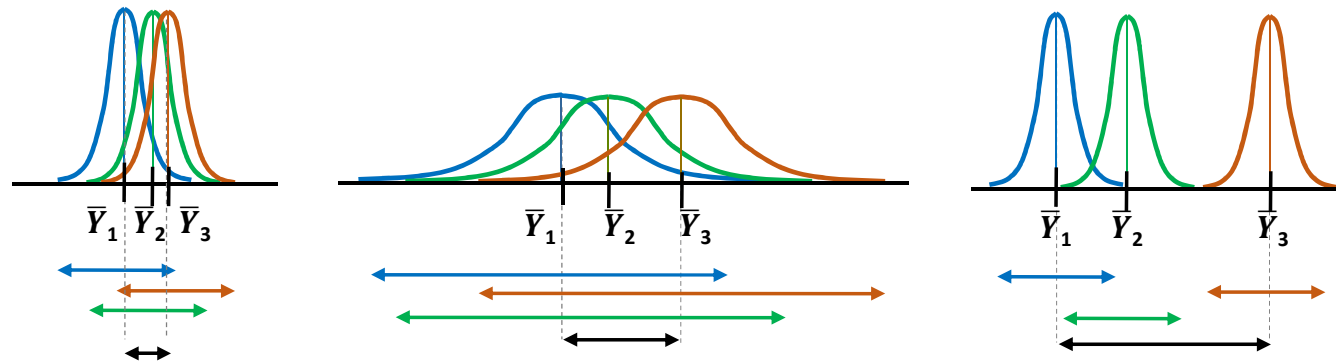
Transformer la VD (\sqrt{x} , $\log(x)$) pour améliorer normalité et/ou homogénéité des résidus.

Quand on ne peut pas pratiquer une ANOVA → test de **Kruskal-Wallis**.

Comparaison >2 échantillons indépendants – Paramétrique

Principe : Comparer la distribution de VD Y au sein et entre échantillons ($k > 2$ modalités de la VI X)
→ **Effet du facteur X sur les variations de Y ?**

Ex avec $k=3$ échantillons :



Moyennes $\bar{Y}_j \rightarrow$ • Proches

• Moins proches

• Eloignées

Recouvrement distributions \rightarrow • Fort

• Fort

• Faible

(variances faibles)

(variances grandes)

*Difficile de dire qu'il y a une différence significative
entre les distributions...*

(l'effet potentiel du facteur X étudié est non perceptible)

*Effet du facteur plus visible
(Distributions différentes)*

Comment le mesurer, le tester statistiquement? \rightarrow analyse de la variance

On va comparer la variabilité des données entre les groupes (inter) à variabilité des données dans les groupes (intra)

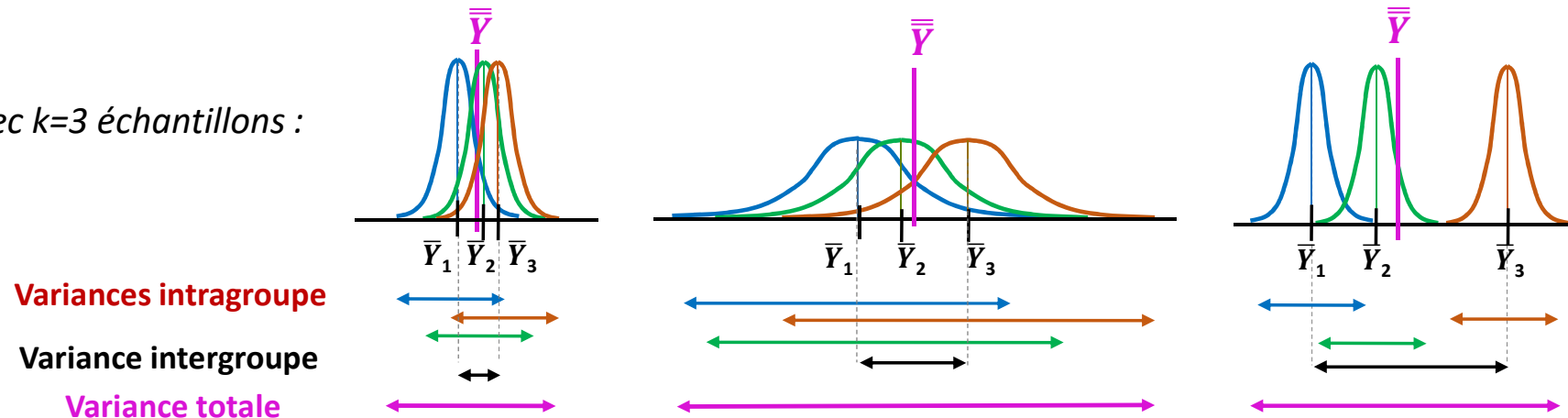
Comparaison >2 échantillons indépendants – Paramétrique

Décision selon :

Variance inter < Variance intra

Variance inter >> Variance intra

Ex avec $k=3$ échantillons :



Décomposition
de la variance :

Variance totale du jeu de données: S_{tot}^2

Dispersion de l'ensemble des observations autour de la moyenne générale $\bar{\bar{Y}}$

Variance intergroupe (factorielle) : S_{inter}^2

Variance intragroupe (résiduelle, erreur) : S_{intra}^2

- Variabilité de Y due aux modalités du facteur étudié
- Variance des \bar{Y}_j des échantillons autour $\bar{\bar{Y}}$
- Part de variation totale de Y expliquée par le modèle (VI)

- Variabilité de Y propre à chaque échantillon (modalité de VI)
- Variance des observations autour de la \bar{Y}_j de chaque éch.
- Part restant non expliquée par le modèle (VI)

Variance totale $S_{tot}^2 = \text{SCE}_{\text{Tot}} / (N-1)$ (avec $N-1=\text{ddl}$)

SCE_{Tot} : somme totale des carrés des écarts entre Y_{ij} et la moyenne générale $\bar{\bar{Y}}$

$$\text{SCE}_{\text{Tot}} = \sum_{j=1}^k \sum_{i=1}^n (Y_{ij} - \bar{\bar{Y}})^2$$

Avec: $i=\text{obs.}$, $n=\text{effectif éch}$, $j=\text{modalité du facteur}$, $k=\text{nb éch}$, $N=\text{nb total obs.}$

Se décompose en: $\text{SCE}_{\text{Tot}} = \text{SCE}_{\text{Inter}} + \text{SCE}_{\text{Intra}}$

Variance inter $S_{\text{inter}}^2 = \text{SCE}_{\text{Inter}} / (k-1)$ (avec $k-1=\text{ddl}$)

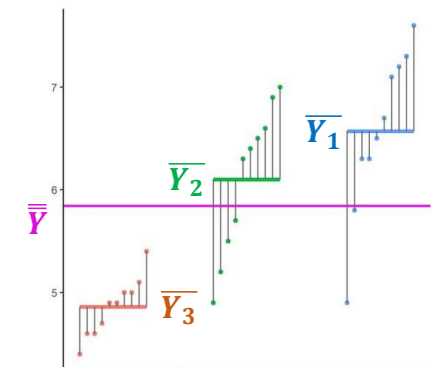
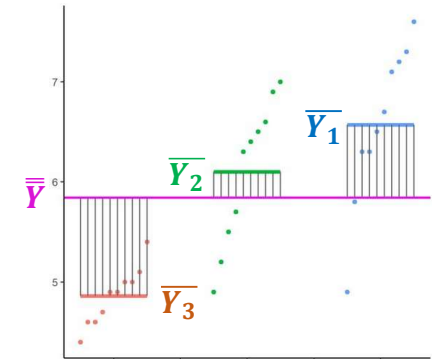
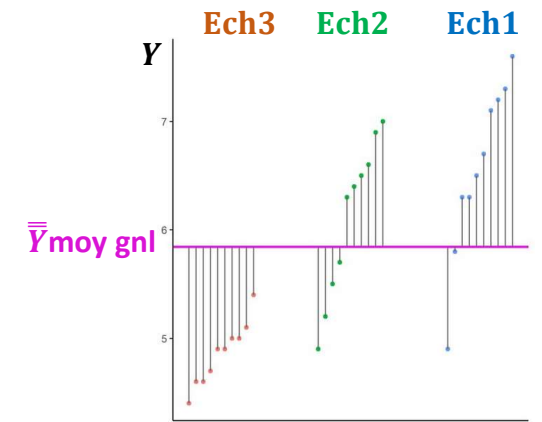
SCE_{Inter} : somme des carrés des écarts factoriels (moy. éch. $\bar{Y}_j \rightarrow$ moy. gnl. $\bar{\bar{Y}}$)

$$\text{SCE}_{\text{Inter}} = \sum_{j=1}^k n_j (\bar{Y}_j - \bar{\bar{Y}})^2$$

Variance intra $S_{\text{intra}}^2 = \text{SCE}_{\text{Intra}} / (N-k)$ (avec $N-k=\text{ddl}$)

SCE_{Intra} : somme des carrés des écarts résiduels ($Y_{ij} \rightarrow$ moy. éch. \bar{Y}_j)

$$\text{SCE}_{\text{Intra}} = \sum_{j=1}^k \sum_{i=1}^n (Y_{ij} - \bar{Y}_j)^2$$



Comparaison >2 échantillons indépendants – Paramétrique

La statistique de test :

Rapport de variances suivant loi de Fisher-Snedecor

Test unilatérale

$$F = S^2_{\text{inter}} / S^2_{\text{intra}}$$

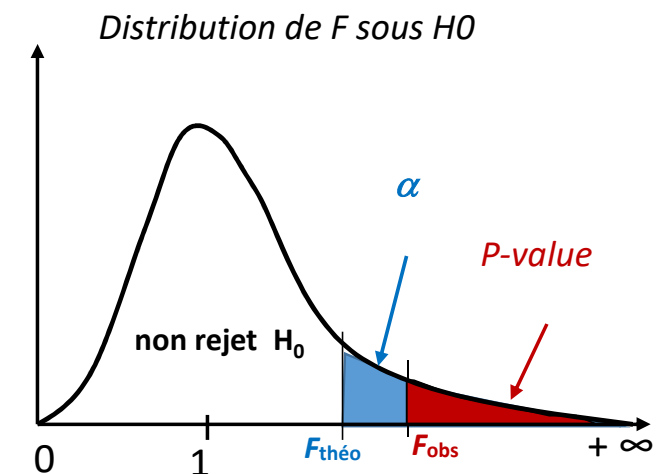
Retour sur les hypothèses :

H_0 : les moyennes des éch. sont similaires (proches)
donc on s'attend à obtenir un S^2_{inter} faible par rapport à S^2_{intra}

→ $F_{\text{obs}} = \text{faible } (\leq 1) \rightarrow \text{p-value forte} > \alpha$

H_1 : les moyennes des éch. sont significativement différentes
on s'attend à obtenir un S^2_{inter} fort et $\gg S^2_{\text{intra}}$

→ $F_{\text{obs}} = \text{fort } (> 1) \rightarrow \text{p-value faible} < \alpha$



Comparaison >2 échantillons indépendants – Paramétrique

$F_{\text{théo}}$ → Table Fisher-Snedecor unilatérale au seuil α , avec:

- Nombre de ddl du numérateur $S^2_{\text{inter}} = k-1$
- Nombre de ddl du dénominateur $S^2_{\text{intra}} = N-k$

Règle de décision :

Si $F_{\text{obs}} > F_{\text{théo}}$ (p-value < seuil α) → Rejet de H_0

- Le facteur étudié (VI) a un effet sur la VD
- Au moins 2 moyennes différent
→ test post-hoc pour trouver la ou lesquelles ≠

Table F: valeurs critiques de la distribution F

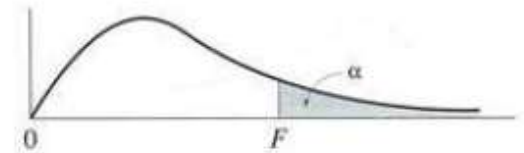


TABLE 1 $\alpha = .05$

		Degrés de liberté pour le numérateur														
		1	2	3	4	5	6	7	8	9	10	15	20	25	30	∞
Degrés de liberté pour le dénominateur	1	161.4	199.5	215.8	224.8	230.0	233.8	236.5	238.6	240.1	242.1	245.2	247.6	249.0	250.0	250.0
	2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40	19.43	19.45	19.46	19.47	19.48
	3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.70	8.64	8.60	8.57	8.55
	4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.86	5.78	5.74	5.71	5.69
	5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.62	4.53	4.50	4.47	4.45
	6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	3.94	3.84	3.81	3.78	3.76
	7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.51	3.41	3.38	3.35	3.33
	8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.22	3.11	3.08	3.05	3.03
	9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.01	2.90	2.87	2.84	2.82
	10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.85	2.74	2.71	2.68	2.66
	11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.72	2.61	2.58	2.55	2.53
	12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.62	2.51	2.48	2.45	2.43
	13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.53	2.42	2.39	2.36	2.34
	14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.46	2.35	2.32	2.29	2.27
	15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.40	2.29	2.26	2.23	2.21
	16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.35	2.24	2.21	2.18	2.16
	17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.31	2.20	2.17	2.14	2.12
	18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.27	2.16	2.13	2.10	2.08
	19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.23	2.12	2.09	2.06	2.04
	20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.20	2.09	2.06	2.03	2.01
	22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30	2.15	2.04	2.01	1.98	1.96
	24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25	2.10	1.99	1.96	1.93	1.91
	26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22	2.07	1.96	1.93	1.90	1.88

Comparaison >2 échantillons indépendants – Paramétrique

Et après ? Si on veut savoir quel(s) échantillon(s) sont différents de quels autres il faut faire un test de comparaisons multiples « post-hoc ».

Il en existe de nombreux :

- Méthode LSD (Least Significant Difference)
- Méthode SNK (Student – Newman – Keuls)
- Méthode de Scheffe
- **Méthode HSD de Tukey**
- **Méthode de Bonferroni**
- **Méthode de Dunnett**
- ...

Le + utilisé pour plans équilibrés

Pour plans déséquilibrés

Pour comparer des traitements à un témoin (ex : contrôle)

Méthode	test	distribution	remarques
Scheffé	S	F	Conservatif, peu puissant, cohérent avec l'ANOVA
Tukey	q	distribution d'étendue	Bon compromis
GT2	q	distribution d'étendue	Bon compromis
Student-Newman-Keuls	d	distribution d'étendue	puissant mais trop libéral
Duncan	q	distribution d'étendue	puissant mais trop libéral. Presque abandonné
Dunnett	q	Distribution d'étendue	compare les traitements à un contrôle
Bonferroni	t	t au seuil $\alpha' = \alpha/k$	très conservatif, peu puissant
Sidak	t	t au seuil $\alpha' = 1 - (1 - \alpha)^{\frac{1}{k}}$	très conservatif, peu puissant
PPDS (Plus Petite Différence Significative) ou LSD (Least Significant Difference) de Fisher	t	t au seuil α	limite le nombre de comparaisons. Trop libéral néanmoins. Presque abandonné.

TABLE 1 – Avantages et inconvénients de quelques procédures de comparaisons multiples (tests post-hoc)

Effectuer le test sous R

Application : (Exercice d'interprétation de sortie R, exam AETPF 2019)

Pour étudier l'influence de l'exposition sur la croissance en mètres de conifères, on a mesuré 10 arbres sur 4 versants opposés (notés Vers_1 à Vers_4) d'une montagne.



Y a-t-il un effet significatif de l'exposition sur la croissance des conifères?

$$H_0 : \bar{Y}_1 = \bar{Y}_2 = \bar{Y}_3 = \bar{Y}_4 = \bar{\bar{Y}}$$

H_1 : au moins une des moyennes est différente

Effectuer le test sous R

Importer le jeu de données « Versant_longFR.csv » sous R (Moodle).

```
Data_long<-read.table("Versant_longFR.csv", sep=";", dec=",", header = T)
```

```
Data_long$versant<-as.factor(Data_long$versant)
```

```
str(Data_long)
```

```
'data.frame':  40 obs. of  3 variables:
 $ station   : int  1 2 3 4 5 6 7 8 9 10 ...
 $ versant   : chr  "Vers_1" "Vers_1" "Vers_1" "Vers_1" ...
 $ croissance: num  23.4 24.4 24.6 24.9 25 26.2 26.3 26.8 26.9 27 ...
```

```
table(Data_long$versant)
```

De quel type de plan s'agit il? →

Vers_1	Vers_2	Vers_3	Vers_4
10	10	10	10

Obs. VI = Facteur (k = 4 modalités) VD quantitative

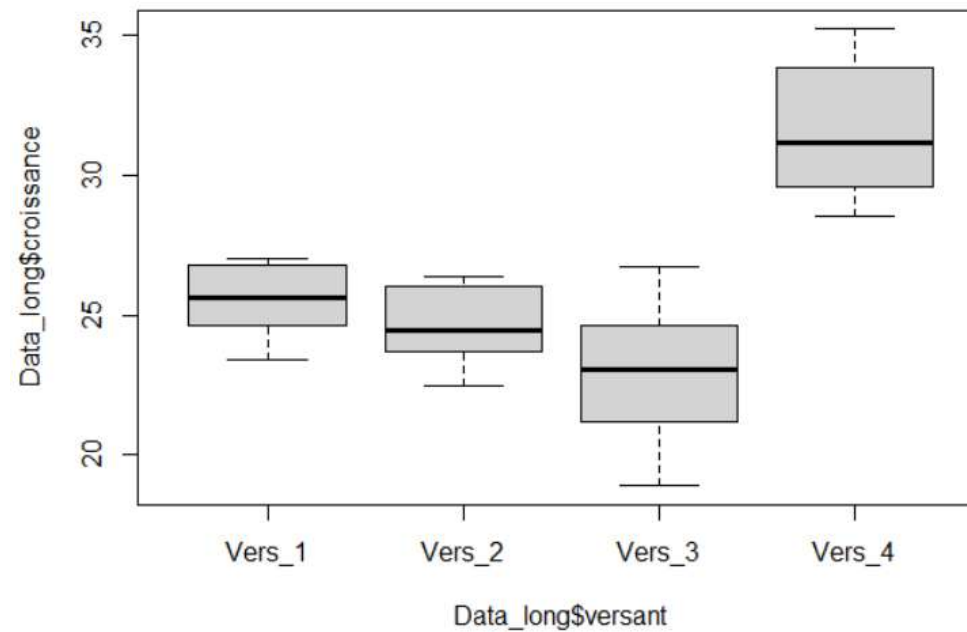
	station	versant	croissance
1	1	Vers_1	23.4
2	2	Vers_1	24.4
3	3	Vers_1	24.6
4	4	Vers_1	24.9
5	5	Vers_1	25.0
6	6	Vers_1	26.2
7	7	Vers_1	26.3
8	8	Vers_1	26.8
9	9	Vers_1	26.9
10	10	Vers_1	27.0
11	1	Vers_2	22.5
12	2	Vers_2	22.9

Effectuer le test sous R

Code :

```
boxplot(croissance~versant,Data_long)
```

Sortie :



Effectuer le test sous R

Code : `aov_Data_long<-lm(croissance~versant,data=Data_long)` ← **Fonction lm() :**
valable pour plan
équilibrés ou non

```
shapiro.test(residuals(aov_Data_long))
```

```
plot(aov_Data_long,2,pch=19)
```

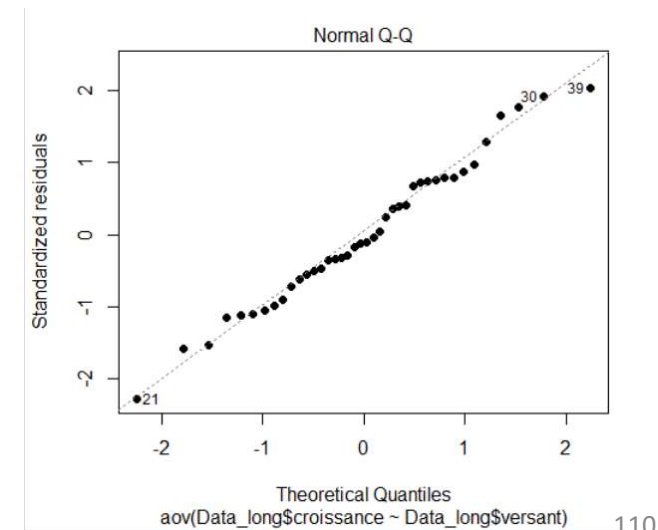
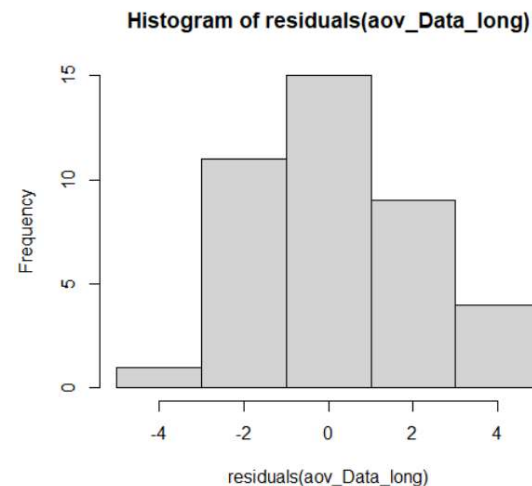
```
hist(residuals(aov_Data_long),breaks = seq(-5,5,2))
```

Qu'en pensez vous ?

Sortie :

Shapiro-Wilk normality test

```
data: residuals(aov_Data_long)  
W = 0.98385, p-value = 0.8276
```



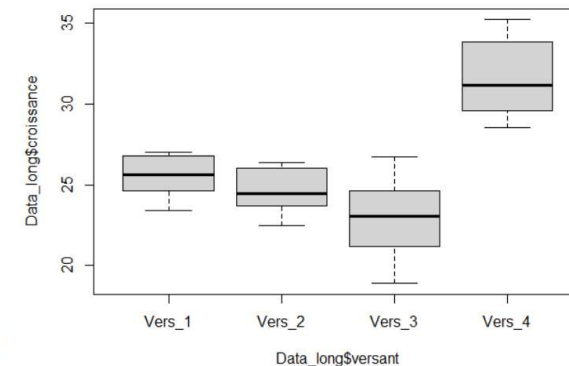
Effectuer le test sous R

Code : `bartlett.test(croissance~versant, data=Data_long)`

Qu'en pensez vous ?

Sortie : Bartlett test of homogeneity of variances

data: croissance by versant
Bartlett's K-squared = 6.2563, df = 3, p-value = 0.09979



Alternative → Test de Levene = moins sensible aux écarts de normalité

```
library(car)  
leveneTest(croissance~versant, data=Data_long)
```

Ici, même conclusion → $p=0.127 > 0.05$

Effectuer le test sous R

Code : `summary(aov_Data_long)`

Sortie :

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	25.5500	0.6165	41.442	< 2e-16	***
versantVers_2	-0.9600	0.8719	-1.101	0.27819	
versantVers_3	-2.4100	0.8719	-2.764	0.00894	**
versantVers_4	5.8800	0.8719	6.744	7.13e-08	***

Estimation de la \neq entre moy. de chaque éch. et moy. de ref.

Moy. de la modalité de référence
(défaut = 1^{er} ordre alphabétique ou fixé manuellement)

Test de significativité \neq entre moy.

Residual standard error: 1.95 on 36 degrees of freedom
Multiple R-squared: 0.7438, Adjusted R-squared: 0.7225
F-statistic: 34.84 on 3 and 36 DF, p-value: 9.61e-11

R^2 = coefficient de détermination multiple [0 ; 1]

→ traduit la qualité d'ajustement du modèle.

→ Représente la part de variation de Y expliquée par la VI (ici = facteur « versant »).

$$R^2 = \text{SCE}_{\text{inter}} / \text{SCE}_{\text{total}} = \text{SCE}_{\text{inter}} / (\text{SCE}_{\text{inter}} + \text{SCE}_{\text{intra}})$$

Résultats test Fisher-Snedecor
(F_{obs} , ddl, p-val)

Effectuer le test sous R

Code :

```
anova(aov_Data_long)
```

anova() (sans majuscule!) permet d'afficher la table classique d'analyse de la variance avec l'effet global du facteur fixe (ici = « versant ») dont on test l'effet sur la VD (fonction équivalente de « drop1() » dans le Mooc)

Sortie :

Analysis of Variance Table

Source de variation de Y = Response: croissance

		Df	Sum Sq	Mean Sq	F value	Pr(>F)
Inter (expliquée)	→ versant	3	397.29	132.430	34.84	9.61e-11 ***
Intra (résiduelle)	→ Residuals	36	136.84	3.801		

↑ ↑ ↑ ↑

ddl SCE S^2 F_{obs}

p-val test Fisher-Snedecor

(même résultat que précédemment)

CCL → H1: Différence hautement significative entre au moins 2 groupes... mais lesquels?

Effectuer le test sous R

Code :

```
library(multcomp)
tukey <- glht(aov_Data_long, linfct=mcp(versant="Tukey"))
summary(tukey)
```

↑
Résultats du
modèle

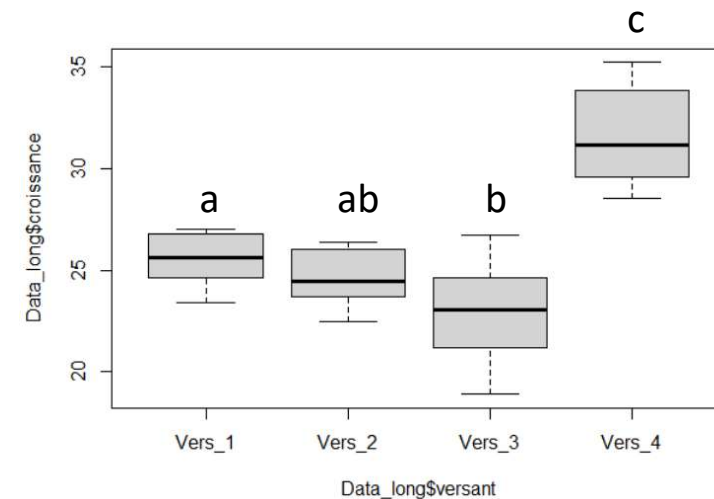
↗
Facteur fixe = VI définissant
les groupes à comparer

↖
Test post-hoc Tukey :
valide pour plan équilibrés

Sortie :

Linear Hypotheses:

	Estimate	Std. Error	t value	Pr(> t)
Vers_2 - Vers_1 == 0	-0.9600	0.8719	-1.101	0.6912
Vers_3 - Vers_1 == 0	-2.4100	0.8719	-2.764	0.0425 *
Vers_4 - Vers_1 == 0	5.8800	0.8719	6.744	<0.001 ***
Vers_3 - Vers_2 == 0	-1.4500	0.8719	-1.663	0.3577
Vers_4 - Vers_2 == 0	6.8400	0.8719	7.845	<0.001 ***
Vers_4 - Vers_3 == 0	8.2900	0.8719	9.508	<0.001 ***



CCL : La croissance des conifères sur le versant 4 est significativement > versant 1 > versant 3 (p-value < 0.001).
Le versant 2 montre une croissance des conifères significativement < versant 4 (p-value < 0.001), mais non significativement différente de celle des versants 1 et 3.

Autres types d'ANOVA

Pour info; Il existe d'autres types d'ANOVA, plus complexes, par exemple :

- **ANOVA à 2 facteurs croisés** pour :

- plan **équilibré** (ANOVA type I) → on visualise les résultats avec `summary()` et `anova()` {base}
- plan **déséquilibré** (ANOVA de type III)
 - rajouter dans fonction `lm()` l'argument optionnel :
`contrasts = list(facteur1 = contr.sum, facteur2 = contr.sum)`
(permet que le calcul de SCEs soient exactes dans un modèle d'ANOVA III)
 - on visualise les résultats avec `summary()` et `Anova()` {car}

(NB: pas de ≠ dans calcul des SCEs entre une ANOVA 1 facteur sur plan équilibré ou déséquilibré)

- **ANOVA sur plan hiérarchisé** (ANOVA type III) → visualise résultats avec `summary()` et `Anova()` {car}
 - si **déséquilibré**, rajouter dans `lm()` `contrasts = list(facteur1 = contr.sum, facteur2 = contr.sum)`
- ANOVA sur mesures répétées, plan mixte ou blocs → utiliser fonction `lme()` {nlme} ou `lme4`

Autres types d'ANOVA : 2 facteurs

Ex.1 : `aov2F <- lm(rendement ~ variete * engrais, data= ble)` #(charger jeu de données «ble» sur Moodle)

`table(ble$variete,ble$engrais)`

	Eng_A	Eng_B	Eng_C
var_1	3	3	3
var_2	3	3	3
var_3	3	3	3
var_4	3	3	3

← De quel type de plan s'agit il?

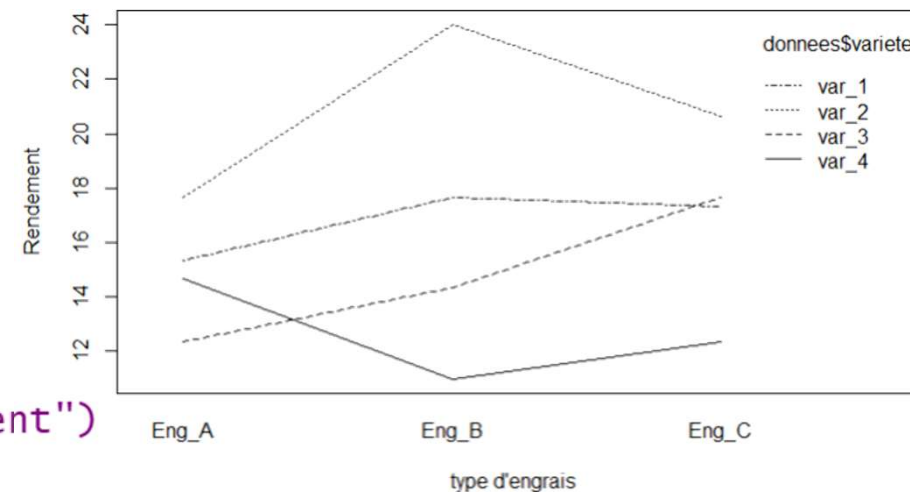
`interaction.plot(ble$engrais,ble$variete,
ble$rendement, fixed = T,
xlab = "type d'engrais",ylab="Rendement")`

`anova(aov2F)`

Response: rendement

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
variete	3	322.08	107.361	124.677	9.043e-15	***
engrais	2	28.50	14.250	16.548	3.042e-05	***
variete:engrais	6	105.50	17.583	20.419	2.447e-08	***
Residuals	24	20.67	0.861			

Les courbes se croisent = interaction ?



3 effets sont testés :

- Effet principal (= seul) variété
- Effet principal type d'engrais
- Interaction variété x type d'engrais

CCL : Interaction hautement signif → l'effet du type d'engrais sur le rendement dépend de la variété.

Autres types d'ANOVA : 2 facteurs

Ex.2 : `aov2_ex2<-lm(Y~V*N, data=oats)` # jeu de données « oats » inclus dans package MASS (à charger)

```
library(MASS) ; oats<-get(data("oats"))
```

```
table(oats$V,oats$N)
```

	0.0cwt	0.2cwt	0.4cwt	0.6cwt
Golden.rain	6	6	6	6
Marvellous	6	6	6	6
Victory	6	6	6	6

← *Quel type de plan?*

```
interaction.plot(oats$N,oats$V,oats$Y, fixed = T,
  xlab = "Concentration azote",ylab="Rendement")
```

```
anova(aov2_ex2)
```

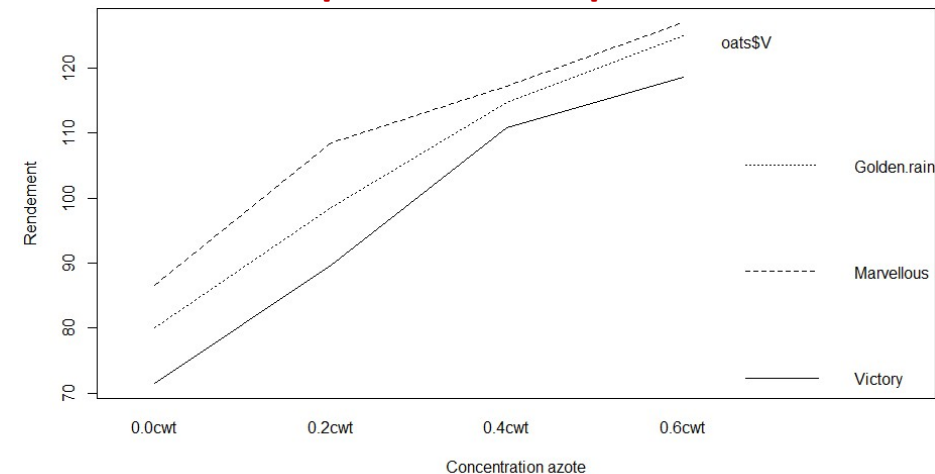
```
Response: Y
      Df Sum Sq Mean Sq F value    Pr(>F)
V       2  1786.4   893.2   1.7949    0.1750
N       3 20020.5  6673.5  13.4108 8.367e-07 ***
V:N     6   321.7    53.6   0.1078    0.9952
Residuals 60 29857.3   497.6
```

```
aov2_ex3<-lm(Y~V+N, data=oats)
```

```
anova(aov2_ex3)
```

```
Response: Y
      Df Sum Sq Mean Sq F value    Pr(>F)
V       2  1786.4   893.2   1.9533    0.1499
N       3 20020.5  6673.5  14.5946 2.155e-07 ***
Residuals 66 30179.1   457.3
```

Courbes « parallèles » = pas d'interaction

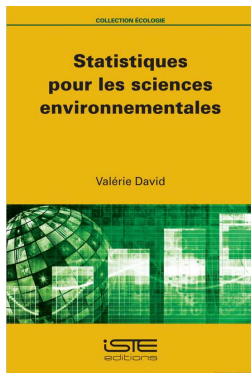


3 effets sont testés :

- Effet principal variété d'avoine
- Effet principal concentration azote
- Interaction variété x concentration azote

Comme interaction NS, on recalcule les p-val pour les effets principaux (seul concentration azote significatif)

Pour aller plus loin ...



Ebook 12 E :

<https://www.istegroup.com/fr/produit/statistiques-pour-les-sciences-environnementales/>

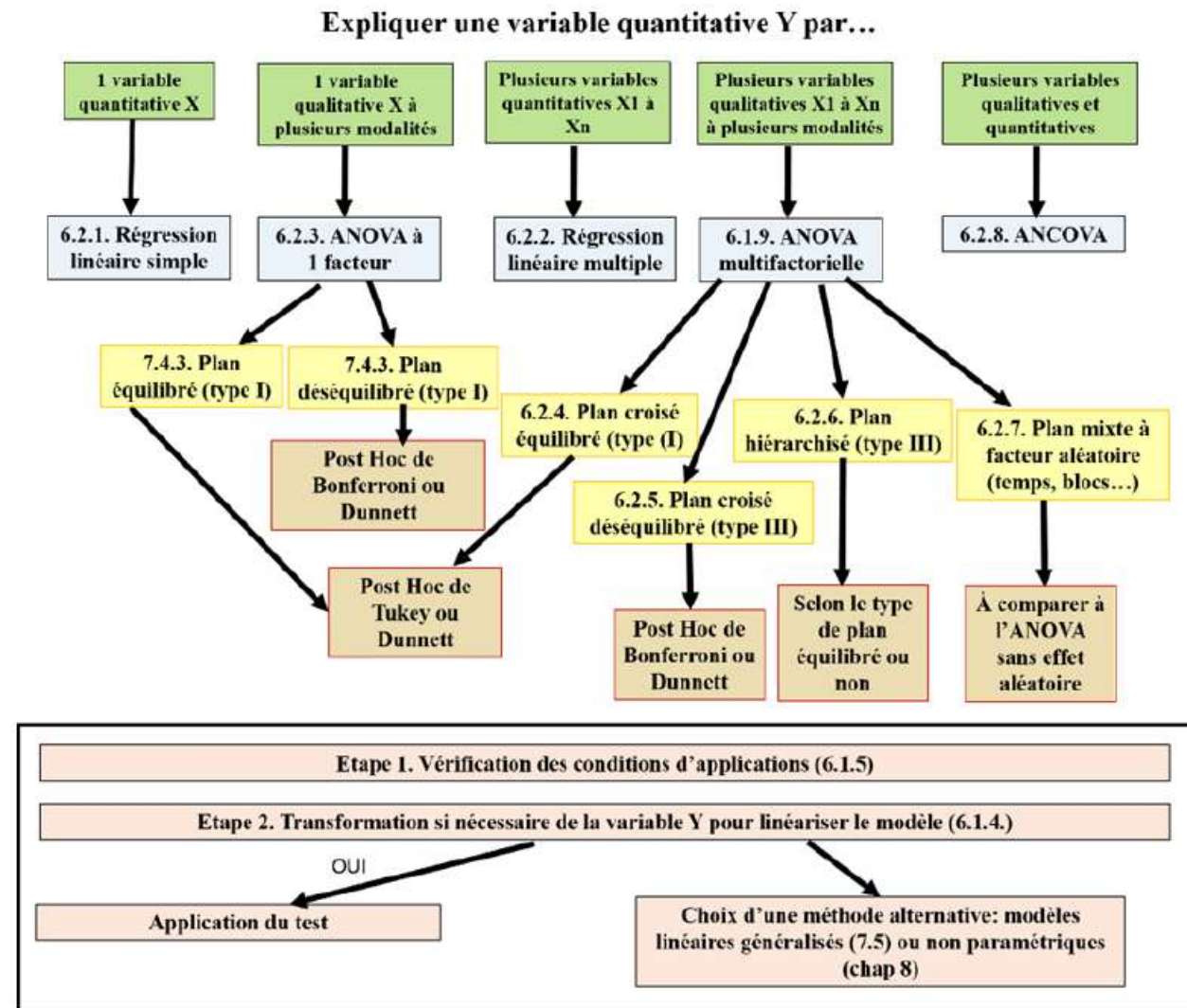


Figure 7.7. Clé de choix du plan de modèle linéaire à mettre en place et étapes principales