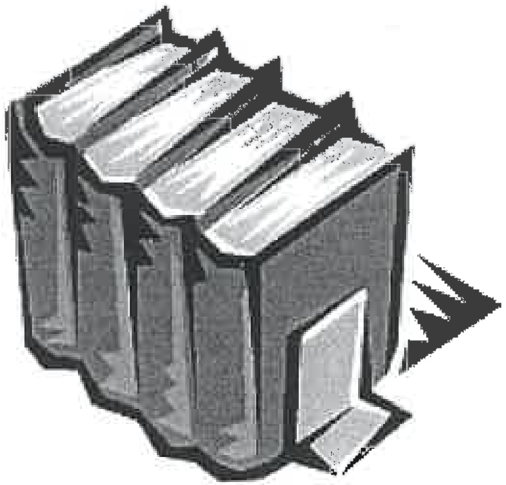
Alexia MONSAVOIR
06 23 42 24 30

RECHERCHE EXPERIMENTALE APPLIQUEE A L'AGRO-ALIMENTAIRE

LES CONNAISSANCES THEORIQUES



SOMMAIRE CONNAISSANCES THEORIQUES

CONNAISSANCES THEORIQUES

1 UN FACTEUR

Un facteur est une variable, ou un état, qui agit sur une mesure étudiée.

Exemples :
- un objet posé sur le plateau d'une balance de Roberval provoque une déviation de l'aiguille (l'objet = facteur; mesure = poids),
- une température de réaction peut stimuler une activité enzymatique (température = facteur; mesure = activité enzymatique),

Les facteurs d'une étude sont les paramètres susceptibles d'influer sur une réponse étudiée.

Leur nombre peut être élevé, ils sont quantitatifs ou qualitatifs et il faut pouvoir leur attribuer 2 états (2 niveaux). Le choix des facteurs se fait en fonction de la connaissance du phénomène qu'a l'expérimentateur mais aussi de son intuition. Les contraintes économiques sont aussi à prendre en considération.
L'expérimentateur devra établir une liste aussi exhaustive que possible des paramètres pouvant influer la mesure étudiée, mais il devra faire la distinction entre :
- les facteurs de l'étude notés X_j en variables codées ou U_j en variables naturelles.
- les facteurs contrôlables, paramètres à fixer, mais non étudiés,
- les facteurs non contrôlables mais identifiables, paramètres responsables de fluctuations aléatoires de la mesure étudiée Y.

2 UN PLAN EXPERIMENTAL

Un plan expérimental sera représenté par un tableau (matrice) à N lignes et k colonnes
Chaque colonne représente un facteur (température, pression, concentration d'une substance).
Chaque ligne correspond à une expérience (un essai) dans laquelle on fixe les niveaux des facteurs et sur laquelle on réalisera une mesure, appelée "réponse" notée Y.

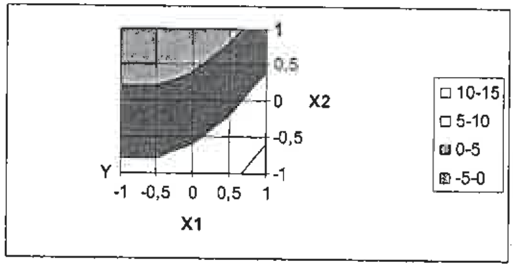
N° essai	Température (°C) = U_1	Pression (bar) = U_2	Concentration (g/l) = U_3
1	15	1	50
2	25	1	30
3	15	1,5	30
4	25	1,5	50

Réponse Y
Y1
Y2
Y3
Y4

Plan expérimental optimal : c'est l'ensemble des expériences le plus efficace qui donnera un ensemble de réponses avec l'incertitude la plus faible.

3 REPONSE Y

Une réponse notée Y est le résultat de l'expérience (déviation de l'aiguille dans une pesée, mesure d'une activité enzymatique). La réponse est quantitative (rendement, coût, nombre de colonies bactériennes, activité enzymatique, poids). C'est la variable dépendante ou variable à expliquer. Les variations de la réponse seront à expliquer par les variations des niveaux des facteurs expérimentaux.
On parle de courbes iso réponses quand on a plusieurs combinaisons des niveaux des facteurs pour une même réponse.



Ici la réponse Y est comprise entre 10 et 5 dans la zone jaune donc pour des combinaisons de X_1 = 0.5 et X2 = - 0.5 par exemple. La ligne qui sépare la zone verte de la zone jaune correspond aux conditions possibles pour X_1 et X_2 pour avoir une réponse Y = 10.

4 SIGNIFICATIVITE D'UN FACTEUR

Un facteur est significatif si, lorsqu'il est modifié, il entraîne une modification de la réponse. Un facteur est non significatif s'il n'a pas d'effet sur la réponse du système étudié ou si l'effet est trop faible pour être apprécié.

Au sens statistique un facteur a un effet non significatif (noté NS) si on a un risque α (ou une p-value) supérieur à 5% de se tromper en rejetant l'absence d'effet.

Au sens statistique un facteur a un effet significatif (noté * ou S) si on a un risque α (ou une p-value) compris entre 1 et 5% de se tromper en rejetant l'absence d'effet.

Au sens statistique un facteur a un effet très significatif (noté ** ou TS) si on a un risque α (ou une p-value) compris entre 0.1 et 1% de se tromper en rejetant l'absence d'effet.

Au sens statistique un facteur a un effet hautement significatif (noté *** ou HS) si on a un risque α (ou une p-value) inférieur à 0.1% de se tromper en rejetant l'absence d'effet.

Le même raisonnement est appliqué au test F de l'analyse de la variance, dans ce cas il ne s'agit pas de tester si un effet de facteur est différent de 0 mais de vérifier si le rapport de 2 variances est supérieur à 1.

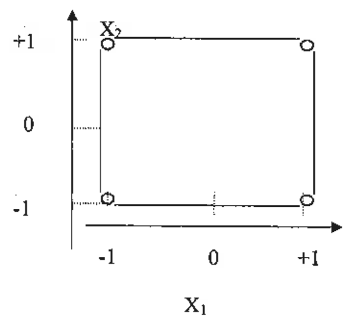
5 LE DOMAINE EXPERIMENTAL

Le domaine expérimental définit les limites entre lesquelles chaque facteur expérimental va varier. Il s'agit du choix de l'expérimentateur. C'est dans ce domaine expérimental que l'on choisira les niveaux du facteur de telle sorte que l'effet soit certain.
Le domaine expérimental ne doit être ni trop grand, ni trop petit de façon à pouvoir mesurer un effet de facteur.

Par exemple dans une étude à 2 facteurs à 2 niveaux
- le facteur 1 : température (10°C = niveau bas et 20°C = niveau haut)
- le facteur 2 : pH à 2 niveaux (1 = niveau bas et 3 = niveau haut)

La combinaison des niveaux donne 4 conditions expérimentales. Si on traduit ces niveaux en variables codées -1 (niveau bas) et +1 (niveau haut) alors les 4 expériences forment les coins d'un carré.
Si l'étude comportait 3 facteurs alors le domaine expérimental en variables codées serait représenté par les sommets d'un cube.

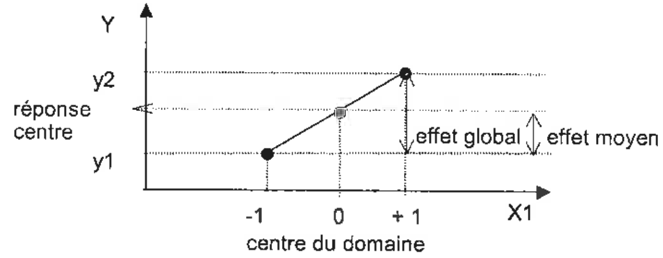
Remarque : le centre du domaine expérimental correspond au niveau 0 c'est à dire 15°C et 2 pour les facteurs 1 et 2 de l'exemple ci-dessus.



6 EFFET GLOBAL - EFFET MOYEN

Pour les déterminations des effets il faut distinguer les notions :
- d'effet global du facteur j, il mesure la variation de Y quand X_j passe du niveau bas (-1) au niveau haut (+1) (= 2* β_j).
- d'effet moyen du facteur j, noté β_j, il mesure la variation de Y quand X_j passe du niveau (0) au niveau haut (+1). Il est égal à la moitié de l'effet global.

Souvent pour les besoins de l'analyse il faudra réaliser des expériences au centre du domaine expérimental c'est à dire tous les facteurs étant réglés au niveau X_j= 0 en variables codées.



Exemple dans le cas d'une étude à 2 facteurs :

N° d'essai	X_1	X_2	Y
1	-1	-1	2
2	+1	-1	6
3	-1	+1	8
4	+1	+1	22

<u>Effet global de X_1</u> : moyenne des valeurs de Y quand X_1 est au niveau + 1 - moyenne des valeurs de Y quand X1 est au niveau -1 = (1/2*(6+22) - 1/2*(8+2)) = 14 - 5 = 9
<u>b1 = Effet moyen de X_1</u> = 1/2*effet global = 4.5 soit 1/4*(6+22-8-2)
Ce qui revient à faire la somme algébrique des Y affectés des signes de la colonne X_1 divisée par 4

<u>Effet global de X_2</u> : moyenne des valeurs de Y quand X_2 est au niveau + 1 - moyenne des valeurs de Y quand X_2 est au niveau -1 = (1/2*(8+22) - 1/2*(6+2)) = 15 - 4 = 11
b2 = Effet moyen de X_2 = 1/2*effet global = 5.5 soit 1/4*(8+22-6-2)
Ce qui revient à faire la somme algébrique des Y affectés des signes de la colonne X_2 divisée par 4

En ce qui concerne $b_0 = \hat{Y}_0$ = Réponse estimée calculée au centre du domaine expérimental (pour $X_1 = 0$ et $X_2 = 0$) :

Moyenne des Y pour X_1 au niveau -1 et au niveau +1 = 1/2*(1/2*(8+2) + 1/2*(6+22)) = 1/2*(5 + 14) = 9.5

Mais aussi :

Moyenne des Y pour X_2 au niveau -1 et au niveau +1 = 1/2*(1/2*(6+2) + 1/2*(8+22)) = 1/2*(4 + 15) = 9.5

Ce qui revient à faire la somme algébrique des Y affectés des signes de la colonne I (vecteur identité) divisée par 4

N° d'essai	I	Y
1	+1	2
2	+1	6
3	+1	8
4	+1	22

7 NIVEAU D'UN FACTEUR

Un niveau de facteur représente la valeur (quantitative ou qualitative) de la modalité que prend ce facteur pour une expérimentation.

Le facteur température a 2 niveaux si on choisit les modalités 15°C et 25°C, on parle de niveau bas (noté -1) et de niveau haut (noté +1). Le niveau 0 représente le centre du domaine expérimental, dans cet exemple il serait égal à 20°C.

Le facteur origine d'une matière première a 2 niveaux si on choisit les modalités fournisseur A et fournisseur B qui constitueront le niveau bas (noté -1) et le niveau haut (noté +1). Dans ce cas il n'y a pas de centre du domaine expérimental.

8 VARIABLES CODEES ET NATURELLES

Pour pouvoir comparer entre eux les effets des facteurs on préfère utiliser les variables codées (notée X_j) à la place des variables naturelles (notées U_j).

Les variables codées sont des nombres sans dimension, indépendantes des unités et comparables entre elles. L'étude de l'effet de la variation d'un facteur revient alors à l'étude de la variable codée qui lui est associée.

Matrice expérimentale				Plan expérimental		
X_1	X_2	X_3		U_1	U_2	U_3
+1	+1	-1	Essai 1	100	40	A
-1	+1	+1	Essai 2	50	40	B
+1	-1	+1	Essai 3	100	30	B
-1	-1	-1	Essai 4	50	30	A
Variables codées				Variables naturelles		

Ces variables codées sont définies de la manière suivante:

X_{ij} : variable codée du facteur j pour le niveau i

$$X_{IJ} = (U_{IJ} - U_{0J}) / \Delta U_J$$

U_{ij} : valeur de la variable naturelle du facteur j au niveau i

U_{0j} : valeur de la variable naturelle du facteur j au centre du domaine expérimental

ΔU_j : pas de variation de la variable naturelle; demi variation entre la valeur maximale et la valeur minimale de la variable naturelle j.

Exemple: U_1: facteur température,

Facteur N°1 (j = 1) à 3 niveaux i = 0 (centre) ,1 (niveau -1) ou 2 (niveau +1)

U_{11}: 50°C	$X_{11} = (50 - 75) / 25 = -1$
U_{21}: 100°C	$X_{21} = (100 - 75) / 25 = +1$
$U_{01} = (100 + 50) / 2 = 75$°C	$X_{01} = (75 - 75) / 25 = 0$
$\Delta U_1 = (100 - 50) / 2 = 25$°C	$\Delta X_1 = (1 - (-1)) / 2 = 1$

Remarque 1 : si la variable est de nature qualitative par convention on affectera les codes +1 et -1 à chacune des modalités.

Exemple : U_2 : type d'additif

Facteur N°2 à 2 niveaux type A = -1 et type B = +1

Remarque 2 : Si on ne connait pas la valeur du pas de variation ΔU_j (situation possible dans les modules 7 à 10) on utilisera le théorème de Thalès.

9 MATRICE EXPERIMENTALE

Une matrice expérimentale est un tableau à N lignes (N expériences) et k colonnes (pour k facteurs) exprimée en variables codées. La dimension de cette matrice est égale à [N; k] où N est le nombre d'essais et k le nombre de facteurs.

Exemple d'une matrice expérimentale pour 2 facteurs, dim = [4; 2]:

Essai	Facteur X_1	Facteur X_2
1	-1	-1
2	+1	-1
3	-1	+1
4	+1	+1

A chaque essai on fixe les conditions expérimentales codées (-1 pour le niveau bas du facteur, +1 pour le niveau haut) à attribuer à chacun des facteurs X_j.

10 LE MODELE MATHEMATIQUE EN VARIABLES CODEES

Les hypothèses de départ d'une étude supposent que chaque facteur X_j agit linéairement et additivement sur la réponse Y selon un modèle. Dans ce cas il s'agit d'un modèle du 1er degré avec k' coefficients (k' = k+1) du type :

$$\hat{y} = \beta_0 + \beta_1 X_1 + \ldots + \beta_j X_j + \ldots \beta_k X_k = \beta_0 + \Sigma\, \beta_j X_j$$

Dans certains cas il peut y avoir des interactions entre 2 ou plusieurs facteurs et le modèle mathématique à k' coefficients s'écrit :

$$\hat{y} = \beta_0 + \Sigma\, \beta_j X_j + \beta_{12} X_1 X_2 + \ldots + \beta_{ij} X_i X_j + \ldots = \beta_0 + \Sigma\, \beta_j X_j + \Sigma\, \beta_{ij} X_i X_j + \ldots$$

On l'appelle modèle synergique si tous les 2^k coefficients ($k' = 2^k$) sont présents dans le modèle.

Dans d'autres cas on devra rechercher les valeurs des k' coefficients d'un modèle quadratique de la forme :

$$\hat{y} = \beta_0 + \Sigma\, \beta_j X_j + \beta_{11} X_1^2 + \ldots + \beta_{jj} X_j^2 + \ldots + \beta_{kk} X_k^2 \ldots + \Sigma\, \beta_{ij} X_i X_j + \ldots$$
$$\hat{y} = \beta_0 + \Sigma\, \beta_j X_j + \Sigma\, \beta_{ij} X_i X_j + \Sigma\, \beta_{jj} X_j^2$$

Dans tous les cas chaque coefficient β_j sera estimé par la valeur calculée de b_j effet moyen du facteur j. Il en est de même en ce qui concerne les interactions.

11 MATRICE DES EFFETS EN VARIABLES CODEES

La matrice des effets est un tableau à N lignes (N expériences de la matrice expérimentale) et k' colonnes (k' coefficients du modèle) en variables codées, elle est notée matrice X ou plus simplement X dans les formules des calculs. Sa dimension est égale à [N, k'] avec k' = nombre de coefficient dans le modèle.

Exemple: soit le modèle à k' = 6 coefficients:

$$\hat{y} = b_0 + b_1 X_1 + b_2 X_2 + b_{12} X_1 X_2 + b_{11} X_1^2 + b_{22} X_2^2$$

Dans ce cas, la matrice expérimentale a une dimension égale à [12; 2], et la matrice des effets a une dimension égale à [12; 6] et sera de la forme:

I	X_1	X_2	X_1X_2	X_1^2	X_2^2
1	-1	-1	+1	1	1
1	+1	-1	-1	1	1
1	-1	+1	-1	1	1
1	+1	+1	+1	1	1
1	0	0	0	0	0
1	0	0	0	0	0
1	0	0	0	0	0
1	0	0	0	0	0
1	-1.414	0	0	2	0
1	+1.414	0	0	2	0
1	0	-1.414	0	0	2
1	0	+1.414	0	0	2

On construit l'alternance des signes des colonnes X_1X_2, X_1^2 et X_2^2 à partir des produits des signes des colonnes X_1 et / ou X_2.
Cette matrice sera utilisée pour déterminer les valeurs numériques des coefficients du modèle en variables codées.
Les variables étant sans dimension, les effets calculés à l'aide de cette matrice seront indépendants des unités choisies et comparables entre eux.

Remarque très importante :
La matrice des effets X contient bien les N essais, y compris les essais répétés, s'il y en a. Dans ce cas elle peut perdre la propriété d'Hadamard ; il faudra alors passer par le calcul matriciel pour déterminer la matrice B.

12 CALCULS DES COEFFICIENTS DU MODELE

Pour déterminer les k' coefficients d'un modèle on utilise le calcul matriciel appliqué à la régression :

$$B = (X'X)^{-1} X'Y$$

N : nombre d'essais
B : matrice des k' coefficients du modèle, dim B = [k', 1]
X : matrice des effets en variables codées dim X = [N, k'] où N est le nombre d'essais dans la matrice expérimentale
X' : transposée de la matrice des effets dim X' = [k', N]
X'X : produit matriciel, dim X'X = [k' ; k']
$(X'X)^{-1}$: inverse de la matrice X'X, dim $(X'X)^{-1}$ = [k' ; k']
Y : matrice des réponses, dim Y = [N, 1]
X'Y : produit matriciel, dim X'Y = [k' ; 1]

Cas particulier des matrices ayant la propriété d'Hadamard (X'X) = NI:

Matrice X			
+1	-1	-1	+1
+1	+1	-1	-1
+1	-1	+1	-1
+1	+1	+1	+1

Matrice X' transposée de X			
+1	+1	+1	+1
-1	+1	-1	+1
-1	-1	+1	+1
+1	-1	-1	+1

Produit X'X = NI			
4	0	0	0
0	4	0	0
0	0	4	0
0	0	0	4

Inverse du produit X'X = I/N			
1/4	0	0	0
0	1/4	0	0
0	0	1/4	0
0	0	0	1/4

Matrice des réponses = matrice Y
y1
y2
y3
y4

Produit X'Y
+ y1 + y2 + y3 + y4
- y1 + y2 - y3 + y4
- y1 - y2 + y3 + y4
+ y1 - y2 - y3 + y4

Matrice B	=	(X'X) $^{-1}$ X'Y
b0	=	$\frac{1}{4}$*(+ y1 + y2 + y3 + y4)
b1		$\frac{1}{4}$*(- y1 + y2 - y3 + y4)
b2		$\frac{1}{4}$*(- y1 - y2 + y3 + y4)
b3		$\frac{1}{4}$*(+ y1 - y2 - y3 + y4)

Chaque coefficient b_j est égal à la somme algébrique des réponses Yi affectées des signes de la colonne X_j de la matrice X divisée par le nombre d'expérience.

Quand N=4 on dit que la matrice est une H_4.
N = 8 → H_8 (H pour Hadamard)
N = 12 → H_{12}

On rajoute des expériences pour bien estimer l'erreur je crois

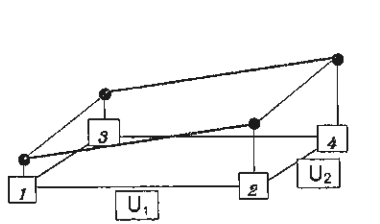
13 Notion d'interaction

Il y a interaction entre 2 facteurs U_1 et U_2 si l'effet du facteur U_1 dépend du niveau de U_2 ou inversement.

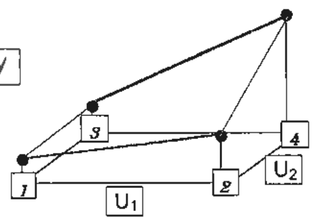
Exemple : il y a interaction entre le facteur U_1 " vitesse de conduite " et le facteur U_2 " pression de gonflage du pneu " si la vitesse de conduite a un effet différent sur Y " l'usure des pneus " selon que ceux-ci sont sous - gonflés ou sur - gonflés.

N° d'essai	U_1 vitesse	U_2 pression pneu	Y Sans interaction	Y Avec interaction
1	lente	sur gonflage	2	2
2	rapide	sur gonflage	8	6
3	lente	sous gonflage	12	8
4	rapide	sous gonflage	18	22

Dans le schéma sans interaction on remarque que les effets du facteur U_1 (pente rouge) sont identiques quel que soit le niveau d'U_2. De même les effets du facteur U_2 (pente bleue) sont les mêmes que U_1 soit au niveau bas ou haut. On n'observe pas ces propriétés dans le schéma avec interaction.



Sans interaction



Avec interaction

> Une interaction entre 2 facteurs est dite interaction du 1er ordre,
> Une interaction entre 3 facteurs est dite interaction du 2ème ordre,
> Une interaction entre 4 facteurs est dite interaction du 3ème ordre....etc.

14 Nombre de coefficients

Dans une étude à k facteurs à 2 niveaux on peut avoir à mesurer <u>au maximum</u> 2^k effets :

C_k^0 = 1 niveau moyen noté bo qui correspond à la réponse estimée au centre du domaine expérimental

C_k^1 = k effets moyens principaux notés bj

C_k^2 = 1/2 * k * (k-1) effets d'interaction du 1er ordre notées bij

Plus généralement : C_k^q = nombre d'interactions d'ordre (q -1)

En règle générale on considère que les interactions d'ordre 2 et plus, sont négligeables devant les effets principaux et les interactions de 1er ordre.

Rappel :

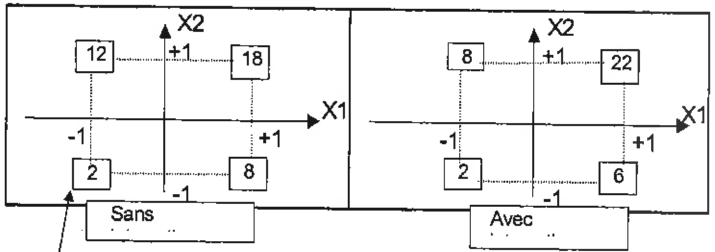
$$C_k^q = k!/[q!(k-q)!]$$

15 Croix d'interaction

C'est la représentation graphique d'une interaction de 1er ordre entre 2 facteurs X_1 et X_2.

Exemple : il y a interaction entre le facteur X_1 " vitesse de conduite " et le facteur X_2 " pression de gonflage du pneu " si la vitesse de conduite a un effet différent sur Y " l'usure des pneus " selon que ceux-ci sont sous - gonflés ou sur - gonflés.

N° d'essai	vitesse	pression	Y (sans)	Y (avec)
1	lente	sur	2	2
2	rapide	sur	8	6
3	lente	sous	12	8
4	rapide	sous	18	22

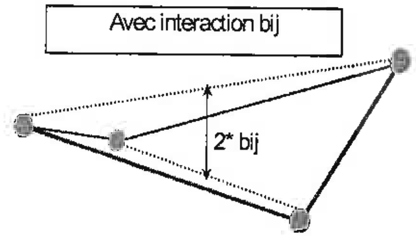


Dans chaque coin du carré figure la réponse expérimentale obtenue dans les conditions X_1 et X_2 de l'essai. Ici, Y = 2 quand la vitesse est lente (niveau -1 pour X_1) et le pneu sur gonflé (niveau -1 pour X_2).

Dans le modèle sans interaction l'effet de X_1 vaut +6 que X_2 soit au niveau -1 ou au niveau +1 (Y passe de 2 à 8 et de 12 à 18), de même l'effet de X_2 vaut +10 que X_1 soit au niveau -1 ou au niveau +1 (Y passe de 2 à 12 et de 8 à 18).

Dans le modèle avec interaction l'effet de X_1 n'est pas le même selon le niveau de X_2 : il vaut +4 (= 6-2) quand X_2 est au niveau -1 alors qu'il vaut +14 (= 22-8) quand X_2 est au niveau +1. De même l'effet de X_2 n'est pas le même selon le niveau de X_1 : il vaut +6 (= 8-2) quand X_1 est au niveau -1 alors qu'il vaut +16 (= 22 - 6) quand X_1 est au niveau +1.

Représentation géométrique de l'effet global d'une interaction de 1er ordre notée $2*b_{ij}$:



16 ALTERNANCE DES SIGNES D'UNE INTERACTION

L'alternance des signes de la colonne de l'interaction permet de calculer l'effet de l'interaction, elle s'obtient en faisant le produit des signes des colonnes de X_1 et de X_2:

N° d'essai	I	X_1	X_2	X_1X_2	Y
1	+1	-1	-1	+1	2
2	+1	+1	-1	-1	6
3	+1	-1	+1	-1	8
4	+1	+1	+1	+1	22

En l'absence d'interaction l'effet moyen de chaque facteur devrait être le même quelque soit le niveau de l'autre facteur :
Effet moyen de X_1 pour X_2 fixé au niveau bas (-1) = 1/2*(6-2) = 2
Effet moyen de X_1 pour X_2 fixé au niveau haut (+1) = 1/2*(22-8) = 7
On compare ces résultats en calculant leur demi variation, reflet de l'interaction = 1/2*(7-2) = 2.5. On peut aussi l'écrire sous la forme : 1/2*(1/2*((22-8)-(6-2)) = 1/4*(+2 -6 -8 +22) = 2.5

Effet moyen de X_2 pour X_1 fixé au niveau bas (-1) = 1/2*(8-2) = 3
Effet moyen de X_2 pour X_1 fixé au niveau haut (+1) = 1/2*(22-6) = 8
On compare ces résultats en calculant leur demi variation, reflet de l'interaction = 1/2*(8-3) = 2.5. On peut aussi l'écrire sous la forme : 1/2*(1/2*((22-6)-(8-2)) = 1/4*(+2 -8 -6 +22) = 2.5

Pour calculer l'effet de l'interaction entre 2 facteurs, <u>si la matrice des effets possède la propriété d'Hadamard</u>, il suffit donc de faire la somme algébrique des Y affectés des signes du produit de ceux des colonnes X_1 et X_2 divisée par 4. L'effet d'interaction de 1er ordre se note b_{ij}.

17 ESTIMATION DE LA VARIANCE EXPERIMENTALE $\sigma^2\varepsilon$

Pour tester les k' coefficients (b_j, b_{ij}, b_{jj}) du modèle (test t sur les b_j et b_{ij}) on a besoin de calculer les écart-types des coefficients $\sigma(b_j)$, $\sigma(b_{ij})$ et $\sigma(b_{jj})$ pour cela il faut déterminer la matrice des variances covariances VarB = $\sigma^2\varepsilon * (X'X)^{-1}$.

La variance expérimentale $\sigma^2\varepsilon$ peut être estimée à partir :
- des fluctuations naturelles (CM répétitions) des facteurs non étudiés (facteurs non contrôlés), c'est la variabilité naturelle du procédé, on la mesure grâce aux répétitions. Dans cette situation on utilisera le ddl des répétitions = r - 1 pour le test t qui suivra
- de la variabilité résiduelle (*CM e*) mesurée à l'aide de la différence entre des réponses estimées et des réponses observées. Dans cette situation on utilisera le ddl résiduel = N - k' pour le test t qui suivra.

- Si un modèle est prévisionnel et structurel (voir ANOVA §21) l'erreur expérimentale $\sigma^2\varepsilon$ est estimée avec *CMe*,
- Si le modèle n'est que structurel, alors l'erreur expérimentale $\sigma^2\varepsilon$ est estimée uniquement avec *CM* répétitions.

Pour estimer la variance expérimentale nous supposerons que:
- les facteurs non contrôlés fluctuent de façon aléatoire; en chaque point i du domaine expérimental, le résultat de mesure yi est une variable aléatoire.
- les résultats de mesure aux différents points du domaine expérimental sont indépendants les uns des autres.
- la variance expérimentale des mesures est la même en tout point du domaine expérimental.

Pour estimer la variance expérimentale 3 méthodes sont possibles:
- doubler le plan expérimental,
- répéter un essai du dispositif,
- répéter l'expérience au centre du domaine expérimental,

17 - 1 - DOUBLER LE PLAN

Dans cette situation la variance expérimentale σ^2_ε ne peut être estimée qu'à partir variance résiduelle (CMe) qui s'accompagne d'un fort ddle ce qui rend l'analyse statistique plus puissante mais le plan expérimental plus coûteux.
Dans ce cas, la matrice des effets conserve la propriété d'Hadamard.
Chacune des expériences du plan est réalisée 2 fois, de façon indépendante.
On a N (= 2*nombre d'essai prévu initialement) essais au total. En chaque point correspondant i du domaine expérimental, nous avons 2 résultats pour chaque expérience i; nous pouvons estimer l'erreur expérimentale par:

CMe = SCEe / ddle

SCEe = SCEyi - SCE $\hat{y}i$
ddle = N - k'
k' : nombre de coefficients
N = nombre total d'essais

Application numérique: exemple à partir d'une matrice des effets d'un synergique à 2 facteurs, on trouve CMe = 13,875 avec un ddl = 4.
Chaque essai i (i = 1 à 4) est doublé,

N° essai	Yi	Yi^	ei
1	172	174,5	-2,5
2	148	145	3
3	78	75,5	2,5
4	170	167,5	2,5
1'	177	174,5	2,5
2'	142	145	-3
3'	73	75,5	-2,5
4'	165	167,5	-2,5
somme	1125	1125	0
SC	170519	170464	55,5
SCE	12315,9	12260,4	55,5
ddl	7	3	4
CM		4086,79	**13,875**

Avec cette méthode on ne peut pas dissocier les valeurs individuelles de SCE adéquation (erreur d'ajustement) et SCE répétitions (erreur naturelle).

17 - 2 - REPETER UN POINT I DU DOMAINE EXPERIMENTAL

On pourra utiliser la méthode des répétitions d'un essai du plan expérimental. Dans cette situation on peut estimer la variance liée aux répétitions CMrepet, elle représente les fluctuations naturelles.
Dans ce cas, la matrice des effets perd la propriété d'Hadamard.

Une des expériences i du plan est réalisée r fois (j = 1...r), de façon indépendante, on estimera CM répétitions à l'aide de la formule :

CM répétitions = SCE répétitions / (r - 1) ν = ddl = (r - 1)

$\bar{y}i$: moyenne des réponses de l'essai répété = $\sum_{j=1}^{j=r}(yij)$ / r

SCE répétitions = SCE yij ($\bar{y}i$) : Somme des Carrés des Ecarts des Yij par rapport à leur moyenne

SCE yij ($\bar{y}i$) = $\sum_{j=1}^{j=r}(yij-\bar{y}i)^2$

Application numérique: 3 répétitions (r = 3) pour l'essai 1
y_{11} =172 y_{12} = 173 y_{13} = 171 $\bar{y}i$ = (172 + 173 + 171) / 3 = 172
CM répétitions = ((172 - 172)² + (173 - 172)² + (171 - 172)²) / 2 = 1
avec ddl = 2

17 - 3 - REPETER L'ESSAI AU CENTRE DU DOMAINE EXPERIMENTAL

On pourra utiliser la méthode des répétitions de l'essai au centre du domaine expérimental. Dans cette situation on peut estimer la variance liée aux répétitions CMrepet, elle représente les fluctuations naturelles.
Dans ce cas, la matrice des effets conserve presque !! la propriété d'Hadamard.

L'expérimentateur effectue plusieurs détermination (r points : j = 1...r) au centre du domaine expérimental (tous les facteurs de l'étude sont réglés au niveau 0 en **variables codées**); ce sont les "points au centre". Outre l'estimation de la variance de la mesure, ces points permettent de vérifier l'absence de dérive au cours de l'expérimentation et permet la recherche de modèle. Le ddl sera alors égal au (nombre de points au centre - 1 donc r - 1).

CM répétitions = SCE répétitions/ (r - 1)	ν = ddl = (r - 1)

$\bar{y}_0$: moyenne des réponses de l'essai au centre répété = $\sum_{j=1}^{j=r}(y_0 j)$ / r

SCE répétitions = SCE y0j ($\bar{y}_0$) : Somme des Carrés des Ecarts des Yij par rapport à leur moyenne = $\sum_{j=1}^{j=r}(y_0 j - \bar{y}_0)^2$

Application numérique :
y_{01} = 64.6 y_{02} = 65.1 y_{03} = 65.5 $\bar{y}_0$ = (64.6 + 65.1 + 65.5) / 3 = 65.07
CM répétitions = SCEy_{0j}($\bar{y}_0$)/ (3-1)

CM répétitions = [(64.6 - 65.07)² + (65.1 - 65.07)² + (65.5 - 65.07)²] /2 = 0.203
ν = ddl = (3 - 1) = 2

18 CONTRASTES

On appelle contraste d'une colonne j, la valeur numérique obtenue en calculant la somme algébrique des réponses Y affectées des signes de la colonne j de la matrice des effets, divisée par le nombre d'expériences N. Un contraste est noté ℓ_j. Le contraste représente selon le dispositif expérimental :

- un effet de facteur principal seul $\ell j = b_j$,
- un effet d'interaction seul $\ell j = b_{ij}$,
- une somme d'effets et d'interactions et/ou d'interactions entre elles

19 ANALYSE STATISTIQUE DES COEFFICIENTS OU DES CONTRASTES

On réalise un test t sur chaque coefficient et à l'aide du t de Student (pour ν = ddl) lu dans la table pour les 3 niveaux de signification (5%, 1% et 0.1%) on conclut sur la mise en évidence d'effet principaux (ou d'interaction) ou non. Si on dispose d'Excel on détermine la p-value.

On a besoin des variances des coefficients :VarB = $\sigma^2\varepsilon$ (X'X)$^{-1}$

Les variances des effets, (§20) des interactions et de la moyenne se trouvent dans la diagonale de: Var B = $\sigma^2\varepsilon$ (X'X)$^{-1}$ ou sont égales à $\sigma^2\varepsilon$ / N (si X a la propriété d'Hadamard).
En disposant d'une estimation de la variance expérimentale notée $\sigma^2\varepsilon$ on pourra donc analyser statistiquement les effets principaux et les interactions afin d'en déduire ceux et celles qui sont significatifs. Pour chaque interaction significative on représentera et analysera la croix (ou le cube) d'interaction (§15) correspondante.

On suppose que la variable aléatoire b_j (ℓ_j) suit une loi Normale (si variance connue) ou de Student (pour ddl = ν si variance estimée).
Pour tester la significativité du contraste on teste l'hypothèse nulle H_0 contre son alternative H_1.

H_0 : ℓj = 0 ou β_j = 0 (pas de mise en évidence d'un effet)
H_1 : $\ell j \neq 0$ ou $\beta_j \neq 0$ (mise en évidence d'un effet)
On calcule le critère statistique noté t calc (variable centrée réduite) :

$$|t\ calc| = \ell j / \sigma(\ell j) \text{ ou } |t\ calc| = bj / \sigma(bj)$$

Puis on détermine le risque de se tromper en rejetant H_0 :
p-value = 2*P(T > |t calc|)

Si on utilise une matrice des effets X ayant la propriété d'Hadamard, à chaque contraste ℓj est associé sa variance $\sigma^2(\ell j)$ ou $\sigma^2(b_j) = \sigma^2\varepsilon / N$
Sinon on utilise la diagonale de la matrice VarB = $\sigma^2\varepsilon\,(X'X)^{-1}$.

Exemple pour $\sigma^2\varepsilon$ = CMe = 13,875 avec un ddl = 4

CMe= 13,875 ddl = 3 $\sigma^2(bi)$ = 13,875 / 8 = 1,734 $\sigma(bj)$ = 1,317

j	12	2	1
bj	30,375	-19,125	15,625
|t calc|	23,06	-14,52	11,86
p-value (Student pour ddl = 4)	0,0000	0,0001	0,0003
Niveau de significativité	***	***	***

Les facteurs X_1 et X_2 ont des effets hautement significatifs, l'interaction X_1X_2 est hautement significative, on fera donc l'interprétation à partir de la croix d'interaction.

20 Calcul de la matrice des variances-covariances VarB

Dans la méthode d'analyse de la régression multiple on détermine la matrice :

$$VarB = \sigma^2\varepsilon * (X'X)^{-1}$$

Il s'agit du produit d'une matrice ($(X'X)^{-1}$) part une constante, la variance expérimentale, $\sigma^2\varepsilon$. Dans ce cas on obtient une matrice symétrique VarB dont les valeurs de la diagonale représentent les variances des coefficients bj notées $\sigma^2(b_j)$; de part et d'autres, les valeurs représentent les covariances entre les régresseurs.

Ci-dessous est représentée la matrice VarB associée au modèle :

$$\hat{y} = b_0 + b_1X_1 + ... + b_jX_j + ...b_kX_k = b_0 + \Sigma\, b_jX_j$$

$$VarB = \begin{bmatrix} \sigma^2(b_0) & \mathrm{cov}(b_0,b_1) & ..... & \mathrm{cov}(b_0,b_j) & ..... & \mathrm{cov}(b_0,b_k) \\ \mathrm{cov}(b_1,b_0) & \sigma^2(b_1) & ..... & \mathrm{cov}(b_1,b_j) & ..... & \mathrm{cov}(b_1,b_k) \\ ..... & ..... & ..... & ..... & ..... & ..... \\ \mathrm{cov}(b_j,b_0) & ..... & ..... & \sigma^2(b_j) & ..... & \mathrm{cov}(b_j,b_k) \\ ..... & ..... & ..... & ..... & ..... & ..... \\ \mathrm{cov}(b_k,b_0) & \mathrm{cov}(b_k,b_1) & ..... & \mathrm{cov}(b_k,b_j) & ..... & \sigma^2(b_k) \end{bmatrix}$$

Dans le cas particulier des matrices des effets ayant la propriété d'Hadamard, $(X'X)^{-1} = I / N$ on obtiendra :

$$VarB = \sigma^2\varepsilon * \begin{bmatrix} 1/N & 0 & ..... & ..... & 0 \\ 0 & 1/N & 0 & ..... & 0 \\ ..... & 0 & ..... & ..... & ..... \\ 0 & ..... & 0 & 1/N & 0 \\ 0 & 0 & ..... & 0 & 1/N \end{bmatrix}$$

$\sigma^2\varepsilon$ = CM répétitions si le modèle est structurel
$\sigma^2\varepsilon$ =CMe si le modèle est structurel et prévisionnel

Ce qui donne en fait une matrice où tous les éléments sont nuls exceptés ceux de la diagonale qui sont tous égaux à $\sigma^2\varepsilon / N$. Donc tous les $\sigma^2(b_j)$ sont égaux entre eux.

21 Analyse de la variance

On peut utiliser l'analyse des la variance (ANOVA) pour calculer la probabilité d'observer un type de lien entre une combinaison de variables explicatives Xj et une réponse à expliquer Y (modèle) alors qu'il n'existe pas.

Modèle à k' coefficients construit à partir de N expériences :

$$\hat{y} = b_0 + \Sigma\, b_jX_j + b_{11}X_1^2 + ... + b_{jj}X_j^2 + ... + b_{kk}X_k^2 ... + \Sigma\, b_{ij}X_iX_j + ...$$
$$\hat{y} = b_0 + \Sigma\, b_jX_j + \Sigma\, b_{ij}X_iX_j + ... + \Sigma\, b_{jj}X_j^2$$

<u>Rappels du cours de 2ème année (ou « ce que fait Excel ! ») :</u>
Pour tester la validité d'un modèle à k' coefficients déterminés à l'aide de N essais, on teste l'hypothèse nulle H_0 contre son alternative H_1.

H_0 : $CM\hat{y} / CMe \leq 1$ (part des variations de Y expliquée par la régression est du même ordre de grandeur que la part non expliquée par la régression)
H_1 : $CM\hat{y} / CMe > 1$ (part des variations de Y expliquée par la régression est supérieure à la part non expliquée par la régression)

On calcule le critère statistique noté F calc : F calc = $CM\hat{y} / CMe$
Puis on détermine le risque de se tromper en rejetant H_0 à l'aide d'Excel (p-value).

Origine de la variation	SCE	ddl	CM	F calculé
Liée à la régression	$SCE\hat{y}(\bar{y})$	k' - 1	$CM\hat{y} = SCE\hat{y} / (k' - 1)$	$CM\hat{y} / CMe$
Liée à la résiduelle	SCEe	N - k'	CMe = SCEe / (N - k')	
Total	$SCEy(\bar{y})$	N - 1		

La quantité F calculé (critère statistique de Fisher à 2 ddl : $\nu_1 = k' - 1$ et $\nu_2 = (N - k')$) nous renvoie à la probabilité α de déclarer à tort que " la liaison est significative ".

SCEe mesure la somme des carrés des résidus ei ; ei = yi - ŷi

MAIS !!! Si vous disposez de répétitions d'un essai du plan ou d'essais au centre, alors vous pourrez valider le modèle sur sa validité prévisionnelle et/ou structurelle !!! Pour cela, dans la SCEe, on sépare :

- la part de la variation liée aux fluctuations naturelles (en calculant la SCE des essais répétés= SCE repet)
- la part de la variation liée à l'adéquation au modèle (calculée par différence entre SCEe et SCErepet).

Dans ce cas on détermine les F calculés pour valider le modèle :
1°) sur le plan prévisionnel, puis 2°) sur le plan structurel.

Hypothèses

Sur le plan prévisionnel
H_0 : CM ad / CMrepet ≤ 1 (l'erreur commise sur les estimations est du même ordre de grandeur que l'erreur naturelle)
H_1 : CM ad / CMrepet > 1 (l'erreur commise sur les estimations est supérieure à l'erreur naturelle)

Sur le plan structurel
H_0 : $CM\hat{y} / \sigma^2_\varepsilon \leq 1$ (part des variations de Y expliquée par la régression est du même ordre de grandeur que l'erreur expérimentale)
H_1 : $CM\hat{y} / \sigma^2_\varepsilon > 1$ (part des variations de Y expliquée par la régression est supérieure à l'erreur expérimentale)

Origine de la variation	SCE	ddl	CM = SCE / ddl	F calculé
Liée à la régression	SCEŷ($\bar{y}$)	k' - 1	CMŷ	CMŷ / σ^2_ε
Liée à la résiduelle	SCEe	N - k'	CMe	
adéquation	SCE ad	N- k'-r+1	CMad	CMad/ CMrepet
répétition	SCE repet	r - 1	CM repet	
Total	SCEy($\bar{y}$)	N - 1		

Si le modèle prévisionnel est valide alors le σ^2_ε = CMe = fluctuations naturelles liées aux répétitions + écarts d'adéquation (ce que fait Excel). Si prévi. valide alors σ^2_ε = CMe = CMrept + CMad

Si le modèle prévisionnel n'est pas valide alors le σ^2_ε = CMrépétitions = fluctuations naturelles liées aux répétitions uniquement (ce que ne fait pas Excel), dans ce cas les courbes iso réponses n'auront pas de raison d'être.

si prévi. pas valide alors σ^2_ε = CMrepet

x	y1	y1^	e
5	12,5	13,05	0,55
6	15,5	15,10	-0,40
	16,8	17,15	0,35
7	17,3	17,15	-0,15
7	17,8	17,15	-0,65
8	19,5	19,21	-0,29
9	20,5	21,26	0,76
10	23,5	23,32	-0,18
SCE	77,175	75,453	1,722

sce repet
0,50

2,055	2,773
0,127	0,953
0,9777	0,536
262,93	6

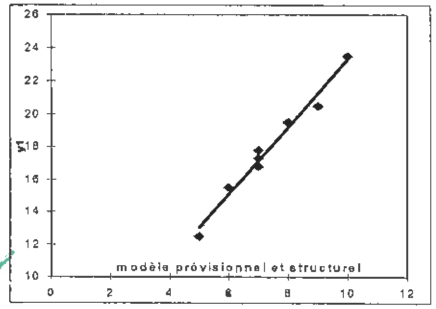
Insertion- fonction- statistique - droitereg- tout utilisable

R²	0,9777
Erreur-type	0,536

ANALYSE DE VARIANCE

	ddl	SCE	CM	F	alpha	
Régression	1	75,45	75,45	262,93	0,0000	***
Résidus	6	1,72	0,287			
adéquation	4	1,222	0,305	1,222	0,4965	NS
répétition	2	0,50	0,250			
Total	7	77,18				

	Coefficients	Erreur-type	Statistique t	Probabilité	
x	2,05	0,13	16,22	0,0000	***



CMe = 0,287
ddl = 6

	x	y2	y2^	e	
	5	10	13,79	3,79	
	6	18	15,62	-2,38	
	7	19	17,44	-1,56	sce repet
	7	18,5	17,44	-1,06	0,50
	7	19,5	17,44	-2,06	
	8	16	19,27	3,27	
	9	23	21,09	-1,91	
	10	21	22,92	1,92	
SCE	17,875	105,375	59,546	45,829	
SP	1102				
moy	7,375	18,125			
$s^2(b)$	0,013986				

1,825	4,664
0,654	4,919
0,5651	2,764
7,80	6

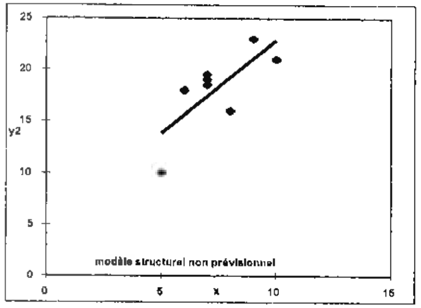
insertion- fonction- statistique - droitereg- parties grisées non utilisable

R^2	0,5651
Erreur-type	0,500

ANALYSE DE VARIANCE

	ddl	SCE	CM	F	alpha	
Régression	1	59,55	59,55	238,19	0,0042	***
Résidus	6	45,83	7,64			
adéquation	4	45,329	11,332	45,329	0,0217	*
répétition	2	0,50	0,250			
Total	7	105,38				

	Coefficients	Erreur-type	Statistique t	Probabilité	
x	1,83	0,12	15,43	0,0000	***



CMrépét = 0,250
ddl = 2

22 COEFFICIENT b_0

Dans un modèle linéaire multiple simple b_0 calculé représente l'estimation de la réponse au centre du domaine expérimental. C'est une valeur calculée représentant l'ordonnée à l'origine dans le cas le plus simple d'une droite d'équation $\hat{y} = b_0 + b_1X$ (avec X = 0).
Mais si on réalise des expériences au centre du domaine expérimental la réponse expérimentale obtenue se note aussi b_0 mais on l'appelle bo vrai.

Si les courbes iso réponses sont planes alors b_0 vrai = b_0 calculé, par contre s'il existe une courbure alors on observera une différence entre ces 2 valeurs.

23 APPROCHE DE LENTH

Cette méthode permet de désigner les facteurs « actifs » en l'absence d'estimation de l'erreur expérimentale.

> Si l'hypothèse de Normalité des effets de facteurs est vérifiée alors il existe un lien entre la médiane des valeurs absolues $|\widetilde{bj}|$ des effets et leur « pseudo » écart-type σ' (bj) tel que : $\sigma_0(bj) = 1.5 * |\widetilde{bj}|$

Sur la liste des k effets des k facteurs, dont les valeurs absolues sont rangées par ordre croissant, on écarte tous les effets « out » c'est à dire ceux dont la valeur absolue est supérieure à 2.5 *σ_0 (bj) ; il reste alors k' effets à partir desquels on estime un nouvel σ_0' (bj)

Puis on accepte de retenir un effet de facteur comme étant actif avec α % de risque d'erreur si sa valeur absolue est supérieure à $t_{1-\alpha/2}(\nu) * \sigma_0'(bj)$ avec :

- $\nu = k' / 3$
- $t_{1-\alpha/2}(\nu)$ pris dans la table de Student
- σ_0' (bj) pseudo écart-type des k' facteurs restant

exemple :

N° facteur	bj	abs bj	série rangée	série restante
1	1,255	1,255	0,002	0,002
2	0,271	0,271	0,011	0,011
3	0,002	0,002	0,011	0,011
4	-0,202	0,202	0,079	0,079
5	0,095	0,095	0,095	0,095
6	0,079	0,079	0,202	0,202
7	0,011	0,011	0,271	0,271
8	-0,011	0,011	**1,255**	
médiane			0,087	0,079
$\sigma_0(bj) = 1,5 *$ médiane			0,131	0,119
$2,5 * \sigma_0(bj)$			0,326	0,296
décision			Ecarter l'effet du facteur 1	Conserver les 7 effets restant
ddl				2,33
$t_{0,975}^{(2,33)}$				4,30
effet actif si bj > à				**0,510**

On a 5% de chance de déclarer à tort que b1, l'effet du facteur U1, est un effet actif.