Analyse de la variance

M2 Statistiques et Econométrie

Fanny MEYER

Morgane CADRAN

Margaux GAILLARD

Analyse de la variance
Plan du cours

I. Introduction

II. Analyse de la variance à un facteur

III. Analyse de la variance à deux facteurs

IV. Analyse de la covariance

V. Problèmes spécifiques

I. Introduction

- Cadre:

- Endogène : Variable quantitative

- Exogène(s) : Variable(s) qualitative(s) appelée(s) facteur(s)

- Objectifs:

- Comparer les moyennes de l'endogène pour chaque modalité des facteurs

- Etudier l'effet de ces facteurs sur la variable réponse

II. Analyse de la variance à un facteur

1) Modèle
2) Vérification des conditions
3) Anova
4) Comparaisons multiples

Analyse de la variance à un facteur

Présentation des données :

- Plantation d'arbres dans 3 forêts
- Comparaison de la hauteur des arbres

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

Analyse de la variance à un facteur

Présentation des données :

- **Les forêts** : Variable qualitative contenant trois modalités, appelée facteur (à effets fixes).
- **Hauteur des arbres** : Réponse, notée Y.

L'analyse de variance à un facteur teste l'effet d'un facteur contrôlé A ayant p modalités sur les moyennes d'une variable quantitative Y.

Analyse de la variance à un facteur

Les échantillons sont de même taille => expérience équilibrée.

- Moyenne de chaque échantillon :

$$\overline{y_i} = \frac{1}{J}\sum_{j=1}^{J} y_{ij}, \quad i = 1,\ldots, I.$$

- Variance de chaque échantillon :

$$s^2{}_i(y) = \frac{1}{J}\sum_{j=1}^{J} (y_{ij} - \overline{y}_i)^2, \quad i = 1,\ldots, I.$$

Analyse de la variance à un facteur

Application à l'exemple :

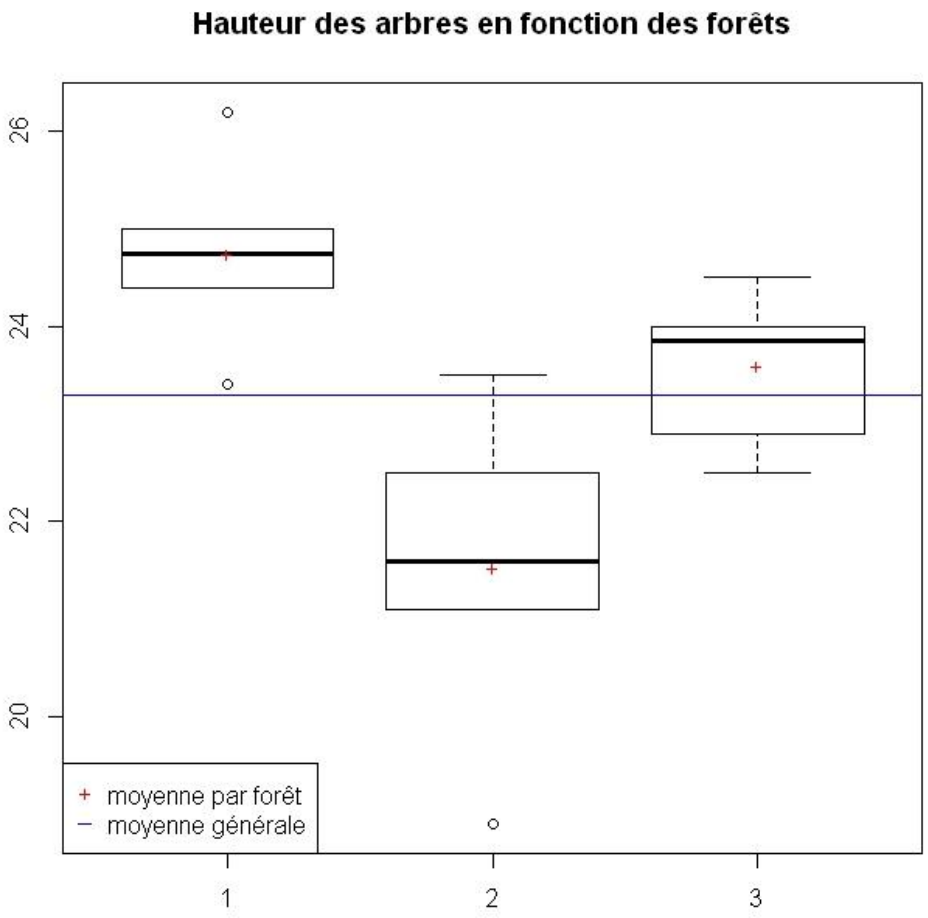
$\overline{y_1} = 24{,}75$ $\overline{y_2} = 21{,}53$ $\overline{y_3} = 23{,}6$

$s_1 = 0{,}83$ $s_2 = 2{,}49$ $s_3 = 0{,}57$

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

Analyse de la variance à un facteur

Application à l'exemple :



Analyse de la variance à un facteur

- Modèle:

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

- Test de comparaison des moyennes :

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

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

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

II. Analyse de la variance à un facteur

1) Modèle
2) Vérification des conditions
3) Anova
4) Comparaisons multiples

Analyse de la variance à un facteur

Les trois conditions pour l'ANOVA:

1. Les p échantillons comparés sont **indépendants**.
2. La variable quantitative étudiée suit une **loi normale** dans les p populations comparées.
3. Les p populations comparées ont même variance : **Homogénéité des variances** ou homoscédasticité.

Analyse de la variance à un facteur

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.

Exemple => Les forêts sont indépendantes.

Analyse de la variance à un facteur

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 :

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

$x_{(i)}$ 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.

Analyse de la variance à un facteur

3. Homogénéité :

Test de **Bartlett** :

- Comparaison multiple de variances

  (H0) : $\sigma^2_1 = \sigma^2_2 = \ldots = {}^2\sigma_I$

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

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

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 → (H0) vraie

Analyse de la variance à un facteur

Retour à l'exemple :

- **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 H0 => 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 H0 => homogénéité des variances

II. Analyse de la variance à un facteur

1) Modèle
2) Vérification des conditions
3) Anova
4) Comparaisons multiples

Tableau ANOVA : Propriétés fondamentales

- La moyenne de toutes les observations est la moyenne des moyennes de chaque échantillon:

$$\bar{y} = \frac{1}{n}\sum_{j=1}^{J}\sum_{i=1}^{I} y_{ij}$$

Exemple : $\bar{y}$ = (24,75+21,53+23,60)/3 = 23,29

- La variance de toutes les observations est la somme de la variance des moyennes et de la moyenne des variances:

$$s^2(y) = \frac{1}{n}\sum_{i}\sum_{j}(y_{ij} - \bar{y})^2 \quad = \frac{1}{I}\sum_{i=1}^{I}(\bar{y_i} - \bar{y})^2 + \frac{1}{I}\sum_{i=1}^{I} s_i{}^2(y) \qquad \textbf{(1)}$$

Exemple : $s^2(y)$ = 3,06

Analyse de la variance à un facteur

- Variance des moyennes = $\frac{1}{I}\sum_{i=1}^{I}(\overline{y_i} - \overline{y})^2$

  = $\frac{1}{3}((24{,}75-23{,}29)^2+(21{,}53-23{,}29)^2+(23{,}60-23{,}29)^2)$

  = 1,77

- Moyenne des variances = $\frac{1}{I}\sum_{i=1}^{I}s^2{}_i(y)$ = $\frac{1}{3}(0{,}83+2{,}49+0{,}57)$ = 1,29

→ Somme = 3,06 → équation précédente vérifiée

On multiplie **(1)** par n : $\sum_{i=1}^{I}\sum_{j=1}^{J}(y_{ij}-\overline{y})^2 = J\sum_{i=1}^{I}(\overline{y_i}-\overline{y})^2+\sum_{i=1}^{I}(\sum_{j=1}^{J}(y_{ij}-\overline{y_i})^2)$

Cette relation s'écrit : $SC_{tot} = SC_F + SC_R$

Analyse de la variance à un facteur

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.

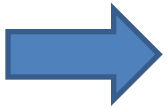
Analyse de la variance à un facteur

Retour à l'exemple : (calculs avec R)

Sc_tot = 51.31

SC_F = 31.88

SC_R = 19.43



On retrouve bien la relation précédente.

Analyse de la variance à un facteur

(H0) : $\mu_1 = \mu_2 = ... = \mu_I$
(H1) : Les μ_i ne sont pas tous égaux.

- Si **(H0)** est vraie alors la variation due au facteur SC_F doit être **petite** par rapport à la variation résiduelle SC_R .
- Par contre, si **(H1)** est vraie alors la variation due au facteur SC_F doit être **grande** par rapport à la quantité SC_R .

Pour comparer ces quantités, Fisher a considéré le rapport des carrés moyens.

Carré moyen associé au facteur : $$CM_F = \frac{SC_F}{I-1}$$

Carré moyen résiduel : $$CM_R = \frac{SC_R}{n-1}$$

=> estimateur sans biais de la variance des erreurs qu'on appelle variation résiduelle notée aussi Sr^2.

Analyse de la variance à un facteur

TEST DE FISHER:

(H0) : $\mu_1 = \mu_2 = ... = \mu_I$
(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 :

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

Alors: si $F_{obs} < c$ → H0 est vraie
si $F_{obs} \geq c$ → H1 est vraie

Analyse de la variance à un facteur

Tableau de l'ANOVA fourni par le test de Fisher:

Variation	SC	ddl	CM	Fobs	Fc
Due au facteur	SC_F	l-1	CM_F	$\frac{CM_F}{CM_R}$	c
Résiduelle	SC_R	n-l	CM_R		
Totale	SC_{tot}	n-1			

Analyse de la variance à un facteur

Tableau de l'ANOVA :

Application à notre exemple :

Variation	SC	ddl	CM	Fobs	Fc
Due au facteur	31.88	2	15.94	12.31	0.0007
Résiduelle	19,43	15	1.29		
Totale	51.31	17			

p-value < 0.05 donc les hauteurs moyennes sont significativement différentes dans chaque forêt.

II. Analyse de la variance à un facteur

1) Modèle
2) Vérification des conditions
3) Anova
4) Comparaisons multiples

Analyse de la variance à un facteur

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

- Test de Tukey : test de la différence franchement significative (HSD= honestly significative 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.

Analyse de la variance à un facteur

Méthode :

- Pour chaque paire i et l de groupes, on calcule un IC de niveau (1-α)% 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

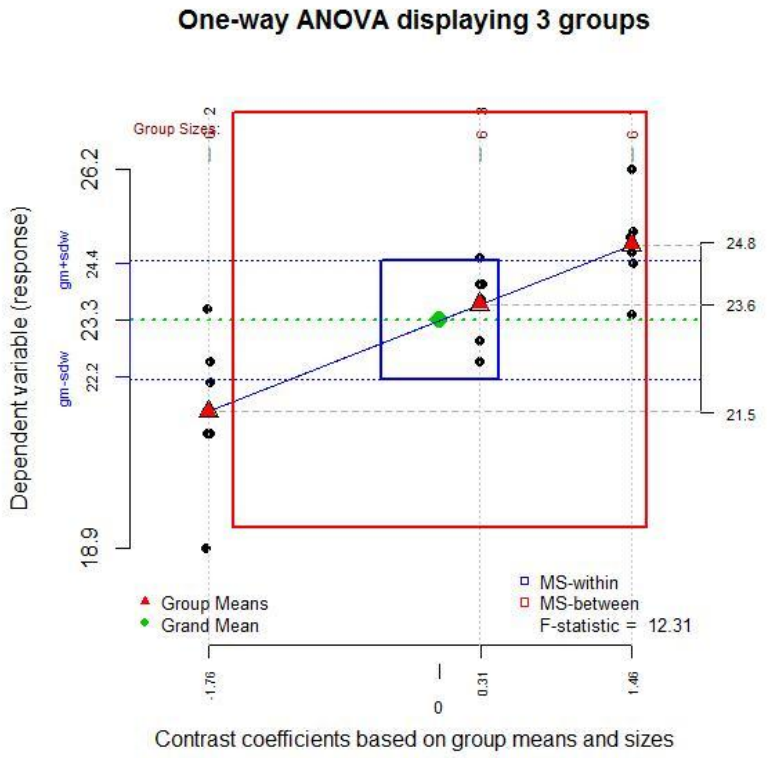
0 est dans l'intervalle de confiance de 3-1 → les hauteurs moyennes dans les forêts 1 et 3 ne sont pas significativement différentes.

Analyse de la variance à un facteur

Représentation graphique de l'ANOVA :

→ *package « granova »*

> granova.1w(hauteur,foret)



III. Analyse de la variance à deux facteurs

1) Modèle

2) Vérification des conditions

3) Anova

4) Comparaisons multiples

5) Facteurs sans répétitions

Analyse de la variance à deux facteurs

- Variables étudiées :

- facteur à I modalités

- facteur à J modalités

- variable quantitative Y

- Dans la population correspondant à la modalité d'ordre i du premier facteur et à la modalité d'ordre j du deuxième facteur :

$$Y \sim N(\mu_{ij}, \sigma^2)$$ pour i=1,...,I et j=1,...,J.

Analyse de la variance à deux facteurs

- Echantillons indépendants de même taille K de la variable Y dans chacune des IJ populations, soit au total un n-échantillon avec **n = IJK**.

- Modèle : $$Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}$$

pour tout i=1,...,I ; j=1,...,J ; k=1,...,K sous contraintes:

$$\sum_{i=1}^{I} \alpha_i = \sum_{j=1}^{J} \beta_j = \sum_{i=1}^{I} (\alpha\beta)_{ij_0} = \sum_{j=1}^{J} (\alpha\beta)_{i_0 j} = 0 \text{ pour } i_0 = 1,\ldots,I \text{ et } j_0 = 1,\ldots,J.$$

- Hypothèse : $$\varepsilon_{ijk} \sim N(0, \sigma^2)$$

Analyse de la variance à deux facteurs

- Autre écriture du modèle :

$$Y_{ijk} = \mu_{ij} + e_{ijk}$$

pour i=1...,I ; j=1,...,J ; k=1,...,K ;

Avec e_{ijk} les erreurs de mesure (inconnues).

Analyse de la variance à deux facteurs

Présentation des données de l'exemple :

- Expérience : des secrétaires tapent un texte pendant 5 minutes sur différentes machines à écrire. L'expérience est répétée le lendemain.
- Premier facteur à 4 modalités : modèles de machines à écrire
- Second facteur à 5 modalités : secrétaires professionnelles
- Variable quantitative : nombre moyen de mots tapés en une minute.

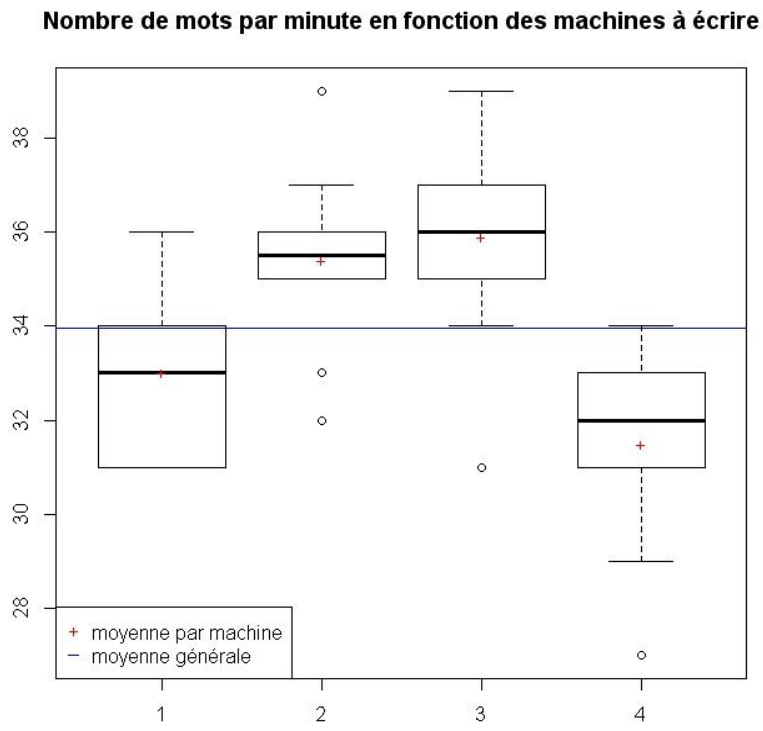
Analyse de la variance à deux facteurs

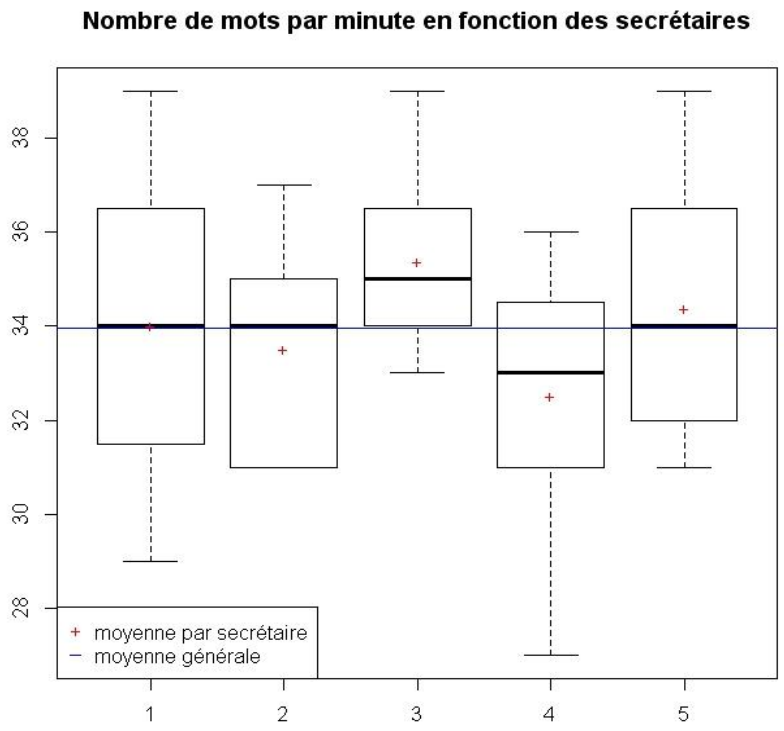
I=4, J=5, K=2 → échantillon de n=40 observations.

Machines à écrire	Secrétaires				
	1	2	3	4	5
1	33	31	34	34	31
	36	31	36	33	31
2	32	37	39	33	35
	35	35	36	36	36
3	37	35	34	31	37
	39	35	37	35	39
4	29	31	33	31	33
	31	33	34	27	33

Analyse de la variance à deux facteurs

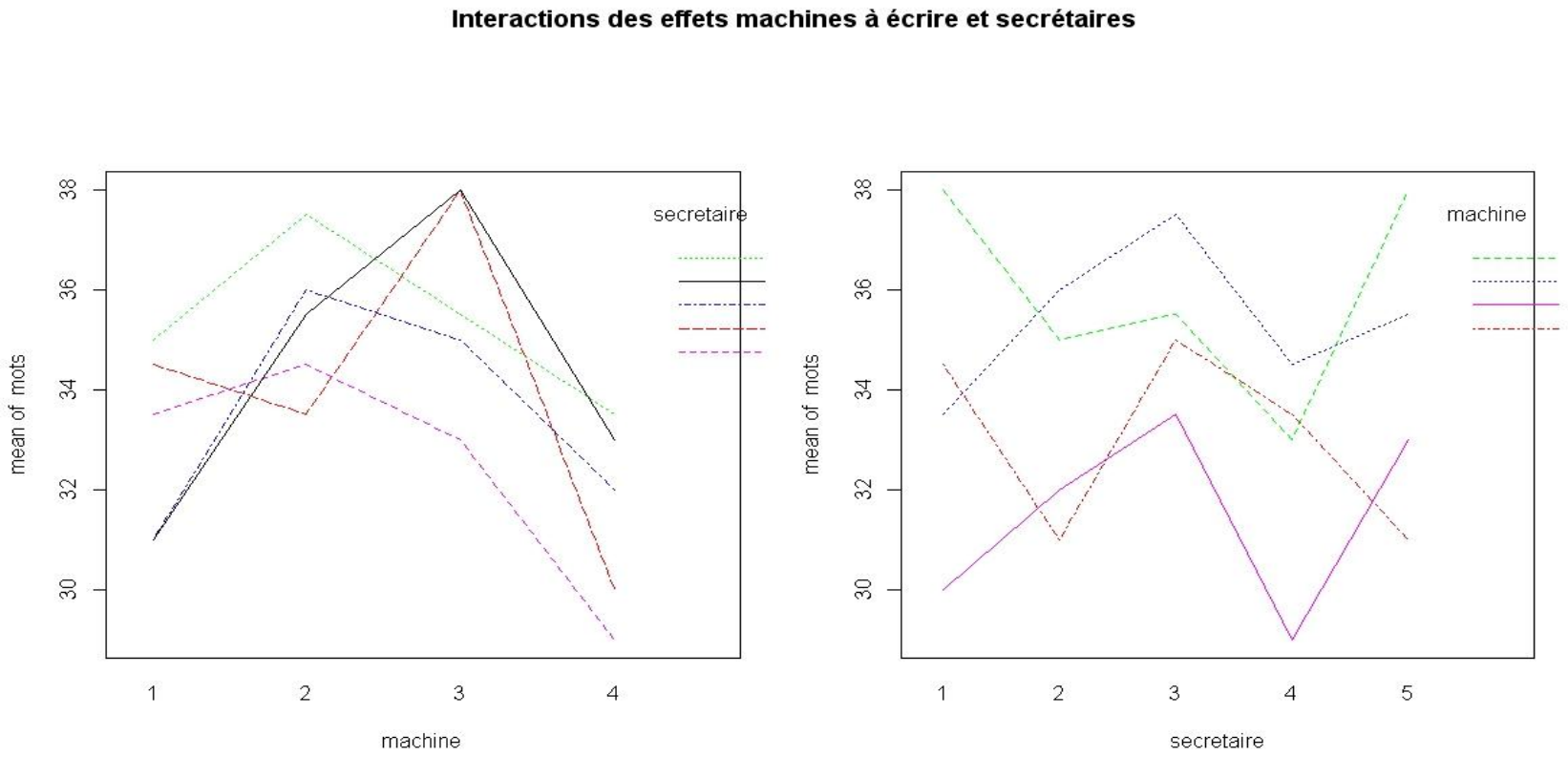
- But : analyser l'influence de la machine à écrire et de la secrétaire sur le nombre moyen de mots tapés en une minute.





Analyse de la variance à deux facteurs

Représentation graphique des interactions :



Le nombre moyen de mots tapés en une minute sur les machines diffère avec les secrétaires, et vice versa.

III. Analyse de la variance à deux facteurs

1) Modèle

2) Vérification des conditions

3) Anova

4) Comparaisons multiples

5) Facteurs sans répétitions

Analyse de la variance à deux facteurs

- Moyennes théoriques μ_{ij} estimées par les moyennes observées $\bar{y}_{ij\bullet}$ (« valeurs ajustées »).
- Résidus : $\hat{e}_{ijk} = y_{ijk} - \bar{y}_{ij\bullet}$

pour i = 1,...,I ; j = 1,..., J ; k = 1,...,K.

- Mêmes conditions à vérifier :

  1- indépendance des données

  2- normalité des résidus

  3- homogénéité des variances (homoscédasticité)

Analyse de la variance à deux facteurs

Exemple : 1) Indépendance

Les données sont indépendantes.

2) Normalité des résidus

> mod.int=lm(mots~machine*secretaire,data=texte)

> residus=residuals(mod.int)

> shapiro.test(residus)

Shapiro-Wilk normality test	
data: residus	
W = 0.9464	p-value = 0.05702

→ Ici on accepte H0 car p-value > 0,05 donc les résidus sont normaux.

Exemple : 3) Homoscédasticité

> bartlett.test(residus~machine,data=texte)

> bartlett.test(residus~secretaire,data=texte)

Bartlett test of homogeneity of variances		
data: residus by machine		
Bartlett's K-squared = 1.8254	df = 3	**p-value = 0.6094**
data: residus by secretaire		
Bartlett's K-squared = 8.9698	df = 4	**p-value = 0.06186**

Ici les p-value > 0,05 donc on accepte H0.

→ Ainsi les **variances** des machines et des secrétaires sont **homogènes**. Ces deux résultats ne nous garantissent pas l'égalité des 20 (4*5) variances théoriques mais sont de bons indicateurs pour l'homoscédasticité.

III. Analyse de la variance à deux facteurs

1) Modèle

2) Vérification des conditions

3) Anova

4) Comparaisons multiples

5) Facteurs sans répétitions

Analyse de la variance à deux facteurs

L'analyse de la variance à deux facteurs avec répétitions permet **trois tests de Fisher** :

- Effet du premier facteur

H0: les paramètres α_i sont tous nuls

H1: les paramètres α_i ne sont pas tous nuls

- Effet du second facteur

H0: les paramètres β_j sont tous nuls

H1: les paramètres β_j ne sont pas tous nuls

- Effet de l'interaction des deux facteurs

H0: les paramètres $(\alpha\beta)_{ij}$ sont tous nuls

H1: les paramètres $(\alpha\beta)_{ij}$ ne sont pas tous nuls

Analyse de la variance à deux facteurs

Les statistiques :

$$\bar{Y} = \frac{1}{n}\sum_{i,j,k} Y_{ijk},$$

$$\bar{Y}_{ij\bullet} = \frac{1}{K}\sum_{k} Y_{ijk}, \quad \bar{Y}_{i\bullet\bullet} = \frac{1}{JK}\sum_{j,k} Y_{ijk}, \quad \bar{Y}_{\bullet j\bullet} = \frac{1}{IK}\sum_{i,k} Y_{ijk}.$$

$$SC_T = \sum_{i,j,k} (Y_{ijk} - \bar{Y})^2, \quad SC_R = \sum_{i,j,k} (Y_{ijk} - \bar{Y}_{ij\bullet})^2,$$

$$SC_\alpha = \sum_{i,j,k} (\bar{Y}_{i\bullet\bullet} - \bar{Y})^2, \quad SC_\beta = \sum_{i,j,k} (\bar{Y}_{\bullet j\bullet} - \bar{Y})^2,$$

$$SC_{\alpha\beta} = \sum_{i,j,k} (\bar{Y}_{ij\bullet} - \bar{Y}_{i\bullet\bullet} - \bar{Y}_{\bullet j\bullet} + \bar{Y})^2 \quad .$$

Analyse de la variance à deux facteurs

Equation d'analyse de la variance :

$$SC_T = SC_R + SC_\alpha + SC_\beta + SC_{\alpha\beta}$$

- SC_R : dispersion due aux fluctuations individuelles
- SC_α : mesure l'effet du premier facteur
- SC_β : mesure l'effet du second facteur
- $SC_{\alpha\beta}$: mesure l'effet de l'interaction entre les deux facteurs

Analyse de la variance à deux facteurs

Propriété sur les lois des statistiques :

$$\frac{SC_\alpha / (I-1)}{SC_R / IJ(K-1)} = \frac{CM_\alpha}{CM_R} \sim F_{(I-1),IJ(K-1)} \text{ sous H0,}$$

$$\frac{SC_\beta / (J-1)}{SC_R / IJ(K-1)} = \frac{CM_\beta}{CM_R} \sim F_{(J-1),IJ(K-1)} \text{ sous H0,}$$

$$\frac{SC_{\alpha\beta} / (I-1)(J-1)}{SC_R / IJ(K-1)} = \frac{CM_{\alpha\beta}}{CM_R} \sim F_{(I-1)(J-1),IJ(K-1)} \text{ sous H0.}$$

Analyse de la variance à deux facteurs

Tableau de l'ANOVA :

Variation	SC	ddl	CM	F_obs	F_c
Due à F_α	SC_α	I-1	CM_α	CM_α / CM_R	C_α
Due à F_β	SC_β	J-1	CM_β	CM_β / CM_R	C_β
Due à $F_{\alpha\beta}$	$SC_{\alpha\beta}$	(I-1)(J-1)	$CM_{\alpha\beta}$	$CM_{\alpha\beta} / CM_R$	$C_{\alpha\beta}$
Résiduelle	SC_R	IJ(K-1)	CM_R		
Totale	SC_T	*n-1*			

→ Quand nous décidons H1, le facteur a un effet significatif sur la réponse.

Analyse de la variance à deux facteurs

Exemple :

> mod.int=lm(mots~machine*secretaire,data=texte)

> anova(mod.int)

Analysis of Variance Table					
Response: mots					
	Df	**Sum Sq**	**Mean Sq**	**F value**	**Pr(>F)**
Machine	3	128,10	42,7	16,42	**1,279e-05**
Secrétaire	4	36,15	9,04	3,48	**0,02603**
Interaction	12	77,65	6,47	2,49	**0,03450**
Résidus	20	52	2,6		

Ici les p-value < 0,05 donc on décide H1.

→ Ainsi les facteurs **machine** et **secrétaire** ainsi que leur **interaction** ont un effet significatif sur le nombre de mots tapés en une minute.

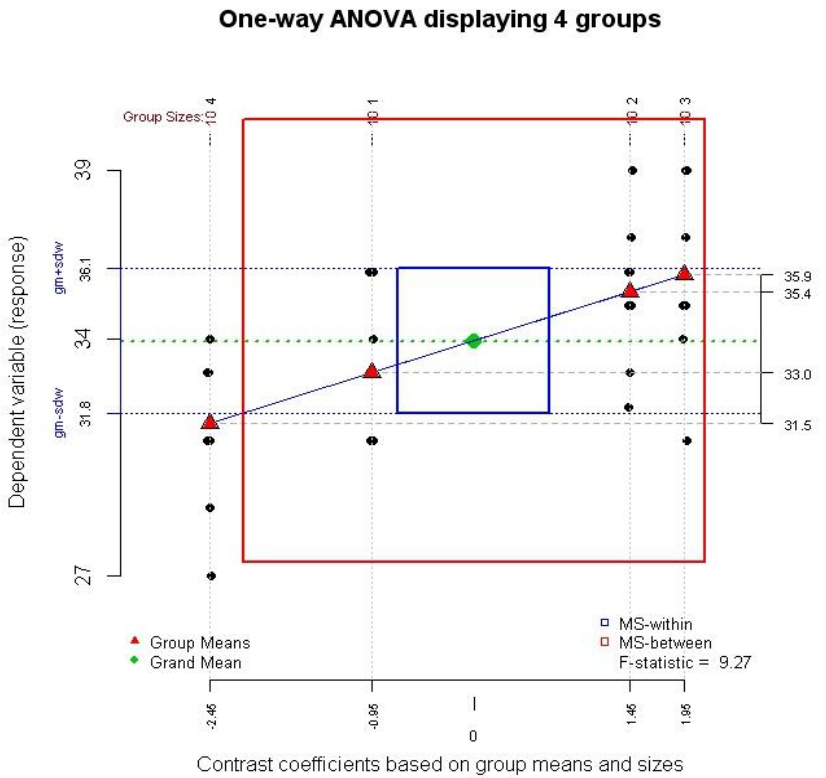
→ L'effet de la secrétaire sur le nombre de mots tapés diffère selon la machine à écrire, et vice versa.

Analyse de la variance à deux facteurs

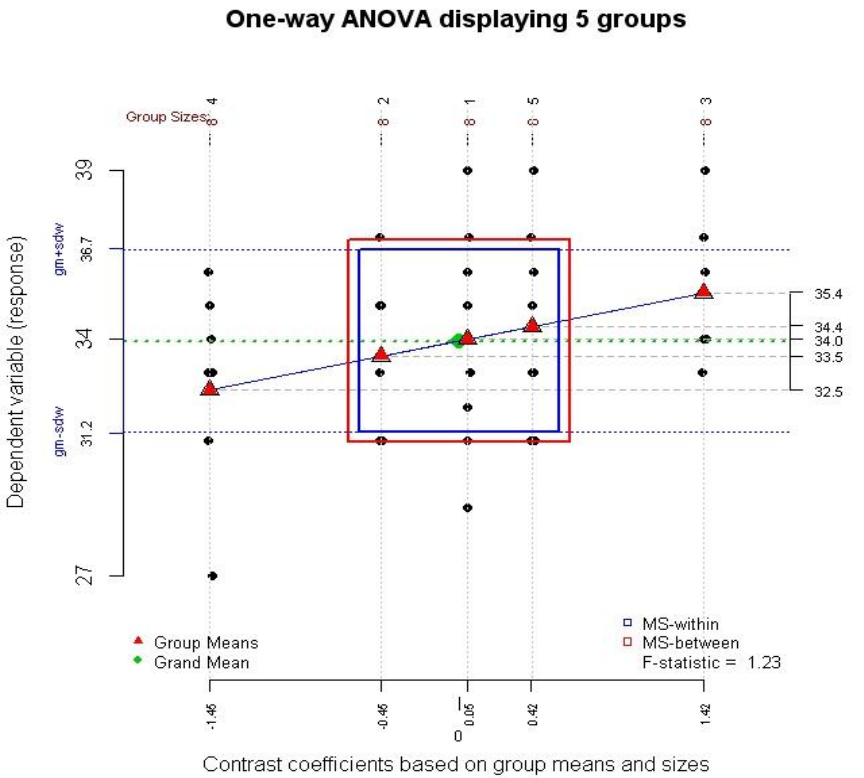
Représentation graphique de l'ANOVA :

→ *package « granova »*

> granova.1w(mots,machine)



> granova.1w(mots,secretaire)



III. Analyse de la variance à deux facteurs

1) Modèle

2) Vérification des conditions

3) Anova

4) Comparaisons multiples

5) Facteurs sans répétitions

Analyse de la variance à deux facteurs

Comparaisons multiples :

- Lorsque l'effet d'un facteur a été mis en évidence : le test de Tukey s'applique.
- Si le nombre d'observations le permet.
- L'objectif est de comparer les moyennes de la variable réponse dans les différents groupes.

Analyse de la variance à deux facteurs

Exemple : > mod = aov(mots~machine*secretaire, data=texte)

> TukeyHSD(mod, "machine", ordered = TRUE)

→Le nombre de mots tapés en une minute n'est en moyenne pas significativement différent pour les machines 1 et 4, ainsi que pour les machines 2 et 3.

Tukey multiple comparisons of means

95% family-wise confidence level

factor levels have been ordered

$ machine

	diff	lower	upper	p adj
1-4	1,5	-0,52	3,52	0,1936
2-4	3,9	1,88	5,92	0,0001
3-4	4,4	2,38	6,42	0,00003
2-1	2,4	0,38	4,42	0,0163
3-1	2,9	0,88	4,92	0,0034
3-2	0,5	-1,52	2,52	0,8984

Analyse de la variance à deux facteurs

Exemple : > TukeyHSD(mod, "secretaire", ordered = TRUE)

→Le nombre de mots tapés en une minute n'est en moyenne pas significativement différent pour les 5 secrétaires dans l'ensemble, sauf pour les secrétaires **3 et 4**.

Tukey multiple comparisons of means

95% family-wise confidence level

factor levels have been ordered

$ secretaire

	diff	lower	upper	p adj
2-4	1	-1,41	3,41	0,7287
1-4	1,5	-0,91	3,91	0,3691
5-4	1,88	-0,54	4,29	0,1779
3-4	2,88	0,46	5,29	0,0148
1-2	0,5	-1,91	2,91	0,9701
5-2	0,88	-1,54	3,29	0,8119
3-2	1,88	-0,54	4,29	0,1779
5-1	0,38	-2,04	2,79	0,9899
3-1	1,38	-1,04	3,79	0,4530
3-5	1	-1,41	3,41	0,7289

III. Analyse de la variance à deux facteurs

1) Modèle

2) Vérification des conditions

3) Anova

4) Comparaisons multiples

5) Facteurs sans répétitions

Analyse de la variance à deux facteurs

- Facteurs sans répétition : deux facteurs à, respectivement, I et J modalités et une seule observation pour chaque population, c'est à dire **K = 1**.
- Les résultats précédents ne sont plus valables.
- Nous devons supposer que **l'interaction entre les deux facteurs est nulle**.

- Modèle additif : $Y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}$

avec les contraintes $\sum_{i=1}^{I} \alpha_i = \sum_{j=1}^{J} \beta_j = 0.$

Analyse de la variance à deux facteurs

Equation d'analyse de la variance :

$$SC_T = SC_R + SC_\alpha + SC_\beta$$

- La somme des carrés correspondant à l'interaction est associée ici à la somme des carrés de la résiduelle.
- Les valeurs ajustées sont données par :

$$\hat{\mu}_{ij} = \bar{y}_{i\bullet} + \bar{y}_{\bullet j} - \bar{y}$$

- Les résidus sont donnés par :

$$\hat{e}_{ij} = y_{ij} - \bar{y}_{i\bullet} - \bar{y}_{\bullet j} + \bar{y}$$

pour i=1,...,I et j=1,...,J.

Analyse de la variance à deux facteurs

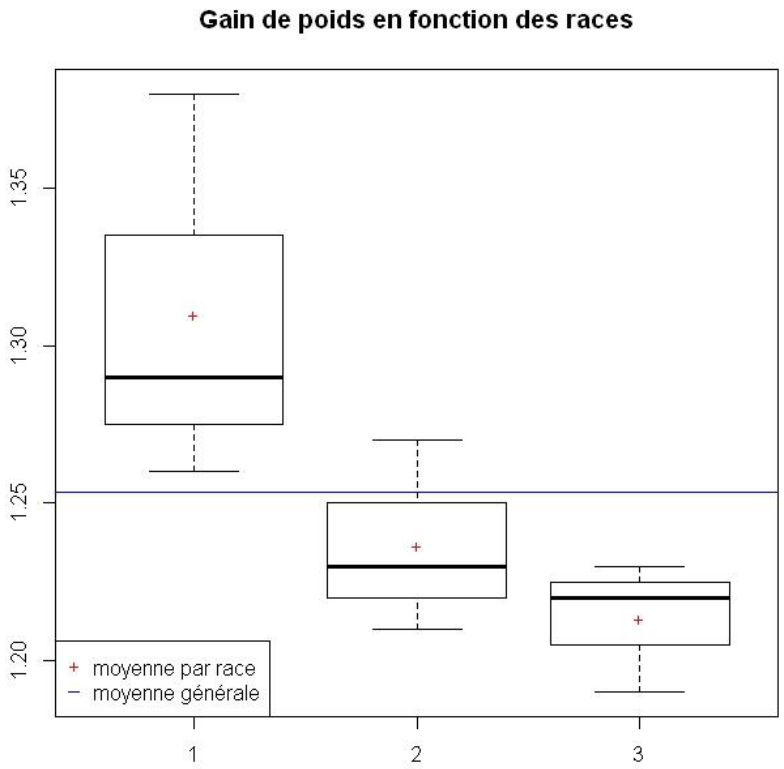
Exemple :

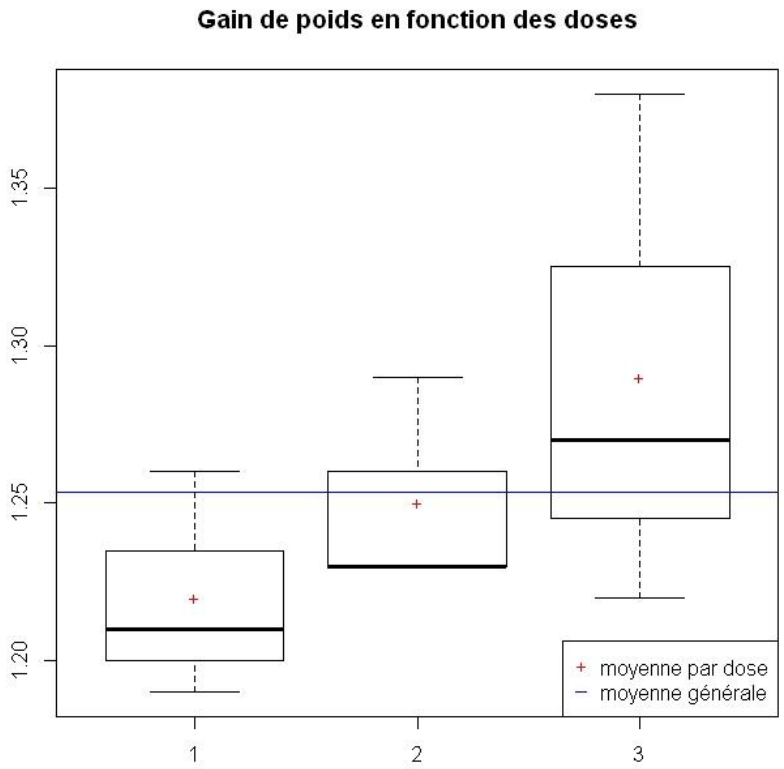
- Expérience : traitement à base de vitamine B12 sur des animaux de races différentes
- Premier facteur : 3 races d'animaux notées R_i
- Second facteur : 3 doses du traitement notées D_j (5, 10 et 15 µg par cm^3)
- Variable quantitative : Y_{ij} = gain moyen de poids par jour à l'issue d'un traitement de 50 jours.
- Un seul animal est utilisé pour chaque couple «race-traitement» → **K=1**.

Analyse de la variance à deux facteurs

Données de l'exemple :

	R_1	R_2	R_3
D_1	1,26	1,21	1,19
D_2	1,29	1,23	1,23
D_3	1,38	1,27	1,22





Analyse de la variance à deux facteurs

La procédure est analogue à celle de l'analyse de la variance à deux facteurs avec répétitions :

- Tester l'effet des races et des doses à partir de tests de Fisher (anova).
- Il faut là encore vérifier les trois conditions fondamentales :

- normalité des résidus : test de Shapiro-Wilk

- homoscédasticité : test de Bartlett

- indépendance des données.

- Effectuer des comparaisons multiples si le facteur a un effet sur la réponse.

Analyse de la variance à deux facteurs

Exemple :

- Vérification des conditions fondamentales :
  - *Normalité* : test de Shapiro → p-value = 0.9632 donc OK
  - *Homoscédasticité* : test de Bartlett → par **race** : p-value = 0.1961

    par **dose**: p-value = 0.5822

  donc les variances sont homogènes. OK
  - *Indépendance* : les données sont indépendantes. OK

- Tester l'effet des facteurs race et dose par Anova :
  - Fisher → **race** : p-value = 0.029 < 0.05 donc la race a un effet significatif sur le gain de poids.
  - Fisher → **dose** : p-value = 0.088 > 0.05 donc la dose n'a pas d'effet significatif sur le gain de poids.

Analyse de la variance à deux facteurs

Exemple :

- Comparaisons multiples : par race

Tukey multiple comparisons of means 95% family-wise confidence level factor levels have been ordered				
Fit: aov(formula = gain ~ race + dose, data = poids)				
$race				
	diff	lower	upper	p adj
2-3	0.023	-0.058	0.104	0.6040
1-3	0.097	0.015	0.178	0.0288
1-2	0.073	-0.008	0.155	0.0687

→ Les gains de poids moyens des races 2 et 3 ne sont significativement pas différents.

IV. Analyse de la covariance

1) Présentation

2) Modèle

3) Procédure d'analyse

4) Exemple d'application

Analyse de la Covariance
Présentation

Qu'est ce que l'analyse de la Covariance?

- Modèle linéaire : - Variables explicatives discrètes (Facteurs)
  - Variables explicatives continues (Covariables)

  ➡ « Mélange de l'analyse de la variance et de la régression »

- Apport de la covariable:

L'ajout d'une covariable dans un modèle d'Anova permet de réduire la variabilité de l'erreur.

IV. Analyse de la covariance

1) Présentation

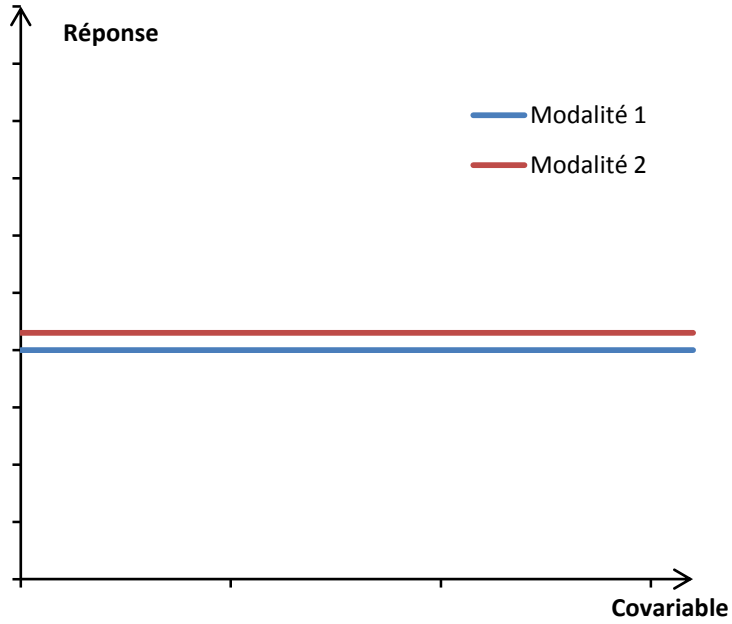
2) Modèle

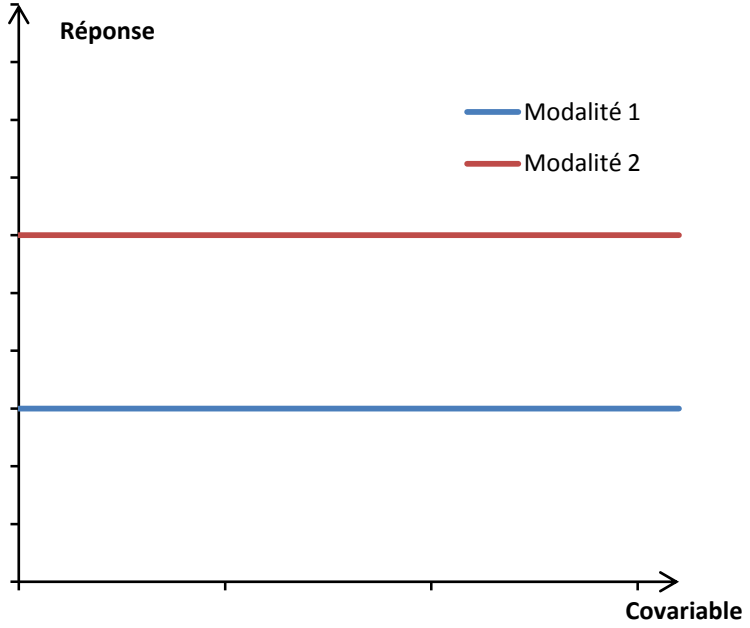
3) Procédure d'analyse

4) Exemple d'application

Analyse de la Covariance
Modèle

Illustration Graphique : modèle avec 1 facteur à 2 modalités et 1 covariable





$$y_{ij} = \mu + \varepsilon_{ij}$$

-> Les moyennes entre les deux modalités ne sont pas significativement différentes.
-> La covariable n'a pas d'effet significatif.

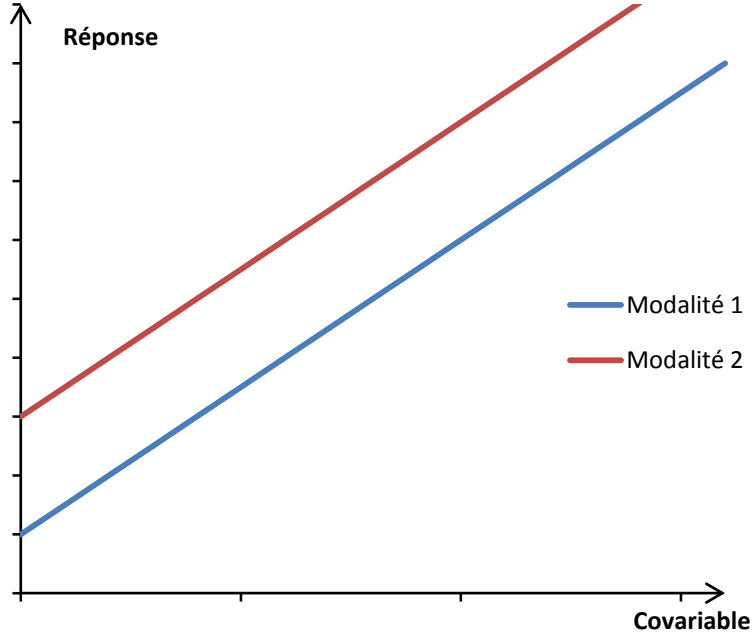
$$y_{ij} = \mu_i + \varepsilon_{ij}$$

-> Les moyennes entre les deux modalités sont significativement différentes.
-> La covariable n'a pas d'effet significatif.

ANOVA

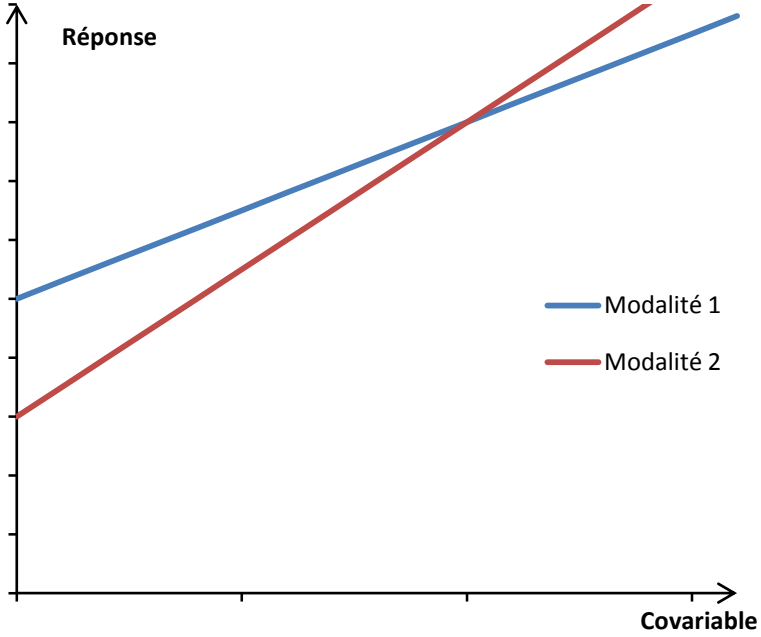
Analyse de la Covariance
Modèle

Illustration Graphique : modèle avec 1 facteur à 2 modalités et 1 covariable



$$y_{ij} = \mu_i + \beta x_{ij} + \varepsilon_{ij}$$

-> La covariable a un effet significatif, mais n'influe pas différemment selon le niveau.



$$y_{ij} = \mu_i + \beta_i x_{ij} + \varepsilon_{ij}$$

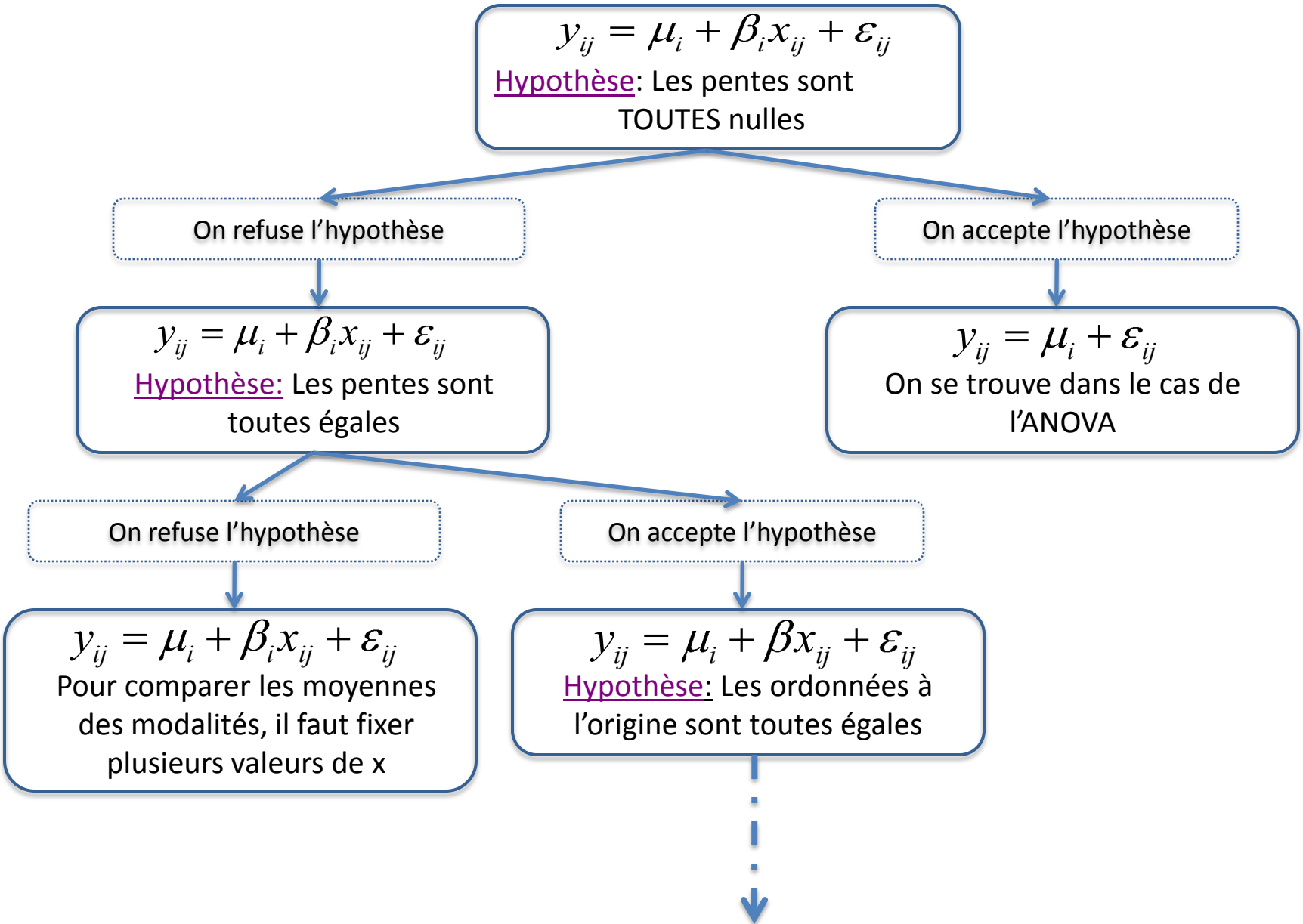
-> La covariable a un effet significatif, et influe différemment selon le niveau.

ANCOVA

IV. Analyse de la covariance

1) Présentation
2) Modèle
3) Procédure d’analyse
4) Exemple d’application

Analyse de la Covariance
Procédure d'analyse



Analyse de la Covariance
Procédure d'analyse

$$y_{ij} = \mu_i + \beta x_{ij} + \varepsilon_{ij}$$

Hypothèse: Les ordonnées à l'origine sont toutes égales

On refuse l'hypothèse

$$y_{ij} = \mu_i + \beta x_{ij} + \varepsilon_{ij}$$

Faire des tests de comparaisons multiples pour savoir quelles moyennes diffèrent

On accepte l'hypothèse

$$y_{ij} = \mu + \beta x_{ij} + \varepsilon_{ij}$$

On se trouve dans le cas d'une régression linéaire simple

Conditions à vérifier :

Ce sont les mêmes que pour l'ANOVA (normalité des résidus, homoscédasticité, indépendance des donnés) et la **linéarité du modèle**.

Analyse de la Covariance
Procédure d'analyse

Modèle 1: $y_{ij} = \mu_i + \beta_i x_{ij} + \varepsilon_{ij}$

Modèle 2: $y_{ij} = \mu_i + \beta x_{ij} + \varepsilon_{ij}$

Modèle 3: $y_{ij} = \mu_i + \varepsilon_{ij}$

- 1ère hypothèse à tester sur la covariable:

$$\begin{cases} H_0 : \beta_1 = \beta_2 = \ldots = \beta_k = 0 \\ H_1 : au\ \ moins\ \ des\ \ \beta_i\ \ est\ \ différent\ \ de\ 0 \end{cases}$$

- Statistique de test:

$$\frac{SC_{\text{modèle 3}} - SC_{\text{modèle 1}} / k}{SC_{\text{modèle 1}} / n - 2k} : \ F(k, n-2k) \text{ sous } \mathrm{H}_0$$

retourDiapositive 68

Analyse de la Covariance
Procédure d'analyse

Modèle 1: $y_{ij} = \mu_i + \beta_i x_{ij} + \varepsilon_{ij}$

Modèle 2: $y_{ij} = \mu_i + \beta x_{ij} + \varepsilon_{ij}$

Modèle 3: $y_{ij} = \mu_i + \varepsilon_{ij}$

- 2ème hypothèse à tester sur la covariable:

$$\begin{cases} H_0 : \beta_1 = \beta_2 = \ldots = \beta_k \\ H_1 : au\ \ moins\ \ des\ \ \beta_i\ \ est\ \ différent\ \ des\ \ autres \end{cases}$$

- Statistique de test:

$$\frac{SC_{\text{modèle 2}} - SC_{\text{modèle 1}} / k-1}{SC_{\text{modèle 1}} / n-2k} : \ F(k-1, n-2k) \text{ sous } \mathrm{H}_0$$

retour

Analyse de la Covariance
Procédure d'analyse

Modèle 1: $y_{ij} = \mu_i + \beta_i x_{ij} + \varepsilon_{ij}$

Modèle 2: $y_{ij} = \mu_i + \beta x_{ij} + \varepsilon_{ij}$

Modèle 3: $y_{ij} = \mu_i + \varepsilon_{ij}$

- <u>Hypothèse à tester sur les modalités:</u>

$$\begin{cases} H_0 : \mu_1 = \mu_2 = \ldots = \mu_k \ (où \ x = 0) \\ H_1 : au \ moins \ des \ \mu_i \ est \ différent \ des \ autres \ (où \ x = 0) \end{cases}$$

- La statistique de test est identique à celle de l'ANOVA.

retour

IV. Analyse de la covariance

1) Présentation
2) Modèle
3) Procédure d’analyse
4) Exemple d’application

Analyse de la Covariance
Exemple d'application

But: Comparer le gain de poids moyen quotidien de bœufs nourris pendant 160 jours selon deux régimes différents. Deux groupes de 8 bœufs sont constitués (Régime = 1 et Régime = 2), et on mesure le poids initial des bêtes (variable poids_ini) en plus de leur gain moyen (variable poids_gain).

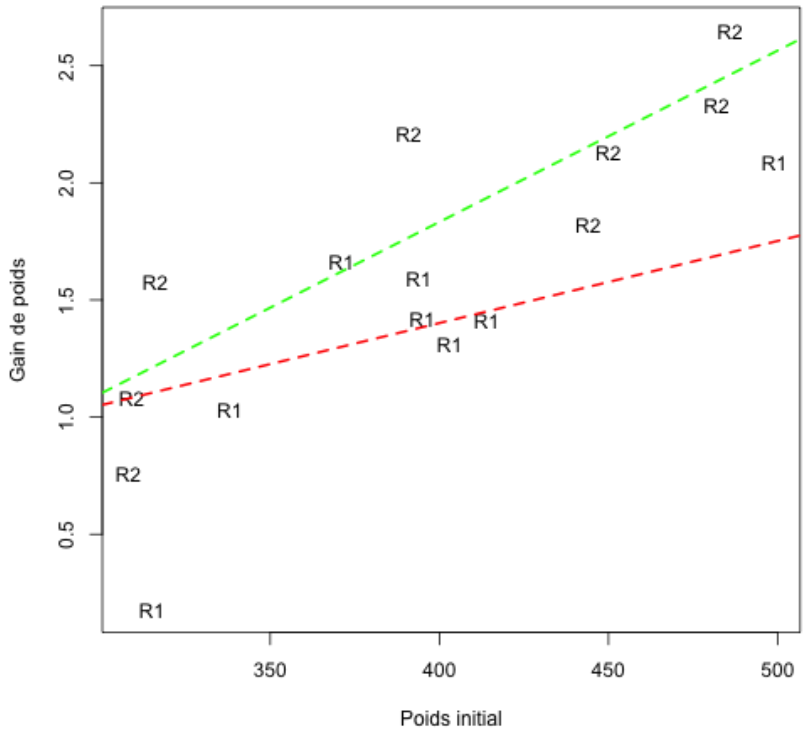
Régime	Poids_gain	Poids_ini
1	1.03	338
1	1.31	403
1	1.59	394
…	…	…
2	1.82	444
2	2.13	450
2	2.33	482
…	…	…

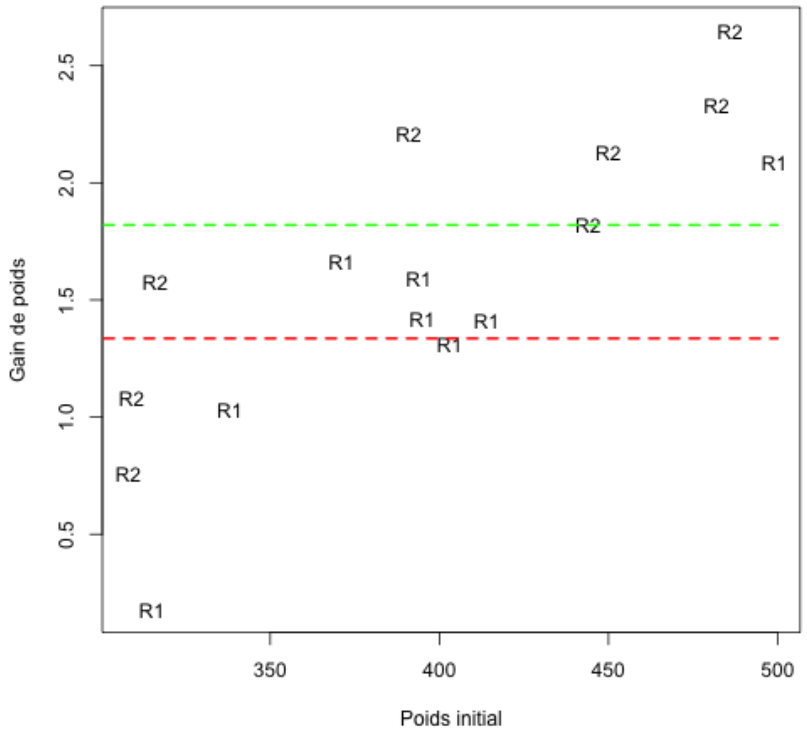
Analyse de la Covariance
Exemple d'application

On teste si la covariable a une influence:

$$poids\_gain_{ij} = régime_i + poids\_ini_i x_{ij} + \varepsilon_{ij}$$

$$poids\_gain_{ij} = régime_i + \varepsilon_{ij}$$





Analyse de la Covariance
Exemple d'application

```
> lm1 = lm (poids_gain ~ regime)
> lm2 = lm (poids_gain ~ regime + poids_ini +regime:poids_ini)
> anova(lm1,lm2)
```

Analysis of Variance Table						
Model 1: poids_gain ~ regime Model 2: poids_gain ~ regime + poids_ini + regime:poids_ini						
	Res Df	**RSS**	**Def**	**Sum of Sq**	**F value**	**Pr(>F)**
1	14	5.10				
2	12	1.29	2	3.81	17.66	**0.000265**

Ici la p-value < 0,05 donc on décide H1

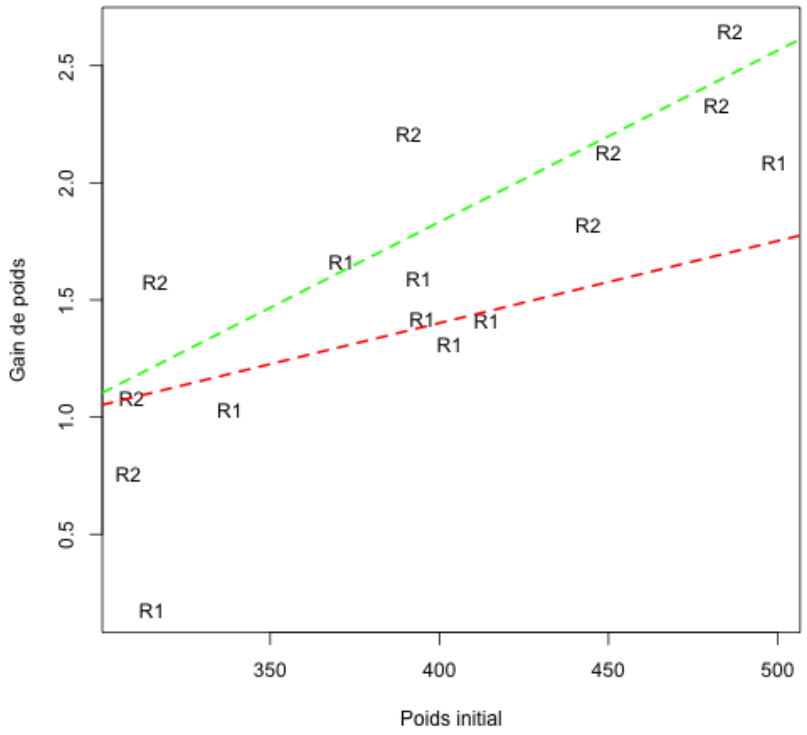
→ Ainsi le poids initial (la covariable) une influence significative sur le gain de poids.

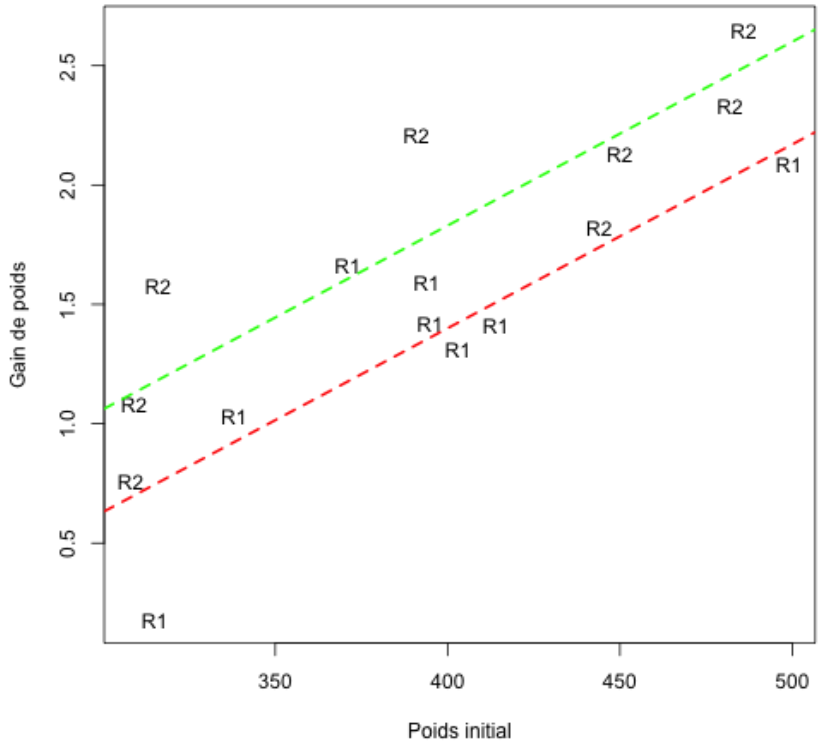
Analyse de la Covariance
Exemple d'application

On teste si les pentes sont toutes égales :

$$poids\_gain_{ij} = régime_i + poids\_ini_i x_{ij} + \varepsilon_{ij}$$

$$poids\_gain_{ij} = régime_i + poids\_ini\ x_{ij} + \varepsilon_{ij}$$





Analyse de la Covariance
Exemple d'application

```
> lm3 = lm (poids_gain ~ regime+ poids_ini)
> lm2 = lm (poids_gain ~ regime + poids_ini + regime:poids_ini)
> anova(lm3,lm2)
```

Analysis of Variance Table						
Model 1: poids_gain ~ regime + poids_ini Model 2: poids_gain ~ regime + poids_ini + regime:poids_ini						
	Res Df	**RSS**	**Def**	**Sum of Sq**	**F value**	**Pr(>F)**
1	13	1.31				
2	12	1.29	1	0.02	0.18	**0.67**

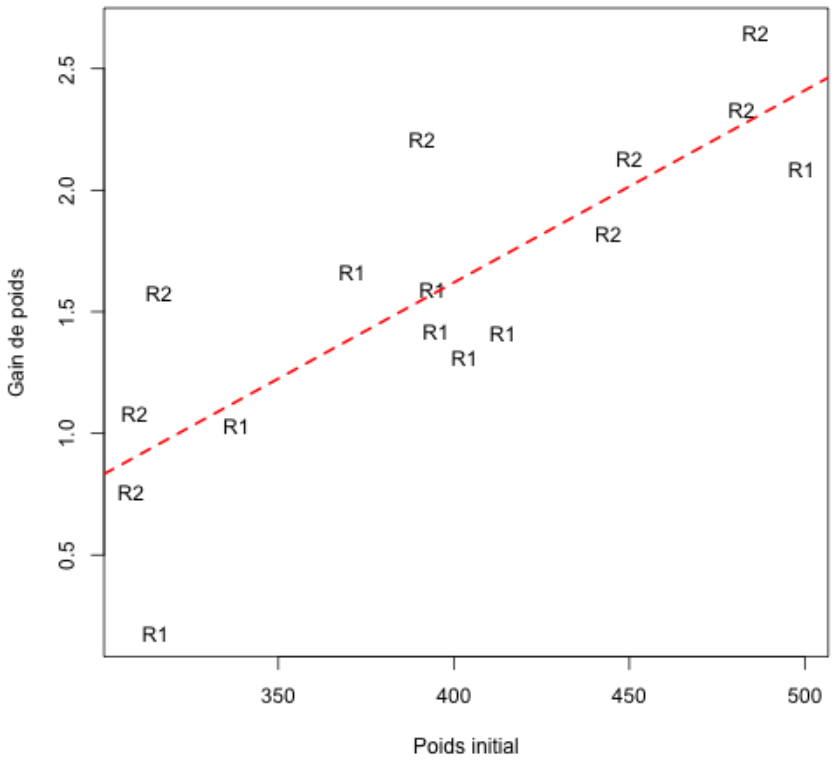
Ici la p-value > 0,05 donc on décide H0

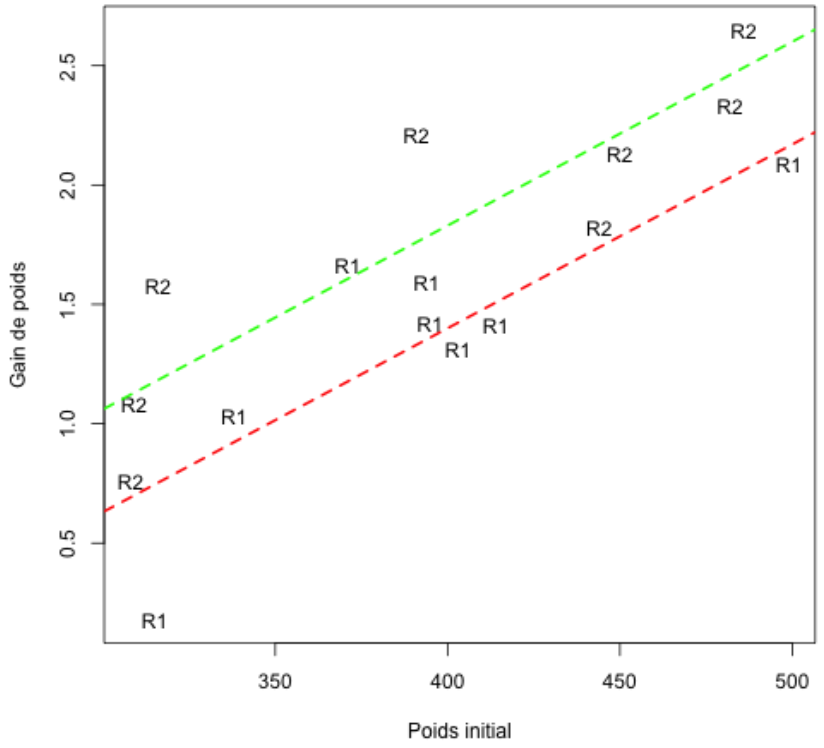
→ Ainsi le poids initial (la covariable) a le même effet quelque soit le régime (les pentes sont les mêmes).

Analyse de la Covariance
Exemple d'application

On teste si les ordonnées à l'origine sont toutes égales :

$$poids\_gain_{ij} = régime + poids\_ini\ x_{ij} + \varepsilon_{ij} \qquad poids\_gain_{ij} = régime_i + poids\_ini\ x_{ij} + \varepsilon_{ij}$$





Analyse de la Covariance
Exemple d'application

```
> lm3 = lm (poids_gain ~ regime+ poids_ini)
> lm4 = lm (poids_gain ~ poids_ini)
> anova(lm4,lm3)
```

Analysis of Variance Table						
Model 1: poids_gain ~ regime Model 2: poids_gain ~ regime+ poids_ini						
	Res Df	**RSS**	**Def**	**Sum of Sq**	**F value**	**Pr(>F)**
1	14	5.10				
2	13	1.31	1	3.79	37.49	**3.643e-05**

Ici la p-value < 0,05 donc on décide H1

→ Ainsi les régimes ont des effets significativement différents. On retient le modèle final: $poids\_gain_{ij} = régime_i + poids\_ini\ x_{ij} + \varepsilon_{ij}$

V. Problèmes spécifiques

1) Hypothèses non vérifiées
2) Modèles à plus de deux facteurs
3) Effets aléatoires

Et si les hypothèses ne sont pas vérifiées?

→ Transformation de la variable Y

Par exemple : log, puissance

→ Test non paramétrique

Par exemple: Kruskal-Wallis

Attention aux valeurs extrêmes ou aberrantes qui peuvent fausser les tests.

V. Problèmes spécifiques

1) Hypothèses non vérifiées

2) Modèles à plus de deux facteurs

3) Effets aléatoires

Analyse de la variance
Problèmes spécifiques

- Pas de problèmes théoriques

- Multiplication des indices et explosion du nombre d'interactions
  ⇒ beaucoup d'expérimentations nécessaires (si plan complet)

- Modèles moins ambitieux : hypothèses sur l'absence d'interactions d'ordres élevés.

- On parle de plans fractionnaires.

V. Problèmes spécifiques

1) Hypothèses non vérifiées

2) Modèles à plus de deux facteurs

3) Effets aléatoires

Analyse de la variance
Problèmes spécifiques

- Effets fixes : Traitements déterminés par l'expérimentateur

Modèle : $Y_{ij} = \mu_i + \varepsilon_{ij}$

- Effets aléatoires : Pas sous le contrôle de l'expérimentateur

Modèle : $Y_{ij} = \mu + \tau_i + \varepsilon_{ij}$ avec $\tau_i \sim N(0, \sigma^2)$

- Différences :
  - Formulation du modèle
  - Effets que l'on peut « généraliser » à la population apparente