Modèles d'analyse de variance

avec

STATISTICA

Bernard Clément, PhD

Génistat Conseils inc, 2010

RÉFÉRENCES

M. Kutner,. C. Nachtsheim, J. Neter, W. Li. (2005) *Applied Linear Statistical Models, 5 ed.* McGraw-Hill, QA 278.2.K87 2005, ISBN 0-07-238688-6

C.E. Lunneborg (1994) *Modeling Experimental and Observational Data*
Duxbury Press

R. O. Kuehl (2000) *Design of Experiments: Statistical Principles of Research Design and Analysis, 2 ed.*
Duxbury Press

G. A. Miliken, D. J. Johnson (1984) *Analysis of Messy Data, vol 1, vol 2*
Lifetime Learning Pubications, Wadsworth

R.R. Hocking (1985) *The analysis of Linear Models*
Brooks/Cole, QA 276 H56 1985, ISBN 053403618X

P. McCullagh (1983) *Generalized Linear Models*
London : Chapman and Hall, QA 276 M38 1983, ISBN 0412238500

S. R. Searle (1987) *Linear Models for Unbalanced Data*
New York : Wiley, QA 279 S42 1987, ISBN 0471840963

R. H. Myers, D. G. Montgomery (2002) *Generalized Linear Models with Applications in Engineering and the Sciences*
New York : Wiley, QA 276 M94 2002, ISBN 0471355739

D. J Hand (1987) *Multivariate Analysis of Variance and Repeated Measures: A Practical Approach to Behavioural Scientists*
Chapman and Hall, 1987, QA 278 H345 1987, ISBN 0412258102

StatSoft Electronic Textbook http://www.statsoft.com/textbook/

StatSoft, Inc. (2009). STATISTICA (data analysis software system), version 9.0.
www.statsoft.com

TABLE DES MATIÈRES

1. CONCEPTS DE BASE DE L'ANALYSE DE LA VARIANCE (ANOVA)

Le but de l'analyse de la variance est de tester la présence de différences significatives ou non entre des moyennes. Le cas classique le plus simple est celui de la comparaison de deux moyennes provenant d'échantillons indépendents. Le test t de Student est la procédure statistique appropriée pour traiter de ce cas. Si on a plus de deux échantillons indépendants, il faut une autre méthode pour tester simultanément l'égalité ou non de toutes les moyennes. L'analyse de la variance est la méthode employée pour tester plusieurs moyennes. Au cœur de cette méthode est la décomposition de la variabilité totale selon les différentes sources présentes dans les données. La variabilité totale est partionnée (décomposée) en deux sources: la variabilité dûe aux écarts entre les moyennes des différentes modalités d'un facteur (Inter groupe) et la variabilité résiduelle (non expliquée, Intra groupe) dûe à l'erreur expérimentale provenant des répétitions. Cette idée se généralise à des études statistiques avec plusieurs facteurs intervenant dans des plans (designs) expérimentaux complexes.

Exemple 1: cas de 2 groupes avec un seul facteur

id	groupe	y-réponse
1	a	1
2	a	2
3	a	3
4	b	5
5	b	6
6	b	7

	groupe a	groupe b
moyenne	2	6
somme de carrés (SS)	2	2
moyenne globale	4	
somme totale de carrés	28	

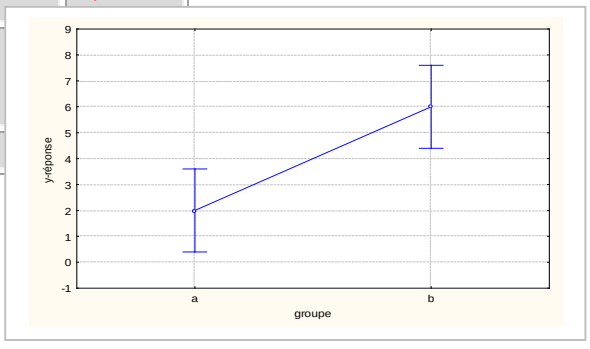
L'aboutissement d'une analyse de la variance (ANOVA) prend toujours la forme d'un tableau dont les éléments sont :

- l'identification des sources de variabilité ;
- le calcul des sommes de carrés (SS) ;
- le caclul des degrés de liberté (df) ;
- le calcul des carrés moyens (MS = SS / df) ;
- le caclul du ou des ratios F_0 = MS (effet)/ MS (erreur) ;
- l'évaluation de la probabilité p (p-value): p = Prob ($F \geq F_0$).

ANOVA

Source variabilité	Degr. de liberté (df)	y-réponse SS	y-réponse MS	y-réponse F	y-réponse p
Intercept	1	96	96	96	0,00061
Inter groupe	1	24	24	24	0,00805
Intra groupe (erreur)	4	4	1		
Totale	5	28			

les groupes sont statistiquement différents car p est « petit »



Test de signification
Sous l'hypothèse nulle (pas de difference entre les deux populations), la variance estimée avec la variabilité à l'intérieur de chaque groupe (intra) devrait être à peu près la même que la variance entre les groupes (inter). On peut comparer ces variances à l'aide d'un test F (distribution de Fisher) basé sur le ratio des deux variances moyennes (MS). Lorsque le ratio est assez grand ou encore lorsque la probabilité (p-value) d'être dépassée est petite (disons 0,05 ou moins), on conclut que les moyennes des populations sont significativement différentes l'une de l'autre.

Variables dépendantes (réponses) et facteurs (variables independantes)
Les variables mesurées sont appelées les *variables dépendantes* (réponses, variable d'intérêt, variable à expliquer). Les variables qui sont manipulées ou contrôlées (fixées ou mesurées) sont appelées les *variables indépendantes (*facteurs, variables explicatives, variables d'entrée).

Plusieurs facteurs
En général, les expériences ont typiquement plusieurs facteurs, généralement 5 ou moins. La méthode d'analyse ANOVA est capable de tenir en compte **plusieurs facteurs** ainsi que des **structures complexes** qui peuvent être présentes dans les données.

Exemple 2 : Supposons que l'on tient en compte un **deuxième facteur**, sexe de l'individu.

id	groupe	sexe	y-réponse
1	a	hom	1
2	a	hom	2
3	a	hom	3
4	b	hom	5
5	b	hom	6
6	b	hom	7
7	a	fem	3
8	a	fem	4
9	a	fem	5
10	b	fem	7
11	b	fem	8
12	b	fem	9

moyenne	groupe a	groupe b
hommes	2	6
femmes	4	8
tous	3	7

On peut partitionner la variabilité selon 3 sources :

(1) variabilité dûe au facteur groupe experimental
(2) variabilité dûe au facteur sexe
(3) erreur (inexpliquée, intra, résiduelle)

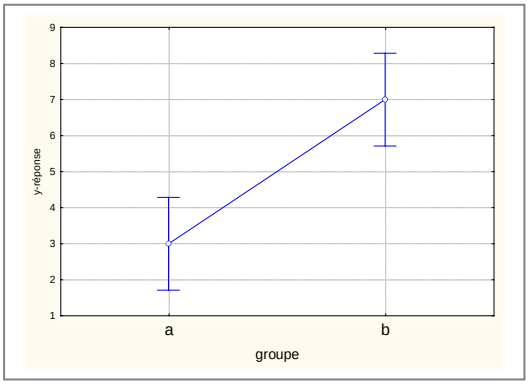
note: il y a une source additionnelle - ***interaction*** – qui sera présentée après.

Si on n'inclut pas le facteur sexe dans l'analyse, on obtient le tableau d'analyse de la variance suivant.

ANOVA

Source	df	y-réponse SS	y-réponse MS	y-réponse F	y-réponse p
Intercept	1	300	300	150	0,000000
Inter groupe	1	48	48	24	0,000624
Erreur	10	20	2		
Totale	11	68			

les groupes sont différents mais une partie de cette différence est dûe au facteur sexe.



L'analyse ANOVA est une méthode puissante car elle permet de :

- **tester chaque facteur** en contrôlant tous les autres ;
- tester des **hypothèses complexes** faisant intervenir les effets **d'interaction** entre les facteurs ;
- augmenter la **sensitivité** (puissance) des tests de signification.
  Les variables qui ne sont pas tenus en compte, augmente la somme de carrés SS de l'erreur résiduelle (expérimentale) et les degrés de liberté (df) associés. Ainsi, le carré moyen de l'erreur (résiduelle) MS peut augmenter et, avoir comme conséquence, de masquer l'effet des autres facteurs.

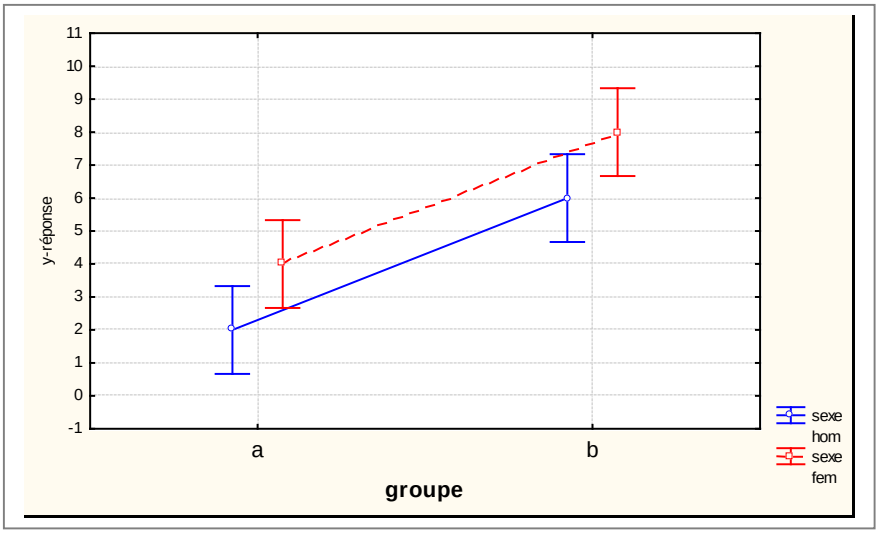
Effects principaux et effets d'interaction

Une analyse correcte pour comparer les 2 groupes devra tenir en compte l'influence du facteur sexe. Il y deux modèles que l'on peut proposer. Un **premier modèle général** (complet) qui incorpore des effets principaux des facteurs et en plus un effet d'interaction entre les eux facteurs. Un **deuxème modèle simple** incorpore des effets principaux seulement. On recommande de considérer un modèle général complet avec des effets d'interaction comme la première modélisation pour interpréter des données. À la suite de cette première analyse on peut enlever des effets principaux et des effets d'intraction qui ne sont pas significatifs en vue d'avoir un modèle parcimonieux. On recommande de conserver dans le modèle final des effets principaux non significatifs si des effets d'interaction significatifs impliquent ces facteurs. C'est le **principe de hiérarchie**.

<u>Analyse 1</u> : modèle complet y = général + groupe + sexe + groupe*sexe

ANOVA

Source	df	SS	MS	F	p
Intercept	1	300	300	300	0.0000
groupe	1	48	48	48	0.0001
sexe	1	12	12	12	0.0085
groupe*sexe	1	0	0	0	1.0000
erreur	8	8	1		
total	11	68			



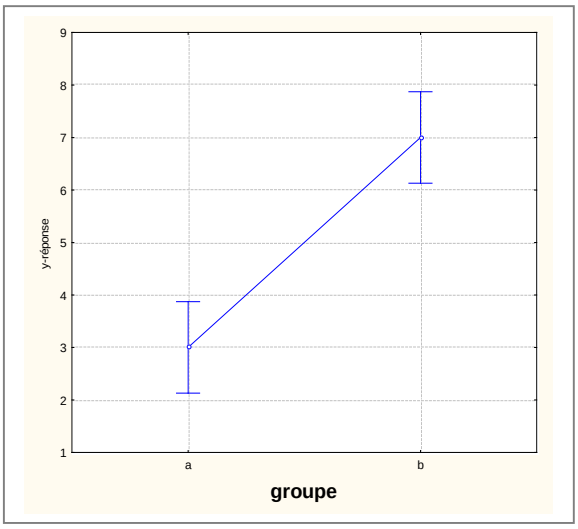
graphique de visualisation des données

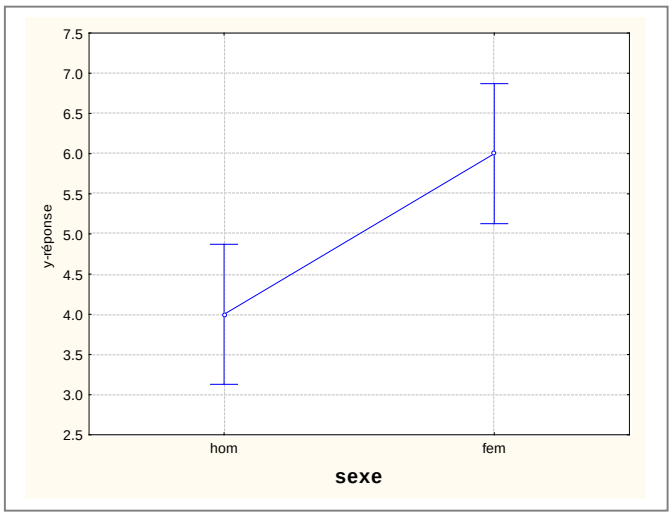
graphique de l'interaction entre les 2 facteurs

parrallélisme = absence d'intraction

Analyse 2 : modèle simplifié **y = général + groupe + sexe**

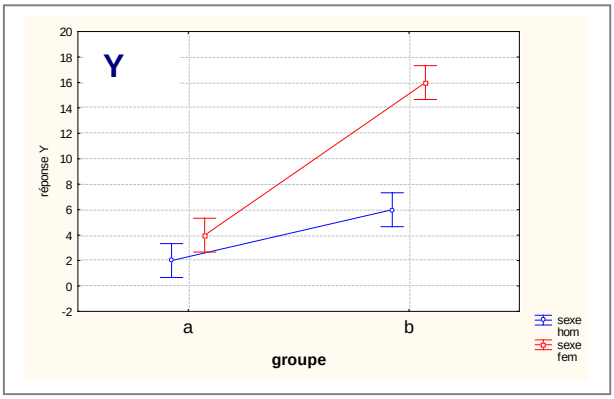
ANOVA					
Source	**SS**	**df**	**MS**	**F**	**p**
Intercept	300	1	300	337.5	0.000000
groupe	48	1	48	54.0	0.000043
sexe	12	1	12	13.5	0.005121
erreur	8	9	0.89		
total	68	11			

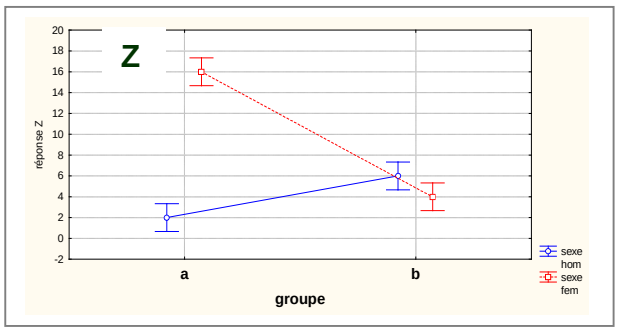




Exemple 3 : présence d'interaction

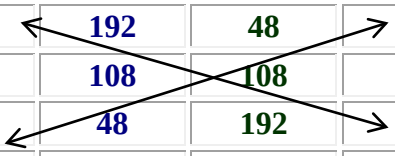
données				
id	**groupe**	**sexe**	**réponse Y**	**réponse Z**
1	a	hom	1	1
2	a	hom	2	2
3	a	hom	3	3
4	b	hom	5	5
5	b	hom	6	6
6	b	hom	7	7
7	a	fem	3	15
8	a	fem	4	17
9	a	fem	5	16
10	b	fem	15	3
11	b	fem	17	4
12	b	fem	16	5





ANOVA

Source	df	réponse Y SS	réponse Y MS	réponse Y F	réponse Z SS	réponse Z MS	réponse Z F
Intercept	1	588	588	588	588	588	588
groupe	1	192	192	192	48	48	48
sexe	1	108	108	108	108	108	108
groupe*sexe	1	48	48	48	192	192	192
erreur	8	8	1		8	1	
Total	11	356			356		



Interprétation des effets d'interactions d'ordre deux et plus

On peut dire qu'un **effet d'intéraction** entre deux facteurs a pour conséquence de modifier l'effet principal d'un des deux facteurs selon la valeur prise par le deuxième facteur. En général, un effet d'intéraction entre trois facteurs a pour conséquence de modifier l'effet d'interaction entre de 2 des 3 facteurs selon la valeur prise par le troisième facteur.

Dans le cas où les facteurs sont continus (modèles de régression), les effets d'interaction sont les coefficients des termes d'ordre 2 et plus du modèle polynomial. Ils sont plus faciles à interpréter que le cas où les facteurs sont qualitatifs (modèles d'analyse de variance).

2. L'APPROCHE PROCESSUS et L'ANALYSE STATISTIQUE

La figure 1 représente un processus et l'approche statistique pour son étude :

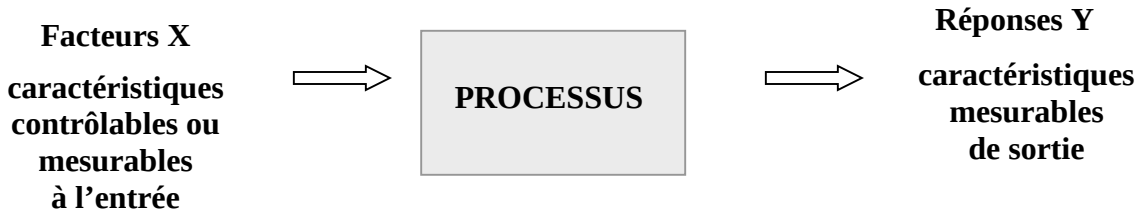


Figure 1 : l'approche processus et l'analyse statistique

La problématique est de dégager des relations et des équations entre les variables d'entrée (facteurs X) et les variables de réponse Y :

$$Y = F(X)$$

On admet a priori que ces relations ne peuvent être parfaites. Les variables facteurs X disponibles seront responsables de l'explication d'une partie seulement de la variabilité des variables de réponse Y.

Nous ne sommes pas dans un monde idéal avec des relations entrées (X)-sorties(Y) complètement prévisibles et connues. Cela s'explique par deux raisons fondamentales. D'une part, le système à l'étude n'est pas déterminé uniquement par les variables mesurées. Il y a toujours des variables inconnues et, par conséquent non mesurées, qui ne sont pas disponibles. D'autre part, la relation entrée-sortie F est inconnue, généralement complexe, et ne peut pas être complètement représentée l'aide d'une fonction relativement simple comme une fonction polynomiale.

Ces constations constituent la règle plutôt que l'exception dans toutes les études statistiques. Mais au delà de ces contraintes, il y moyen d'affiner notre connaissance afin de dégager des résultats pratiques et exploitables pour l'étude des processus.

Il y a deux méthodes statistiques pour l'étude des relations entrées-sorties. *L'analyse de régression* et *l'analyse de la variance*. Il est utile de connaître les distinctions et les limites de chaque méthode afin de savoir reconnaître le contexte approprié de leur application.

En **l'analyse de régression**, la structure des données est généralement simple et, elle est généralement constituée, de données historiques et observationnelles sans que celles-ci soient sous un contrôle précis comme dans les expériences planifiées. *L'objectif principal est le développement d'une équation (modèle) reliant les réponses et les prédicteurs*. Les prédicteurs du modèle à développer sont quantitatifs et prennent quasiment autant de valeurs distinctes que le nombre d'observation dont on dispose. L'équation est exploitée pour construire des tableaux de prédiction dans l'espace des prédicteurs. De plus, en général, un objectif d'optimisation de la réponse est souvent rattaché à ces études : déterminer un ou plusieurs ensembles de valeurs des prédicteurs qui permettent d'optimiser (maximum, minimum, nominal) la réponse. L'emphase de l'analyse est mise sur la qualité de l'équation développée : l'analyse des résidus, l'indentification des observations influentes, l'analyse de sensibilité, la recherche d'un modèle parcimonieux, etc.

En **analyse de variance**, les variables d'entrée sont des variables catégoriques (qualitatives) ayant un nombre relativement restreint de valeurs ou modalités. La source des données provient d'études expérimentales planifiées. *L'objectif visé est d'identifier si les effets principaux de chaque facteur et les effets d'interaction impliquant plusieurs facteurs ont une influence réelle sur la réponse au delà de la variabilité expérimentale*. La structure expérimentale est généralement le résultat d'un plan reflétant une certaine complexité. La difficulté principale est de proposer un modèle qui reflète correctement les subtilités du plan pour faire l'analyse de la variabilité du système. Le plan expérimental doit être décrit précisément : méthode d'assignation des traitements aux unités expérimentales, présence de contraintes, structure factorielle ou autre des traitements, participation possible des unités expérimentales à la collecte des données à plusieurs reprises, (mesures répétées dans le temps), etc. Le tableau résume les éléments distinctifs entre les deux méthodes d'analyse.

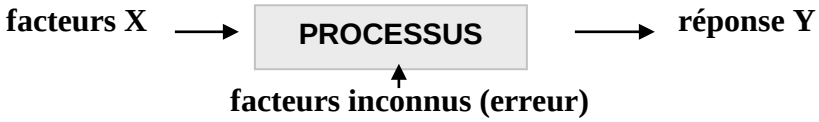
Tableau : comparaison des méthodes d'analyse

Elément de comparaison	**Analyse de régression**	**Analyse de variance**
Source des données	observationnelles ou historiques	résultat d'un plan d'expérimentation
Nombre d'observations	grand nombre : centaines, milliers ou plus	petit nombre : dizaines
Variables d'entrée	quantitatives	catégoriques
Nombre de valeurs distinctes des variables d'entrée	autant qu'il y a d'observations	nombre restreint généralement moins de 10
Utilisation des variables codage indicatrices (1 et 0) ou codage à effets (1/0/-1)	occasionnelle	employées systématique pour représenter les modalités
But	développement d'un modèle prédictif de la réponse	identification des effets significatifs sur la réponse
Emphase et difficulté	forme et la qualité du modèle	spécification du modèle reflétant la complexité du plan expérimental
Structure des données	simple	complexe

Dans les applications, il arrive très souvent que nous soyons en présence de variables continues et de variables catégoriques et que l'on s'interroge sur leur influence sur la variable de réponse. Dans ce dernier cas on utilise la méthode dite de **régression généralisée.** Cette méthode est aussi connue sous le nom d'**analyse de covariance** et elle combine l'analyse de régression et l'analyse de variance.

Types d'analyse et les désigns expérimentaux

L'analyse de la variance (ANOVA) est une méthode pour analyser l'influence de **facteurs qualitatifs** sur une (ou plusieurs) variables de réponse. Si tous les *facteurs (variables) sont quantitatif*s, **l'analyse de régression** est la méthode recommandée pour analyser les données. Le but recherché de l'analyse de régression est de développer des équations de prédiction et optimiser la variable de réponse. Si on a un mélange de facteurs (variables) qualitatifs et quantitatifs, la méthode est connue sous le nom **d'analyse de covariance** ou de **régression généralisée**.



X : variables / facteurs	ANALYSE
quantitatifs / numériques / continus	régression
qualitatifs / catégoriques	variance
quantitatifs et qualitatifs	covariance

La nature des variables X n'est pas le seul et unique élément à tenir en compte pour décider du type d'analyse statistique à employer avec un ensemble de données. Le tableau suivant définit les autres éléments à considérer.

élément	qualificatif	commentaire
type d'étude statistique	▪ expérimentale (mode actif) ▪ observationnelle (passif)	
contrôle des X	▪ fixés (étudiés, principaux) ▪ mesurés ▪ bloqués (secondaires) ▪ inconnus	*inconnus* : tout ce que l'on ne connait pas ; s'appelle *l'erreur expérimentale* même si les données n'ont pas été générée par un plan expérimental structuré
nombre de X	▪ 1 : simple (one way) ▪ ≥ 2 : multifacteurs (multi way)	
nature des X	▪ fixes ▪ aléatoires ▪ hybrides (mixed)	*fixe* : modalités (valeurs) choisies délibérement *aléatoire* : échantillon de modalités extraites de manière aléatoire (aucun contrôle)
nombre de Y	▪ 1 : ANOVA ▪ ≥ 2 : MANOVA	
design expérimental : (plan, devis) méthode d'assignation des traitements aux unités (sujets) expérimentales	▪ aléatoire (sans restriction) ▪ blocs ▪ mesures répétées ▪ fractionnaires / incomplet ▪ emboités (nested) ▪ parcelles divisées (split –plot) ▪ permutation (crossover) ▪ **…..**	*mesures répétées* : la variable de réponse est mesurée plusieurs fois sur la même unité statistique

Remarques

- **étude observationnelle** (rétrospective) **:** on étudie l'influence de facteurs sans avoir affaire à des données provenant d'un expérience ;

  étude expérimentale (prospective)**:** les données proviennent d'une expérience planifiée où les modalités des facteurs sont affectées aux unités au hasard selon un plan spécifique;

- **modèle linéaire général :** la régression et l'analyse de variance sont des cas particuliers de ce modèle qui est sans doute le plus fréquemment employé en analyse statistique ;

- **test t de Student :** est un cas particulier de l'analyse de la variance : 1 facteur avec 2 modalités;

- en général, les unités (sujets, individus statistiques) sont choisies au hasard parmi une population statistique ; lorsque l'on contrôle certains facteurs, on fait **l'assignation au hasard des traitements** (combinaison des facteurs contrôlés) aux unités : c'est la **randomisation** ; si nécessaire on fixe (contrôle) des **facteurs secondaires** sur les unités pour améliorer l'efficacité de l'étude (**plans en blocs**, …) : c'est le principe du **blocage ;**

- **facteur intra** (within) : si la même unité statistique est mesurée plusieurs fois (temps, différentes conditions), on a un **plan en mesures répétées** ; le facteur enfoui dans la réponse est appelé *facteur intra* (« within ») ; on traite ce cas en définissant plusieurs variables de réponse et en analysant avec une analyse de variance multidimensionnelle (MANOVA) ; les réponses sont naurellement dépendantes car elles sont mesurées sur le même sujet ; on décompose la variabilité totale en *variabilité intra sujet* et en *variabilité inter sujet (bettween) ;*

- **facteur inter** (between) **:** varie sur des groupes distints de sujets ;

  - **analyse de covariance** à chaque fois que l'on mesure (contrôle indirect) une variable quantitative jouant un rôle de facteur et que l'on est en présence de d'autres facteurs qualitatifs ;

- **facteurs emboités** (nested) **:** si les modalités prisent par un facteur sont spécifiques aux modalités d'un autre facteur, le concept d'interaction entre les 2 facteurs emboités n'est pas défini;

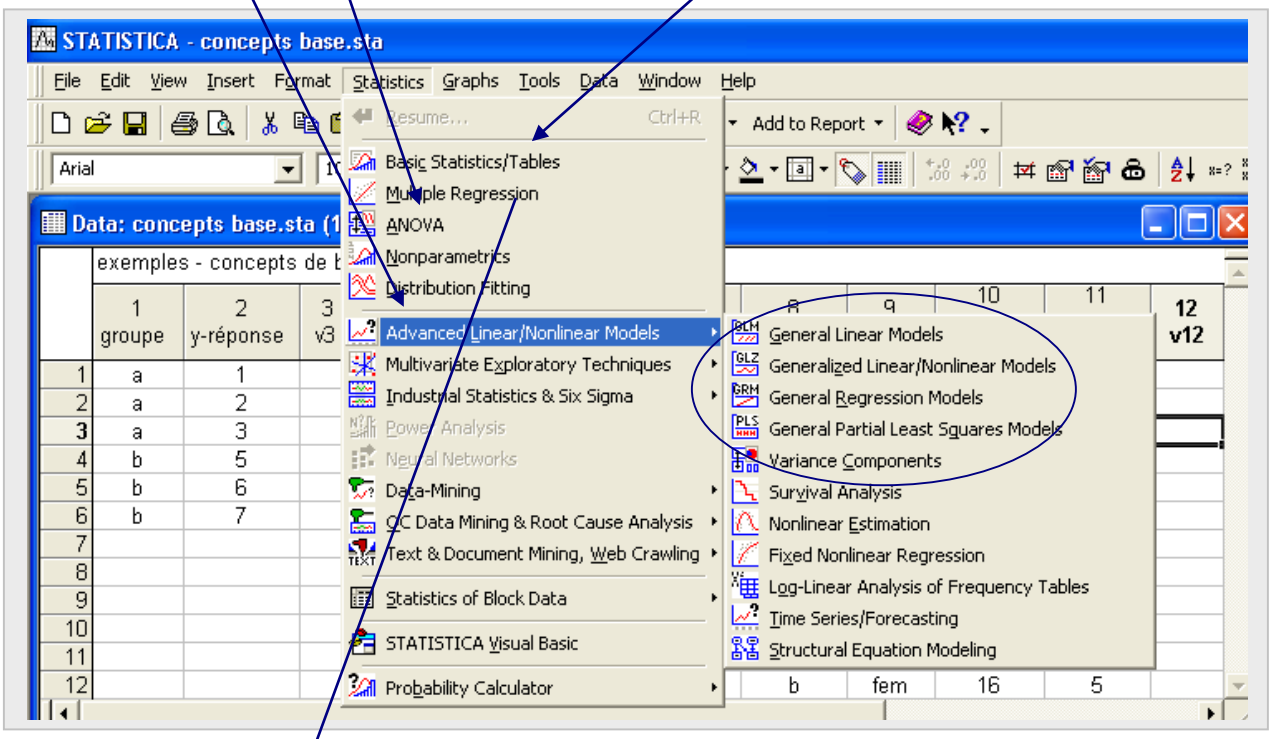
- reconnaître les **structures** (traitements, assignation) présentent dans les données est nécessaire si l'on veut réaliser la « bonne » analyse statistique ; il est plus facile si on a participé à la planification de l'étude et l'on connait le domaine (contexte) d'application ; l'utilisation d'un logiciel statistique de donne pas la réponse à ces questions ;

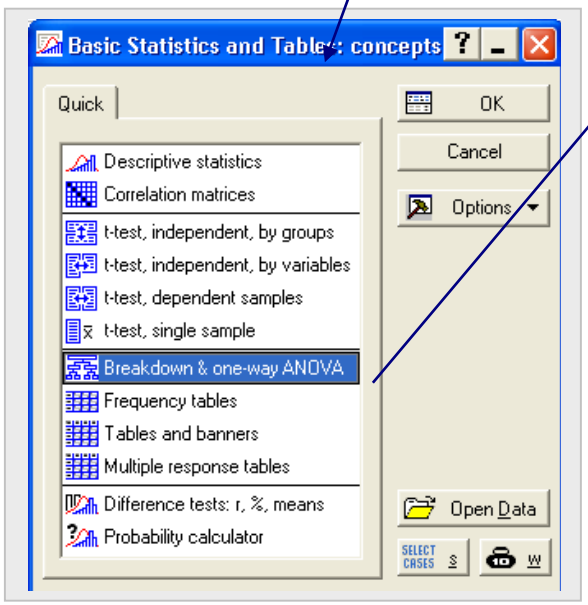
- les **designs (plans) expérimentaux** sont généralement identifiés par la structure des traitements ou **la structure d'assignation des traitements aux unités expérimentales**.
  Exemples : plan factoriel complet, plan fractionnaire, plan complètement aléatoire, plan en bloc, plan incomplet, plan en carré Latin, plan en mesures répétées, plan avec facteurs aléatoires, ….
  On distingue aussi les modèles par la méthode d'analyse statistique employée : ANOVA multifactorielle, analyse de covariance, analyse en mesures répétées, facteurs aléatoires,….

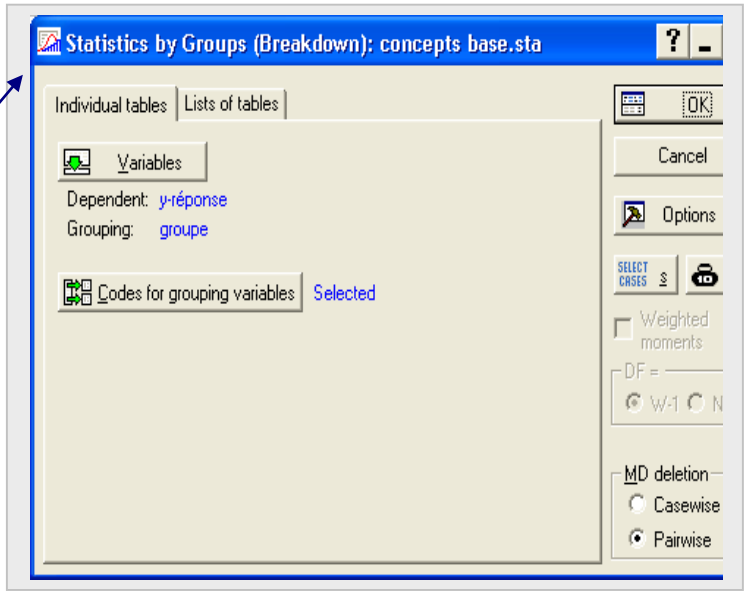
3. Le logiciel *STATISTICA* pour l'analyse de la variance

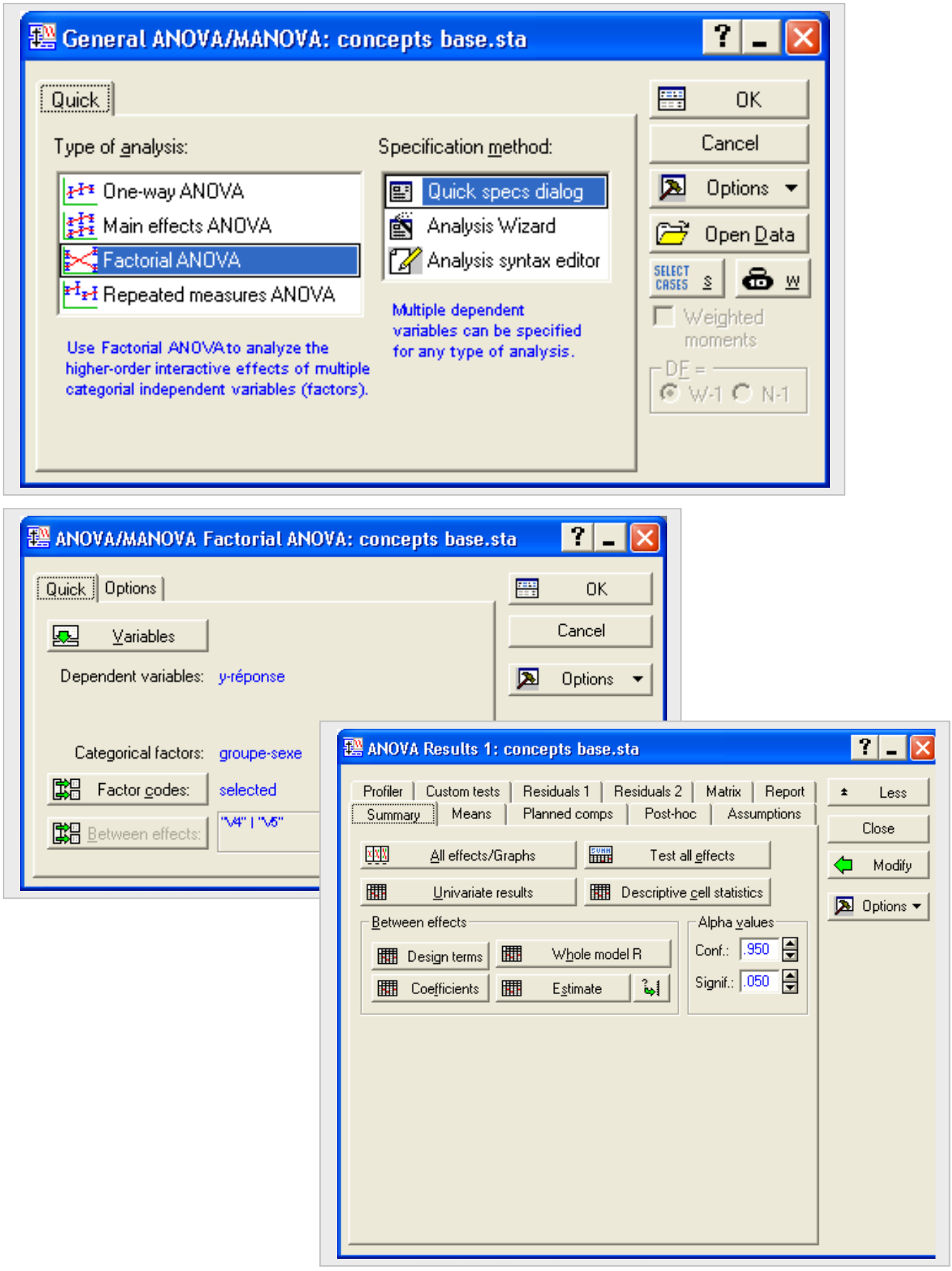
La mise en œuvre d'une analyse de variance peut se faire avec plusieurs modules de *STATISTICA*.

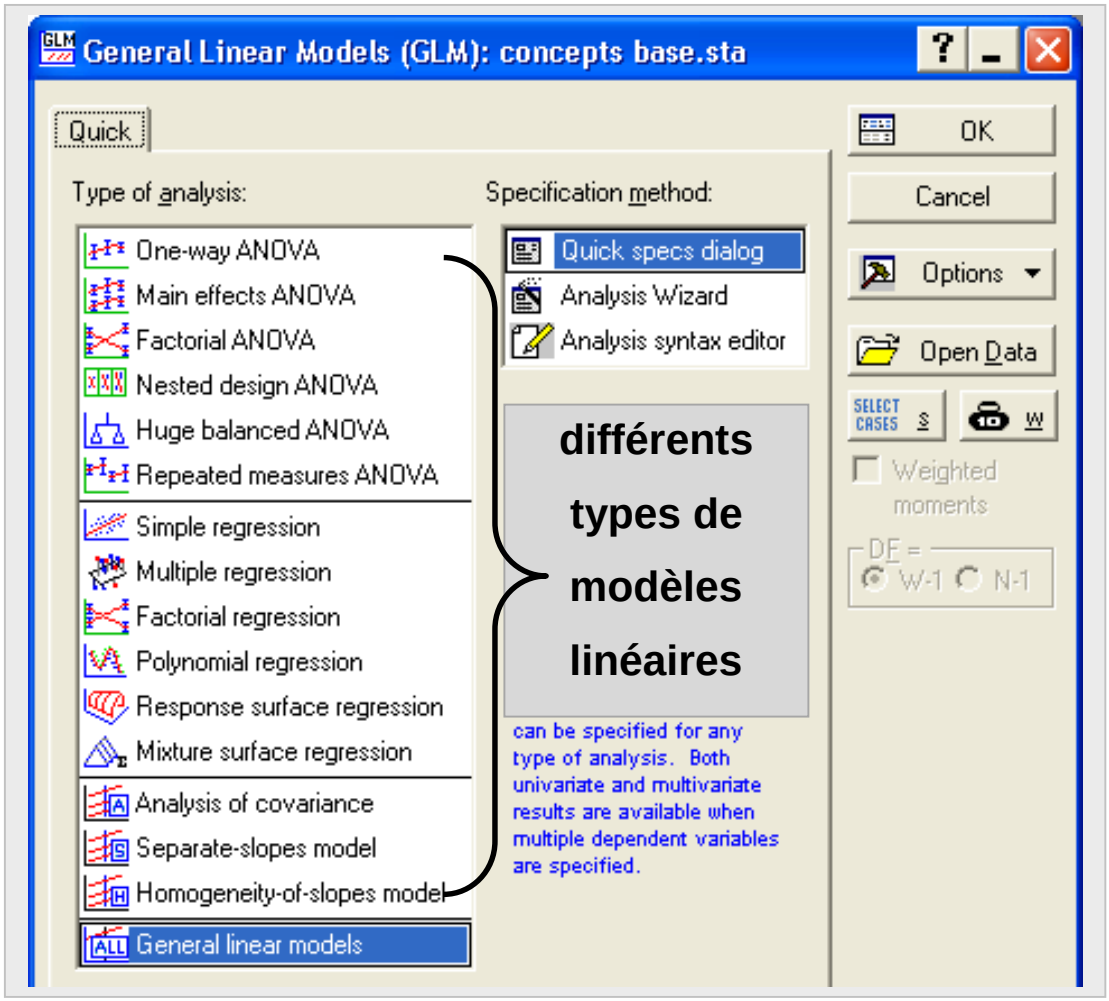
- Basic Statistics / Tables …… Breakdown & **one-Way ANOVA**
- **ANOVA**
- **Advanced Linear** / Non linear **Models**
- Experimental design (DOE)

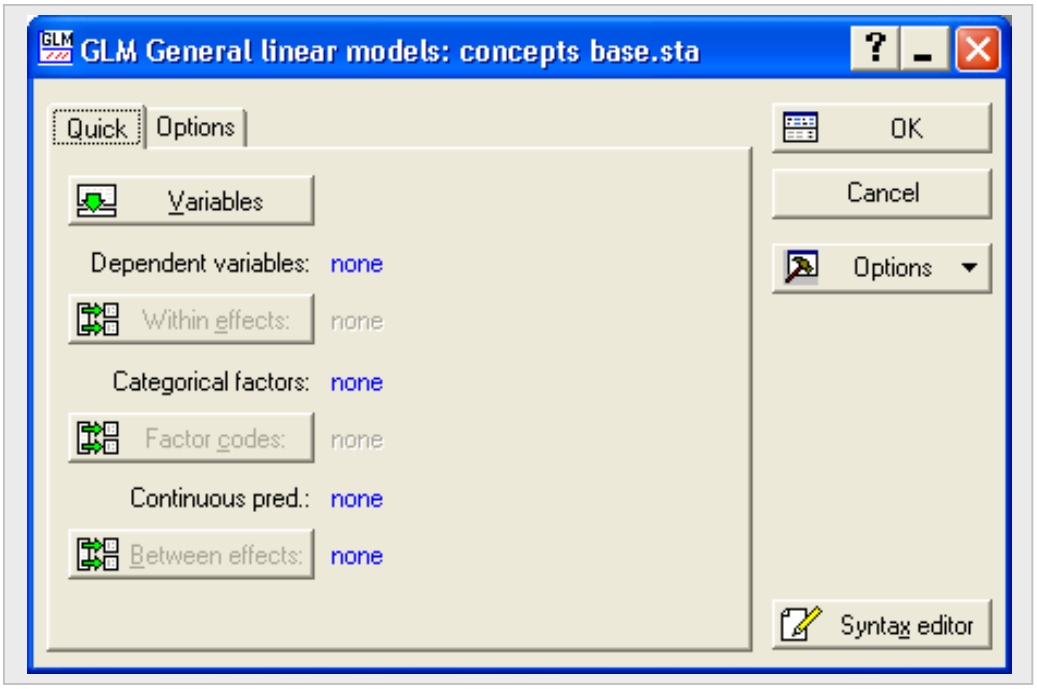












4. MODÈLES LINÉAIRES STATISTIQUES

Les *modèles d'analyse de variance* sont employés avec des *facteurs qualitatifs (catégoriques)* et les *modèles de régression* sont employés avec des *facteurs quantitatifs (continus)*. Les deux catégories de modèles sont des cas particuliers du modèle linéaire statistique.

Il est possible d'employer les modèles de régression pour analyser des données avec des facteurs qualitatifs. La stratégie consiste à définir des variables indicatrices pour représenter les modalités des variables qualitatives. Une variable indicatrice prend les valeurs 0 ou 1 pour représenter l'absence ou la présence de la modalité dans les données. On peut aussi employer un *codage à effet* avec des valeurs -1 / 0 / 1 si on est en présence d'un facteur ayant 3 modalités ou plus. Il faut comprendre les opérations de codage et leurs conséquences afin de spécifier un

Modèle linéaires

$$\mathbf{Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \varepsilon} \qquad \mathbf{(1)}$$

Y : réponse

$X_1, X_2, \ldots, X_k$: variables explicatives

$\beta_0, \beta_1, \beta_2, \ldots, \beta_k$: coefficients du modèle

ε : terme d'erreur $\varepsilon \sim N(0, \sigma^2)$

Si les variables X de l'équation (1) ne prenent que des valeurs 0 ou 1 pour représenter des variables de codage associées à des facteurs qualitatifs, l'equation (1) est appelé **modèle d'analyse de la variance.** Si l'équation (1) contient des facteurs continus et des variables à valeurs 0 et 1, on a affaire à un **modèle d'analyse de covariance.** Il ne fait pas de sens d'employer directement l'équation (1) si les facteurs X sont qualitatifs car les modalités des facteurs X ne sont pas des nombres mais des identificateurs de classe (groupe). Par exemple, si X représente le sexe du répondant dans une étude, les modalités « homme » et « femme » ne sont pas numériques. Toutefois, il est possible d'utiliser l'équation (1) si on associe des variables indicatrices (binaires) à chacune des modalités d'un facteur qualitatif. La méthode est connue sous le nom de ***codage disjonctif***. La méthode consiste à créer une variable indicatrice dont les seules valeurs sont 1 ou 0 pour chacune des modalités sauf une.

une variable qualitative avec c modalités (classes) peut être représentée

par c – 1 variables indicatrices, chacune prenant les valeurs 0 et 1

Exemple : Y : usure outil X_1 : vitesse d'opération M : manufacturier (M_1, M_2, M_3, M_4)

codage de M

Posons $X_2 = 1$ si M_1 et $X_2 = 0$ autrement

$X_3 = 1$ si M_2 et $X_3 = 0$ autrement

$X_4 = 1$ si M_3 et $X_4 = 0$ autrement

M	X_1	X_2	X_3	X_4	$X_1 + X_2 + X_3 + X_4$
M_1	x_{i1}	1	0	0	1
M_2	x_{i1}	0	1	0	1
M_3	x_{i1}	0	0	1	1
M_4	x_{i1}	0	0	0	0

modèle $\mathbf{Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4}$

M_4 : $Y = \beta_0 + \beta_1 X_1$ **(2)**

M_1 : $Y = (\beta_0 + \beta_2) + \beta_1 X_1$ **(3)**

M_2 : $Y = (\beta_0 + \beta_3) + \beta_1 X_1$ **(4)**

M_3 : $Y = (\beta_0 + \beta_4) + \beta_1 X_1$ **(5)**

L'usure de l'outil dépend de la vitesse X_1 avec la même pente β_1 pour tous les manufacturieurs. L'influence du facteur manufacturier se traduit uniquement par un changement d'ordonnées à l'origine. C'est le modèle *d'analyse de covariance* aussi appelé *modèle à pentes égales.*

Autre assignation dans le cas ou la variable catégorique prend 2 modalités seulement

On peut aussi remplacer les valeurs 0 et 1 par les valeurs -1 et 1.

Exemple: Y : ventes

X_1 : dépenses publicité F : incorporation (oui, non) M : expérience management (oui, non)

Posons $X_2 = 1$ si incorporée et $X_2 = -1$ sinon

$X_3 = 1$ si expérimenté et $X_3 = -1$ sinon

modèle du premier ordre $\mathbf{Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3}$ **(8)**

incorporation	exp. management	modèle	
oui	oui	$Y = (\beta_0 + \beta_2 + \beta_3) + \beta_1 X_1$	**(9)**
oui	non	$Y = (\beta_0 + \beta_2 - \beta_3) + \beta_1 X_1$	**(10)**
non	oui	$Y = (\beta_0 - \beta_2 + \beta_3) + \beta_1 X_1$	**(11)**
non	non	$Y = (\beta_0 - \beta_2 - \beta_3) + \beta_1 X_1$	**(12)**

Codage disjonctif complet

On utilise c variables indicatrices pour les c modalités de la variable qualitative et on **enlève le terme d'intercepte β_0** dans le modèle de régression. Cela dans le but d'éviter un problème de multi colinéarité.

Exemple : Y : usure outil X_1 : vitesse d'opération M : manufacrurier (M_1, M_2, M_3, M_4)

Posons $X_2 = 1$ si M_1 et $X_2 = 0$ autrement

$X_3 = 1$ si M_2 et $X_3 = 0$ autrement

$X_4 = 1$ si M_3 et $X_4 = 0$ autrement

$X_5 = 1$ si M_4 et $X_5 = 0$ autrement

Modèle $\mathbf{Y = \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5}$ **(13)**

M_1 : $Y = \beta_2 + \beta_1 X_1$ **(14)**

M_2 : $Y = \beta_3 + \beta_1 X_1$ **(15)**

M_3 : $Y = \beta_4 + \beta_1 X_1$ **(16)**

M_4 : $Y = \beta_5 + \beta_1 X_1$ **(17)**

Il faut faire l'ajustement de moindre carrés du **modèle de régression (13) sans l'intercepte β_0**.
Si on ajoute l'intercepte dans le modèle (13), on est en situation de multi collinéarité car

$$\mathbf{X_2 + X_3 + X_4 + X_5 = 1}$$

Une méthode pour éviter le problème de multi colinéarité est d'employer le codage à effet.

Codage à effet 1 / 0 / -1: cas de variable catégorique avec 3 modalités et plus

Lorsque le nombre de modalités de la variable catégorique prend *3 modalités et plus*, on utilise le codage à effet. Par exemple, si le facteur A prend quatre modalités a, b, c, d, on définit trois variables X_1, X_2, X_3 selon le tableau :

A	X_1	X_2	X_3
a	1	0	0
b	0	1	0
c	0	0	1
d	-1	-1	-1

On choisit arbitrairement une des modalités, ici d, pour fin de comparaison avec les autres modalités.

Exemple : Y : usure outil X_1 : vitesse d'opération M : manufacturier (M_1, M_2, M_3, M_4)

On définit 3 variables X_2, X_3, X_4 pour les 4 modalités M_1, M_2, M_3, M_4

$X_2 = 1$ si $M = M_1$ et $X_2 = 0$ si $M = M_2$ ou M_3 et $X_2 = -1$ si $M = M_4$
$X_3 = 1$ si $M = M_2$ et $X_3 = 0$ si $M = M_1$ ou M_3 et $X_3 = -1$ si $M = M_4$
$X_4 = 1$ si $M = M_3$ et $X_4 = 0$ si $M = M_1$ ou M_2 et $X_3 = -1$ si $M = M_4$

manufacurier	X_1	X_2	X_3	X_4
M_1	x_{i1}	1	0	0
M_2	x_{i1}	0	1	0
M_3	x_{i1}	0	0	1
M_4	x_{i1}	-1	-1	-1

modèle $\mathbf{Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4}$

manufacurier	modèle		
M_1	$Y = \beta_0 + \beta_2 + \beta_1 X_1$	$= \gamma_1 + \beta_1 X_1$	**(3a)**
M_2	$Y = \beta_0 + \beta_3 + \beta_1 X_1$	$= \gamma_2 + \beta_1 X_1$	**(4a)**
M_3	$Y = \beta_0 + \beta_4 + \beta_1 X_1$	$= \gamma_3 + \beta_1 X_1$	**(5a)**
M_4	$Y = \beta_0 - \beta_2 - \beta_3 - \beta_4 + \beta_1 X_1$	$= \gamma_4 + \beta_1 X_1$	**(2a)**

Les équations (2a), (3a), (4a), (5a) sont equivalentes aux équations (2), (3), (4), (5).
L'usure de l'outil dépend de la vitesse X_1 avec la même pente β_1 pour tous les manufacturieurs.
On peut mesurer l'influence du facteur manufacturier sur Y en comparant les coefficients des modèles.

comparaison	coefficients γ	coefficients β
M_1 vs M_4	$\gamma_1 - \gamma_4$	$\beta_3 + \beta_4$
M_2 vs M_4	$\gamma_2 - \gamma_4$	$\beta_2 + \beta_4$
M_3 vs M_4	$\gamma_3 - \gamma_4$	$\beta_2 + \beta_3$

Le codage à effet est employé par *STATISTICA* dans le le traitement des variables catégoriques dans les modèle linéaires.

Exemple: 2 variables catégoriques A et B avec 2 modalités chacune + effet d'interaction

X_1 variable de codage (1/-1) représente la variable catégorique A avec modalités A_1 et A_2
X_2 variable de codage (1/-1) représente la variable catégorique B avec modalités B_1 et B_2
$X_3 = X_1 * X_2$ représente l'effet l'interaction
la matrice X pour ce design est

X =		X_0	X_1	X_2	X_3
	A_1B_1	1	1	1	1
	A_1B_2	1	1	-1	-1
	A_2B_1	1	-1	1	-1
	A_2B_2	1	-1	-1	1

Exemple : 3 variables catégoriques A, B, C avec 2 modalités chacune (A_1, A_2) (B_1, B_2) (C_1, C_2)

design factoriel 2 x 2 x 2 (aussi noté 2^3) avec les effets d'ineraction simple (2 à 2)
la matrice X pour ce design est

X =

	X_0	X_1	X_2	X_3	X_1*X_2	X_1*X_3	X_2*X_3
$A_1B_1C_1$	1	1	1	1	1	1	1
$A_1B_1C_2$	1	1	1	-1	1	-1	-1
$A_1B_2C_1$	1	1	-1	1	-1	1	-1
$A_1B_2C_2$	1	1	-1	-1	-1	-1	1
$A_2B_1C_1$	1	-1	1	1	-1	-1	1
$A_2B_1C_2$	1	-1	1	-1	-1	1	-1
$A_2B_2C_1$	1	-1	-1	1	1	-1	-1
$A_2B_2C_2$	1	-1	-1	-1	1	1	1

Exemple : présence de variables catégoriques et de variables continues

Le modèle souvent employé est le modèle ***d'analyse de covariance*** aussi appelé ***modèle à pentes parallèles.*** Cela implique qu'il n'y a pas d'interaction en les variables catégoriques et les variables continues. Par exemple, supposons que nous avons une variable continue X_1 et une variable catégorique A avec 3 catégories A_1, A_2, A_3 et qu'il y a 6 observations:
$X_1 = 7$ et 4, avec $A = A_1$ $X_1 = 9$ et 3 avec $A = A_2$ X1 = 6 et 8, avec $A = A_3$.
La variable A est représentée par 2 variables de codage à effet (1/0/-1) X_2 et X_3.
Le modèle s'écrit $\mathbf{Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3}$
et la matrice X est

X =

X_0	X_1	X_2	X_3
1	**7**	**1**	**0**
1	**9**	**0**	**1**
1	**6**	**-1**	**-1**
1	**4**	**1**	**0**
1	**3**	**0**	**1**
1	**8**	**-1**	**-1**

Le modèle prend les expressions particulières suivantes:

$$\mathbf{Y = \beta_0 + \beta_1 X_1 + \beta_2 = (\beta_0 + \beta_2) + \beta_1 X_1} \quad \textbf{si } X_2 = 1 \textbf{ et } X_3 = 0$$
$$\mathbf{Y = \beta_0 + \beta_1 X_1 + \beta_3 = (\beta_0 + \beta_3) + \beta_1 X_1} \quad \textbf{si } X_2 = 0 \textbf{ et } X_3 = 1$$
$$\mathbf{Y = \beta_0 + \beta_1 X_1 - \beta_2 - \beta_3 = (\beta_0 - \beta_2 - \beta_3) + \beta_1 X_1} \quad \textbf{si } X_2 = -1 \textbf{ et } X_3 = -1$$

Exemple: supposons qu'il y a une interaction entre X_1 et la variable catégorique A

On ajoute 2 des variables d'interaction $X_4 = X_1 * X_2$ $X_5 = X_1 * X_3$

Le modèle s'écrit $\mathbf{Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5}$

et la matrice X est

$X =$

X_0	X_1	X_2	X_3	X_4	X_5
1	7	1	0	7	0
1	9	0	1	0	9
1	6	-1	-1	-6	-6
1	4	1	0	4	0
1	3	0	1	0	3
1	8	-1	-1	-8	-8

Le modèle prend les expressions particulières suivantes:

$$\mathbf{Y = \beta_0 + \beta_1 X_1 + \beta_2 + \beta_4 X_4 = (\beta_0 + \beta_2) + (\beta_1 + \beta_4) X_1 = \gamma_{01} + \gamma_{11} X_1} \quad \textbf{si } X_2 = 1 \textbf{ et } X_3 = 0$$

$$\mathbf{Y = \beta_0 + \beta_1 X_1 + \beta_3 + \beta_5 X_5 = (\beta_0 + \beta_3) + (\beta_1 + \beta_5) X_1 = \gamma_{02} + \gamma_{12} X_1} \quad \textbf{si } X_2 = 0 \textbf{ et } X_3 = -1$$

$$\mathbf{Y = \beta_0 + \beta_1 X_1 - \beta_2 - \beta_3 - \beta_4 X_1 - \beta_5 X_1 = (\beta_0 - \beta_2 - \beta_3) + (\beta_1 - \beta_4 - \beta_5) X_1}$$
$$\mathbf{= \gamma_{03} + \gamma_{13} X_1} \quad \textbf{si } X_2 = -1 \textbf{ et } X_3 = -1$$

Les pentes $\gamma_{11}, \gamma_{12}, \gamma_{13}$ ainsi que les ordonnées à l'origine $\gamma_{01}, \gamma_{02}, \gamma_{03}$ sont différentes.

Autres applications des variables indicatrices

- Modèles pour facteurs emboités
- modèle de régression par morceaux ;
- discontinuité dans un modèle de régression ;
- séries chronologiques ;
- remplacement de variables quantitatives avec des banques de données abondantes :
  plusieurs centaines / milliers d'observations

5. ANALYSE DE LA VARIANCE AVEC UN FACTEUR

Contexte

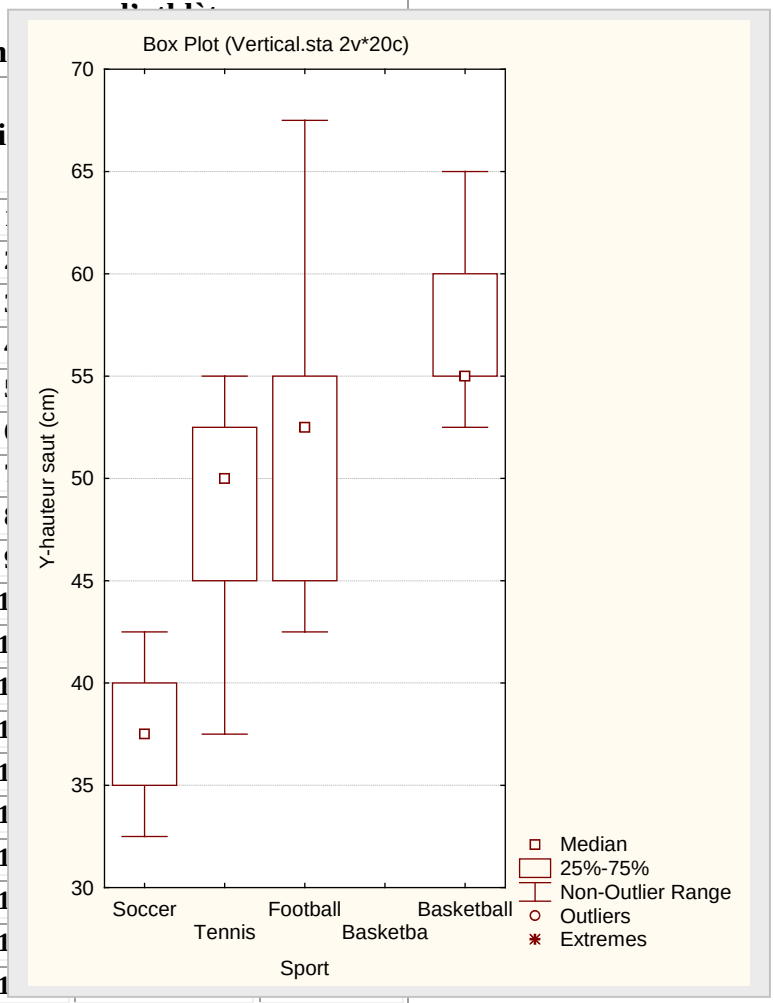
facteur A : modalités (niveaux) 1, 2, … , g g : nombre de modalités (niveaux, groupes)

cas 1 : classification des unités : étude observationnelle / rétrospective
cas 2 : expérimentation : modalités sont affectés au hasard aux unités

facteur fixe : les conclusions s'appliquent à ces modalités - **modèle à effets fixes**
on s'intéresse aux moyennes corespondant à ces modalités
facteur aléatoire : échantillon au hasard d'une population de modalités
les conclusions s'appliquent à cette population – **modèle à effets aléatoires**
on s'intéresse aux composantes de la variance
réponse Y : variable quantitative

Exemple 4 : saut en hauteur

comparaison de 4 groupes



20	**Basketball**	**55**

données

niveau	observations y_{ij}			moyennes	# obs.	variances
1	y_{11}	y_{12} ...	$y_{1\,n1}$	$\overline{y}_{1.}$	n_1	s_1^2
2	y_{21}	y_{22}	$y_{2\,n2}$	$\overline{y}_{2.}$	n_2	s_2^2
.	.....					
i	y_{i1}	y_{i2} ...	$y_{i\,ni}$	$\overline{y}_{i.}$	n_i	s_i^2
.	.....					
g	y_{g1}	y_{g2} ...	$y_{g\,ng}$	$\overline{y}_{g.}$	n_g	s_g^2
tous				$\overline{y}_{..}$	N	

$$y_{i.} = \sum_j y_{ij}$$

$$y_{..} = \sum_i \sum_j y_{ij}$$

$$\overline{y}_{i.} = y_{i.} / n_i$$

$$N = \sum n_i$$

$$\overline{y}_{..} = y_{..} / N$$

$$SS_i = \sum (y_{ij} - \overline{y}_{i.})^2$$

$$s_i^2 = SS_i / (n_i - 1)$$

Modèle à moyennes de cellules (pas d'effet général)

$$Y_{ij} = \mu_i + \varepsilon_{ij} \qquad i = 1, 2, \ldots, g \qquad j = 1, 2, \ldots, n_i \qquad (20)$$

Y_{ij} : valeur de la variable de réponse j-ième essai modalité i du facteur

μ_i : paramètre - moyenne de la cellule i

ε_{ij} : erreurs aléatoires indépendantes distribuées $N(0, \sigma^2)$

conséquences

$$E(Y_{ij}) = \mu_i$$

$$Var(Y_{ij}) = Var(\varepsilon_{ij}) = \sigma^2$$

$Y_{ij} \sim N(\mu_i, \sigma^2)$ ~ : symbole pour « distribuée comme »

Modèle linéaire général : notation matricielle

$$Y = X\beta + \varepsilon \qquad (21)$$

Y : vecteur N x 1 d'observations (données)

X : matrice du modèle N x p
fonctions des k variables (facteurs) explicatives

β : vecteur p x 1 de paramètres (statistiques) à estimer

ε : vecteur N x 1 d'erreur + hypothèse de normalité

remarque : la *linéarité* est relative à β et non pas aux variables $X_1, X_2, \ldots$

questions : plan collecte données, estimation de β, valiation du modèle, tests d'hypothèses, etc.

Modèle linéaire GÉNÉRALISÉ : $Z = g(Y) = X\beta + \varepsilon$ g : fonction de lien

distribution	fonction de lien	nom
normale	$Z = g(Y) = Y$	identité
log normale	$Z = g(y) = \log(Y)$	log
	$Z = g(Y) = Y^a$	puissance
binaire (0-1)	$Z = g(Y) = \log(Y / (1-Y))$	logit
	$Z = g(Y) = \text{invnorm}(Y)$	probit
	$Z = g(Y) = \log(-\log(1-Y))$	logit complémentaire

Ces transformations permettent d'étendre l'appplication du modèle linéaire à des variables de réponse Y dont la nature et la distribution ne suit pas le cas classique.

Exemple 4 (suite) : hauteur modèle sans intercepte (sans effet général)

$g = 4$ $n_i = n = 5$ $N = 20$ $Y : 20 \times 1$ $X : 20 \times 4$ $\varepsilon : 20 \times 1$

$$
\begin{bmatrix} Y11 \\ Y12 \\ . \\ . \\ Y15 \\ Y21 \\ . \\ . \\ . \\ . \\ . \\ . \\ . \\ . \\ . \\ Y41 \\ . \\ . \\ Y44 \\ Y45 \end{bmatrix}
=
\begin{bmatrix} 38 \\ 43 \\ . \\ . \\ . \\ . \\ . \\ . \\ . \\ . \\ . \\ . \\ . \\ . \\ . \\ . \\ . \\ . \\ 53 \\ 55 \end{bmatrix}
=
\begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \end{bmatrix}
\begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \\ \mu_4 \end{bmatrix}
+
\begin{bmatrix} \varepsilon_{11} \\ \varepsilon_{12} \\ . \\ . \\ . \\ . \\ . \\ . \\ . \\ . \\ . \\ . \\ . \\ . \\ . \\ . \\ . \\ . \\ \varepsilon_{44} \\ \varepsilon_{45} \end{bmatrix}
$$

$$\mathbf{Y = X\beta + \varepsilon}$$

<u>Ajustement du modèle</u> : principe des moindres carrés

$$\textbf{Minimum} \quad Q = \sum\sum (Y_{ij} - \mu_i)^2$$

$$\textbf{Solution} \quad \widehat{\mu_i} = \overline{Y}_{i.} \quad \textbf{et} \quad \widehat{Y_{ij}} = \overline{Y}_{i.}$$

<u>Tableau d’analyse de la variance</u> : ANOVA

SOURCE	SOMME CARRÉS (SS)	deg. lib. (df)	CARRÉ MOYEN (MS)	F
Traitements	$SS_{trait} = \sum n_i \sum (\overline{y}_{i.} - \overline{\overline{y}}_{..})^2$	$g - 1$	MS_{trait}	$F_0 = MS_{trait} / MSE$
Erreur	$SS_{erreur} = \sum \sum (y_{ij} - \overline{y}_{i.})^2$	$N - g$	$MSE = \widehat{\sigma^2}$	
Total	$SS_{total} = \sum \sum (y_{ij} - \overline{\overline{y}}_{..})^2$	$N - 1$	$(N = \sum n_i)$	

$H_0 : \mu_1 = \mu_2 = \quad = \mu_g$ hypothèse nulle d’égalité des moyennes

Rejet de H_0 si $F_0 > F(g - 1, N - g, 1 - \alpha)$ (loi F de Fisher-Snedecor)

où $F(g - 1, N - g, 1 - \alpha)$ est le $(1 - \alpha)^{ième}$ percentile loi F (g -1, N-g)
avec $g - 1$ degrés de liberté au numérateur
et $N - g$ degrés de liberté au dénominateur

Équation fondamentale de la décomposition de la variabilité

$$\sum\sum (y_{ij} - \overline{y}_{..})^2 = \sum n_i \sum (\overline{y}_{i.} - \overline{y}_{..})^2 + \sum\sum (y_{ij} - \overline{y}_{i.})^2$$

$$SS_{total} = SS_{traitement} + SS_{erreur}$$

$$df_{total} = N - 1 = (g - 1) + (N - g) = df_{traitement} + df_{erreur}$$

$$E(MSE) = \sigma^2$$

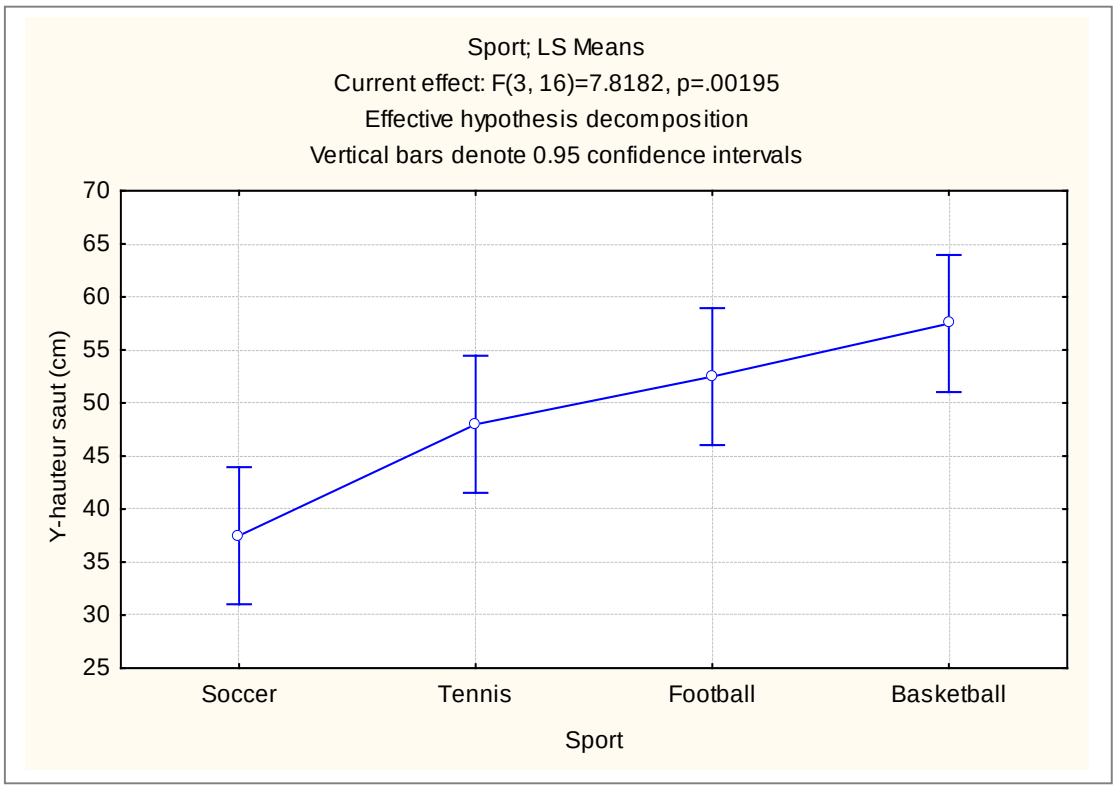
$$E(MS_{trait}) = \sigma^2 + \sum n_i (\mu_i - \overline{\mu})^2 / (g - 1)$$

$$\text{où} \quad \overline{\mu} = \sum n_i \mu_i / (N - 1)$$

Exemple 4 : suite

ANOVA

	Degr. Of Freedom	Y-hauteur saut SS	Y-hauteur saut MS	Y-hauteur saut F	Y-hauteur saut p
Intercept	1	47775,31	47775,31	1029,502	0,000000
Sport	3	1088,44	362,81	7,818	0,001952
Erreur	16	742,50	46,41		
Total	19	1830,94			



Modèle à type d'effets : effet général + effet différentiel

$$\mu_i = \mu + (\mu_i - \mu) = \mu + \tau_i$$
$$Y_{ij} = \mu + \tau_i + \varepsilon_{ij} \qquad i = 1, 2, \ldots, g \qquad j = 1, 2, \ldots, n_i \qquad (22)$$

Y_{ij} : valeur de la variable de réponse j-ième essai modalité i du facteur
μ : effet général
τ_i : effet différentiel de la modalité i du facteur
ε_{ij} : erreurs aléatoires indépendantes distribuées $N(0, \sigma^2)$

conséquences
$$E(Y_{ij}) = \mu + \tau_i$$
$$Var(Y_{ij}) = Var(\varepsilon_{ij}) = \sigma^2$$
$$Y_{ij} \sim N(\mu + \tau_i, \sigma^2)$$

Définition de μ : 2 possibilités

définition 1
$$\mu = \sum \mu_i / g \qquad (23)$$
$$\sum \tau_i = 0 \qquad (24)$$

définition 2
$$\mu = \sum \omega_i \mu_i \qquad \sum \omega_i = 1 \qquad (25)$$
$$\sum \omega_i \tau_i = 0 \qquad (26)$$

exemples de la définition 2

exemple A : parc véhicules automobiles composée de
50 % compactes 30% berlines 20% VUS
Y : consommation essence
$$\mu = 0.5 * \mu_1 + 0.3 * \mu_2 + 0.2 * \mu_3$$

exemple B : $\omega_i = n_i / N$
si $n_i = n$ alors $\omega_i = 1/g$ alors (25) donne (23)

Hypothèse $H_0 : \mu_1 = \mu_2 = \ldots = \mu_g$ $H_a : \mu_i$ pas tous égaux

équivalent à $H_0 : \tau_1 = \tau_2 = \ldots = \tau_g = 0$ $H_a : \tau_i \neq 0$ pour au moins un i

Approche par régression avec un codage à effet
$$Y_{ij} = \mu + \tau_i + \varepsilon_{ij} \qquad \sum \tau_i = 0 \qquad (27)$$
$$\tau_g = -\tau_1 - \tau_2 - \ldots - \tau_{g-1}$$

Exemple 4 (suite): saut en hauteur $g = 4$ $n_1 = n_2 = n_3 = n_4 = 5$ $N = 20$

Posons
$$X_{ijt} = \begin{cases} 1 & \text{si observation provient du groupe } i = 1, 2, \ldots, g-1 \\ -1 & \text{si observation provient groupe g} \\ 0 & \text{autrement} \end{cases}$$

(27) devient
$$Y_{ij} = \mu + \tau_1 X_{ij1} + \tau_2 X_{ij2} + \tau_3 X_{ij3} + \varepsilon_{ij} \qquad (28)$$

	X_{ij0}	X_{ij1}	X_{ij2}	X_{ij3}		
38	1	1	0	0	μ	ε_{11}
43	1	1	0	0	τ_1	.
33	1	1	0	0		.
40	1	1	0	0	τ_2	.
35	1	1	0	0		.
45	1	0	1	0	τ_3	.
53	1	0	1	0		.
38	1	0	1	0		.
55	1	0	1	0		.
50	1	0	1	0		.
55	1	0	1	0		.
68	1	0	0	1		.
43	1	0	0	1		.
45	1	0	0	1		.
53	1	0	0	1		.
60	1	0	0	1		.
65	1	-1	-1	-1		.
55	1	-1	-1	-1		.
53	1	-1	-1	-1		.
55	1	-1	-1	-1		.
	1	-1	-1	-1		ε_{45}

$$Y = X \beta + \varepsilon$$

id	Sport	i	j	X1	X2	X3	Y-hauteur saut (cm)
1	Soccer	1	1	1	0	0	38
2	Soccer	1	2	1	0	0	43
3	Soccer	1	3	1	0	0	33
4	Soccer	1	4	1	0	0	40
5	Soccer	1	5	1	0	0	35
6	Tennis	2	1	0	1	0	45
7	Tennis	2	2	0	1	0	53
8	Tennis	2	3	0	1	0	38
9	Tennis	2	4	0	1	0	55
10	Tennis	2	5	0	1	0	50
11	Football	3	1	0	0	1	55
12	Football	3	2	0	0	1	68
13	Football	3	3	0	0	1	43
14	Football	3	4	0	0	1	45
15	Football	3	5	0	0	1	53
16	Basketball	4	1	-1	-1	-1	60
17	Basketball	4	2	-1	-1	-1	65
18	Basketball	4	3	-1	-1	-1	55
19	Basketball	4	4	-1	-1	-1	53
20	Basketball	4	5	-1	-1	-1	55

Analyse par régression : exemple 4

R= 0,77	R²= 0,594	Adjusted R²= 0,518	F(3,16) = 7,82	p = 0,00195		
	Beta	Std.Err.	B	Std.Err.	t(16)	p-level
Intercept			48.88	1.52	32.09	0.0000
X1	-0.841	0.195	-11.38	2.64	-4.31	0.0005
X2	-0.065	0.195	-0.88	2.64	-0.33	0.7445
X3	0.268	0.195	3.63	2.64	1.37	0.1884

Analysis of Variance; DV: Y-hauteur saut (cm)					
	Sums of	df	Mean	F	p-level
Regress.	1088.44	3	362.81	7.8182	0.0020
Residual	742.50	16	46.41		
Total	1830.94				

Modèle à type d'effet pondérés : $\omega_i = n_i / N$

$$Y_{ij} = \mu + \tau_1 X_{ij1} + \tau_2 X_{ij2} + \tau_3 X_{ij3} + \varepsilon_{ij} \qquad (29)$$

$$\sum (n_i / N)\, \tau_i = 0$$

$$\tau_g = (-n_1/n_g)\,\tau_1 + (-n_2/n_g)\,\tau_2 + \ldots + (-n_{g-1}/n_g)\,\tau_{g-1} \qquad (30)$$

Posons

$$X_{ijt} = \begin{cases} 1 \text{ si l'observation provient du groupe } i = 1, 2, .., g-1 \\ (-n_i/n_g) \text{ si l'observation provient groupe } i = g \\ 0 \text{ autrement} \end{cases}$$

Exemple 5: ventes selon 4 designs d'emballage (Kutner et all 5 ed. p 686)

design 1 = 3 couleurs + 0 personnages BD design 3 = 3 couleurs + personnages BD
design 2 = 5 couleurs + 0 personnages BD design 4 = 5 couleurs + personnages BD

$g = 4$ $n_1 = 5$ $n_2 = 5$ $n_3 = 4$ $n_4 = 5$ $N = 19$

	design	magasin	Y-caisses vendues	X1	X2	X3
1	1	1	11	1	0	0
2	1	2	17	1	0	0
3	1	3	16	1	0	0
4	1	4	14	1	0	0
5	1	5	15	1	0	0
6	2	1	12	0	1	0
7	2	2	10	0	1	0
8	2	3	15	0	1	0
9	2	4	19	0	1	0
10	2	5	11	0	1	0

	design	magasin	Y-caisses vendues	X1	X2	X3
11	3	1	23	0	0	1
12	3	2	20	0	0	1
13	3	3	18	0	0	1
14	3	4	17	0	0	1
15	4	1	27	-1	-1	-0.8
16	4	2	33	-1	-1	-0.8
17	4	3	22	-1	-1	-0.8
18	4	4	26	-1	-1	-0.8
19	4	5	28	-1	-1	-0.8

Regression Summary for Dependent Variable: Y-caisses vendues R=0.8877 R²=0.7880 Adjusted R²= 0.7456 F(3,15)=18.59 p = 0,00003						
	Beta	Std.Err.	B	Std.Err.	t(15)	p-level
Intercept			18.63	0.75	25.01	0.0000
X1	-0.47	0.1443	-4.03	1.25	-3.23	0.0056
X2	-0.61	0.1443	-5.23	1.25	-4.20	0.0008
X3	0.09	0.1417	0.87	1.44	0.60	0.5562

Analysis of Variance; DV: Y-caisses vendues					
	Sums of	df	Mean	F	p-level
Regress.	588.22	3	196.07	18.59	0.00003
Residual	158.20	15	10.55		
Total	746.42				

Nombre d'obsevations pour des études avec un facteur

La puissance du test F = probabilité de rejeter H_0 si H_a est vraie

$$\text{Puissance} = \text{Prob}\,(F > F_{1-\alpha,\, g-1,\, N-g} \mid \Phi) \qquad (31)$$

$$= H(\Phi, g, N, \alpha,) \quad \text{distribution F non centrale de paramètre } \Phi$$

$$\Phi = (1/\sigma)\left[\sum n_i (\mu_i - \bar{\mu})^2 / g\right]^{0.5} \qquad \bar{\mu} = \sum \mu_i / g \qquad (32)$$

Φ : paramètre de non centralité

$$\text{Si } n_i = n \qquad \Phi = (1/\sigma)\left[(n/g)\sum (\mu_i - \bar{\mu})^2 / g\right]^{0.5}$$

$$\Delta = \max(\mu_i) - \min(\mu_i) \qquad n = h(g, \alpha, \Delta/\sigma)$$

puissance = 0,90 n = nombre d'observations dans chaque groupe

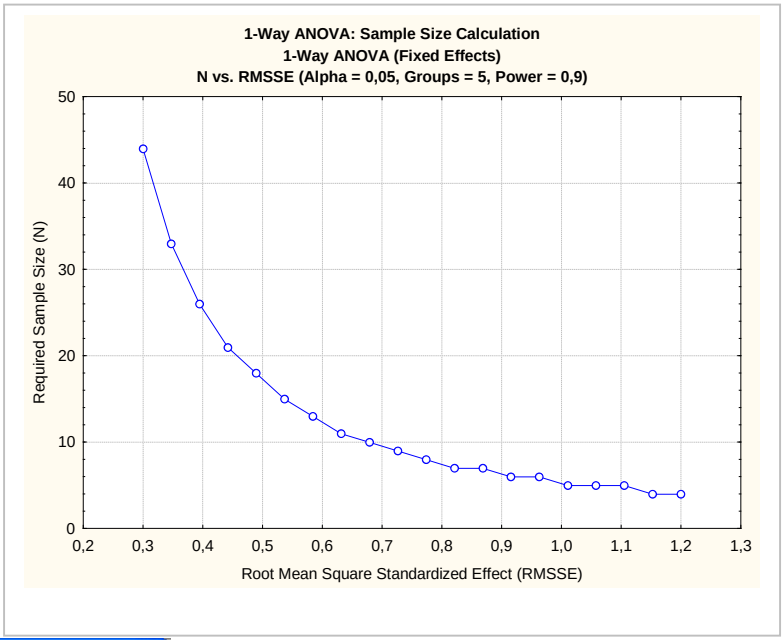
0,90 est souvent la valeur utilisée

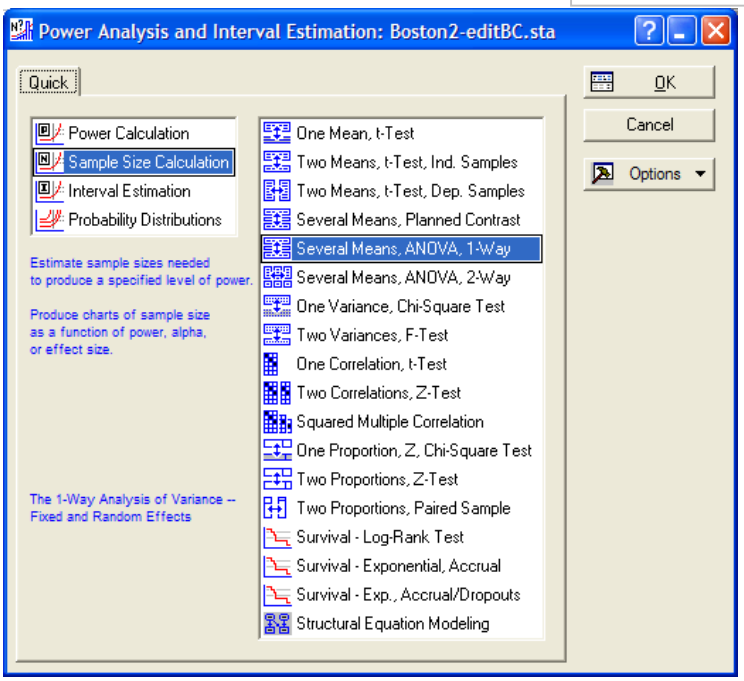
	Δ/σ = 1	Δ/σ = 1	Δ/σ = 1	Δ/σ = 1.5	Δ/σ = 1.5	Δ/σ = 1.5	Δ/σ = 2	Δ/σ = 2	Δ/σ = 2
g	α = 0,1	α = 0,05	α = 0,01	α = 0,1	α = 0,05	α = 0,01	α = 0,1	α = 0,05	α = 0,01
2	18	23	32	9	11	15	6	7	10
3	22	27	37	11	13	18	7	8	11
4	25	30	40	12	14	19	7	9	12
5	27	32	43	13	15	20	8	9	12
6	29	34	46	14	16	21	8	10	13
7	31	36	48	14	17	22	9	10	13
8	32	38	50	15	18	23	9	11	14
9	33	40	52	16	18	24	9	11	14
10	35	41	54	16	19	25	10	11	15

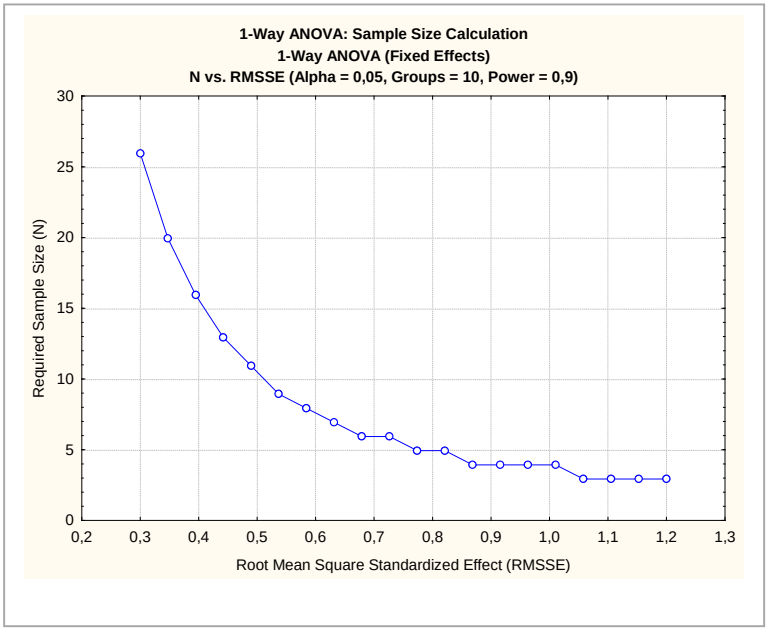
puissance = 0,95 **n = nombre d'observations dans chaque groupe**

	Δ/σ=1	Δ/σ=1	Δ/σ=1	Δ/σ=1.5	Δ/σ=1.5	Δ/σ=1.5	Δ/σ=2	Δ/σ=2	Δ/σ=2
g	**α = 0,1**	**α = 0,05**	**α = 0,01**	**α = 0,1**	**α = 0,05**	**α = 0,01**	**α = 0,1**	**α = 0,05**	**α = 0,01**
2	**23**	**27**	**38**	**11**	**13**	**18**	**7**	**8**	**11**
3	**27**	**32**	**43**	**13**	**15**	**20**	**8**	**9**	**12**
4	**30**	**36**	**47**	**14**	**17**	**22**	**9**	**10**	**13**
5	**33**	**39**	**51**	**15**	**18**	**23**	**9**	**11**	**14**
6	**35**	**41**	**53**	**16**	**19**	**25**	**10**	**11**	**15**
7	**37**	**43**	**56**	**17**	**20**	**26**	**10**	**12**	**15**
8	**39**	**45**	**58**	**18**	**21**	**27**	**11**	**12**	**16**
9	**40**	**47**	**60**	**19**	**22**	**28**	**11**	**13**	**16**
10	**42**	**48**	**62**	**19**	**22**	**29**	**11**	**13**	**17**

Utilisation de *STATISTICA* pour le calul de la taille échantillonnale n







6. ANALYSE DES MOYENNES ET COMPARAISONS MULTIPLES

Si le test F est significatif cela veut dire que les moyennes sont statistiquement différentes. Peut-on dire plus ? Sont-elles toutes différentes ? Sinon, quelle moyenne diffère de quelle autre?
Peut-on faire des comparaisons (contrastes) entre des groupes de moyennes?

Toutes ces questions constituent l'analyse a posteriori (post-hoc) des moyennes. Elles font intervenir le problème de comparaisons multiples sur le même ensemble de données. Il faut contrôler les risques associés à ces comparaisons multiples.

On veut contrôler le risque et avoir un coefficient de confiance global de $1 - \alpha$ sur l'ensemble des comparaisons (tests). En effet, si on fait un nombre de k comparaisons, chacune avec un coefficient de confiance de $1 - \alpha$, alors le coefficient de confiance global sur l'ensemble des k comparaisons diminue. En d'autres termes, plus on augmente le nombre de comparaisons (tests), plus on augmente les chances de conclure à tort.

Le tableau suivant illustre le problème.

nombre de modalités g	nombre de comparaisons $k = g*(g-1)/2$	coefficent de confiance global $(1 - \alpha)^k$	$1 - \alpha = 0,95$
2	1	$1 - \alpha$	0,95
3	3	$(1 - \alpha)^3$	0,86
4	6	$(1 - \alpha)^6$	0,735
5	10	$(1 - \alpha)^{10}$	0,60
6	15	$(1 - \alpha)^{15}$	0,46
8	28	$(1 - \alpha)^{28}$	0,24
10	45	$(1 - \alpha)^{45}$	0,10

Il est impératif que le risque global soit contrôlé et équivalent au risque de l'ANOVA. On distingue 2 catégories de tests :

- **tests (comparaisons) planifiés <u>avant</u> l'exécution des calculs**
- **tests suggérés <u>après</u> l'analyse (post hoc, a posteriori) (« data snooping »)**

<u>A - intervalle de confiance pour une moyenne particulière</u>

$$\mu_i : \overline{Y}_{i.} \pm t(1 - \alpha/2, N - g) * MSE^{0.5} * [1/n_i]^{0,5} \qquad (33)$$

où $t(1 - \alpha/2, N - g)$: $(1 - \alpha/2)^{\text{ième}}$ **percentile loi T de Student avec (N - g) degrés de liberté**

$1 - \alpha$ **: coefficient de confiance**

<u>exemple</u> : données sur le design 1 d'emballage

$\overline{Y}_{1.} = 14,6$ $\quad MSE = 10,55$ $\quad 1- \alpha = 0.95$ $\quad t(0.95, 15) = 2,13$

$$11,5 \leq \mu_1 \leq 17,7$$

B – intervalle de confiance pour la différence entre 2 moyennes

$i \neq i'$

$$\mu_i - \mu_{i'} : (\bar{Y}_{i.} - \bar{Y}_{i'.}) \pm t(1 - \alpha/2, N - g) * MSE^{0.5} * [(1/n_i) + (1/n_{i'})]^{0.5} \qquad (34)$$

exemple : données sur le design d'emballage

design	1	2	3	4
moyenne	14,6	13,4	19,5	27,2
nombre obs.	5	5	4	5

$1 - \alpha = 0,95$ $\quad \mu_3 - \mu_4 : (19,5 - 27,2) \pm 2,13 * 10,55^{0.5} * [(1/5) + (1/4)]^{0.5}$

$$-12,3 \leq \mu_3 - \mu_4 \leq -3,7$$

C – contraste

$$L = \sum c_i \mu_i \qquad \sum c_i = 0 \quad \text{def. contraste} \qquad (35)$$

la différence entre 2 moyennes est un cas particulier de contraste

$$\hat{L} = \sum c_i \bar{Y}_{i.}$$

$$s(\hat{L}) = MSE^{0.5} * [\sum c_i^2 / n_i]^{0.5}$$

$(\hat{L} - L) / s(\hat{L})$ suit une loi de Student avec $(N - g)$ degrés de liberté (36)

application : $H_0 : L = 0$ vs $H_a : L \neq 0$

rejet de H_0 si $[\hat{L} / (s(L)\hat{\ })] > t(1 - \alpha/2, N - g)$ (37)

exemple : données sur le design d'emballage

$L = 0,5*(\mu_1 + \mu_2) - 0,5*(\mu_3 + \mu_4)$

L : comparaison 3 couleurs vs 5 couleurs

$\hat{L} = -9,35$ $\qquad \sum c_i^2 / n_i = 0,2125$

$s(\hat{L}) = 2,24$

$-12,5 \leq L \leq -6,2$

Procédures d'inférences simultanées (comparaisons multiples)

Les méthodes A – B – C ont deux limitations :

1. **le coefficient de confiance 1 – α et le seuil α d'un test s'applique à UN test seulement.**

2. **le test ou la comparaison n'a pas été suggéré par les données (« data snooping »).**

La solution de ce problème est d'utiliser une procédure de comparaison multiple qui inclut toutes les inférences possibles qui peuvent être anticipées et d'intérêt après que les données furent examinées. Par exemple, on peut s'inéresser à toutes les comparaisons définies par les différences entre toutes les paires de moyennes. Il existe 3 procédures pour faire de l'inférence après avoir vu les données sans affecter le coefficient de confiance:

- **méthode de Tukey («HSD = Honest Significant Differences »)**
- **méthode de Scheffé pour les contrastes**
- **méthode de Bonferronni pour les comparaisons prédéfinies**

Méthode de Tukey

La méthode est dédiée sur les comparaisons (contrastes spécifiques) définies par les différences entre toutes les moyennes prises 2 à 2 :

$$H_0 : \mu_i - \mu_{i'} = 0 \quad \text{vs} \quad H_a : \mu_i - \mu_{i'} \neq 0$$

Le test repose sur la distribution « Studentized Range » dont la définition suit.

$Y_1, Y_2, \ldots, Y_g$: **g observations indépendantes d'une population** $N(\mu, \sigma^2)$

$W = \max(Y_1, Y_2, .., Y_g) - \min(Y_1, Y_2, .., Y_g)$: **étendue ("range")**

S^2 : **estimation de** σ^2 **basée sur** υ **degrés de liberté**

$Q(g, \upsilon) = W / S$: **« studentized range »**

Extrait du tableau complet du 95ième percentile de la distribution: q (0,95 ; g , υ)
(Kutner et all 5 ed. p. 1334)

g : nombre de groupes **υ : degrés de liberté**

υ	g = 2	g = 3	g = 4	g = 5	g = 10	g = 15	g = 20
2	6.08	8.33	9.80	10.9	14.0	15.7	16.8
5	3.64	4.60	5.22	5.67	6.80	7.72	8.21
10	3.15	3.88	4.33	4.65	5.60	6.11	6.47
20	2.95	3.58	3.96	4.23	5.01	5.43	5.71
40	2.86	3.44	3.79	4.04	4.73	5.11	5.36
60	2.83	3.40	3.74	3.98	4.65	5.00	5.24
120	2.80	3.36	3.68	3.92	4.56	4.90	5.13
infini	2.77	3.31	3.63	3.86	4.47	4.80	5.01

Méthode de Tukey

$$D = \mu_i - \mu_{i'} \qquad \hat{D} = \bar{Y}_{i.} - \bar{Y}_{i'.}$$

$$s^2(\hat{D}) = MSE * [(1/n_i) + (1/n_{i'})] \quad (38)$$

$$T = 0.707 * q(1 - \alpha; g, N - g) \quad (39)$$

Intervalle de confiance simultané de toutes les différences avec coefficient confiance $1 - \alpha$

$$D : \hat{D} \pm T * s(D) \quad (40)$$

exemple : données de design d'emballage

Tukey HSD test; variable Y-caisses vendues
probabilities for Post Hoc Tests Error: Between MS = 10.547, df = 15

	empaquettage	{1} 14.6	{2} 13.4	{3} 19.5	{4} 27.2
1	1		0.9354	0.1550	0.0003
2	2	0.9354		0.0584	0.0002
3	3	0.1550	0.0584		0.0143
4	4	0.0003	0.0002	0.0143	

Méthode de Scheffé

$$L = \sum c_i \mu_i \qquad \sum c_i = 0$$

$$H_0 : L = 0 \quad vs \quad H_a : L \neq 0$$

$$\hat{L} = \sum c_i \bar{Y}_{i.}$$

$$s(\hat{L}) = MSE^{0.5} * [\sum c_i^2 / n_i]^{0.5}$$

$$S = (g - 1) * F(1-\alpha, g - 1, N - g) \quad (41)$$

Intervalle de confiance simultané de tous les contrastes avec coefficient de confiance $1 - \alpha$

$$L : \hat{L} \pm S * s(\hat{L}) \quad (42)$$

exemple : données de design d'emballage

Scheffé test; variable Y-caisses vendues
probabilities for Post Hoc Tests Error: Between MS = 10.547, df = 15.000

	design	{1}	{2}	{3}	{4}
1	1		0.9507	0.2125	0.0002
2	2	0.9507		0.0895	0.0001
3	3	0.2125	0.0895		0.0248
4	4	0.0002	0.0001	0.0248	

Méthode de Bonferroni

La famille d'intérêt comprend les comparaisons pairées, les contrastes et toute combinaison linéaire quelconque. L'utilisateur doit spécifier la comparaison avant de faire l'analyse de la variance.

$$L : \hat{L} \pm B * s(\hat{L}) \qquad (43)$$

$$B = t(\upsilon, N - g) \qquad \upsilon = 1 - (\alpha/2g) \qquad (44)$$

exemple : données de design d'emballage

Bonferroni test; variable Y-caisses vendues
probabilities for Post Hoc Tests Error: Between MS = 10.547, df = 15

	design	{1}	{2}	{3}	{4}
1	1		1.0000	0.2397	0.0001
2	2	1.0000		0.0808	0.0000
3	3	0.2397	0.0808		0.0180
4	4	0.0001	0.0000	0.0180	

Comparaison des méthodes

- Si on veut seulement faire des comparaisons entre les paires, la procédure de Tukey est supérieure et elle est recommandée.
- Si le test F rejette l'égalité des moyennes alors il existe au moins un contraste qui diffère de zéro parmi tous les contrastes.
- La procédure de Bonferroni est préférable à la procédure de Scheffé si le nombre de contrastes d'intérêt est à peu près le même que le nombre de modalités.
- Il existe d'autres procédures pour des fonctions spécialisées. Par exemple, la procédure de Dunnett pour comparer chaque traitement vis-à-vis un contrôle ;
- procédure de Hsu : choisir le « meilleur » traitement.

ANOM : Analyis Of Means (Ott)

C'est une méthode alternative au test F. Elle est basée sur l'ensemble des tests de l'effet différentiel de chaque modalité. La méthode a aussi l'avantage d'une représentation graphique semblable à une carte Xbar de Shewhart. On fait plusieurs tests:

$$H_{0\,i} : \tau_i = 0 \qquad \text{vs} \qquad H_a : \tau_i \neq 0 \qquad i = 1, 2, .., g$$

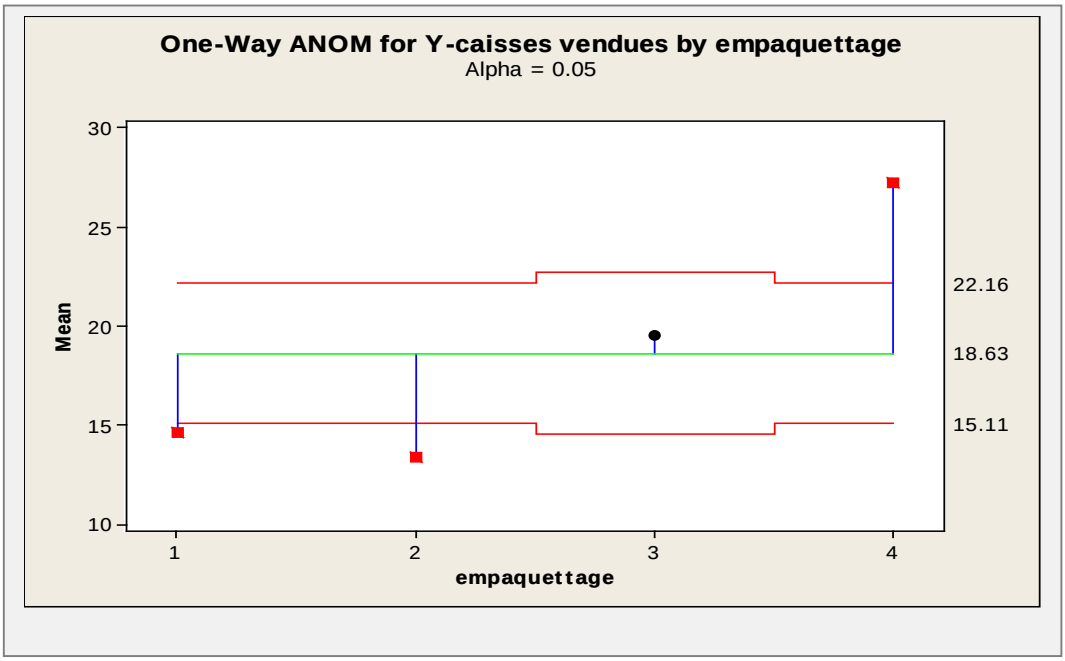
$$\hat{\tau}_i = \overline{Y}_{i.} - \overline{\overline{Y}}_{..} \qquad (45)$$

$$s^2(\hat{\tau}_i) = MSE \left[((g-1)/g)^2 (1/n_i) + (1/g^2) \left(\sum_{h \neq i} (1/n_h) \right) \right] \qquad (46)$$

ANOM : test si le moyennes diffèrent de la moyenne globale

ANOVA : test si les moyennes sont différentes

exemple : données de design d'emballage



Analyse de variance si le facteur est continu

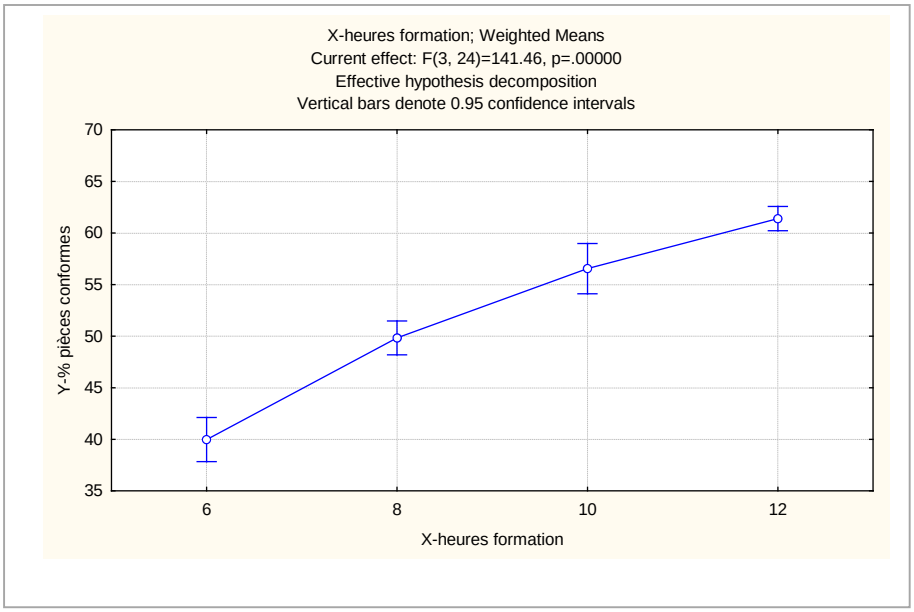
Exemple 6 : pourcentage de pièces produites conformes VS heures de formation

	X-heures formation	Y-% pièces conformes
1	6	40
2	6	39
3	6	39
4	6	36
5	6	42
6	6	43
7	6	41
8	8	53
9	8	48
10	8	49
11	8	50
12	8	51
13	8	50
14	8	48

	X-heures formation	Y-% pièces conformes
15	10	53
16	10	58
17	10	56
18	10	59
19	10	53
20	10	59
21	10	58
22	12	63
23	12	62
24	12	59
25	12	61
26	12	62
27	12	62
28	12	61

On peut faire plus qu'une comparaison de l'influence du facteur : on peut développer une équation de prédiction.

Première analyse : modèle d'analyse de variance



ANOVA

	df	SS	MS	Ratio F	p-value
Intercept	**1**	**75608.04**	**75608.04**	**17740.43**	**0.000000**
X-heures formation	**3**	**1808.68**	**602.89**	**141.46**	**0.000000**
Error	**24**	**102.29**	**4.26**		
Total	**27**	**1910.96**			

Tukey HSD test; variable Y-% pièces conformes
Approximate Probabilities for Post Hoc Tests Error:
Between MS = 4.2619, df = 24.000

	X-heures formation	{1}	{2}	{3}	{4}
1	**6**		**0.0002**	**0.0002**	**0.0002**
2	**8**	**0.0002**		**0.0002**	**0.0002**
3	**10**	**0.0002**	**0.0002**		**0.0012**
4	**12**	**0.0002**	**0.0002**	**0.0012**	

Première analyse : modèle de régression

M1 modèle de prédiction d'ordre 1 : $Y = \beta_0 + \beta_1 * X + \varepsilon$

Regression Summary for Dependent Variable: Y-% pièces conformes
R= 0,961 R^2= 0,923 Adjusted R^2= 0,920 F(1,26)=312,88 p = 0,000

	Beta	Std.Err.	B	Std.Err.	t(26)	p-level
Intercept			20.014	1.8612	10.754	0.0000
X-heures formation	0.9609	0.0543	3.550	0.2007	17.688	0.0000

Analysis of Variance; DV: Y-% pièces conformes

	SS	df	MS	F	p-level
Regress.	1764.35	1	1764.35	312.88	0.0000
Residual	146.61	26	5.64		
Total	1910.96				

M2 modèle de prédiction d'ordre 2 : $Y = \beta_0 + \beta_1 * X + \beta_2 * X^2 + \varepsilon$

Regression Summary for Dependent Variable: Y-% pièces conformes
R= 0,973 R^2= 0,946 Adjusted R^2= .942 F(2,25)=219.72 p = 0, 0000

	Beta	Std.Err.	B	Std.Err.	t(25)	p-level
Intercept			-3.736	7.4549	-0.501	0.6207
X-heures formation	2.483	0.4692	9.175	1.7335	5.293	0.0000
X**2	-1.530	0.4692	-0.312	0.0958	-3.261	0.0032

Analysis of Variance; DV: Y-% pièces conformes

	SS	df	MS	F	p-level
Regress.	1808.10	2	904.050	219.72	0.0000
Residual	102.86	25	4.115		
Total	1910.96				

7. DIAGNOSTICS ET MESURES CORRECTIVES

- **Diagnostic : écarts importants par rapport aux hypothèses de base?**
- **Si oui, qu'elles sont les mesures correctives?**

L'analyse diagnostique est basée sur les graphiques des résidus.
Il y 4 types de résidus :

$$e_{ij} = Y_{ij} - \hat{Y}_{ij} = Y_{ij} - Y_{i.} \qquad \text{résidu brut} \qquad (47)$$

$$e_{ij}^* = e_{ij} / MSE^{0.5} \qquad \text{résidu semi studentisé} \qquad (48)$$

$$r_{ij} = e_{ij}^* / [(n_i - 1) / n_i]^{0.5} \qquad \text{résidu studentisé} \qquad (49)$$

$$t_{ij} = e_{ij}^* [(N - g - 1) / (SSE [1 - (1/n_i)] - e_{ij}^2)]^{0.5} \qquad (50)$$

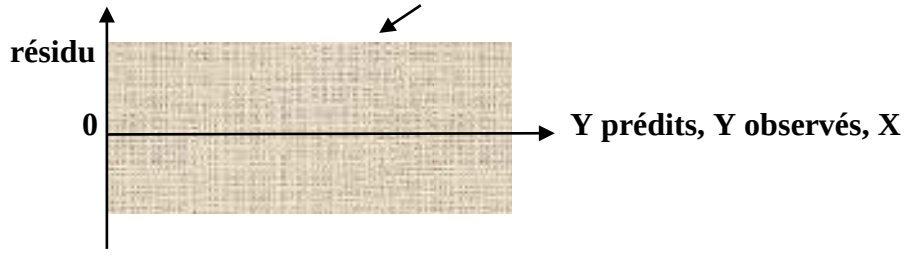
résidu studentisé supprimé
résidu studentisé avec observation supprimée

Les écarts du modèle d'ANOVA en ordre d'importance décroissante sont :

- **variance non constante**
- **erreurs (observations) non indépendantes**
- **présence de valeurs aberrantes**
- **normalité du terme d'erreur**
- **omission de variables explicatives importantes**

HYPOTHÈSES	VÉRIFICATION	DIAGNOSTIC
variance non constante	graphique de résidus studentisés VS valeurs prédites	- bande horizontale - tests : Hartley, Brown-Forsythe
non indépendance	si l'ordre temporel est connu	- résidus VS temps - test d'indépendance sérielle
valeurs aberrantes	t_{ij} VS valeurs prédites	-
normalité	résidus sur échelle de probabilité gaussienne	écart par rapport à la droite-
omisssion	résidus VS valeurs prédites	résisdus corréllés avec autres facteurs non tenu en compte

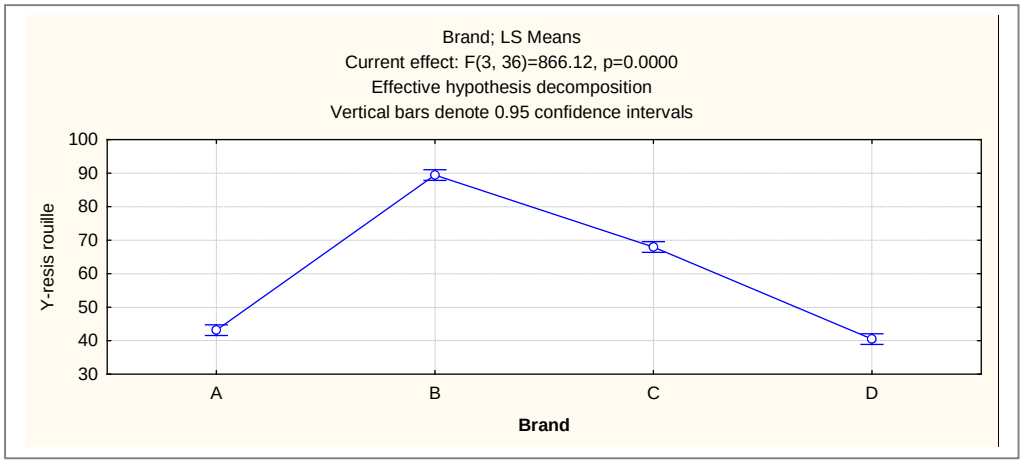
graphique des résidus (axe vertical) VS : $\hat{Y}$ prédits, Y observés, chaque facteur X
allure = bande horizontale



Exemple 7 : inhibiteur de rouille

Kutner et all - 5 ed p. 735			
id	Brand	rep	Y-resis rouille
1	A	1	43.9
2	A	2	39.0
3	A	3	46.7
4	A	4	43.8
5	A	5	44.2
6	A	6	47.7
7	A	7	43.6
8	A	8	38.9
9	A	9	43.6
10	A	10	40.0
11	B	1	89.8
12	B	2	87.1
13	B	3	92.7
14	B	4	90.6
15	B	5	87.7
16	B	6	92.4
17	B	7	86.1
18	B	8	88.1
19	B	9	90.8
20	B	10	89.1

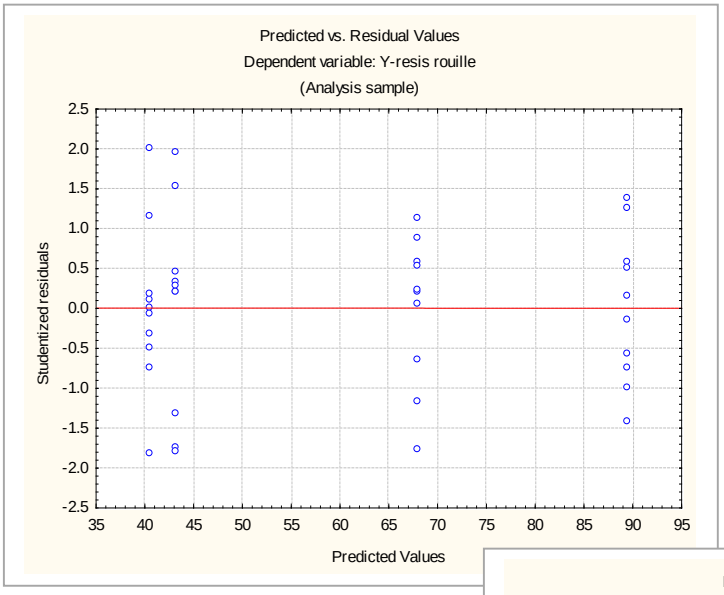
	Brand	rep	Y-resis rouille
21	C	1	68.4
22	C	2	69.3
23	C	3	68.5
24	C	4	66.4
25	C	5	70.0
26	C	6	68.1
27	C	7	70.6
28	C	8	65.2
29	C	9	63.8
30	C	10	69.2
31	D	1	36.2
32	D	2	45.2
33	D	3	40.7
34	D	4	40.5
35	D	5	39.3
36	D	6	40.3
37	D	7	43.2
38	D	8	38.7
39	D	9	40.9
40	D	10	39.7

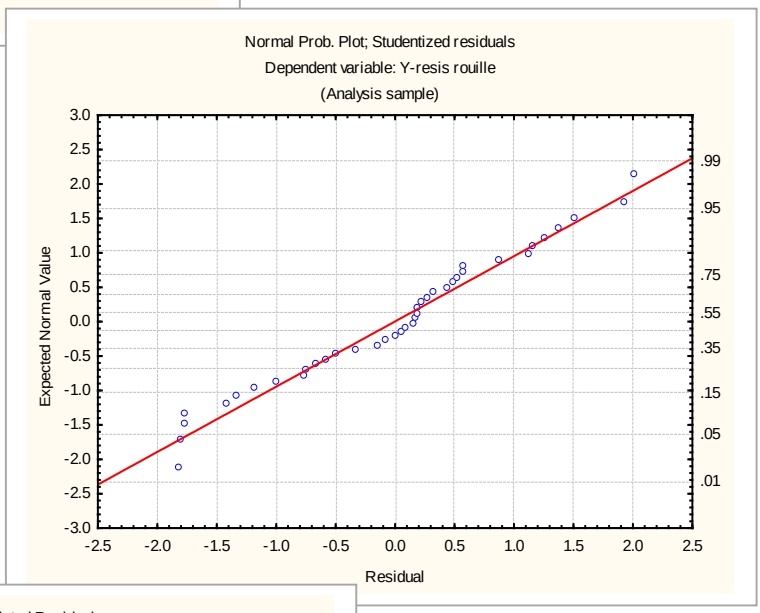


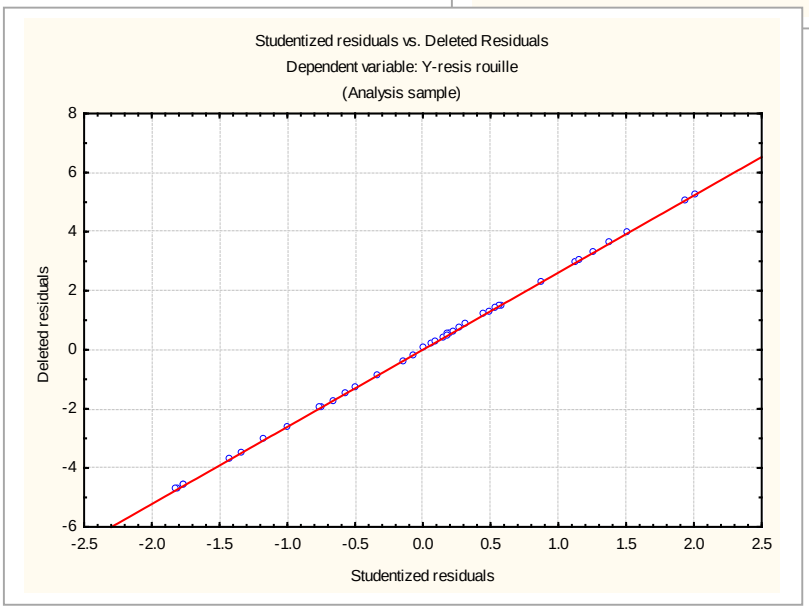
	df	SS	MS	F	p
Intercept	1	145202.5	145202.5	23649.26	0.0000
Brand	3	15953.5	5317.8	866.12	0.0000
Error	36	221.0	6.1		
Total	39	16174.5			

Tableau des résidus

		Y obs	Y préd	rédidu brute	Résidu semi stu e*	résidu studentisér	résidu supprimét
1		43,90	43,14	0,76	0,308	0,323	0,319
2		39,00	43,14	-4,14	-1,676	-1,761	-1,817
3		46,70	43,14	3,56	1,441	1,514	1,543
4		43,80	43,14	0,66	0,267	0,281	0,277
5		44,20	43,14	1,06	0,429	0,451	0,446
6		47,70	43,14	4,56	1,846	1,940	2,021
7		43,60	43,14	0,46	0,186	0,196	0,193
8		38,90	43,14	-4,24	-1,717	-1,804	-1,865
9		43,60	43,14	0,46	0,186	0,196	0,193
10		40,00	43,14	-3,14	-1,271	-1,336	-1,351
11		89,80	89,44	0,36	0,146	0,153	0,151
12		87,10	89,44	-2,34	-0,947	-0,995	-0,995
13		92,70	89,44	3,26	1,320	1,387	1,406
14		90,60	89,44	1,16	0,470	0,493	0,488
15		87,70	89,44	-1,74	-0,704	-0,740	-0,736
16		92,40	89,44	2,96	1,198	1,259	1,270
17		86,10	89,44	-3,34	-1,352	-1,421	-1,442
18		88,10	89,44	-1,34	-0,543	-0,570	-0,565
19		90,80	89,44	1,36	0,551	0,579	0,573
20		89,10	89,44	-0,34	-0,138	-0,145	-0,143
21		68,40	67,95	0,45	0,182	0,191	0,189
22		69,30	67,95	1,35	0,547	0,574	0,569
23		68,50	67,95	0,55	0,223	0,234	0,231
24		66,40	67,95	-1,55	-0,628	-0,659	-0,654
25		70,00	67,95	2,05	0,830	0,872	0,869
26		68,10	67,95	0,15	0,061	0,064	0,063
27		70,60	67,95	2,65	1,073	1,127	1,132
28		65,20	67,95	-2,75	-1,113	-1,170	-1,176
29		63,80	67,95	-4,15	-1,680	-1,765	-1,822
30		69,20	67,95	1,25	0,506	0,532	0,526
31		36,20	40,47	-4,27	-1,729	-1,816	-1,879
32		45,20	40,47	4,73	1,915	2,012	2,106
33		40,70	40,47	0,23	0,093	0,098	0,096
34		40,50	40,47	0,03	0,012	0,013	0,013
35		39,30	40,47	-1,17	-0,474	-0,498	-0,492
36		40,30	40,47	-0,17	-0,069	-0,072	-0,071
37		43,20	40,47	2,73	1,105	1,161	1,167
38		38,70	40,47	-1,77	-0,717	-0,753	-0,748
39		40,90	40,47	0,43	0,174	0,183	0,180
40		39,70	40,47	-0,77	-0,312	-0,328	-0,323
	MEAN case 1-40			0	0	0	-0,00142
	SD case 1-40			2,38	0,964	1,013	1,035
	SUM case 1-40			-0,00	-0,000	-0,000	-0,057
	MIN case 1-40			-4,27	-1,729	-1,816	-1,879
	MAX case 1-40			4,73	1,915	2,012	2,106







Les résidus ont un comportement acceptable.

Tests homogénéité de la variance : Hartley, Bartlett, Cochran, Brown-Forsythe, Levene

Test de Hartley exigence : $n_i = n$ + normalité

H_0 : $\sigma_1^2 = \sigma_2^2 = \ldots = \sigma_g^2$ (51)

H_a : les variances ne sont pas toutes égales (52)

$$\text{Hartley : } H^* = \max(s_i^2) \,/\, \min(s_i^2) \quad (53)$$

$$\text{Rejet de } H_0 \text{ si } H > H(1-\alpha, g, n-1) \quad (54)$$

$H(1-\alpha, g, df)$: $(1-\alpha)$ percentile distribution de H

Exemple 7 : inibiteur de rouille

marque	# obs	moyenne	écart type	variance
tous	40	60.25	20.36	414.53
A	10	43.14	3.00	9.00
B	10	89.44	2.22	4.93
C	10	67.95	2.17	4.71
D	10	40.47	2.44	5.95

$H^* = 9.00 / 4.17 = 1.91$: on ne rejette pas H_0 car $p = 0.7532$

Tests of Homogeneity of Variances					
	Hartley	Cochran	Bartlett	df	p
Y-resis rouille	1.91	0.366	1.199	3	0.7532

Test de Brown-Forsythe n_i peuvent être inégaux et le test est robuste à la non normalité

On analyse les écarts d_{ij} des observations y_{ij} par rapport à leur médiane

$$d_{ij} = |y_{ij} - \text{med}(y_{ij})| \qquad \text{med(y) = médiane(y)} \quad (55)$$

$$F_{BF} = MS_{TR} / MSE \quad (56)$$

$$MS_{TR} = \sum n_i * (\overline{d_{i.}} - \overline{d_{..}})^2 / (g-1) \quad (57)$$

$$MSE = \sum\sum (d_{ij} - \overline{d_{i.}})^2 / (N-g) \quad (58)$$

$$\overline{d_{i.}} = \sum d_{ij} / n_i \qquad \overline{d_{..}} = \sum\sum d_{ij} / N \quad (59)$$

F_{BF} suit approximativement loi $F(g-1, N-g)$

On rejette H_0 si $F_{BF} > F(1-\alpha, g-1, N-g)$

Test de Levene $d_{ij} = |y_{ij} - moy(y_{ij})|$ **moy(y) = moyenne(y)**

Le test de Brown-Forsythe constitue une modification du test de Levene

Test de Cochran $n_i = n$ **(égaux)**

$$C = \max(s_i^2) / \sum s_i^2 \qquad (60)$$

loi d'échantillonange de C dépend de g et de n

on rejette H_0 si $C > C(1 - \alpha ; n, g)$

tableau des percentiles de la distribution de C :$C(1 - \alpha ; n, g)$

n	percentile 1 - α	g = 2	g = 3	g = 4	g = 5	g = 8	g = 10
2	0.95	0.9985	0.9669	0.9065	0.8412	0.6798	0.6020
	0.99	0.9999	0.9933	0.9676	0.9279	0.7945	0.7175
3	0.95	0.9750	0.8709	0.7679	0.6838	0.5157	0.4450
	0.99	0.9950	0.9423	0.8643	0.7885	0.6152	0.5358
4	0.95	0.9392	0.7977	0.6841	0.5981	0.4377	0.3733
	0.99	0.9794	0.8831	0.7814	0.6957	0.5209	0.4469
5	0.95	0.9057	0.7457	0.6287	0.5441	0.3910	0.3311
	0.99	0.9586	0.8335	0.7212	0.6329	0.4627	0.3934
6	0.95	0.8772	0.7071	0.5895	0.5065	0.3595	0.3029
	0.99	0.9373	0.7933	0.6761	0.5875	0.4226	0.3572
8	0.95	0.8332	0.6530	0.5365	0.4564	0.3185	0.2666
	0.99	0.8988	0.7335	0.6129	0.5229	0.3704	0.3106
10	0.95	0.8010	0.6167	0.5017	0.4387	0.2926	0.2439
	0.99	0.8674	0.6912	0.5702	0.5037	0.3373	0.2813
17	0.95	0.7341	0.5466	0.4366	0.3645	0.2462	0.2032
	0.99	0.7949	0.6059	0.4884	0.4094	0.2779	0.2297
37	0.95	0.6602	0.4748	0.3720	0.3066	0.2022	0.1655
	0.99	0.7067	0.5153	0.4057	0.3351	0.2214	0.1811

(Statistical Principles in Experimental Design, 2 ed., B.J. Winer, 1971, Mc Graww-Hill, p. 876)

Test de Bartlett **les n_i peuvent être inégaux**

$$c = 1 + (1/3*(g-1))*[\sum (1/(n_i-1) - (1/N)] \qquad (61)$$

$$B = (2.303/c)*[(N - g)*\log(MSE) - \sum (n_i - 1)*\log(s_i^2)] \qquad (62)$$

B suit approximativement loi khi-deux avec (g – 1) degrés de liberté

On rejette H_0 si $B > \chi^2 (1 - \alpha ; g - 1)$

Exemple 8 : données de flux – soudure de composants électroniques

Kutner et all - 5 ed. p. 783			
id	type flux	rep	Y-force soudure
1	A	1	14.87
2	A	2	16.81
3	A	3	15.83
4	A	4	15.47
5	A	5	13.60
6	A	6	14.76
7	A	7	17.40
8	A	8	14.62
9	B	1	18.43
10	B	2	18.76
11	B	3	20.12
12	B	4	19.11
13	B	5	19.81
14	B	6	18.43
15	B	7	17.16
16	B	8	16.40
17	C	1	16.95
18	C	2	12.28
19	C	3	12.00
20	C	4	13.18

id	Type flux	rep	Y-force soudure
21	C	5	14.99
22	C	6	15.76
23	C	7	19.35
24	C	8	15.52
25	D	1	8.59
26	D	2	10.90
27	D	3	8.60
28	D	4	10.13
29	D	5	10.28
30	D	6	9.98
31	D	7	9.41
32	D	8	10.04
33	E	1	11.55
34	E	2	13.36
35	E	3	13.64
36	E	4	12.16
37	E	5	11.62
38	E	6	12.39
39	E	7	12.05
40	E	8	11.95

flux	n	moyenne	variance
tous	40	14.21	10.96
A	8	15.42	1.531
B	8	18.53	1.570
C	8	15.00	6.185
D	8	9.74	0.667
E	8	12.34	0.592

Hartley	Cochran	Bartlett	df	p
10.445	0.5865	12.985	4	0.0113

Test	SS Effect	df Effect	MS Effect	SS Error	Df Error	MS Error	F	p
Levene	8,69	4	2,17	24,8	35	0,71	3,07	0,029
Brown-Forsythe	9,35	4	2,34	27,9	35	0,80	2,94	0,034

Tous les tests concordent : les variances sont inégales.

Mesures correctives

VARIANCES	NORMALITÉ	MESURE CORRECTIVE
hétérogènes	oui	régresssion pondérée
hétérogènes	non	transformation de Box-Cox
« gros » écarts	« gros » écarts	ANOVA non paramétrique Kruskall-Wallis

Exemple 8 : suite données de flux

$$Y_{ij} = \mu_i + \varepsilon_{ij} \qquad \varepsilon_{ij} \sim N(0, \sigma_i^2) \qquad (63)$$

Modèle à cellules

poids w $w_{ij} = 1 / s_i^2$ (64)

on remplace l'ANOVA par un modèle de régresssion avec des variables indicatrices et on fait l'ajustement de moindres carrés pondérés avec les poids w

modèle complet (F) : $Y_{ij} = \mu_1 X_{ij1} + \mu_2 X_{ij2} + ... + \mu_g X_{ijg} + \varepsilon$ (65)

$$X_{ijt} = \begin{cases} 1 \text{ si le cas provient du niveau i du facteur} \\ 0 \text{ autrement} \end{cases}$$

modèle réduit (R) : $Y_{ij} = \mu X_{ij1} + \mu X_{ij2} + \mu X_{ijg}$ (66)

sous $H_0 : \mu_1 = \mu_2 = ... = \mu_g = \mu$

SSE w (R) : somme de carrés erreur modèle réduit R (67)

SSE w (F) : somme de carrés erreur modèle complet F (68)

$$Fw = [(SSEw(R) - SSEw(F)) / SSEw(F)] * (N - g) / (g - 1) \qquad (69)$$

suit approximativement loi $F(g - 1, N - g)$

on rejette H_0 **si** $Fw > F(1 - \alpha, g - 1, N - g)$

Exemple : suite

groupe	i = 1	i = 2	i = 3	i = 4	i = 5
w_{ij}	0.653	0.637	0.162	1.449	1.689
s_i^2	1,531	1,570	6,185	0,667	0,592

Données Flux 2

id	flux	rep	indA	indB	indC	indD	indE	Gen	poids	Y-soudure
1	A	1	1	0	0	0	0	1	0.653	14.87
2	A	2	1	0	0	0	0	1	0.653	16.81
3	A	3	1	0	0	0	0	1	0.653	15.83
4	A	4	1	0	0	0	0	1	0.653	15.47
5	A	5	1	0	0	0	0	1	0.653	13.60
6	A	6	1	0	0	0	0	1	0.653	14.76
7	A	7	1	0	0	0	0	1	0.653	17.40
8	A	8	1	0	0	0	0	1	0.653	14.62
9	B	1	0	1	0	0	0	1	0.637	18.43
10	B	2	0	1	0	0	0	1	0.637	18.76
11	B	3	0	1	0	0	0	1	0.637	20.12
12	B	4	0	1	0	0	0	1	0.637	19.11
13	B	5	0	1	0	0	0	1	0.637	19.81
14	B	6	0	1	0	0	0	1	0.637	18.43
15	B	7	0	1	0	0	0	1	0.637	17.16
16	B	8	0	1	0	0	0	1	0.637	16.40
17	C	1	0	0	1	0	0	1	0.162	16.95
18	C	2	0	0	1	0	0	1	0.162	12.28
19	C	3	0	0	1	0	0	1	0.162	12.00
20	C	4	0	0	1	0	0	1	0.162	13.18
21	C	5	0	0	1	0	0	1	0.162	14.99
22	C	6	0	0	1	0	0	1	0.162	15.76
23	C	7	0	0	1	0	0	1	0.162	19.35
24	C	8	0	0	1	0	0	1	0.162	15.52
25	D	1	0	0	0	1	0	1	1.499	8.59
26	D	2	0	0	0	1	0	1	1.499	10.90
27	D	3	0	0	0	1	0	1	1.499	8.60
28	D	4	0	0	0	1	0	1	1.499	10.13
29	D	5	0	0	0	1	0	1	1.499	10.28
30	D	6	0	0	0	1	0	1	1.499	9.98
31	D	7	0	0	0	1	0	1	1.499	9.41
32	D	8	0	0	0	1	0	1	1.499	10.04
33	E	1	0	0	0	0	1	1	1.689	11.55
34	E	2	0	0	0	0	1	1	1.689	13.36
35	E	3	0	0	0	0	1	1	1.689	13.64
36	E	4	0	0	0	0	1	1	1.689	12.16
37	E	5	0	0	0	0	1	1	1.689	11.62
38	E	6	0	0	0	0	1	1	1.689	12.39
39	E	7	0	0	0	0	1	1	1.689	12.05
40	E	8	0	0	0	0	1	1	1.689	11.95

modèle complet (F) : $\hat{Y} = 15.4 * indA + 18.5 * indB + 15.0 * indC + 9.7 * indD + 12.3 * indE$

$SSEw(F) = 35.0$ **avec 35 degrés de liberté**

modèle réduit (R) : $\hat{Y} = 12.88 * gen$ $SSw\ (R) = 359.2$ **avec 39 degrés de liberté**

$Fw = (359.2 - 35.0) / 35 * (35 / (39 - 35)) = 81.05$

$Fw > F\ (0.99, 4, 35) = 3.91$ **rejet de** H_0

Transformations de la variable de réponse : cas de variances inégales

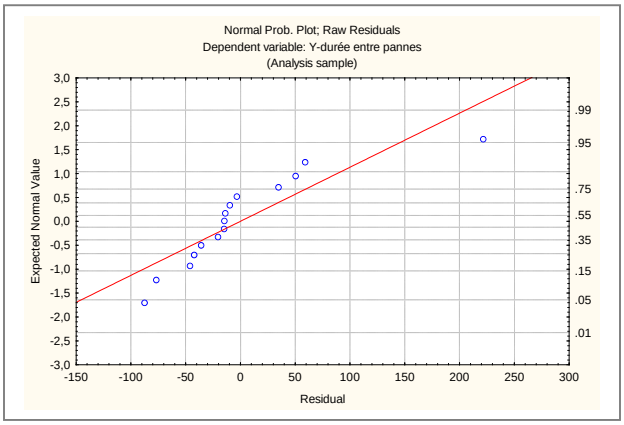
CONDITION	TRANSFORMATION
réponse Y est un comptage : distribution Poisson	$Y' = \sqrt{Y}$ ou $Y' = \sqrt{Y} + \sqrt{Y + 1}$
réponse Y est une proportion : distribution binomiale	$Y' = 2 \arcsin(\sqrt{Y})$
$(\text{ecart type})^2$ proportionnel à la moyenne	$Y' = \sqrt{Y}$
écart type proportionnel à la moyenne	$Y' = \log(Y)$
écart type proportionnel à la $(\text{moyenne})^2$	$Y' = 1 / Y$

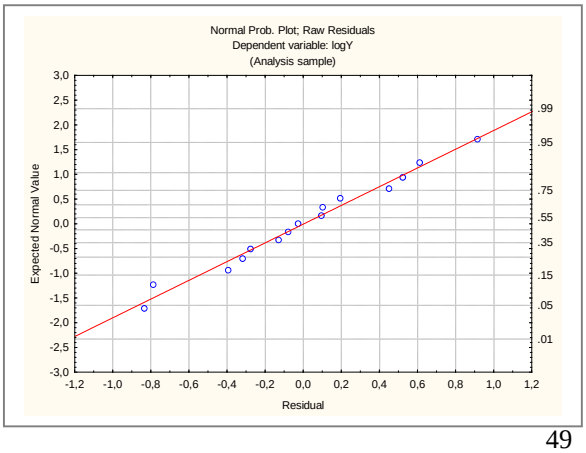
Recommandation: examiner les quantités $s_i^2 / \bar{Y}_{i.}$, $s_i / \bar{Y}_{i.}$, $s_i / \bar{Y}_{i.}^2$ pour chaque niveau du facteur et choisir la transformation dont le coefficient de variation (CV) est le plus petit

Exemple 9 : temps entre les pannes d'ordinateur dans trois villes

ville	obs	m moyenne	s écart type	s^2 / m	s / m	s / moy^2
A	5	50.37	42.29	35.51	0.84	0.017
B	5	22.13	33.22	49.86	1.50	0.068
C	5	121.21	127.15	133.39	1.05	0.009
			moyenne	72,92	1,13	0,031
			std dev	52,854	0,338	0,032
			CV(%)	72	30	103

choix : transformation logarithmique





Transformation de Box-Cox

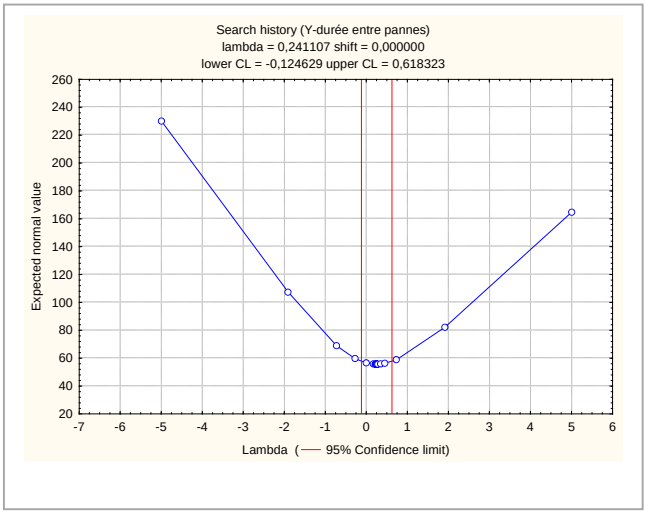
$$Y' = Y^{\lambda} \qquad -2 < \lambda < 2 \qquad (70)$$

$\lambda = ?$ on choisit λ tel que SSE(λ) soit minimum

on prend un valeur arondie

si $\lambda = 0$ $Y' = \log(Y)$

Exemple 9 : suite



$\lambda = 0{,}24$

$-0{,}125 < \lambda < 0{,}618$

On peut prendre $\lambda = 0$

donc une $Y' = \log(Y)$

transformation logarithmique

Importance des écarts d'hypothèses de base sur le modèle d'ANOVA

1. Le **manque de normalité n'est pas très important pour le cas de modèles à effets fixes**. Dans un récent article, (*Six Sigma Forum magazine*, vol 4, no 3, May 2005), John Sall et Bradley Jones traitent de la **peur irrationnelle de la non normalité** (« Leptokurtosiphobia »). Ils concluent que tester la normalité des résidus est une étape non nécessaire car
   (a) pour de « grands échantillons » la non normalité est facile à détecter mais elle est sans conséquence,
   (b) pour de « petits échantillons », la non normalité pourrait avoir des conséquences, mais la non normalité est quasiment impossible à détecter : aucun test est suffisamment puissant.
   Pour le cas de modèles à effets aléatoires, les conséquences sont plus importantes.

2. Le **test F est robuste** si les tailles n_i ne sont pas trop inégales.

3. **Indépendance** : conséquences importantes pour l'inférence. Par exemple, une forte auto corrélation dans les valeurs de la réponse Y a comme conséquence pratique que les tailles sont plus faibles en réalité qu'elles le paraissent, rendant ainsi plus difficile la détection des différences signiticatives. Les mesures répétées sur une même unité d'observation constituent un cas fréquent de dépendance. Il est important de savoir reconnaître cette situation lorsqu'elle est présente dans la structure des données et de faire une analyse appropriée. Cette méthode sera vue plus loin.

ANOVA non paramétrique : test de Kruskall-Wallis

Si on ne peut pas transformer la variable de réponse afin d'avoir une distribution qui s'approche de la normale, on peut faire un test non paramétrique qui ne dépend pas de la forme de la distribution. Les méthodes paramétriques sont basées sur les rangs de la variable de réponse plutôt que les valeurs observées. Dans le cas de l'ANOVA, la procédure porte le nom de Kruskall-Wallis.

Procédure

On assigne aux observations Y_{ij} le rang R_{ij} des valeurs ordonnées en ordre croissant de 1 à N. On procède comme dans le test F usuel que l'on applique aux rangs R_{ij}.

Test est est basé sur $F_{KW} = MSTR / MSE$ **(71**

$$MSTR = \sum n_i (\overline{R}_{i.} - \overline{R}_{..})^2 / (g - 1) \quad (72)$$

$$MSE = \sum\sum (R_{ij} - \overline{R}_{i.})^2 / (N - g) \quad (73)$$

où

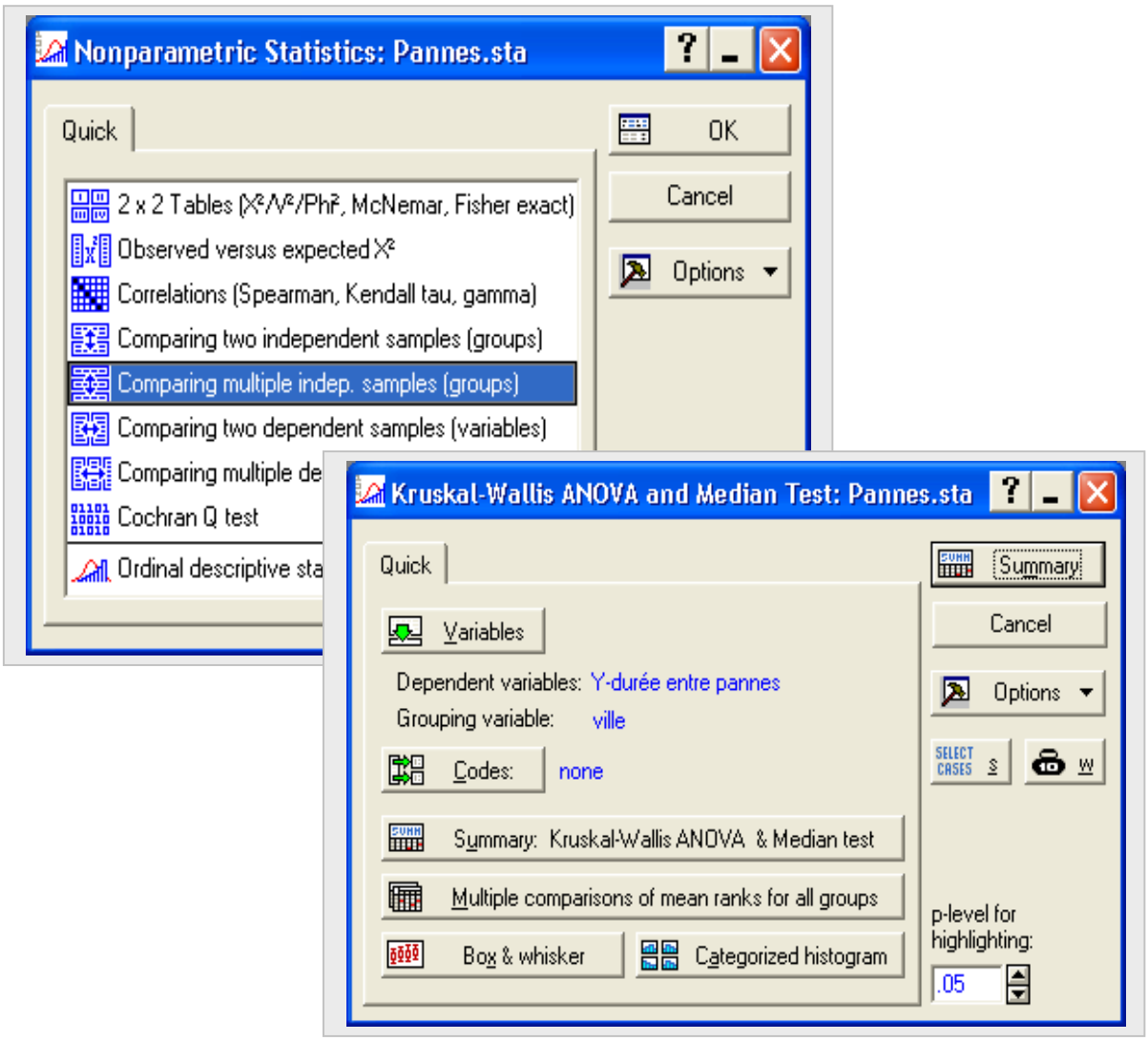
$$\overline{R}_{i.} = \sum R_{ij} / n_i \quad (74)$$

$$\overline{\overline{R}}_{..} = \sum\sum R_{ij} / N = (N + 1) / 2 \quad (75)$$

Exemple 9 : suite durées des pannes

	ville	échantillon	Y-durée entre pannes	Y-Rang	logY
1	A	1	4.41	2	0.644
2	A	2	100.65	13	2.003
3	A	3	14.45	6	1.160
4	A	4	47.13	9	1.673
5	A	5	85.21	12	1.930
6	B	1	8.24	4	0.916
7	B	2	81.16	11	1.909
8	B	3	7.35	3	0.866
9	B	4	12.29	5	1.090
10	B	5	1.61	1	0.207
11	C	1	106.19	14	2.026
12	C	2	33.83	7	1.529
13	C	3	78.88	10	1.897
14	C	4	342.81	15	2.535
15	C	5	44.33	8	1.647

Analyse non paramétrique avec *STATISTICA*



Test F basé sur les rangs Y-rang					
	SS	**DF**	**MS**	**F**	**p**
Intercept	**960.00**	**1**	**960.00**	**61.02**	**0.0000**
ville	**91.20**	**2**	**45.60**	**2.90**	**0.0940**
Error	**188.80**	**12**	**15.73**		

H_0 pas rejetée au seuil de 0.05 confirmation par le test de Kruskal-Wallis

Kruskal-Wallis ANOVA - Y-durée entre panes Kruskal-Wallis test: H (2, N= 15) =4.56 p =.1023				
	Ville	**Code**	**Valid - N**	**Sum of - Ranks**
	A	**1**	**5**	**42**
	B	**2**	**5**	**24**
	C	**3**	**5**	**54**

8. ANOVA avec 2 facteurs à effets fixes

Eléments de définition de la problématique et de l'analyse:

- **tailles échantillonnales dans les cellules n_{ij} : égales, inégales, = 1 , = 0 (certaines cellules);**
- **facteurs : qualitatifs, quantitatifs, fixes, aléatoires, type bloc (secondaire) croisés, emboités (hiérarchie);**
- **design / modèles : complètement aléatoire (CA), à bloc complet (RCBD), à effets fixes, à effets aléatoires, mixtes, analyse de covariance ANCOVA,**

EXEMPLE	FACTEURS	DOMAINE	RÉPONSE	COMMENTAIRE
publicité	A : média B : prix	A : radio, journal B : 55 -60 - 65	Y –volume ventes	CA facteurs fixes
médecine	A : traitement B : genre	A : nouveau, placébo B : homme, femme	Y- tension artérielle	CA A : facteur fixe B : facteur aléatoire
bois traité	A : solution B : concentration	A : CCA, organique B : 0.25- 0.80-2.50	Y- test à la rupture	CA facteurs fixes
vente produit alimentaire	A : hauteur tablette B : largeur tablette	A : bas-millieu-haut B : régulière- large	Y-volume ventes	CA facteurs fixes
méthode d'enseignement	A : méthode B =X= moy cumul	A : M – D X : varie entre 2 et 4 (mesuré)	Y – taux de réussite cours	observationnelle analyse de covariance : ANCOVA

CAS : 2 facteurs A, B fixes A : niveaux i = 1, 2,.., a B : niveaux j = 1, 2, …, b

Y_{ijk} k-ième répétition réponse Y de chaque combinaison (cellule) (i, j)
k = 1, 2,…, $n_{ij} = n > 1$ N = n a b : nombre total d'observations
assignation au hasard (sans restriction) des traitements aux unités d'observation (expérimentales)

Autre cas : $n_{ij} = n = 1$ et $n_{ij} > 0$ et inégales

Modèle de type MOYENNES DES CELLULES (pas d'effet général)

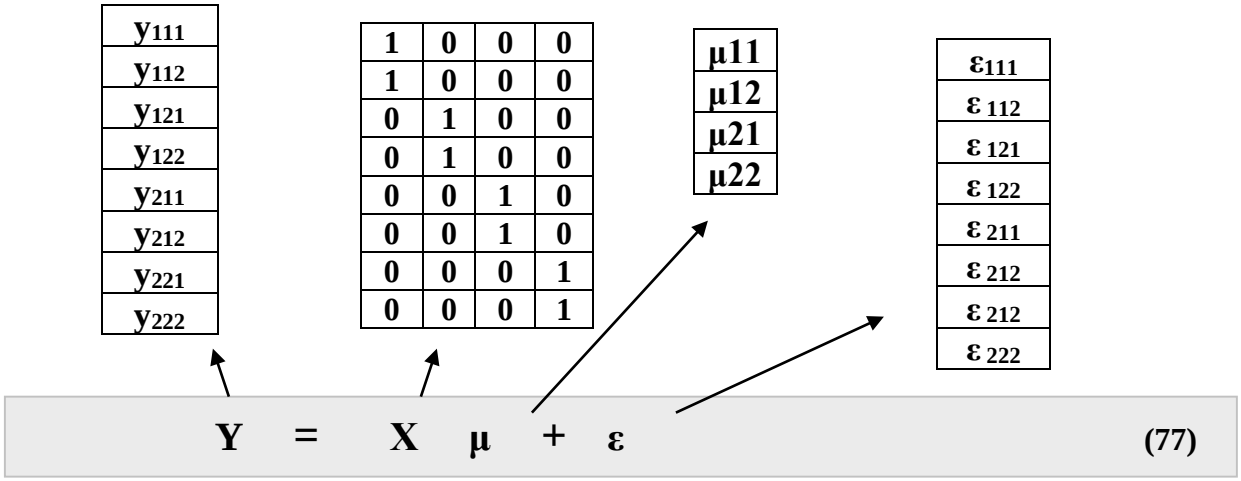
$$Y_{ijk} = \mu_{ij} + \varepsilon_{ijk} \qquad (76)$$

μ_{ij} : paramètres $\varepsilon_{ijk} \sim N(0, \sigma^2)$

i = 1, 2,.., a j = 1, 2,..., b k = 1, 2, ... , $n_{ij} = n$

$Y_{ijk} \sim N(\mu_{ij}, \sigma^2)$

Écriture matricielle : **exemple avec a = b = n = 2**



(77)

Modèle de type EFFETS

$$Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk} \quad (78)$$

$$\mu = \sum\sum \mu_{ij} / ab \quad (79)$$

$$\mu_{i.} = \sum \mu_{ij} / b \qquad \mu_{.j} = \sum \mu_{ij} / a \quad (80)$$

$$\alpha_i = \mu_{i.} - \mu \qquad \sum_i \alpha_i = 0 \quad (81)$$

$$\beta_j = \mu_{.j} - \mu \qquad \sum_j \beta_j = 0 \quad (82)$$

$$(\alpha\beta)_{ij} = \mu_{ij} - \mu_{i.} - \mu_{.j} + \mu = \mu_{ij} - (\mu + \alpha_i + \beta_j) \quad (83)$$

$$\sum_j (\alpha\beta)_{ij} = 0 \qquad \sum_i (\alpha\beta)_{ij} = 0 \quad (84)$$

$$Y_{ijk} \sim N(\mu_{..} + \alpha_i + \beta_j + (\alpha\beta)_{ij}, \sigma^2)$$

<u>Ajustement du modèle</u> : modèle à moyennes de cellules

minimum $Q = \sum\sum (Y_{ijk} - \mu_{ij})^2$ **principes des moindres carrés**

solution: $\hat{\mu}_{ij} = \overline{Y}_{ij.}$ **et** $\hat{Y}_{ijk} = \overline{Y}_{ij.}$ (85)

résidu : $e_{ijk} = Y_{ijk} - \hat{Y}_{ij.} = Y_{ijk} - \overline{Y}_{ij.}$ (86)

<u>Ajustement du modèle</u> : modèle à effets

minimum $Q = \sum\sum (Y_{ijk} - \mu - \alpha_i - \beta_j - (\alpha\beta)_{ij})^2$

$$\sum_i \alpha_i = 0 \qquad \sum_j \beta_j = 0 \qquad \sum_i (\alpha\beta)_{ij} = 0 \qquad \sum_j (\alpha\beta)_{ij} = 0$$

paramètre	estimateur
μ	$\hat{\mu} = \overline{Y}_{...}$ (87)
$\alpha_i = \mu_{i.} - \mu$	$\hat{\alpha}_i = \overline{Y}_{i..} - \overline{Y}_{...}$ (88)
$\beta_j = \mu_{.j} - \mu$	$\hat{\beta}_j = \overline{Y}_{.j.} - \overline{Y}_{...}$ (89)
$(\alpha\beta)_{ij} = \mu_{ij} - \mu_{i.} - \mu_{.j} + \mu$	$\widehat{(\alpha\beta)}_{ij} = \overline{Y}_{ij.} - \overline{Y}_{i..} - \overline{Y}_{.j.} + \overline{Y}_{...}$ (90)

ANOVA : analyse de la variance 2 facteurs avec n répétitions

Source	SS	df	MS	F	p-value
A	SSA	a - 1	MSA = SSA / (a-1)	MSA / MSE	
B	SSB	b - 1	MSB = SSB / (b – 1)	MSB / MSE	
AB	SSAB	(a – 1)(b – 1)	MSAB = SSAB / (a-1)(b-1)	MSAB / MSE	
erreur	SSE	ab(n-1)	MSE = SSE / ab(n-1)	----	
totale	SStotale	abn - 1	----	-----	

$$SSA = nb \sum (\overline{Y}_{i..} - \overline{Y}_{...})^2 \qquad SSB = na \sum (\overline{Y}_{.j.} - \overline{Y}_{...})^2$$

$$SSAB = n \sum\sum (\overline{Y}_{ij.} - \overline{Y}_{i..} - \overline{Y}_{.j.} + \overline{Y}_{...})^2$$

$$SSE = \sum\sum\sum (Y_{ijk} - \overline{Y}_{ij.})^2$$

$$SStotale = \sum\sum\sum (Y_{ijk} - \overline{Y}_{...})^2$$

test de H_0 : $(\alpha\beta)_{ij} = 0$ pour tout (i, j) vs H_a : certains $(\alpha\beta)_{ij}$ ne sont pas zéro

on rejette H_0 si $F = MSAB / MSE \geq F(1 - \alpha ; (a-1)*(b-1) , (n-1)*ab)$ (91)

test de $H_0 : \alpha_1 = \alpha_1 = \ldots = \alpha_a = 0$ vs H_a : certains α_i ne sont pas zéro.

on rejette H_0 si $F = MSA / MSE \geq F(1 - \alpha ; a-1 , (n-1)*ab)$ (92)

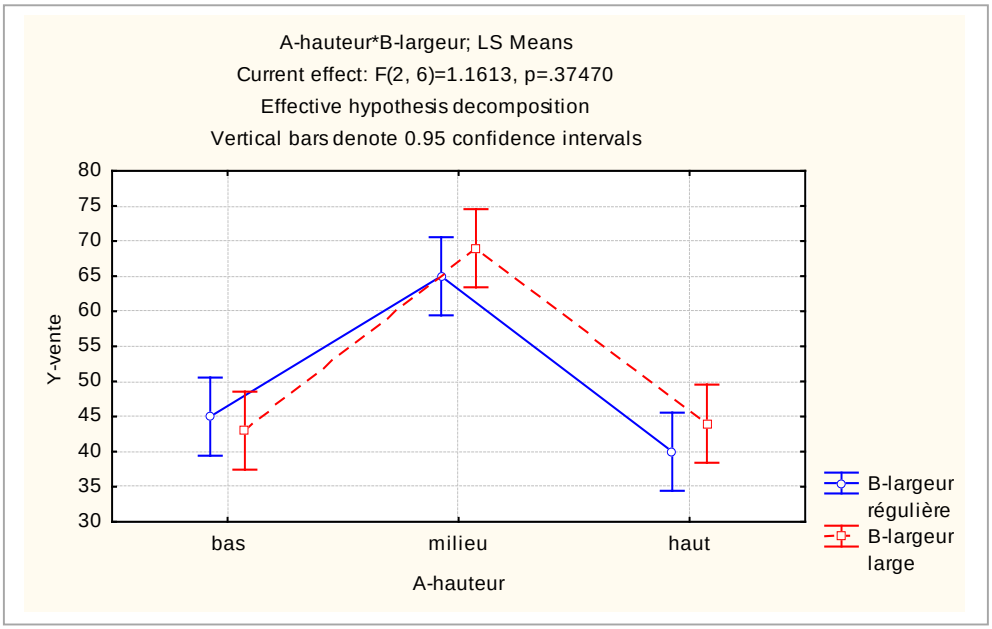
test de $H_0 : \beta_1 = \beta_1 = \ldots = \beta_b = 0$ vs H_a : certains β_j ne sont zéro.

on rejette H_0 si $F = MSA / MSE \geq F(1 - \alpha ; b-1 , (n-1)*ab)$ (93)

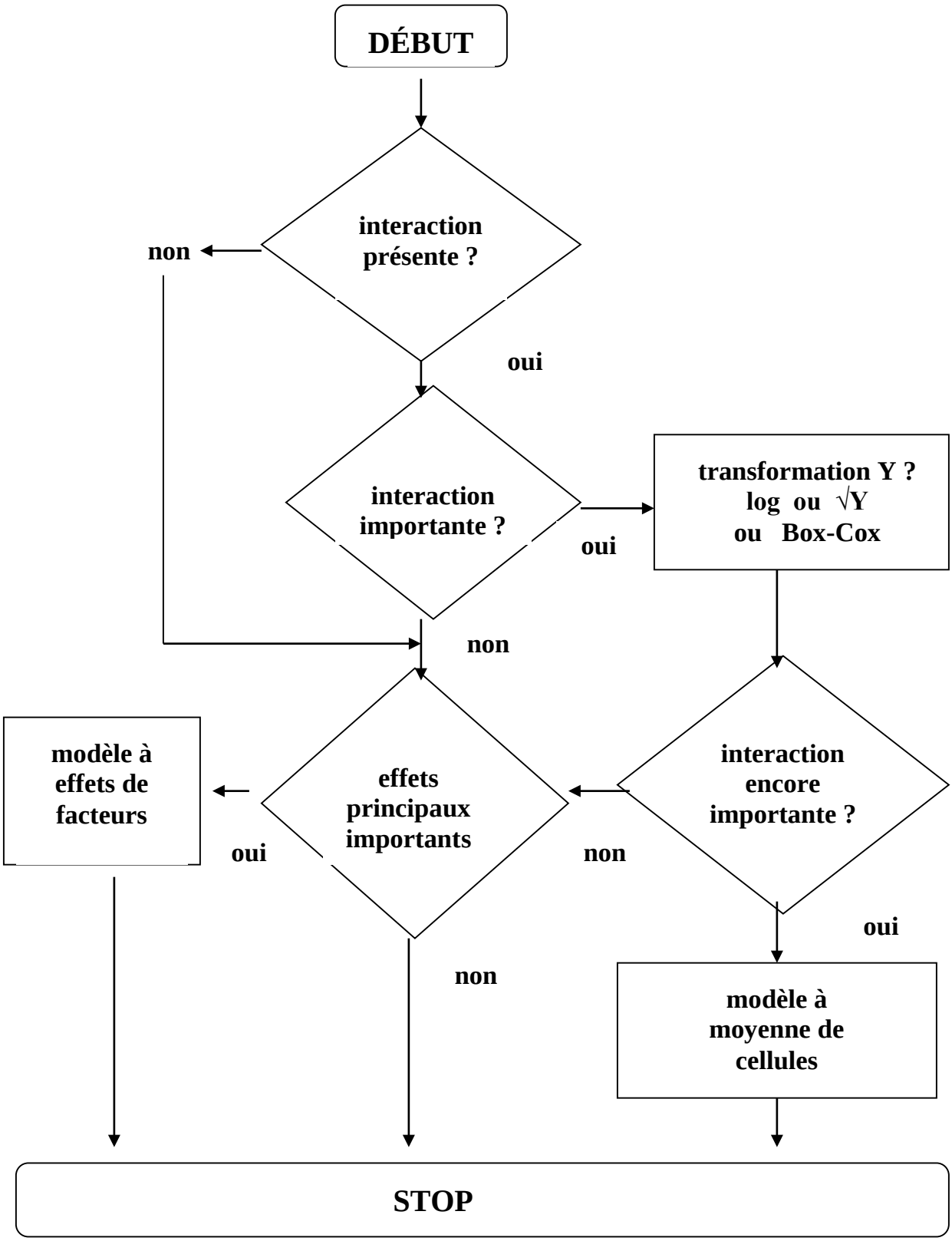
Exemple 10 : étalage

Kutner 5 ed. p. 833				
	A-hauteur	B-largeur	rep	Y-vente
1	bas	régulière	1	47
2	bas	régulière	2	43
3	bas	large	1	46
4	bas	large	2	40
5	milieu	régulière	1	62
6	milieu	régulière	2	68
7	milieu	large	1	67
8	milieu	large	2	71
9	haut	régulière	1	41
10	haut	régulière	2	39
11	haut	large	1	42
12	haut	large	2	46

ANOVA					
	DF	SS	MS	F	p
Intercept	1	31212	31212.0	3020.52	0.00000
A-hauteur	2	1544	772.0	74.71	0.00006
B-largeur	1	12	12.0	1.16	0.32261
A-hauteur*B-largeur	2	24	12.0	1.16	0.37470
Error	6	62	10.3		
Total	11	1642			



Stratégies pour l'étude de 2 facteurs (Kutner & all p. 848)



INTERACTION : présente (significative)? importante (« grande »)? transformation?

<u>exemple</u> 11: Y : durée (minute) apprentissage tâche
facteur : genre (homme, femme) - age (jeune, moyen, agé)
3 versions de la réponse : Y1 aucune interaction
Y2 interaction importante (forte)
Y3 interaction faible

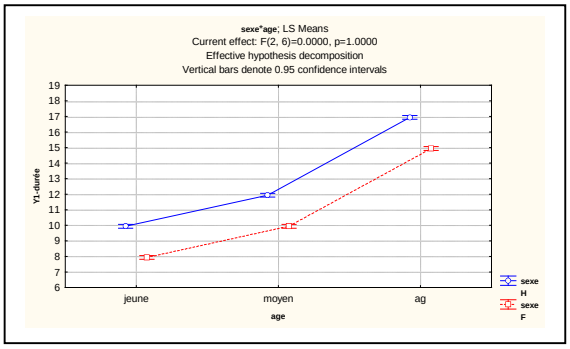
données

exemples :	Y1 aucune interaction				Y2 interaction importante					Y3 interaction faible
	sexe	age	Y1 durée	Y2 durée	Y3 durée	sexe	age	Y1 moy	Y 2 moy	Y3 moy
1	H	jeune	9.95	8.95	9.75	H	jeune	10	9	9.8
2	H	moyen	11.95	11.95	11.95	H	moyen	12	12	12.0
3	H	agé	16.95	17.95	17.15	H	agé	17	18	17.2
4	F	jeune	7.95	8.95	8.15	F	jeune	8	9	7.8
5	F	moyen	9.95	9.95	9.95	F	moyen	10	10	10.0
6	F	agé	14.95	13.95	14.75	F	agé	15	14	15.2
7	H	jeune	10.05	9.05	9.85					
8	H	moyen	12.05	12.05	12.05	H		13	13	13
9	H	agé	17.05	18.05	17.25	F		11	11	11
10	F	jeune	8.05	9.05	8.25		jeune	9	9	9
11	F	moyen	10.05	10.05	10.05		moyen	11	11	11
12	F	agé	15.05	14.05	14.85		agé	16	16	16
						gen		12	12	12

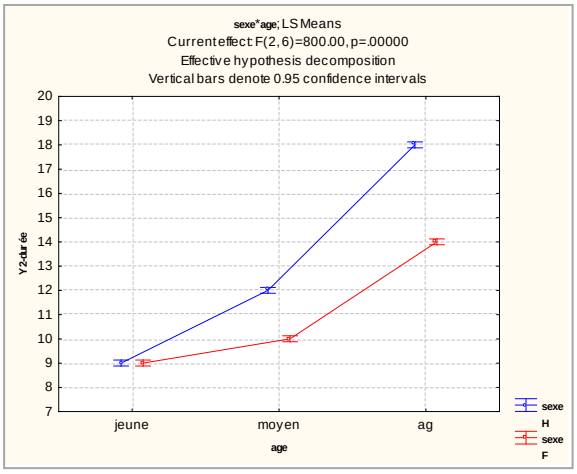
$$\text{inter } Y = \overline{Y}_{ij.} - \overline{Y}_{i..} - \overline{Y}_{.j.} + \overline{Y}_{...}$$

	sexe	age	Y1 moy	Y2 moy	Y3 moy		inter Y1	inter Y2	inter Y3
1	H	jeune	10	9	9.8		0	-1	-0.2
2	H	moyen	12	12	12.0		0	0	0
3	H	agé	17	18	17.2		0	1	0.2
4	F	jeune	8	9	8.2		0	1	-0.2
5	F	moyen	10	10	10.0		0	0	0
6	F	agé	15	14	14.8		0	-1	0.2

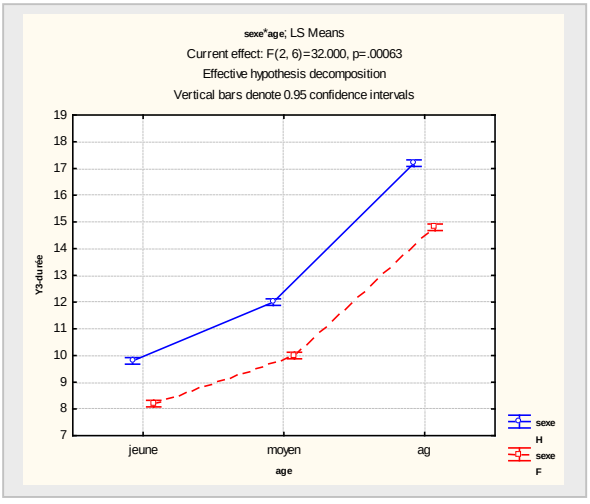
ANOVA : Y1					
Source	DF	SS	MS	F	p-value
Intercept	1	1728.00	1728	345600	0.0000
sexe	1	12.00	12	2400	0.0000
age	2	104.00	52	10400	0.0000
sexe*age	2	0.00	0	0	1.0000
Error	6	0.03	0.005		
Total	11	116.030			



ANOVA : Y2					
Source	DF	SS	MS	F	p-value
Intercept	1	1728	1728	345600	0.00000
sexe	1	12	12	2400	0.00000
age	2	104	52	10400	0.00000
sexe*age	2	8	4	800	0.00000
Error	6	0.03	0.005		
Total	11	124.03			



ANOVA : Y 3					
	DF	SS	MS	F	p-value
Intercept	1	1728	1728	345600	0.00000
sexe	1	12	12	2400	0.00000
age	2	104	52	10400	0.00000
sexe*age	2	0.32	0.16	32	0.00063
Error	6	0.03	0.005		
Total	11	116.35			



Élimination d'interaction par transformation: 2 cas spéciaux

modèle multiplicatif $\mu_{ij} = \mu\, \alpha_i\, \beta_j$ **(94)**

$$\log(\mu_{ij}) = \log(\mu) + \log(\alpha_i) + \log(\beta_j) \quad (95)$$

$$\mu'_{ij} = \mu' + \alpha'_i + \beta'_j \quad (96)$$

Avec un modèle multiplicatif, une analyse de Y détecterait la présence d'interaction. Une analyse de Y' = log(Y) donne un modèle additif sans interaction.

modèle racine $\mu_{ij} = (\sqrt{\alpha_i} + \sqrt{\beta_j})^2$ **(97)**

$= \alpha_i + \beta_j + 2\sqrt{\alpha_i}\sqrt{\beta_j}$: **modèle avec interaction**

$$\mu'_{ij} = (\mu_{ij})^{0.5} = \sqrt{\alpha_i} + \sqrt{\beta_j} = \alpha'_i + \beta'_j \quad (98)$$

modèle sans interaction

Une analyse de $Y' = \sqrt{Y}$ donne un modèle additif sans interaction.

Transformation de Box-Cox : $Y' = Y^{\lambda}$ **peut « élimimer » une interaction.**

Analyse des effets avec un modèle additif : facteurs sans interaction sur Y

$$E(Y_{ijk}) = \mu_{ij} = \mu + \alpha_{i.} + \beta_{.j} \quad (99)$$

modèle ajusté : $\hat{Y}_{ij} = \hat{\mu} + \hat{\alpha}_{i.} + \hat{\beta}_{.j}$

$$= \hat{\mu} + (\hat{\mu}_{i.} - \hat{\mu}_{..}) + (\hat{\mu}_{.j} - \hat{\mu}_{..}) \quad (100)$$

$$= \overline{Y}_{...} + (\overline{Y}_{i..} - \overline{Y}_{...}) + (\overline{Y}_{.j.} - \overline{Y}_{...}) \quad (101)$$

contraste facteur A: $L = \sum c_i \mu_i \qquad \sum c_i = 0$ **(102)**

$$\hat{L} = \sum c_i \hat{\mu}_{i.} = \sum c_i \overline{Y}_{i..} \quad (103)$$

$$s^2(\hat{L}) = (MSE / bn) \sum c_i^2 \quad (104)$$

$(\hat{L} - L) / s(\hat{L}) \sim$ **Student avec (n-1)*ab degrés de liberté**

Intervalle confiance L : $\hat{L} \pm t(1 - \alpha ; (n-1)*ab)$ **(105)**

comparaison pairée : $D = \mu_{i..} - \mu_{i'.} \qquad i \neq i'$

procedure de Tukey : $H_0 : D = 0$ **vs** $H_a : D \neq 0$

rejet de H_0 si $|\hat{D}| = |\overline{Y}_{i..} - \overline{Y}_{i'..}| > (MSE/bn)^{0.5}\, q(1 - \alpha ; a ; (n-1)*ab)$ **(106)**

$q(1 - \alpha; a; (n-1)*ab)$: "studentized range"

autres procedures : Bonferronni, Scheffé

Aussi : formules analogues pour le facteur B

formules analogues pour le facteur A et le facteur B combiné
référence : Kutner et all 5 ed. p. 851-855

Exemple 11 : données sur la durée Y1 de l'apprentissage
6 groupes définis par les combinaisons de age X sexe

Tukey HSD test; variable Y1-durée Approximate Probabilities for Post Hoc Tests
Error: Between MS = .00500, df = 6

	sexe	age	{1} 10	{2} 12	{3} 17	{4} 8	{5} 10	{6} 15
1	H	jeune		0.0002	0.0002	0.0002	1.0000	0.0002
2	H	moyen	0.0002		0.0002	0.0002	0.0002	0.0002
3	H	agé	0.0002	0.0002		0.0002	0.0002	0.0002
4	F	jeune	0.0002	0.0002	0.0002		0.0002	0.0002
5	F	moyen	1.0000	0.0002	0.0002	0.0002		0.0002
6	F	agé	0.0002	0.0002	0.0002	0.0002	0.0002	

Analyse des effets en présence d'interactions importantes

Si on ne peut éliminer une interaction importante entre 2 facteurs par une transformation (logaritmique ou racine) alors on peut analyser les effets des facteurs avec le modèle à moyenne de cellules.

comparaison pairée $D = \mu_{ij} - \mu_{i'j'} \quad ij \neq i'j'$ (104)

procedure de Tukey $D : \hat{D} \pm T\, s(\hat{D})$ (105)

$\hat{D} = \bar{Y}_{ij.} - \bar{Y}_{i'j'.} \qquad s^2(\hat{D}) = 2*MSE / n$ (106)

$T = (1/\sqrt{2})\, q\,(1 - \alpha;\ ab,\ (n-1) * ab)$ (107)

Test de $H_0 : D = 0$ vs $H_a : D \neq 0$

Rejet de H_0 si $|\hat{D}| > (1/\sqrt{2})\, s(\hat{D})\, q\,(1 - \alpha;\ ab,\ (n-1)*ab)$ (108)

contrastes multiples $L = \sum\sum c_{ij}\, \mu_{ij} \qquad \sum\sum c_{ij} = 0$ (109)

procédure de Scheffé $L : \hat{L} \pm S\, s(\hat{L})$ (110)

$\hat{L} = \sum\sum c_{ij}\, \bar{Y}_{ij.}$ (111)

$s(\hat{L}) = (MSE / n) * \sum\sum c_{ij}^2$ (112)

$S^2 = (ab-1) * F\,(1 - \alpha;\ ab-1,\ (n-1)*ab)$ (113)

Test de $H_0 : L = 0$ vs $H_a : L \neq 0$

Rejet de H_0 si $(\hat{L}^2 / ((ab-1)\, s^2(\hat{L})) > F\,(1 - \alpha;\ ab-1,\ (n-1)*ab)$ (114)

Exemple 11 : données d'apprentissage Y2

Scheffe test; variable Y2-durée Probabilities for Post Hoc Tests
Error: Between MS = .00500, df = 6

	sexe	age	{1} 9	{2 12	{3} 18	{4} 9	{5} 10	{6} 14
1	H	jeune		0.0000	0.0000	1.0000	0.0002	0.0000
2	H	moyen	0.0000		0.0000	0.0000	0.0000	0.0000
3	H	agé	0.0000	0.0000		0.0000	0.0000	0.0000
4	F	jeune	1.0000	0.0000	0.0000		0.0002	0.0000
5	F	moyen	0.0002	0.0000	0.0000	0.0002		0.0000
6	F	agé	0.0000	0.0000	0.0000	0.0000	0.0000	

« Pooling » lorsque l'interaction n'est pas significative

Avec 2 facteurs on postule toujours pour commencer un modèle avec un effet d'interaction (voir eq. (78)). Si à l'analyse on constate que l'interaction n'est pas significative, on peut réviser le modèle et postuler un modèle sans interaction c-à-d sans le terme $(\alpha\beta)_{ij}$ dans l'équation (78). Cette opération peut se faire à condition que le ratio F soit petit, disons < 2. Les conséquences de cette opération sont:

$$\text{SSE (modèle révisé)} = \text{SSE (modèle complet)} + \text{SSAB} \qquad (115)$$

$$\text{df (modèle révisé)} = (a-1)(b-1) + (n-1)ab = nab - a - b - 1 \qquad (116)$$

$$\text{MSE (modèle révisé)} = \text{SSE (modèle révisé)} / \text{df (modèle révisé)} \qquad (117)$$

En général, le MSE (modèle révisé) devrait être « voisin » du MSE (modèle complet). Les résultats des tests de signification des effets principaux des facteurs ne devraient pas changer suite à cette opération.

CAS où le nombre de répétition n = 1

Si n = 1, il n'est plus possible d'estimer la variance σ^2 avec les répétitions. La seule possibilité est de postuler un modèle sans interaction :

$$Y_{ij} = \mu + \alpha_i + \beta_j + + \varepsilon_{ij} \qquad (118)$$

Le tableau d'analyse de la variance prend une forme plus simple comparativement à celui avec $n_{ij} = n > 1$ (page 52).

ANOVA : 2 facteurs avec n = 1

Source	SS	df	MS	F	p-value
A	SSA	a-1	MSA = SSA / (a -1)	MSA / MSAB	
B	SSB	b-1	MSB = SSB / (b-1)	MSB / MSAB	
erreur	SSAB	(a-1)(b-1)	MSAB = SSAB / (a-1)(b-1)	------	
totale	SStotale	ab - 1	------	------	

$$SSA = b \sum (Y_{i.} - Y_{..})^2 \qquad SSB = a \sum (Y_{.j} - Y_{..})^2 \qquad (119)$$

$$SSAB = \sum\sum (Y_{ij} - \overline{Y}_{i.} - \overline{Y}_{.j} + \overline{Y}_{..})^2 \qquad (120)$$

$$SStotale = \sum\sum (Y_{ij} - \overline{Y}_{..})^2 \qquad (121)$$

La somme de carrés SSAB est maintenant employée pour estimer l'erreur expérimentale.

Le test d'hypothèse de la nullité du facteur A est basé sur le ratio F = SSA /SSAB qui suit une loi F (a – 1, (a-1)(b-1)). Un résultat analogue s'applique pour le facteur B.

Le test de Tukey pour détecter la présence d'une interaction avec n = 1

Il est possible de tester la présence d'une interaction avec n = 1
On suppose que l'interaction est de la forme d'un produit comme dans un polynôme :

$$(\alpha\beta)_{ij} = D\,\alpha_i\,\beta_j \qquad (122)$$

où D est une constante. L'estimateur de D est

$$\hat{D} = [\sum\sum \hat{\alpha}_i\,\hat{\beta}_j\,Y_{ij}] \;/\; [\sum \hat{\alpha}_i^2 \sum \hat{\beta}_j^2] \qquad (123)$$

$$= [\sum\sum (\overline{Y}_{i.} - \overline{Y}_{..})(\overline{Y}_{.j} - \overline{Y}_{..})\,Y_{ij}] \;/\; C \qquad (124)$$

où $$C = \sum (\overline{Y}_{i.} - \overline{Y}_{..})^2 \sum (\overline{Y}_{.j} - \overline{Y}_{..})^2 \qquad (125)$$

La somme de carrés associée à cette forme d'interaction est:

$$SSAB^* = [\sum\sum (\overline{Y}_{i.} - \overline{Y}_{..})(\overline{Y}_{.j} - \overline{Y}_{..})\,Y_{ij}]^2 / C \qquad (126)$$

La somme de carrés résiduelle SSresid est :

$$SSresid = SStotale - SSA - SSB - SSAB^* \qquad (127)$$

On peut monter que $$F^* = SSAB^* / (SSresid / (ab - a - b)) \qquad (128)$$

suit une loi de fisher-Snedecor F (1, ab – a –b)

test de H_0 **:** $D = 0$ **vs** **Ha :** $D \neq 0$ (126)

on rejette H_0 **si** $F^* > F(1 - \alpha ; 1, ab - a - b)$ (127)

Exemple : données d'apprentissage (exemple 11) Y = Y2 et les 6 premières observations

$\overline{Y}_{..} = 11.95$ $\overline{Y}_{1.}$ **(homme)=** 12.95 $\overline{Y}_{2.}$ **(femme)** $= 10.95$

$\overline{Y}_{.1}$ **(jeune)** $= 8.95$ $\overline{Y}_{.2}$ **(moyen)** $= 10.95$ $\overline{Y}_{.3}$ **(agé)** $= 15.95$

$\sum\sum (\overline{Y}_{i.} - \overline{Y}_{..})^2 = 2$ $\sum\sum (\overline{Y}_{.j} - \overline{Y}_{..})^2 = 26$

$\sum\sum (Y_{i.} - Y_{..})(Y_{.j} - Y_{..})\,Y_{ij} = 0$

$\hat{D} = 0$ $SSAB^* = 0^2 / 2*26 = 0$

SSresid = 61.95 - 6 - 52 - 0 = 3.95 F* = 0 / 3.95 = 0

H_0 n'est pas rejetée.

Remarques

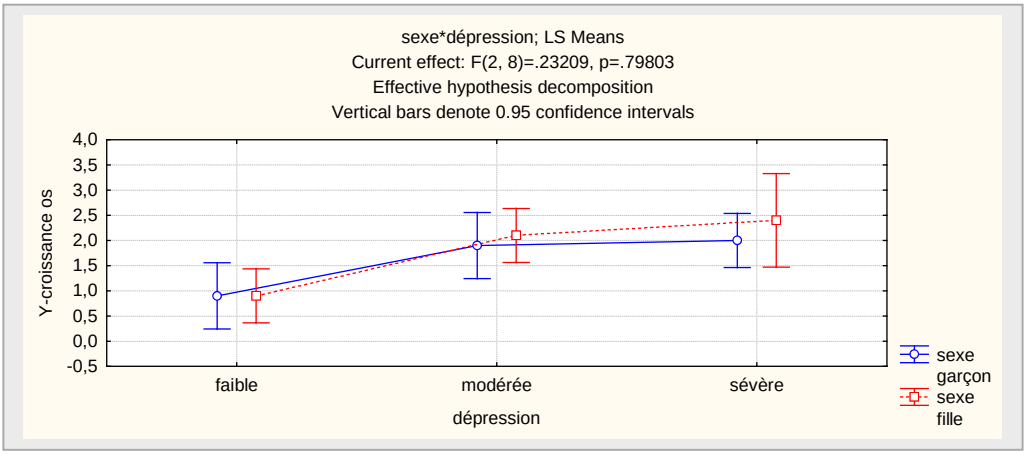
- **Si le test de l'interaction est significatif on peut employer une transformation de type Box-Cox pour essayer de l'éliminer. Cela élargit l'éventail des transformations log et racine.**
- **Le test de Tukey est probablement plus utile lorsque les 2 facteurs sont quantitatifs.**

Cas où le nombre de répétitions n_{ij} sont inégaux et > 0

Dans les études observationnelles et, occasionnellement dans les études expérimentales, les tailles échantillonnles n_{ij} sont généralement inégales. Cela a pour conséquence que l'orthogonalité de la décomposition des sommes de carrés de l'ANOVA n'existe plus : l'adition des sommes de carrés des effets (principaux et interactions) n'est plus égale à la somme totale des carrés SSTO. La situation est analogue à celle de la méthode régression où l'on a rarement l'orthogonalité dans la structure des données. Les tailles inégales ont pour conséquence que l'unicité du modèle n'est plus assurée. On analyse les données avec le modèle linéaire général en introduisant des variables indicatrices pour les modalités.

Exemple 12 : données sur la croissance des os (Y) Kutner et all 5 ed. p. 954
facteurs : sexe enfant (gaçon, fille) dépression (sévère, modérée, faible)

	sexe	dépression	rep	Y-croissance os		X1	X2	X3	X1X2	X1X3
1	garçon	sévère	1	1.4		1	1	0	1	0
2	garçon	sévère	2	2.4		1	1	0	1	0
3	garçon	sévère	3	2.2		1	1	0	1	0
4	garçon	modérée	1	2.1		1	0	1	0	1
5	garçon	modérée	2	1.7		1	0	1	0	1
6	garçon	faible	1	0.7		1	-1	-1	-1	-1
7	garçon	faible	2	1.1		1	-1	-1	-1	-1
8	fille	sévère	1	2.4		-1	1	0	-1	0
9	fille	modérée	1	2.5		-1	0	1	0	-1
10	fille	modérée	2	1.8		-1	0	1	0	-1
11	fille	modérée	3	2.0		-1	0	1	0	-1
12	fille	faible	1	0.5		-1	-1	-1	1	1
13	fille	faible	2	0.9		-1	-1	-1	1	1
14	fille	faible	3	1.3		-1	-1	-1	1	1



On introduit une variable indicatrice (X1) pour le facteur sexe et 2 variables indicatrices (X2, X3) pour le facteur dépresssion :

X1 = 1	**si sexe = garçon**	**X1 = -1**	**si sexe = fille**
X2 = 1	**si dépresssion = sévère**	**X2 = 0**	**si modérée** X2 = -1 si faible
X3 = 1	**si dépresssion = modérée**	**X3 = 0**	**si sévère** X2 = -1 si faible

Le modèle à effets principaux et d'interaction est :

$$Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk} \qquad (128)$$

$$i = 1, 2 \qquad j = 1, 2, 3 \qquad k = 1, 2,..., n_{ij}$$
$$n_{11} = 3 \qquad n_{12} = 2 \qquad n_{13} = 2$$
$$n_{21} = 1 \qquad n_{22} = 3 \qquad n_{23} = 3$$

Le modèle équivalent (modèle complet) écrit avec les variables indicatrices est:

$$MC : Y_{ijk} = \mu + \alpha_1 X_{ijk1} + \beta_1 X_{ijk2} + \beta_2 X_{ijk3} + (\alpha\beta)_{11} X_{ijk1} X_{ijk2} + (\alpha\beta)_{12} X_{ijk1} X_{ijk3} + \varepsilon_{ijk} \qquad (129)$$

La procédure pour l'examen des effets de A (sexe), de B (dépression) et de l'interaction AB est basée sur l'ajustement de 4 modèles de régression : MC, MR1, MR2, MR3:

1. **test de la nullité des effets d'interaction $(\alpha\beta)_{11}$ et de $(\alpha\beta)_{12}$**

$$H_1 : (\alpha\beta)_{11} = (\alpha\beta)_{12} = 0 \text{ vs } H_a : \text{non } H_0$$

$$MC + H1 : MR1 : Y_{ijk} = \mu + \alpha_1 X_{ijk1} + \beta_1 X_{ijk2} + \beta_2 X_{ijk3} + \varepsilon_{ijk} \qquad (130)$$

2. **test de la nullité de l'effet principal du facteur A**

$$H_2 : \alpha_1 = 0 \text{ vs } H_a : \text{non } H_0$$

$$MC + H2 : MR2 : Y_{ijk} = \mu + \beta_1 X_{ijk2} + \beta_2 X_{ijk3} + (\alpha\beta)_{11} X_{ijk1} X_{ijk2} + (\alpha\beta)_{12} X_{ijk1} X_{ijk3} + \varepsilon_{ijk} \qquad (131)$$

3. **test de la nullité de l'effet principal du facteur B**

$$H_3 : \beta_1 = \beta_2 = 0 \text{ vs } H_a : \text{non } H_0$$

$$MC + H3 : MR3 : Y_{ijk} = \mu + \alpha_1 X_{ijk1} + (\alpha\beta)_{11} X_{ijk1} X_{ijk2} + (\alpha\beta)_{12} X_{ijk1} X_{ijk3} + \varepsilon_{ijk} \qquad (132)$$

MR1, MR2, MR3 constituent des modèles réduits; ils contiennent moins de paramètres que le modèle complet MC. Les tests d'hypothèses H_1, H_2, H_3 sont réalisés par un test F basé sur la réduction de la somme de carrés de l'erreur (résiduelle) SSE calculée sous le modèle complet SSE(C) et la somme de carrés de l'erreur SSE calculée sous le modèle réduit SS(R).

$$F^* = [(SSE(R) - SSE(C)) / (df(R) - df(C))] / [SSE(C) / df(C)] \qquad (133)$$
$$= [\text{delta SSE / delta (df)}] / MSE(C)$$

Le ratio F* suit une loi F (df(R) - df(C) , df(C))

Les résultats de ces analyses de modélisation sont résumés dans le tableau suivant.

modèle	constante	α_1	β_1	β_2	$(\alpha\beta)_{12}$	$(\alpha\beta)_{12}$
MC	1.70	0.1	0.10	0.50	0.30	0.0
MR1	1.68	0.086	0.467	0.327	0	0
MR2	1.69	0	0.444	0.328	- 0.067	- 0.017
MR3	1.63	0.019	0	0	0.067	- 0.193

SSTO = 5.77 avec 13 degrés de liberté

modèle	SSmodèle	SSE	df(mod)	df(C)	R^2	R^2 ajusté
MC	4.47	1.30	5	8	0.77	0.63
MR1	4.40	1.38	3	10	0.76	0.69
MR2	4.35	1.42	4	9	0.75	0.64
MR3	0.28	5.49	3	10	0.55	0.35

test	delta SSE / delta (df)	MSE	Ratio F	p value
H_1	0.037 / 2	1.3 / 8	0.23	0.80
H_2	0.120 / 1	1.3 / 8	0.74	0.41
H_3	2.09 / 2	1.3 / 8	12.89	0.003

H_3 est rejetée : le facteur dépression est significatif sur Y.
H_2 n'est pas rejetée : le facteur sexe n'est pas signicatif sur Y.
H_1 n'est pas rejetée : il n'y a pas d'effet d'interaction.

ANOVA

	df	SS	MS	F	p-value
Intercept	1	34.680	34.680	213.42	0.0000
sexe	1	0.120	0.120	0.74	0.4152
dépression	2	4.190	2.095	12.89	0.0031
sexe*dépression	2	0.075	0.038	0.23	0.7980
Error	8	1.300	0.162		
Total	13	5.774			

A la suite du tableau d'analyse de la variance, il est utile de compléter l'analyse avec des intervalles de confiance :

- **les moyennes de niveau (modalité) de facteurs**
- **les différences entre les paires de moyennes**
- **des contrastes ou combinaisons linéaires de moyennes**
- **des comparaisons multiples**

Consulter Kutner et all 5 ed. p. 961-962.

Dans certaines applications les moyennes μ_{ij} $\mu_{i.}$ $\mu_{.j}$ n'ont pas la même importance. Les inférences se font sur des combinaisons linéaires de moyennes dont les coefficients réflètent l'importance des moyennes.

Consulter Kutner et all 5 ed. p. 970-974.

9. Analyse de la variance avec un facteur bloc

Voir les notes de cours sur la planification d'expérience MTH6301

http://www.cours.polymtl.ca/mth6301/mth6301-cours/DOE2011-chap07-blocs.pdf

10. Analyse de la variance avec des facteurs aléatoires

Notes de cours sont placées sur le site WEB

http://www.cours.polymtl.ca/mth6301/mth6302B/Modeles_d'analyse_de_variance_2012-partie_6.pdf

11. Analyse de covariance

Cette méthode d'analyse combine les éléments des modèles de régression et les modèles d'analyse de la variance. L'idée de base est d'augmenter les modèles d'analyse de variance contenant les effets de facteurs catégoriques (qualitatifs) avec une ou plusieurs variables continues (quantitatives) qui sont reliées à la variable de réponse. Cette augmentation a pour objectif de réduire la variance du terme d'erreur dans le modèle augmentant ainsi la sensibilité de l'analyse à détecter des effets significatifs. Les modèles d'analyse de covariance sont des cas particuliers des modèles d'analyse de régression avec un mélange de variables quantitatives et de variables qualitatives représentées par des variables indicatrices de type 0-1.

Variables concomitantes (covariables)

Chaque variable quantitative est appelée une *covariable.* Ces variables doivent avoir une certaine correlation avec la variable de réponse. Ces variables doivent être observées durant ou après l'étude et ne doivent pas être reliées aux facteurs catégoriques (traitements) de l'étude.

Exemple 13 : campagne promotionnelle de vente d'un produit alimentaire
facteur : type de promotion
1 : échantillonnage du produit par les clients
2 : espace additionnel dans les étagères habituelles
3 : étalage additionnel dans les allées
réponse Y : volume des ventes
covariable X : volume des ventes de la période précédente (X-vente avant)

Kutner et all 5 ed. p. 926

	promotion	magasin	X-vente avant	Y-vente
1	1	11	21	38
2	1	12	26	39
3	1	13	22	36
4	1	14	28	45
5	1	15	19	33
6	2	21	34	43
7	2	22	26	38
8	2	23	29	38
9	2	24	18	27
10	2	25	25	34
11	3	31	23	24
12	3	32	29	32
13	3	33	30	31
14	3	34	16	21
15	3	35	29	28

Exemple 14 : bois traité

Contexte et données de l'étude CCA : « Chromate Copper Arsenate »

Le CCA préserve le bois des dommages causés par l'eau et les insectes. Le CCA présente des risques pour la santé et l'environnement. Une nouvelle solution dite « organic » pourrait remplacer le CCA. Une étude statistique fut conduite pour comparer les deux solutions. Les données du fichier représentent 360 planches de pin traitées avec les 2 solutions à 3 niveaux de concentration: 0.25, 0.80 et 2.25 (lbs/pi cu). Ces niveaux de concentration sont employés pour 3 applications typiques: bois exposé à l'air, bois de fondation, bois en eau salée. Les planches traitées furent placées dans des chambres à un vieillissement accéléré durant plusieurs heures. Chaque heure représente l'équivalent d'une exposition d'une année aux éléments. Suite à l'application de la solution et du processus de vieillissement, chaque planche fut soumise à un test de bris à la rupture représenté par la variable « Load ». Cette variable de réponse est du type « larger the better ». Les facteurs contrôlés sont : Concentration, Solution. Heures est un facteur continu mesuré (covariable). Objectif de l'expérience : comparer la performance (LOAD) de la solution « organic » avec la solution CCA. La solution « organic » est-elle aussi bonne que la solution CCA?

Données

	Concentration Facteur 1	Solution Facteur 2	X-heure covariable	Y-Load réponse
1	0.25	CCA	1.0	15.597
2	0.25	ORG	1.0	11.025
3	0.80	CCA	1.0	12.329
4	0.80	ORG	1.0	14.275
5	2.50	CCA	0.9	18.739
6	2.50	ORG	1.0	16.668
7	0.25	CCA	2.0	14.826
8	0.25	ORG	2.0	9.447
9	0.80	CCA	2.1	14.553
10	0.80	ORG	2.1	12.894
.	.	.	.	.
358	0.80	ORG	60.1	9.183
359	2.50	CCA	59.9	15.035
360	2.50	ORG	60.1	15.569

Exemple 15 : comparaison de 2 méthodes d'enseignement

facteur : méthode d'enseignement (M = magistral, D = distance)
réponse Y : résultats à un examen commun
covariables X1 : cote R (CEGEP)
X2 : moyenne cumulative de l'étudiant à Polytechnique

Modèle d'analyse de covariance avec un facteur

$$Y_{ij} = \mu + \tau_i + \gamma (X_{ij} - \overline{X}_{..}) + \varepsilon_{ij} \qquad (134)$$

où μ : effet général (moyenne globale)

τ_i : effet différentiel du facteur avec la contrainte

$$\sum \tau_i = 0 \qquad (135)$$

γ : coefficient de régression (pente) entre X et Y indépendante du facteur

X_{ij} : constantes connues de la covariable X

ε_{ij} : terme d'erreur $\sim N(0, \sigma^2)$

$i = 1, 2, \ldots, g$ $\quad j = 1, 2, \ldots, n_i$

conséquence $Y_{ij} \sim N(\mu_{ij}, \sigma^2)$ $\quad \mu_{ij} = \mu_. + \tau_i + \gamma (X_{ij} - \overline{X}_{..})$

remarques

- l'hypothèse que la pente γ est constante est cruciale : il n'y a pas d'interaction entre la covariable X et le facteur d'intérêt. Sinon il faut employer des équations de régresssions distinctes : une pour chque modalité du facteur.
- On peut avoir plusieurs covariables : X, W,..
- On utilisera des variables indicatrices pour le facteur qualitif : on introduit (g – 1) variables I_i prenant les valeurs 1, -1, 0

$$I_i = \begin{cases} 1 & \text{si l'observation provient du traitement } i = 1, 2, .., g-1 \\ -1 & \text{si l'observation provient du traitement } g \\ 0 & \text{autrement} \end{cases}$$

- l'analyse du modèle de covariance devient un cas particulier du modèle linéaire général.

L'équation (135) devient

$$Y_{ij} = \mu + \tau_1 I_{ij1} + \tau_2 I_{ij2} + \ldots \tau_{g-1} I_{ijg-1} + \gamma x_{ij} + \varepsilon_{ij} \qquad (136)$$

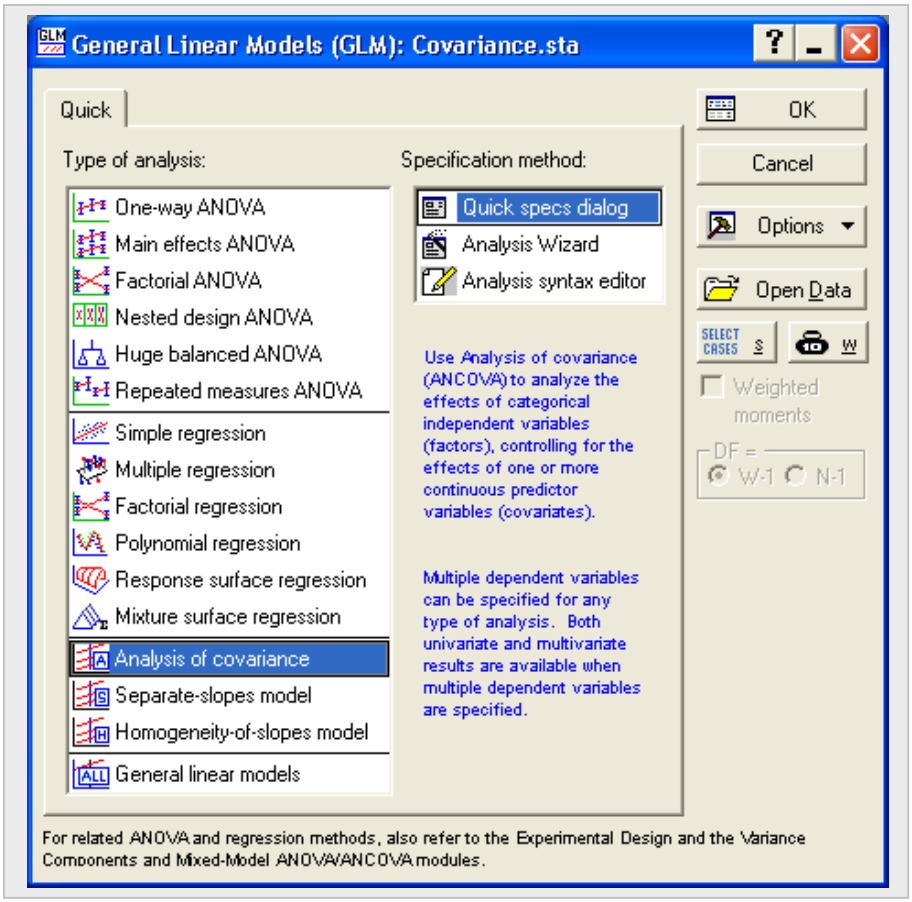
où $x_{ij} = X_{ij} - \overline{X}_{..}$ (137)

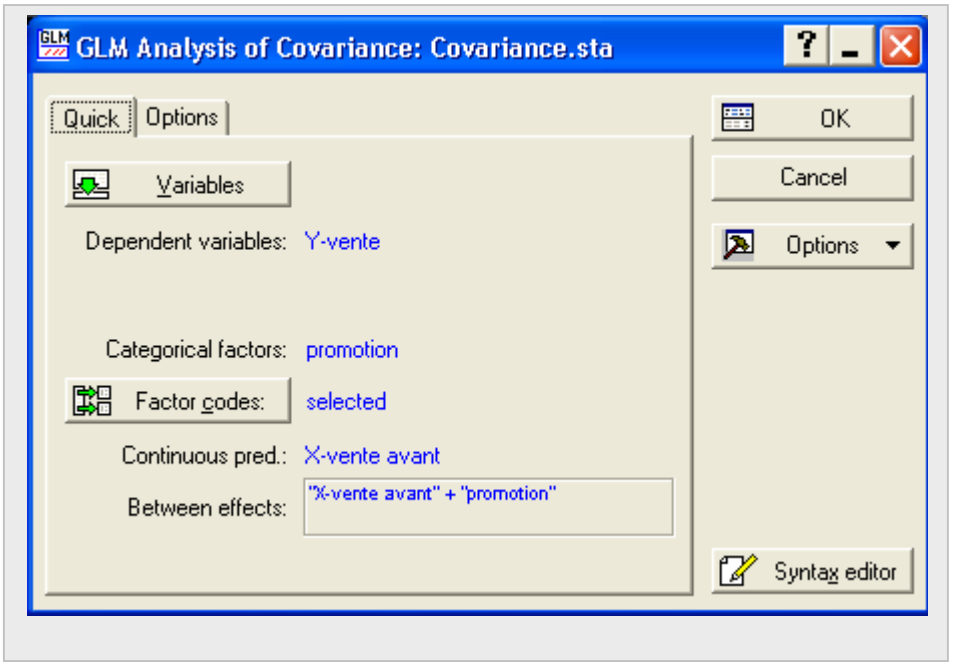
inférence d'intérêt $H_0 : \tau_1 = \tau_2 = \ldots = \tau_{g-1} = 0$ (138)

vs H_a : les τ_i ne sont pas tous nuls

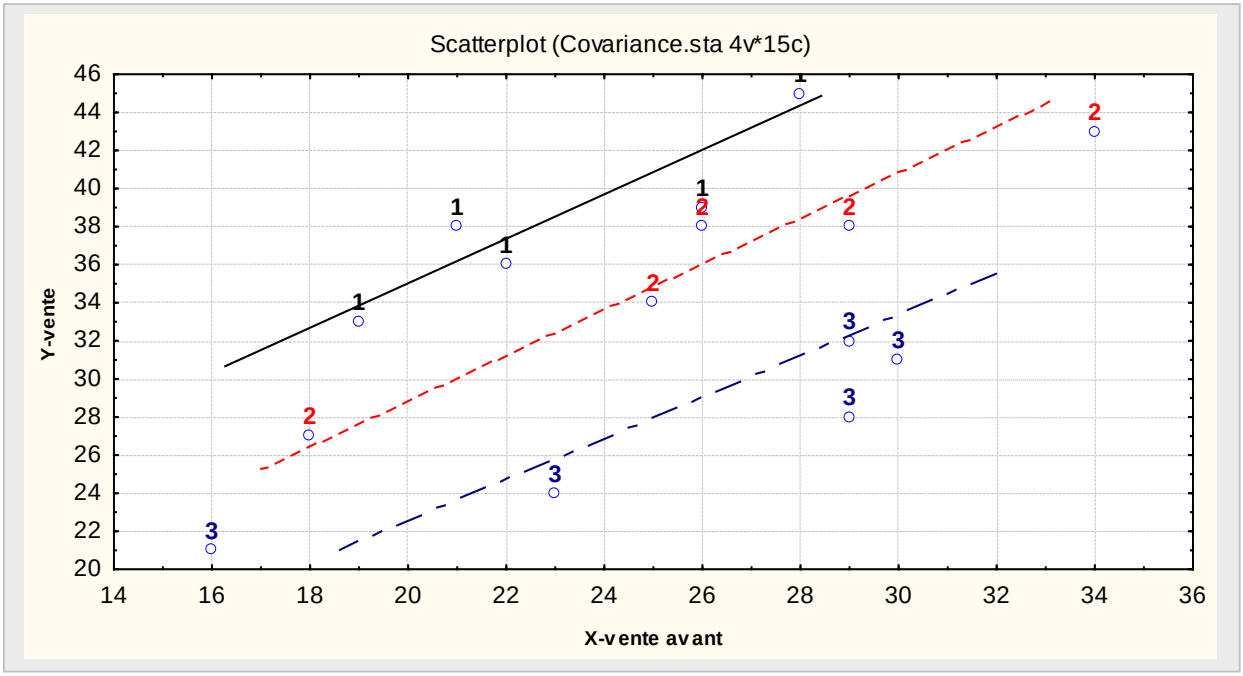
le test est conduit en comparant le résultat de l'ajustement du modèle complet avec le résultat de l'ajustement du modèle réduit sous l'hypothèse H_0.

L'analyse de covariance avec Statistica utilise le module « Linear general model »



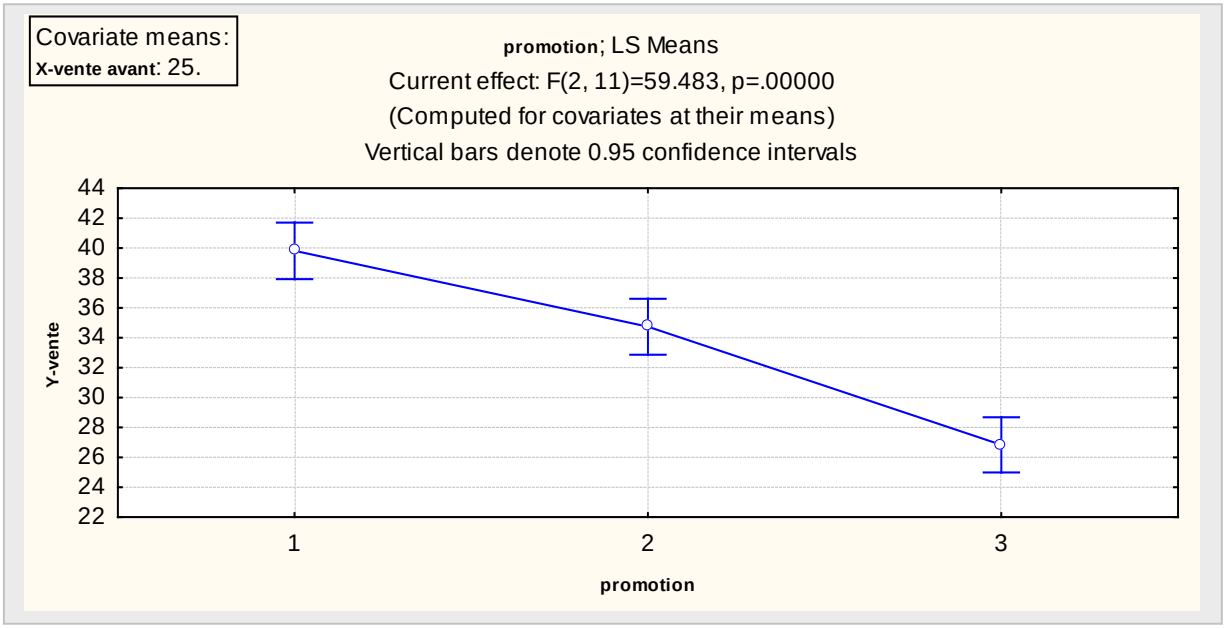


Exemple 13 : campagne promotionnelle



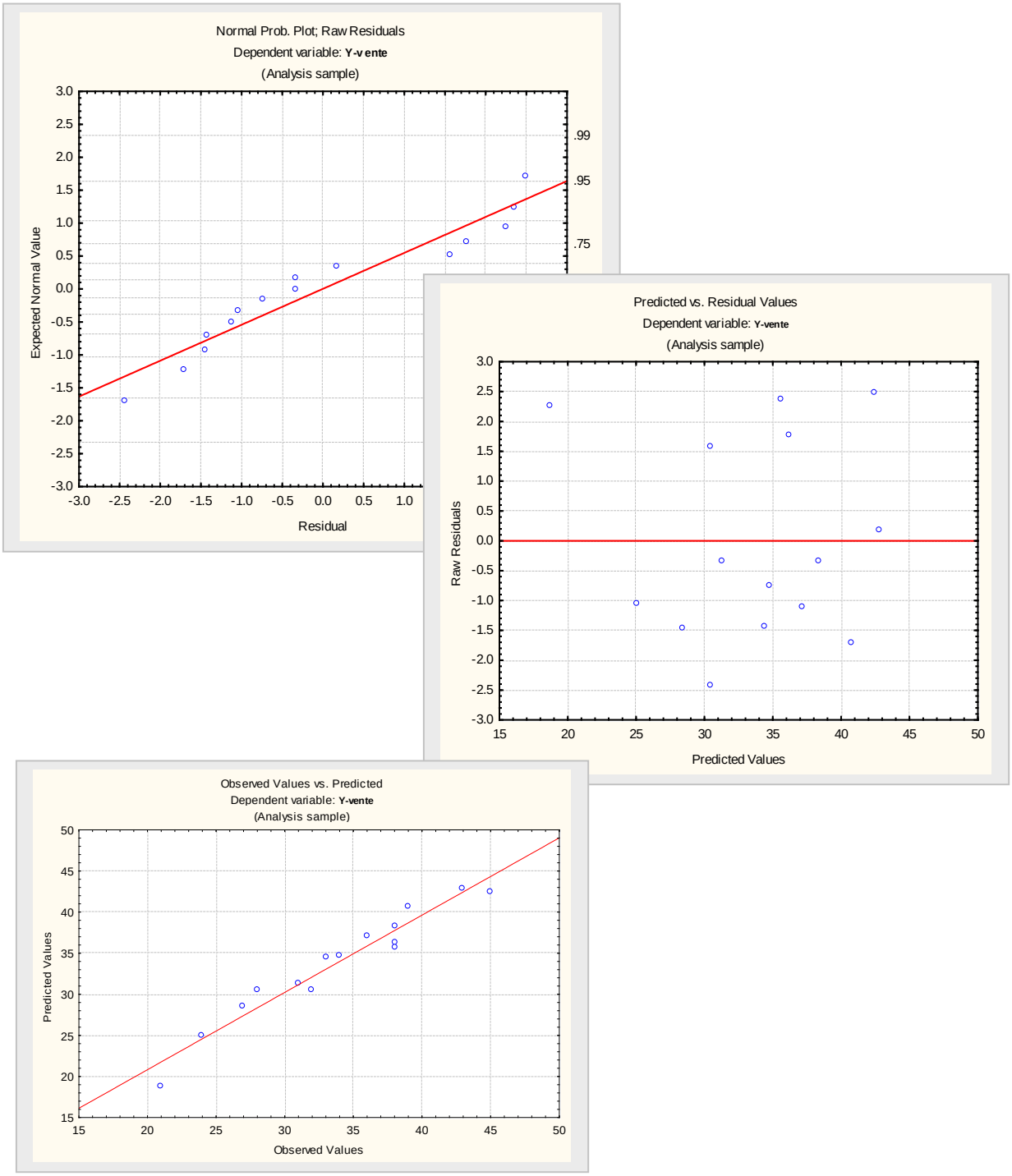
ANCOVA

	df	Y-vente SS	Y-vente MS	Y-vente F	Y-vente p-value
Intercept	**1**	**66.16**	**66.16**	**18.9**	**0.001168**
X-vente avant	**1**	**269.03**	**269.03**	**76.7**	**0.000003**
promotion	**2**	**417.15**	**208.58**	**59.5**	**0.000001**
Error	**11**	**38.57**	**3.51**		
Total	**14**	**646.40**			



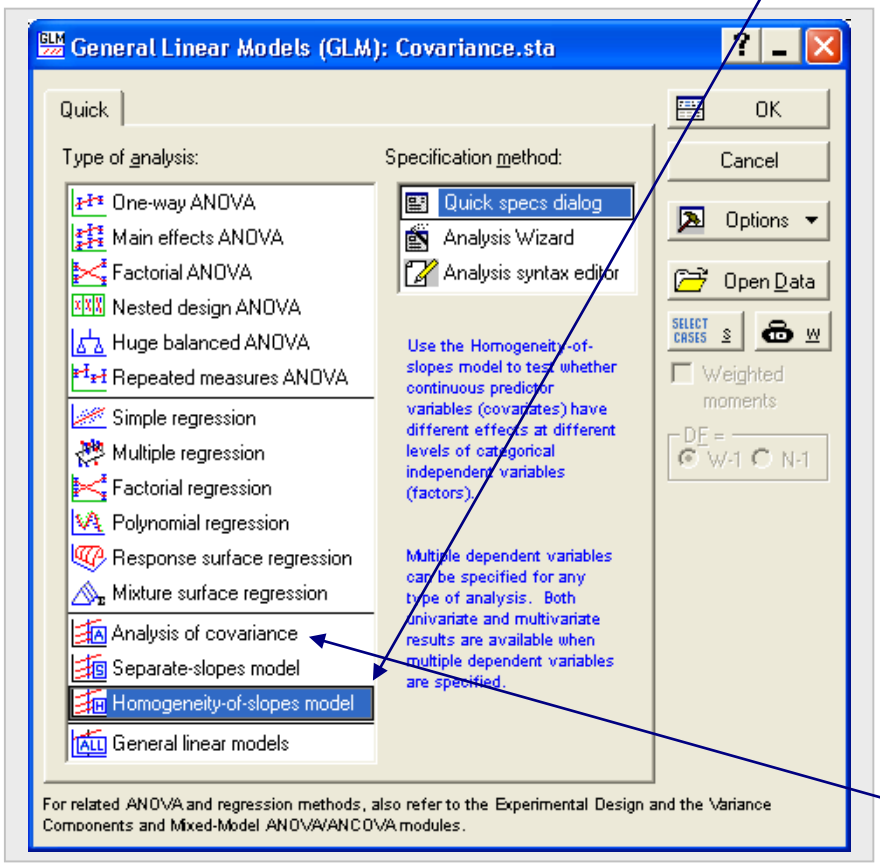
Tukey HSD test; MS = 3.5065, df = 11.000				
	promotion moyenne	{1} 38.2	{2} 36.0	{3} 27.2
1	1		0.1970	0.0002
2	2	0.1970		0.0002
3	3	0.0002	0.0002	

Analyse des résidus



Avait-on raison de penser que la pente γ est identique pour les 3 types de promotion?

On peut tester cette hypothèse avec la procédure « Homogeneity of slopes model »

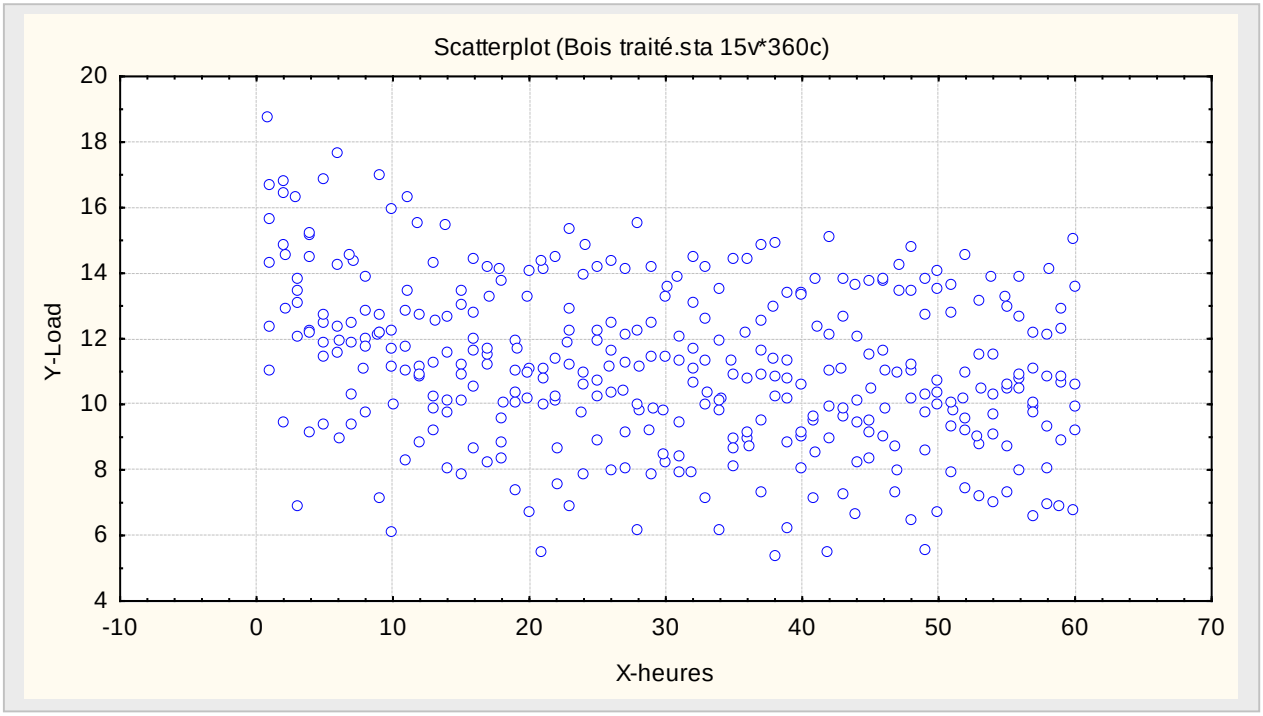


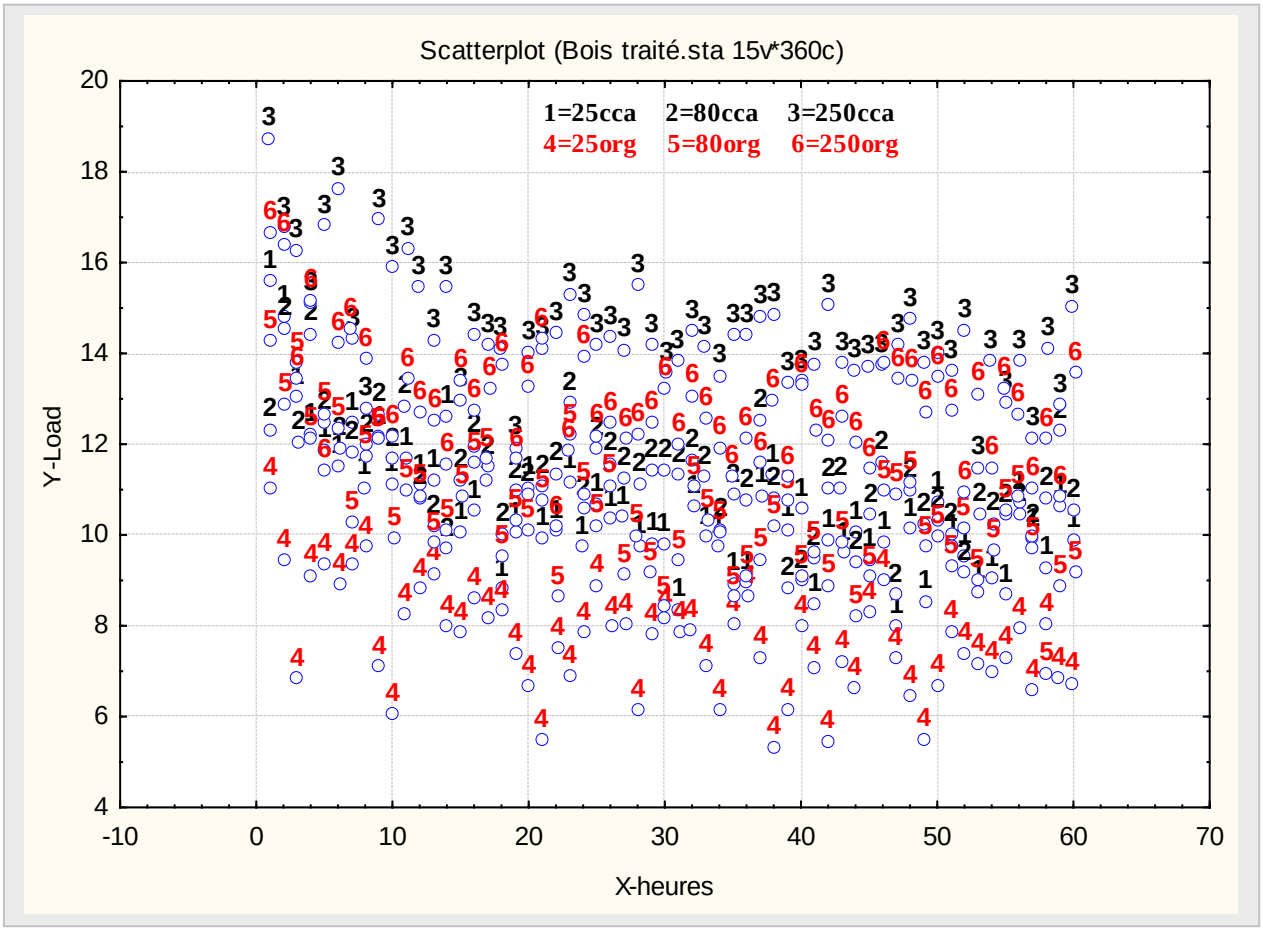
pentes égales ? : test de l'interaction covariable X facteur

	DF	SS	MS	F	p-value
Intercept	**1**	**48.74**	**48.74**	**13.917**	**0.0047**
promotion	**2**	**1.26**	**0.63**	**0.180**	**0.8379**
X-vente avant	**1**	**243.14**	**243.14**	**69.423**	**0.0000**
promotion*X-vente avant	**2**	**7.05**	**3.53**	**1.007**	**0.4032**
Error	**9**	**31.52**	**3.50**		
Total	**14**	**646.40**			

Le modèle à pentes égales est acceptable car l'interaction n'est pas significative.
Lorsque l'on rejette l'hypothèse, on emploie le modèle à pentes distinctes.

Exemple 14 : bois traité **Analyse de covariance avec 2 facteurs**





Test de l'homogénité des pentes

Test d'homogénité pentes

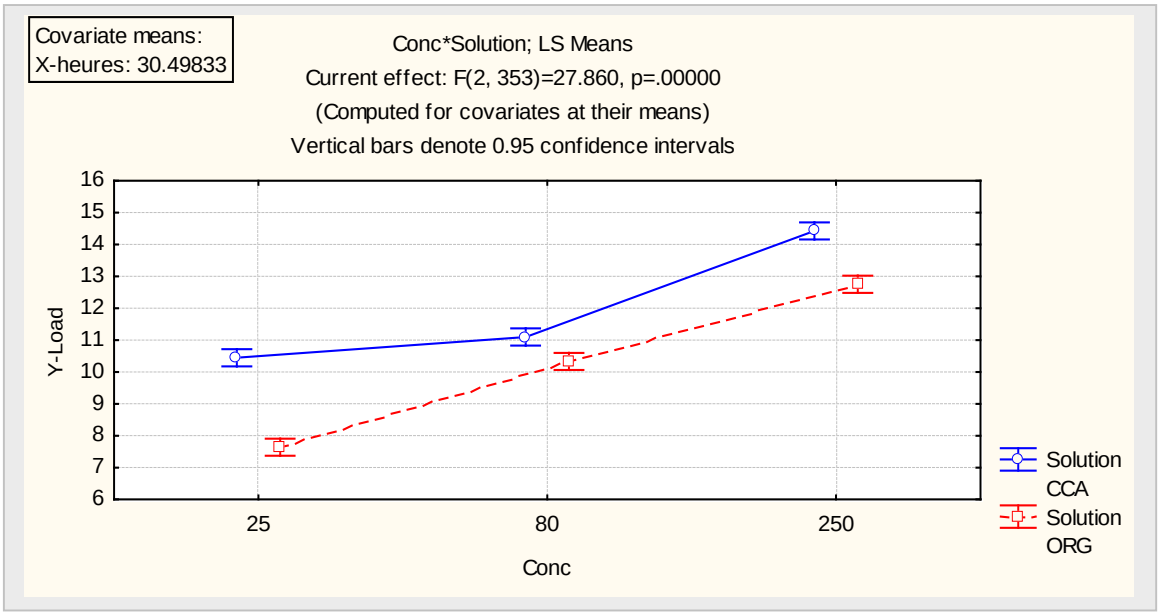
	DF	SS	MS	F	p-value
Intercept	1	13411.29	13411.29	12037.08	0.000000
Conc	2	276.09	138.04	123.90	0.000000
Solution	1	90.91	90.91	81.59	0.000000
X-heures	1	180.89	180.89	162.35	0.000000
Conc*Solution	2	33.40	16.70	14.99	0.000001
Conc*X-heures	2	1.52	0.76	0.68	0.506072
Solution*X-heures	1	2.36	2.36	2.12	0.146050
Conc*Solution*X-heures	2	5.07	2.53	2.27	0.104406
Error	348	387.73	1.11		
Total	359	2184.32			

Il n'y a pas d'interaction de la covariable X-heures avec les facteurs Conc et Solution.

Modèle avec pente homogène (analyse de covariance)

ANCOVA

	DF	SS	MS	F	p-value
Intercept	1	13411.64	13411.64	11934.61	0.000000
X-heures	1	180.93	180.93	161.01	0.000000
Conc	2	1268.41	634.21	564.36	0.000000
Solution	1	275.61	275.61	245.26	0.000000
Conc*Solution	2	62.62	31.31	27.86	0.000000
Error	353	396.69	1.12		
Total	359	2184.32			



12. ANOVA avec trois facteurs

La méthode d'ANOVA avec trois facteurs (multifacteurs) est une généralisation de ce qui a été vu avec deux facteurs. Toutefois lorsqu'on a le même nombre d'observations pour chaque combinaison des facteurs (cellule), l'analyse statistique est grandement simplifiée. Dans les autres cas, par exemple avec des tailles très inégales et des cellules sans observation (données manquantes), le traitement statistique est plus compliqué. L'approche régression est est recommandée.

Exemple 15 : influence de 3 facteurs sur le temps d'apprentissage Y :
3 exemples de variable de réponse : Y1, Y2, Y3
facteur 1 : genre (homme, femme)
facteur 2 : age (jeune, moyen, agé)
facteur 3 : IQ (élevé, normal)

Kutner et all 5 ed. p 994, 999, 1001

id	Genre	Age	IQ	Y1	Y2	Y3
1	homme	jeune	élevé	9	10.5	10
2	homme	moyen	élevé	12	12.5	12
3	homme	agé	élevé	18	16	15.5
4	homme	jeune	normal	19	17.5	18
5	homme	moyen	normal	20	19.5	20
6	homme	agé	normal	21	23	23.5
7	femme	jeune	élevé	9	9.5	10
8	femme	moyen	élevé	10	10.5	11
9	femme	agé	élevé	14	13	13.5
10	femme	jeune	normal	19	18.5	18
11	femme	moyen	normal	20	19.5	19
12	femme	agé	normal	21	22	21.5

Modèle de TYPE CELLULE avec le même nombre d'observation dans chaque cellule

$$Y_{ijk\ell} = \mu_{ijk} + \varepsilon_{ijk\ell} \qquad (139)$$

μ_{ijk} : paramètres - moyenne de la cellule (i, j, k)

$\varepsilon_{ijkl} \sim N(0, \sigma^2)$

$i = 1, 2, ..., a \quad j = 1, 2,.., b \quad k = 1, 2,...,c$

$\ell = 1, 2, ..., n_{ijk} = n \quad N = abcn$

Notation remarque : la sommation est faite sur l'indice représenté par un point (.)

$$\mu_{ij.} = (1/c)\sum \mu_{ijk} \qquad \mu_{i.k} = (1/b)\sum \mu_{ijk}$$
$$\mu_{.j.k} = (1/a)\sum \mu_{ijk} \qquad \mu_{i..} = (1/bc)\sum\sum \mu_{ijk} \qquad (140)$$
$$\mu_{.j..} = (1/ac)\sum\sum \mu_{ijk} \qquad \mu_{..k} = (1/ab)\sum\sum \mu_{ijk}$$
$$\mu_{...} = (1/abc)\sum\sum\sum \mu_{ijk}$$

Modèle de TYPE EFFETS

$$Y_{ijkl} = \mu + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk} + (\alpha\beta\gamma)_{ijk} + \varepsilon_{ijkl} \quad (141)$$

effets principaux (142)

$$\alpha_i = \mu_{i..} - \mu \qquad \sum \alpha_i = 0$$

$$\beta_j = \mu_{.j.} - \mu \qquad \sum \beta_j = 0$$

$$\gamma_k = \mu_{..k} - \mu \qquad \sum \gamma_k = 0$$

Interactions doubles (143)

$$(\alpha\beta)_{ij} = \mu_{ij.} - \mu_{i..} - \mu_{.j.} + \mu \qquad \sum_i (\alpha\beta)_{ij} = 0 \qquad \sum_j (\alpha\beta)_{ij} = 0$$

$$(\alpha\gamma)_{ik} = \mu_{i.k} - \mu_{i..} - \mu_{..k} + \mu \qquad \sum_i (\alpha\gamma)_{jk} = 0 \qquad \sum_k (\alpha\gamma)_{jk} = 0$$

$$(\beta\gamma)_{jk} = \mu_{.j.k} - \mu_{.j.} - \mu_{..k} + \mu \qquad \sum_j (\beta\gamma)_{jk} = 0 \qquad \sum_k (\beta\gamma)_{jk} = 0$$

Interaction triple (144)

$$(\alpha\beta\gamma)_{ijk} = \mu_{ijk} - [\mu_{.} + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk}]$$

$$= \mu_{ijk} - (\mu_{ij.} + \mu_{i.k} + \mu_{.j.k}) + (\alpha_i + \beta_j + \gamma_k) - \mu_{...}$$

$$\sum_i (\alpha\beta\gamma)_{ijk} = 0 \quad \text{pour tout } j, k$$

$$\sum_j (\alpha\beta\gamma)_{ijk} = 0 \quad \text{pour tout } i, k$$

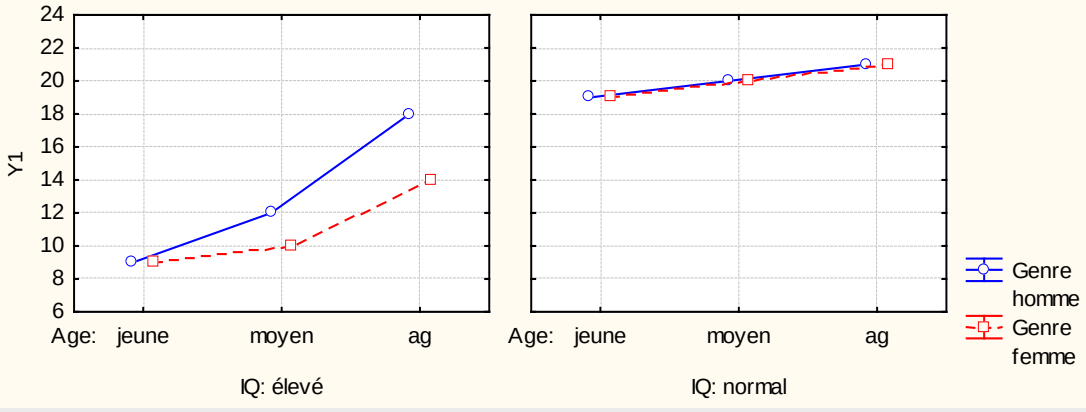
$$\sum_k (\alpha\beta\gamma)_{ijk} = 0 \quad \text{pour tout } i, j$$

remarques

- **les formes (139) et (141) sont équivalentes**
- **la formulation (141) est plus utile sauf si, exceptionnellement, on peut monter que la présence d'interactions doubles et triple est un artefact du choix de l'échelle pour mesurer la variable de réponse. Cela est plutôt exceptionnel.**
- **C'est toujours une bonne idée de faire un examen graphique des données préalablement à l'analyse. Les graphiques d'interaction sont particulièrment révélateurs des effets.**

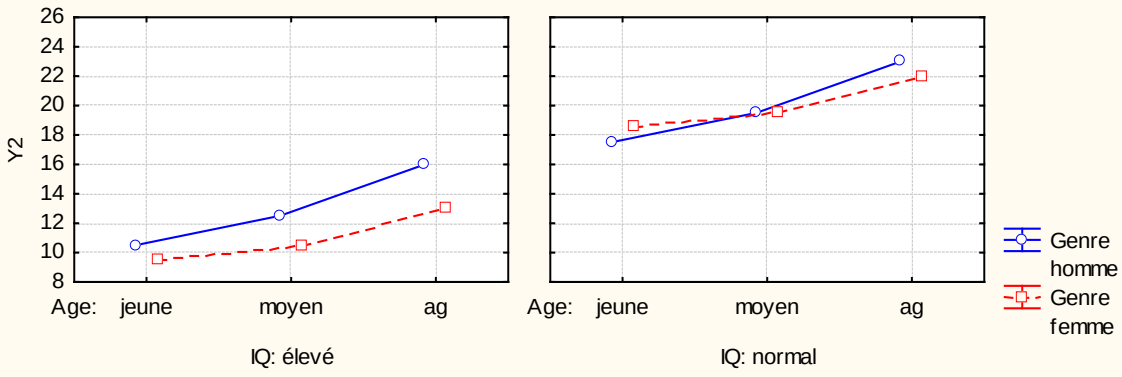
Example 15 : examen des données et graphiques d'interaction – Y1, Y2, Y3

Y1



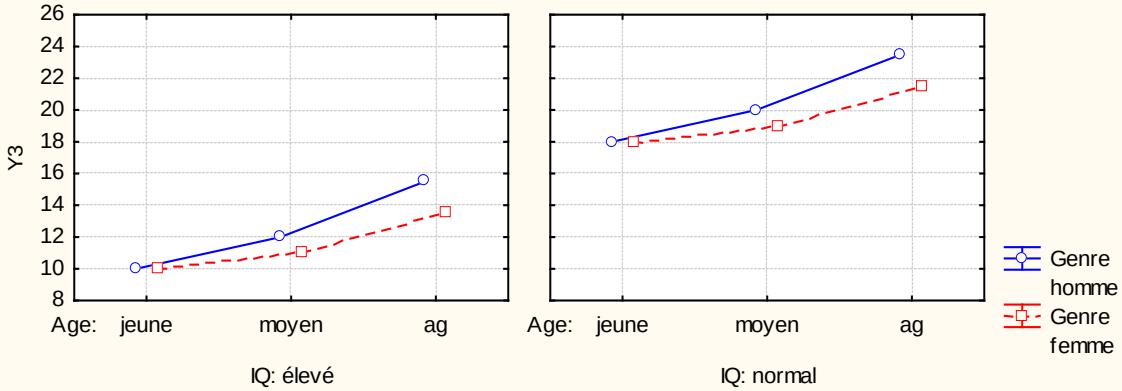
interaction triple : oui

Y2



interaction triple : non
interaction double : oui

Y3

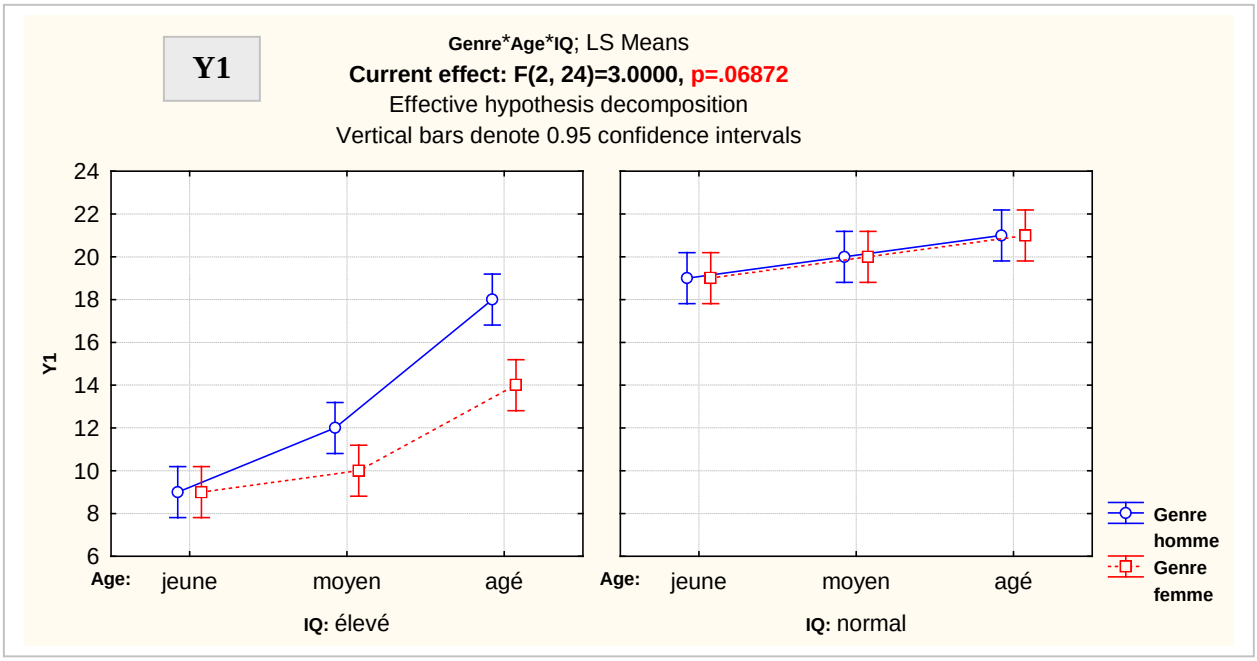


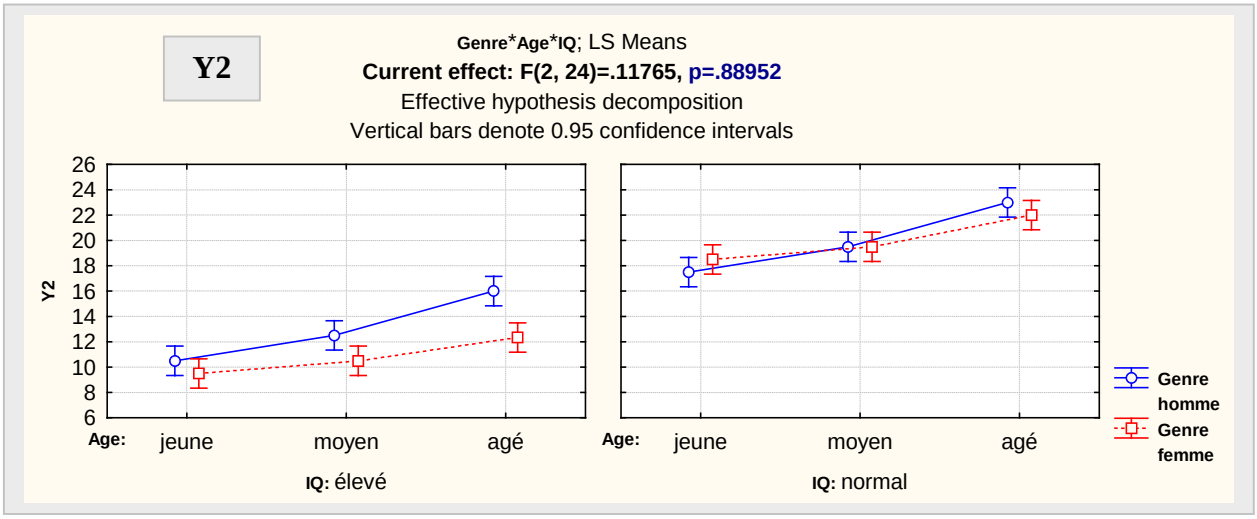
interaction triple: non
interaction double: non

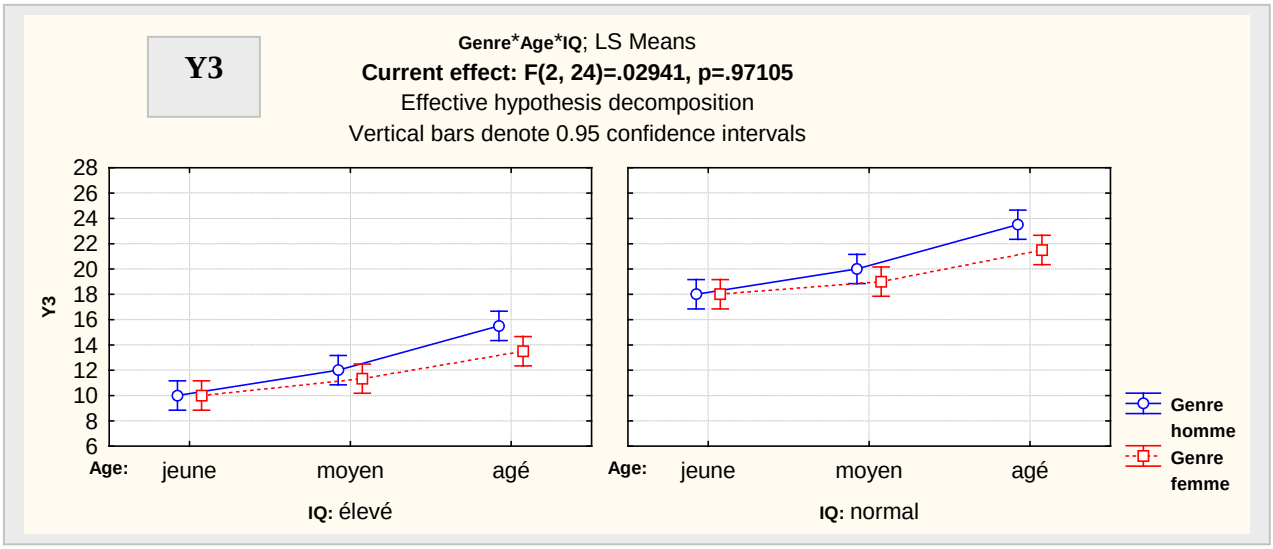
<u>Exemple</u> 16 : influence de 3 facteurs sur le temps d'apprentissage Y :
données avec 2 répétitions sur 2 autres groupes d'individus
3 exemples de variable de réponse : Y1, Y2, Y3
facteur 1 : genre (homme, femme)
facteur 2 : age (jeune, moyen, agé)
facteur 3 : IQ (élevé, normal)

Kutner et all 5 ed. p 994, 999, 1001 - rep = 1 représentent les données d'origine
rep = 2 et rep = 3 sont des données ajoutées
rep = 2 : Y(rep=2) = Y(rep=1) - 1 et rep=3 : Y(rep=3) = Y(rep=1) + 1

	rep	Genre	Age	IQ	Y1	Y2	Y3
1	1	homme	jeune	élevé	9	10.5	10.0
2	1	homme	moyen	élevé	12	12.5	12.0
3	1	homme	agé	élevé	18	16.0	15.5
4	1	homme	jeune	normal	19	17.5	18.0
5	1	homme	moyen	normal	20	19.5	20.0
6	1	homme	agé	normal	21	23.0	23.5
7	1	femme	jeune	élevé	9	9.5	10.0
8	1	femme	moyen	élevé	10	10.5	11.0
9	1	femme	agé	élevé	14	13.0	13.5
10	1	femme	jeune	normal	19	18.5	18.0
11	1	femme	moyen	normal	20	19.5	19.0
12	1	femme	agé	normal	21	22.0	21.5
13	2	homme	jeune	élevé	8	9.5	9.0
14	2	homme	moyen	élevé	11	11.5	11.0
15	2	homme	agé	élevé	17	15.0	14.5
16	2	homme	jeune	normal	18	16.5	17.0
17	2	homme	moyen	normal	19	18.5	19.0
18	2	homme	agé	normal	20	22.0	22.5
19	2	femme	jeune	élevé	8	8.5	9.0
20	2	femme	moyen	élevé	9	9.5	11.0
21	2	femme	agé	élevé	13	12.0	12.5
22	2	femme	jeune	normal	18	17.5	17.0
23	2	femme	moyen	normal	19	18.5	18.0
24	2	femme	agé	normal	20	21.0	20.5
25	3	homme	jeune	élevé	10	11.5	11.0
26	3	homme	moyen	élevé	13	13.5	13.0
27	3	homme	agé	élevé	19	17.0	16.5
28	3	homme	jeune	normal	20	18.5	19.0
29	3	homme	moyen	normal	21	20.5	21.0
30	3	homme	agé	normal	22	24.0	24.5
31	3	femme	jeune	élevé	10	10.5	11.0
32	3	femme	moyen	élevé	11	11.5	12.0
33	3	femme	agé	élevé	15	12.0	14.5
34	3	femme	jeune	normal	20	19.5	19.0
35	3	femme	moyen	normal	21	20.5	20.0
36	3	femme	agé	normal	22	23.0	22.5







Ajustement du modèle : tableau des estimateurs des paramètres

paramètre	estimateur
μ_{ijk}	$\overline{Y}_{ijk.} = (1/n) \sum_{l} Y_{ijkl}$
$\mu_{ij.}$	$\overline{Y}_{ij..} = (1/nc) \sum_{k} \sum_{l} Y_{ijkl}$
$\mu_{i.k}$	$\overline{Y}_{i.k.} = (1/nb) \sum_{j} \sum_{l} Y_{ijkl}$
$\mu_{.jk}$	$\overline{Y}_{.jk.} = (1/na) \sum_{i} \sum_{l} Y_{ijkl}$
$\mu_{i..}$	$\overline{Y}_{i...} = (1/nbc) \sum_{j} \sum_{k} \sum_{l} Y_{ijkl}$
$\mu_{.j.}$	$\overline{Y}_{.j..} = (1/nac) \sum_{i} \sum_{k} \sum_{l} Y_{ijkl}$
$\mu_{..k}$	$\overline{Y}_{..k.} = (1/nab) \sum_{i} \sum_{j} \sum_{l} Y_{ijkl}$
μ	$\overline{Y}_{....} = (1/nabc) \sum_{i} \sum_{j} \sum_{k} \sum_{l} Y_{ijkl}$
α_i	$\overline{Y}_{i...} - \overline{Y}_{....}$
β_j	$\overline{Y}_{.j..} - \overline{Y}_{....}$
γ_k	$\overline{Y}_{..k.} - \overline{Y}_{....}$
$(\alpha\beta)_{ij}$	$\overline{Y}_{ij..} - \overline{Y}_{i...} - \overline{Y}_{.j..} + \overline{Y}_{....}$
$(\alpha\gamma)_{ik}$	$\overline{Y}_{i.k.} - \overline{Y}_{i...} - \overline{Y}_{..k.} + \overline{Y}_{....}$
$(\beta\gamma)_{jk}$	$\overline{Y}_{.jk.} - \overline{Y}_{.j..} - \overline{Y}_{..k.} + \overline{Y}_{....}$
$(\alpha\beta\gamma)_{ijk}$	$\overline{Y}_{ijk.} - \overline{Y}_{ij..} - \overline{Y}_{i.k.} - \overline{Y}_{.jk.} + \overline{Y}_{i...} + \overline{Y}_{.j..} + \overline{Y}_{..k.} - \overline{Y}_{....}$

$$\widehat{Y}_{ijkl} = \overline{Y}_{ijk.} \qquad \textbf{(145}$$

$$\textbf{résidu } e_{ijkl} = Y_{ijkl} - \widehat{Y}_{ijkl} \qquad \textbf{(146)}$$

ANOVA

Source	SS	Df	MS	F	p-value
A	SSA	a – 1	MSA	MSA / MSE	
B	SSB	b – 1	MSB	MSB / MSE	
C	SSC	c – 1	MSC	MSC /MSE	
AB	SSAB	(a-1)(b-1)	MSAB	MSAB / MSE	
AC	SSAC	(a-1)(c -1)	MSAC	MSAC / MSE	
BC	SSBC	(b-1)(c-1)	MSBC	MSCB / MSE	
ABC	SSABC	(a-1)(b-1)(c-1)	MSABC	MSABC / MSE	
erreur	SSE	(n-1)abc	MSE	----------	
totale	SSTO	abcn - 1	----------	----------	

$$SSA = nbc \sum (\overline{Y}_{i\ldots} - \overline{Y}_{\ldots\ldots})^2 \quad (147a)$$

$$SSB = nac \sum (\overline{Y}_{.j..} - \overline{Y}_{\ldots\ldots})^2 \quad (147b)$$

$$SSC = nab \sum (\overline{Y}_{..k.} - \overline{Y}_{\ldots\ldots})^2 \quad (147c)$$

$$SSAB = nc \sum\sum (\overline{Y}_{ij..} - \overline{Y}_{i\ldots} - \overline{Y}_{.j..} + \overline{Y}_{\ldots\ldots})^2 \quad (148ab)$$

$$SSAC = nb \sum\sum (\overline{Y}_{i.k.} - \overline{Y}_{i\ldots} - \overline{Y}_{..k.} + \overline{Y}_{\ldots\ldots})^2 \quad (148ac)$$

$$SSBC = na \sum\sum (\overline{Y}_{.jk.} - \overline{Y}_{.j..} - \overline{Y}_{..k.} + \overline{Y}_{\ldots\ldots})^2 \quad (148bc)$$

$$SSABC = n \sum\sum\sum (\overline{Y}_{ijk.} - \overline{Y}_{ij..} - \overline{Y}_{i.k.} - \overline{Y}_{.jk.} + \overline{Y}_{i\ldots} + \overline{Y}_{.j..} + \overline{Y}_{..k.} - \overline{Y}_{\ldots\ldots})^2 \quad (149)$$

$$SSE = \sum\sum\sum\sum (Y_{ijkl} - \overline{Y}_{.ijk.})^2 \quad (150)$$

$$SSTO = \sum\sum\sum\sum (Y_{ijkl} - \overline{Y}_{\ldots\ldots})^2 \quad (151)$$

Tests d'hypothèses

$H_0 : (\alpha\beta\gamma)_{ijk} = 0$ vs H_a : pas tous les $(\alpha\beta\gamma)_{ijk}$ sont nuls (152)

On rejette H_0 si $MSABC / MSE > F(1 - \alpha ; (a-1)(b-1)(c-1) , (n-1)abc)$

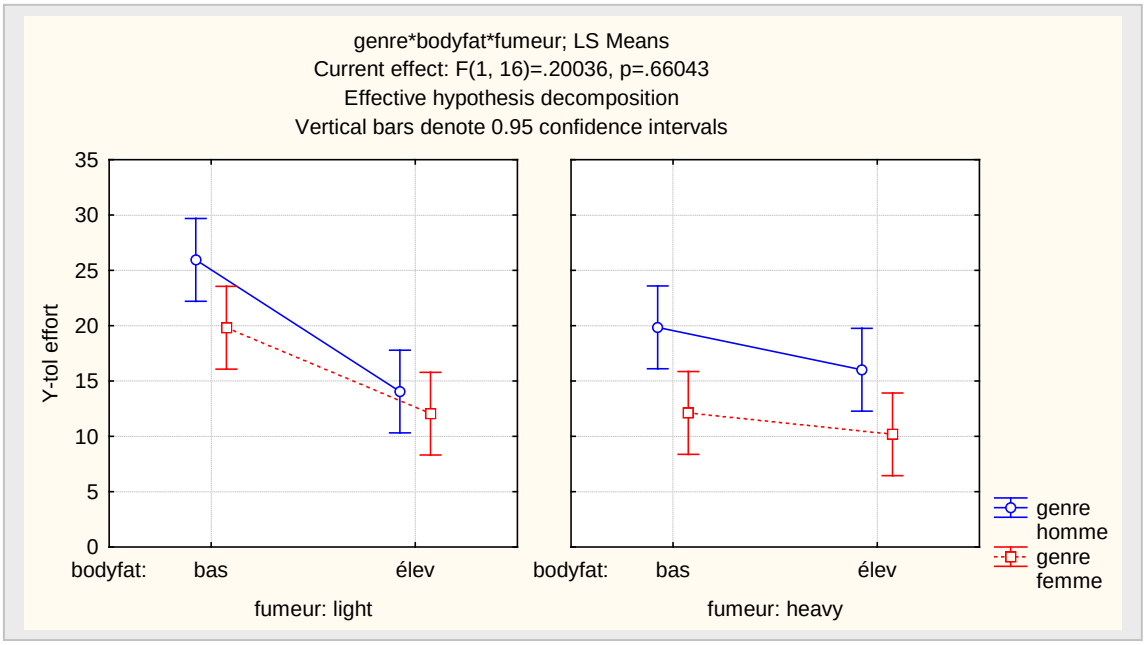
On peut aussi tester les hypothèses sur la nullité des intractions doubles avec les ratios F du tableau d'analyse de la variance.

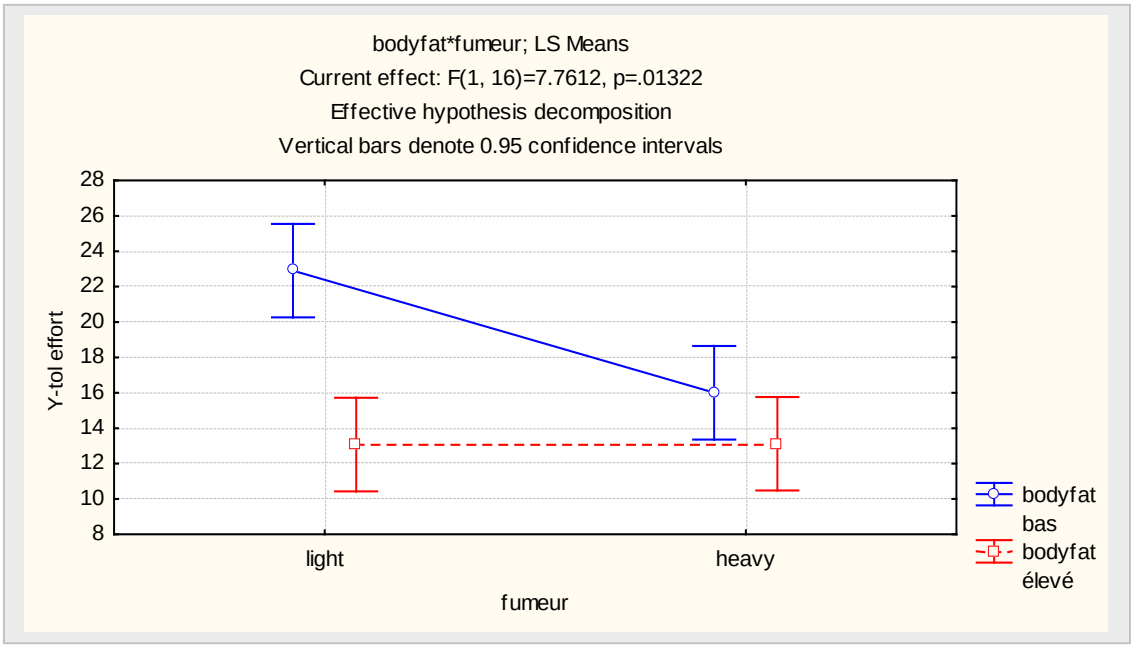
Exemple 17 : effet de 3 facteurs sur la tolérance à l'exercice

Kutner et all 5 ed. p 1004					
	genre	bodyfat	fumeur	rep	Y-tol effort
1	homme	bas	light	1	24.1
2	homme	bas	light	2	29.2
3	homme	bas	light	3	24.6
4	femme	bas	light	1	20.0
5	femme	bas	light	2	21.9
6	femme	bas	light	3	17.6
7	homme	élevé	light	1	14.6
8	homme	élevé	light	2	15.3
9	homme	élevé	light	3	12.3
10	femme	élevé	light	1	16.1
11	femme	élevé	light	2	9.3
12	femme	élevé	light	3	10.8
13	homme	bas	heavy	1	17.6
14	homme	bas	heavy	2	18.8
15	homme	bas	heavy	3	23.2
16	femme	bas	heavy	1	14.8
17	femme	bas	heavy	2	10.3
18	femme	bas	heavy	3	11.3
19	homme	élevé	heavy	1	14.9
20	homme	élevé	heavy	2	20.4
21	homme	élevé	heavy	3	12.8
22	femme	élevé	heavy	1	10.1
23	femme	élevé	heavy	2	14.4
24	femme	élevé	heavy	3	6.1

ANOVA

	DF	SS	MS	F	p-value
Intercept	1	6353.76	6353.76	680.61	0.0000
genre	1	176.58	176.58	18.92	0.0005
bodyfat	1	242.57	242.57	25.98	0.0001
fumeur	1	70.38	70.38	7.54	0.0144
genre*bodyfat	1	13.65	13.65	1.46	0.2441
genre*fumeur	1	11.07	11.07	1.19	0.2923
bodyfat*fumeur	1	72.45	72.45	7.76	0.0132
genre*bodyfat*fumeur	1	1.87	1.87	0.20	0.6604
Error	16	149.37	9.34		
Total	23	737.95			





Stratégie et modélisation

M1 : modèle complet avec interaction doubles et interaction triple - équation (141)
M2 : modèle réduit avec 3 interactions doubles mais sans interaction triple
M3 : modèle réduit avec 1 ou 2 intractions double
M4 : modele avec effets principaux seulement

INTERACTION TRIPLE SIGNIFICATIVE ET IMPORTANTE?

OUI : la variable de réponse doit être analysée avec le modèle à moyenne de cellules : M1

NON : INTERACTIONS DOUBLES SIGNIFICATIVES ET IMPORTANTES ?

OUI : la variable de réponse doit être analysée avec le modèle réduit (sans le terme d'interaction triple) de moyennes à cellules : M2 ou M3

NON : la variable de réponse peut être interprétée avec le modèle à effets principaux seulement : M4

le modèle final retenu devrait respecter le principe de la hiérarchie
si une interaction est présente (car significative et importante) il est recommandé de mettre dans l'équation de prédiction les effets principaux correspondants même s'il ne sont pas tous significatifs.

Exemple 17 : suite

ANOVA

	SS	DF	MS	F	p
Intercept	6353.76	1	6353.76	714.20	0.0000
genre	176.58	1	176.58	19.85	0.0003
bodyfat	242.57	1	242.57	27.27	0.0001
fumeur	70.38	1	70.38	7.91	0.0120
genre*bodyfat	13.65	1	13.65	1.53	0.2323
genre*fumeur	11.07	1	11.07	1.24	0.2801
bodyfat*fumeur	72.45	1	72.45	8.14	0.0110
Error	151.24	17	8.90		

Modèle final : Y = gen + genre + bodyfat + fumeur + bodyfat*fumeur

Les procédures de Tukey, Scheffé et Bonferronni s'appliquent pour tester des contrastes pour les effets. Consulter Kutner et all 5 ed. pp. 1014-1019

Quoi faire si les tailles n_{ijk} sont (très) inégales mais toutes > 0?

Réponse : employer l'approche par régression avec des variables indicatrices

Exemple 17 : données de l'exemple 16 avec des données non équilibrées Y2 (fichier stress)

Kutner et all 5 ed. p 1004

id	genre	bodyfat	fumeur	rep	Y tol effort	Y2	v7	Xg	Xb	Xf	XgXb	XgXf	XbXf	XgXbXf
1	homme	bas	light	1	24.1	24.1		1	-1	-1	-1	-1	1	1
2	homme	bas	light	2	29.2	29.2		1	-1	-1	-1	-1	1	1
3	homme	bas	light	3	24.6			1	-1	-1	-1	-1	1	1
4	femme	bas	light	1	20.0	20.0		-1	-1	-1	1	1	1	-1
5	femme	bas	light	2	21.9			-1	-1	-1	1	1	1	-1
6	femme	bas	light	3	17.6			-1	-1	-1	1	1	1	-1
7	homme	élevé	light	1	14.6	14.6		1	1	-1	1	-1	-1	-1
8	homme	élevé	light	2	15.3	15.3		1	1	-1	1	-1	-1	-1
9	homme	élevé	light	3	12.3	12.3		1	1	-1	1	-1	-1	-1
10	femme	élevé	light	1	16.1	16.1		-1	1	-1	-1	1	-1	1
11	femme	élevé	light	2	9.3	9.3		-1	1	-1	-1	1	-1	1
12	femme	élevé	light	3	10.8			-1	1	-1	-1	1	-1	1
13	homme	bas	heavy	1	17.6	17.6		1	-1	1	-1	1	-1	-1
14	homme	bas	heavy	2	18.8			1	-1	1	-1	1	-1	-1
15	homme	bas	heavy	3	23.2			1	-1	1	-1	1	-1	-1
16	femme	bas	heavy	1	14.8	14.8		-1	-1	1	1	-1	-1	1
17	femme	bas	heavy	2	10.3	10.3		-1	-1	1	1	-1	-1	1
18	femme	bas	heavy	3	11.3			-1	-1	1	1	-1	-1	1
19	homme	élevé	heavy	1	14.9	14.9		1	1	1	1	1	1	1
20	homme	élevé	heavy	2	20.4	20.4		1	1	1	1	1	1	1
21	homme	élevé	heavy	3	12.8	12.8		1	1	1	1	1	1	1
22	femme	élevé	heavy	1	10.1	10.1		-1	1	1	-1	-1	1	-1
23	femme	élevé	heavy	2	14.4	14.4		-1	1	1	-1	-1	1	-1
24	femme	élevé	heavy	3	6.1			-1	1	1	-1	-1	1	-1

Regression Summary for Dependent Variable: Y2
R= 0.885 R^2= 0.783 Adjusted R^2= 0.594 F(7,8)=4.1382 p

	Beta	Std.Err.	B	Std.Err.	t(8)	p-level
Intercept			16.4813	0.91187	18.0741	0.00000
Xg	0.40712	0.17626	2.1063	0.91187	2.3098	0.04970
Xb	-0.51285	0.17201	- 2.7188	0.91187	-2.9815	0.01756
Xf	-0.36488	0.17765	- 1.8729	0.91187	-2.0539	0.07406
XgXb	-0.15826	0.17626	- 0.8188	0.91187	-0.8979	0.39547
XgXf	0.01973	0.17626	0.1021	0.91187	0.1119	0.91362
XbXf	0.43875	0.17765	2.2521	0.91187	2.4697	0.03873
XgXbXf	0.09705	0.17626	0.5021	0.91187	0.5506	0.59694

Conclusion : l'interprétation est analogue à celle basée sur le modèle retenu avec toutes les données.

13. Modélisation avec quatre facteurs et plus

On peut supposer que dans ces circonstances on est en présence de données observationnelles historiques et non pas de données résultant d'une expérimentation planifiée dont le but est de tester l'influence de quelques facteurs contrôlés. En présence de nombreux facteurs, l'objectif de l'étude est plutôt de nature exploratoire et on cherche à identifier un ou plusieurs « bons » modèles pour représenter et synthétiser l'ensemble des données. Cette recherche de modèles est aussi connue sous le nom de « méthodes de sélection de variables » et elles ont été abondamment développées dans la littérature statistique.

Problème de la multicollinéarité

Avec de nombreux facteurs (prédicteurs ou variables explicatives) il y a une possibilité très réelle que plusieurs d'entre elles présentent des corrélations assez élevées ; cela conduit à de mauvaises estimations des coefficients du modèle. Le problème se présente lorsque le modèle de régression multiple est significatif mais aucun des coefficients du modèle n'est significatif. On remarque alors une grande différence entre le R^2 et le R^2 ajusté.

Exemple 18 : huit facteurs affectant le prix de voitures (Y)
référence : http://cedric.cnam.fr/~saporta/multicol.pdf

label	NOM	CYL	PUIS	LON	LAR	POIDS	VITESSE	NAT	FINITION	Y-PRIX
ALFAS	ALFASUD-TI-1350	1350	79	393	161	870	165	I	B	30570
AUDI	AUDI-100-L	1588	85	468	177	1110	160	D	TB	39990
SIMCA	SIMCA-1307-GLS	1294	68	424	168	1050	152	F	M	29600
CITRO	CITROEN-GS-CLUB	1222	59	412	161	930	151	F	M	28250
FIAT	FIAT-132-1600GLS	1585	98	439	164	1105	165	I	B	34900
LAN	LANCIA-BETA-1300	1297	82	429	169	1080	160	I	TB	35480
PEU	PEUGEOT-504	1796	79	449	169	1160	154	F	B	32300
REN16	RENAULT-16-TL	1565	55	424	163	1010	140	F	B	32000
REN30	RENAULT-30-TS	2664	128	452	173	1320	180	F	TB	47700
TOY	TOYOTA-COROLLA	1166	55	399	157	815	140	J	M	26540
ALFET	ALFETTA-1.66	1570	109	428	162	1060	175	I	TB	42395
PRIN	PRINCESS-1800-HL	1798	82	445	172	1160	158	GB	B	33990
DAT	DATSUN-200L	1998	115	469	169	1370	160	J	TB	43980
TAU	TAUNUS-2000-GL	1993	98	438	170	1080	167	D	B	35010
RAN	RANCHO	1442	80	431	166	1129	144	F	TB	39450
MAZ	MAZDA-9295	1769	83	440	165	1095	165	J	M	27900
OPEL	OPEL-REKORD-L	1979	100	459	173	1120	173	D	B	32700
LADA	LADA-1300	1294	68	404	161	955	140	U	M	22100

Analysis of Variance; DV: PRIX

	SS	df	MS	F	p-level
Regress.	520591932	6	86765322	4.469	0.015603
Residual	213563858	11	19414896		
Total	734155790				

Regression Summary for Dependent Variable: Y-PRIX (Prix voitures.sta) R= .84208242 R^2= 0.70910281 Adjusted R^2= 0.55043161 F(6,11)=4.469 p= 0.0156						
	Beta	Std.Err.	B	Std.Err.	t(11)	p-level
Intercept			-8239.36	42718.42	-0.192876	0.850571
CYL	-0.199449	0.315835	-3.51	5.55	-0.631497	0.540617
PUIS	0.874912	0.542254	282.17	174.88	1.613472	0.134938
LON	-0.050588	0.436482	-15.04	129.75	-0.115899	0.909821
LAR	0.168747	0.333176	208.69	412.05	0.506479	0.622518
POIDS	0.262068	0.513149	12.57	24.62	0.510705	0.619651
VITESSE	-0.205272	0.410598	-111.11	222.26	-0.499934	0.626971

Globalement la régresssion est significative (0.0156) mais aucun coefficient du modèle est significatif !

On peut détecter la multicolinéarité avec

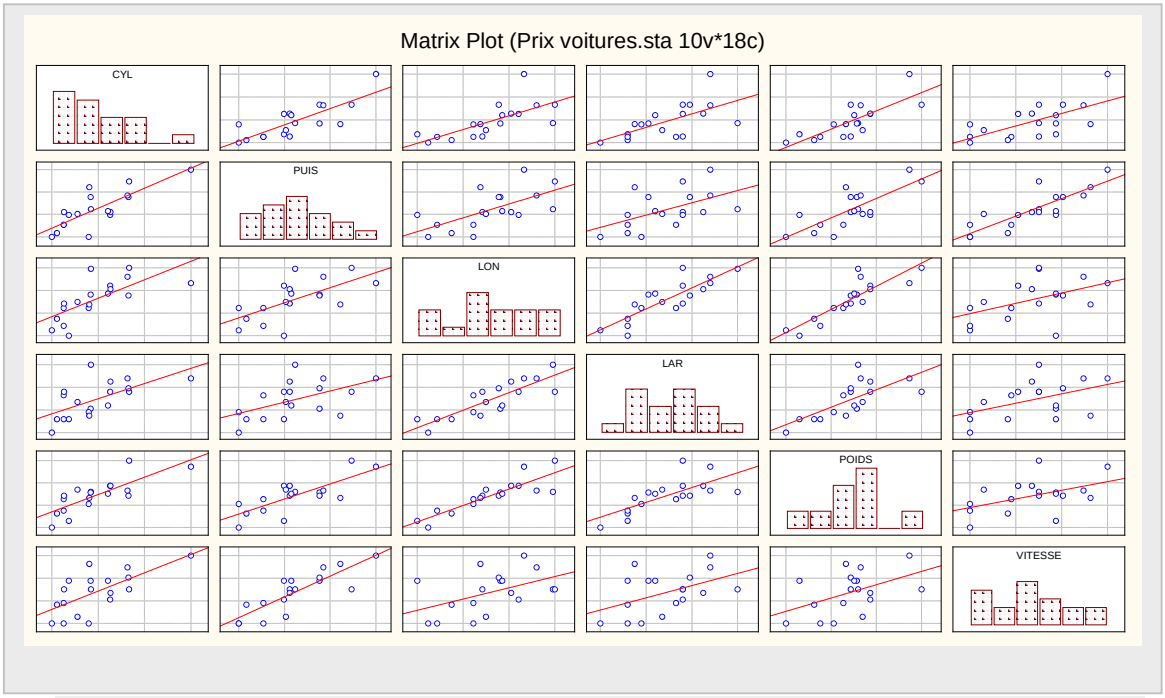
- **l'étude de la matrice des corrélations**
- **le calclul des facteurs d'inflation de la variance VIF**

$$\text{VIF(variable j)} = 1 / (1 - R^2 \text{ (variable j | autres variables)}) \quad (153)$$

R^2 : corrélation multiple de la variable j sur les autres variables

correlations						
	CYL	PUIS	LON	LAR	POIDS	VITESSE
CYL	1.00	0.80	0.70	0.63	0.79	0.66
PUIS	0.80	1.00	0.64	0.52	0.77	0.84
LON	0.70	0.64	1.00	0.85	0.87	0.48
LAR	0.63	0.52	0.85	1.00	0.72	0.47
POIDS	0.79	0.77	0.87	0.72	1.00	0.48
VITESSE	0.66	0.84	0.48	0.47	0.48	1.00

On remarque une corrélation assez élevée (> 0.70) entre plusieurs paires de variables.



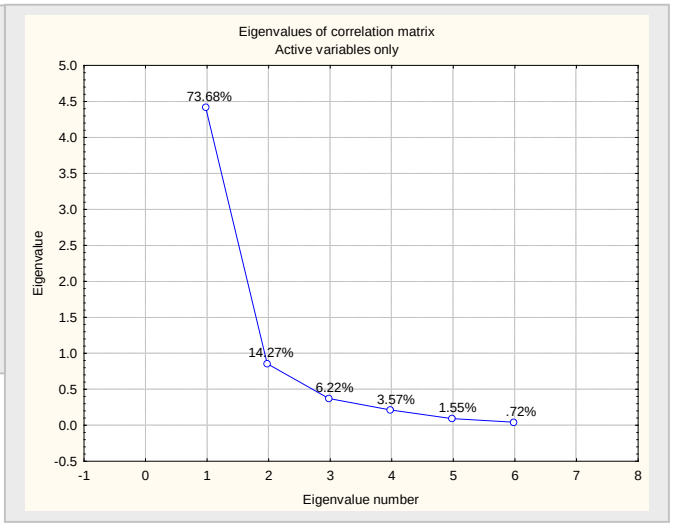
Redundancy of Independent Variables; DV: PRIX
R-square column contains R-square of respective variable with all other independent variables

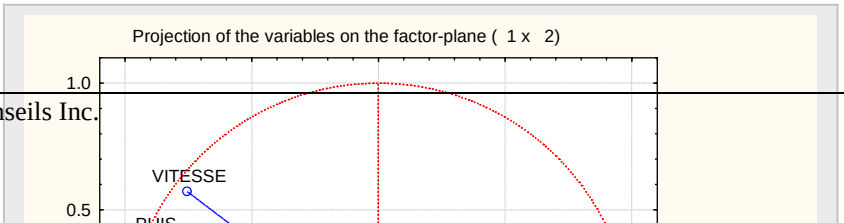
	Toleran.	R-square	Partial	Semipart
CYL	0.27	0.73	-0.19	-0.10
PUIS	0.09	0.91	0.44	0.26
LON	0.14	0.86	-0.03	-0.02
LAR	0.24	0.76	0.15	0.08
POIDS	0.10	0.90	0.15	0.08
VITESSE	0.16	0.84	-0.15	-0.08

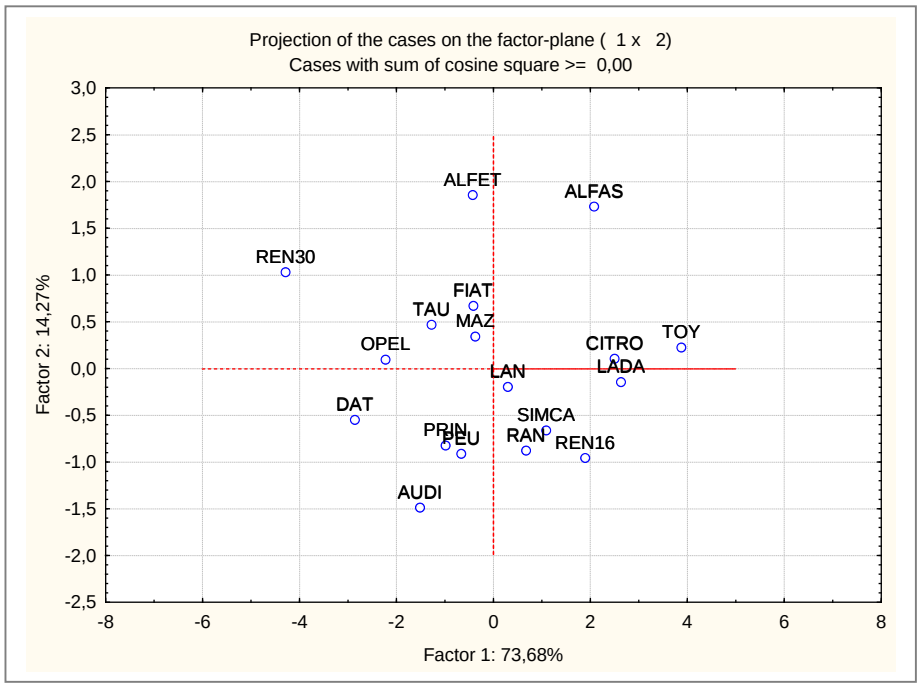
Une analyse en composante principale (ACP) montrent que ces 6 variables sont très bien expliquées par les 2 premières composantes principales avec un pourcentage de 88%.

Eigenvalues of correlation matrix, and related statistics
Active variables only

	Eigenvalue	% Total	Cumulative	Cumulative
1	4.421	73.68	4.42	73.7
2	0.856	14.27	5.28	87.9
3	0.373	6.22	5.65	94.2
4	0.214	3.57	5.86	97.7
5	0.093	1.55	5.96	99.3
6	0.043	0.72	6.00	100.0







Critères pour comparer des modèles avec k effets

- **maximiser R^2 ajusté** $= 1 - [(n-1)/(n-k)](1-R^2)$ **(154)**

  $= 1 - (\hat{\sigma}^2 / s^2_y)$ **équivalent à minimiser** $\hat{\sigma}^2$

- **viser un Cp de Mallows proche de k + 1**

  $$Cp = (SSE / \hat{\sigma}^2) + 2k + 2 - n \quad \textbf{(155)}$$

- **Maximiser AIC (Akaike Information Criterion)**

  $$AIC = -2\ln(L) + 2(k+1) = n \ln(SSE/n) + 2(k+1) + n(l n\pi + 1) \quad \textbf{(156)}$$

  L : fonction de vraisemblance

Algorithmes de sélection

- **dénombrement exhaustif : $2^p - 1$ modèles**

p	6	10	15	20	30	50
$2^p - 1$	63	1023	32767	1 048 575	10^{10}	10^{15}

- **méthodes pas à pas (« stepwise »)**
  ascendante (« forward »), descendante (« backward »), …

Exemple 18 : suite

Regression Summary for Dependent Variable: PRIX
R= .7987 R^2= .6379 Adjusted R^2= .6153 F(1,16)=28.189
backward

	Beta	Std.Err.	B	Std.Err.	t(16)	p-level
Intercept			12364	4215.9	2.9326	0.0098
PUIS	0.7987	0.1504	258	48.5	5.3094	0.0001

Regression Summary for Dependent Variable: Y-PRIX
R= .8286 R^2= .6866 Adjusted R^2= .6448 F(2,15)=16.433
forward

	Beta	Std.Err.	B	Std.Err.	t(15)	p-level
Intercept			1775.6	8031.0	0.2211	0.8280
PUIS	0.5363	0.2246	173.0	72.4	2.3884	0.0305
POIDS	0.3429	0.2246	16.5	10.8	1.5269	0.1476

Autres stratégies de modélisation avec toutes les variables

- **régression sur les composantes principales :**
  prédiction appochée en éliminant les petites valeurs propres
  les composantes principales font l'extraction du maximum de variance des prédicteurs sans tenir compte de Y ;

- **régression PLS (Partial Least Square) (H. S. Wold)**
  semblable à la régression sur composantes principales mais les composantes PLS sont optimisées pour être prédictives de Y ;
  la régression PLS donne en pratique de bonnes prévisions même si le nombre d'observations n est petit et le nombre de variables p est grand

- **régression de type ridge (Hoerl et Kennard)**

- **Méthodes du DATA MINING.**

Régression sur variables qualitatives

On recode les modalités en variables indicatrices.
Puisque la somme des indcatrices vaut 1, on élimine une modalité.
On peut introduire des variables d'interaction avec le produit des variables indicatrices associées aux différentes variables qualitatives.

14. Design (plan) en mesures répétées

Le *design en mesures répétées* utilise le même sujet ou la même unité statistique à plusieurs reprises. Celle-ci peut être une personne, un animal, un morceau de terrain, des plantes etc. Les sujets servent de blocs et les unités expérimentales à l'intérieur d'un bloc peuvent être vues comme différentes occasions lorsque l'on applique un traitement à un sujet. Les designs en mesures répétées sont beaucoup employés dans les sciences du comportement et dans les sciences du vivant. Parmi les designs très employés, il y a le design « crossover » et le design à parcelles divisées (« split-plot »)

Exemple : on étudie l'impact de 2 campagnes publicitaires dans 15 régions géographiques. Elles constituent les sujets. Dans chaque région, on randomise l'ordre de la capagne publicaitaire. On attend une certaine période de temps (« washout period ») entre les deux campagnes.

Exemple : des individus (sujets) souffrant de migraines recoivent 2 traitements: un nouveau médicament (A) et un traitement placébo (P). L'ordre dans lequel ils recoivent les 2 traitements est randomisé. La moitié des sujets recoivent les traitements dans l'ordre AP tandis que l'autre moitié ils recoivent les traitments dans l'ordre PA.

Le design est identifié sous le vocable de *design en mesures répétées* car le même sujet reçoit plusieurs traitements et il est mesuré pour chaque traitement. Il ne faut pas confondre le concept de mesures répétées avec le concept de répétition. Dans ce dernier cas le sujet est mesuré plusieurs fois pour le même traitement, tandis que dans un design à mesures répétées, le même sujet est mesuré dans différentes conditions expérimentales.

Avantages du design en mesures répétées

- **La variabilité inter sujet est exclue de l'erreur expérimentale, donc il est plus facile de comparer les traitements entre eux.**
- **Chaque sujet sert comme son propre contrôle.**
- **Économie du nombre de sujets.**

Désavantages du design en mesures répétées

- **Il exige une période d'attente entre les traitements à cause des phénomènes d'accoutumance (posologie en medecine), d'apprentissage (tests sur des humains), ou d'accumulation (traitements chimique en agriculture).**

- **Effets d'interférence :**
  - **effet de l'ordre**
  - **effet de passage entre deux traitements consécutifs (« carryover »).**

- **Les variables de réponses sont dépendantes.**

- **Pour contrer les effets d'interférence: on randomise, pour chaque sujet, l'ordre d'assignation des traitements.**

Tableau général des données d'un design à mesures répétées

unité ou sujet	temps (époque)	indicatrice donnée manquante	réponse Y	covariables continues X W….	facteurs catégoriques A B…..
1	1	δ_{11}	Y_{11}	X_{11} W_{11} ...	A_{11} B_{11}….
	2	δ_{12}	Y_{12}	X_{12} W_{12} ...	A_{12} B_{12}….
	….	….	….	….	….
	j	δ_{1j}	Y_{1j}	X_{1j} W_{1j} ...	A_{1j} B_{1j}….
	….	….	….	….	….
	t_1	$\delta_{1\,t1}$	Y_{1t1}	$X_{1\,t1}$ $W_{1\,t1}$ ...	$A_{1\,t1}$ $B_{1\,t1}$…
…….	…….	……..	…….	……..	…….
i	1	δ_{i1}	Y_{i1}	X_{i1} W_{i1} ...	A_{i1} B_{i1}….
	….	….	….	….	….
	j	$\delta_{i\,j}$	$Y_{i\,j}$	$X_{i\,j}$ $W_{i\,j}$ ...	$A_{i\,j}$ $B_{i\,j}$….
	….	….	….	….	….
	t_i	$\delta_{i\,ti}$	$Y_{i\,ti}$	$X_{i\,ti}$ $W_{i\,ti}$ ...	$A_{i\,ti}$ $B_{i\,ti}$….
…..	…….	……..	…….	…….	…….
n	1	δ_{n1}	$Y_{n\,t1}$	X_{n1} W_{n1} ...	A_{n1} B_{n1}….
	….	….	….	….	….
	j	δ_{nj}	$Y_{n\,tj}$	$X_{n\,tj}$ $W_{n\,tj}$ ...	$A_{n\,tj}$ $B_{n\,tj}$…..
	….	….	….	….	….
	t_n	$\delta_{n\,tn}$	$Y_{n\,tn}$	X_{tn} $W_{n\,tn}$ ...	$A_{n\,tn}$ $B_{n\,tn}$….

Le temps (observations en occasions multiples) peut être remplacé par des observations sous plusieurs conditions; par exemple l'espace.

$$\delta_{i\,j} = \begin{cases} 1 & \text{si toutes les données de Y, X, W, …, A, B, … sont disponibles} \\ 0 & \text{autrement (données manquantes)} \end{cases}$$

observations dépendantes

rôle : sert à classer les données en sous groupes

2 approches pour faire l'analyse statistique :

- **<u>approche unidimensionnelle</u> : les données répétées de la réponse de chaque sujet sont résumées une seule nouvelle variable de réponse par une transformation appropriée: par exemple une pente de moindres carrés sur les données (t,Y) ;**
- **<u>approche multidimensionnelle</u> : les données répétées sont considérées comme plusieurs variables de réponse dépendantes; dans ce cas il est préférable d'organiser les données avec autant de variables de réponse Y qu'il y a de mesures répétées. Les exemples qui suivent montrent l'organisation des fichiers de données pour faire l'analyse.**

Exemple 19 : Y = évaluation (sur 40) de 4 bouteilles de vin par 6 juges A, B, C, D, E, F
Elle est transformée en 4 variables de réponse dépendantes
Y-vin1, Y-vin2, Y-vin3 Y-vin4
facteur 1 = juge (6 modalités) facteur 2 = vin (4 modalités)

Kutner et all 5 ed. p. 1132 – fichier WINE								
id	juge	vin id	Y-rang	new	Y-vin1	Y-vin2	Y-vin3	Y-vin4
1	A	vin1	20		20	24	28	28
2	A	vin2	24					
3	A	vin3	28					
4	A	vin4	28					
5	B	vin1	15		15	18	23	24
6	B	vin2	18					
7	B	vin3	23					
8	B	vin4	24					
9	C	vin1	18		18	19	24	23
10	C	vin2	19					
11	C	vin3	24					
12	C	vin4	23					
13	D	vin1	26		26	26	30	30
14	D	vin2	26					
15	D	vin3	30					
16	D	vin4	30					
17	E	vin1	22		22	24	28	26
18	E	vin2	24					
19	E	vin3	28					
20	E	vin4	26					
21	F	vin1	19		19	21	27	25
22	F	vin2	21					
23	F	vin3	27					
24	F	vin4	25					

Exemple 20 : poids de 16 rats de laboratoire jour 1, 8, 15, …, 64 - 3 diètes (A, B, C)

Crowder and Hand 1990 p.19 - Analysis of Repeated Measures
expérience sur 16 rats - 3 diètes - Y-poids en grammes à jour = 1, 8, .., 64
unités statistiques = 16 rats de laboatoire
chaque unité est mesurée à 11 reprises
facteur 1 = diète (3 modalités) = A – B - C
facteur 2 = jour (11 modalités) enfoui dans la réponse Y = poids
fichier = BodyWeights

id	diète	sujet	Y-poids jour 1	Y-poids jour 8	Y-poids jour 15	Y-poids jour 22	Y-poids jour 29	Y-poids jour 36	y-poids jour 43	Y-poids jour 44	Y-poids jour 50	Y-poids jour 57	Y-poids jour 64	pente régression
1	A	1	240	250	255	260	262	258	266	266	265	272	278	0.484
2	A	2	225	230	230	232	240	240	243	244	238	247	245	0.330
3	A	3	245	250	250	255	262	265	267	267	264	268	269	0.398
4	A	4	260	255	255	265	265	268	270	272	274	273	275	0.330
5	A	5	255	260	255	270	270	273	274	273	276	278	280	0.406
6	A	6	260	265	270	275	275	277	278	278	284	279	281	0.318
7	A	7	275	275	260	270	273	274	276	271	282	281	284	0.202
8	A	8	245	255	425	268	270	265	265	267	273	274	278	0.409
9	B	9	410	415	425	428	448	443	442	446	456	468	478	1.011
10	B	10	405	420	430	440	448	460	458	464	475	484	496	1.341
11	B	11	445	445	450	452	455	455	451	450	462	466	472	0.363
12	B	12	555	560	565	580	590	597	595	595	612	618	628	1.148
13	C	13	470	465	475	485	487	493	493	504	507	518	525	0.919
14	C	14	535	525	530	533	535	540	525	530	543	544	559	0.315
15	C	15	520	525	530	540	543	546	538	544	553	555	548	0.493
16	C	16	510	510	520	515	530	538	535	542	550	553	569	0.905

Exemple 21 : Y mesure afflux de sang sur 5 parties du corps: bone, brain, skin, muscle, heart
8 unité statistiques = 8 rats de laboratoire ; 4 font de l'exercice et 4 n'en font pas
Facteur 1 = exercice (oui, non)
Facteur 2 = body (5 modalités) enfoui dans la réponse Y blood flow

utner et all 5ed p. 1150 - fichier = BodyParts

id	Exercice	sujet	body	Y-blod flow	new	Y-bf bone	Y-bf brain	Y-bf skin	Y-bf muscle	Y-bf heart
1	non	1	bone	4		4	3	5	5	4
2	non	1	brain	3						
3	non	1	skin	5						
4	non	1	muscle	5						
5	non	1	heart	4						
6	non	2	bone	1		1	3	6	3	8
7	non	2	brain	3						
8	non	2	skin	6						
9	non	2	muscle	3						
10	non	2	heart	8						
11	non	3	bone	3		3	1	4	4	7
12	non	3	brain	1						
13	non	3	skin	4						
14	non	3	muscle	4						
15	non	3	heart	7						
16	non	4	bone	1		1	4	3	2	7
17	non	4	brain	4						
18	non	4	skin	3						
19	non	4	muscle	2						
20	non	4	heart	7						
21	oui	5	bone	3		3	6	12	22	11
22	oui	5	brain	6						
23	oui	5	skin	12						
24	oui	5	muscle	22						
25	oui	5	heart	11						
26	oui	6	bone	3		3	5	8	18	12
27	oui	6	brain	5						
28	oui	6	skin	8						
29	oui	6	muscle	18						
30	oui	6	heart	12						
31	oui	7	bone	4		4	7	10	20	14
32	oui	7	brain	7						
33	oui	7	skin	10						
34	oui	7	muscle	20						
35	oui	7	heart	14						
36	oui	8	bone	2		2	4	7	16	8
37	oui	8	brain	4						
38	oui	8	skin	7						
39	oui	8	muscle	16						
40	oui	8	heart	8						

Expérience en mesures répétées sur tous les traitements d'un facteur fixe

Modèle $Y_{ij} = \mu + \rho_i + \tau_j + \varepsilon_{ij}$ **(157)**

$i = 1, 2, \ldots, s$ et $j = 1, 2, \ldots, n$

où μ : effet général

ρ_i : : effet aléatoire du sujet i - indépendantes et $N(0, \sigma_\rho^2)$

τ_j : effet différentiel de la modalité j du facteur fixe et $\sum \tau_j = 0$

ε_{ij} : erreur aléatoire indépendantes et $N(0, \sigma^2)$

ρ_i , ε_{ij} sont indépendantes

Il s'agit d'un modèle mixte: facteur avec effet fixe + facteur avec effet aléatoire

Conséquences

$$E(Y_{ij}) = \mu + \tau_j \quad (158)$$

$$Var(Y_{ij}) = \sigma_\rho^2 + \sigma^2 \quad (159)$$

$$Cov(Y_{ij}, Y_{ij'}) = \sigma_\rho^2 \qquad j \neq j' \quad (160)$$

$$Cov(Y_{ij}, Y_{i'j}) = 0 \qquad i \neq i' \quad (161)$$

$$Corr(Y_{ij}, Y_{ij'}) = \omega = \sigma_\rho^2 / (\sigma_\rho^2 + \sigma^2) \quad (162)$$

ω : coefficient de corrélation intra classe

hypothèse clé : ω est constant (symmétrie composée)

vérification : test de sphéricité de Mauchley

Analyse de la variance

totale	$SSTOT = \sum\sum (Y_{ij} - \overline{Y}_{..})^2$	(163)
sujet	$SSS = n \sum (\overline{Y}_{i.} - \overline{Y}_{..})^2$	(164)
traitement	$SSTR = s \sum (\overline{Y}_{.j} - \overline{Y}_{..})^2$	(165)
erreur	$SSTR.S = \sum\sum (Y_{ij} - \overline{Y}_{i.} - \overline{Y}_{.j} + \overline{Y}_{..})^2$	(166)
	$SSW = \sum\sum (Y_{ij} - \overline{Y}_{i.})^2$	

Equation de décomposition

$$SSTOT = SSS + \underbrace{SSTR + SSTR.S} \quad (167)$$

totale = inter sujet + intra sujet (SSW)

ANOVA

Source	SS	df	MS	E(MS)
Sujets	SSS	s – 1	MSS	$\sigma^2 + n\sigma_\rho^2$
Traitements	SSTR	n – 1	MSTR	$\sigma^2 + (s/(n-1)) \sum \tau_j^2$
Erreur	SSTR.S	(n - 1)(s - 1)	MSTR.S	σ^2
Totale	SSTOT	sn - 1		

Test de l'effet de traitement

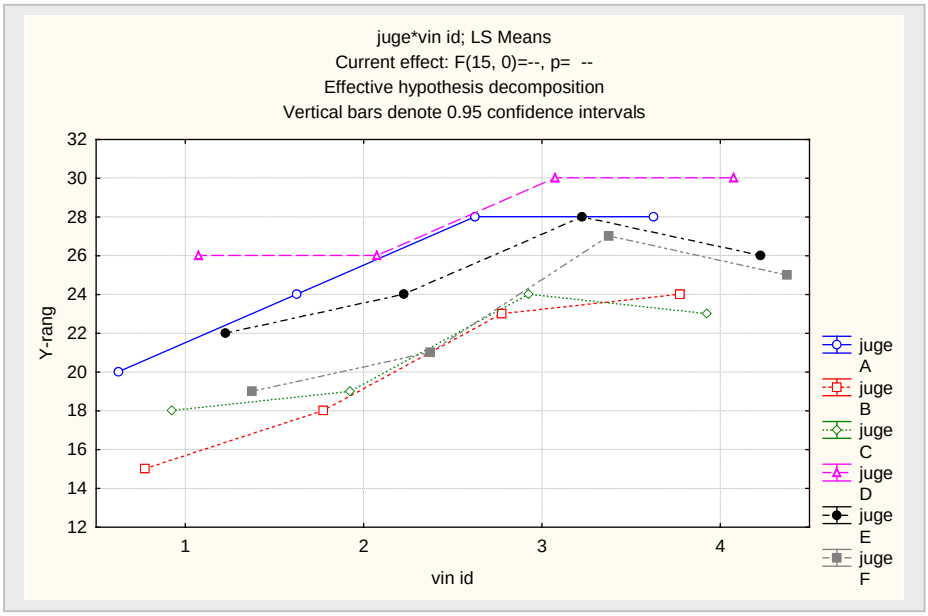
$$H_0 : \tau_j = 0 \quad \text{pour } j = 1, 2, \ldots, n$$
$$H_a : \text{les } \tau_j \text{ ne sont pas tous nuls}$$

On rejette H_0 si $F^* = MSTR / MSTR.S > F(1 - \alpha ; n - 1, ((n - 1)(s - 1))$ (168)
où α est le seuil du test

remarques

- **on peut analyser les données avec une ANOVA pour 2 facteurs sans répétition;**
- **si le facteur n'est pas fixe mais aléatoire, le test (168) reste valide;**
- **le design en mesures répétées est plus efficace que le design complètement aléatoire;**
- **on peut aussi tester l'hypothèse de l'effet du facteur sujet en utilisant le ratio MSS / MSTR.S**

Exemple 19 : (suite) analyse avec 2 facteurs sans répétition -procédure GLM ;
l'interaction sert à estimer l'erreur expérimentale
le facteur « juge » peut être considéré comme un facteur « bloc »

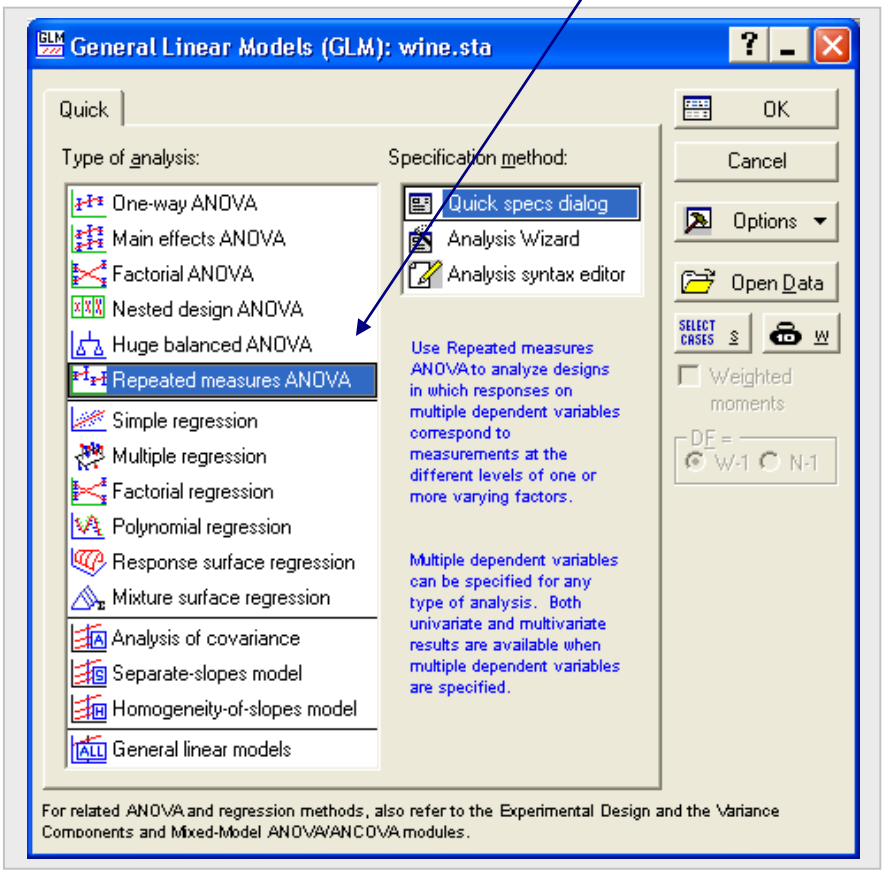


ANOVA erronnée car Y est dépendante

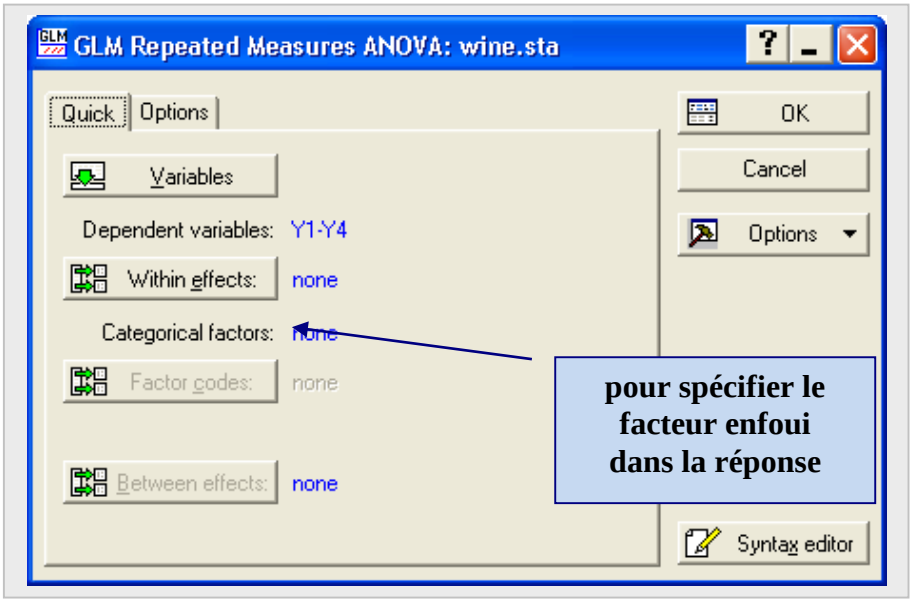
	DF	SS	MS	F	p-value
Intercept	1	13442.67	13442.67	12602.50	0.000000
juge	5	173.33	34.67	32.50	0.000000
vin id	3	184.00	61.33	57.50	0.000000
Error	15	16.00	1.07		
Total	23	373.33			

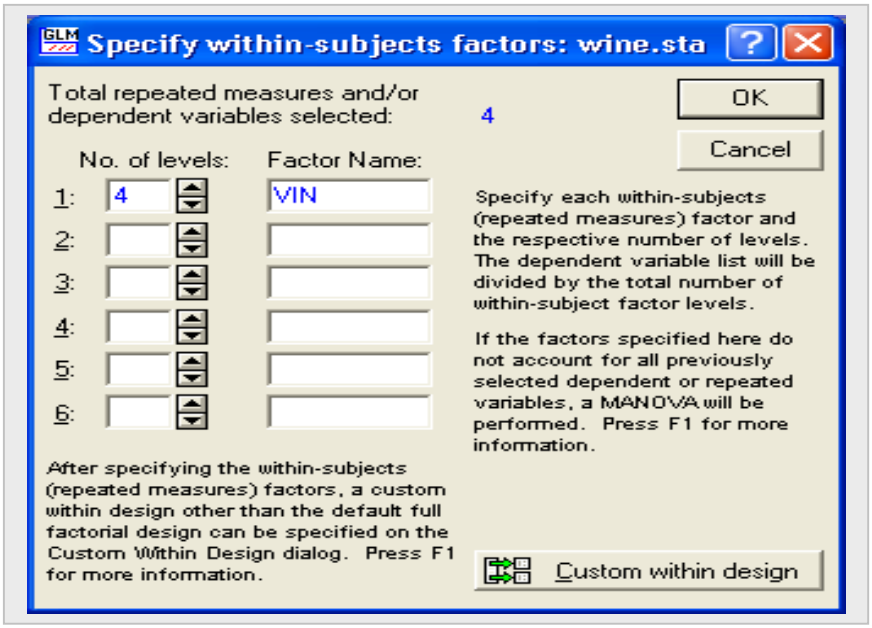
Utilisation de *STATISTICA* pour l'analyse du design en mesures répétées

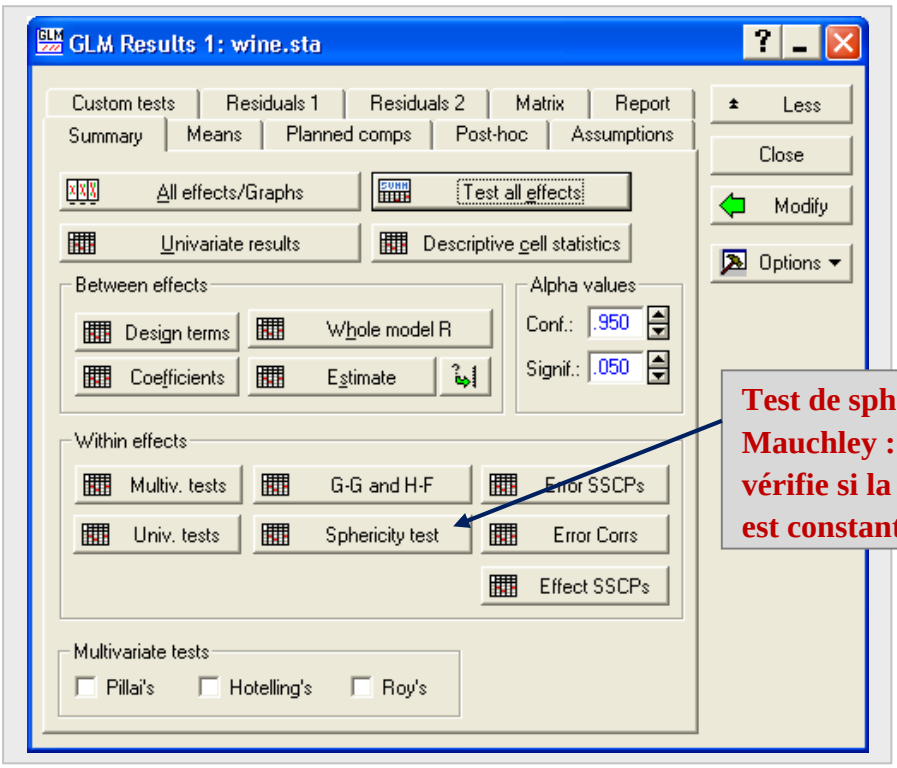
avec des facteurs intra enfoui dans la réponse (« whithin factors »)
et des facteurs inter (croisés) (« between factors »



Exemple 19 (suite) évaluation de 4 bouteilles de vin par 6 juges





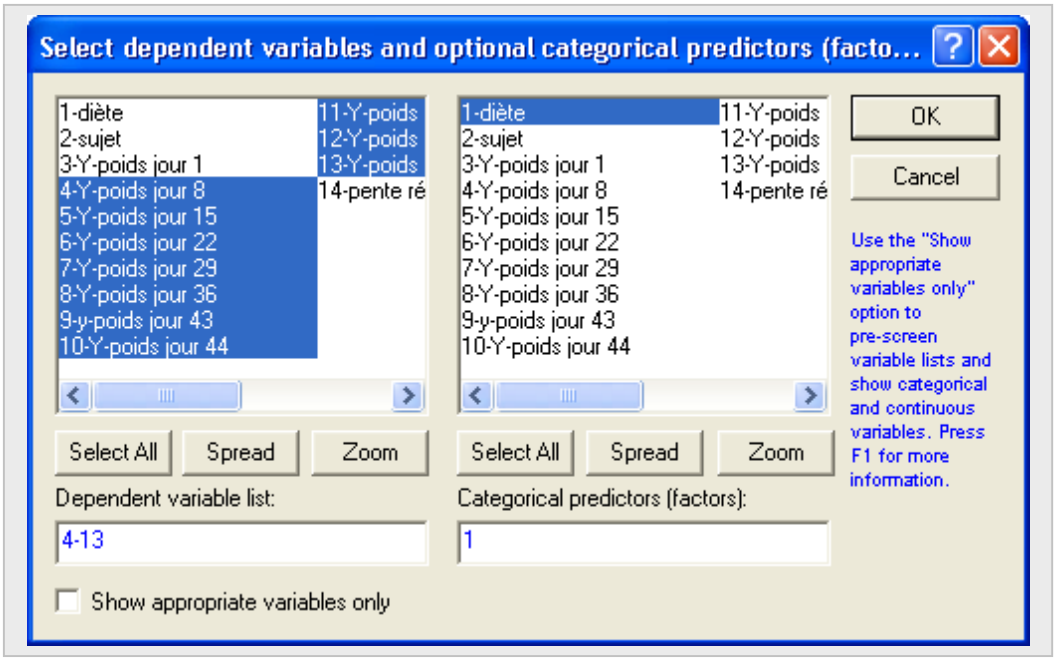


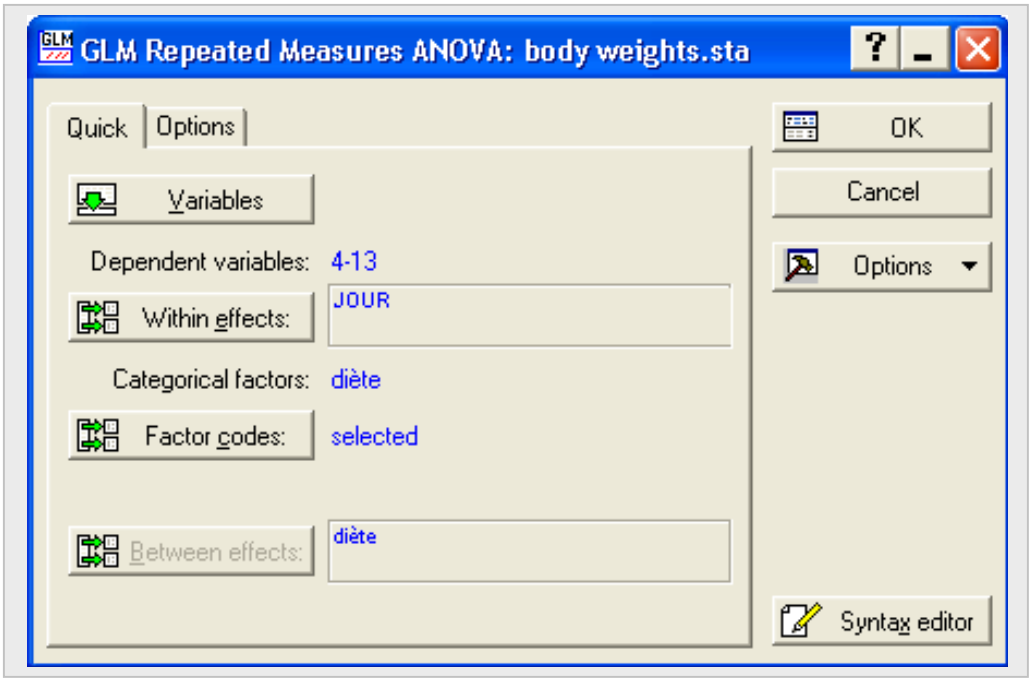
Repeated Measures Analysis of Variance					
	SS	df	MS	F	p
Intercept	13442.67	1	13442.67	387.7692	0.000006
Error	173.33	5	34.67		
VIN	184.00	3	61.33	57.5000	0.000000
Error	16.00	15	1.07		

<u>Test de sphéricité de Mauchley</u> : le coefficient de corrélation entre les Y est-il constant?
réponse : oui

Mauchley Sphericity Test				
	W	**Chi-Sqr.**	**df**	**p**
VIN	**0.351563**	**3.891091**	**5**	**0.5652**

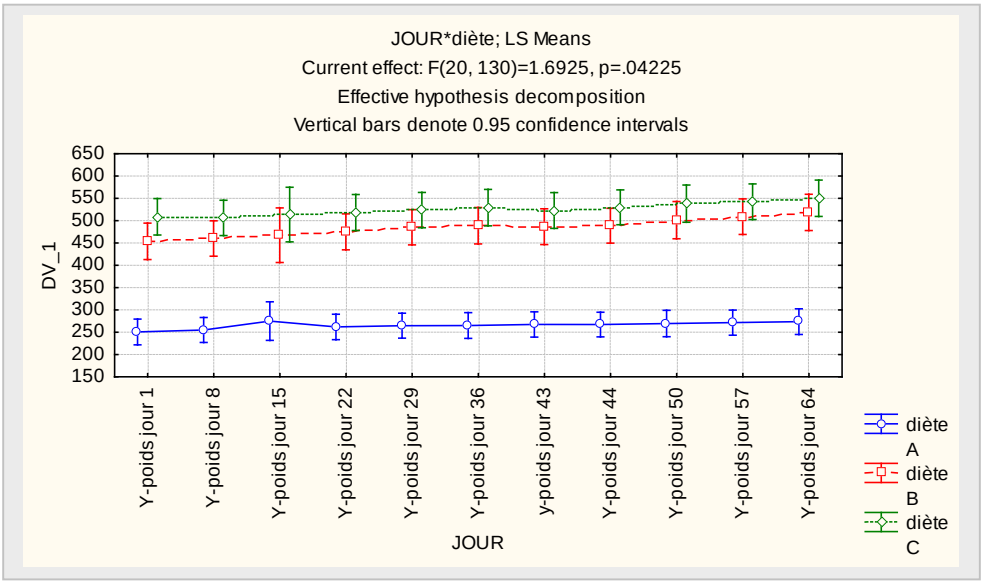
<u>Exemple</u> 20 : données de poids de 16 rats de laboratoire selon 3 diètes
poids de l'annimal au jour 1, 8, 15,…, 64

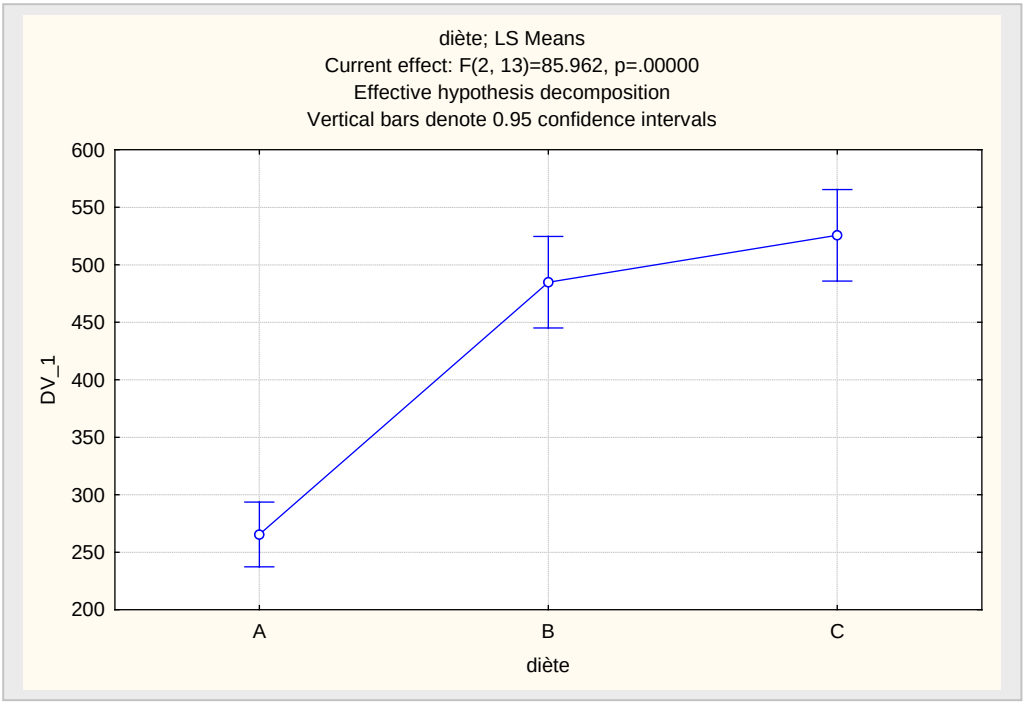




Repeated Measures Analysis of Variance

Source	SS	DF	MS	F	p
Intercept	28670191	1	28670191	1920.662	0.000000
diète	2566339	2	1283169	85.962	0.000000
Error	194054	13	14927		
JOUR	25041	10	2504	11.560	0.000000
JOUR*diète	7332	20	367	1.693	0.042250
Error	28160	130	217		

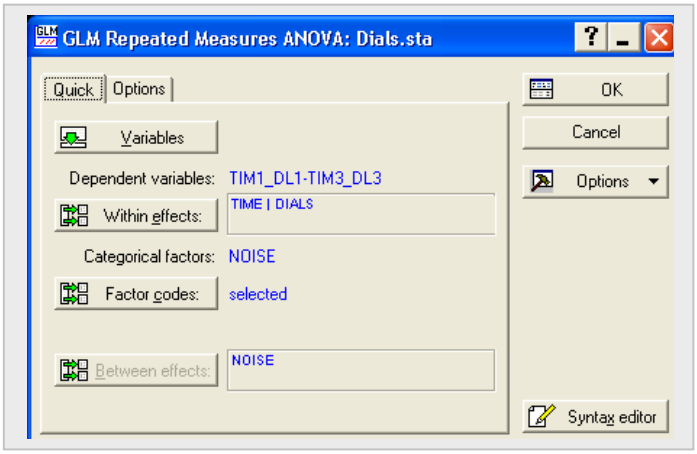




Exemple 22 : mesure de l'habileté d'un opérateur (sujet) de machine à ajuster 3 contrôles (« dials ») : DL1, DL2, DL3,
Y = nombre d'erreurs durant 3 périodes consécutives de 10 minutes
Les ajustements de l'opérateur sont faits selon 2 conditions qui lui sont inconnues: noise = « white » et « meaningful »
facteurs inter: 2 facteurs font enfuis dans la réponse : « dials » et « time »

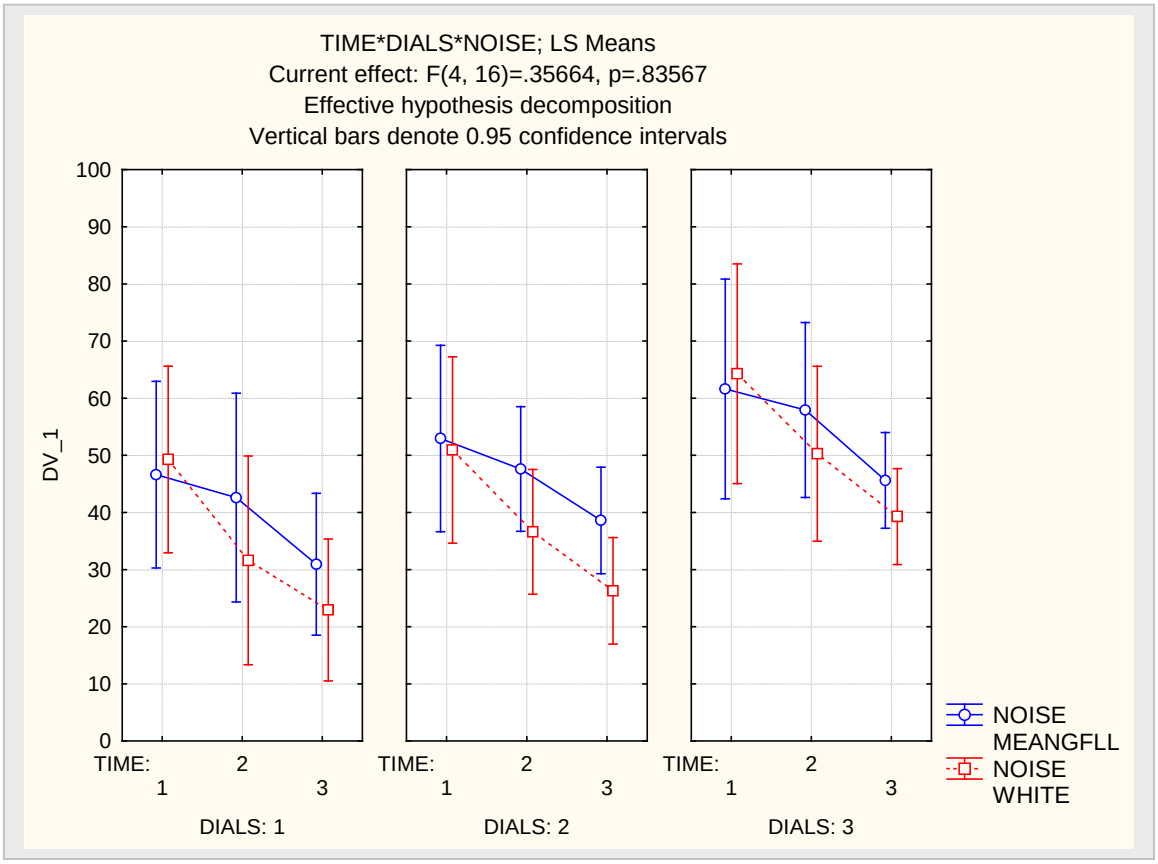
Example data file for repeated measures ANOVA

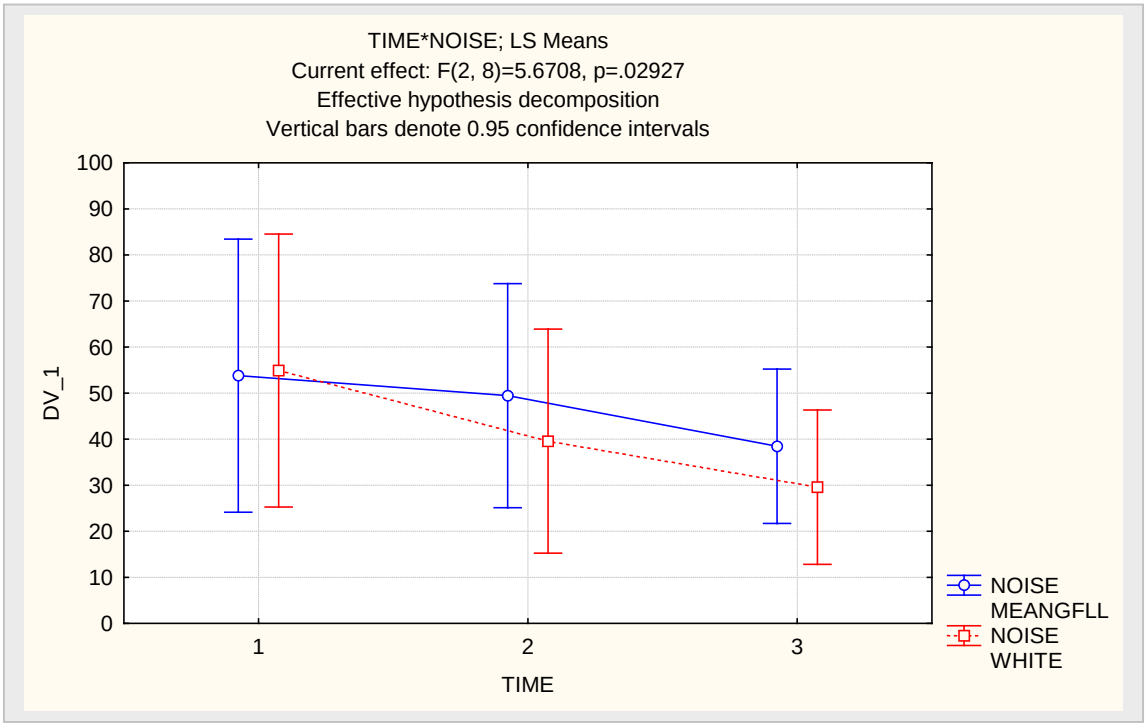
	NOISE	TIM1_DL1	TIM1_DL2	TIM1_DL3	TIM2_DL1	TIM2_DL2	TIM2_DL3	TIM3_DL1	TIM3_DL2	TIM3_DL3
1	MEANGFLL	45	53	60	40	52	57	28	37	46
2	MEANGFLL	35	41	50	30	37	47	25	32	41
3	MEANGFLL	60	65	75	58	54	70	40	47	50
4	WHITE	50	48	61	25	34	51	16	23	35
5	WHITE	42	45	55	30	37	43	22	27	37
6	WHITE	56	60	77	40	39	57	31	29	46



Repeated Measures Analysis of Variance

	SS	df	MS	F	p
Intercept	105868.2	1	105868.2	169.9935	0.000200
NOISE	468.2	1	468.2	0.7517	0.434841
Error	2491.1	4	622.8		
TIME	3722.3	2	1861.2	63.3888	0.000012
TIME*NOISE	333.0	2	166.5	5.6708	0.029268
Error	234.9	8	29.4		
DIALS	2370.3	2	1185.2	89.8232	0.000003
DIALS*NOISE	50.3	2	25.2	1.9074	0.210215
Error	105.6	8	13.2		
TIME*DIALS	10.7	4	2.7	0.3357	0.849917
TIME*DIALS*NOISE	11.3	4	2.8	0.3566	0.835669
Error	127.1	16	7.9		

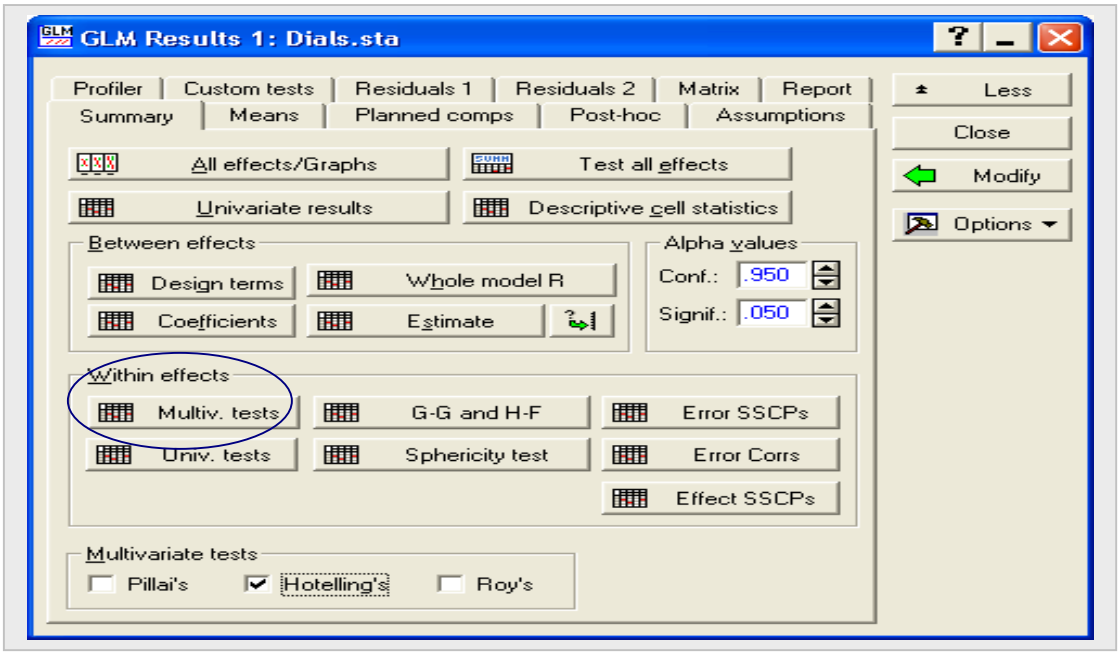




Interprétation : effet significatif de TIME et de DIALS
pas d'effet du facteur NOISE
interaction significative entre NOISE et TIME mais relativement faible.

MANOVA : Analyse de variance multivariée Exemple 22 : suite

Les 9 variables de réponse génèrent un vecteur en 9 dimensions. Les vecteurs de moyennes de ces variables diffèrent-ils selon les facteurs TIME, NOISE, et DIALS ? Il y a 4 tests multivariés (Wilk, Pillai, Hotelling, Roy) pour tester les effets principaux et d'interaction de ces facteurs.



Multivariate tests for repeated measure

	Test	Value	F	Effect	Error	p
TIME	Wilks	0.051	28.1453	2	3	0.011382
	Hotellng	18.764	28.1453	2	3	0.011382
TIME*NOISE	Wilks	0.156	8.1110	2	3	0.061657
	Hotellng	5.407	8.1110	2	3	0.061657
DIALS	Wilks	0.016	91.4562	2	3	0.002050
	Hotellng	60.971	91.4562	2	3	0.002050
DIALS*NOISE	Wilks	0.565	1.1549	2	3	0.424671
	Hotellng	0.770	1.1549	2	3	0.424671
TIME*DIALS	Wilks	0.001	331.4450	4	1	0.041170
	Hotellng	1325.780	331.4450	4	1	0.041170
TIME*DIALS*NOISE	Wilks	0.000	581.8750	4	1	0.031081
	Hotellng	2327.500	581.8750	4	1	0.031081

Expérience à 2 facteurs avec mesures répétées sur un seul facteur

Dans plusieurs situations expérimentales faisant intervenir 2 facteurs, des mesures répétées peuvent être faites seulement sur un des deux facteurs. Par exemple, supposons que l'on veuille étudier l'effet de deux types de stimuli (facteur A) sur l'habileté d'un individu à résoudre deux types de problèmes (facteur B). Les modalités du facteur B sont : problème concret, problème abstrait.

Chaque sujet devra résoudre chaque type de problème mais il ne pourra être exposé aux 2 conditions de stimuli (facteur A) à cause de l'effet résiduel causé par l'apprentissage. L'organisation des données est présentée par le schéma suivant. Le facteur B est répété et chaque sujet constitue un bloc. Certains sujets recoivent B1 pour commencer et B2 ensuite tandis que les autres recoivent la séquence B2 au début et B1 ensuite. (« Cross over design »

stimulus	sujet	traitement A	ordre 1 traitement B	ordre 2 traitement B
A1	1	A1	B1	B2
	.	.	.	.
	s	A1	B2	B1
……..	…….	……	……..	…….
A2	s + 1	A2	B2	B1
	.	.	.	.
	2s	A2	B1	B2

Modèle : S : facteur aléatoire sujet $i = 1, 2, \ldots, s$
A : facteur fixe $j = 1, 2, \ldots, a$
B : facteur fixe $k = 1, 2, \ldots, b$
le facteur S est emboité dans le facteur A

Y_{ijk} : réponse

$$Y_{ijk} = \mu + \rho_{i(j)} + \alpha_j + \beta_k + (\alpha\beta)_{jk} + \varepsilon_{ijk} \qquad (169)$$

où μ : effet général

$\rho_{i(j)}$: effet aléatoire du sujet (bloc) i emboité dans la modalité j du facteur A
distribution $N(0, \sigma_p^2)$

α_j : effet différentiel du facteur A $\quad \sum \alpha_j = 0$

β_k : effet différentiel du facteur B $\quad \sum \beta_k = 0$

$(\alpha\beta)_{jk}$: effet d'interaction AB $\quad \sum (\alpha\beta)_{jk} = 0 \quad \sum (\alpha\beta)_{jk} = 0$

ε_{ijk} : erreur aléatoire distribuée $\quad N(0, \sigma^2)$

ε_{ijk}, $\rho_{i(j)}$ sont indépendantes

Conséquences

$$E(Y_{ijk}) = \mu + \rho_{i(j)} + \alpha_j + \beta_k + (\alpha\beta)_{jk} \quad (170)$$

$$Var(Y_{ijk}) = \sigma_p^2 + \sigma^2 \quad (171)$$

$$Cov(Y_{ijk}, Y_{ijk'}) = \sigma_p^2 \qquad k \neq k' \quad (172)$$

$$Cov(Y_{ijk}, Y_{i'j'k'}) = 0 \qquad i \neq i' \text{ ou/et } j \neq j' \quad (173)$$

Analyse de la variance

facteur A	$SSA = bs \sum (\overline{Y}_{.j.} - \overline{Y}_{...})^2$	(174)
facteur B	$SSB = as \sum (\overline{Y}_{.k.} - \overline{Y}_{...})^2$	(175)
interaction AB	$SSAB = s \sum\sum (\overline{Y}_{.jk} - \overline{Y}_{.j.} - \overline{Y}_{..k} + \overline{Y}_{...})^2$	(176)
facteur S sujet (emboité A)	$SSS(A) = b \sum\sum (\overline{Y}_{ij.} - \overline{Y}_{.j.})^2$	(177)
erreur	$SSB.S(A) = \sum\sum\sum (Y_{ijk} - \overline{Y}_{.jk} - \overline{Y}_{ij.} + \overline{Y}_{.j.})^2$	(178)
totale	$SSTOT = \sum\sum\sum (Y_{ijk} - \overline{Y}_{...})^2$	(179)

ANOVA

Source	SS	df	MS	F	test : rejet H_0 si
A	SSA	a – 1	MSA	MSA / MSS(A)	> F(1 - α; a-1; a(s-1))
B	SSB	b – 1	MSB	MSB / MSB.S(A)	> F(1 - α; a-1; a(s-1)(b-1))
AB	SSAB	(a-1)(b-1)	MSAB	MSAB / MSB.S(A)	> F(1 - α; a-1; a(s-1)(b-1))
S(A)	SSS(A)	a(s – 1)	MSS(A)	-------	----------
Erreur	SSB.S(A)	a(s-1)(b-1)	MSB.S(A)	-------	
Totale	SSTOT	abs – 1	------	--------	----------

Espérance des carrés moyens MS : constitue la base des tests sur les coefficients du modèle

$$E(MSA) = \sigma^2 + b\sigma_p^2 + (bs/(a-1)) \sum \alpha_j^2 \quad (180)$$

$$E(MSB) = \sigma^2 + (as/(b-1)) \sum \beta_k^2 \quad (181)$$

$$E(MSAB) = \sigma^2 + (s/(a-1)(b-1)) \sum\sum (\alpha\beta)^2_{jk} \quad (182)$$

$$E(MSS(A)) = \sigma^2 + b\sigma_p^2 \quad (183)$$

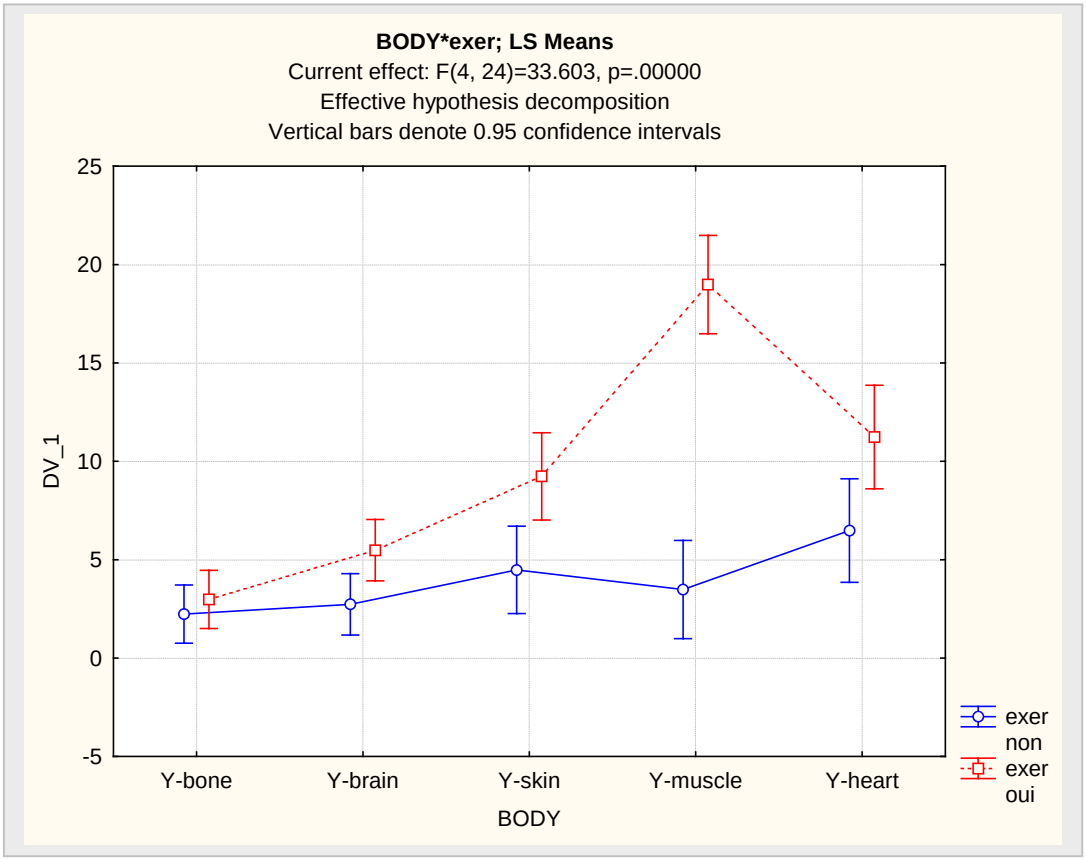
$$E(MSB.S(A)) = \sigma^2 \quad (184)$$

Exemple 21 : afflux de sang (bloodflow) - il y a 2 termes d'erreur

Repeated Measures Analysis of Variance

	SS	DF	MS	F	p
Intercept	1822.5	1	1822.5	247.40	0.0000
exer	324.9	1	324.9	324.9 / 7.4 = 44.1	0.0006
erreur	44.2	6	7.4	--	--
BODY	389.5	4	97.4	97.4 / 1.9 = 49.9	0.0000
BODY*exer	262.1	4	65.5	65.5 / 1.9 =33.6	0.0000
erreur	46.8	24	1.9	--	--
total	1067.5	39	--	--	--

S(A) = erreur = 7.4 **B.S(A) = erreur = 1.9**



Expérience à 2 facteurs avec mesures répétées sur deux facteurs

Exemple 23 : mesure de afflux de sang suite à la prise de 2 médicaments A et B

A1B1 : placebo, placebo **A1B2 :placebo, médicament B**

A2B1 : médicament A, placebo **A2B2 : médicament A et médicament B**

Kutner et all 5 ed. p. 1158					
	sujet	Y-A1B1	Y-A1B2	Y-A2B1	Y-A2B2
1	s1	2	10	9	25
2	s2	-1	8	6	21
3	s3	0	11	8	24
4	s4	3	15	11	31
5	s5	1	5	6	20
6	s6	2	12	9	27
7	s7	-2	10	8	22
8	s8	4	16	12	30
9	s9	-2	7	7	24
10	s10	-2	10	10	28
11	s11	2	8	10	25
12	s12	-1	8	6	23

Modèle S : facteur sujet (aléatoire) (= facteur bloc) $i = 1, 2, \ldots, s$

A : facteur1 fixe $j = 1, 2, \ldots, a$

B : facteur2 fixe $k = 1, 2, \ldots, b$

chaque sujet S recoit toutes les combinaisons (j, k) des facteurs A et B

Y_{ijk} : réponse du sujet i recevant modalité j de A et la modalité k de B

$$Y_{ijk} = \mu + \rho_i + \alpha_j + \beta_k + (\alpha\beta)_{jk} + (\rho\alpha)_{ij} + (\rho\beta)_{ik} + \varepsilon_{ijk} \qquad (185)$$

où μ : effet général

ρ_i : effet aléatoire du facteur sujet $\sim N(0, \sigma_\rho^2)$

α_j : effet différentiel du facteur fixe A $\sum \alpha_j = 0$

β_k : effet différentiel du facteur fixe B $\sum \beta_k = 0$

$(\alpha\beta)_{jk}$: effet d'interaction AB et $\sum_j (\alpha\beta)_{jk} = 0$ tout k et $\sum_k (\alpha\beta)_{jk} = 0$ tout j

$(\rho\alpha)_{ij}$: effet aléatoire d'interaction SA $\sim N(0, ((a-1)/a)\,\sigma^2_{\rho\alpha})$ $\sum_j (\rho\alpha)_{ij} = 0$ tout i

$\text{cov}((\rho\alpha)_{ij}, (\rho\alpha)_{ij'}) = (-1/a)\,\sigma^2_{\rho\alpha}$ $j \neq j'$

$(\rho\beta)_{ik}$: effet aléatoire d'intraction SB $\sim N(0, ((b-1)/b)\,\sigma^2_{\rho\beta})$ $\sum_i (\rho\beta)_{ik} = 0$ tout i

$\text{cov}((\rho\beta)_{ik}, (\rho\beta)_{ik'}) = (-1/b)\,\sigma^2_{\rho\beta}$ $k \neq k'$

ρ_i, $(\rho\alpha)_{ij}$, $(\rho\beta)_{ik}$ sont indépendantes 2 à 2

ε_{ijk}, ρ_i, $(\rho\alpha)_{ij}$, $(\rho\beta)_{ik}$ sont indépendantes de ρ_i, $(\rho\alpha)_{ij}$, $(\rho\beta)_{ik}$

$\varepsilon_{ijk} \sim N(0, \sigma^2)$

Conséquences

$$E(Y_{ijk}) = \mu \ldots + \alpha_j + \beta_k + (\alpha\beta)_{jk} \quad (186)$$

$$Var(Y_{ijk}) = \sigma^2_\rho + ((a-1)/a)\,\sigma^2_{\rho\alpha} + ((b-1)/b)\,\sigma^2_{\rho\beta} + \sigma^2 \quad (187)$$

Analyse de la variance

facteur sujet S $$SSS = ab \sum (\overline{Y}_{i..} - \overline{Y}_{...})^2 \quad (188)$$

facteur A $$SSA = sb \sum (\overline{Y}_{.j.} - \overline{Y}_{...})^2 \quad (189)$$

facteur B $$SSB = sa \sum (\overline{Y}_{..k} - \overline{Y}_{...})^2 \quad (190)$$

interaction AB $$SSAB = s \sum\sum (\overline{Y}_{.jk.} - \overline{Y}_{.j.} - \overline{Y}_{..k} + \overline{Y}_{...})^2 \quad (191)$$

interaction AS $$SSAS = b \sum\sum (\overline{Y}_{ij.} - \overline{Y}_{.i.} - \overline{Y}_{.j.} + \overline{Y}_{...})^2 \quad (192)$$

interaction BS $$SSBS = a \sum\sum (\overline{Y}_{i.k} - Y_{i..} - \overline{Y}_{..k} + \overline{Y}_{...})^2 \quad (193)$$

interaction ABS

$$SSABS = \sum\sum\sum (\overline{Y}_{ijk} - \overline{Y}_{ij.} - \overline{Y}_{i.k} - \overline{Y}_{.jk} + \overline{Y}_{i..} + \overline{Y}_{.j.} + \overline{Y}_{..k} - \overline{Y}_{...})^2 \quad (194)$$

totale $$SSTO = \sum\sum\sum (Y_{ijk} - \overline{Y}_{...})^2 \quad (195)$$

Espérance des carrés moyens MS : constitue la base des tests sur les coefficients du modèle

$$E(MSS) = \sigma^2 + ab\,\sigma_\rho^2 \quad (196)$$

$$E(MSA) = \sigma^2 + b\,\sigma_{\rho\alpha}^2 + ((bs/(a-1)) \sum \alpha_j^2 \quad (197)$$

$$E(MSB) = \sigma^2 + a\,\sigma_{\rho\beta}^2 + ((as/(b-1)) \sum \beta_k^2 \quad (198)$$

$$E(MSAB) = \sigma^2 + ((s/(a-1)(b-1)) \sum\sum (\alpha\beta)_{jk}^2 \quad (199)$$

$$E(MSAS) = \sigma^2 + b\,\sigma_{\rho\alpha}^2 \quad (200)$$

$$E(MSBS) = \sigma^2 + b\,\sigma_{\rho\beta}^2 \quad (201)$$

$$E(MSABS) = \sigma^2 \quad (202)$$

ANOVA

Source	SS	df	MS	F	test: rejet H_0 si
Sujets S	SSS	s – 1	MSS	----------	----------
Facteur A	SSA	a – 1	MSA	F3=MSA / MSAS	> F(1-α; a-1;(a-1)(s-1))
Facteur B	SSB	b – 1	MSB	F2=MSB / MSBS	> F(1-α; b-1;(b-1)(s-1))
Inter AB	SSAB	(a-1)(b-1)	MSAB	F1=MSAB / MSABS	> F(1- α; (a-1)(b-1); (a-1)(b-1)(s-1)
Inter AS	SSAS	(a-1)(s-1)	MSAS	---------	----------
Inter BS	SSBS	(b-1)(s-1)	MSBS	---------	----------
Erreur	SSABS	(a-1)(b-1)(s-1)	MSABS	---------	----------
Totale	SSTO	abs -1	--------		

Hypothèses

$H_0 : (\alpha\beta)_{jk} = 0$ pour tout (j, k) — ratio F1 plus haut

$H_0 : \alpha_j = 0$ pour tout j — ratio F3 plus haut

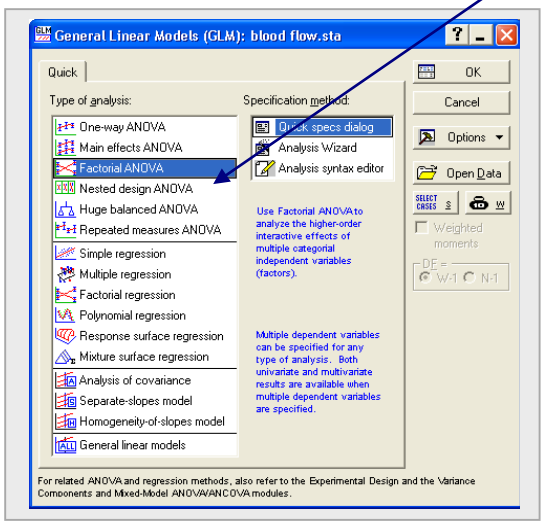
$H_0 : \beta_k = 0$ pour tout k — ratio F2 plus haut

Exemple 23 : données de « blood flow » : deux méthodes pour l'analyse

méthode 1 : considérer que les 3 facteurs A, B, S interviennent dans une expérience factorielle complète sans répétition ; on emploie « Factorial ANOVA »
remarque : il est important de définir correctement les ratios pour les tests

méthode 2 : approche à mesures répétées avec 4 variables de réponse

Méthode 1 procédure Factorial ANOVA



données : format classique

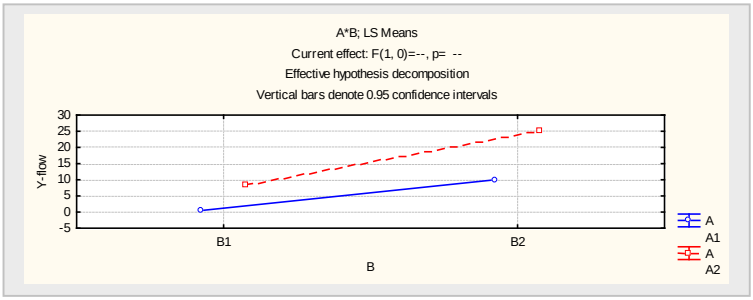
Kutner et all 5 ed. p. 1158

	sujet	A	B	Y-flow
1	s1	A1	B1	2
2	s1	A1	B2	10
3	s1	A2	B1	9
4	s1	A2	B2	25
.	.	.	.	.
43	s11	A2	B1	10
44	s11	A2	B2	25
45	s12	A1	B1	-1
46	s12	A1	B2	8
47	s12	A2	B1	6
48	s12	A2	B2	23

Univariate Tests of Significance for Y Effective hypothesis decomposition

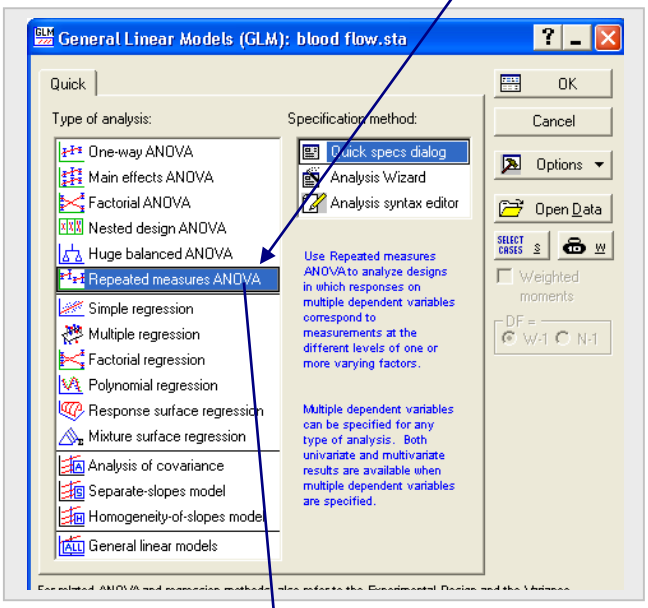
	SS	DF	MS	F	p
Intercept	5808.0	1	5808.0		
sujet	258.5	11	23.5		
A	1587.0	1	1587.0		
B	2028.0	1	2028.0		
sujet*A	22.5	11	2.045		
sujet*B	42.50	11	3.864		
A*B	147.0	1	147.0		
sujet*A*B	12.5	11	1.136		
Error		0			
Total	4098.0				

Test de l'interaction AB : F3 = 147.000 / 1.136 = 129.4 significatif
Test de l'effet de A : F1 = 1587.00 / 2.045 = 776.04 significatif
Test de l'effet de B : F2 = 2028.00 / 3.864 = 524.85 significatif



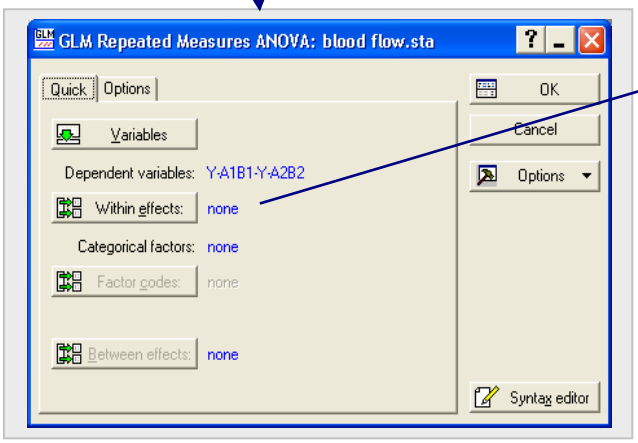
Méthode 2 : approche à mesures répétées avec 4 variables de réponse

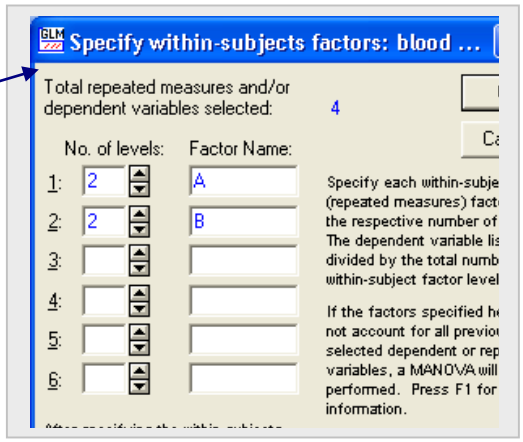
données : facteurs enflouis dans la réponse



Kutner et all 5 ed. p. 1158

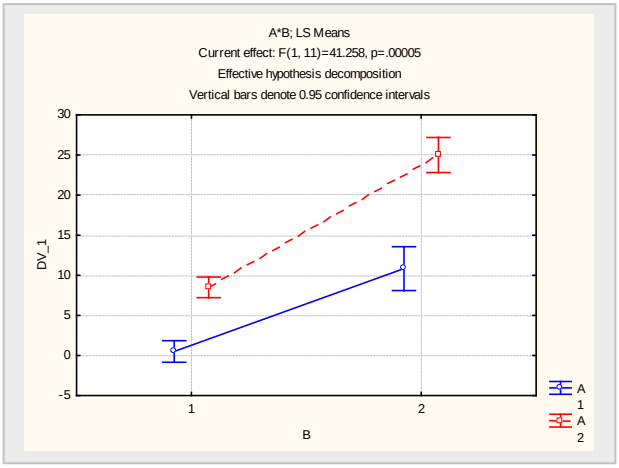
	sujet	Y-A1B1	Y-A1B2	Y-A2B1	Y-A2B2
1	s1	2	10	9	25
2	s2	-1	8	6	21
3	s3	0	11	8	24
4	s4	3	15	11	31
5	s5	1	5	6	20
6	s6	2	12	9	27
7	s7	-2	10	8	22
8	s8	4	16	12	30
9	s9	-2	7	7	24
10	s10	-2	10	10	28
11	s11	2	8	10	25
12	s12	-1	8	6	23





Repeated Measures Analysis of Variance

	SS	Degr. of Freedom	MS	F	p
Intercept	5808,000	1	5808,000	247,149	0,000000
Error	258,500	11	23,500		
A	1587,000	1	1587,000	775,867	0,000000
Error	22,500	11	2,045		
B	2028,000	1	2028,000	524,894	0,000000
Error	42,500	11	3,864		
A*B	147,000	1	147,000	129,360	0,000000
Error	12,500	11	1,136		



Remarque

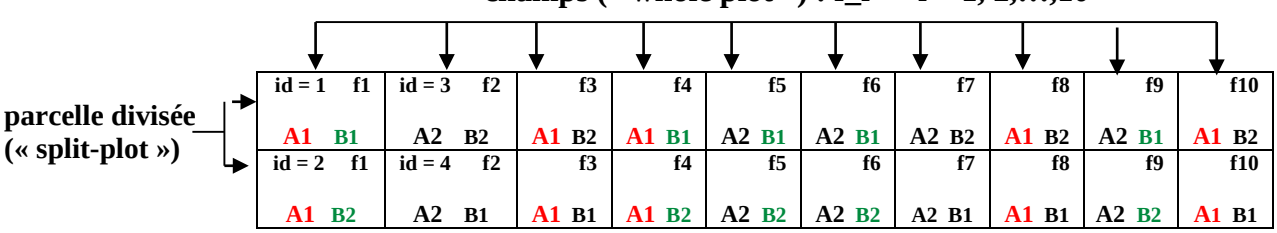
Si le design n'est pas équilibré (nombre inégal dans chaque cellule) ou s'il y a des données absentes dans certaines cellules, if faut employer une approche par régression avec des variables indicatrices pour faire l'analyse.

15. Le plan à parcelles divisées (« Split-Plot design »)

Le plan à parcelles divisées est fréquemment employé dans les expériences en sciences agronomiques, en laboratoire, dans les expériences industrielles, en sciences de la vie et en sciences sociales. Nous présentons le cas d'expériences avec deux facteurs mais on peut appliquer les plans à parcelles divisées avec plus de deux facteurs.

Le plan a été originalement développé pour les expériences agronomiques. On considère une étude pour évaluer l'effet de deux facteurs sur le rendement Y d'une production d'une variété de blé. Les deux facteurs sont : la méthode d'irrigation (facteur A) et le fertilisant (facteur B). On suppose que chaque facteur prend deux modalités ou plus. Si chaque facteur varie avec deux modalités, on a donc un plan factoriel complet avec quatre traitements : (A, B) = (A1, B1), (A1, B2), (A2, B1), (A2, B2).

La caractéristique fondamentale du *plan à parcelles divisées* est la méthode d'attribution des traitements aux unités expérimentales. Il y a un premier niveau d'attribution à des unités expérimentales qui sont appelées les parcelles (« whole plot »). Celles-ci recoivent, au hasard, l'une des deux modalités du facteur A. Les parcelles sont ensuite subdivsées en de plus petites unités appelées parcelles divisées (« split-plot »). Ces dernières recoivent, au hasard, l'une des deux modalités du facteur B. Dans l'exemple, 10 champs reçurent une méthode d'irrigation (A) et, après subdivision des champs en unité plus petite (« split-plot »), on attribua un fertilisant (B). La figure illustre le processus d'attribution des traitements A et B aux champs et parcelles.



Les champs f_i reçurent (dans l'ordre) la méthode d'irrigation : A1 / A2 / A1 / A1 / A2 / A2 / A2 / A1 / A2 / A1.
Les paires de parcelles recurent les modalités (B1, B2), (B2, B1), (B2, B1), (B1, B2), (B1, B2), (B1, B2), (B2, B1), (B2, B1), (B1, B2), (B2, B1)

Exemple 24 : rendement (Y-yield) d'une variété de blé

Kutner et all 5 ed. p. 1170 - données design split-plot / parcelles divisées

id	field	A irrigat	B fertilisant	Y yield		id	field	A irrigat	B fertilisant	Y yield
1	f1	irrig1	fert1	43		11	f6	irrig2	fert1	45
2	f1	irrig1	fert2	48		12	f6	irrig2	fert2	48
3	f2	irrig2	fert2	70		13	f7	irrig2	fert2	51
4	f2	irrig2	fert1	63		14	f7	irrig2	fert1	47
5	f3	irrig1	fert2	43		15	f8	irrig1	fert1	27
6	f3	irrig1	fert1	40		16	f8	irrig1	fert2	30
7	f4	irrig1	fert1	31		17	f9	irrig2	fert1	54
8	f4	irrig1	fert2	36		18	f9	irrig2	fert2	57
9	f5	irrig2	fert1	52		19	f10	irrig1	fert2	39
10	f5	irrig2	fert2	53		20	f10	irrig1	fert1	36

Structure

Dans cet exemple, la structure du plan est formé par un facteur externe (ou facteur inter unité) (« between factor ») constitué par irrigation et d'un facteur enfoui dans la réponse (ou facteur intra unité) (« within factor ») constitué par le facteur fertilisant. Il ya deux types d'unité : les champs constituant les grandes unités et, les parcelles, qui forment les unités plus petites.

Modèle

Cette structure a déjà été vue :(voir eq. (169): *deux facteurs avec mesure répétées sur un facteur*

$$Y_{ijk} = \mu + \rho_{i(j)} + \alpha_j + \beta_k + (\alpha\beta)_{jk} + \varepsilon_{ijk} \qquad (203)$$

α_j : effet de la méthode d'irrigation (facteur A) – sur l'unité parcelle
β_k : effet du fertilisant (facteur B) – traitement parcelle divisée
$\rho_{i(j)}$: effet de la modalité i du facteur parcelle (=sujet) emboité dans la modalité j du facteur A (méthode d'irrigation)
$(\alpha\beta)_{jk}$ effet d'interaction entre le facteur A et le facteur B

Le tableau d'ANOVA est identique à celui déjà présenté. Nous le reproduisons ici en changeant l'ordre des lignes et en renommant SSS(A) par SSW(A) et SSB.S(A) par SSB.W(A). Les tests d'hypothèses concernent les l'effet principal du facteur A, celui de facteur B ainsi que l'effet d'interaction AB.

ANOVA

Source	SS	df	MS	F	test : rejet H_0 si
grande parcelle (whole plot)					
A	SSA	a – 1	MSA	MSA / MSW(A)	> F(1 - α; a-1; a(s-1))
S(A)(erreur)	SSW(A)	a(s – 1)	MSW(A)	-------	----------
petite parcelle (split plot)					
B	SSB	b – 1	MSB	MSB / MSB.W(A)	> F(1 - α; a-1; a(s-1)(b-1))
AB	SSAB	(a-1)(b-1)	MSAB	MSAB / MSB.W(A)	> F(1 - α; a-1; a(s-1)(b-1))
Erreur	SSB.W(A)	a(s-1)(b-1)	MSB.W(A)-------		
Totale	SSTO	abs – 1	------	--------	----------

réorganisation de données en mesures répétées

Kutner et all 5 ed. p. 1170 - données design split-plot / parcelles divisées

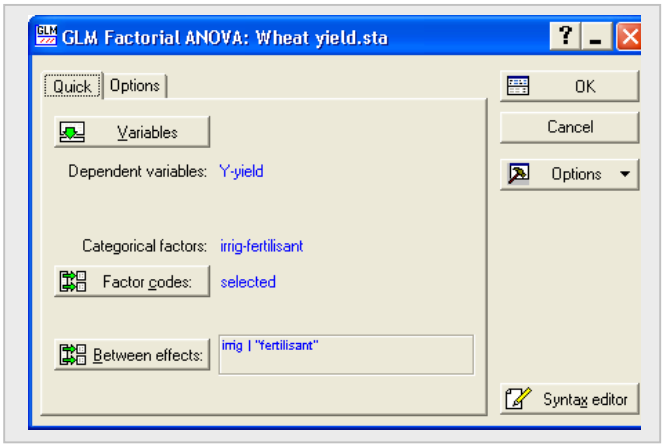
	field2	irrigat2	Y-fert1	Y-fert2
1	f1	irrig1	43	48
2	f2	irrig1	40	43
3	f3	irrig1	31	36
4	f4	irrig1	27	30
5	f5	irrig1	36	39
6	f6	irrg2	63	70
7	f7	irrg2	52	53
8	f8	irrg2	45	48
9	f9	irrg2	47	51
10	f10	irrg2	54	57

Analyse avec *STATISTICA*

Une analyse erronnée : basée sur une interprétation erronnée du plan expérimental :
plan factoriel de 2 facteurs et assignation des traitements aléatoirement aux unités expérimentales. Il y a 2 erreurs dans cette interprétation :

- **l'assignation est en mode parcelles divisées et non pas aléatoire ;**
- **il y a deux tailles d'unités expérimentales.**

Voyons le résultat de la mise en œuvre GLM factorial ANOVA.



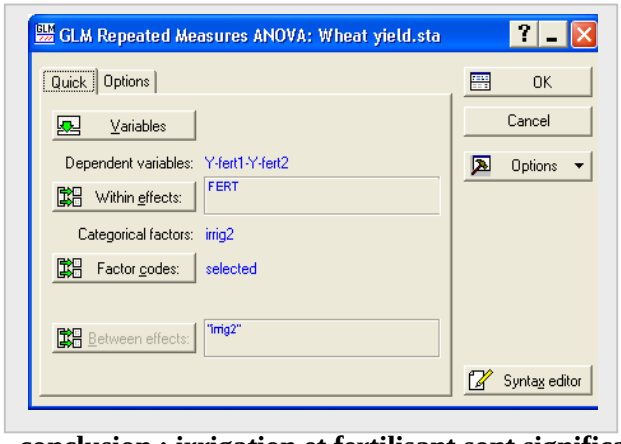
ANOVA

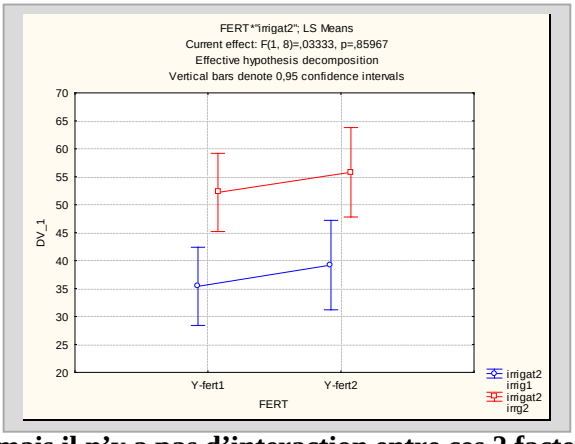
	DF	SS	MS	F	p-value
Intercept	**1**	**41678.45**	**41678.45**	**784.9049**	**0.000000**
irrig	**1**	**1394.45**	**1394.45**	**26.2608**	**0.000102**
fertilisant	**1**	**68.45**	**68.45**	**1.2891**	**0.272940**
irrig*fertilisant	**1**	**0.05**	**0.05**	**0.0009**	**0.975900**
Error	**16**	**849.60**	**53.10**		
Total	**19**	**2312.55**			

conclusion : facteur irrigation est significatif
facteur fertilisant et l'interaction irrig*fertilisant non significatifs

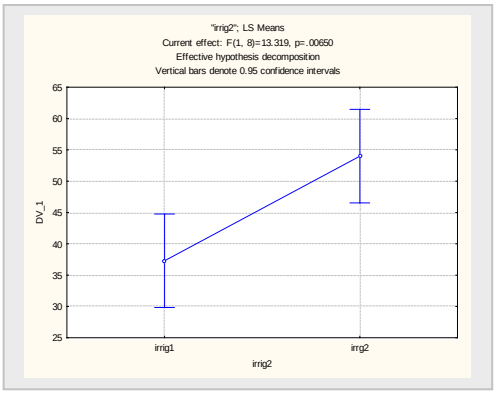
Mais cette analyse est erronnée pour les raisons mentionnées plus haut.

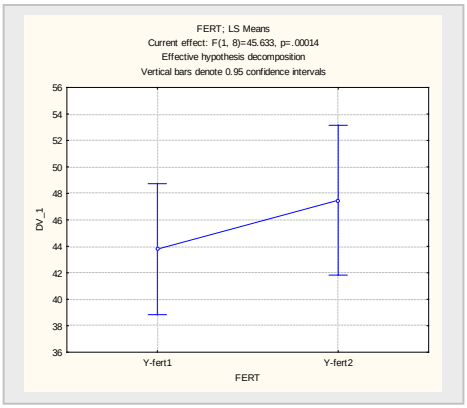
Une analyse correcte : basée sur le design « Split Plot » - mesures répétées





conclusion : irrigation et fertilisant sont significatifs mais il n'y a pas d'interaction entre ces 2 facteurs.





ANOVA : Repeated Measures Analysis of Variance

	SS	DF	MS	F	p
Intercept	41678.45	1	41678.45	398.0750	0.000000
"irrig2"	1394.45	1	1394.45	13.3185	0.006499
Error	837.60	8	104.70		
FERT	68.45	1	68.45	45.6333	0.000144
FERT*"irrig2"	0.05	1	0.05	0.0333	0.859674
Error	12.00	8	1.50		

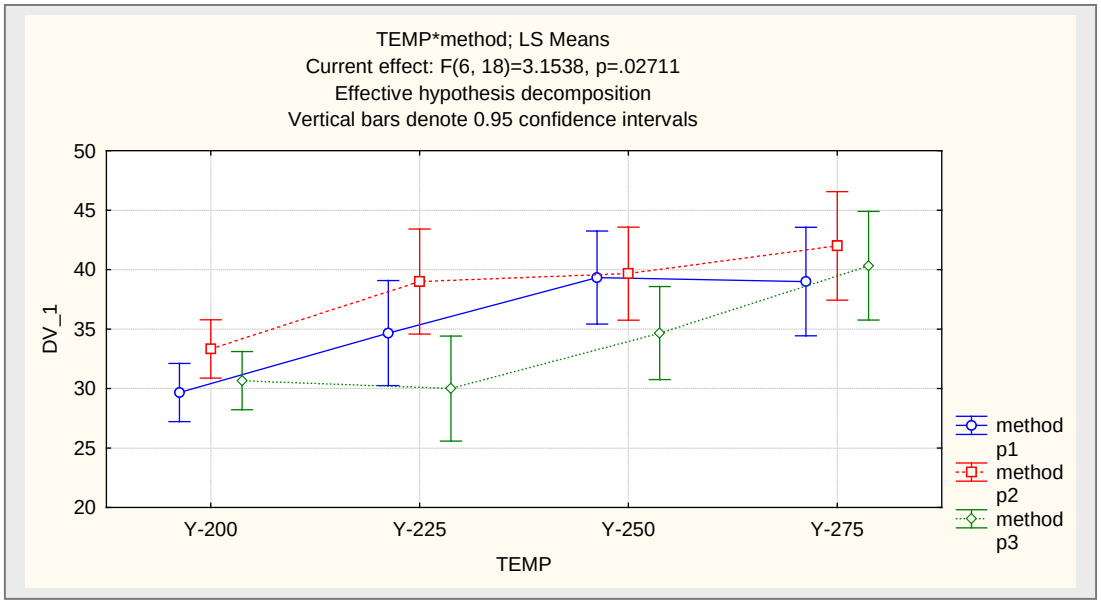
Exemple 25 : fabrication papier

- 3 méthodes préparation pulpe : P1 – P2 -P3
- 4 niveaux de température : 200 - 225 - 250 - 275 (deg. F)
- 3 x 4 = 12 traitements
- réponse Y : force tension papier
- 2 répétitions : n = 3
- contrainte ressources : 12 essais par jour
- exécution des 12 traitements par jour : répétition ou bloc
- lot («plot») pulpe préparé selon une méthode
- lot divisé en 4 échantillons («subplot» / «split-plot» /split unit)

données	répétition (bloc) 1			répétition 2			répétition 3		
méthode prép. pulpe	P1	P2	P3	P1	P2	P3	1	2	3
température									
200	30	34	29	28	31	31	31	35	32
225	35	41	26	32	36	30	37	40	34
250	37	38	33	40	42	32	41	39	39
275	36	42	36	41	40	40	40	44	45

Repeated Measures Analysis of Variance - ANOVA

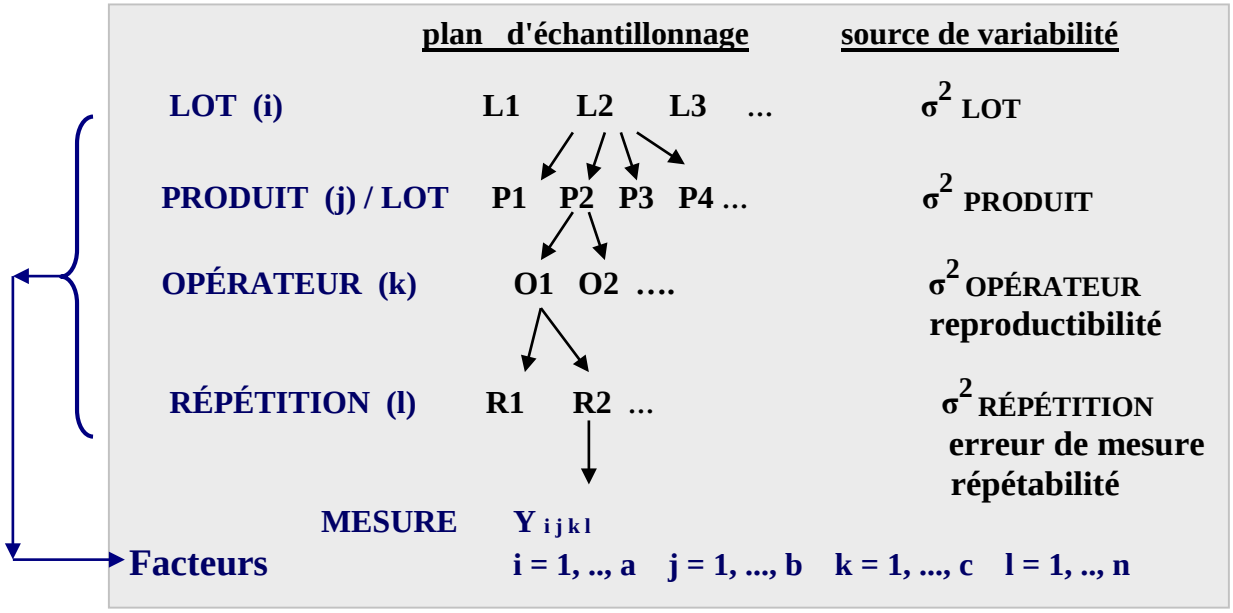
	SS	DF	MS	F	p
Intercept	46728.03	1	46728.03	2462.971	0.0000
method	128.39	2	64.19	3.384	0.1038
Error	113.83	6	18.97		
TEMP	434.08	3	144.69	36.427	0.000
TEMP*method	75.17	6	12.53	3.154	0.0271
Error	71.50	18	3.97		



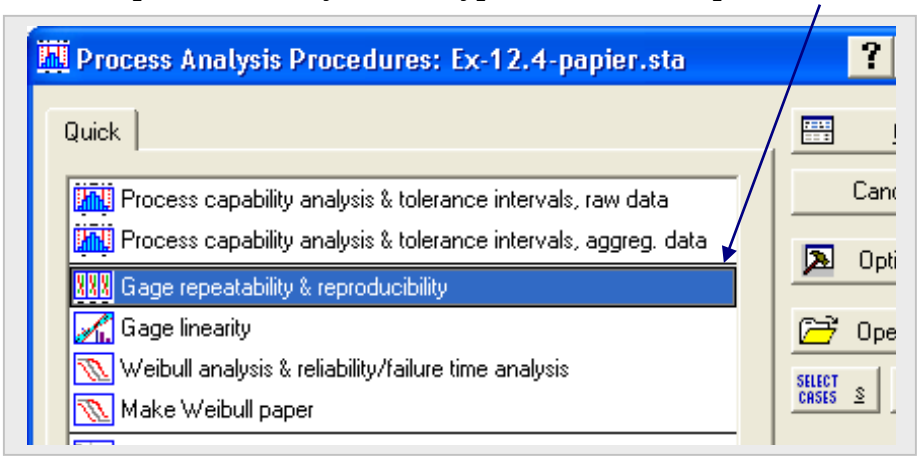
16. Facteurs emboités

Lorsque les niveaux d'un facteur B sont spécifiques aux niveaux d'un facteur A, les facteurs sont dits *emboités*. Cette structure des traitements est différente du cas de facteurs *croisés* ou toutes les combinaisons des niveaux des facteurs A et B sont présentes. Avec plus de deux facteurs, on peut avoir deux types de structures des traitements : facteurs *complètement* emboités ou facteurs *partiellement* emboités (partiellement croisés). Les *mesures répétées* constitue un cas de facteurs emboités. Il est fréquent que dans un design avec facteurs emboités que plusieurs soient des facteurs aléatoires.

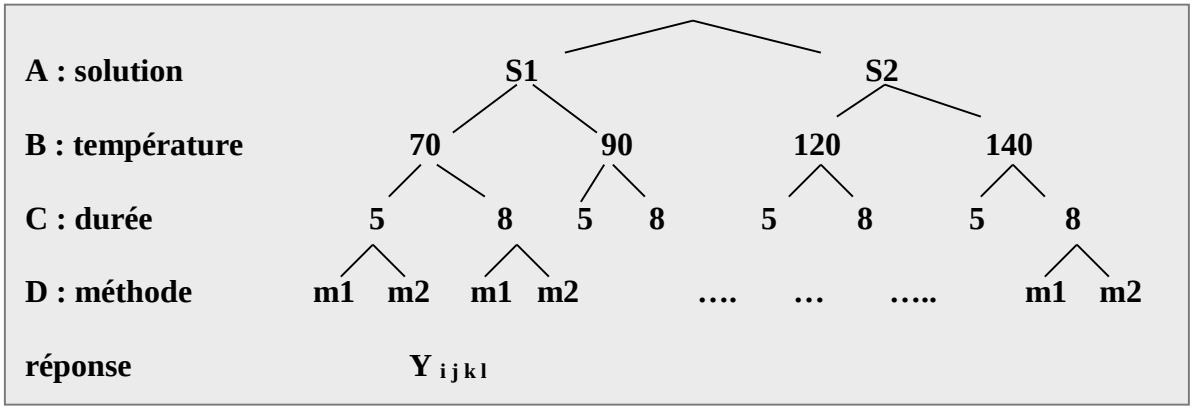
Exemple 26 : évaluation d'un processus de mesurage
étude de répétabilité (erreur appareil) et de reproductibilité (opérateur)



Le facteur PRODUIT est emboité dans le facteur LOT.
Le facteur OPÉRATEUR peut être emboité ou croisé avec le facteur PRODUIT.
Le facteur RÉPÉTITION est toujours emboité dans le facteur OPÉRATEUR.
L'objectif principal d'une telle étude est d'estimer les composants de la variance : σ^2 LOT , σ^2 PRODUIT, σ^2 OPÉRATEUR, σ^2 RÉPÉTITION.
Tous les facteurs sont aléatoires. Le plan d'échantillonnage souvent employé est
a = 1 b = 10 à 20 c = 2 ou 3 n = 2 ou 3
La conception et l'analyse de ce type d'étude est disponible avec *STATISTICA*.



Exemple 27 : 4 facteurs partiellement emboités



B emboité dans A **C croisé avec B** **D croisé avec C**

Exemple 28 : 3 facteurs complètement emboités – Kutner et all 5 ed. p. 1089

Instructeur	A		B		C		D		E		F	
	groupe		groupe		groupe		groupe		groupe		groupe	
Ville	1	2	1	2	1	2	1	2	1	2	1	2
Atlanta	25	29	14	11	x		x		x		x	
Chicago	x		x		11	6	22	18	x		x	
San Francisco	x		x		x		x		17	20	5	2

Le facteur Instructeur est emboité dans le facteur Ville.
Les groupes (cours) constituent des répétitions.

Kutner 5ed. p 1089 - instructeur est emboité dans ville

	Ville	Instructeur	groupe	Y-test
1	Atlanta	A	1	25
2	Atlanta	A	2	29
3	Atlanta	B	1	14
4	Atlanta	B	2	11
5	Chicago	C	1	11
6	Chicago	C	2	6
7	Chicago	D	1	22
8	Chicago	D	2	18
9	San Francisco	E	1	17
10	San Francisco	E	2	20
11	San Francisco	F	1	5
12	San Francisco	F	2	2

Modèle pour 2 facteurs fixes emboités

facteur A fixe avec a modalités ; facteur B fixe emboité dans A avec b modalités. plan équilibré avec n répétitions

$$Y_{ijk} = \mu + \alpha_i + \beta_{j(i)} + \varepsilon_{ijk} \qquad (204)$$

$i = 1, 2,.., a \quad j = 1, 2,\ldots, b \quad k = 1, 2,\ldots, n$

où μ : effet général

α_i : effet du facteur A et $\sum \alpha_i = 0$

$\beta_{j(i)}$: effet du facteur B et $\sum_j \beta_{j(i)} = 0$ pour tout i

ε_{ijk} : terme d'erreur distribué $N(0, \sigma^2)$

Conséquences

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

$Var(_{ijk}) = \sigma^2$

Il n'y a pas de terme d'interaction dans le modèle

Estimation des paramètres

$$\hat{\mu}_{\cdot} = \bar{Y}_{\ldots} \qquad \hat{\alpha}_i = \bar{Y}_{i..} - \bar{Y}_{\ldots} \qquad \hat{\beta}_{j(i)} = \bar{Y}_{ij.} - \bar{Y}_{\ldots} \qquad (205)$$

résidus $e_{ijk} = Y_{ijk} - \hat{Y}_{ijk} = Y_{ijk} - \bar{Y}_{ij.}$ (206)

Analyse de la variance

totale $SSTO = \sum\sum\sum (Y_{ijk} - \bar{Y}_{\ldots})^2$ (207)

facteur A $SSA = bn \sum (Y_{i..} - \bar{Y}_{\ldots})^2$ (208)

facteur B (emboité A)) $SSB(A) = n \sum\sum (Y_{ij.} - \bar{Y}_{i..})^2$ (209)

erreur $SSE = \sum\sum\sum (Y_{ijk} - \bar{Y}_{ij.})^2$ (210)

ANOVA

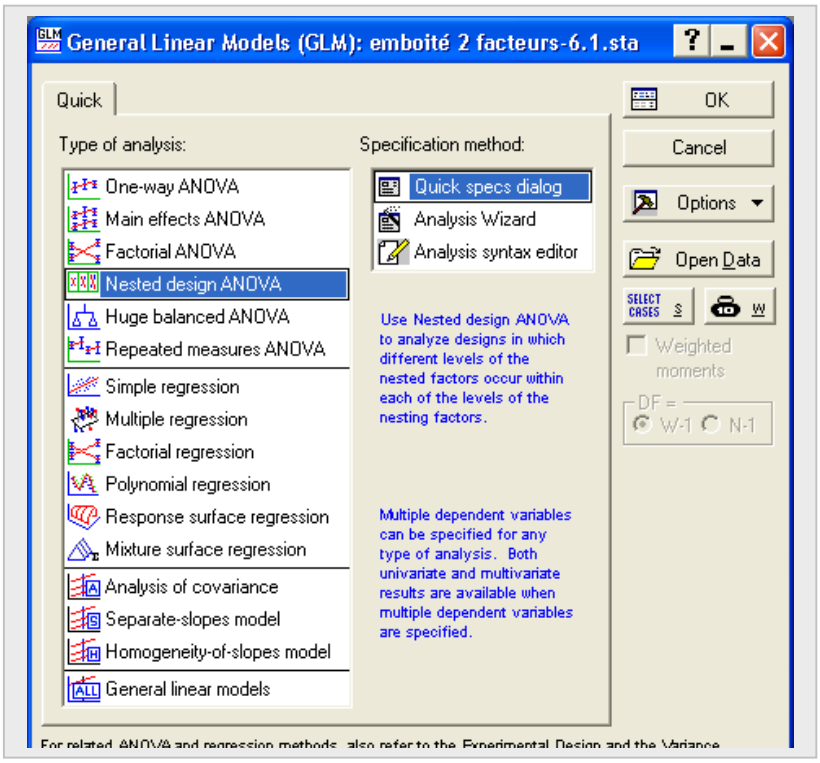
Source	SS	df	MS	E()MS)
Facteur A	SSA	a – 1	MSA	$\sigma^2 + (bn/(a-1)) \sum \alpha_i^2$
Facteur B(A)	SSB(A)	b – 1	MSB(A)	$\sigma^2 + (n/b(a-1)) \sum\sum \beta_{j(i)}^2$
Erreur	SSE	ab(n-1)	MSE	σ^2
Totale	SSTO	abn – 1	------	--------

Tests

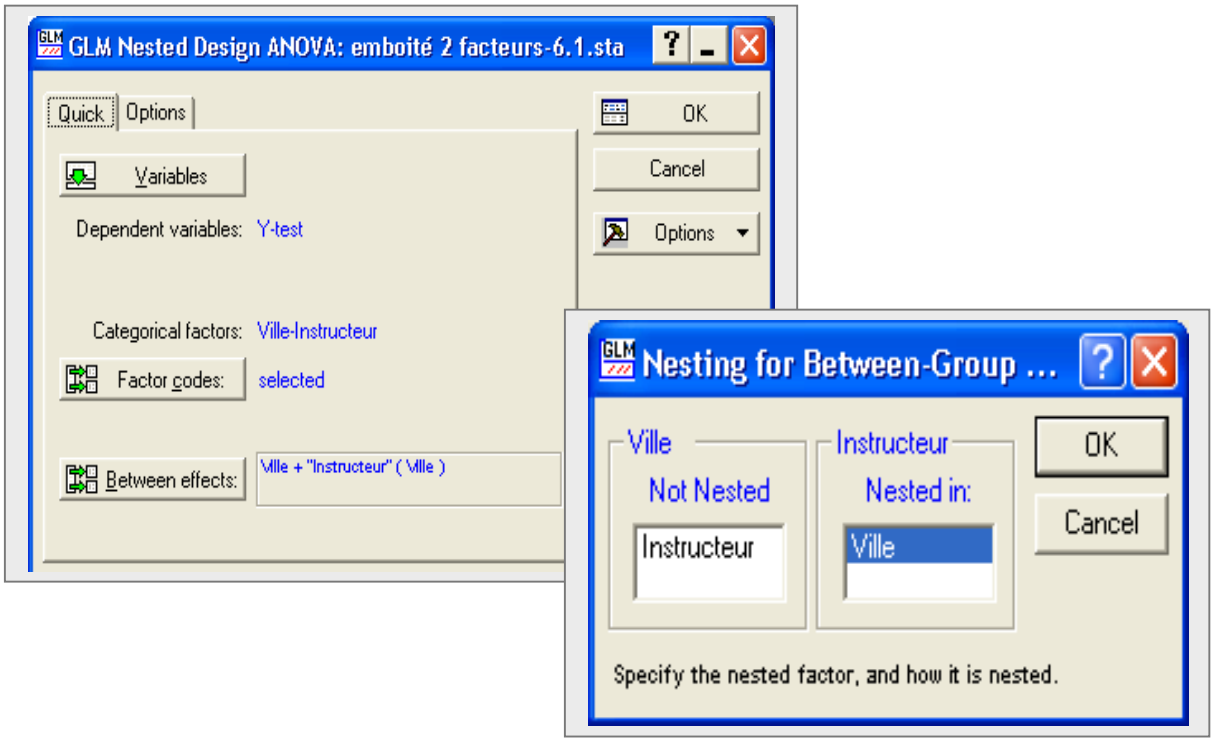
H_0 : tous les $\alpha_i = 0$	FA = MSA / MSE	$F(1 - \alpha ; a - 1 ; ab(n-1))$
H_0 : tous les $\beta_{j(i)} = 0$	FB = MSB(A) / MSE	$F(1 - \alpha ; b - 1 ; ab(n-1))$

On emploie la méthode de Tukey pour identifier les différences de moyennes significatives lorsque les hypothèses sont rejetées.

Utilisation de *STATISTICA* pour les designs avec facteurs emboités

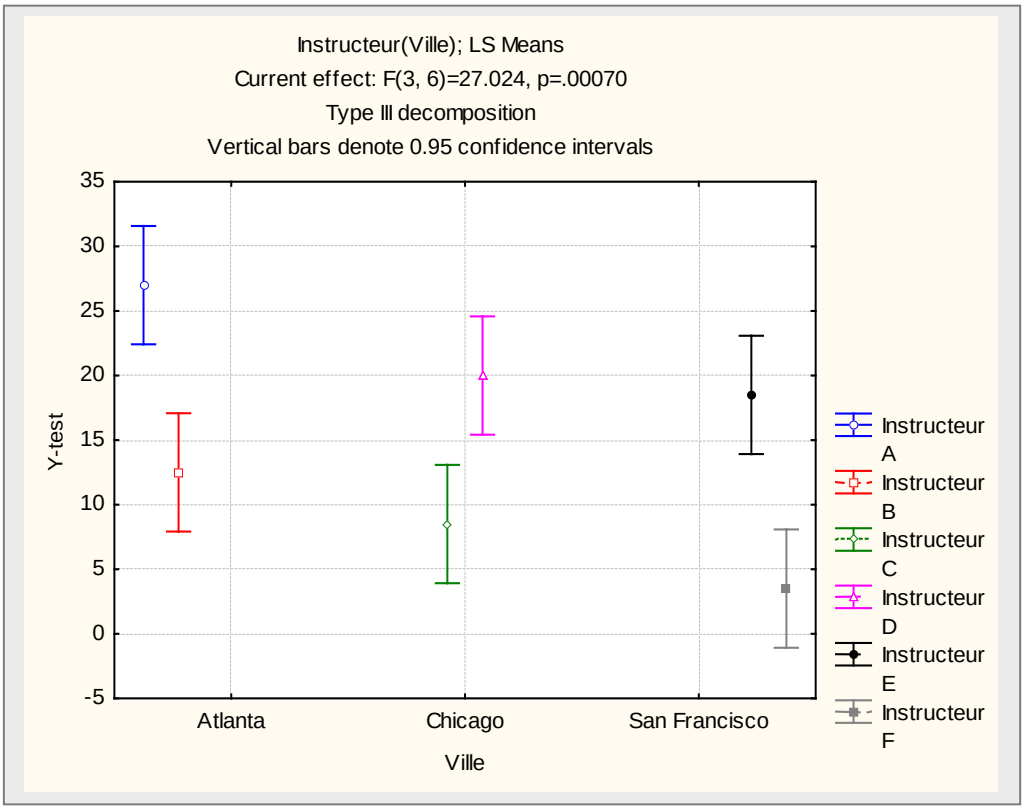


Exemple 28



ANOVA

	DF	SS	MS	F	p-value
Intercept	1	2700.00	2700.00	385.71	0.00000
Ville	2	156.50	78.25	11.18	0.00947
Instructeur(Ville)	3	567.50	189.17	27.02	0.00070
Error	6	42.00	7.00		
Total	11	766.00			



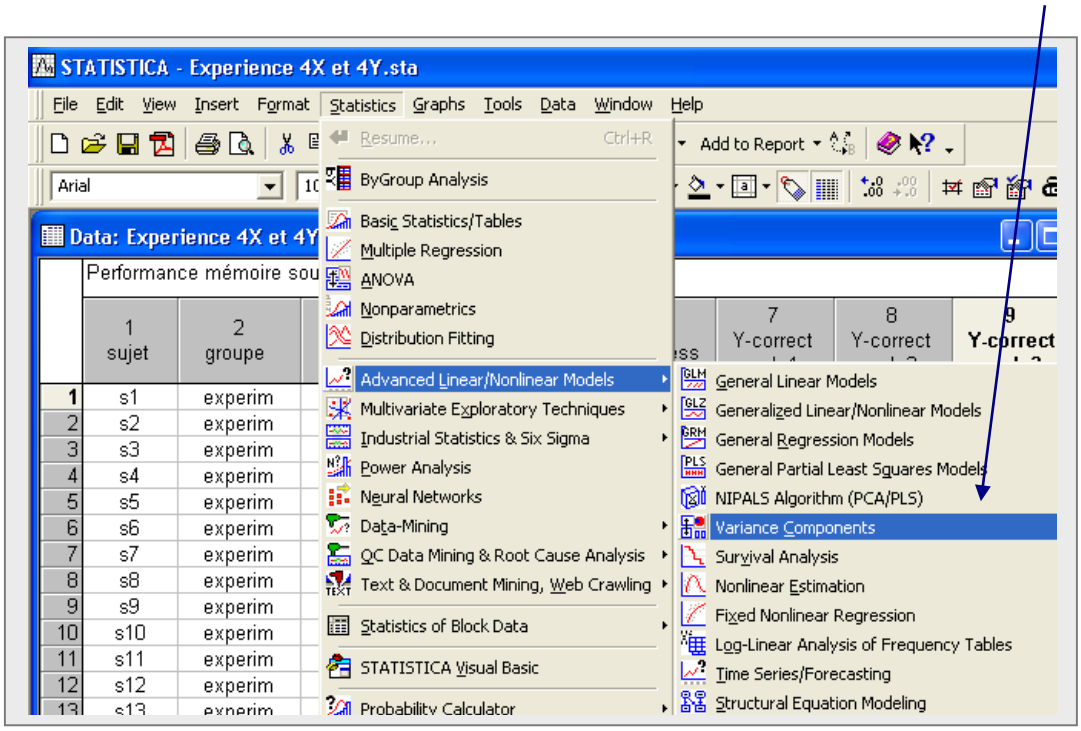
Tukey HSD test; variable Y-test (emboité 2 facteurs-6.1.sta)
Approximate Probabilities for Post Hoc Tests Error: Between MS = 7.0000, df = 6.0000

	Ville	Instructeur	{1}	{2}	{3}	{4}	{5}	{6}
1	Atlanta	A		0.0116	0.0034	0.2180	0.1156	0.0010
2	Atlanta	B	0.0116		0.6713	0.1765	0.3295	0.0937
3	Chicago	C	0.0034	0.6713		0.0342	0.0620	0.4837
4	Chicago	D	0.2180	0.1765	0.0342		0.9899	0.0061
5	San Francisco	E	0.1156	0.3295	0.0620	0.9899		0.0098
6	San Francisco	F	0.0010	0.0937	0.4837	0.0061	0.0098	

Remarques

- **Si le plan de données n'est pas équilibré (« balanced »), on analyse les données avec une approche par régression et l'utilisation de variables indicatrices.**

- **Si les facteurs A et B sont aléatoires on s'intéresse à l'estimation des composants de la variance. Il existe une procédure de *STATISTICA* pour faire cette analyse.**



17. MODÈLES AVEC FACTEURS ALÉATOIRES

18. MODÈLES AVEC FACTEURS ALÉATOIRES ET FACTEURS FIXES

19. PLANS EN BLOCS INCOMPLETS ÉQUILIBRÉS : conception et analyse

20. PLANS CROSSOVER