TECHNIQUES STATISTIQUES POUR L'EXPERIMENTATION

BORDEAUX SCIENCES AGRO

Semestre 7

Florent ARNAL

florent.arnal@u-bordeaux.fr

Septembre 2018

PRESENTATION DES MODALITES

Formation - semaine 1				
03/09/18				07/09/18
L	M	M	J	V
CM1 intro RH + recrutement	CM 1	CM 2	Droit du travail	TD 3 gérer les travaux à la vigne
TD 1 Table ronde	Anglais renforcé	CM 2 Conduite de réunion	Droit du travail	CM 3
Anglais	TD autonomie	TD 2 conduite de réunion	Sport	TD 2
	TPNE	TD 1		TD4 Motiv. Au travail

Formation - semaine 2				
10/09/18				14/09/18
L	M	M	J	V
TPNE		TD7 Manager les équipes en GEA	CM 4	CM 5
TD5 manag. Situationnel	Anglais renforcé	TPNE	TPNE	TD 4
Anglais	TD 6 Entretiens	Droit du travail	Sport	TPNE
	TD 3	Droit du travail		TD8 entretien pro

1

Organisation du module

Formation - semaine 3				
17/09/18				21/09/18
L	M	M	J	V
TD 5	TPNE	TPNE	TD 6 (1 h) TPNE	
TPNE	Anglais renforcé	TPNE	TPNE	Eval RH
Anglais	TD9 CV+LM + entreprises libérées	TD 6 (1 h) TPNE	Sport	Eval Stats
	TPNE	TPNE		Eval RH

2

Objectif du module

Traiter des données issues d'expérimentations pour mettre en évidence d'éventuels effets liés à des éléments contrôlés ou provoqués.

Un exemple tiré d'un mail :

Bonjour,
Je suis actuellement en stage en … sur le thème de la résistance de la pomme de terre OGM au mildiou et son impact sur l'agro-biodiversité.
L'expérimentation se présente ainsi :
3 variétés de pommes de terre
3 régimes de pulvérisation de fongicide
Blocs en randomisation
Le premier sujet … était de tester les résistances des pommes de terre en fonction de la variété et du régime de pulvérisation …

Plan du cours

CM 1 : Présentation du module
Rappels et compléments de statistique liés au cours de 1A

CM 2 : Classe inversée sur l'anova

CM 3 : Tests post hoc, Dispositifs expérimentaux

CM 4 :
Généralisation de l'Anova, Tests non paramétriques

CM 5 : Transformation de variables, Prolongements

INTRODUCTION : La démarche expérimentale

Deux approches fondamentalement différentes

Approche expérimentale

Détection des relations causales
Modèle défini à priori

Conditions contrôlées
Phénomènes provoqués

Difficultés liées à l'inférence
(Estimation de paramètres inconnus,
Comparaisons, ...)

Approche descriptive

Description d'une population,
d'un système
(à l'aide de graphiques et paramètres)

Grande résolution
Absence de biais

Modèle a postériori

5

Schéma général d'une expérimentation

Définition des objectifs

Définition des individus
(unités expérimentales)

Définition des facteurs
(testés ou contrôlés)

Définition des variables

Dispositif expérimental

Traitement

Interprétation

6

Vocabulaire

Facteur = Variable qualitative (catégorielle) dont les différentes valeurs prises sont appelés niveaux, modalités voire catégories.

- *Facteurs provoqués*

introduits volontairement

- *Facteurs aléatoires*

inhérents au milieu (terrain, environnement de l'essai) appelés facteurs contrôlés lorsque le dispositif expérimental utilisé les prend en compte

Vocabulaire

Facteurs aléatoires

=> Gradient d'hétérogénéité en expérimentations végétales

Liés au terrain :
-Hétérogénéités **naturelles** (pente, cailloux, ...)
-Hétérogénéités liées à **l'homme** (haie par exemple)

Liés aux interventions sur l'expérimentation
-hétérogénéités liées au **travail du sol** (réglage des outils, conditions des interventions)
-hétérogénéités liées aux **interventions culturales** (applications de produits phytosanitaires, de fertilisants)

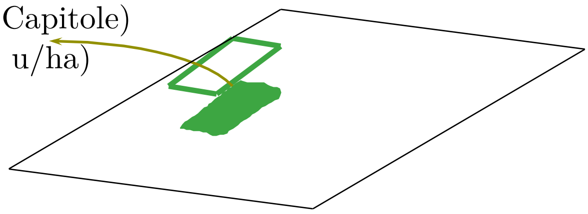
Hétérogénéités potentielles ou réelles ?

Vocabulaire

Traitement = combinaison des modalités, niveaux des facteurs étudiés
Ex : Essai de fertilisation azotée sur blé d'hiver
2 Facteurs étudiés : La *fumure* et les *variétés de blé*
4 variétés, 3 niveaux de fumure : 12 traitements

Unité expérimentale (U.E.) = entité qui reçoit un traitement

1 variété (ex : Capitole)
1 dose (ex : 60 u/ha)



9

Mise en place d'un essai

Choix des niveaux du Facteur
→ Facteurs qualitatifs / quantitatifs

Définition du domaine d'étude
Progression liée aux usages voire arithmétique ou géométrique

Mise en place d'un éventuel témoin
→ non traité / traitement de référence

Nombre de répétitions
→ Puissance de l'essai, CV

Expérience à plusieurs facteurs
→ Mise en place d'un dispositif expérimental
Ex : dispositif complet, avec randomisation, ...

10

PRESENTATION DU PRINCIPE DE L'ANALYSE DE VARIANCE (ANOVA)

Cas de dispositifs sans facteurs emboités, à effets fixes (Anova de type I)

On considère un facteur prenant p niveaux de moyennes $\mu_1, \cdots, \mu_p$. On teste l'hypothèse selon laquelle le facteur étudié n'a pas d'effet. Ainsi, on a

$\mathcal{H}_0 : \mu_1 = \cdots = \mu_p$

$\mathcal{H}_1$: "deux moyennes au moins sont différentes"

Dans le cas de 4 modalités :
6 tests de Student a priori à mettre en place

Risque alpha au global trop élevé

Répétition de tests et Risque Alpha

Calcul du risque α_{global} lorsque l'on effectue 6 tests :
On considère que $\mathcal{H}_0$ est vraie c'est-à-dire qu'il n'y a pas d'effet significatif du facteur étudié (toutes les moyennes sont égales).

$1-\alpha_{global}$ correspond à la probabilité qu'aucun des 6 tests ne mette en évidence (à tort) un effet significatif.
Avec un seuil de signification de 5 %, pour 6 tests indépendants, on a :

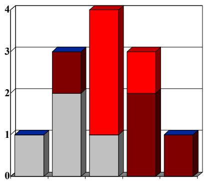
$$1-\alpha_{global} = 0,95^6$$

$$\alpha_{global} \simeq 26,5\%$$

On ne peut donc envisager de réaliser plusieurs tests de Student ...

Plus généralement, pour n tests, avec un risque α, on a

$$\alpha_{global} = 1-(1-\alpha)^n$$



Variation totale

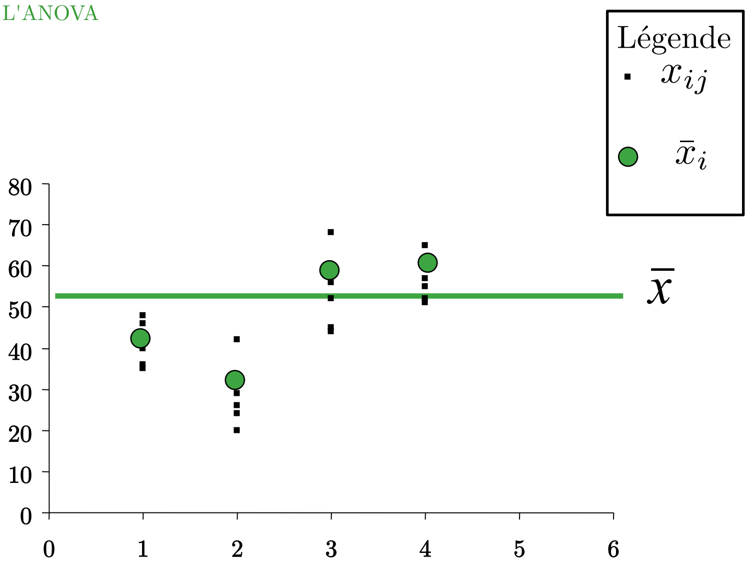
Variabilité
inter
(dûe aux traitements)

Variabilité
intra
(résiduelle)

Si la variabilité inter (liée au facteur)
est **supérieure** à la variabilité résiduelle

$\mathcal{H}_0$ est rejetée (effet factoriel significatif)

Principe de L'ANOVA



15

Principe de L'ANOVA

$$x_{ij} - \bar{x} = (x_{ij} - \bar{x}_i) + (\bar{x}_i - \bar{x}) \text{ avec } e_{ij} = x_{ij} - \bar{x}_i$$

donc $$x_{ij} = \bar{x} + (\bar{x}_i - \bar{x}) + e_{ij}$$

Observation

Part expliquée par l'effet du traitement

Moyenne globale

Résidu e_{ij}

Les résidus $e_{ij} = x_{ij} - \bar{x}_i$, par niveau (et donc au global), sont de somme et de moyenne nulle.

16

Principe de L'ANOVA

MODELE LINEAIRE

$$X_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

où

- μ est la moyenne (globale) associée aux p populations.
- X_{ij} est une variable aléatoire (réponse quantitative). Les X_{ij} sont indépendantes.
- ε_{ij} est une variable aléatoire d'erreur (non observées). On a : $\varepsilon_{ij} \sim \mathcal{N}(0;\sigma)$, σ étant un paramètre inconnu (à estimer).
- α_i correspond à l'effet associé à la modalité i avec $\sum_{i=1}^{p} \alpha_i = 0$.

  Une estimation de cet effet est donnée par

$$\hat{\alpha}_i = \bar{x}_i - \bar{x}$$

Le modèle peut également s'écrire

$$X_{ij} = \mu_i + \varepsilon_{ij}$$

où $\mu_i = \mu + \alpha_i$ dont une estimation est $\hat{\mu}_i = \bar{x}_i$.

17

En posant

- $SCE_{fact} = SCE_{inter} = \sum_{i,j} (\bar{x}_i - \bar{x})^2$
- $SCE_{res} = SCE_{intra} = \sum_{i,j} (x_{ij} - \bar{x}_i)^2$
- $SCE_{totale} = \sum_{i,j} (x_{ij} - \bar{x})^2$.

On a

$$SCE_{totale} = SCE_{fact} + SCE_{res}$$

Source	Df	SCE	CM	F_{obs}	p-valeur
Factorielle	$p-1$	SCE_{fact}	$CM_{fact} = \dfrac{SCE_{fact}}{p-1}$	f_{obs}	$\mathbb{P}(F_{obs} > f_{obs})$
Résiduelle	$N-p$	SCE_{res}	$CM_{res} = \dfrac{SCE_{res}}{N-p}$		
Totale	$N-1$	SCE_{totale}			

pour p modalités, N observations. Sous $\mathcal{H}_0$,

$$F_{obs} = \frac{CM_{fact}}{CM_{res}} \sim \mathcal{F}(p-1; N-p)$$

18

Le tableau de L'ANOVA

Avec R :

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Facteur	...	...	...	...	...
Residuals	...	...	...		

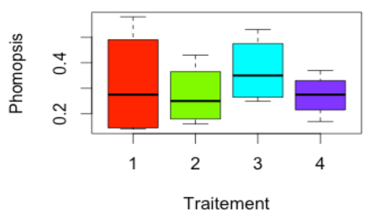
Code R :

```
anova = aov(variable ~ facteur)
summary(anova)
```


Un exemple

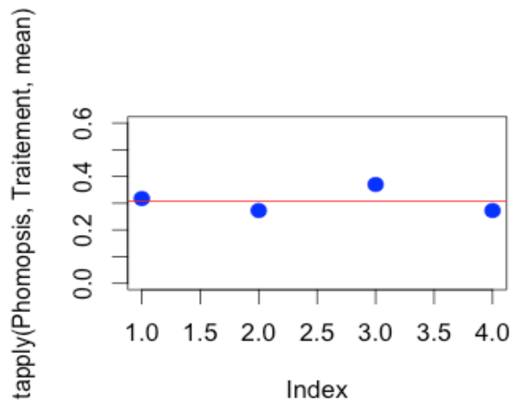
Observation des pourritures sur châtaignes – essai fertilisation

Traitement	Bloc	% obs Phomopsis
1	1	58%
1	2	14%
1	3	40%
1	4	15%
2	1	30%
2	2	43%
2	3	20%
2	4	16%
3	1	25%
3	2	42%
3	3	53%
3	4	28%
4	1	26%
4	2	37%
4	3	17%
4	4	29%



Principe de L'ANOVA

58	14	40	15	$\bar{x}_1$ =	31,75
30	43	20	16	$\bar{x}_2$ =	27,25
25	42	53	28	$\bar{x}_3$ =	37
26	37	17	29	$\bar{x}_4$ =	27,25
				$\bar{x}$ =	30,8



21

Principe de L'ANOVA

Code R

```
Traitement = as.factor(Traitement)
anova = aov(Phomopsis ~ Traitement)
summary(anova)
```

Résultat R

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Traitement	3	0.025	0.008606	0.413	0.746
Residuals	12	0.249	0.020819		

22

CONDITIONS POUR LA MISE EN ŒUVRE D'UNE ANOVA

23

Conditions

$$X_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

où

- X_{ij} est une variable aléatoire (réponse quantitative).
  Les X_{ij} sont indépendantes et gaussiennes.

  $$X_{ij} \sim \mathcal{N}(\mu_i; \sigma)$$

- ε_{ij} est une variable aléatoire d'erreur (non observées). On a :

  $$\varepsilon_{ij} \sim \mathcal{N}(0; \sigma)$$

  σ étant un paramètre inconnu avec

  $$\hat{\sigma}^2 = CM_{res}$$

Les hypothèses de normalité et homoscédasticité se vérifient sur les résidus.
L'indépendance est liée au contexte de l'expérimentation.

24

Pour vérifier les conditions d'utilisation :

1. Égalité des variances :
   Méthodes graphiques, tests de Bartlett, Levene ou Brown-Forsythe avec

$$\mathcal{H}_0 : \sigma_1^2 = \cdots = \sigma_p^2$$

2. Normalité de la distribution des résidus :
   Graphique Quantile-Quantile, test de Shapiro avec

$$\mathcal{H}_0 : \text{" La distribution est normale "}$$

Autour de l'homoscédasticité :

- Test de Bartlett (couramment utilisé, notamment en cas de distributions gaussiennes) avec une statistique distribuée suivant une loi du Khi-deux.

- Test de Levene correspondant à une anova sur les valeurs absolues des résidus.

- Test de Brown-Forsythe, variante du test de Levene sur les médianes et non les moyennes.
  Gain de puissance en cas de non normalité.

Conditions

Normalité et petits échantillons
Graphique Quantile-Quantile (Droite de Henry)

Notons x_i les différentes valeurs à considérer et f_i les fréquences cumulées définies par :

$$f_i = \frac{\text{nb val} \leq x_i}{N+1}$$

Soit le quantile u_i^* tel que

$$\mathbb{P}(X^* \leq u_i^*) = f_i \text{ ie } u_i^* = \phi^{-1}(f_i)$$

Si $X \sim \mathcal{N}(\mu, \sigma)$, on doit avoir

$$\mathbb{P}\left(X \leq x_i\right) \simeq f_i$$

ie

$$\mathbb{P}\left(X^* \leq \frac{x_i - \mu}{\sigma}\right) \simeq \mathbb{P}(X^* \leq u_i^*)$$

Il en résulte que

$$\frac{x_i - \mu}{\sigma} \simeq u_i^*$$

En conséquence, les points de coordonnées (x_i, u_i^*) doivent être sensiblement alignés.

27

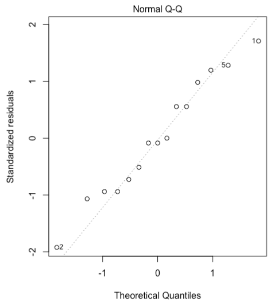
Conditions

Graphiques diagnostiques sous R

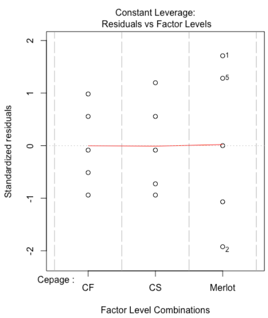
Code R :

```
par(mfrow=c(2,1))
plot(anova, which=c(2,5))
```

QQ-norm des résidus

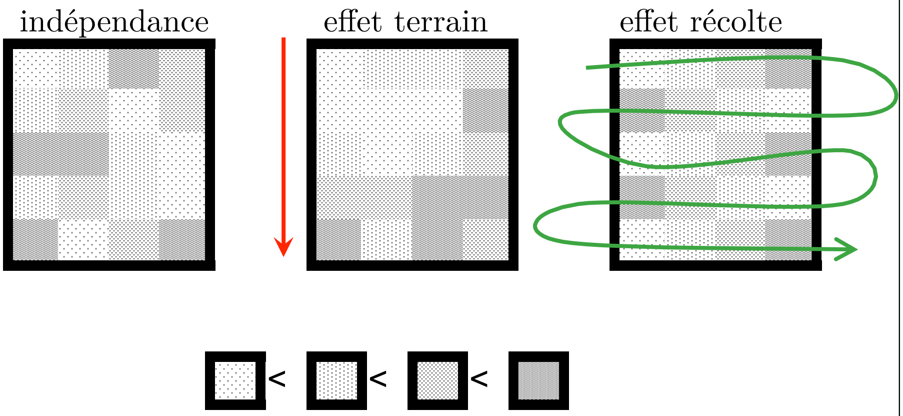


Résidus studentisés voisins de $\frac{e_{ij}}{\hat{\sigma}}$.
La plupart des résidus doivent appartenir à $[-2; 2]$ et les nuages doivent avoir des allures similaires.



28

Indépendance



Exemple de cartographie des résidus
Essai avec effet terrain très localisé

	1	2	3	4	5	6	7	8	9	10
1	F1n7 b1	F1n13 b1	F1n11 b1		F1n2 b1		F1n7 b2	F1n12 b2	F1n9 b2	F1n6 b2
2			F1n12 b1		F1n10 b1		F1n5 b2	F1n3 b2	F1n15 b2	F1n13 b2
3		F1n9 b1	F1n15 b1	F1n14 b1	F1n5 b1	F1n1 b2		F1n10 b2	F1n8 b2	
4	F1n6 b1	F1n8 b1	F1n1 b1	F1n3 b1	F1n4 b1	F1n4 b2	F1n2 b2	F1n14 b2	F1n11 b2	
5	F1n11 b4	F1n3 b4	F1n5 b4	F1n13 b4	F1n7 b4	F1n3 b3	F1n5 b3	F1n6 b3	F1n14 b3	F1n2 b3
6	F1n1 b4		F1n8 b4		F1n6 b4	F1n7 b3	F1n1 b3		F1n15 b3	F1n9 b3
7	F1n14 b4		F1n9 b4	F1n4 b4	F1n2 b4	F1n13 b3		F1n11 b3	F1n12 b3	F1n10 b3
8	F1n10 b4		F1n12 b4		F1n15 b4			F1n8 b3		F1n4 b3

< -.4685   < 0   < .4684   < 999999

COMPARAISONS MULTIPLES DE MOYENNES

TESTS POST HOC

Tester un effet factoriel conduit à tester

$\mathcal{H}_0 : \mu_1 = \mu_2 = \cdots = \mu_p$
contre
$\mathcal{H}_1 : \exists (i,j)$ tel que $\mu_i \neq \mu_j$

$\mathcal{H}_0$ non rejetée

Pas d'effet significatif

$\mathcal{H}_0$ rejetée

Effet significatif

Comparaison multiple de moyennes

Principe basé sur les tests de Student

Cas d'une comparaison de 2 moyennes :
Sous $\mathcal{H}_0 : \mu_i = \mu_j$, on a :

$$T_{obs} = \frac{\overline{X}_i - \overline{X}_j}{\sqrt{CM \times \left(\frac{1}{n_i} + \frac{1}{n_j}\right)}} \sim \mathcal{T}(ddl_{CM})$$

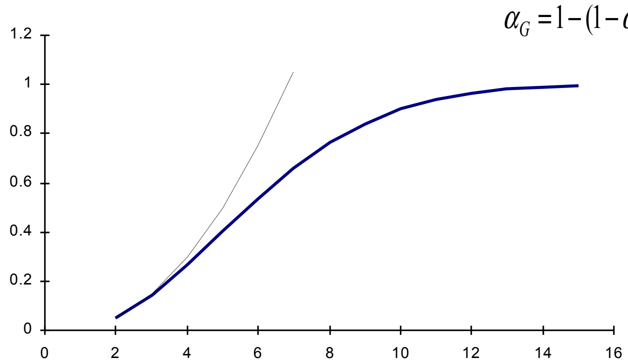
Pour un test post hoc, CM est remplacé par CM_{res}

Sous $\mathcal{H}_0 : \mu_i = \mu_j$, on a donc :

$$T_{obs} = \frac{\overline{X}_i - \overline{X}_j}{\sqrt{CM_{res} \times \left(\frac{1}{n_i} + \frac{1}{n_j}\right)}} \sim \mathcal{T}(N - p)$$

Inconvénient : chaque test est réalisé au risque α

$$\alpha_G = 1 - (1-\alpha)^{\frac{p(p-1)}{2}}$$



Test de Bonferonni

Adaptation du risque de première espèce α de chaque test pour minimiser le risque global α_{global} basée sur la relation

$$\alpha_{global=1-(1-\alpha)^{\text{nb tests}}} \underset{\alpha \to 0}{\sim} (\text{nb tests}) \times \alpha$$

On a donc (usuellement)

$$\alpha = \frac{\alpha_{global}}{\text{nombre de tests}} = \frac{0,05}{\frac{p(p-1)}{2}}$$

Ainsi, lorsque l'on compare trois populations, on considère $\alpha = \dfrac{0,05}{3} \simeq 1,7\%$.

Avec R :
La fonction *pairwise.t.test( )* renvoie les p-valeurs en les multipliant par le nombre de comparaisons.
Il faut donc les comparer au seuil de signification (5% usuellement).

Exemple

Expérimentation sur l'éclaircissage chimique / GALA

Témoin	89	94	110	123
Wilthin	78	91	91	116
Prog classique	45	53	57	67

Résultat

	Df	SumSq	Mean Sq	F value	Pr(>F)
methode		5246			0.0019 **
Residuals		1731			

Test de Bonferonni

pairwise.t.test(var,methode,p.adj="bonferroni",alternative="two.sided")

	classique	temoin
temoin	**0.0024**	-
wilthin	**0.0104**	1.0000

Témoin	104	a
Wilthin	94	a
Classique	55.5	b

Critique du test :
Risque élémentaire choisi faible induisant une puissance souvent insuffisante

Autres tests post hoc (**plus puissants**)

- Test de Bonferroni-Holm
  Adaptation du risque alpha suivant le nombre de tests restants (après avoir ordonné les p-valeurs)

- Test de Newman-Keuls (SNK)
  SNK.test() du package *agricolae*

- Test de Tukey

- Méthode des Contrastes :
  ANOVA (modèle linéaire) permettant de tester (p-1) groupements

pairwise.t.test(var,methode,p.adj="bonferroni",alternative="two.sided")

	classique	temoin
temoin	0.0024	-
wilthin	0.0104	1.0000

pairwise.t.test(x=var, g=methode, p.adjust.method = "holm")

	classique	temoin
temoin	0.0024	-
wilthin	0.007	0.3345

39

ANOVA À PLUSIEURS FACTEURS & MESURES RÉPÉTÉES

40

ANOVA à 2 facteurs *avec interactions (répétitions)*

Exemple introductif :
Etude de la durée entre la date de fin de chargement en sucre des raisins (date déterminée par différentes mesures et calculs) et la date de vendange suivant deux variétés et 4 millésimes.

Effet des variétés :
les variétés ont-elles des comportements différents ?

Effet du millésime :
les quatre millésimes donnent-ils des résultats différents ?

Interactions :
les variétés ont-elles des comportements différents suivant le millésime ?

41

ANOVA à 2 facteurs avec interactions (répétitions)

	Variété 1	Variété 2	
Millésime 1	$x_{111}, \cdots, x_{11n}$ (moyenne $\bar{x}_{11}$)	$x_{121}, \cdots, x_{12n}$ (moyenne $\bar{x}_{12}$)	moyenne $\bar{x}_{1\bullet}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
Millésime 4	$x_{411}, \cdots, x_{41n}$ (moyenne $\bar{x}_{41}$)	$x_{421}, \cdots, x_{42n}$ (moyenne $\bar{x}_{42}$)	moyenne $\bar{x}_{4\bullet}$
	moyenne $\bar{x}_{\bullet 1}$	moyenne $\bar{x}_{\bullet 2}$	moyenne globale $\bar{x}$

Plan équilibré avec n répétitions par traitement (Variété*Millésime)

Plus généralement, on note p le nombre de niveau du facteur 1 et q celui du facteur 2.
Il y a ainsi $N = pqn$ observations.

42

ANOVA à 2 facteurs avec interactions (répétitions)

Modalité facteur 1

Modalité facteur 2

Effet traitement *ij*

$$x_{ijk} = \bar{x} + (\bar{x}_{ij} - \bar{x}) + (x_{ijk} - \bar{x}_{ij})$$

résidu

répétition

et

$$(\bar{x}_{ij} - \bar{x}) = (\bar{x}_{i\bullet} - \bar{x}) + (\bar{x}_{\bullet j} - \bar{x}) + (\bar{x}_{ij} - \bar{x}_{i\bullet} - \bar{x}_{\bullet j} + \bar{x})$$

Effet lié facteur 1 (modalité *i*)

Effet lié facteur 2 (modalité *j*)

Terme d'interaction

43

ANOVA à 2 facteurs avec interactions (répétitions)

Modèle linéaire pour deux facteurs avec interactions

$$X_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ijk} + \varepsilon_{ijk}$$

Effet du traitement *i* du Facteur 1

Effet du traitement *j* du Facteur 2

Effet des interactions

Résidus

44

ANOVA à 2 facteurs avec interactions (répétitions)

RELATION ENTRE LES SCE :
Cas deux facteurs avec *n* répétitions

$$\sum_{ijk}\left(x_{ijk}-\bar{x}\right)^2=\sum_{ijk}\left(\bar{x}_{i.}-\bar{x}\right)^2+\sum_{ijk}\left(\bar{x}_{.j}-\bar{x}\right)^2+\sum_{ijk}\left(\bar{x}_{ij}-\bar{x}_{i.}-\bar{x}_{.j}+\bar{x}\right)^2+\sum_{ijk}\left(x_{ijk}-\bar{x}_{ij}\right)^2$$

$$\mathrm{SCE_{tot}} = \mathrm{SCE_{F1}} + \mathrm{SCE_{F2}} + \mathrm{SCE_{inter}} + \mathrm{SCE_{res}}$$

45

ANOVA à 2 facteurs avec interactions (répétitions)

RELATION ENTRE LES ddl :
Cas deux facteurs avec *n* répétitions

Totale (npq unités) : npq-1
Facteur 1 (p modalités) : p-1
Facteur 2 (q modalités) : q-1
Interactions : $(p$-$1)(q$-$1)$
Résiduelle : $pq(n$-$1)$

$$\mathrm{ddl_{tot}} = \mathrm{ddl_{F1}} + \mathrm{ddl_{F2}} + \mathrm{ddl_{inter}} + \mathrm{ddl_{res}}$$

46

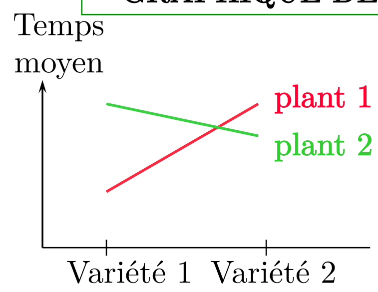
Hypothèses et principe de l'ANOVA
Cas deux facteurs avec n répétitions

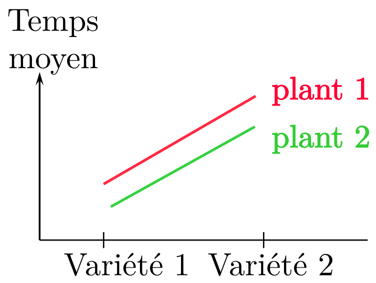
On suppose la normalité des résidus, leur indépendance et l'homoscédasticité des populations.

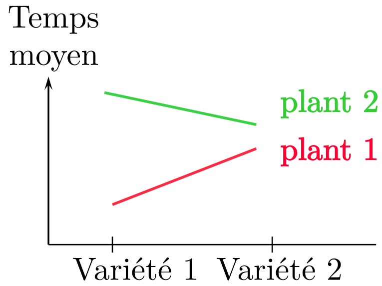
$\mathcal{H}_0$: « il n'y a pas d'effet du facteur 1 »
$\mathcal{H}_0$' : « il n'y a pas d'effet du facteur 2 »
$\mathcal{H}_0$'' : « il n'y a pas d'effet des interactions ».

Les différents tests vont s'effectuer en **comparant les carrés moyens concernés et le carré moyen résiduel.**
Ces tests seront toujours des tests **unilatéraux.**

GRAPHIQUE DES INTERACTIONS







Code R :
interaction.plot(facteur1, facteur2, variable)

ANOVA à 2 facteurs avec interactions (répétitions)

Code R :

% Définition des facteurs

```
millesime <- as.factor(millesime)
variete <- as.factor(variete)
```

% Anova

```
summary( aov( rdt ~ millesime * variete))
```


ANOVA à 2 facteurs avec interactions (répétitions)

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Variete	1	29376	29376	18.192	0.000165 ***
Millesime	3	1246	415	0.257	0.855685
Variete:Millesime	3	61435	20478	12.682	1.26e-05 ***
Residuals	32	51673	1615		

Effet des interactions <u>significatif</u>

➔ Effet variétal différent suivant le millésime

L'interprétation seule des effets principaux n'a aucun sens.

ANOVA à 2 facteurs avec interactions (répétitions)

Cas d'interaction traitements-blocs (parcelles) significative

Les blocs sont peut être :

mal choisis (trop différents) ou

mal disposés

Recherche d'éventuels résidus suspects permettant de déceler des parcelles suspectes (maladies par ex)

51

ANOVA à 2 facteurs *sans répétition*

Exemple :
L'influence de 3 traitements, à base de vitamines, est étudiée sur des animaux de 3 races différentes.
Le GMQ est mesuré au bout de 50 jours sur un seul animal pour chaque couple "race-traitement".

	Race A	Race B	Race C
T1	1260	1210	1190
T2	1290	1230	1230
T3	1380	1270	1220

52

RELATION ENTRE LES ddl : Cas deux facteurs sans répétition

Totale (pq unités) : pq-1
Facteur 1 (p modalités) : p-1
Facteur 2 (q modalités) : q-1
Résiduelle : (p-1)(q-1)

$$\boxed{ddl_{tot} = ddl_{F1} + ddl_{F2} + ddl_{res}}$$

Code R :

```
summary(aov(GMQ~traitement+race))
```

Résultat R :

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
traitement	2	7400.0	3700.0	4.7234	0.0885 .
race	2	15266.7	7633.3	9.7447	0.0290 *
Residuals	4	3133.3	783.3		

ANOVA à mesures répétées

L'anova à mesures répétées est utilisée lorsque l'on effectue des mesures successives sur différents sujets (individus) répartis en plusieurs groupes de traitement.

On est confronté à ce type d'analyse lorsque l'on considère, par exemple :
• l'évolution du pH de différents vins (sujets) durant plusieurs semaines consécutives
• l'évolution des notes pour différents vins notés par un jury constitué de plusieurs sujets (personnes).

La répétition des mesures induit au appariemment des données nécessitant un taritement particulier.

Usuellement, pour une anova à 1 facteur, la variabilié totale est décomposée en la variabilté inter (factorielle) et la variabilité résiduelle (within).

Dans un plan d'expérience à mesures répétées (à un facteur intra-sujet), la variabilité within peut être scindée en :
• la variabilité inter-sujets due aux différences entre les individus ;
• la variabilité intra-sujets, dans les mêmes conditions, un même sujet ne reproduit que rarement le même résultat.

55

ANOVA à mesures répétées

Le principe de l'analyse de variance à mesures répétées (ou à échantillons dépendants) consiste à éliminer cette variabilité due aux différences entre sujets et donc à soustraire la variabilité inter-sujets de la variabilité (résiduelle) "within".

La variabilité erreur par rapport à laquelle on teste l'effet du traitement est réduite, il est plus facile d'identifier les effets du traitement.
Ce type d'expérience (plans intra-sujets) est donc plus puissant que les plans où chaque sujet est soumis à une seule observation (plans inter-sujets).

Une ANOVA à mesures répétées à 1 critère de classification revient à effectuer **une anova à 2 critères de classification dans laquelle le second critère serait le critère sujet avec une seule observation** (Anova à 2 facteurs sans répétition).

Code R :
anova = aov(Variable ∼ Sujet+Semaine)
summary(anova)

56

DISPOSITIFS EXPERIMENTAUX

57

Objectif 1 : mesure d'un phénomène

↓

Il faut contrôler les facteurs de variations potentiels :

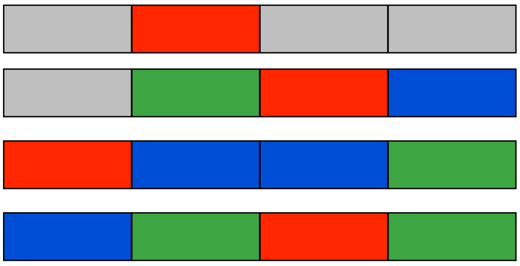
- effet terrain
- effet localisation (au sens large)
- effet juge

Objectif 2 : optimisation du coût expérimental

58

Randomisation totale

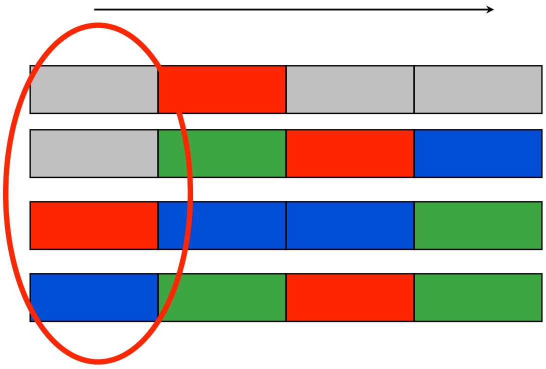
Modèle de base - Disposition aléatoire des parcelles
1 facteur étudié

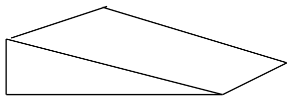


Plan factoriel à 1 facteur étudié, avec répétition

Dispositif en blocs

Ex : Effet terrain

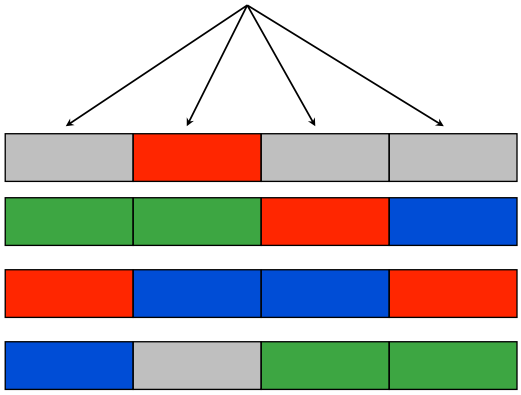




Ex : terrain pentu

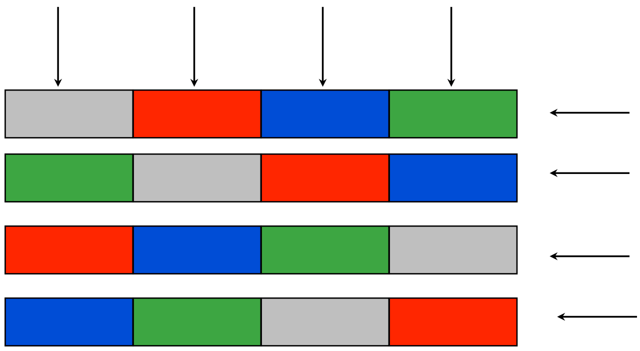
Zone favorable : sur-représentation d'un traitement

Blocs : zones homogènes (facteur contrôlé)



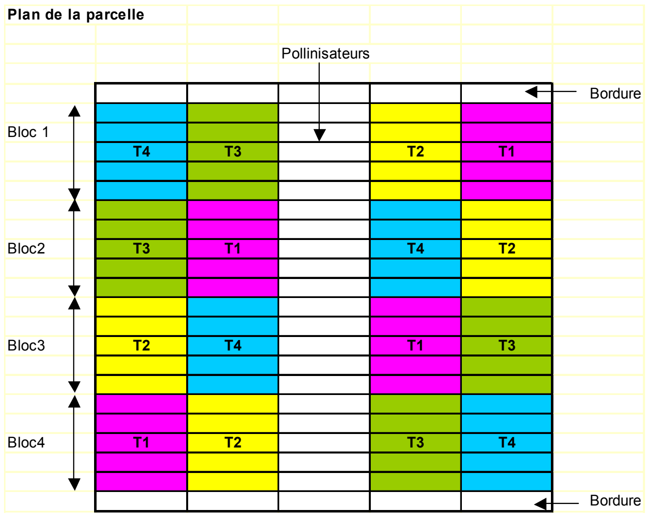
Carré latin : 1 facteur étudié + 2 facteurs aléatoires à prendre en compte.

Chaque traitement est présent
une seule une seule fois dans
chacune des lignes et chacune des colonnes.



Dispositifs

Essai fertilisation châtaigne



63

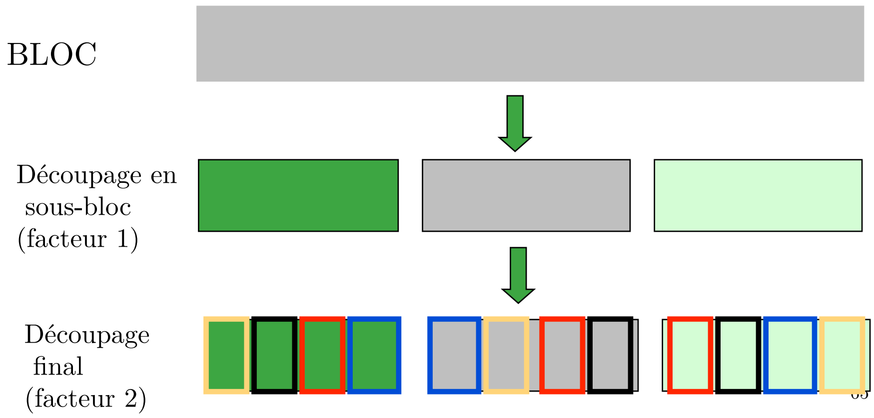
Traitement des données à l'aide d'une ANOVA à 3 facteurs

```
Lignes = as.factor(Lignes)
Colonnes = as.factor(Colonnes)
T = as.factor(Traitement)
anova = aov(Hauteur~Lignes+Colonnes+T)
summary(anova)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Lignes	3	661	220	2.088	0.203
Colonnes	3	2833	944	8.945	0.012 *
T	3	13616	4539	42.994	0.00018 ***
Residuals	6	633	106		

Dispositif en parcelles partagées (Split plot)

Principe : Hiérarchisation des facteurs
(1 principal et 1 secondaire=subsidiaire)
Découper les blocs en sous-blocs



Exemple de dispositif en split-plot

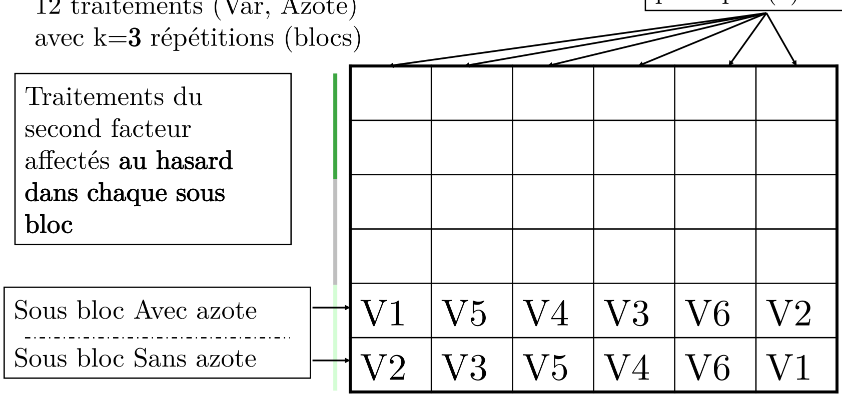
Facteur principal = Variété (p=6 variétés)

Facteur secondaire = dose d'azote (avec ou sans soit q=2 niveaux)

12 traitements (Var, Azote)
avec k=**3** répétitions (blocs)

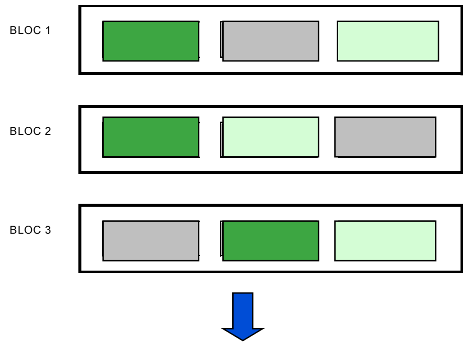
Chaque bloc est divisé en autant de sous blocs que de modalités du facteur principal (6)

Traitements du second facteur affectés **au hasard dans chaque sous bloc**



Dispositifs

Analyse de variance - Split plot
Anova sur Facteur 1

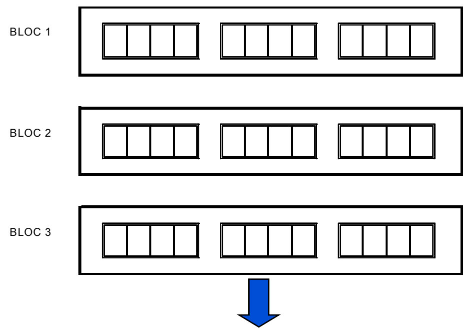


Équivalent à un dispositif 1 Facteur en blocs

Source des variations	*ddl*	
Totale 1	pk-1	17
Facteur 1	p-1	5
Blocs	k-1	2
Résiduelle	(p-1)(k-1)	10

Dispositifs

Analyse de variance - Split plot
Anova sur Facteur 2 et Interactions



Source des variations	*ddl*
Totale	pqk-1
Facteur 1	pk-1
Facteur 2	q-1
Interaction	(p-1)(q-1)
Résiduelle	p(q-1)(k-1)

Analyse de variance - Split plot
Facteur secondaire

Code R :

```
Var = as.factor(Var)
Bloc = as.factor(Bloc)
Azote = factor(Azote)

anova <- aov( Rdt~Var*Azote + Error(Bloc + Bloc:Var) )
summary(anova)
```


TRANSFORMATION
DE
VARIABLE

Lorsque la nature des données ne permet pas une réalisation correcte de l'Analyse de Variance, on peut être amené à transformer la variable pour améliorer

- la normalité
- égaliser les variances (homoscédasticité).

Les causes de non respect des hypothèses sont souvent connues a priori

⇓

Interprétation facilitée

Autre solution : tests non paramétriques

Type de problèmes et solutions

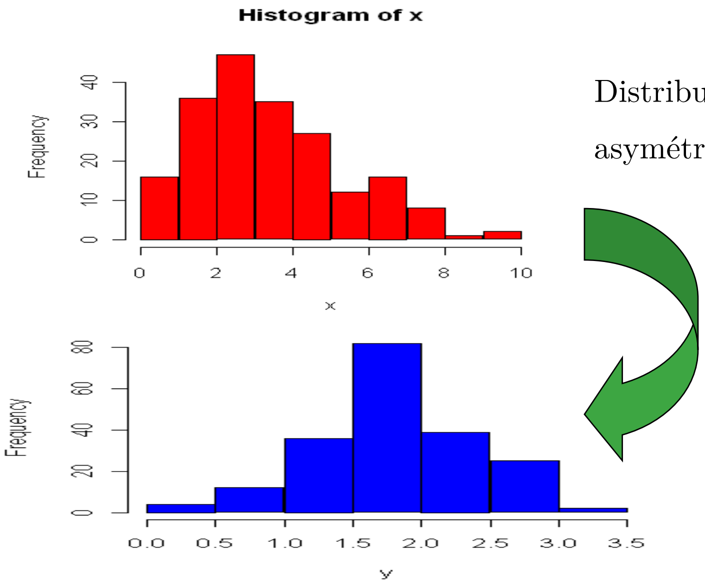
dénombrement : attaque de ravageurs, défauts de production

⟶ Loi de Poisson (variance et moyennes égales)

$$Y = \sqrt{X}$$

$$Y = \sqrt{X + \tfrac{1}{2}} \quad \text{également envisageable}$$

Si phénomène d'agrégation ⟶ log



Distribution asymétrique

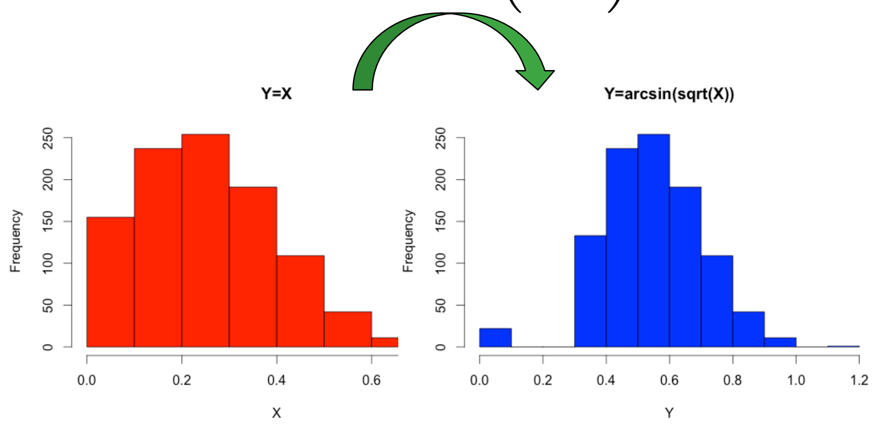
$$Y = \sqrt{X}$$

73

Etude d'une proportion : proportion d'attaque, d'éléments défectueux

$\longrightarrow$ Loi Binomiale

$$Y = \arcsin\left(\sqrt{X}\right)$$



Variable Log normale :
phénomène exponentiel, multiplicatif

$$X_{ij} = \alpha_i \times \mu \times \varepsilon_{ij}$$

$$\Downarrow$$

$$\ln X_{ij} = \ln \alpha_i + \ln \mu + \ln \varepsilon_{ij}$$

TESTS NON PARAMÉTRIQUES

Par opposition avec les méthodes “gaussiennes”

⇩

Aucune hypothèse sur les distributions *(Distribution free)*

Tests souvent fondés sur des transformations en rangs

⇩

Perte d’information et donc de puissance
(Risque β plus important)

Comparaison de 2 échantillons appariés

Test des signes :

On dénombre les différences positives (on négatives)
Sous H_0, l’espérance du nombre de différences + est égale à celle du nombre des différences -
P(+) = P(-) = 0,5
On effectue un test binomial

Test de Wilcoxon (signes et rang) :

Test plus puissant que le test des signes
Rangement des différences par ordre croissant de leur valeur absolue. Sous H_0, l’espérance du rang des différences + est égale à celle du nombre des différences -
$E(Y_+)=E(Y_-)=n(n+1)/4$

Exemple de test des signes : Hauteurs d'arbres par deux méthodes

Arbres debout	20,4	25,4	25,6	25,6	**26,6**	**28,6**	28,7	29	29,8	30,5	30,9	31,1
Arbres abattus	21,7	26,3	26,8	28,1	**26,2**	**27,3**	29,5	32	30,9	32,3	32,3	31,7

Sous $\mathcal{H}_0$, le nombre X de différences positives (ou négatives) suit la loi binomiale $\mathcal{B}(n; 0,5)$.

Le test des signes se ramène à un test binomial dont la probabilité du succès est 0,5.

Ici : X est égale à 2. A l'aide de R, on obtient :
\> binom.test(2,12, p = 0.5,alternative = "less")
Exact binomial test
data: 2 and 12
number of successes = 2, number of trials = 12,
p-value = 0.019

Avec le test de Wilcoxon
\> wilcox.test(x, y,paired=TRUE, alternative = "less")
data: x and y
p-value = 0.009

Comparaison de 2 échantillons indépendants
Test de Mann-Whitney
(appelé également test de Mann-Whitney-Wilcoxon)

Les notes moyennes attribuées par un jury d'expert sont les suivantes :
Vin 1 : 22 31 14 19 24 28 27 28
Vin 2 : 25 13 20 11 23 16 21 18 17 26

On réordonne les 18 observations par ordre croissant :
11 13 **14** 16 17 18 **19** 20 21 **22** 23 **24** 25 26 **27** **28** **28** **31**

On considère W égale au nombre de couples (x_i, y_j) où $x_i > y_j$.
Il y a $8 \times 10 = 80$ couples donc $\mathbb{E}(W) = \dfrac{8 \times 10}{2} = 40$

Sur ces échantillons, la réalisation de W est $w = 2 + 5 + 7 + 8 + 10 + 10 + 10 + 10 = 62$.

A l'aide du logiciel R, on obtient :
> wilcox.test(x, y, alternative = "greater")
W = 62, p-value = 0.028

Et si les conditions d'application pour une ANOVA à 1 facteur ne sont pas remplies ....

Test de Kruskal-Wallis (basé sur les rangs) :

$$K_{obs} = \frac{12}{N(N+1)} \sum_{i=1}^{p} n_i \left(R_i - \frac{N+1}{2} \right)^2 \sim \chi^2(p-1)$$

où R_i correspond à la somme des rangs des observations associées à la modalité i

Test post hoc associé :
Test de Mann-Whitney-Wilcoxon pour chaque couple