

Chapitre 5b – Analyse de la variance à deux facteurs

STT-1920 Méthodes statistiques

Département de mathématiques et de statistique



UNIVERSITÉ
LAVAL

Plan

Expériences factorielles à deux facteurs

Le modèle

Analyse de la variance

Validation des postulats du modèle

Expériences factorielles à deux facteurs

Exemple 1

On s'intéresse à l'effet de la **température** et de la **concentration** d'un catalyseur sur le **rendement** d'une réaction chimique, ainsi qu'à leur **interaction**.

On mène une seule expérience où l'on fait varier la température et la concentration simultanément.

Température	Concentration		
	0,01	0,02	0,03
5 °C	$Y_{111}, Y_{112}, Y_{113},$ Y_{114}, Y_{115}	$Y_{121}, Y_{122}, Y_{123},$ Y_{124}, Y_{125}	$Y_{131}, Y_{132}, Y_{133},$ Y_{134}, Y_{135}
10 °C	$Y_{211}, Y_{212}, Y_{213},$ Y_{214}, Y_{215}	$Y_{221}, Y_{222}, Y_{223},$ Y_{224}, Y_{225}	$Y_{231}, Y_{232}, Y_{233},$ Y_{234}, Y_{235}

Contexte et objectifs

- On veut vérifier si l'effet de A dépend du niveau de B, et vice versa : **interaction** potentielle entre A et B.
- S'il n'y a pas d'interaction, on pourra estimer ou tester :
 - l'effet du facteur A (à a niveaux), et
 - l'effet du facteur B (à b niveaux).
- On prend n mesures de Y pour chacune des ab combinaisons de traitements
 $\Rightarrow abn$ observations.
- L'assignation des unités aux combinaisons de traitements doit se faire de façon **complètement aléatoire**.

Pourquoi une expérience factorielle ?

Approche « un facteur à la fois » : une expérience pour A (B fixé), puis une autre pour B (A fixé).

- Si l'effet de A dépend du niveau de B, on ne le verra **jamais** ;
- il faut plus d'unités expérimentales pour la même précision.

Expérience factorielle (toutes les combinaisons de niveaux)

- permet de détecter et d'estimer l'**interaction** entre A et B ;
- chaque observation sert à estimer **à la fois** l'effet de A et celui de B.

Les principes d'expérimentation du chapitre 5a s'appliquent toujours : **contrôle**, **répétition** (au moins $n = 2$ par combinaison, sinon on ne peut pas séparer l'interaction de l'erreur) et **randomisation**.

QUIZ!

Dans l'exemple 1,

- Quels sont les facteurs?
- Quelle est la variable réponse?
- Que valent a , b et n ? Combien y a-t-il d'observations en tout?

Le modèle

Le modèle

Modèle d'ANOVA, expérience factorielle à deux facteurs

$$Y_{ijk} = \underbrace{\mu + \tau_i + \beta_j + (\tau\beta)_{ij}}_{\mu_{ij}} + \varepsilon_{ijk}, \quad i = 1, \dots, a; j = 1, \dots, b; k = 1, \dots, n$$

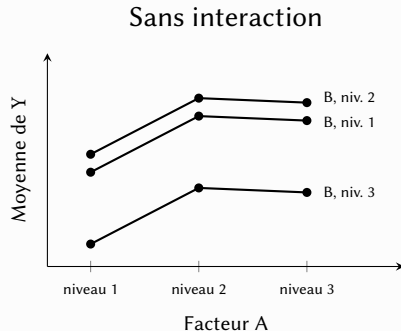
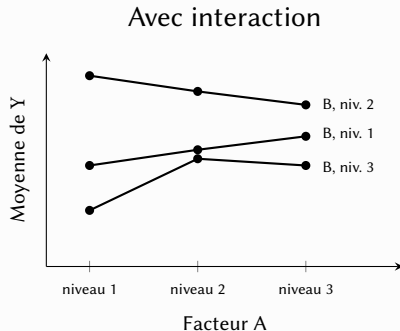
- μ = moyenne globale de la réponse dans la population;
- τ_i = effet du niveau i du facteur A (estimé par $\bar{y}_{i..} - \bar{y}_{...}$);
- β_j = effet du niveau j du facteur B (estimé par $\bar{y}_{.j.} - \bar{y}_{...}$);
- $(\tau\beta)_{ij}$ = effet de l'**interaction** entre A et B au niveau i de A et au niveau j de B (estimé par $\bar{y}_{ij.} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{y}_{...}$);
- ε_{ijk} = erreurs aléatoires indépendantes de loi $N(0, \sigma^2)$.

Les postulats

Les conditions d'application du modèle à deux facteurs sont les mêmes qu'à un facteur :

1. les observations Y_{ijk} sont **indépendantes**
(donc les erreurs ε_{ijk} sont indépendantes);
2. la **variance** des Y_{ijk} est la même (σ^2) pour chaque combinaison de facteurs
(donc la variance des ε_{ijk} est constante);
3. les Y_{ijk} proviennent de lois **normales** $N(\mu_{ij}, \sigma^2)$
(donc les ε_{ijk} sont de loi $N(0, \sigma^2)$).

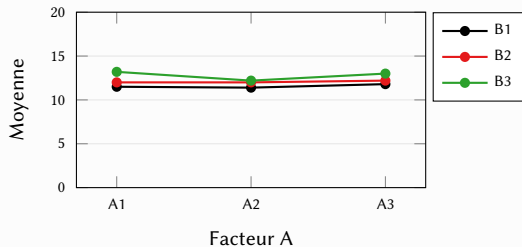
Graphique des moyennes : illustrer les effets des facteurs



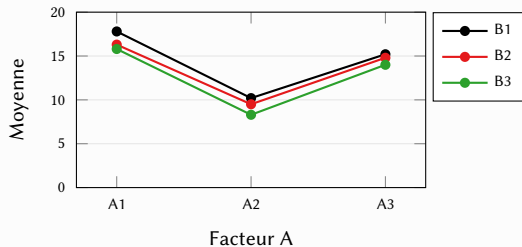
Interaction

En présence d'interaction, l'effet de A **change selon le niveau de B** : les courbes ne sont pas parallèles.

Exercice : quels effets semblent importants ?

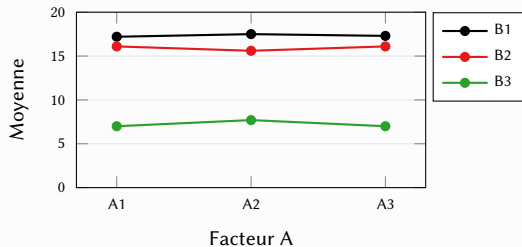


- Interaction ?
- Effet de A ?
- Effet de B ?

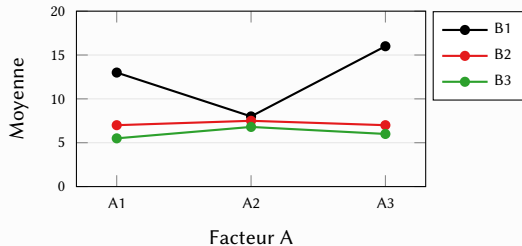


- Interaction ?
- Effet de A ?
- Effet de B ?

Exercice (suite)



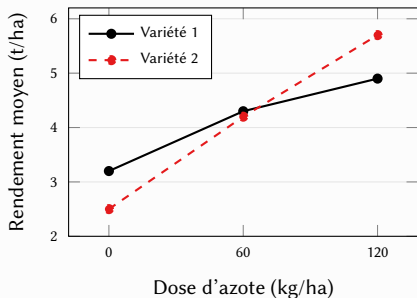
- Interaction ?
- Effet de A ?
- Effet de B ?



- Interaction ?
- Effet de A ?
- Effet de B ?

QUIZ!

Rendement moyen (t/ha) de deux variétés de blé selon la dose d'azote :



Moyennes des variétés (sur les 3 doses) :
4,13 et 4,13 t/ha.

- Les courbes sont-elles parallèles ? Y a-t-il interaction ?
- Peut-on conclure que la variété n'a pas d'effet sur le rendement ?
- Quelle variété recommander sans azote ? Avec 120 kg/ha ?

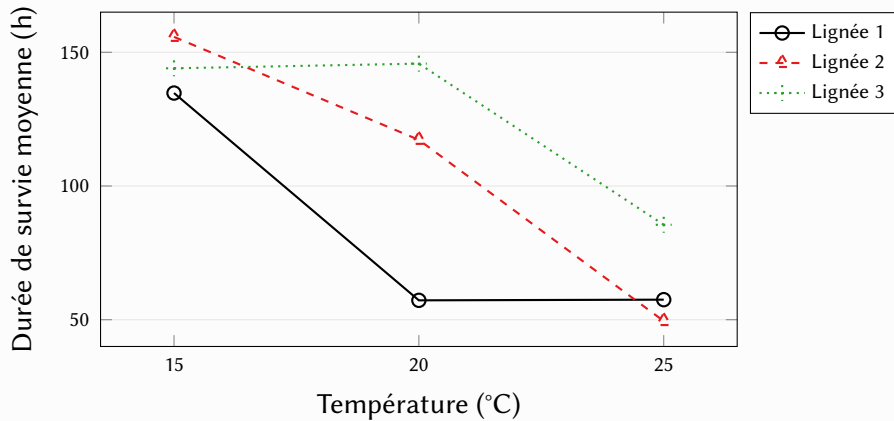
Exemple 2 : résistance au jeûne chez la drosophile

On étudie l'effet de la **lignée** génétique (A) et de la **température** d'élevage (B) sur la **durée de survie** (en heures) de drosophiles privées de nourriture. On place $n = 4$ mouches (chacune dans sa fiole, attribuée au hasard) à chacune des 9 combinaisons lignée $\times$ température ($a = b = 3$).

Lignée	Température					
	15 °C		20 °C		25 °C	
1	130	155	34	40	20	70
	74	180	80	75	82	58
2	150	188	126	122	25	70
	159	126	106	115	58	45
3	138	110	174	120	96	104
	168	160	150	139	82	60

Données fictives, adaptées d'un exemple classique de D. C. Montgomery, *Design and Analysis of Experiments*.

Exemple 2 : graphique des moyennes



Exemple 2 : moyennes des cellules et moyennes marginales

Notation : $\bar{y}_{ij\cdot}$ = moyenne de la combinaison (i, j) ; $\bar{y}_{i\cdot\cdot}$ = moyenne du niveau i de A; $\bar{y}_{\cdot j\cdot}$ = moyenne du niveau j de B; $\bar{y}_{\dots}$ = moyenne globale.

Lignée	Température			$\bar{y}_{i\cdot\cdot}$
	15 °C	20 °C	25 °C	
1	134,75	57,25	57,50	83,17
2	155,75	117,25	49,50	107,50
3	144,00	145,75	85,50	125,08
$\bar{y}_{\cdot j\cdot}$	144,83	106,75	64,17	$\bar{y}_{\dots} = 105,25$

Le graphique des moyennes trace les $\bar{y}_{ij\cdot}$ (le centre du tableau).

QUIZ!

À l'aide du tableau des moyennes de l'exemple 2, estimez :

- l'effet de la lignée 3, τ_3 ;
- l'effet de la température de 15 °C, β_1 ;
- l'interaction $(\tau\beta)_{11}$ (lignée 1 à 15 °C). Comment l'interpréter?

Analyse de la variance

Décomposition de la variabilité observée

$$\underbrace{\sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^n (Y_{ijk} - \bar{Y}_{...})^2}_{SS_T} = \underbrace{bn \sum_{i=1}^a (\bar{Y}_{i..} - \bar{Y}_{...})^2}_{SS_A} + \underbrace{an \sum_{j=1}^b (\bar{Y}_{.j.} - \bar{Y}_{...})^2}_{SS_B} \\ + \underbrace{n \sum_{i=1}^a \sum_{j=1}^b (\bar{Y}_{ij.} - \bar{Y}_{i..} - \bar{Y}_{.j.} + \bar{Y}_{...})^2}_{SS_{AB}} + \underbrace{\sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^n (Y_{ijk} - \bar{Y}_{ij.})^2}_{SS_E}$$

où SS_T est la variabilité **totale**, SS_A et SS_B sont les variabilités dues aux facteurs A et B, SS_{AB} est la variabilité due à l'**interaction** entre A et B, et SS_E est la variabilité due à l'**erreur**.

Table d'analyse de la variance

Source	S.C.	d.l.	Carré moyen	F_0	Valeur-p
Facteur A	SS_A	$a - 1$	$MS_A = \frac{SS_A}{a - 1}$	$F_0^A = \frac{MS_A}{MS_E}$	$P(F_{a-1; ab(n-1)} > F_0^A)$
Facteur B	SS_B	$b - 1$	$MS_B = \frac{SS_B}{b - 1}$	$F_0^B = \frac{MS_B}{MS_E}$	$P(F_{b-1; ab(n-1)} > F_0^B)$
Interaction	SS_{AB}	$(a - 1)(b - 1)$	$MS_{AB} = \frac{SS_{AB}}{(a - 1)(b - 1)}$	$F_0^{AB} = \frac{MS_{AB}}{MS_E}$	$P(F_{(a-1)(b-1); ab(n-1)} > F_0^{AB})$
Erreur	SS_E	$ab(n - 1)$	$MS_E = \frac{SS_E}{ab(n - 1)}$		
Totale	SS_T	$abn - 1$			

QUIZ!

Une agronome compare le rendement de 4 variétés de blé (A) sous 3 régimes d'engrais (B), avec 5 parcelles par combinaison, attribuées au hasard.

- Combien de parcelles lui faut-il en tout ?
- Donnez les degrés de liberté de chaque source de variation (A, B, interaction, erreur, totale).
- S'il n'y a pas d'interaction, quelle loi suit la statistique F_0^{AB} ?

Stratégie d'analyse : l'interaction d'abord !

Ordre d'interprétation

1. On teste d'abord l'**interaction** : $H_0 : (\tau\beta)_{ij} = 0$ pour tous i, j .
2. Si l'interaction est **significative** : l'effet de A dépend du niveau de B. On **n'interprète pas** A et B séparément ; on compare les moyennes des combinaisons $\bar{y}_{ij}$. (graphique des moyennes).
3. Sinon : on interprète les **effets principaux** de A et de B (tests F, moyennes marginales).

Attention! En présence d'interaction, un effet principal non significatif ne veut pas dire que le facteur n'a pas d'effet (souvenez-vous du quiz sur les variétés de blé).

Exemple 2 : ANOVA sur la durée de survie

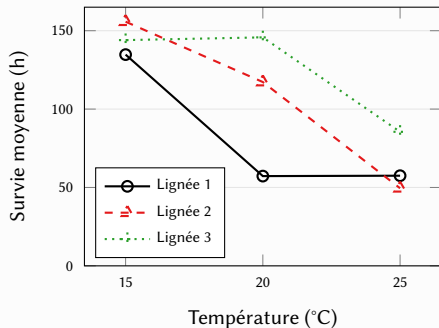
```
> AnovaModel.1 <- lm(Survie ~ Lignee*Temperature, data=drosophiles)
> Anova(AnovaModel.1)
Anova Table (Type II tests)
```

Response: Survie

	Sum Sq	Df	F value	Pr(>F)	
Lignee	10633	2	7.9834	0.001888	**
Temperature	39083	2	29.3438	0.0000001694	***
Lignee:Temperature	9438	4	3.5429	0.018973	*
Residuals	17981	27			

- Quelles conclusions pouvez-vous tirer **au seuil de 1 %** ?
- Quelles conclusions pouvez-vous tirer **au seuil de 5 %** ?

Exemple 2 : interpréter l'interaction (seuil de 5 %)



L'interaction est significative ($p = 0,019$) : l'effet de la température **dépend de la lignée**.

- À 15 °C, les trois lignées survivent de 135 à 156 h en moyenne.
- À 20 °C, la survie de la lignée 1 chute (57 h), alors que la lignée 3 se maintient (146 h).
- À 25 °C, toutes les lignées survivent moins longtemps (50 à 86 h).

Dire simplement « la lignée 3 survit en moyenne 125 h » masquerait ces différences : on interprète les combinaisons, pas les effets principaux seuls.

Validation des postulats du modèle

Postulats et analyse des résidus

1. **Indépendance** des observations :
randomisation ? informations sur la collecte ?
2. **Normalité** des ε_{ijk} :
histogramme et graphique quantile-quantile normal des résidus, test de Shapiro-Wilk;
3. **Variance homogène** des ε_{ijk} pour chaque combinaison de facteurs :
graphique des résidus en fonction des moyennes $\bar{Y}_{ij\cdot}$ (ou diagrammes en boîtes juxtaposés), test de Bartlett.

Résidus

Les erreurs ε_{ijk} sont estimées par les résidus $\hat{\varepsilon}_{ijk} = y_{ijk} - \bar{y}_{ij\cdot}$.

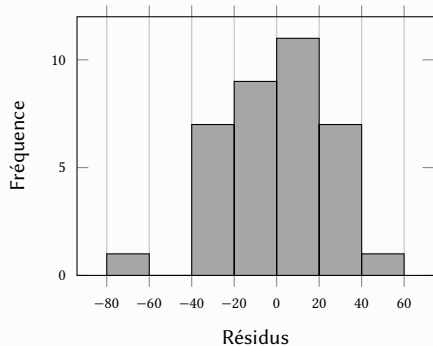
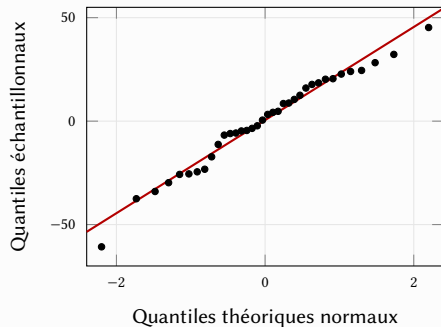
Exemple 2 : R Commander

- Pour commander une analyse de variance à plusieurs facteurs :
menu Statistiques → Moyennes → ANOVA à plusieurs facteurs.
- Pour ajouter les résidus au jeu de données :
menu Modèles → Ajouter les statistiques des observations aux données
(cocher Valeurs ajustées et Résidus).

Directement en R :

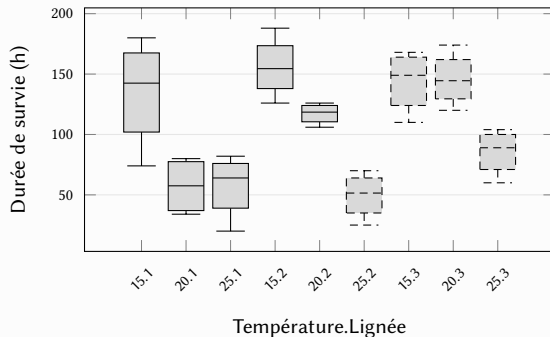
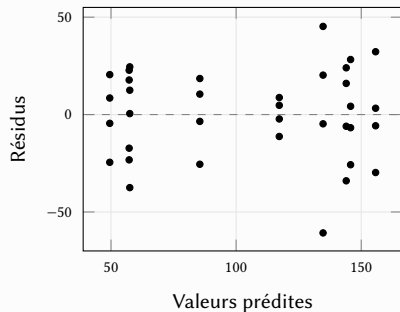
```
modele <- lm(Survie ~ Lignee*Temperature, data=drosophiles)
drosophiles$residus <- residuals(modele)
drosophiles$ajustees <- fitted(modele)
```

Exemple 2 : analyse des résidus (normalité)



Shapiro-Wilk normality test
data: residuals.AnovaModel.1
 $W = 0.9777$, $p\text{-value} = 0.6672$

Exemple 2 : analyse des résidus (égalité des variances)



```
Bartlett test of homogeneity of variances
data: residuals.AnovaModel.1 by interaction(Lignée, Temperature)
Bartlett's K-squared = 6.8636, df = 8, p-value = 0.5514
```

Réponses aux exemples

- **Exemple 1** : facteurs = température (A) et concentration (B); réponse = rendement; $a = 2$, $b = 3$, $n = 5$, donc $abn = 30$ observations.
- **Exercice** : (1) pas d'interaction, peu d'effet de A et de B; (2) pas d'interaction, effet de A, peu d'effet de B; (3) pas d'interaction, peu d'effet de A, effet de B; (4) interaction (les effets de A et B ne s'interprètent pas séparément).
- **Quiz (blé)** : les courbes se croisent : il y a interaction. En moyenne, les deux variétés donnent le même rendement (4,13 t/ha), mais on ne peut pas conclure que la variété n'a pas d'effet : sans azote, la variété 1 est meilleure (3,2 contre 2,5 t/ha); à 120 kg/ha, c'est la variété 2 (5,7 contre 4,9).
- **Quiz (estimation des effets)** : $\hat{\tau}_3 = 125,08 - 105,25 = 19,83$ h; $\hat{\beta}_1 = 144,83 - 105,25 = 39,58$ h;
 $(\widehat{\tau\beta})_{11} = 134,75 - 83,17 - 144,83 + 105,25 = 12,00$ h : à 15 °C, la lignée 1 survit 12 h de plus que ce que prédiraient les seuls effets de la lignée et de la température.
- **Quiz (d.l.)** : $4 \times 3 \times 5 = 60$ parcelles; d.l. A : 3, B : 2, interaction : 6, erreur : $4 \cdot 3 \cdot 4 = 48$, totale : 59. Sous H_0 (pas d'interaction), $F_0^{AB} \sim F_{6,48}$.

Réponses aux exemples (suite)

- **Exemple 2, seuil de 1 %** : l'interaction n'est pas significative ($0,019 > 0,01$); la lignée ($p = 0,002$) et la température ($p < 0,001$) ont chacune un effet significatif sur la durée de survie moyenne.
- **Exemple 2, seuil de 5 %** : l'interaction est significative ($0,019 < 0,05$) : l'effet de la température dépend de la lignée. On interprète à l'aide du graphique des moyennes (p. ex., à 20 °C, la lignée 3 survit beaucoup plus longtemps que la lignée 1 : 146 h contre 57 h).
- **Postulats** : Shapiro-Wilk ($p = 0,67$) et Bartlett ($p = 0,55$) : pas d'évidence contre la normalité ni contre l'égalité des variances.