A PROPOS DES QUALIFICATIFS COMPLET, ORTHOGONAL ET ÉQUILIBRÉ EN ANALYSE DE LA VARIANCE

J.J. CLAUSTRIAUX[1] et A.F. IEMMA[2]

RÉSUMÉ

Les termes complet, orthogonal et équilibré qui caractérisent un ensemble de données soumis à l'analyse de la variance sont définis et illustrés.

SUMMARY

With examples, we introduce the words complete, orthogonal and balanced used in analysis of variance.

1. INTRODUCTION

L'objectif de cette publication est de préciser les termes **complet, orthogonal** et **équilibré**, qualificatifs rencontrés dans l'exposé de principes statistiques théoriques relatifs, notamment, à l'analyse de la variance comportant plusieurs critères de classification. Ils sont introduits en considérant le cas à deux critères de classification.

Cette note est justifiée par le fait que la pratique de la consultation statistique montre que ces notions sont rarement bien comprises par les chercheurs non spécialistes, alors que les logiciels statistiques utilisés en "libre service" proposent des options d'analyse qui s'y réfèrent en offrant des résultats dont les interprétations peuvent être incompatibles, voire inexactes, avec les objectifs poursuivis.

Après la présentation des notations et de deux exemples (paragraphe 2), les notions sont définies (paragraphe 3). Toutes les combinaisons possibles des trois termes sont ensuite présentées (paragraphe 4). Enfin, la démarche à suivre pour analyser pratiquement un tableau de données à deux facteurs fixes au moyen des procédures GLM des logiciels statistiques Minitab et SAS (paragraphe 5) précède quelques conclusions (paragraphe 6).

---

(1) Professeur à la Faculté universitaire des Sciences agronomiques de Gembloux (Belgique).

(2) Professeur à l'*Escola Superior de Agricultura Luiz De Queiroz*, Piracicaba, São Paulo (Brésil).

Signalons encore que parmi les nombreux ouvrages statistiques approfondissant ces termes, une référence utile est certainement celle de SEARLE [1987]. Une publication [IEMMA et CLAUSTRIAUX, 1999] contribue, notamment, à étayer des considérations du paragraphe 5.

2. NOTATIONS ET EXEMPLES

Avant tout calcul, les données issues d'un dispositif expérimental étudiant deux sources de variation peuvent être rassemblées dans un tableau à deux entrées. En considérant l'indice i pour la première source de variation (i = 1, ..., p) et l'indice j pour la deuxième source de variation (j = 1, ..., q), on peut établir la distribution de fréquences à deux dimensions des effectifs n_{ij} par combinaison ou cellule ij, les fréquences marginales et totale étant notées $n_{i.}$, $n_{.j}$ et $n_{..}$.

Le tableau 1 illustre deux situations. Le premier exemple est relatif à la distribution de 34 lots de graines d'orge traitées par quatre doses différentes de filtrats et extraits de cultures de trois souches différentes d'un champignon pathogène [DAGNELIE, CLAUSTRIAUX, 1981]. Le deuxième exemple concerne la répartition par sexe d'animaux ayant reçu trois alimentations différentes [DAGNELIE, CLAUSTRIAUX, 1981]. Les données de ces deux exemples figurent en annexe.

Tableau 1. Fréquences des observations pour les deux exemples.

Exemple 1	Dose 1	Dose 2	Dose 3	Dose 4	$n_{i.}$
Souche 1	3	3	3	3	12
Souche 2	3	3	3	2	11
Souche 3	3	2	3	3	11
$n_{.j}$	9	8	9	8	34

Exemple 2	Alimentation 1	Alimentation 2	Alimentation 3	$n_{i.}$
Mâle	4	4	4	12
Femelle	6	6	6	18
$n_{.j}$	10	10	10	30

3. DÉFINITIONS

3.1. Terme complet

Le qualificatif **complet** (*complete*) est utilisé lorsque la condition suivante est satisfaite:

$$n_{ij} > 0, \text{ pour tout i et j.}$$

Dans le cas contraire, le terme **incomplet** (*incomplete*) est employé.

Souvent pour l'analyse statistique, un tableau de données complet est aussi qualifié de cas **équirépété** (*equireplicate*) si les fréquences n_{ij} sont constantes et de cas **non équirépété** sinon. A un tableau incomplet correspond aussi un dispositif ou ensemble de données avec **cellule(s) vide(s)** (*empty cell*).

Les deux exemples présentés (tableau 1) sont complets et non équirépétés.

3.2. Terme orthogonal

Pour chacune des cellules de la distribution de fréquence, la quantité suivante peut être calculée:

$$k_{ij} = \frac{n_{i.}\, n_{.j}}{n_{..}};$$

elle correspond dans une table de contingence au calcul d'une fréquence attendue $\hat{n}_{ij}$.

Si $k_{ij} = n_{ij}$, le qualificatif **orthogonal** (*orthogonal data, proportionate sub-class numbers*) est d'application, sinon on utilise le terme **non orthogonal** (*non orthogonal data, disproportionate sub-class numbers*).

Les expressions anglaises font parfaitement apparaître que ce terme qualifie la répartition des observations et non le modèle d'analyse: on parle de données orthogonales ou non orthogonales; certes, dans le second cas, elles engendrent un modèle d'analyse dont les différentes sommes des carrés des écarts ne sont pas indépendantes; il est souvent appelé modèle non orthogonal.

On note également qu'en planification d'expériences, au qualificatif orthogonal est associée l'expression **plan orthogonal** (*orthogonal design*).

Pour le premier exemple, les valeurs k_{ij} sont différentes des n_{ij} (exemple pour $n_{11} : k_{11} = 3{,}18$); elles sont égales dans le deuxième exemple. On conclut que les données sont dites respectivement non orthogonales et orthogonales.

Signalons encore que la condition nécessaire et suffisante d'orthogonalité définie ci-dessus engendre que les fréquences des cellules sont proportionnelles, c'est-à-dire :

$$\frac{n_{ij}}{n_{ij'}} = \frac{n_{i'j}}{n_{i'j'}}.$$

Ceci permet de comprendre la raison pour laquelle le terme proportionnel est parfois préféré en langue anglaise au mot orthogonal.

Si l'orthogonalité n'est pas vérifiée, il convient d'être très prudent aux méthodes d'estimations des sommes des carrés des écarts.

3.3. Terme équilibré

L'examen des fréquences n_{ij} permet d'introduire le terme **équilibré** (*balanced*) qui qualifie un tableau si elles sont égales. Sinon, le terme **déséquilibré** ou **non équilibré** (*unbalanced*) est utilisé.

Cependant, il faut noter le cas particulier du déséquilibre uniforme ou planifié, rencontré, notamment, en expérimentation sous l'expression **plan équilibré** (*balanced design*) et entraînant ainsi une certaine confusion dans les définitions.

Pour un tableau complet et orthogonal, le déséquilibre uniforme se traduit pas des fréquences n_{ij} constantes pour chacune des variantes d'une des deux sources de variation; les fréquences marginales pour l'autre source de variation sont égales. Pour un tel tableau, aussi appelé tableau **à effectifs proportionnels** (*proportionate subclass numbers*), signalons dès à présent que l'analyse est identique à celle d'un tableau complet, orthogonal et équilibré.

Pour un tableau incomplet, c'est-à-dire non orthogonal, déséquilibré uniformément, les fréquences n_{ij} non nulles sont égales et les cellules vides sont uniformément réparties *a priori*, de telle sorte que chaque combinaison des variantes des deux sources de variation soient également représentées (tableaux 2 et 3 : figure 2.2.2'). L'analyse de ce type de tableau se simplifie partiellement [IEMMA, CLAUSTRIAUX, 1999].

Si on se réfère au tableau 1, on peut dire que les deux exemples sont non équilibrés; en particulier, le deuxième exemple est un cas à déséquilibre uniforme.

4. COMBINAISON DES QUALIFICATIFS

4.1. Au moins une fréquence n_{ij} supérieure à un

Pour les trois qualificatifs variant théoriquement selon deux modalités, il est utile de présenter leurs combinaisons factorielles, accompagnées chaque fois d'un exemple simple reprenant uniquement les fréquences n_{ij}, $n_{i.}$, $n_{.j}$ et $n_{..}$.

Toutefois, il y a lieu tout d'abord de considérer le cas général où au moins une fréquence n_{ij} est supérieure à l'unité (tableau 2). Comme on peut le constater, plusieurs cas sont impossibles.

Tableau 2. Cas possibles (au moins une fréquence $n_{ij} > 1$).

1. Complet

1.1. Orthogonal

1.1.1. Equilibre

3	3	3	3	12
3	3	3	3	12
3	3	3	3	12
9	**9**	**9**	**9**	**36**

1.1.2. Déséquilibre

6	1	8	15
6	1	8	15
12	**2**	**16**	**30**

1.1.2'. Déséquilibre uniforme

2	2	2	6
4	4	4	12
6	6	6	18
12	**12**	**12**	**36**

1.2. Non orthogonal

1.2.1. Equilibre

Impossible

1.2.2. Déséquilibre

3	3	3	3	12
3	3	3	2	11
3	2	3	3	11
9	**8**	**9**	**8**	**34**

1.2.2'. Déséquilibre uniforme

3	3	2	8
2	3	3	8
3	2	3	8
8	**8**	**8**	**24**

2. Incomplet

2.1. Orthogonal

2.1.1. Equilibre

Impossible

2.1.2. Déséquilibre

Impossible

2.2. Non orthogonal

2.2.1. Equilibre

Impossible

2.2.2. Déséquilibre

0	3	3	3	9
1	2	2	1	6
3	0	3	3	9
4	**5**	**8**	**7**	**24**

2.2.2'. Déséquilibre uniforme

0	3	3	3	9
3	3	0	3	9
3	0	3	3	9
3	3	3	0	9
9	**9**	**9**	**9**	**36**

4.2. Cas particulier: fréquences n_{ij} unitaires

On rencontre aussi des tableaux complets ou incomplets avec des fréquences nulles ou unitaires.

Il a semblé utile de présenter l'ensemble des cas possibles (tableau 3); évidemment, six situations sont impossibles, à savoir :

- un tableau complet, orthogonal et déséquilibré;
- un tableau complet, non orthogonal et équilibré;
- un tableau complet, non orthogonal et déséquilibré;
- un tableau incomplet, orthogonal et équilibré;
- un tableau incomplet, orthogonal et déséquilibré;
- un tableau incomplet, orthogonal et équilibré.

En expérimentation, le tableau complet, orthogonal et équilibré correspond, par exemple, à la description d'un dispositif en blocs aléatoires complets et le tableau incomplet, non orthogonal et déséquilibré uniformément à celle d'un dispositif en blocs aléatoires incomplets équilibrés, d'un lattice équilibré ou même d'un carré latin.

5. ANALYSE

5.1. Au moins une fréquence n_{ij} supérieure à un

L'analyse statistique des données d'un tableau à deux critères croisés fixes peut se réaliser, par exemple, grâce aux deux procédures disponibles dans le logiciel Minitab [X, 1996], destinées aux calculs des sommes des carrés des écarts (*sequential sum of square* ou *Seq SS* et *adjusted sum of square* ou *Adj SS*) ou aux quatre propositions du logiciel SAS [X, 1990] notées types 1, 2, 3 ou 4. Il faut signaler que les deux types proposés par Minitab correspondent respectivement aux types 1 (*Seq SS*) et 3 (*Adj SS*) de SAS.

En considérant, d'une part, un tableau à deux critères (C1 et C2 ou A et B) comprenant au moins une fréquence n_{ij} supérieure à l'unité et, d'autre part, qu'il y a intérêt à étudier l'interaction totale ou partielle entre les deux sources croisées de variation, les résultats sont obtenus en appliquant les procédures suivantes, respectivement pour Minitab et SAS :

```
GLM C3 = C1 C2 C1*C2
```

ou

```
PROC GLM;
     CLASS A B;
     MODEL Y = A B A*B/options;
```

Les options à choisir dans ce second cas sont en principe SS1, SS2, SS3 ou SS4, en fonction du type choisi (1, 2, 3 ou 4).

1° Pour tester uniquement l'absence d'interaction totale ou partielle (cas incomplet), il faut considérer l'option *Seq SS* ou *Adj SS* pour Minitab et l'un des quatre types pour SAS, puisque les résultats fournis sont identiques.

Tableau 3. Cas possibles $\left(n_{ij} = 1 \text{ ou } 0\right)$.

1. Complet

1.1. Orthogonal

1.1.1. Equilibre

1	1	1	1	4
1	1	1	1	4
1	1	1	1	4
3	3	3	3	12

1.1.2. Déséquilibre

Impossible

1.2. Non orthogonal

1.2.1. Equilibre

Impossible

1.2.2. Déséquilibré

Impossible

2. Incomplet

2.1. Orthogonal

2.1.1. Equilibre

Impossible

2.1.2. Déséquilibre

Impossible

2.2. Non orthogonal

2.2.1. Equilibre

Impossible

2.2.2. Déséquilibre

0	1	1	1	3
1	1	1	1	4
1	0	1	1	3
2	2	3	3	10

2.2.2'. Déséquilibre uniforme

0	1	1	1	3
1	1	0	1	3
1	0	1	1	3
1	1	1	0	3
3	3	3	3	12

En cas de rejet de l'hypothèse sur l'interaction ou interaction significative, il est suggéré de réaliser une analyse de la variance à un critère de classification pour chacune des variantes de l'un ou l'autre des deux critères ou encore de structurer les moyennes des combinaisons ij ou $\hat{m}_{ij}$, par exemple grâce aux méthodes de comparaisons multiples des moyennes si les critères sont qualitatifs, en étant néanmoins attentif aux estimations des erreurs-standards en cas d'un tableau non orthogonal.

2° Si l'hypothèse d'absence d'interaction totale ou partielle est acceptée, les résultats sont obtenus en omettant dans la procédure C1*C2 ou A*B.

Pour avoir des estimations correctes des sommes des carrés des écarts des effets principaux, il faut simplement déterminer le caractère orthogonal ou non orthogonal du tableau des fréquences (paragraphe 3.2).

Pour un tableau orthogonal, on se réfère à l'option *Seq SS* ou *Adj SS* pour Minitab ou à l'un des quatre types pour SAS.

Par contre, si le tableau est non orthogonal, l'option *Adj SS* de Minitab ou l'un des types 2, 3 ou 4 de SAS est considéré. L'ordre d'entrée des critères dans la procédure est important en présence de cellule vide pour obtenir une somme des carrés des écarts ajustée qui soit correcte (pour A : A B; pour B : B A).

3° En supposant respectées les conditions d'application, les résultats de l'analyse de la variance pour le premier exemple (tableau 1), traité par le logiciel SAS, figurent au tableau 4 pour le modèle A B A*B et pour le modèle B A B*A. L'interaction n'étant pas significative, le tableau 5 rassemble les informations équivalentes pour le modèle A B et pour le modèle B A.

La source de variation des doses étant très hautement significative ($P_r > F = 0{,}0001$), l'hypothèse nulle relative aux souches étant aussi acceptée, l'analyse de la variance peut encore se réduire à une analyse de la variance à un critère de classification (tableau 6).

Dès lors, les estimations des moyennes et des erreurs-standards pour les doses se déterminent simplement comme suit (tableau 7) :

$$\hat{m}_{i.} = \frac{1}{n_{i.}} \sum_{j=1}^{q} \sum_{k=1}^{n_{ij}} x_{ijk},$$

$$\hat{\sigma}_{\bar{X}_{i.}} = \sqrt{\frac{\hat{\sigma}^2}{n_{i.}}},$$

$\hat{\sigma}^2$ = carré moyen de l'erreur.

Tableau 4. Premier exemple : modèles A B A*B et B A B*A.

Sources de variation	Degrés de liberté	Sommes des carrés des écarts	Carrés moyens	F_{obs}
		Type 1		
Doses	3	18.999,95	6.333,32	
Souches	2	250,78	125,39	
Doses × Souches	6	123,79	20,63	0,35
		Type 2		
Doses	3	19.071,98	6.357,32	
Souches	2	250,78	125,39	
Doses × Souches	6	123,79	20,63	0,35
		Type 3		
Doses	3	18.991,45	6.330,48	
Souches	2	256,84	128,42	
Doses × Souches	6	123,79	20,63	0,35
		Type 4		
Doses	3	18.991,45	6.330,48	
Souches	2	256,84	128,42	
Doses × Souches	6	123,79	20,63	0,35
		Type 1		
Souches	2	178,75	89,38	
Doses	3	19.071,98	6.357,32	
Souches × Doses	6	123,79	20,63	0,35
		Type 2		
Souches	2	250,78	125,39	
Doses	3	19.071,98	6.357,32	
Souches × Doses	6	123,79	20,63	0,35
		Type 3		
Souches	2	256,84	128,42	
Doses	3	18.991,45	6.330,48	
Souches × Doses	6	123,79	20,63	0,35
		Type 4		
Souches	2	256,84	128,42	
Doses	3	18.991,45	6.330,48	
Souches × Doses	6	123,79	20,63	0,35
Erreur	22	1.295,22	58,87	
Total	33	20.669,74		

Tableau 5. Premier exemple : modèles A B et B A.

Sources de variation	Degrés de liberté	Sommes des carrés des écarts	Carrés moyens	F_{obs}
		Type 1		
Doses	3	18.999,95	6.333,32	
Souches	2	250,78	125,39	
		Type 2		
Doses	3	19.071,98	6.357,32	125,44
Souches	2	250,78	125,39	2,47
		Type 3		
Doses	3	19.071,98	6.357,32	
Souches	2	250,78	125,39	
		Type 4		
Doses	3	19.071,98	6.357,32	
Souches	2	250,78	125,39	
		Type 1		
Souches	2	170,75	85,38	
Doses	3	19.071,98	6.357,32	
		Type 2		
Souches	2	250,78	125,39	2,47
Doses	3	19.071,98	6.357,32	125,44
		Type 3		
Souches	2	250,78	125,39	
Doses	3	19.071,98	6.357,32	
		Type 4		
Souches	2	250,78	125,39	
Doses	3	19.071,98	6.357,32	
Erreur	28	1.419,01	50,68	
Total	33	20.669,74		

Tableau 6. Premier exemple : modèle A.

Sources de variation	Degrés de liberté	Somme des carrés des écarts	Carrés moyens	F_{obs}
Doses	3	18.999,95	6.333,32	113,79
Erreur	30	1.669,79	55,66	
Total	33	20.699,74		

Tableau 7. Estimations des moyennes et des erreurs-standards pour les doses (modèle A).

Doses	Moyennes	Erreurs-standards
1	78,66	2,49
2	56,29	2,64
3	33,03	2,49
4	16,30	2,64

Pour les souches, les paramètres se confondent à ceux de la moyenne générale :

$$\hat{m} = \frac{1}{n_{..}} \sum_{i=1}^{p} \sum_{j=1}^{q} \sum_{k=1}^{n_{ij}} x_{ijk} = 46{,}64,$$

$$\hat{\sigma}_{\overline{X}} = \sqrt{\frac{\hat{\sigma}^2}{n_{..}}} = 1{,}28.$$

4° A titre d'information, en considérant le modèle A B A*B et en supposant la présence de l'interaction, les estimations des moyennes et des erreurs-standards pour l'un des deux critères, par exemple le facteur dose, se déterminent comme suit :

$$\hat{m}_{i.} = \frac{1}{q} \sum_{j=1}^{q} \hat{m}_{ij},$$

$$\hat{\sigma}_{\overline{X}_{i.}} = \sqrt{\frac{\hat{\sigma}^2}{q^2} \sum_{j=1}^{q} \frac{1}{n_{ij}}},$$

$\hat{\sigma}^2$ = carré moyen de l'erreur.

Ces estimations sont présentées au tableau 8. La simple transposition des indices permet de calculer $\hat{m}_{.j}$ et $\hat{\sigma}_{\overline{X}_{.j}}$, à savoir les paramètres pour les souches.

Tableau 8. Estimations des moyennes et des erreurs-standards pour les doses (modèles A B A*B).

Doses	Moyennes	Erreurs-standards
1	78,66	2,56
2	56,83	2,76
3	33,03	2,56
4	15,88	2,76

Si on considère toujours le modèle à deux critères, cette fois-ci sans interaction (modèle A B), les estimations des moyennes et des erreurs-standards ne sont pas aussi évidentes à calculer. Il faut tenir compte des solutions des équations normales pour les estimations de la moyenne générale, des effets relatifs au premier facteur et au second facteur [SEARLE, SPEED, MILLIKEN, 1980]. Le tableau 9 rassemble les résultats proposés par SAS (*LSMEAN* et *Std Ers LSMEAN*) pour les doses.

Tableau 9. Estimations des moyennes et des erreurs-standards pour les doses (modèles A B).

Doses	Moyennes	Erreurs-standards
1	78,66	2,37
2	56,76	2,53
3	33,03	2,37
4	16,14	2,53

5° Pour le second exemple (tableau 1), l'hypothèse d'absence d'interaction étant acceptée, on aboutit au tableau d'analyse de la variance du modèle complet, orthogonal et équilibré (tableau 10).

Tableau 10. Second exemple : modèle A B ou B A.

Sources de variation	Degrés de liberté	Somme des carrés des écarts	Carrés moyens	F_{obs}
Sexes	1	118.682,69	118.682,69	50,53
Alimentations	2	1.014,87	507,44	< 1
Erreur	26	61.071,91	2.348,92	
Total	29	180.769,47		

5.2. Cas particulier: fréquences n_{ij} unitaires

Pour le cas particulier d'un tableau complet ou incomplet avec fréquences n_{ij} unitaires, la règle énoncée au 2° du paragraphe 5.1 est d'application.

Dans ce cas, on suppose ne pas s'intéresser à l'interaction des deux critères et l'un des deux critères est au moins considéré comme aléatoire.

6. EN GUISE DE CONCLUSION

Finalement, en pratique, en l'absence d'interaction, le qualificatif essentiel à déterminer est orthogonal (ou non).

Rappelons aussi que le cas le plus facile à analyser est de toute façon un tableau complet, équilibré et orthogonal. De plus, des méthodes simples d'estimations de données manquantes peuvent être très utiles dans certains cas pour se rapprocher de la norme énoncée ci-dessus.

Enfin, en présence d'un tableau à plusieurs dimensions incomplet, non orthogonal et déséquilibré, dont les facteurs étudiés sont fixes, une méthode simple pour analyser les données est de reconstruire un tableau à une entrée pour y appliquer ensuite l'analyse de la variance à un critère de classification, éventuellement suivie d'une structuration des moyennes s'il y a rejet de l'hypothèse nulle supposant l'égalité des effets relatifs aux objets étudiés. Dans ce cas, les moyennes ne seront pas biaisées. Mais, il sera plus difficile de mettre en évidence ou d'interpréter les interactions.

7. REMERCIEMENTS

Au terme de cette note, nous exprimons nos remerciements à Madame J. AUSTRAET, qui a largement contribué à sa mise au point.

8. RÉFÉRENCES

DAGNELIE P., CLAUSTRIAUX J.J. [1981]. *Théorie et méthodes statistiques, exercices*. Gembloux, Presses agronomiques, 186 p.

IEMMA A., CLAUSTRIAUX J.J. [1999]. Etude des hypothèses de l'analyse de la variance à deux critères de classification : approche par l'exemple. Gembloux, *Notes Stat. Inform.* (Gembloux) 99/3, 36 p.

SEARLE S.R. [1987]. *Linear models for unbalanced data*. New York, John Wiley, 536 p.

SEARLE S.R., SPEED F.M., MILLIKEN G.A. [1980]. Population marginal means in the linear model : an alternative to least squares means. *Amer. Stat.* 34, 216-221.

X [1990]. *SAS/STAT User's guide*. Cary, SAS Institute, vol. 2, 891-1686.

X [1996]. *Minitab reference manual, release 11 for Windows, Windows 95, Windows NT*. PA State College, Minitab.

9. ANNEXES

Données de l'exemple 1 : pourcentages de germination.

Souches	Doses			
	40	50	60	70
1	76,0 78,4 77,0	52,2 48,9 52,0	33,3 24,0 42,0	13,0 13,1 13,7
2	76,6 86,3 71,4	65,3 67,3 42,2	23,4 42,0 24,0	12,2 12,8
3	73,4 82,7 86,1	71,4 51,0	42,0 35,4 31,2	25,0 24,0 16,6

Données de l'exemple 2 : gains de poids journaliers (g).

Sexes	Alimentations		
	1	2	3
Mâle	728 685 625 721	636 697 662 691	686 648 605 602
Femelle	610 513 496 472 518 566	486 571 508 428 606 518	567 543 523 600 567 576