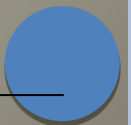

Chapitre 4 ANALYSE DE LA VARIANCE (ANOVA)



L'Analyse de Variance à un facteur (ANOVA 1)

- A. Introduction**
- B. Formulation du modèle**
- C. Conditions d'application de l'ANOVA**
- D. Mesure de la décomposition de la SCE**
- E. Test de Tukey**



Introduction

21 candidats, 3 examinateurs (resp. 6,8 et 7 étudiants)

Examineur	A	B	C
Notes	10,11,11 12,13,15	8,11,11,13 14,15,16,16	10,13,14,14 15,16,16
Effectif	6	8	7
Moyenne	12	13	14

"effet d'examineur"?

Forêt 1	Forêt 2	Forêt 3
23,3	18,9	22,5
24,4	21,1	22,9
24,6	21,1	23,7
24,9	22,1	24,0
25,0	22,5	24,0
26,2	23,5	24,5

"effet de plantation"?



Introduction

Objectif = Quand utiliser
l'ANOVA?

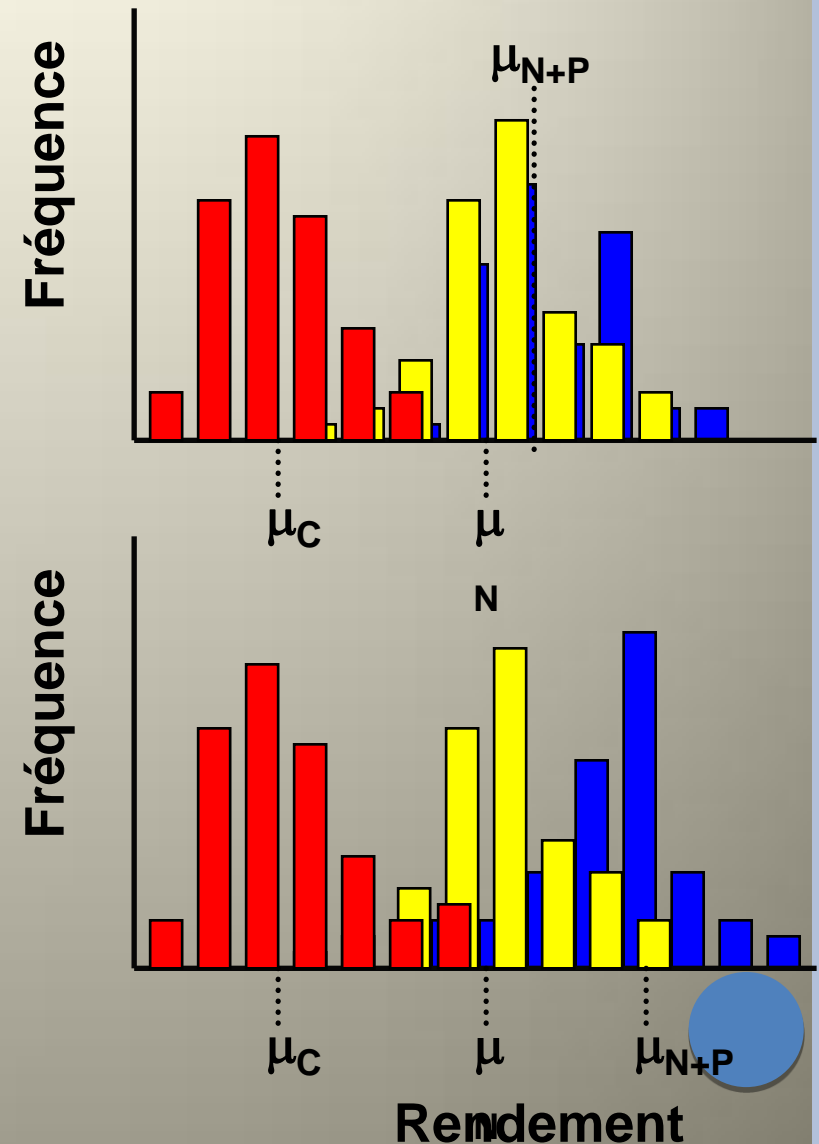
- Pour tester l'effet d'une variable explicative dite **facteur** contrôlé (chaque facteur a k niveaux ou modalités) sur les moyennes d'une variable quantitative Y
- l'ANOVA teste si toutes les moyennes sont égales

ANOVA 1

Possibilités et limites

Permet de tester si toutes les moyennes sont égales (au niveau α)...

...mais si on rejette H_0 , l'ANOVA ne dit pas lesquelles



		Statistical Decision	
		Reject Null	Retain Null
True Population Status	Null is True	Type I Error α	Correct Decision $1-\alpha$
	Null is False	Correct Decision $1-\beta$	Type II Error β

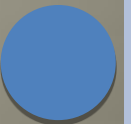


Types d'ANOVA

Fixe : les traitements sont déterminés (manipulés) par le chercheur

Aléatoires : les modalités sont choisies au hasard dans une population de modalités: on peut estimer l'effet du facteur pour d'autres modalités non étudiées

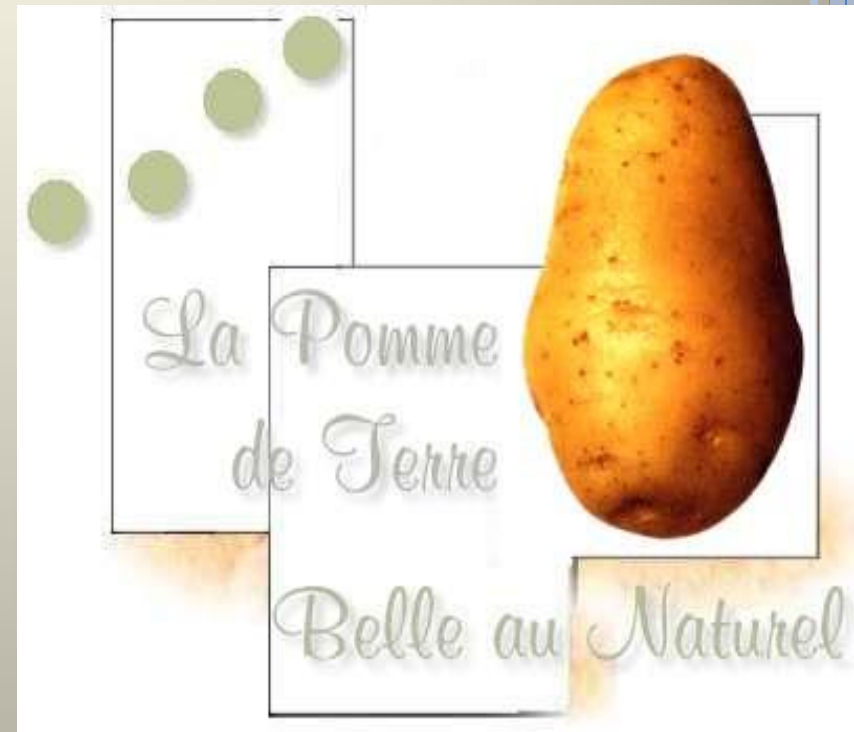
Données identiques, modèles différents, calculs identiques mais seulement pour l'ANOVA à un critère de classification!



ANOVA fixe : rendement agricole

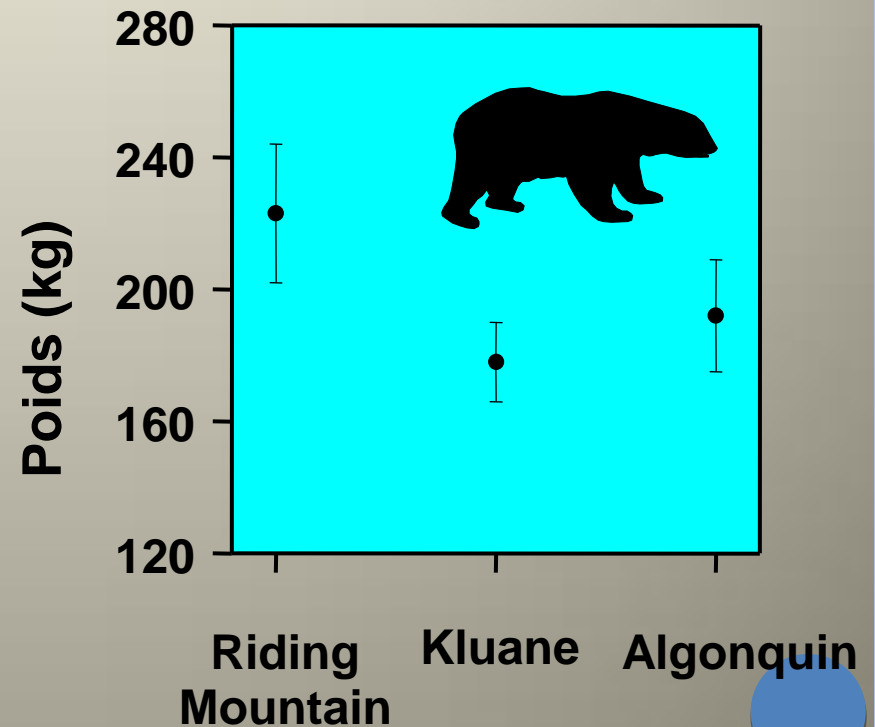
	sable	argile	terreau
	21	16	23
	20	18	31
	16	11	24

n_i	3	3	3	9	= N
$T_{i.}$	57	45	78	180	= T
$\bar{y}_{i.}$	19	15	26	20	



ANOVA aléatoire: poids de l'ours noir

- variable dépendante est le poids,
- facteur (X) = site, $p=3$
- Question = effet site, au-delà des sites étudiés



Un seul facteur F

k niveaux

k échantillons de tailles respectives n_1, \dots, n_k

Effectif total

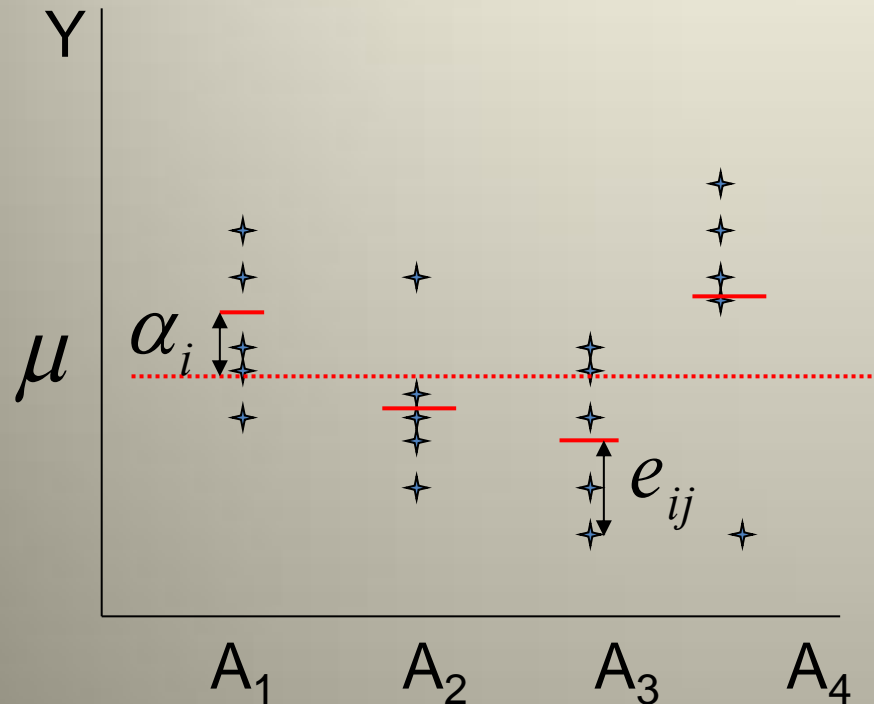
$$n = \sum_{i=1}^k n_i$$

$n_i = n_j \forall i,j \rightarrow$
expérience équilibrée

À chaque expérience, on mesure la valeur de la variable Y.

Données

Niveau (population)	Nb. obs.	Valeurs de Y
1	n_1	$y_{11}, y_{12}, \dots, y_{1n_1}$
2	n_2	$y_{21}, y_{22}, \dots, y_{2n_2}$
\vdots	\vdots
k	n_k	$y_{k1}, y_{k2}, \dots, y_{kn_k}$



Les p moyennes sont-elles identiques?

Les modalités de A influencent-elles Y ?

$$Y_{ij} = \mu + \alpha_i + e_{ij}$$

μ moyenne de Y

α_i effet de la $i^{\text{ème}}$ modalité (constante). $H_0 : \alpha_i = 0$

e_{ij} erreur aléatoire

- Modèle:

$$y_{ij} = \mu_i + \varepsilon_{ij}, \quad i = 1, \dots, I \text{ et } j = 1, \dots, J$$

- Test de comparaison des moyennes :

Hypothèse nulle (H0) : $\mu_1 = \mu_2 = \dots = \mu_I$

Contre (H1) : Les μ_i ne sont pas tous égaux.

=> Utilisation de **l'analyse de la variance à un facteur.**



Conditions d'application de l'ANOVA

les k échantillons sont indépendants et de loi Normale.

Les y_{ij} sont des réalisations de la v.a. $Y_{ij} \sim \mathcal{N}(m_i, \sigma^2)$ et $Y_{ij}, Y_{i'j'}$ indépendantes pour $i \neq i'$ ou $j \neq j'$.

Test de shapiro-wilkson (sur les résidus)

Autrement dit, pour chaque i , $(y_{ij})_{j \leq n_i}, \dots, y_{in_i}$ est un échantillon standard.

L'écart-type (théorique) est le même pour tous les niveaux. La moyenne (théorique) peut varier avec le niveau.

Homogénéité des variances ou homoscédasticité.

Test de Bartlett



1. Indépendance :

- Pas de test statistique simple pour étudier l'indépendance.
- Les conditions de l'expérience choisie nous déterminent si nous sommes dans le cas de l'indépendance.

2. Normalité :

Test de **Shapiro-Wilk** sur l'ensemble des résidus

(H0) : les résidus suivent une loi normale

(H1) : les résidus ne suivent pas une loi normale

- Statistique de test :

$$W = \frac{\left(\sum_{i=1}^{\lfloor n/2 \rfloor} a_i (x_{(n-i+1)} - x_{(i)}) \right)^2}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

$x_{(i)}$ correspond à la série des données triées, et a_i sont des constantes fournies par des tables spécifiques.

- Décision : On rejette H0 si $W < W_{crit}$.

Les valeurs seuils W_{crit} pour différents risques α et effectifs n sont lues dans la table de Shapiro-Wilk.



3. Homogénéité :

Test de Bartlett :

- Comparaison multiple de variances

$$(H_0) : \sigma^2_1 = \sigma^2_2 = \dots = \sigma^2_I$$

(H1) : les σ^2_I ne sont pas toutes égales

- Statistique de test :
$$B_{obs} = \frac{1}{C} \left[(n-1) \ln(s^2_R) - \sum_{i=1}^I (n_i - 1) \ln(s^2_{c,i}) \right]$$

avec
$$C = 1 + \frac{1}{3(I-1)} \left(\left(\sum_{i=1}^I \frac{1}{n_i - 1} \right) - \frac{1}{n-1} \right)$$

et B_{obs} suit une loi du Khi-Deux à $I-1$ ddl.

- Décision : Si $B_{obs} < c \rightarrow (H_0)$ vraie

Exemple : forêt

Application à l'exemple :

$$\overline{y_1} = 24,75$$

$$\overline{y_2} = 21,53$$

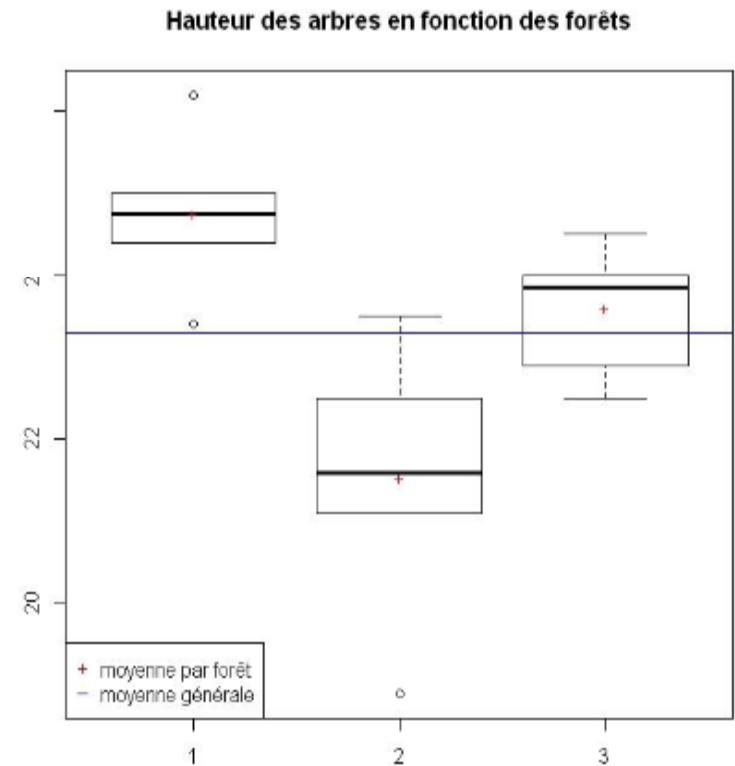
$$\overline{y_3} = 23,6$$

$$s_1 = 0,83$$

$$s_2 = 2,49$$

$$s_3 = 0,57$$

Nombre d'observations : $n = I * J = 6 * 3 = 18$



- **Normalité (Shapiro)** : nombre d'observations trop faible pour tester sur chaque forêt donc on va tester sur tout l'échantillon.

Test de Shapiro-Wilk	
W=0.9748	P-value=0.882

p-value = 0.882 > 0.05 donc on accepte H_0 => normalité.

- **Homogénéité (Bartlett)** : nombre d'observations trop faible pour tester sur chaque forêt donc on va tester sur tout l'échantillon.

Test de Bartlett		
B=2.8279	Df=2	P-value= 0.2432

p-value = 0.2432 donc on accepte H_0 => homogénéité des variances



ANOVA 1

Propriété fondamentale

Dans une ANOVA, la variance totale est répartie en deux composantes:

intergroupe: variance des moyennes des différents groupes (modalités)

intragroupe (erreur): variance des observations autour de la moyenne du groupe



Propriété fondamentale

Variation due au facteur :

dispersion des moyennes autour de la moyenne générale.



$$SC_{tot} = SC_F + SC_R$$



Variation totale :

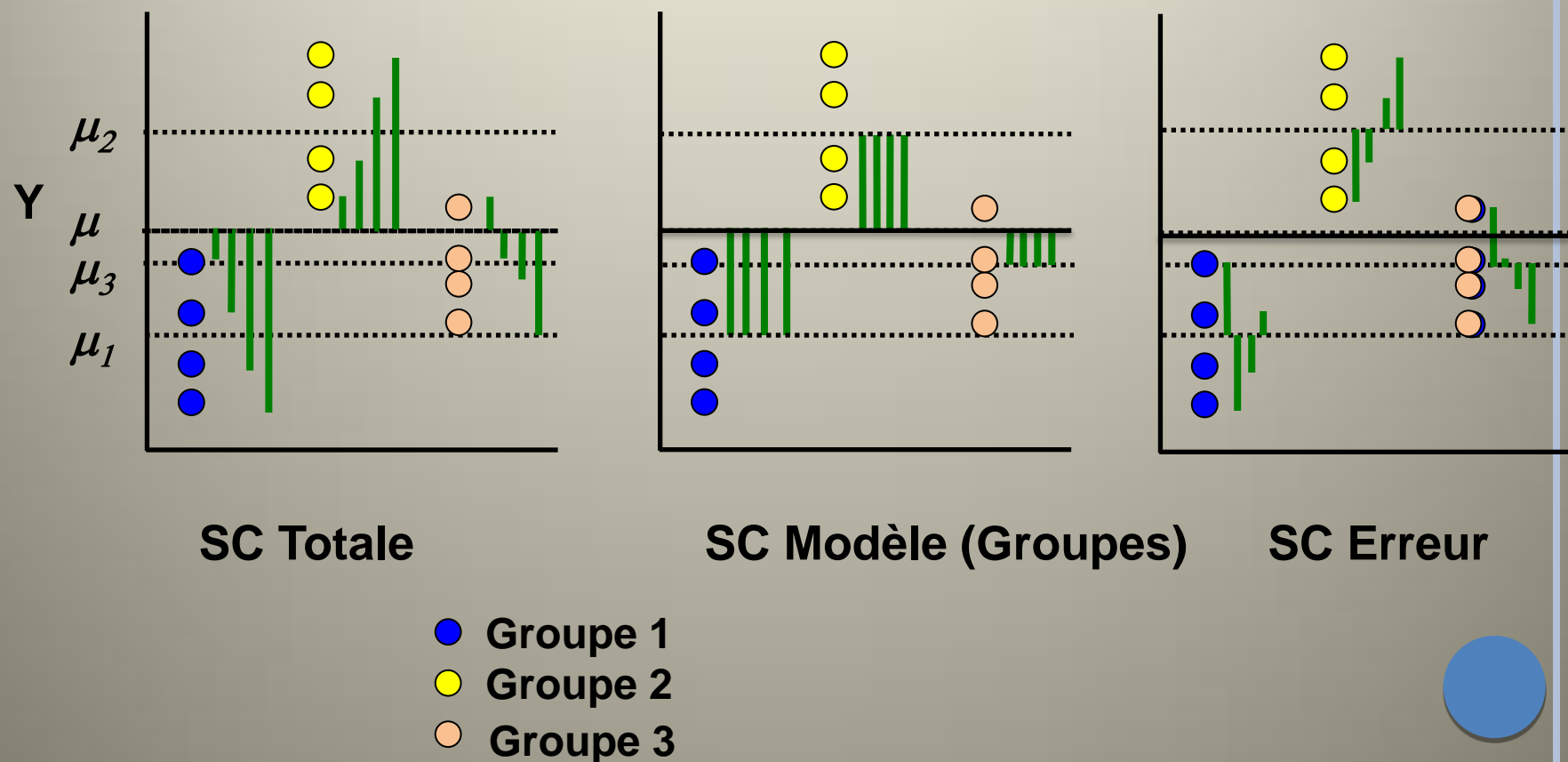
dispersion des données autour de la moyenne générale.



Variation résiduelle :

dispersion des données à l'intérieur de chaque échantillon autour de sa moyenne.

Répartition de la somme des carrés totale



$$\text{SCE}_T = \text{SCE}_A + \text{SCE}_{R(=E)}$$


$$\sum_{i=1}^p \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{..})^2 = \sum_{i=1}^p \sum_{j=1}^{n_i} (\bar{Y}_{i.} - \bar{Y}_{..})^2 + \sum_{i=1}^p \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{i.})^2$$

$$= \text{SCE}_{\text{inter}} + \text{SCE}_{\text{intra}}$$

$$= \text{SCE}_B + \text{SCE}_W$$



Tableau d'ANOVA

Sources de variation	Somme des carrés	Degré de liberté (<i>ddl</i>)	Carré moyen (CM)	<i>F</i>
Totale	$\sum_{i=1}^k \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y})^2$	$n - 1$	SCE_T / dd <i>l</i>	$\frac{CM_A}{CM_R}$ 
Facteu r	$\sum_{i=1}^k n_i (\bar{Y}_i - \bar{Y})^2$	$p - 1$	SCE_A / d <i>dl</i>	
Résidu s	$\sum_{i=1}^k \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_i)^2$	$n - p$	SCE_R / d <i>dl</i>	

TEST DE FISHER:

$$(H_0) : \mu_1 = \mu_2 = \dots = \mu_I$$

(H1) : Les μ_i ne sont pas tous égaux.

Si les 3 conditions (Indépendance, Normalité et Homogénéité) sont vérifiées et si (H0) est vraie,

Alors :

$$F_{obs} = \frac{CM_F}{CM_R} \sim F_{I-1, n-I}$$

Décision : Pour un seuil donné α (5% en général) les tables de Fisher nous fournissent une valeur critique c telle que :

$$P_{H_0}(F_{I-1, n-1} < c) = 1 - \alpha$$

Alors:

si $F_{obs} < c \rightarrow H_0$ est vraie

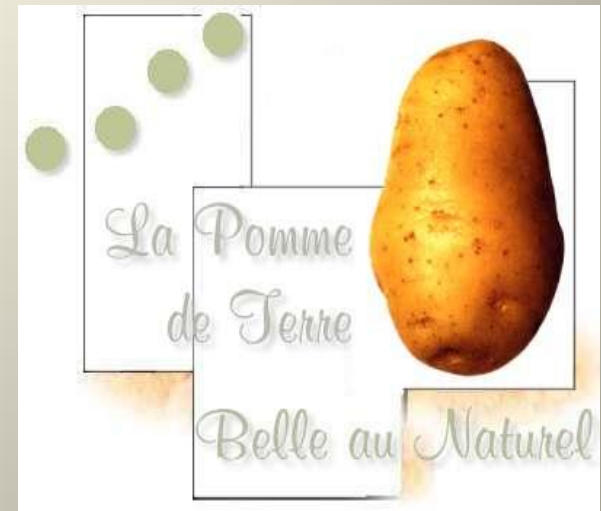
si $F_{obs} \geq c \rightarrow H_1$ est vraie

Exemple

	sable	argile	terreau	
	21	16	23	
	20	18	31	
	16	11	24	
$T_{i.}$	57	45	78	180

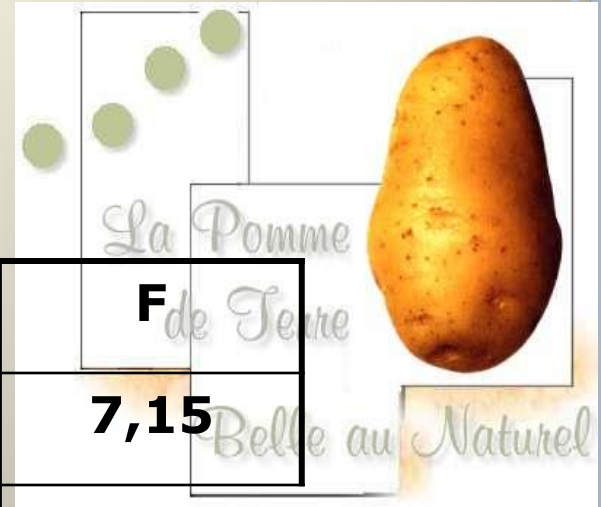
$$SCE_T = 21^2 + 20^2 + \dots - \frac{180^2}{9} = 264$$

$$SCE_A = \frac{57^2}{3} + \frac{45^2}{3} + \dots - \frac{180^2}{9} = 186$$



Exemple

SV	SCE	ddl	CM	F
A	186	2	93	7,15
R	78	6	13	
T	264	8		



$$F_{6,\alpha=0,05}^2 = 5,14$$

→ Conclusion ?



Comparaison multiple

But : classer les traitements par groupes qui sont significativement différents.

- Test de Tukey : test de la différence franchement significative (HSD= honestly significant difference)

- S'applique sur un facteur si :
 - Les 3 conditions fondamentales sont vérifiées,
 - Le facteur est à effet fixe, avec au moins 3 modalités,
 - Le facteur a un effet significatif sur la réponse.

Méthode :

- Pour chaque paire i et l de groupes, on calcule un IC de niveau $(1-\alpha)\%$ de la différence $(\mu_i - \mu_l)$.
- Si zéro appartient à l'IC, les moyennes ne sont pas jugées significativement différentes au niveau α .

Exemple :

	Diff	Lower	Upper	P-value
2-1	-3.22	-4.92	-1.51	0.0005
3-1	-1.15	-2.86	0.56	0.22
3-2	2.07	0.36	3.77	0.02

0 est dans l'intervalle de confiance de 3-1 → les hauteurs moyennes dans les forêts 1 et 3 ne sont pas significativement différentes.