Traitement des données en Psychologie

UV PSY38X2

Présentation du cours 2003/2004

Organisation matérielle

Cours magistral :
mardi 16h-17h - Salle A216
Travaux dirigés :
1 groupe de TD de statistiques.
Gr. 1 : mardi 17h-18h - Salle A216
2 sous-groupes de TD en Informatique.
salle A204 ou A206, 2 h. par quinzaine
sous-gr. A - mercredi 18h-20h - Sem. A
sous-gr. B - mercredi 18h-20h - Sem. B
Monitorat informatique

Contrôle des connaissances : (contrôle continu)
70 % Examen écrit (3 heures)
30 % Evaluation de TD

Bibliographie

- G. Mialaret. Statistiques appliquées aux sciences humaines. PUF
- B. Beaufils. Statistiques appliquées à la psychologie. Tomes 1 et 2. LEXIFAC Bréal
- N. Guéguen. Manuel de statistiques pour psychologues
- D.C. Howell. Méthodes statistiques en sciences humaines
- D. Laveault, J. Grégoire. Introduction aux théories des tests en sciences humaines. De Boeck Université
- J.P. Rossi. La méthode expérimentale en Psychologie
- H. Abdi. Introduction au traitement statistique des données expérimentales. PUG
- P. Rateau. Méthode et statistique expérimentales en sciences humaines. Ellipses

Contenu

Documents fournis :

Copie des transparents en statistiques
Fiches de TD
Sites internet permettant de télécharger ces documents :
Depuis les salles de TD d'informatique :
`http ://letsamba.univ-brest.fr/~carpenti/`
Depuis le reste de l'Université (domaine univ-brest.fr) :
`http ://infolettres.univ-brest.fr/~carpenti/`
Depuis l'extérieur, ou depuis le domaine univ-brest.fr :
`http ://geai.univ-brest.fr/~carpenti/`
Voir aussi :
`http ://cours.univ-brest.fr/Discipline/informatique/carpentier/`

N.B. Documents (autres que les fiches de TD d'informatique) au format .pdf lisible par Acrobat Reader

Programmes, contrôle des connaissances, documents plus anciens :
`http ://geai.univ-brest.fr/enseignements.html`

F de Fisher. Analyse de variance à un facteur.
Introduction aux plans d'expériences.
Analyse de variance à plusieurs facteurs.
Corrélation linéaire. Régression linéaire.
Introduction à l'analyse des données multidimensionnelles :
Analyse en composantes principales.
Analyse en facteurs principaux.
Analyse factorielle des correspondances.
Analyse des correspondances multiples.
Classification ascendante hiérarchique.

Comparaison de deux variances
F de Fisher

Exemple. Deux tests mesurant la même aptitude
– Groupe 1 : 25 sujets. $\overline{x_1} = 40$; $s^2_{1c} = 65$
– Groupe 2 : 30 sujets. $\overline{x_2} = 38$; $s^2_{2c} = 30$

La précision est-elle la même pour les deux tests ?

Cas général

Deux échantillons de tailles n_1 et n_2 extraits de deux populations. Moyennes égales ou différentes. Distribution normale de la variable dans les populations parentes.

Problème : Les *variances* dans les populations parentes sont-elles égales ?

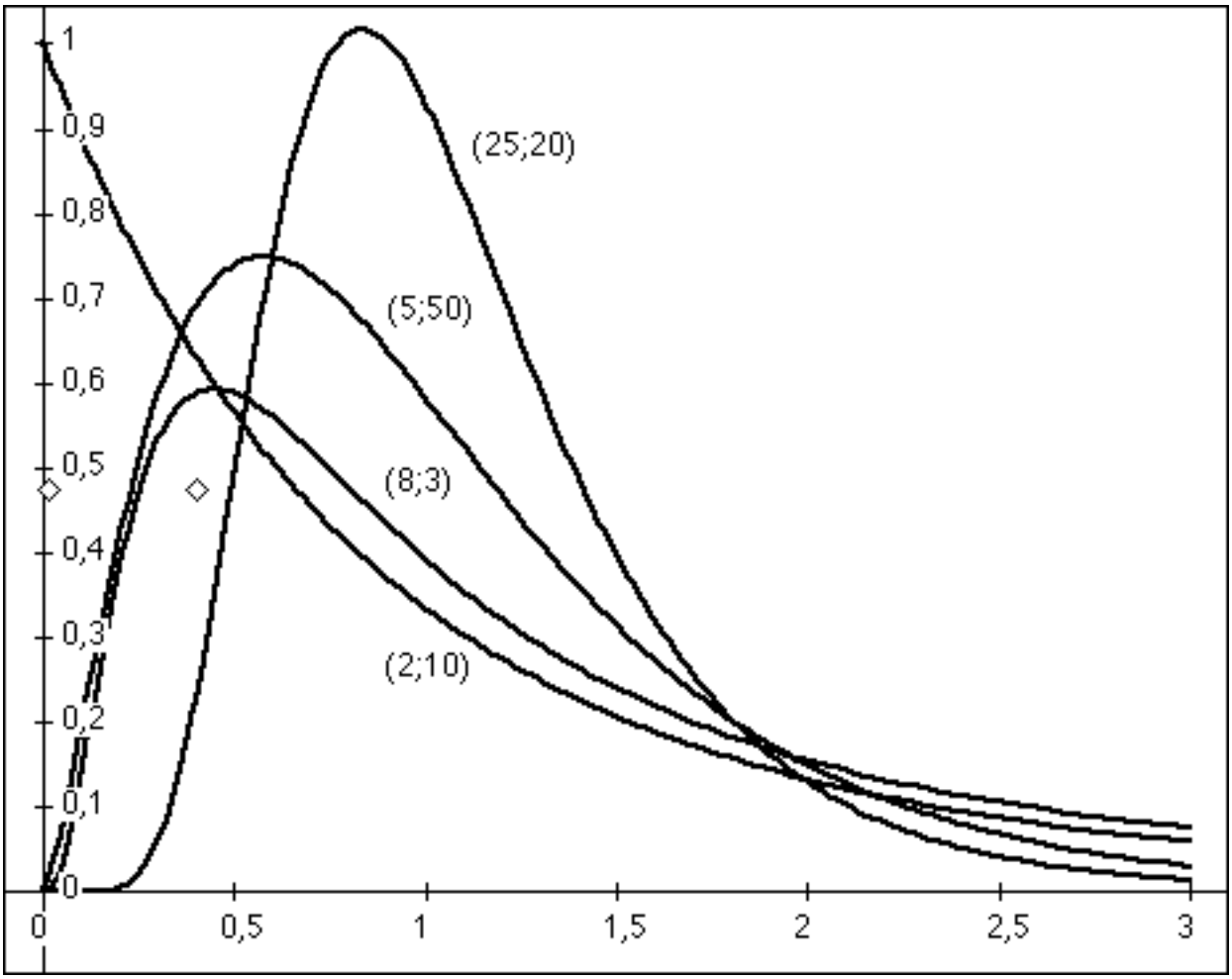
H_0 : Les variances sont égales
H_1 : La première variance est supérieure à la deuxième.

Statistique de test

$$F = \frac{s^2_{1,c}}{s^2_{2,c}}$$

F suit une **loi de Fisher à $n_1 - 1$ et $n_2 - 1$ degrés de liberté.**

Distributions du F de Fisher



Analyse de Variance à un facteur

Exemple introductif : Test commun à trois groupes d'élèves. Moyennes observées dans les trois groupes : $\overline{x}_1 = 8$, $\overline{x}_2 = 10$,$\overline{x}_3 = 12$.

Question : s'agit-il d'élèves "tirés au hasard" ou de groupes de niveau ?

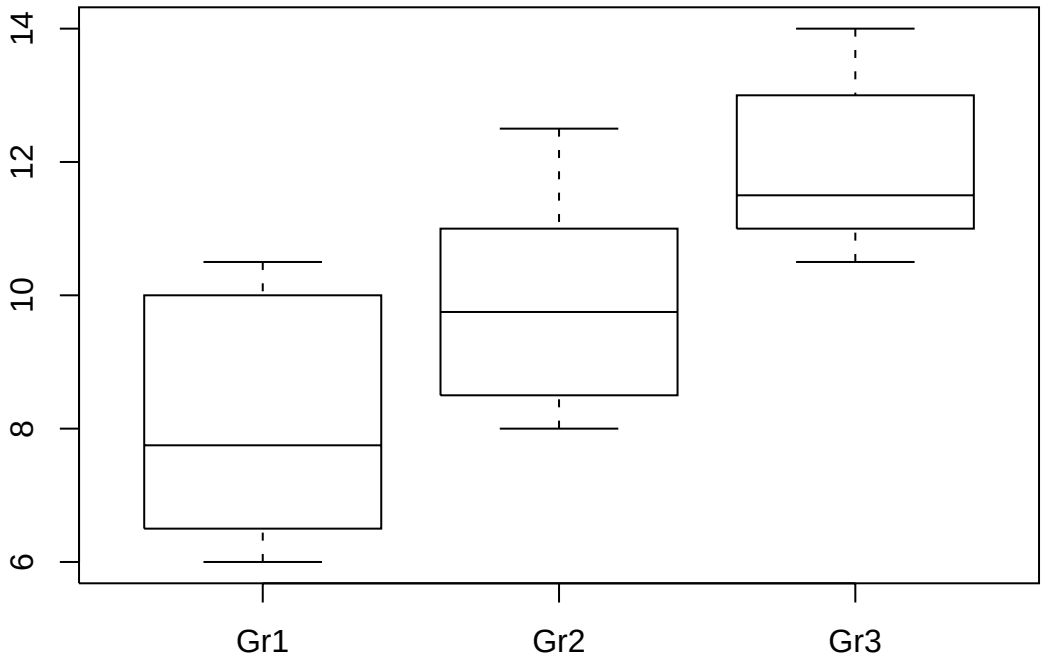
Première situation :

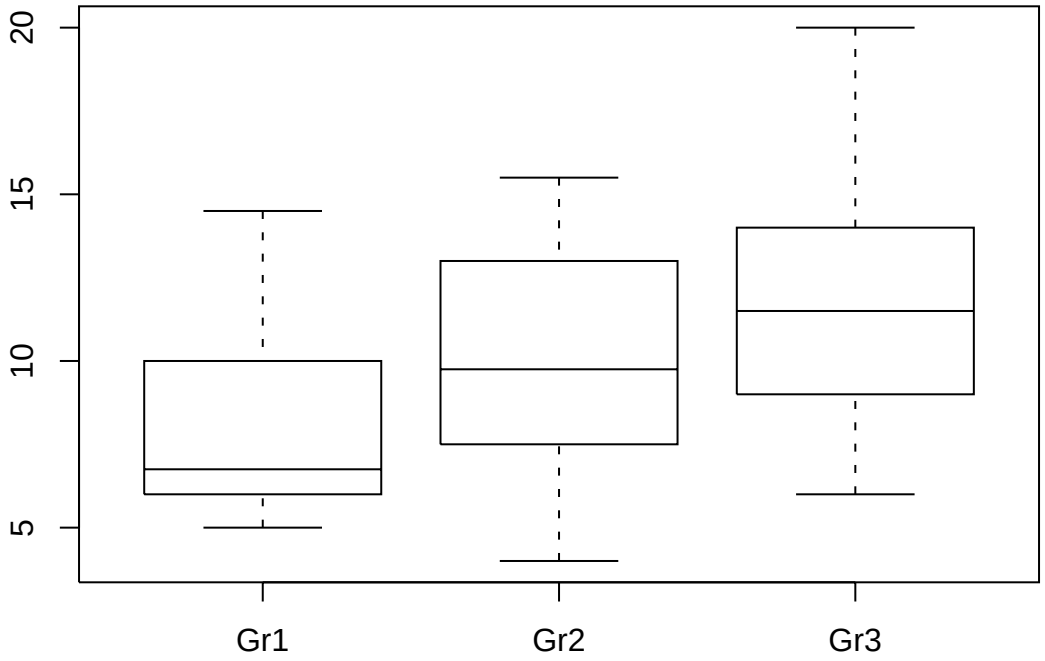
	Gr1	Gr2	Gr3
	6	8	10.5
	6.5	8.5	10.5
	6.5	8.5	11
	7	9	11
	7.5	9.5	11
	8	10	12
	8	11	13
	10	11	13
	10	12	14
	10.5	12.5	14
$\overline{x}_i$	8	10	12

Deuxième situation :

	Gr1	Gr2	Gr3
	5	4	6
	5.5	5.5	7
	6	7.5	9
	6	9	10
	6.5	9.5	11
	7	10	12
	7.5	11	13
	10	13	14
	12	15	18
	14.5	15.5	20
$\overline{x}_i$	8	10	12

Boîtes à moustaches pour les deux situations proposées





Démarche utilisée : nous comparons la dispersion des moyennes (8, 10, 12) à la dispersion à l'intérieur de chaque groupe.

Comparer a moyennes sur des groupes indépendants

Plan d'expérience : $\mathcal{S} < \mathcal{A}_a >$

Une variable $\mathcal{A}$, de modalités $A_1, A_2, \ldots, A_a$ définit a groupes indépendants.

Variable dépendante X mesurée sur chaque sujet.
x_{ij} : valeur observée sur le i-ème sujet du groupe j.

Problème : La variable X a-t-elle la même moyenne dans chacune des sous-populations dont les groupes sont issus ?

H_0 : $\mu_1 = \mu_2 = \ldots = \mu_a$
H_1 : Les moyennes ne sont pas toutes égales.

Construction de la statistique de test :

Notations :
n_1, n_2, . . ., n_a : effectifs des groupes.
N : effectif total
$T_{\cdot 1}$, . . ., $T_{\cdot a}$: sommes des observations pour chacun des groupes.
$T_{\cdot\cdot}$ ou T_G : somme de toutes les observations.

Somme des carrés totale ou variation totale :

$$SC_T = \sum_{i,j} x_{ij}^2 - \frac{T_G^2}{N}$$

Elle se décompose en une variation "intra-groupes" et une variation "inter-groupes" :

$SC_T = SC_{inter} + SC_{intra}$ avec :

$$SC_{inter} = \sum_{j=1}^{a} \frac{T_{.j}^2}{n_j} - \frac{T_G^2}{N}$$

$$SC_{intra} = \sum_{i,j} x_{ij}^2 - \sum_{j=1}^{a} \frac{T_{.j}^2}{n_j}$$

Carrés moyens :

$$CM_{inter} = \frac{SC_{inter}}{a-1} \quad ; \quad CM_{intra} = \frac{SC_{intra}}{N-a}$$

Statistique de test : $F = \dfrac{CM_{inter}}{CM_{intra}}$

F suit une loi de Fisher à $(a-1)$ et $(N-a)$ ddl.

Présentation des résultats

Source de variation	SC	ddl	CM	F
$\mathcal{A}$ (inter-groupes)	SC_{inter}	$a-1$	CM_{inter}	F_{obs}
Résiduelle (intra-gr.)	SC_{intra}	$N-a$	CM_{intra}	
Total	SC_T	$N-1$		

Organisation des calculs

i j	1	2	3	Total
1 ...	x_{11} ...	...	...	
$T_{\cdot j}$	$T_{\cdot 1}$			T_G
$T_{\cdot j}^2$				
n_j				N
$\frac{T_{\cdot j}^2}{n_j}$				
$\sum x_{ij}^2$				

Exemple :

15 sujets évaluent 3 couvertures de magazine. Sont-elles équivalentes ?

	C1	C2	C3
	14	16	14
	6	14	16
	12	8	14
	10	8	14
	8	14	12
$\overline{x}_i$	10	12	14

Calculs

i j	1	2	3	Total
1	$x_{11} = 14$	16	14	
2	$x_{21} = 6$	14	16	
...	...	...	...	
$T_{.j}$	50	60	70	$T_G = 180$
$T_{.j}^2$	2500	3600	4900	
n_j	5	5	5	$N = 15$
$\frac{T_{.j}^2}{n_j}$	500	720	980	2200
$\sum x_{ij}^2$	540	776	988	2304

$$SC_{inter} = \sum_{j=1}^{a} \frac{T_{.j}^2}{n_j} - \frac{T_G^2}{N} = 2200 - \frac{180^2}{15} = 40$$

$$SC_{intra} = \sum_{i,j} x_{ij}^2 - \sum_{j=1}^{a} \frac{T_{.j}^2}{n_j} = 2304 - 2200 = 104$$

$$SC_T = \sum_{i,j} x_{ij}^2 - \frac{T_G^2}{N} = 144$$

$$CM_{inter} = \frac{SC_{inter}}{a-1} = 20 \ ; \ CM_{intra} = \frac{SC_{intra}}{N-a} = 8.67$$

$$F_{obs} = \frac{CM_{inter}}{CM_{intra}} = 2.31$$

F suit une loi de Fisher avec $ddl_1 = a - 1 = 2$ et $ddl_2 = N - a = 12$.

Résultats

Source	Somme carrés	ddl	Carré Moyen	F
$\mathcal{C}$	40	2	20	2.31
Résid.	104	12	8.67	
Total	144	14		

Pour α=5%, $F_{crit} = 3.88$: H_0 est acceptée

Remarques

– SC_{inter} : c'est la somme des carrés (totale) que l'on obtiendrait si toutes les observations d'un groupe étaient égales à la moyenne de ce groupe.
CM_{inter} : variance corrigée de cet ensemble de données.

– SC_{intra} : c'est la somme des carrés (totale) que l'on obtiendrait en "décalant" chaque observation de façon à avoir la même moyenne dans chaque groupe.
CM_{intra} : "moyenne pondérée" des trois variances corrigées ainsi obtenues.

– Hypothèses "a priori" :
distribution normale de X dans chacun des groupes
Egalité des variances dans les trois populations.

– Si 2 groupes, équivaut à un T de Student. $F = T^2$

Pour les deux situations proposées en introduction :

Situation 1

Analysis of Variance Table

Response : x1

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
group	2	80.000	40.000	17.008	1.659e-05 ***
Residuals	27	63.500	2.352		

Situation 2

Analysis of Variance Table

Response : x2

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
group	2	80.00	40.00	2.7136	0.08436 .
Residuals	27	398.00	14.74		

Vocabulaire des plans d'expérience

Variable dépendante

– On formule une hypothèse : "telle variable a tel effet sur le comportement des sujets"

– On choisit une **variable dépendante**

Définir une variable mesurable (numérique) caractérisant le comportement du sujet.

Qualités d'une bonne variable dépendante : *pertinence, sensibilité.*

Variables indépendantes ou facteurs

– Recherche des **variables indépendantes** ou **facteurs de variation**

Indépendant : indépendant du sujet, manipulé ou contrôlé par l'expérimentateur

Causes susceptibles d'entraîner une variation de la variable dépendante.

Les facteurs sont des variables nominales ou ordinales. Les valeurs prises par un facteur sont ses *modalités* ou *niveaux*.

Les *Facteurs principaux* sont ceux dont on désire étudier l'effet. Ils sont aussi appelés *facteurs d'intérêt*.

Les *Facteurs secondaires* sont les autres causes susceptibles d'influer sur le comportement des sujets. Deux manières de les prendre en compte :

– Contrôle
– Neutralisation.

Facteur systématique ou fixe : l'ensemble des modalités possibles est fini (et petit). Toutes les modalités sont présentes dans l'expérience.

Facteur aléatoire : l'ensemble des modalités est grand (infini). On choisit alors (par tirage au sort) un ensemble de modalités.

Le facteur sujet : généralement, c'est un facteur aléatoire et secondaire. Il est souvent assimilé à une incertitude sur une mesure.

Facteur étiquette : par exemple, le sexe, ou le milieu socio-culturel.

Groupe contrôle

Souvent, il existe un niveau particulier ou "état nul" de la VI. Un groupe de sujets soumis à ce niveau de la VI constitue un *groupe contrôle*.

Permet notamment de contrôler ou de mettre en évidence d'éventuelles *variables parasites*.

Interaction entre facteurs

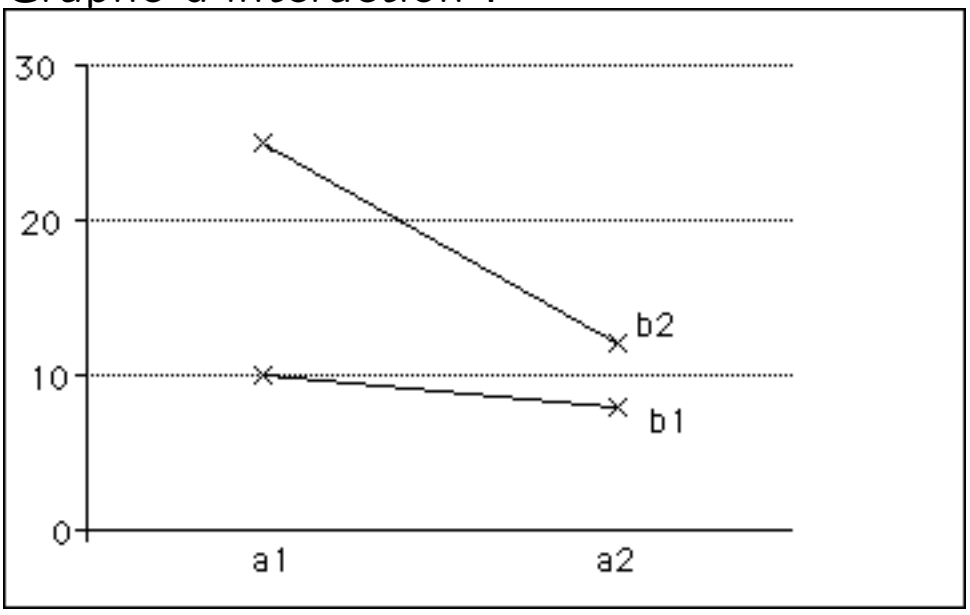
Exemple : Mémorisation d'une liste de mots.
VD : nombre de mots mémorisés

Facteur $\mathcal{A}$: a_1 sujets normaux, a_2 déficit mnésique
Facteur $\mathcal{B}$: b_1 liste de 12 mots, b_2 30 mots

Moyennes observées sur 4 groupes indépendants :

	b_1	b_2
a_1	10	25
a_2	8	12

Graphe d'interaction :



Effet simple d'un facteur $\mathcal{A}$: effet observé de $\mathcal{A}$ lorsque les autres facteurs sont fixés.

Effet principal d'un facteur $\mathcal{A}$: effet observé de $\mathcal{A}$, sans tenir compte des autres conditions.

Définition et écriture d'un plan d'expérience

En général, plusieurs facteurs, avec interaction. Donc : étude simultanée.

Plan factoriel :

Un plan factoriel est un plan dans lequel chaque modalité d'un facteur est combinée avec chaque combinaison de modalités des autres facteurs.

Plan en carré latin

Exemple : 3 facteurs comportant le même nombre de modalités.

On croise les deux premiers facteurs. Les modalités du troisième sont distribuées de façon à réaliser des permutations sur les lignes et les colonnes.

Exemple.

$a_1b_1c_1$	$a_1b_2c_2$	$a_1b_3c_3$
$a_2b_1c_2$	$a_2b_2c_3$	$a_2b_3c_1$
$a_3b_1c_3$	$a_3b_2c_1$	$a_3b_3c_2$

Plans quasi-complets (Rouanet - Lépine 1976)

Croisement : deux ou plusieurs facteurs sont croisés si chaque niveau de l'un des facteurs est combiné avec chaque niveau de chacun des autres facteurs.
Notation : $\mathcal{A}_3 * \mathcal{B}_5$ par exemple.

Emboîtement : Un facteur $\mathcal{A}$ est emboîté dans un facteur $\mathcal{B}$ si chaque niveau de $\mathcal{A}$ est combiné avec un seul niveau de $\mathcal{B}$.
Notation : $\mathcal{A} < \mathcal{B} >$

Emboîtement équilibré : pour chaque niveau du facteur emboîtant, on a le même nombre de niveaux du facteur emboîté.

Les plans *quasi-complets* sont les plans qui peuvent être décrits à l'aide de relations de croisement et d'emboîtement.

Définition :

Un plan est dit *quasi-complet* s'il possède les deux propriétés suivantes :

– Tous les facteurs croisés deux à deux sont croisés ou emboîtés

– Les facteurs croisés deux à deux sont croisés dans leur ensemble.

Exemple : $\mathcal{S}_4 < \mathcal{A}_2 > * \mathcal{B}_2$

Lorsque le facteur *sujet* est croisé avec d'autres facteurs : *plan à mesures répétées*

Dériver un plan d'expérience
Déterminer les sources de variation

1. Ecrire la formule du plan en termes de $*$ et $<>$

Exemple : $\mathcal{S} < \mathcal{A} > * \mathcal{B} * \mathcal{C}$

2. Ecrire les facteurs élémentaires, avec la règle suivante :
Lorsqu'un facteur est emboîté, le facteur élémentaire est accompagné de l'ensemble des facteurs emboîtants, entre parenthèses.

Exemple : $\mathcal{A}$, $\mathcal{B}$, $\mathcal{C}$, $\mathcal{S}(\mathcal{A})$

3. Termes d'interaction : autant d'étapes que de symboles $*$ dans le plan.

Etape 1 : Interactions d'ordre 1. Croiser tous les facteurs élémentaires deux à deux. Rassembler les termes entre parenthèses. Supprimer le terme d'interaction si un facteur y apparaît deux fois.

Exemple : $\mathcal{AB}$, $\mathcal{AC}$, $\mathcal{BC}$, $\mathcal{BS}(\mathcal{A})$, $\mathcal{CS}(\mathcal{A})$

Etape 2 : Interactions d'ordre 2. Croiser les facteurs élémentaires et les termes d'interaction précédents, en appliquant la même règle

Exemple : $\mathcal{ABC}$, $\mathcal{BCS}(\mathcal{A})$

Ainsi, pour l'exemple proposé : 11 sources de variation.

Modèle de score

Hypothèse : additivité des effets.

A chaque source de variation correspond un effet. Si le plan comporte des mesures répétées, il y a aussi un terme d'erreur ; sinon, c'est le terme d'interaction comportant toutes les lettres du plan qui joue ce rôle.

Exemple 1. Pour un plan $\mathcal{S} < \mathcal{A} >$:

$$Y_{as} = \mu + \alpha_a + e_{s(a)}$$

Exemple 2. Pour un plan $\mathcal{S} * \mathcal{A}$

$$Y_{as} = \mu + \alpha_a + s_s + \alpha s_{as} + e_{as}$$

Analyse de variance à plusieurs facteurs

Plan $\mathcal{S}_n * \mathcal{A}_a$

Notations
a : nombre de conditions expérimentales.
n : nombre de sujets.
x_{ij} : valeur de la VD pour le i-ième individu dans la condition expérimentale j.

Hypothèses du test
H_0 : Dans la population parente, les moyennes correspondant aux a conditions expérimentales sont égales.
H_1 : Les moyennes sont différentes.

Présentation des résultats

Source	S. carrés	ddl	C. moyen	F
$\mathcal{A}$	SC_A	$a-1$	CM_A	$\frac{CM_A}{CM_{AS}}$
$\mathcal{S}$	SC_S	$n-1$	CM_S	
Résid.	SC_{AS}	$(n-1)(a-1)$	CM_{AS}	
Total	SC_T	$N-1$		

Les carrés moyens sont obtenus en divisant la somme des carrés de la ligne par le nombre de ddl de la même ligne.
F suit une loi de Fisher-Snedecor à $a-1$ et $(n-1)(a-1)$ degrés de liberté.

Exemple : Effet du bruit sur la discrimination perceptive
Facteur : bruit (3 niveaux)
VD : nombre d'erreurs commises.

Sujets	Absence	Intermittent	Continu
1	117	119	127
2	130	126	131
3	122	118	129
4	123	117	134
5	126	120	137
6	116	120	128

Source	S. carrés	ddl	C. moyen	F
Bruit	403.11	2	201.56	19.98 **
Sujets	164.44	5	32.89	
Résid.	100.89	10	10.09	
Total	668.44	17		

Plan $\mathcal{S} < \mathcal{A}_a * \mathcal{B}_b >$

$\mathcal{A}$ et $\mathcal{B}$: facteurs fixes.

Notations
a, b, n, x_{ijk}, N

Interaction entre les facteurs $\mathcal{A}$ et $\mathcal{B}$

Tableau d'analyse de variance

Source	S. carrés	ddl	C. moyen	F
$\mathcal{A}$	SC_A	$a-1$	CM_A	$\frac{CM_A}{CM_{S(AB)}}$
$\mathcal{B}$	SC_B	$b-1$	CM_B	$\frac{CM_B}{CM_{S(AB)}}$
$\mathcal{AB}$	SC_{AB}	$(a-1)(b-1)$	CM_{AB}	$\frac{CM_{AB}}{CM_{S(AB)}}$
Résid.	$SC_{S(AB)}$	$ab(n-1)$	$CM_{S(AB)}$	
Total	SC_T	$N-1$		

Comme précédemment, chaque carré moyen est calculé en divisant la somme des carrés de la ligne par le nombre de ddl correspondant.

Les statistiques F_A, F_B, F_{AB} suivent des lois de Fisher Snedecor, avec des nombres de ddl différents. Le nombre de degrés de liberté du numérateur est respectivement $(a-1)$, $(b-1)$ et $(a-1)(b-1)$. Celui du dénominateur est $ab(n-1)$.

Exemple : facteurs : sexe, statut socio-économique.
VD : mesure du "locus of control"

	statut socio-économique		
	Bas	Moyen	Elevé
Hommes	10	16	18
	12	12	14
	8	19	17
	14	17	13
	10	15	19
	16	11	15
	15	14	22
	13	10	20
Femmes	8	14	12
	10	10	18
	7	13	14
	9	9	21
	12	17	19
	5	15	17
	8	12	13
	7	8	16

Sources de var.	ddl	SC	CM	F
Sexe	1	65.33	65.33	7.73**
Statut soc-éco	2	338.67	169.33	20.03**
$X \times C$	2	18.67	9.33	1.10 NS
Résidu	42	355.0	8.45	
Total	47	777.67		

Conclusion : les deux facteurs (sexe et statut socio-économique) ont des effets significatifs. En revanche, on n'a pas observé d'interaction entre ces facteurs.

Plan $\mathcal{S} < \mathcal{A}_a > * \mathcal{B}_b$

Plan à mesures partiellement répétées ou plan split-plot

$\mathcal{A}$ et $\mathcal{B}$: facteurs fixes.

Notations
a, b, n, x_{ijk}

Présentation des résultats

Source	S. carrés	ddl	C. moyen	F
	Entre les sujets			
$\mathcal{A}$	SC_A	$a-1$	CM_A	$\dfrac{CM_A}{CM_{S(A)}}$
$\mathcal{S}(\mathcal{A})$	$SC_{S(A)}$	$a(n-1)$	$CM_{S(A)}$	
	Dans les sujets			
$\mathcal{B}$	SC_B	$b-1$	CM_B	$\dfrac{CM_B}{CM_{BS(A)}}$
Int. $\mathcal{AB}$	SC_{AB}	$(a-1)(b-1)$	CM_{AB}	$\dfrac{CM_{AB}}{CM_{BS(A)}}$
Résid.	$SC_{BS(A)}$	$a(n-1)(b-1)$	$CM_{BS(A)}$	
Total	SC_T	$N-1$		

F_A : loi de Fisher à $a-1$ et $a(n-1)$ ddl
F_B : loi de Fisher à $b-1$ et $a(n-1)(b-1)$ ddl
F_{AB} : loi de Fisher à $(a-1)(b-1)$ et $a(n-1)(b-1)$ ddl

Exemple : Expérimentation de Bahrick (reconnaissance de portraits)

Facteurs : sexe du sujet, sexe du portrait
VD : nombre de portraits reconnus

Nom du sujet	Portrait masculin	Portrait féminin
Albert	6	6
Henri	6	6
Jules	5	5
Paul	5	5
Octave	5	6
Albertine	6	8
Henriette	7	8
Julie	6	6
Paule	7	7
Octavie	6	6

Source	S. carrés	ddl	C. moyen	F
	Entre les sujets			
$\mathcal{X}$	7.2	1	7.2	10.28*
$\mathcal{S}(\mathcal{X})$	5.6	8	0.7	
	Dans les sujets			
$\mathcal{P}$	0.8	1	0.8	3.2 NS
Int. $\mathcal{XP}$	0.2	1	0.2	0.8 NS
Résid.	2	8	0.25	
Total	15.8	19		

Remarques et conclusion

Modèle basé sur l'hypothèse d'additivité des effets

Conditions théoriques d'application de la méthode :

– Normalité de la VD dans les populations parentes

– Egalité des variances dans les populations parentes

Tests permettant de vérifier que les conditions sont remplies :

– Test de normalité de Lilliefors ou test d'Anderson Darling

– Tests de O'Brien ou de Bartlett sur les variances

La méthode est robuste : elle fournit des résultats corrects, même si les conditions ne sont qu'approximativement vérifiées.

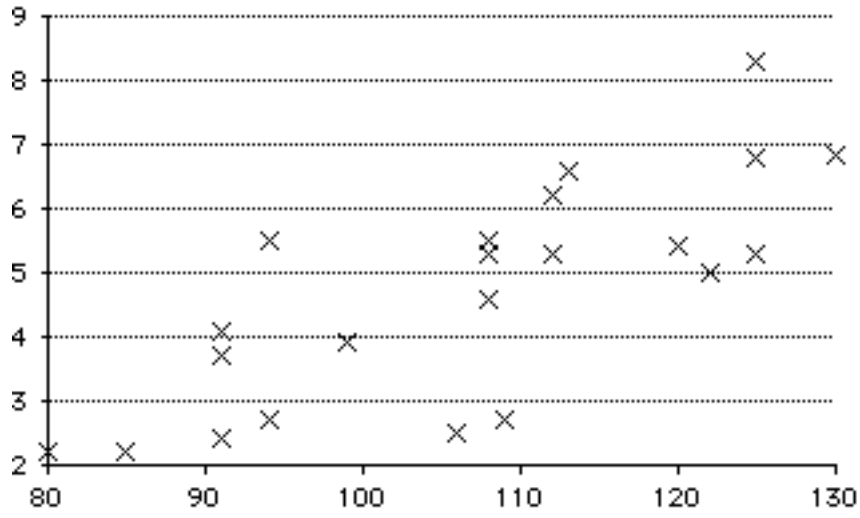
Il existe également des méthodes non paramétriques : travail sur des rangs (test de Kruskal-Wallis)

Corrélation linéaire

Données :

	X	Y
s_1	x_1	y_1
s_2	x_2	y_2
...	...	...

Nuage de points : points (x_i, y_i)



Covariance des variables X et Y

$$Cov(X,Y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})$$

ou

$$Cov(X,Y) = \left(\frac{1}{n}\sum_{i=1}^{n} x_i y_i\right) - \overline{x}\ \overline{y}$$

Coefficient de corrélation de Bravais Pearson

$$r = \frac{Cov(X,Y)}{s(X)s(Y)}$$

Remarques

– Formules analogues avec covariance et écarts types corrigés. La valeur de r est la même dans les deux cas.

– Il existe des relations non linéaires

– Corrélation n'est pas causalité

Significativité du coefficient de corrélation

- Les données (x_i, y_i) constituent un échantillon
- r est une statistique
- ρ : coefficient de corrélation sur la population

H_0 : Indépendance sur la population ; $\rho = 0$
H_1 : $\rho \neq 0$ (bilatéral) ou $\rho > 0$ ou $\rho < 0$ (unilatéral)

Statistique de test

– Petits échantillons : tables spécifiques. $ddl = n - 2$
– Grands échantillons :

$$T = \sqrt{n-2}\ \frac{r}{\sqrt{1-r^2}}$$

T suit une loi de Student à $n - 2$ degrés de liberté.

Régression linéaire

Rôle "explicatif" de l'une des variables par rapport à l'autre. Les variations de Y peuvent-elles (au moins en partie) être expliquées par celles de X ? Peuvent-elles être prédites par celles de X ?

Modèle permettant d'estimer Y connaissant X

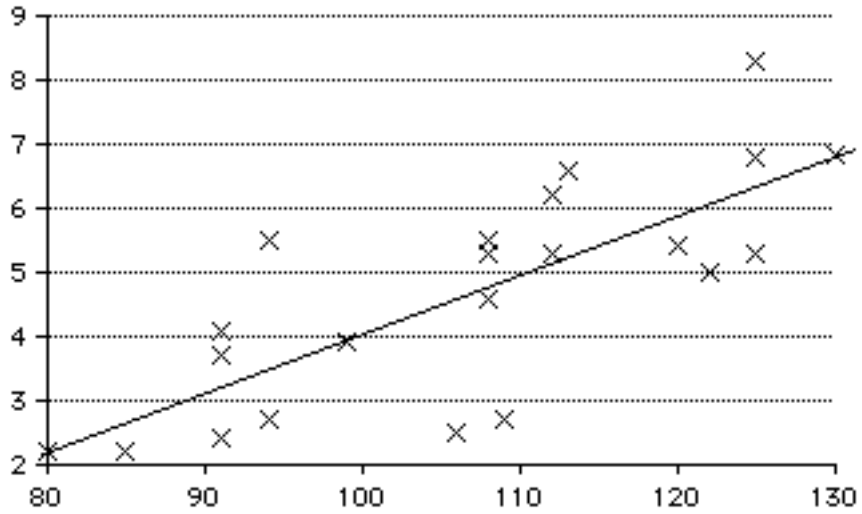
Droite de régression de Y par rapport à X :

La droite de régression de Y par rapport à X a pour équation :

$$y = ax + b$$

avec :

$$a = \frac{Cov(X,Y)}{s^2(X)} \quad ; \quad b = \overline{Y} - a\overline{X}$$



Comparaison des valeurs observées et des valeurs estimées

Valeurs estimées : $\widehat{y}_i = ax_i + b$: variable $\widehat{Y}$
Erreur (ou résidu) : $e_i = y_i - \widehat{y}_i$: variable E

Les variables $\widehat{Y}$ et E sont indépendantes et on montre que :

$$s^2(Y) = s^2(\widehat{Y}) + s^2(E)$$

avec :

$$\frac{s^2(E)}{s^2(Y)} = 1 - r^2 \quad ; \quad \frac{s^2(\widehat{Y})}{s^2(Y)} = r^2$$

$s^2(\widehat{Y})$: variance *expliquée* (par la variation de X, par le modèle)

$s^2(E)$: variance *perdue* ou *résiduelle*

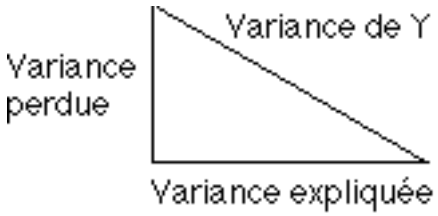
r^2 : part de la variance de Y qui est expliquée par la variance de X. r^2 est appelé *coefficient de détermination*.

Exemple : $r = 0.86$

$r^2 = 0.75$; $1 - r^2 = 0.25$; $\sqrt{1 - r^2} = 0.5$.

• La part de la variance de Y expliquée par la variation de X est de 75%.

• L'écart type des résidus est la moitié de l'écart type de Y.



Remarque : test du coefficient de corrélation

Rappel

Valeurs estimées : $\widehat{y}_i = ax_i + b$
Erreur (ou résidu) : $e_i = y_i - \widehat{y}_i$

$$s^2(Y) = s^2(\widehat{Y}) + s^2(E)$$

avec :

$$\frac{s^2(E)}{s^2(Y)} = 1 - r^2 \quad ; \quad \frac{s^2(\widehat{Y})}{s^2(Y)} = r^2$$

La plupart des logiciels de statistiques utilisent une analyse de variance pour tester la significativité du coefficient de corrélation.

Test du coefficient de corrélation à l'aide d'une analyse de variance

Source	SC	ddl	CM	F
Modèle	$ns^2(\widehat{Y})$	1	CM_1	F_{obs}
Résiduelle	$ns^2(E)$	$n-2$	CM_2	
Total	$ns^2(Y)$	$n-1$		

$F_{obs} = \dfrac{CM_1}{CM_2} = (n-2)\dfrac{r^2}{1-r^2}$ suit une loi de Fisher à 1 et $n-2$ ddl.
On retrouve : $F_{obs} = T_{obs}^2$

Régression linéaire multiple

Position du problème

Une population (ou un échantillon) sur laquelle on a observé un ensemble de variables numériques.

	X_1	X_2	...	X_p
s_1	x_{11}	x_{12}	...	x_{1p}
...	...	...	...	...

Exemple avec trois variables

x_i : âge de la mère
y_i : rang de l'enfant dans la fratrie
z_i : poids de l'enfant à la naissance
n : nombre d'observations (ici : $n = 200$)

	X	Y	Z
s_1	26.5	1	2100
...	...	...	...
s_{200}	34.5	2	4500

Nuage de points

Pour trois variables : représentation dans l'espace.

Pour plus de trois variables, détermination des directions de "plus grande dispersion du nuage" : analyse en composantes principales.

Paramètres associés aux données

Matrice des covariances, matrice des corrélations.

Exemple : *Coefficients de corrélation des variables prises 2 à 2 :*

$$\begin{array}{c} \\ X \\ Y \\ Z \end{array} \begin{bmatrix} X & Y & Z \\ 1 & 0.60 & 0.24 \\ 0.60 & 1 & 0.28 \\ 0.24 & 0.28 & 1 \end{bmatrix}$$

$r_{xz} = 0.24$ $**$: âge et poids sont corrélés
$r_{yz} = 0.28$ $**$: rang et poids sont corrélés
$r_{xy} = 0.60$ $**$: rang et âge sont fortement corrélés

"Hyperplan" de régression

L'une des variables (Z) est la variable "à prévoir". Les autres ($X_1, X_2, \ldots, X_p$) sont les variables "prédicatives".

$$Z = a_0 + a_1 X_1 + \ldots + a_p X_p$$

Avec trois variables :

$$Z = c + aX + bY$$

Passe par le point moyen, c'est-à-dire : $c = \overline{Z} - a\overline{X} - b\overline{Y}$

Coefficient de corrélation multiple

$\hat{Z}$: valeurs estimées à l'aide de l'équation précédente.

$$R = r_{Z\hat{Z}} = \frac{Cov(Z, \hat{Z})}{s(Z)s(\hat{Z})}$$

Dans l'exemple proposé : $R = 0.29$

Comme précédemment, R^2 est la part de la variance "expliquée par le modèle".

Coefficients de corrélation partielle

Corrélations obtenues en contrôlant la troisième variable. Pour calculer $r_{yz.x}$, par exemple :
– On calcule les résidus de la régression de Z par rapport à X
– On calcule les résidus de la régression de Y par rapport à X
– On calcule le coefficient de corrélation entre les deux séries obtenues.

$r_{yz.x} = 0.18$ $**$: A âge constant, rangs et poids sont corrélés
$r_{xz,y} = 0.09$ NS : A rang constant, pas de corrélation entre âge et poids.

Seul le rang de naissance intervient. L'âge de la mère n'est lié au poids de l'enfant que par le rang de naissance.

Analyse en Composantes Principales

Position du problème :
On a observé p variables sur n individus : protocole multivarié.

On cherche à remplacer ces p variables par q nouvelles variables résumant au mieux le protocole, avec $q <= p$, et si possible $q = 2$.

Mini-exemple : 6 sujets décrits par 4 variables :

Données :

Suj	X_1	X_2	X_3	X_4
s1	-11	-60	110	40
s2	-12	-62	93	25
s3	-15	-80	113	39
s4	-14	-75	94	25
s5	-14.5	-82	100	30
s6	-13	-72	102	32

Corrélations des variables prises deux à deux :

	X_1	X_2	X_3	X_4
X_1	1.0000	0.9701	-0.0635	0.0940
X_2	0.9701	1.0000	-0.1018	0.0373
X_3	-0.0635	-0.1018	1.0000	0.9856
X_4	0.0940	0.0373	0.9856	1.0000

Nuage de points obtenu en prenant 2 variables, 3 variables . . .

Données centrées réduites (sans correction ou avec correction) :

Suj	Z_1	Z_2	Z_3	Z_4
s1	1.5993	1.4197	1.0722	1.3648
s2	0.8885	1.1798	-1.2063	-1.1420
s3	-1.2439	-0.9798	1.4743	1.1977
s4	-0.5331	-0.3799	-1.0722	-1.1420
s5	-0.8885	-1.2198	-0.2681	-0.3064
s6	0.1777	-0.0200	0.0000	0.0279

Suj	Z_{1c}	Z_{2c}	Z_{3c}	Z_{4c}
s1	1.4600	1.2960	0.9788	1.2459
s2	0.8111	1.0770	-1.1012	-1.0425
s3	-1.1356	-0.8944	1.3459	1.0933
s4	-0.4867	-0.3468	-0.9788	-1.0425
s5	-0.8111	-1.1135	-0.2447	-0.2797
s6	0.1622	-0.0183	0.0000	0.0254

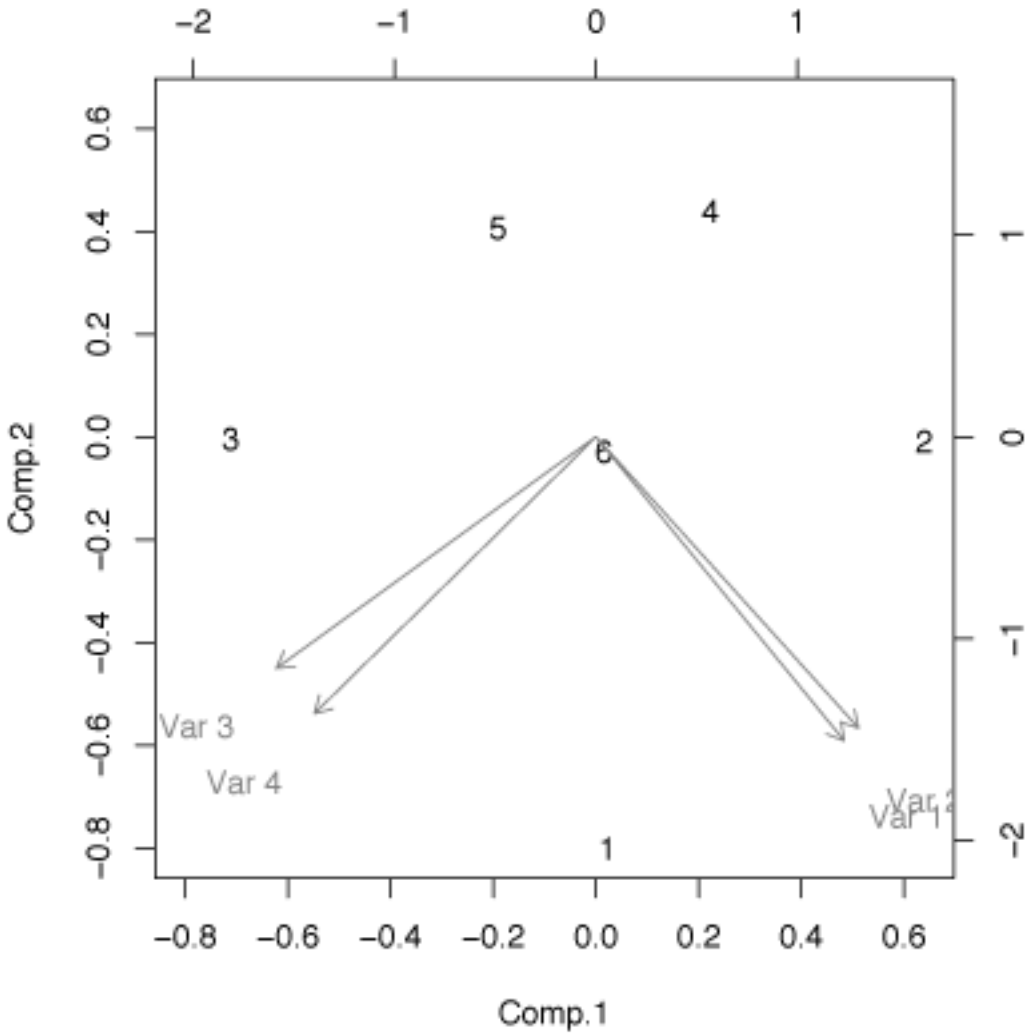
N.B. La matrice des corrélations reste inchangée.

Inertie totale du nuage avec les variables centrées réduites :

$I = \sum z_{ij}^2 = 6 \times 4 = 24$

Le principal résultat fourni par l'ACP :

Projection du nuage selon les deux premières composantes principales



Composantes principales : variables CP_1, CP_2, CP_3, CP_4 telles que :

• CP_1 représente la direction de plus grande dispersion du nuage de points
• CP_2 représente la direction de plus grande dispersion des résidus, une fois l'effet de CP_1 éliminé
• idem pour CP_3 et CP_4
• Les variables CP_j sont des combinaisons des variables Z_j
• Les variables CP_j ne sont en général pas réduites
• Les variables CP_j sont deux à deux indépendantes : pour $j \neq k$, $\rho(CP_j, CP_k) = 0$.

Résultats numériques attendus :

– Valeurs propres

– Scores des individus
– Contributions des individus
– Inertie relative des individus
– Qualité de la représentation des individus

– Saturations des variables
– Contributions des variables
– Qualité de la représentation des variables

Valeurs propres : Chaque valeur propre représente la variance prise en compte par la composante principale correspondante

	CP_1	CP_2	CP_3	CP_4
Valeur propre	2.0011	1.8668	0.0317	0.0003
Prop. variance	0.5003	0.4917	0.0079	0.0001
Prop. cumulée	0.5003	0.9920	0.9999	1.0000

Ici, les deux premiers axes rendent comptent de 99.2% de la variance totale.

Scores des individus : valeurs prises par les variables CP_j sur les individus.

Suj	CP_1	CP_2	CP_3	CP_4
s1	0.0771	-2.7515	-0.0935	0.0166
s2	2.2153	-0.0327	0.1778	-0.0095
s3	-2.4608	-0.0173	0.2445	-0.0036
s4	0.7734	1.5097	0.0664	0.0219
s5	-0.6606	1.3926	-0.2592	0.0064
s6	0.0556	-0.1008	-0.1360	-0.0319

Valeurs propres : variances des variables CP_j.

Expressions des composantes principales comme combinaisons linéaires des variables de départ :

Var	CP_1	CP_2	CP_3	CP_4
Z_1	0.445	-0.548	-0.656	-0.267
Z_2	0.470	-0.525	0.690	0.166
Z_3	-0.572	-0.418	0.232	-0.667
Z_4	-0.504	-0.499	-0.199	0.676

Par exemple :

$CP_1 = 0.445\ Z_1 + 0.470\ Z_2$ - $0.572\ Z_3$ - $0.504\ Z_4$

et, par exemple, pour le premier sujet :

$0.0771 = 0.445 \times 1.5993 + 0.470 \times 1.4197 - 0.572 \times 1.0722 - 0.504 \times 1.3648$

Ce tableau peut aussi être lu dans l'autre sens :

$Z_1 = 0.445\ CP_1$ - $0.548\ CP_2$ - $0.656\ CP_3$ - $0.267\ CP4$

et, pour le premier sujet :

$1.5993 = 0.445 \times 0.0771 - 0.548 \times (-2.7515) - 0.656 \times (-0.0935) - 0.267 \times 0.0166$

Saturations des variables : coefficients de corrélation entre Z_j et CP_k

Var	CP_1	CP_2	CP_3	CP_4
Z_1	0.6288	-0.7687	-0.1169	-0.0048
Z_2	0.6651	-0.7366	0.1228	0.0030
Z_3	-0.8094	-0.5857	0.0413	-0.0119
Z_4	-0.7129	-0.7002	-0.0355	0.0121

Il existe un lien entre les deux tableaux précédents : Par exemple : $0.6288 = \sqrt{2.0011} \times 0.445$

Inertie relative d'un individu :

Carré de la distance de l'individu à l'origine divisé par l'inertie totale du nuage ;

Contribution (relative) d'un individu à la formation d'une composante principale :

Par exemple, pour s1 et CP_1 :

$$CTR = \frac{0.0771^2}{0.0771^2 + \cdots + 0.0556^2} = \frac{0.0771^2}{6 \times 2.0011} = 0.64\%$$

Qualité de la représentation d'un individu par une composante principale, par les composantes principales retenues :

Pour s1 et CP_2 :

$$QLT = \frac{2.7515^2}{0.0771^2 + 2.7515^2 + 0.0935^2 + 0.0166^2} = 0.9980$$

Pour s1 et CP_1, CP_2

$$QLT = \frac{0.0771^2 + 2.7515^2}{0.0771^2 + 2.7515^2 + 0.0935^2 + 0.0166^2} = 0.9988$$

Pour tous les individus :

Suj	QLT
s1	0.9988
s2	0.9936
s3	0.9902
s4	0.9983
s5	0.9725
s6	0.4044

Tous les individus, sauf le dernier, sont très bien représentés par les deux premières composantes principales.

Contribution d'une variable à la formation d'une composante principale :

Exemple : contribution de la première variable à la formation de la première composante principale

$$CTR = \frac{0.6288^2}{0.6288^2 + 0.6651^2 + 0.8094^2 + 0.7129^2}$$

$$CTR = \frac{0.6288^2}{2.011} = 0.1976$$

Qualité de la représentation d'une variable par une composante principale : carré du coefficient correspondant.

Qualité de la représentation d'une variable par les composantes principales retenues :

Pour Z_1 et CP_1, CP_2 :

$$QLT = \frac{0.6288^2 + 0.7687^2}{0.6288^2 + 0.7687^2 + 0.1169^2 + 0.0048^2}$$

$$QLT = 0.6288^2 + 0.7687^2 = 0.9863$$

Pour les 4 variables :

Var	QLT
Z_1	0.9863
Z_2	0.9849
Z_3	0.9982
Z_4	0.9985

Interprétation graphique : carré de la longueur du "vecteur" représentant la variable.

Interprétation des résultats

Scores des individus et saturations ne sont pas exprimées avec la même unité de mesure.

Interpréter chaque axe : part de la variance dont il rend compte, variables avec lesquelles il est corrélé.

Proximités entre individus : à interpréter avec prudence, ils peuvent prendre des valeurs très différentes sur des variables non représentées.

Individus proches de l'origine : ils ont, de toutes façons, peu contribué à l'inertie.

Interpréter plutôt les oppositions marquées entre individus.

Effet de masse ou de taille : lorsque toutes les variables ont entre elles des corrélations positives, le premier axe classe simplement les individus par valeurs des variables.

Variantes

– ACP non normée

– ACP pondérée : on affecte des poids aux individus

Analyse Factorielle des Correspondances

Position du problème :
On dispose d'un tableau de contingence comportant un grand nombre de lignes et de colonnes. On veut faire une étude de ces données, plus précise qu'un simple χ^2.

Pour chacune des variables, quelles sont les modalités qui se ressemblent, quelles sont celles qui s'opposent ?

Pour les couples de modalités des deux variables, quelles sont les modalités qui s'attirent, quelles sont celles qui se repoussent ?

Exemple "historique" : Données Caith

Couleur des yeux et couleur des cheveux pour 5387 enfants du comté de Caithness (Student, 1940)

	HFAI	HRED	HMEDIUM	HDARK	HBLACK
EBLUE	326	38	241	110	3
ELIGHT	688	116	584	188	4
EMEDIUM	343	84	909	412	26
EDARK	98	48	403	681	85

Valeur du Chi2 à 12 ddl : 1240.04
Niveau de significativité : inférieur à 10^{-5}

Etude descriptive : profils des lignes, des colonnes, taux de liaison

Fréquences conjointes f_{ij} et marginales $f_{i.}$, $f_{.j}$

	HFAIR	HRED	HMEDIUM	HDARK	HBLACK	$f_{i.}$
EBLUE	0,0605	0,0071	0,0447	0,0204	0,0006	0,1333
ELIGHT	0,1277	0,0215	0,1084	0,0349	0,0007	0,2933
EMEDIUM	0,0637	0,0156	0,1687	0,0765	0,0048	0,3293
EDARK	0,0182	0,0089	0,0748	0,1264	0,0158	0,2441
$f_{.j}$	0,2701	0,0531	0,3967	0,2582	0,0219	1

Profil des lignes

	HFAIR	HRED	HMEDIUM	HDARK	HBLACK	Total
EBLUE	0,454	0,053	0,336	0,153	0,004	1
ELIGHT	0,435	0,073	0,370	0,119	0,003	1
EMEDIUM	0,193	0,047	0,512	0,232	0,015	1
EDARK	0,075	0,037	0,306	0,518	0,065	1
Masse	0,270	0,053	0,397	0,258	0,022	1

Profil des colonnes

	HFAIR	HRED	HMEDIUM	HDARK	HBLACK	Masse
EBLUE	0,224	0,133	0,113	0,079	0,025	0,133
ELIGHT	0,473	0,406	0,273	0,135	0,034	0,293
EMEDIUM	0,236	0,294	0,425	0,296	0,220	0,329
EDARK	0,067	0,168	0,189	0,490	0,720	0,244
Total	1	1	1	1	1	1

Taux de liaison t_{ij} :

	HFAIR	HRED	HMEDIUM	HDARK	HBLACK
EBLUE	0,6810	-0,0031	-0,1539	-0,4067	-0,8093
ELIGHT	0,6122	0,3829	-0,0683	-0,5392	-0,8844
EMEDIUM	-0,2841	-0,1081	0,2917	-0,1006	-0,3309
EDARK	-0,7241	-0,3125	-0,2275	1,0056	1,9509

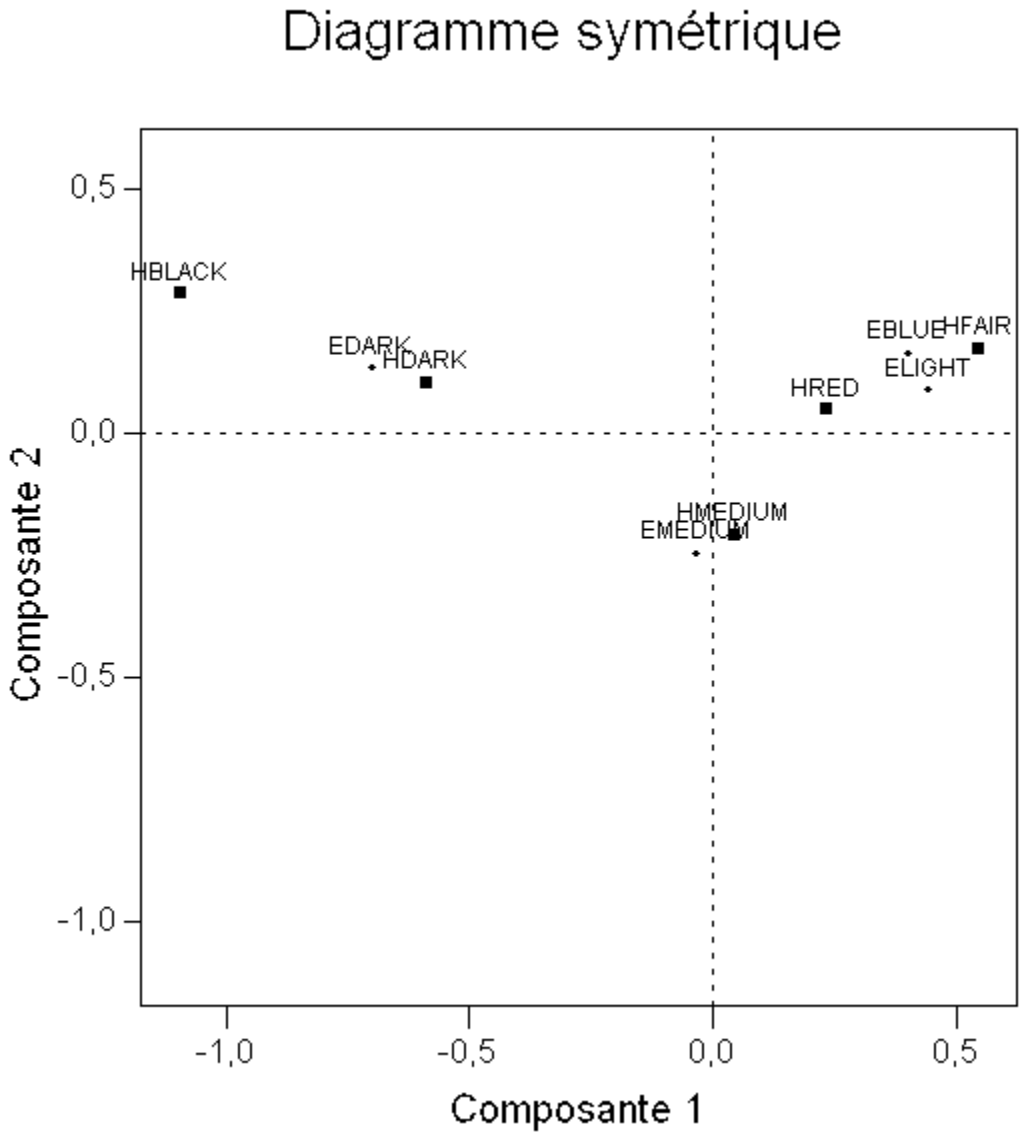
Définition : $t_{ij} = \dfrac{f_{ij} - f_{i.}f_{.j}}{f_{i.}f_{.j}}$

La moyenne des t_{ij}, pondérés par les coefficients $f_{i.}f_{.j}$, est nulle.

La moyenne des t_{ij}^2, pondérés par les coefficients $f_{i.}f_{.j}$, est le *carré moyen de contingence* Φ^2. On a :

$$\Phi^2 = \sum\sum \frac{(f_{ij} - f_{i.}f_{.j})^2}{f_{i.}f_{.j}} = \frac{\chi^2}{N}$$

Analyse des correspondances proprement dite :



Le calcul des valeurs propres est fait à partir d'un tableau calculé à partir des profils ligne et des profils colonne. Nous ne le détaillerons pas ici.

Valeurs propres associées aux axes principaux :

La première valeur propre vaut 1, et n'est pas mentionnée dans les résultats. Les autres valeurs propres sont inférieures à 1.
Somme des valeurs propres : Φ^2.

Axe	Inertie	Proportion	Cumulé
1	0,1992	0,8656	0,8656
2	0,0301	0,1307	0,9963
3	0,0009	0,0037	1,0000
Total	0,2302		

Contributions relatives des modalités à la formation de l'inertie : *Inerties relatives*

	HFAIR	HRED	HMEDIUM	HDARK	HBLACK	Total
EBLUE	0,073	0,000	0,005	0,025	0,008	0,111
ELIGHT	0,129	0,010	0,002	0,096	0,022	0,259
EMEDIUM	0,031	0,001	0,048	0,004	0,003	0,088
EDARK	0,150	0,005	0,022	0,277	0,088	0,543
Total	0,383	0,016	0,078	0,401	0,122	1,000

Pour les deux premiers axes factoriels :

Contributions des lignes

Nom	Qual	Mass	Inert
EBLUE	0,979	0,133	0,111
ELIGHT	0,995	0,293	0,259
EMEDIUM	0,999	0,329	0,088
EDARK	1,000	0,244	0,543

$$\text{Qualité} = \frac{(\text{Coord. selon CP1})^2 + (\text{Coord. selon CP2})^2}{\sum(\text{Coord. selon les CP})^2}$$

$$\text{Exemple : } 0.979 = \frac{0.400^2 + 0.165^2}{0.400^2 + 0.165^2 + 0.064^2}$$

	—Composante 1—			—Composante 2—		
Nom	Coord	Corr	Contr	Coord	Corr	Contr
EBLUE	0,400	0,836	0,107	0,165	0,143	0,121
ELIGHT	0,441	0,956	0,286	0,088	0,039	0,076
EMEDIUM	-0,034	0,018	0,002	-0,245	0,981	0,657
EDARK	-0,703	0,965	0,605	0,134	0,035	0,145

Corr : qualité de la représentation de l'individu par sa projection sur l'axe.

$$\text{Qualité suivant CP i} = \frac{(\text{Coord. selon CP i})^2}{\sum(\text{Coord. selon les CP})^2}$$

$$\text{Exemple : } 0.836 = \frac{0.400^2}{0.400^2 + 0.165^2 + 0.064^2}$$

Contr : contribution relative d'un individu à la formation de l'inertie d'un axe.

$$\text{Contr(Individu i, Axe k)} = \frac{\text{Masse ligne i} \times (\text{Coord. indiv. i selon CP k})^2}{\text{Valeur propre relative à l'axe k}}$$

$$\text{Exemple : } 0.107 = \frac{0.400^2 \times 0.133}{0.1992}$$

Contribution des colonnes

Nom	Qual	Mass	Inert
HFAIR	1,000	0,270	0,383
HRED	0,803	0,053	0,016
HMEDIUM	1,000	0,397	0,078
HDARK	1,000	0,258	0,401
HBLACK	0,998	0,022	0,122

	—Composante 1—			—Composante 2—		
Nom	Coord	Corr	Contr	Coord	Corr	Contr
HFAIR	0,544	0,907	0,401	0,174	0,093	0,271
HRED	0,233	0,770	0,014	0,048	0,033	0,004
HMEDIUM	0,042	0,039	0,004	-0,208	0,961	0,572
HDARK	-0,589	0,969	0,449	0,104	0,030	0,093
HBLACK	-1,094	0,934	0,132	0,286	0,064	0,060

Interprétation des résultats

– Valeur propre proche de 1 : forte liaison entre lignes et colonnes
– Pour chaque axe : points-lignes et points-colonnes dont les contributions sont fortes. Par exemple : points dont la contribution est supérieure à la contribution moyenne. Voir aussi les points bien représentés
– Une modalité (ou un groupe de modalités proches) bien représentée : axe spécifique
– Points proches : à interpéter avec précaution. Intéressant si les points sont bien représentés.
– Etudier les associations points-lignes / points-colonnes proches