

Sommaire

Analyse de la variance à deux facteurs emboîtés

Frédéric Bertrand¹ & Myriam Maumy¹

¹IRMA, Université de Strasbourg
Strasbourg, France

Master 1^{re} Année
2016-2017

Références

Ce cours s'appuie essentiellement sur

- 1 le livre David C. Howell, **Méthodes statistiques en sciences humaines** traduit de la sixième édition américaine aux éditions de Boeck, 2008.
- 2 le livre de Pierre Dagnelie, **Statistique théorique et appliquée**, Tome 2, aux éditions de Boeck, 1998.
- 3 le livre de Hardeo Sahai et Mohammed I. Ageel, **The Analysis of Variance : Fixed, Random and Mixed Models**, aux éditions Birkhäuser, 2000.

Introduction

- Nous sommes dans la situation particulière où les effets des niveaux du facteur B n'ont pas de signification concrète, par exemple ces niveaux dépendent du niveau du facteur A considéré et une étude des effets principaux du facteur B n'a pas de pertinence.
- Nous ne pouvons nous servir d'un modèle où les facteurs sont emboîtés^a, que si nous disposons de répétitions. Dans le cas contraire où les essais ne seraient pas répétés, l'effet dû au facteur B ne pourra être étudié et le modèle que nous devrions utiliser pour analyser les données sera l'un de ceux exposés au chapitre de l'analyse de la variance à un facteur.

a. Ces types de modèles sont également appelés des modèles hiérarchiques ou en anglais *hierarchical* ou *nested models*.



Exemple

Ainsi par exemple un fabricant de détergents alimente plusieurs chaînes de distribution : A_1, A_2, \dots, A_I . Nous pensons que les boîtes de produit livrées à certaines chaînes de distribution contiennent une masse de détergent inférieure à celle des autres chaînes de distribution. Pour étudier cette situation, nous décidons de prélever K boîtes dans J magasins de chaque chaîne.



Sommaire

Exemple - Suite

Ainsi le second facteur B_j , associé au j -ème magasin dans la chaîne, est un repère qui n'a aucune signification réelle : il n'y a, par exemple aucune relation entre le magasin n° 3 de la chaîne 1 et le magasin n° 3 de la chaîne 4. Il n'y a donc aucun intérêt à introduire un terme dans le modèle caractérisant l'effet principal du facteur B .

Pour indiquer la dépendance des niveaux du second facteur B aux niveaux du premier facteur A nous notons les niveaux du second facteur $B : B_{j(i)}, 1 \leq i \leq I$ et $1 \leq j \leq J$.

1 Introduction

- Exemple

2 Modèle à effets fixes

- Avec répétitions

3 Modèle à effets aléatoires

- Avec répétitions

4 Modèle à effets mixtes

- Avec répétitions



Exemple (Damon et Harvey, 1987)

L'expérience consiste à évaluer le gain de masse, en grammes, entre la dixième et la vingtième semaine de poulets soumis à quatre régimes alimentaires obtenus en combinant des niveaux faibles ou élevés de Calcium et de Lysine. Deux enclos de six poulets ont été utilisés pour chacun des quatre traitements étudiés.

Remarque

Les deux facteurs, Régime et Enclos, sont contrôlés par l'expérimentateur.

Tableau des données

	Régime							
	LoCaLoL		LoCaHiL		HiCaLoL		HiCaHiL	
	1	2	1	2	1	2	1	2
Enclos	573	1041	618	943	731	416	518	416
Gain	636	814	926	640	845	729	782	729
de	883	498	717	373	866	590	938	590
masse	550	890	677	907	729	552	755	552
(en g)	613	636	659	734	770	776	672	776
	901	685	817	1050	787	657	576	657

Le modèle

Le modèle statistique s'écrit de la façon suivante :

$$Y_{ijk} = \mu + \alpha_i + \beta_{j(i)} + \varepsilon_{ijk}$$

où $i = 1, \dots, I, j = 1, \dots, J, k = 1, \dots, K$,

avec les deux contraintes supplémentaires :

$$\sum_{i=1}^I \alpha_i = 0 \quad \text{et} \quad \sum_{j=1}^J \beta_{j(i)} = 0, \quad \text{pour tout } i \in \{1, \dots, I\}$$

où Y_{ijk} est la valeur prise par la réponse Y dans les conditions $(\alpha_i, \beta_{j(i)})$ lors du k -ème essai. Nous notons $n = I \times J \times K$ le nombre total de mesures ayant été effectuées.

Contexte

- Un facteur contrôlé α se présente sous I modalités, chacune d'entre elles étant notée α_i .
- Un facteur contrôlé β se présente sous J modalités, chacune d'entre elles dépendant du niveau α_j du facteur α et étant alors notée $\beta_{j(i)}$.
- Pour chacun des couples de modalités $(\alpha_i, \beta_{j(i)})$ nous effectuons $K \geq 2$ mesures d'une réponse Y qui est une variable continue.

Conditions classiques de l'ANOVA

Nous postulons les hypothèses classiques de l'ANOVA pour les variables erreurs \mathcal{E}_{ijk} :

- 1 les erreurs sont indépendantes
- 2 les erreurs ont même variance σ^2 inconnue
- 3 les erreurs sont de loi gaussienne.

Relation fondamentale de l'ANOVA

Nous supposons que les conditions d'utilisation de ce modèle sont bien remplies.

Nous utilisons les quantités SC_α , SC_β , $SC_{\alpha\beta}$, SC_R , SC_{TOT} déjà introduites au chapitre précédent.

Nous posons $SC_{\beta|\alpha} = SC_\beta + SC_{\alpha\beta}$.

Nous rappelons la relation fondamentale de l'ANOVA :

$$SC_{TOT} = SC_\alpha + SC_{\beta|\alpha} + SC_R.$$

Tableau de l'ANOVA

Variation	SC	ddl	CM	F_{obs}	F_c
Due au facteur α	SC_α	$I - 1$	cm_α	$\frac{cm_\alpha}{cm_R}$	C_α
Due au facteur β dans α	$SC_{\beta \alpha}$	$I(J - 1)$	$cm_{\beta \alpha}$	$\frac{cm_{\beta \alpha}}{cm_R}$	$C_{\beta \alpha}$
Résiduelle	SC_R	$IJ(K - 1)$	cm_R		
Totale	SC_{TOT}	$n - 1$			

Tests d'hypothèses

L'analyse de la variance à deux facteurs emboîtés à effets fixes avec répétitions permet deux tests de Fisher.

Retour à l'exemple - Sortie avec MINITAB

Analyse de la variance pour Gain de masse, avec utilisation de la somme des carrés ajustée pour les tests

Source	DL	SomCar	séq	CM	ajust	F	P
Régime	3	53943		17981	0,73	0,539	
Enclos							
(Régime)	4	125688		31422	1,28	0,294	
Erreur	40	982654		24566			
Total	47	1162286					
S = 156,737 R carré = 15,46% R carré (ajust) = 0,66 %							

Remarque

Nous supposons que les conditions du modèle sont bien remplies. Ce que nous vérifierons par la suite.

Analyse des résultats

- 1 Pour le premier test, $P\text{-value} = 0,539$, nous décidons de ne pas refuser l'hypothèse nulle (\mathcal{H}_0). Par conséquent, nous n'avons pas réussi à mettre en évidence d'effet du facteur à effets fixes « Régime ». Le risque associé à cette décision est un risque de deuxième espèce. Pour l'évaluer, il resterait à calculer la puissance de ce test.
- 2 Pour le deuxième test, $P\text{-value} = 0,294$, nous décidons de ne pas refuser l'hypothèse nulle (\mathcal{H}_0). Par conséquent, nous n'avons pas réussi à mettre en évidence d'effet du facteur à effets fixes « Enclos » dans le facteur « Régime ». Le risque associé à cette décision est un risque de deuxième espèce. Pour l'évaluer, il resterait à calculer la puissance de ce test.

Sommaire

- 1 Introduction
 - Exemple
- 2 Modèle à effets fixes
 - Avec répétitions
- 3 **Modèle à effets aléatoires**
 - Avec répétitions
- 4 Modèle à effets mixtes
 - Avec répétitions

Exemple (Box et al., 1978)

Box et al. ont récolté les données d'une expérience conçue pour estimer la moisissure contenue dans une pâte de piment produite par une entreprise agro-alimentaire. Pour ce faire, 15 lots de pots de pâte de piment ont été sélectionnés au hasard dans la production de l'entreprise et dans chacun de ces lots, deux pots de pâte ont été à nouveau sélectionnés au hasard. Deux prélèvements distincts de pâte ont été analysés pour chacun de ces pots.

Remarque

Les deux facteurs, Lot et Échantillon, sont tous deux considérés comme des facteurs à effets aléatoires.

Tableau des données

Lot	1	2	3	4	5					
Échan.	1	2	1	2	1	2				
Analyses	40	30	26	25	29	14	30	24	19	17
	39	30	28	26	28	15	31	24	20	17
Lot	6	7	8	9	10					
Échan.	1	2	1	2	1	2	1	2	1	2
Analyses	33	26	23	32	34	29	27	31	13	27
	32	24	24	33	34	29	27	31	16	24
Lot	11	12	13	14	15					
Échan.	1	2	1	2	1	2	1	2	1	2
Analyses	25	25	29	31	19	29	23	25	39	26
	23	27	29	32	20	30	24	25	37	28

Le modèle

Le modèle statistique s'écrit de la façon suivante :

$$Y_{ijk} = \mu + A_j + B_{j(i)} + \mathcal{E}_{ijk}$$

avec $i = 1, \dots, I, j = 1, \dots, J, k = 1, \dots, K$ et où Y_{ijk} est la valeur prise par la réponse Y dans les conditions $(A_j, B_{j(i)})$ lors du k -ème essai. Nous notons $n = I \times J \times K$ le nombre total de mesures ayant été effectuées.

Contexte

- Les termes A_j représentent un échantillon de taille I prélevé dans une population importante. Nous admettrons que les effets des A_j sont distribués suivant une loi normale centrée de variance σ_A^2 .
- Les termes $B_{j(i)}$ représentent un échantillon de taille J prélevé dans une population importante dépendant du niveau A_j du facteur A . Nous admettrons que les effets des $B_{j(i)}$, sont distribués suivant une loi normale centrée de variance $\sigma_{B|A}^2$.
- Pour chacun des couples de modalités $(A_j, B_{j(i)})$ nous effectuons $K \geq 2$ mesures d'une réponse Y qui est une variable continue.

Conditions liées à ce type d'analyse

Nous supposons que

$$\mathcal{L}(A_i) = \mathcal{N}(0, \sigma_A^2), \text{ pour tout } i, \quad 1 \leq i \leq I,$$

$$\mathcal{L}(B_{j(i)}) = \mathcal{N}(0, \sigma_{B|A}^2), \text{ pour tout } j, \quad 1 \leq j \leq J,$$

ainsi que l'indépendance des effets aléatoires :

- les effets aléatoires A_i sont indépendants
- les effets aléatoires $B_{j(i)}$ sont indépendants
- les effets aléatoires A_i et $B_{j(i)}$ sont indépendants.

Conditions classiques de l'ANOVA

Nous postulons les hypothèses classiques de l'ANOVA pour les variables erreurs \mathcal{E}_{ijk} :

- 1 les erreurs sont indépendantes
- 2 les erreurs ont même variance σ^2 inconnue
- 3 les erreurs sont de loi gaussienne.

Ajout de conditions

Nous ajoutons l'indépendance des effets aléatoires et des erreurs due à ce type d'analyse :

- les effets aléatoires A_i et les erreurs \mathcal{E}_{ijk} sont indépendants
- les effets aléatoires $B_{j(i)}$ et les erreurs \mathcal{E}_{ijk} sont indépendants.

Relation fondamentale de l'ANOVA

Nous supposons que les conditions d'utilisation de ce modèle sont bien remplies.

Nous utilisons les quantités SC_A , SC_B , SC_{AB} , SC_R , SC_{TOT} déjà introduites au chapitre précédent.

Nous posons $SC_{B|A} = SC_B + SC_{AB}$.

Nous rappelons la relation fondamentale de l'ANOVA :

$$SC_{TOT} = SC_A + SC_{B|A} + SC_R.$$

Tableau de l'ANOVA

Variation	SC	ddl	CM	F_{obs}	F_c
Due au facteur A	SC_A	$I - 1$	cm_A	$\frac{cm_A}{cm_{B A}}$	C_A
Due au facteur B dans A	$SC_{B A}$	$I(J - 1)$	$cm_{B A}$	$\frac{cm_{B A}}{cm_R}$	$C_{B A}$
Résiduelle	SC_R	$IJ(K - 1)$	cm_R		
Totale	SC_{TOT}	$n - 1$			

Tests d'hypothèses

L'analyse de la variance à deux facteurs emboîtés à effets aléatoires avec répétitions permet deux tests de Fisher.

Premier test

Nous testons l'hypothèse nulle

$$(\mathcal{H}_0) : \sigma_A^2 = 0$$

contre l'hypothèse alternative

$$(\mathcal{H}_1) : \sigma_A^2 \neq 0.$$

Sous l'hypothèse nulle (\mathcal{H}_0) précédente d'absence d'effet du facteur A et lorsque les conditions de validité du modèle sont respectées, $F_{A,obs}$ est la réalisation d'une variable aléatoire qui suit une loi de Fisher à $I - 1$ et $I(J - 1)$ degrés de liberté.

Deuxième test

Nous testons l'hypothèse nulle

$$(\mathcal{H}_0) : \sigma_{B|A}^2 = 0$$

contre l'hypothèse alternative

$$(\mathcal{H}_1) : \sigma_{B|A}^2 \neq 0.$$

Sous l'hypothèse nulle (\mathcal{H}_0) précédente d'absence d'effet du facteur B dans le facteur A et lorsque les conditions de validité du modèle sont respectées, $F_{B|A,obs}$ est la réalisation d'une variable aléatoire qui suit une loi de Fisher à $I(J - 1)$ et $IJ(K - 1)$ degrés de liberté.

Retour à l'exemple - Sortie avec MINITAB

Analyse de la variance pour Analyse, avec utilisation de la somme des carrés ajustée pour les tests

Source	DL	SomCar séq	CM ajusté	F	P
Lot	14	1210,933	86,495	1,49	0,226
Echan.					
(Lot)	15	869,750	57,983	63,25	0,000
Erreur	30	27,500	0,917		
Total	59	2108,183			
S = 957427 R carré = 98,70% R carré (ajusté) = 97,43 %					

Tableau des données

	Père			Mère		
	1	2	3	1	2	3
Gain	2,77	2,58	2,28	3,01	2,36	2,72
masse	2,38	2,94	2,22	2,61	2,71	2,74

	Père				
	4	5	1	2	3
Gain	2,87	2,31	2,74	2,50	2,50
masse	2,46	2,24	2,56	2,48	2,48

Le modèle

Le modèle statistique s'écrit de la façon suivante :

$$Y_{ijk} = \mu + \alpha_i + B_{j(i)} + \mathcal{E}_{i,j,k}$$

où $i = 1, \dots, I, j = 1, \dots, J, k = 1, \dots, K$,
avec les contraintes supplémentaires :

$$\sum_{i=1}^I \alpha_i = 0,$$

où Y_{ijk} est la valeur prise par la réponse Y dans les conditions $(\alpha_i, B_{j(i)})$ lors du k -ème essai.
Nous notons $n = I \times J \times K$ le nombre total de mesures ayant été effectuées.

Contexte

- 1 Un facteur contrôlé α se présente sous I modalités, chacune d'entre elles étant notée α_i .
- 2 Les termes $B_{j(i)}$ représentent un échantillon de taille J prélevé dans une population importante. Nous admettrons que les effets des $B_{j(i)}$ sont distribués suivant une loi normale centrée de variance $\sigma_{B|\alpha}^2$.
- 3 Pour chacun des couples de modalités $(\alpha_i, B_{j(i)})$ nous effectuons $K \geq 2$ mesures d'une réponse Y qui est une variable continue.

Conditions liées à ce type d'analyse

Nous supposons que

- $\mathcal{L}(B_j(i)) = \mathcal{N}(0, \sigma_{B|\alpha}^2)$, pour tout $j, 1 \leq j \leq J$,
- les effets aléatoires $B_{j(i)}$ sont indépendants.

Conditions classiques de l'ANOVA

Nous postulons les hypothèses classiques de l'ANOVA pour les variables erreurs \mathcal{E}_{ijk} :

- 1 les erreurs sont indépendantes
- 2 les erreurs ont même variance σ^2 inconnue
- 3 les erreurs sont de loi gaussienne.

Ajout de conditions

Nous ajoutons l'indépendance des effets aléatoires et des erreurs due à ce type d'analyse :

- les effets aléatoires $B_{j(i)}$ et les erreurs \mathcal{E}_{ijk} sont indépendants.

Tableau de l'ANOVA

Variation	SC	ddl	CM	F_{obs}	F_c
Due au facteur α	sc_α	$I - 1$	cm_α	$\frac{cm_\alpha}{cm_{B \alpha}}$	c_α
Due au facteur B dans α	$sc_{B \alpha}$	$I(J - 1)$	$cm_{B \alpha}$	$\frac{cm_{B \alpha}}{cm_R}$	$c_{B \alpha}$
Résiduelle	sc_R	$IJ(K - 1)$	cm_R		
Totale	sc_{TOT}	$n - 1$			

Relation fondamentale de l'ANOVA

Nous supposons que les conditions d'utilisation de ce modèle sont bien remplies.

Nous utilisons les quantités SC_α , SC_B , $SC_{\alpha B}$, SC_R , SC_{TOT} déjà introduites au chapitre précédent.

Nous posons $SC_{B|\alpha} = SC_B + SC_{\alpha B}$.

Nous rappelons la relation fondamentale de l'ANOVA :

$$SC_{TOT} = SC_\alpha + SC_{B|\alpha} + SC_R.$$

Tests d'hypothèses

L'analyse de la variance à 2 facteurs emboîtés à effets mixtes avec répétitions permet deux tests de Fisher.

Premier test

Nous souhaitons tester l'hypothèse nulle

$$(\mathcal{H}_0) : \alpha_1 = \alpha_2 = \dots = \alpha_J = 0$$

contre l'hypothèse alternative

$$(\mathcal{H}_1) : \text{Il existe } i_0 \in \{1, 2, \dots, J\} \text{ tel que } \alpha_{i_0} \neq 0.$$

Sous l'hypothèse nulle (\mathcal{H}_0) précédente d'absence d'effet du facteur α et lorsque les conditions de validité du modèle sont respectées, $F_{\alpha, obs}$ est la réalisation d'une variable aléatoire qui suit une loi de Fisher à $I - 1$ et $J(J - 1)$ degrés de liberté.

Décision

Nous concluons alors à l'aide de la p -valeur, rejet si elle est inférieure ou égale au seuil α du test, ou à l'aide d'une table, rejet si la valeur $F_{\alpha, obs}$ est supérieure ou égale à la valeur critique issue de la table.

Comparaisons multiples

Lorsque l'hypothèse nulle (\mathcal{H}_0) est rejetée, nous pouvons procéder à des comparaisons multiples des différents effets des niveaux du facteur. Nous renvoyons au chapitre 1 qui traite des principales méthodes de comparaisons multiples.

Deuxième test

Nous testons l'hypothèse nulle

$$(\mathcal{H}_0) : \sigma_{B|\alpha}^2 = 0$$

contre l'hypothèse alternative

$$(\mathcal{H}_1) : \sigma_{B|\alpha}^2 \neq 0.$$

Sous l'hypothèse nulle (\mathcal{H}_0) précédente d'absence d'effet du facteur B dans α et lorsque les conditions de validité du modèle sont respectées, $F_{B|\alpha}$ est la réalisation d'une variable aléatoire qui suit une loi de Fisher à $I(J - 1)$ et $J(K - 1)$ degrés de liberté.

Retour à l'exemple - Sortie avec MINITAB

Analyse de la variance pour Analyse, avec utilisation de la somme des carrés ajustée pour les tests

Source	DL	SomCar séq	CM ajust	F	P
Père	4	0,09973	0,02493	0,22	0,916
Mère					
(Père)	5	0,56355	0,11271	2,91	0,071
Erreur	10	0,38700	0,03870		
Total	19	1,05028			

$$S = 0,196723 \quad R \text{ carré} = 63,15\% \quad R \text{ carré (ajust)} = 29,99\%$$

