Analyse de Variance

Analyse de Variance à 1 Facteur

L'objectif de l'analyse de variance à 1 facteur est de tester l'égalité des moyennes théoriques d'une variable quantitative de différents groupes ou de différents niveaux du facteur considéré.

Les observations sont sous la forme :

Groupe 1	.....	Groupe K
$x_{1,1}$	.....	$x_{K,1}$
. . . . .	. . . . .	. . . . .
x_{1,n_1}	.....	x_{K,n_K}
T_1		T_K
$m_1 = \frac{T_1}{n_1}$		$m_K = \frac{T_K}{n_K}$

$$N = \sum_{i=1}^{K} n_i \qquad N = \sum_{i=1}^{K} T_i$$

Le modèle de cette analyse est le suivant :

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

Les hypothèses testées :

H_0 : égalité des moyennes, les groupes sont homogènes ou tous les α_i prennent pour valeur 0

H_1 : les groupes ne sont pas homogènes ou au moins un des α_i est différent des autres

Les conditions d'applications :

Tous les x_{ij} suivent des lois Normales de même variance σ^2 (estimée par s_e^2), ou, ce qui est identique, les e_{ij} sont "Normaux", indépendants et de même variance σ^2 (estimée par s_e^2).

Le tableau d'Analyse de Variance est le suivant :

Origine	**Σ des carrés** (a)	**ddl** (b)	**Variance** (a)/(b)	**F**
Intergroupe	$\sum_{i=1}^{K} \frac{T_i^2}{n_i} - \frac{T_g^2}{N}$	K-1	S^2	$\frac{S^2}{s_e^2}$
Intragroupe	$\sum_{i=1}^{K} \sum_{j=1}^{n_i} x_{ji}^2 - \sum_{i=1}^{K} \frac{T_i^2}{n_i}$	N-K	s_e^2	
Totale	$\sum_{i=1}^{K} \sum_{j=1}^{n_i} x_{ji}^2 - \frac{T_g^2}{N}$	N-1		

Un programme d'ANOVA avec MATLAB :

```
function [F,F0,Pvalue] = ANOVA1(x,k);
% [F,F0,Pvalue] = ANOVA1(x,k)
%
% ANOVA oneway (10/01/2001)
%    x : data matrix
%    k : number of observation by groupe
%    soit x est une matrice de "size(n K)", avec K groupe et n
%           observations par groupe (ie chaque groupe dans une colonne)
%    soit x est un vecteur et k indique la repartition des observations
%    dans les K groupes  (e.g. k(1) observation pour le premier groupe
%                         k(length(k)) observation pour le Keme groupe)
%    F : F value
%    F0 : Fseuil pour alpha = 0.05
%    Pvalue
%
%   ATTENTION : Dans le cas de decomposition orthogonale de la variance,
%   =========   il faudrait utiliser "la variance intra" calculee avec
%               toutes les datas => MODIFICATION de la FONCTION de
%               facon a introduire "varintra" et "df2"
%

N = length(x(:));
[n K] = size(x);
somx2 = sum(x(:).^2);

if K==1
   n = k;
   K = length(n);
   na = 1;
   nb = 0;
   for i=1:K
      if i>1
           na = na + n(i-1);
      end
        nb = nb + n(i);
      Ti(i) = sum(x(na:nb));
   end
   Tg = sum(Ti);
   somCinter = sum(Ti.^2./n) - Tg^2/N;
   somCintra = somx2 - sum(Ti.^2./n);
else
   [n K] = size(x);
   Ti = sum(x);
   Tg = sum(Ti);
   somCinter = sum(Ti.^2/n) - Tg^2/N;
   somCintra = somx2 - sum(Ti.^2/n);
end
somCtotal = somx2 - Tg^2/N;

%
% somCintra = -999 si varintra est donnee
% df2 = N-K ou df2 = df2
%

df1 = K-1;
df2 = N-K;

varinter = somCinter / df1;
varintra = somCintra / df2;
vartotal = somCtotal / (N-1);

F = varinter/varintra;
F0 = qf(0.95,df1,df2);
Pvalue = 1-pf(F,df1,df2);
```

```
clc;
fprintf('\n\n');
fprintf('===========================\n');
fprintf(' Analysis of Variance Table \n');
fprintf('==============================================================\n');
fprintf(' Source     d.f.     Sums of Squares     Variances       F\n');
fprintf('--------------------------------------------------------------\n');
fprintf(' Inter      %2.0f       %8.2f           %8.2f
%6.3f\n',df1,somCinter,varinter,F);
fprintf(' Intra      %2.0f       %8.2f           %8.2f\n',df2,somCintra,varintra
);
fprintf(' Total      %2.0f       %8.2f           %8.2f\n',N-
1,somCtotal,vartotal);
fprintf('=====================================================\n');
fprintf('  F         Fseuil      Pvalue\n');
fprintf('-----------------------------------\n');
fprintf(' %6.3f    %6.3f      %4.7f\n',F,F0,Pvalue);
fprintf('===================================\n\n\n');
```

Exercices :

```
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%
% ===================
% ANALYSE DE VARIANCE
% ===================
%
%
%
% FONCTIONS UTILISEES :
% =====================
% ANOVA1.m   Bartlett.m
%
% UTILISER "help name_function" pour obtenir des informations
% sur les fonctions utilisees
%
%
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%
% Exercice 1 :
% ------------
% On mesure le nombre de colonies bactériennes en fonction de 5 milieux de
% culture différents (k=5). Pour chaque milieu de culture on a fait 10
% comptages (ni=10). Les caractéristiques des milieux de culture ont-ils une
% influence sur le nombre de bactéries qui se développent ? Répondre à la
% question précédente avec un seul test statistique.
%   A    B    C    D     E
%   10   11   7    12    7
%   12   18   14   9     6
%   8    12   10   11    10
%   10   15   11   10    7
%   6    13   9    7     7
%   13   8    10   8     5
%   9    15   9    13    6
%   10   16   11   14    7
%   8    9    7    10    9
%   9    13   9    11    6
%
%
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
```

```
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%
% Exercice 2 :
% ------------
% Dans une étude sur la pollution par insecticides dans les grands lacs
% américains, on a suivi l'évolution de la concentration de pesticides
% contenue dans des poissons carnivores (perches, sandres et brochets).
% Les résultats exprimés en mg/g (DDT, DDD et DDE) en fonction de l'âge des
% poissons sont les suivants :
%    2 ans    3 ans    4 ans    5 ans    6 ans
%    0.144    0.285    0.418    0.675    1.13
%    0.171    0.295    0.441    0.685    1.18
%    0.178    0.321    0.451    0.726    1.21
%    0.184    0.354    0.451    0.736    1.23
%    0.193    0.359    0.458
%    0.197    0.361    0.461
%    0.198    0.362    0.464
%    0.199    0.364    0.465
%    0.199    0.373    0.465
%    0.206    0.382    0.469
%    0.216    0.403    0.475
%    0.258    0.407    0.480
%             0.413
%
% La quantité moyenne de pesticides (DDE+DDD+DDT) varie-t-elle selon l'âge
% des poissons ?
%
%
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%
% Exercice 3 :
% ------------
% Une étude a porté sur l'effet mutagène de 6 substances (labellées de A à
% F). 54 cultures cellulaires ont été répartis par tirage au sort dans les
% 6 groupes et pour chaque groupe on a donc effectué 9 fois la même
% expérience. Les chiffres du tableau suivant ont été obtenus :
%      A     B     C     D     E     F
%     -----------------------------------
%      32.   39.   38.   39.   34.   25.
%      34.   32.   39.   38.   32.   32.
%      33.   47.   48.   45.   33.   28.
%      32.   38.   48.   45.   33.   33.
%      34.   49.   47.   38.   40.   25.
%      41.   47.   42.   34.   33.   23.
%      34.   33.   38.   38.   35.   25.
%      38.   37.   39.   38.   38.   28.
%      42.   48.   42.   36.   41.   31.
%     ---------------------------------------------------
%      320   370   381   351   319   250  (somme des x)
%      11406 15570 16275 13799 11397 7046 (somme des x^2)
%
% Analyser les résultats obtenus en regardant si les 6 substances on un
% effet identique. Après cette première analyse on s'est aperçu que l'on
% avait en fait, analyser l'effet de 2 substances X1 et X2. D'après le
% cahier d'expérience, la substance X1 aurait été utiliser pour les groupes
% A et E et la substance X2 pour les groupes B, C, D. Dans le cas F on
% n'aurait utiliser aucune substance. Ces résultats permettent ils de
% retrouver des effets homogènes des substances X1 et X2 ? Les substances
% X1 et X2 ont elles des effets identiques ?
%
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
```

Chapitre 16

Analyse de la variance

OBJECTIF DE L'ANALYSE DE VARIANCE

Au Chap. 8, nous avons utilisé la théorie de l'échantillonnage pour tester la signification des différences entre deux moyennes observées à partir d'échantillons. On supposait que les deux échantillons étaient tirés de populations qui avaient la même variance. Dans de nombreux cas, il est nécessaire de tester la signification de différences entre trois moyennes observées ou plus, ou autrement dit, tester l'hypothèse nulle selon laquelle les moyennes observées sont égales.

EXEMPLE 1. On suppose que dans une expérience en agriculture, quatre traitements chimiques différents du sol ont donné lieu respectivement à une production de 28, 22, 18 et 24 boisseaux par acre. Existe-il une différence significative entre ces moyennes ou la diversité de la production est-elle seulement due au hasard ?

On peut résoudre des problèmes de ce type avec une méthode importante connue sous le nom d'*analyse de variance* et développée par Fisher. Elle utilise la distribution F déjà présentée au Chap. 11.

EXPÉRIENCES UNIFACTORIELLES

Dans une *expérience à un seul facteur*, les mesures (ou observations) sont obtenues pour des groupes d'échantillons indépendants, où le nombre de mesures dans chaque groupe est b. On parle de *traitements*, chacun ayant b *répétitions*. Dans l'Exemple 1, $a = 4$.

Les résultats d'une expérience à un facteur peuvent être présentés dans un tableau à a lignes et b colonnes, comme le montre le Tableau 16.1. Ici X_{jk} est la mesure de la ligne j et la colonne k, où $j = 1, 2, 3, \ldots, a$ et où $k = 1, 2, 3, \ldots, b$. Par exemple, X_{35} désigne la cinquième mesure du troisième traitement.

Tableau 16.1

Traitement 1	$X_{11}, X_{12}, \ldots, X_{1b}$	$\bar{X}_{1.}$
Traitement 2	$X_{21}, X_{22}, \ldots, X_{2b}$	$\bar{X}_{2.}$
⋮	⋮	⋮
Traitement a	$X_{a1}, X_{a2}, \ldots, X_{ab}$	$\bar{X}_{a.}$

On notera $\bar{X}_{j.}$ la moyenne des mesures de la j^e ligne. On a

$$\bar{X}_{j.} = \frac{1}{b}\sum_{k=1}^{b} X_{jk} \qquad j = 1, 2, \ldots, a \tag{1}$$

Le point de $\bar{X}_{j.}$ indique que l'on a fait la somme sur l'indice k. Les valeurs $\bar{X}_{j.}$ sont appelées *moyennes de groupe, moyennes des traitements,* ou *moyennes de ligne*. La *moyenne totale,* ou *moyenne globale,* est la moyenne de toutes les mesures de tous les groupes et est notée $\bar{X}$:

$$\bar{X} = \frac{1}{ab} \sum_{j=1}^{a} \sum_{k=1}^{b} X_{jk} \tag{2}$$

ÉCART TOTAL, ÉCART INTRA-TRAITEMENT, ET ÉCART ENTRE TRAITEMENTS

On définit l'*écart total*, noté V, comme la somme des carrés des écarts de chaque mesure à la moyenne totale $\bar{X}$

$$\text{Écart total} = V = \sum_{j,k} (X_{jk} - \bar{X})^2 \tag{3}$$

En écrivant l'identité

$$X_{jk} - \bar{X} = (X_{jk} - \bar{X}_{j.}) + (\bar{X}_{j.} - \bar{X}) \tag{4}$$

en mettant au carré, et en sommant sur j et k, on a (voir Problème 16.1)

$$\sum_{j,k} (X_{jk} - \bar{X})^2 = \sum_{j,k} (X_{jk} - \bar{X}_{j.})^2 + \sum_{j,k} (\bar{X}_{j.} - \bar{X})^2 \tag{5}$$

ou

$$\sum_{j,k} (X_{jk} - \bar{X})^2 = \sum_{j,k} (X_{jk} - \bar{X}_{j.})^2 + b \sum_{j} (\bar{X}_{j.} - \bar{X})^2 \tag{6}$$

La première somme de la partie droite des Éqs (*5*) et (*6*) est appelée *écart intra-traitements* (puisqu'elle implique les carrés des écarts de X_{jk} aux moyennes des traitements $\bar{X}_{j.}$) et notée V_W. Ainsi

$$V_W = \sum_{j,k} (X_{jk} - \bar{X}_{j.})^2 \tag{7}$$

La deuxième somme de la partie droite des Éqs (*5*) et (*6*) est appelée *écart entre traitements* (puisqu'elle implique les carrés des écarts des différentes moyennes de traitements $\bar{X}_{j.}$ à la moyenne totale $\bar{X}$) et notée V_B. Ainsi

$$V_B = \sum_{j,k} (\bar{X}_{j.} - \bar{X})^2 = b \sum_{j} (\bar{X}_{j} - \bar{X})^2 \tag{8}$$

Les Éqs (*5*) et (*6*) peuvent donc s'écrire

$$V = V_W + V_B \tag{9}$$

MÉTHODES RAPIDES DE CALCUL DES ÉCARTS

Les formules suivantes sont utiles pour minimiser le travail de calcul des écarts ci-dessus :

$$V = \sum_{j,k} X_{jk}^2 - \frac{T^2}{ab} \tag{10}$$

$$V_B = \frac{1}{b} \sum_{j} T_{j.}^2 - \frac{T^2}{ab} \tag{11}$$

$$V_W = V - V_B \tag{12}$$

où T est le total de toutes les valeurs X_{jk} et $T_{j.}$ est le total de toutes les valeurs dans le j^e traitement :

$$T = \sum_{j,k} X_{jk} \qquad T_{j.} = \sum_{k} X_{jk} \tag{13}$$

En pratique, il est commode de soustraire toutes les données du tableau par une valeur déterminée pour simplifier les calculs ; cela n'a pas d'effet sur les résultats finals.

MODÈLE MATHÉMATIQUE POUR L'ANALYSE DE VARIANCE

On peut considérer que chaque ligne du Tableau 16.1 est une variable aléatoire de taille b tirée de la population pour le traitement donné. Pour le j^e traitement, les X_{jk} différeront de la moyenne μ_j de la population d'une *erreur aléatoire*, que l'on note ε_{jk} ; ainsi

$$X_{jk} = \mu_j + \varepsilon_{jk} \tag{14}$$

On suppose que ces erreurs sont distribuées normalement avec une moyenne 0 et une variance σ^2. Si μ est la moyenne de la population pour tous les traitements et si l'on pose $\alpha_j = \mu_j - \mu$, si bien que $\mu_j = \mu + \alpha_j$, alors l'Éq. (*14*) devient

$$X_{jk} = \mu + \alpha_j + \varepsilon_{jk} \tag{15}$$

où $\Sigma_j\, \alpha_j = 0$ (voir Problème 16.9). À partir de l'Éq. (*15*) et sous l'hypothèse que les ε_{jk} sont distribués normalement avec une moyenne 0 et une variance σ^2, on conclut que les X_{jk} peuvent être considérés comme des variables aléatoires distribuées normalement avec une moyenne μ et une variance σ^2.

L'hypothèse nulle selon laquelle les moyennes des traitements sont égales s'écrit ($H_0 : \alpha_j = 0 ; j = 1, 2, \ldots, a$) ou, de la même manière ($H_0 : \mu_j = \mu ; j = 1, 2, \ldots, a$). Si H_0 est vraie, les populations des traitements auront toutes la même distribution normale (i.e. avec la même moyenne et la même variance). Dans ce cas, il n'existe qu'une seule population traitement (i.e. tous les traitements sont statistiquement identiques) ; en d'autres termes, il n'y a pas de différence significative entre les traitements.

VALEURS ATTENDUES DES ÉCARTS

On peut montrer (voir Problème 16.10) que les valeurs attendues de V_W, V_B et V sont données par

$$E(V_W) = a(b-1)\sigma^2 \tag{16}$$

$$E(V_B) = (a-1)\sigma^2 + b\sum_j \alpha_j^2 \tag{17}$$

$$E(V) = (ab-1)\sigma^2 + b\sum_j \alpha_j^2 \tag{18}$$

À partir de l'Éq. (*16*), on obtient

$$E\left[\frac{V_W}{a(b-1)}\right] = \sigma^2 \tag{19}$$

si bien que

$$\hat{S}_W^2 = \frac{V_W}{a(b-1)} \tag{20}$$

est toujours la meilleure estimation (non biaisée) de σ^2 que H_0 soit vraie ou non. D'un autre côté, on voit, d'après les Éqs (*16*) et (*18*), que si H_0 est vraie (i.e. $\alpha_j = 0$) on aura

$$E\left(\frac{V_B}{a-1}\right) = \sigma^2 \text{ et } E\left(\frac{V}{ab-1}\right) = \sigma^2 \tag{21}$$

si bien que

$$\hat{S}_B^2 = \frac{V_B}{a-1} \text{ et } \hat{S}^2 = \frac{V}{ab-1} \tag{22}$$

ne fournissent des estimations non biaisées de σ^2 que dans ce cas. Cependant, si H_0 n'est pas vraie, on obtient d'après l'Éq. (*16*)

$$E(\hat{S}_B^2) = \sigma^2 + \frac{b}{a-1}\sum_j \alpha_j^2 \tag{23}$$

DISTRIBUTIONS DES ÉCARTS

En utilisant la propriété d'additivité du chi-deux (page 264), on peut démontrer les théorèmes fondamentaux suivants sur les distributions des écarts V_W, V_B, et V.

Théorème 1 : V_W/σ^2 a une distribution du chi-carré avec $a(b-1)$ degrés de liberté.

Théorème 2 : Sous l'hypothèse nulle H_0, V_B/σ^2 et V/σ^2 ont une distribution du chi-carré avec respectivement $a-1$ et $ab-1$ degrés de liberté.

Il est important de réaliser que le théorème 1 est valide que H_0 soit vraie ou non, tandis que le théorème 2 n'est valide que si H_0 est vraie.

LE TEST *F* DE L'HYPOTHÈSE NULLE DE L'ÉGALITÉ DES MOYENNES

Si l'hypothèse nulle H_0 n'est pas vraie (i.e. si les moyennes de traitements ne sont pas égales), on tire de l'Éq. (*23*) que l'on peut s'attendre à ce que $\hat{S}_B^2$ soit supérieur à σ^2, et que l'effet s'accentue en même temps que la dispersion entre les moyennes augmente. D'un autre côté, d'après les Éqs (*19*) et (*20*), on peut s'attendre à ce que $\hat{S}_W^2$ soit égal à σ^2, que les moyennes soient égales ou non. Il s'ensuit qu'une statistique correcte pour tester une hypothèse H_0 est fournie par $\hat{S}_B^2 / \hat{S}_W^2$. Si la statistique est significativement grande, on peut conclure qu'il existe une différence significative entre les moyennes des traitements et l'on peut donc rejeter l'hypothèse H_0 ; dans le cas contraire, on peut soit accepter H_0, soit réserver son jugement, en attendant une analyse complémentaire.

Afin d'utiliser la statistique $\hat{S}_B^2 / \hat{S}_W^2$, on doit connaître sa distribution d'échantillonnage. Elle est donnée dans le théorème 3.

Théorème 3 : La statistique $F = \hat{S}_B^2 / \hat{S}_W^2$ a une distribution F avec $a-1$ et $a(b-1)$ degrés de liberté.

Le Théorème 3 nous permet de tester l'hypothèse nulle avec un degré de signification spécifié en utilisant un test unilatéral de la distribution F (discuté au Chap. 11).

TABLEAUX D'ANALYSE DE VARIANCE

Les calculs requis pour le test ci-dessus sont résumés dans le Tableau 16.2, que l'on appelle *tableau d'analyse de variance*. En pratique, on calculerait V et V_B en utilisant aussi bien par la méthode longue [Éqs (*3*) et (*8*)] que par la méthode courte [Éqs (*10*) et (*11*)] et en calculant alors $V_W = V - V_B$. Il faut remarquer que les degrés de liberté de l'écart total (i.e. $ab-1$) sont égaux à la somme des degrés de liberté des écarts entre traitements et intra-traitements.

Tableau 16.2

Variance	Degrés de liberté	Écart	F
Entre traitements $V_B = b\sum_j (\overline{X}_{j.} - \overline{X})^2$	$a-1$	$\hat{S}_B^2 = \dfrac{V_B}{a-1}$	$\dfrac{\hat{S}_B^2}{\hat{S}_W^2}$ avec $a-1$ et $a(b-1)$ degrés de liberté
Intra-traitements $V_W = V - V_B$	$a(b-1)$	$\hat{S}_W^2 = \dfrac{V_W}{a(b-1)}$	
Total $V = V_B + V_W = \sum_{j,k} (X_{jk} - \overline{X})^2$	$ab-1$		

MODIFICATIONS POUR UN NOMBRE INÉGAL D'OBSERVATIONS

Dans le cas ou les traitements 1, ..., a ont différents nombres d'observations – respectivement égaux à N_1, ..., N_a – les résultats sont facilement modifiés. Ainsi on obtient

$$V = \sum_{j,k} (X_{jk} - \overline{X})^2 = \sum_{j,k} X_{jk}^2 - \frac{T^2}{N} \tag{24}$$

$$V_B = \sum_{j,k} (\overline{X}_{j.} - \overline{X})^2 = \sum_j N_j (\overline{X}_{j.} - \overline{X})^2 = \sum_j \frac{T_{j.}^2}{N_j} - \frac{T^2}{N} \tag{25}$$

$$V_W = V - V_B \tag{26}$$

où $\Sigma_{j,k}$ désigne la somme sur k de 1 à N_j puis la somme sur j de 1 à a. Le Tableau16.3 présente l'analyse de variance pour ce cas.

Tableau 16.3

Variance	Degrés de liberté	Écart	F
Entre traitements $V_B = \sum_j N_j (\overline{X}_{j.} - \overline{X})^2$	$a-1$	$\hat{S}_B^2 = \dfrac{V_B}{a-1}$	$\dfrac{\hat{S}_B^2}{\hat{S}_W^2}$ avec $a-1$ et $N-a$ degrés de liberté
Intra-traitements $V_W = V - V_B$	$N-a$	$\hat{S}_W^2 = \dfrac{V_W}{N-a}$	
Total $V = V_B + V_W = \sum_{j,k} (X_{jk} - \overline{X})^2$	$N-1$		

EXPÉRIENCES À DEUX FACTEURS

Les principes de l'analyse de variance à un facteur peuvent être généralisés. L'Exemple 2 illustre la procédure pour *l'expérience à deux facteurs*.

EXEMPLE 2. On suppose qu'une expérience agricole consiste à examiner les rendements par hectare de 4 différents types de blé, où chaque variété est cultivée sur 5 sols différents. Ainsi un total de (4)(5) = 20 terrains sont nécessaires. Il est plus pratique dans un tel cas d'arranger les terrains par *blocs*, disons 4 terrains par blocs, avec une variété différente de blé pour chaque terrain dans un bloc. Donc 5 blocs seraient nécessaires dans ce cas.

Dans ce cas, il y a deux facteurs puisqu'il peut y avoir des différences de rendement par hectare dues à (*a*) le type particulier de blé cultivé ou (*b*) le bloc considéré (qui peut avoir une fertilité différente, etc.).

Par analogie avec l'expérience agricole de l'Exemple 2, on désigne souvent ces deux facteurs de l'expérience comme *traitements* et *blocs*, mais on pourrait évidemment s'y référer comme le facteur 1 et le facteur 2.

NOTATION DANS LE CAS D'EXPÉRIENCES À DEUX FACTEURS

En supposant que l'on possède a traitements et b blocs, on construit le Tableau 16.4, où l'on suppose qu'il n'y a qu'une seule valeur expérimentale (comme le rendement par hectare) pour chaque traitement et bloc. Pour le traitement j et le bloc k, on note la valeur par X_{jk}. La moyenne des valeurs de la j^e ligne est notée par $\overline{X}_{j.}$, où $j = 1, \ldots, a$, tandis que la moyenne des valeurs de la k^e colonne est notée $\overline{X}_{.k}$, où $k = 1, \ldots, b$. La moyenne globale, ou générale, est notée par $\overline{X}$. En symboles,

$$\overline{X}_{j.} = \frac{1}{b}\sum_{k=1}^{b} X_{jk} \quad \overline{X}_{.k} = \frac{1}{a}\sum_{j=1}^{a} X_{jk} \quad \overline{X} = \frac{1}{ab}\sum_{j,k} X_{jk} \tag{27}$$

Tableau 16.4

	Bloc				
	1	2	⋯	b	
Traitement 1	X_{11}	X_{12}	⋯	X_{1b}	$\overline{X}_{1.}$
Traitement 2	X_{21}	X_{22}	⋯	X_{2b}	$\overline{X}_{2.}$
⋮	⋮	⋮	⋮	⋮	⋮
Traitement a	X_{a1}	X_{a2}	⋯	X_{ab}	$\overline{X}_{a.}$
	$\overline{X}_{.1}$	$\overline{X}_{.2}$		$\overline{X}_{.b}$	

ÉCARTS DANS LE CAS D'EXPÉRIENCES À DEUX FACTEURS

Comme dans le cas des expériences à un facteur, on peut définir les écarts pour des expériences à deux facteurs. On définit d'abord l'*écart total*, comme dans l'Éq. (*3*),

$$V = \sum_{j,k}(X_{jk} - \overline{X})^2 \tag{28}$$

En écrivant l'identité,

$$X_{jk} - \overline{X} = (X_{jk} - \overline{X}_{j.} - \overline{X}_{.k} + \overline{X}) + (\overline{X}_{j.} - \overline{X}) + (\overline{X}_{.k} - \overline{X}) \tag{29}$$

en mettant au carré, et en sommant sur j et k, on peut montrer que

$$V = V_E + V_R + V_C \tag{30}$$

où V_E = écart dû à l'erreur ou au hasard $= \sum_{j,k}(X_{jk} - \overline{X}_{j.} - \overline{X}_{.k} + \overline{X})^2$

V_R = écart entre lignes (traitements) $= b\sum_{j=1}^{a}(\overline{X}_{j.} - \overline{X})^2$

V_C = écart entre colonnes (blocs) $= a\sum_{k=1}^{b}(\overline{X}_{.k} - \overline{X})^2$

L'écart dû à l'erreur ou au hasard est aussi appelé l'*écart résiduel* ou *écart aléatoire*.

Les équations suivantes, analogues aux précédentes (*10*) à (*12*), sont les formules de calculs condensées :

$$V = \sum_{jk} X_{jk}^2 - \frac{T^2}{ab} \tag{31}$$

$$V_R = \frac{1}{b}\sum_{j=1}^{a} T_{j.}^2 - \frac{T^2}{ab} \tag{32}$$

$$V_C = \frac{1}{a}\sum_{k=1}^{b} T_{.k}^2 - \frac{T^2}{ab} \tag{33}$$

$$V_E = V - V_R - V_C \tag{34}$$

où $T_{j.}$ est le nombre total de cellules de la j^e ligne, $T_{.k}$ est le nombre total de cellules de la k^e colonne, et T représente le nombre total de cellules.

ANALYSE DE VARIANCE DANS LE CAS D'EXPÉRIENCES À DEUX FACTEURS

La généralisation du modèle mathématique des expériences à un facteur est donnée par l'Éq. (*15*) et nous conduit à considérer que pour les expériences à deux facteurs

$$X_{jk} = \mu + \alpha_j + \beta_k + \varepsilon_{jk} \tag{35}$$

où $\Sigma\alpha_j = 0$ et $\Sigma\beta_k = 0$. Ici, μ est la moyenne théorique, α_j est la partie de X_{jk} due aux différents traitements (appelés quelquefois les *effets traitements*), β_k est la partie de X_{jk} due aux différents blocs (appelés quelquefois les *effets blocs*), et ε_{jk} est la partie de X_{jk} due au hasard ou à l'erreur. Comme précédemment, on suppose que les ε_{jk} sont distribués normalement avec une moyenne de 0 et une variance σ^2, si bien que les X_{jk} sont aussi distribués normalement avec une moyenne μ et une variance σ^2.

D'après les résultats (*16*) à (*18*), on peut démontrer que les espérances des écarts sont données par

$$E(V_E) = (a-1)(b-1)\sigma^2 \tag{36}$$

$$E(V_R) = (a-1)\sigma^2 + b\sum_j \alpha_j^2 \tag{37}$$

$$E(V_C) = (b-1)\sigma^2 + a\sum_k \beta_k^2 \tag{38}$$

$$E(V) = (ab-1)\sigma^2 + b\sum_j \alpha_j^2 + a\sum_k \beta_k^2 \tag{39}$$

Il y a deux hypothèse nulles que l'on voudrait tester :

$H_0^{(1)}$: toutes les moyennes des traitements (ligne) sont égales ; c'est-à-dire que $\alpha_j = 0$ et $j = 1, \ldots, a$.
$H_0^{(2)}$: toutes les moyennes des blocs (colonne) sont égales ; c'est-à-dire que $\beta_k = 0$ et $k = 1, \ldots, b$.

On voit d'après l'Éq. (*38*) que, sans considérer $H_0^{(1)}$ ou $H_0^{(2)}$, le meilleur estimateur (non biaisé) de σ^2 est fourni par

$$\hat{S}_E^2 = \frac{V_E}{(a-1)(b-1)} \quad \text{c'est-à-dire } E(\hat{S}_E^2) = \sigma^2 \tag{40}$$

Si les hypothèses $H_0^{(1)}$ et $H_0^{(2)}$ sont vraies, il découle aussi que

$$\hat{S}_R^2 = \frac{V_R}{a-1} \qquad \hat{S}_C^2 = \frac{V_C}{b-1} \qquad \hat{S}^2 = \frac{V}{ab-1} \tag{41}$$

seront des estimations non biaisées de σ^2. Si $H_0^{(1)}$ et $H_0^{(2)}$ ne sont pas vraies, les Éqs (*36*) et (37), donneront respectivement

$$E(\hat{S}_R^2) = \sigma^2 + \frac{b}{a-1}\sum_j \alpha_j^2 \tag{42}$$

$$E(\hat{S}_C^2) = \sigma^2 + \frac{a}{b-1}\sum_k \beta_k^2 \tag{43}$$

Les théorèmes suivants sont similaires aux Théorèmes 1 et 2 :

> **Théorème 4 :** V_E/σ^2 a une distribution du chi-carré avec $(a-1)(b-1)$ degrés de liberté, quelles que soient $H_0^{(1)}$ et $H_0^{(2)}$.

> **Théorème 5 :** Sous l'hypothèse nulle $H_0^{(1)}$, V_R/σ^2 a une distribution du chi-carré avec $a-1$ degrés de liberté. Sous l'hypothèse nulle $H_0^{(2)}$, V_C/σ^2 a une distribution du chi-carré avec $b-1$ degrés de liberté. Sous les deux hypothèses $H_0^{(1)}$ et $H_0^{(2)}$, V/σ^2 a une distribution du chi-carré avec $ab-1$ degrés de liberté.

Pour tester l'hypothèse $H_0^{(1)}$, il est naturel de considérer la statistique $\hat{S}_R^2/\hat{S}_E^2$ car on peut voir d'après l'Éq. (*42*) que $\hat{S}_R^2$ sera statistiquement différent de σ^2 si les moyennes des lignes (traitement) sont statistiquement différentes. De la même façon, pour tester l'hypothèse $H_0^{(2)}$, on considère la statistique $\hat{S}_C^2/\hat{S}_E^2$. Les distributions de $\hat{S}_R^2/\hat{S}_E^2$ et $\hat{S}_C^2/\hat{S}_E^2$ sont données dans le Théorème 6, qui est l'analogue du Théorème 3.

> **Théorème 6** : Sous l'hypothèse nulle $H_0^{(1)}$, la statistique $\hat{S}_R^2/\hat{S}_E^2$ a une distribution F avec $a-1$ et $(a-1)(b-1)$ degrés de liberté. Sous l'hypothèse nulle $H_0^{(2)}$, la statistique $\hat{S}_C^2/\hat{S}_E^2$ a une distribution F avec $b-1$ et $(a-1)(b-1)$ degrés de liberté.

Le Théorème 6 nous permet d'accepter ou de rejeter $H_0^{(1)}$ ou $H_0^{(2)}$ à des seuils de signification spécifiés. Par commodité, comme dans le cas de l'analyse à un facteur, un tableau d'analyse de variance peut être construit comme dans le Tableau 16.5.

EXPÉRIENCE À DEUX FACTEURS AVEC RÉPÉTITIONS

Dans le Tableau 16.4, il n'y a qu'une seule entrée pour chaque traitement et chaque bloc. On peut souvent obtenir plus d'information sur les facteurs par *répétition* de l'expérience. Dans un tel cas, il y aura plus d'une entrée pour un traitement et un bloc donnés. On supposera qu'il y a c entrées pour chaque position ; on peut apporter des changements appropriés quand les nombres de répétitions ne sont pas égaux.

À cause des répétitions, il faut utiliser un modèle approprié pour remplacer celui de l'Éq. (*35*). On utilise

$$X_{jkl} = \mu + \alpha_j + \beta_k + \gamma_{jk} + \varepsilon_{jkl} \tag{44}$$

où les indices j, k et l de X_{jkl} correspondent respectivement à la ligne j (ou traitement), la colonne k (ou bloc), et la l^e répétition. Dans l'Éq. (*44*), les μ, α_j, et β_k sont définis comme précédemment ;

Tableau 16.5

Variance	Degrés de liberté	Écart	F
Entre traitements $V_R = b\sum_j (\overline{X}_{j.} - \overline{X})^2$	$a-1$	$\hat{S}_R^2 = \dfrac{V_R}{a-1}$	$\hat{S}_R^2 / \hat{S}_E^2$ avec $a-1$ et $(a-1)(b-1)$ degrés de liberté
Entre blocs $V_C = a\sum_k (\overline{X}_{.k} - \overline{X})^2$	$b-1$	$\hat{S}_C^2 = \dfrac{V_C}{b-1}$	$\hat{S}_C^2 / \hat{S}_E^2$ avec $b-1$ et $(a-1)(b-1)$ degrés de liberté
Résiduel ou aléatoire $V_E = V - V_R - V_C$	$(a-1)(b-1)$	$\hat{S}_E^2 = \dfrac{V_E}{(a-1)(b-1)}$	
Total $V = V_R + V_C + V_E = \sum_{j,k} (X_{jk} - \overline{X})^2$	$ab-1$		

ε_{jkl} est le terme d'erreur ou aléatoire, tandis que le γ_{jk} désigne les *effets d'interaction* lignes-colonnes (ou traitements-blocs), appelés simplement *interactions*. On a les restrictions

$$\sum_j \alpha_j = 0 \qquad \sum_k \beta_k = 0 \qquad \sum_j \gamma_{jk} = 0 \qquad \sum_k \gamma_{jk} = 0 \tag{45}$$

et les X_{jkl} sont supposés être distribués normalement avec une moyenne μ et une variance σ^2.

Comme précédemment, l'écart total V de l'ensemble des données peut être divisé en plusieurs écarts dus aux lignes V_R, aux colonnes V_C, l'interaction V_I et l'erreur résiduelle ou aléatoire V_E.

$$V = V_R + V_C + V_I + V_E \tag{46}$$

où

$$V = \sum_{j,k,l} (X_{jkl} - \overline{X})^2 \tag{47}$$

$$V_R = bc \sum_{j=1}^{a} (\overline{X}_{j..} - \overline{X})^2 \tag{48}$$

$$V_C = ac \sum_{k=1}^{b} (\overline{X}_{.k.} - \overline{X})^2 \tag{49}$$

$$V_I = c\sum_{j,k} (\overline{X}_{jk.} - \overline{X}_{j..} - \overline{X}_{.k.} + \overline{X})^2 \tag{50}$$

$$V_E = \sum_{j,k,l} (X_{jkl} - \overline{X}_{jk.})^2 \tag{51}$$

Dans ces résultats les points en indices ont la même signification que celle donnée auparavant (page 362) ; ainsi

$$\overline{X}_{j..} = \frac{1}{bc}\sum_{k,l} X_{jkl} = \frac{1}{b}\sum_{k} \overline{X}_{jk.} \qquad (52)$$

Les valeurs attendues des écarts peuvent être obtenues comme précédemment. En utilisant les nombres de degrés de liberté pour chaque source d'écart, on peut mettre en place le tableau d'analyse de variance comme au Tableau 16.6. On peut utiliser les rapports F de la dernière colonne du Tableau 16.6 pour tester l'hypothèse nulle :

$H_0^{(1)}$: toutes les moyennes des traitements (ligne) sont égales ; c'est-à-dire que $\alpha_j = 0$.

$H_0^{(2)}$: toutes les moyennes des blocs (colonne) sont égales ; c'est-à-dire que $\beta_k = 0$.

$H_0^{(3)}$: il n'y a pas d'interaction entre traitements et blocs ; c'est-à-dire que $\gamma_{jk} = 0$.

Tableau 16.6

Variance	Degrés de liberté	Écart	F
Entre traitements V_R	$a-1$	$\hat{S}_R^2 = \dfrac{V_R}{a-1}$	$\hat{S}_R^2 / \hat{S}_E^2$ avec $a-1$ et $ab(c-1)$ degrés de liberté
Entre blocs V_C	$b-1$	$\hat{S}_C^2 = \dfrac{V_C}{b-1}$	$\hat{S}_C^2 / \hat{S}_E^2$ avec $b-1$ et $ab(c-1)$ degrés de liberté
Interaction V_I	$(a-1)(b-1)$	$\hat{S}_I^2 = \dfrac{V_I}{(a-1)(b-1)}$	$\hat{S}_I^2 / \hat{S}_E^2$ avec $(a-1)(b-1)$ et $ab(c-1)$ degrés de liberté
Résiduel ou aléatoire V_E	$ab(c-1)$	$\hat{S}_E^2 = \dfrac{V_E}{ab(c-1)}$	
Total, V	$abc-1$		

D'un point de vue pratique, on doit décider d'abord si l'hypothèse $H_0^{(3)}$ peut être rejetée ou non avec un degré de signification approprié en utilisant le rapport F $\hat{S}_I^2 / \hat{S}_E^2$ du Tableau 16.6. Deux cas sont à envisager :

1. **On ne peut pas rejeter $H_0^{(3)}$.** Dans ce cas, on peut conclure que les interactions ne sont pas trop importantes. On peut alors tester $H_0^{(1)}$ et $H_0^{(2)}$ en utilisant respectivement les rapports F $\hat{S}_R^2 / \hat{S}_E^2$ et $\hat{S}_C^2 / \hat{S}_E^2$, comme au Tableau 16.6. Quelques statisticiens recommandent dans ce cas de grouper les écarts en prenant le total $V_I + V_E$, en le divisant par le total correspondant aux degrés de liberté $(a-1)(b-1) + ab(c-1)$ et en utilisant cette valeur pour remplacer le dénominateur $\hat{S}_E^2$ dans le test F.

2. **On peut rejeter $H_0^{(3)}$.** Dans ce cas, on peut conclure que les interactions sont significativement grandes. Les différences des facteurs ne seront alors importantes que si elles sont grandes comparativement à de telles interactions. Pour cette raison, beaucoup de statisticiens recommandent que $H_0^{(1)}$ et $H_0^{(2)}$ soient testés en utilisant respectivement les rapports F $\hat{S}_R^2 / \hat{S}_I^2$ et $\hat{S}_C^2 / \hat{S}_I^2$, plutôt que ceux donnés dans le Tableau 16.6. Nous devrons aussi utiliser cette procédure alternative.

La méthode la plus facile de procéder à l'analyse de variance avec répétition est d'abord de totaliser les valeurs des répétitions qui correspondent aux traitements (lignes) et blocs (colonnes) par-

ticuliers. Cela donne un tableau à deux facteurs avec une simple entrée, qui peut être analysée comme dans le Tableau 16.5. Cette procédure est illustrée dans le Problème 16.6.

DÉMARCHE EXPÉRIMENTALE

Les techniques d'analyse de variance abordées ci-dessus sont employées après obtention des résultats d'une expérience. Cependant, afin d'avoir autant d'informations que possible, l'expérience doit être planifiée avec soin ; c'est ce que l'on désigne souvent comme le *plan d'expérience*. Les exemples suivants sont d'importantes applications de plans d'expérience :

1. **Randomisation complète.** On suppose que l'on a une expérience agricole comme dans l'Exemple 1. Pour mettre en place une telle expérience, on pourrait diviser le sol en $4 \times 4 = 16$ terrains (indiqués dans la Fig. 16-1 par des carrés) et affecter chaque traitement (indiqués par A, B, C et D) à quatre blocs de façon aléatoire. L'objectif de la randomisation est d'éliminer les différentes sources d'erreur, comme les différences de fertilité du sol.

D	A	C	C
B	D	B	A
D	C	B	D
A	B	C	A

Randomisation Complète

Fig. 16-1

I	C	B	A	D
II	A	B	D	C
III	B	C	D	A
IV	A	D	C	B

Randomisation par blocs

Fig. 16-2

D	B	C	A
B	D	A	C
C	A	D	B
A	C	B	D

Carré latin

Fig. 16-3

B_γ	A_β	D_δ	C_α
A_δ	B_α	C_γ	D_β
D_α	C_δ	B_β	A_γ
C_β	D_γ	A_α	B_δ

Carré gréco-latin

Fig. 16-4

2. **Randomisation par blocs.** Quand, comme dans l'Exemple 2, il est nécessaire d'avoir un jeu complet de traitements pour chaque bloc, les traitements A, B, C et D sont introduits de façon aléatoire dans chaque bloc : I, II, III, et IV (i.e. les lignes de la Fig. 16-2), et pour cette raison les blocs sont désignés comme des *blocs randomisés*. On utilise ce type de plan quand on veut contrôler *une source de variabilité ou d'erreur* : soit la différence entre les blocs.

3. **Carrés latins.** Pour certains objectifs, il est nécessaire de contrôler *deux sources d'erreurs ou de variabilité* en même temps, comme la différence entre les lignes et la différence entre les colonnes. Dans l'expérience de l'Exemple 1, les erreurs des différentes lignes et colonnes pourraient être dues aux changements de fertilité du sol dans différentes parties du pays. Dans un tel cas, il est préférable que chaque traitement apparaisse une fois dans chaque ligne et une fois dans chaque colonne, comme dans la Fig. 16-3. L'arrangement s'appelle le *carré latin* car les lettres latines A, B, C et D sont utilisées.

4. **Carrés gréco-latins.** S'il est nécessaire de contrôler trois sources d'erreurs ou de variabilité, on utilise un *carré gréco-latin*, comme le montre la Fig. 16-4. Un tel carré comprend essentiellement deux carrés latins superposés l'un sur l'autre, avec les lettres A, B, C et D pour l'un et α, β, γ, δ pour l'autre. La condition supplémentaire qui doit être rencontrée est que chaque lettre latine doit être utilisée une seule et unique fois avec chaque lettre grecque ; quand cette condition est rencontrée, le carré est dit *orthogonal*.