

Analyse multidimensionnelle des données

F.-G. Carpentier

Exemples de données relevant de l'analyse multidimensionnelle

Consommations annuelles de 8 types de denrées alimentaires pour 8 catégories socio-professionnelles

	PAO	PAA	VIO	VIA	POT	LEC	RAI	PLP
AGRI	167	1	163	23	41	8	6	6
SAAG	162	2	141	12	40	12	4	15
PRIN	119	6	69	56	39	5	13	41
CSUP	87	11	63	111	27	3	18	39
CMOY	103	5	68	77	32	4	11	30
EMPL	111	4	72	66	34	6	10	28
OUVR	130	3	76	52	43	7	7	16
INAC	138	7	117	74	53	8	12	20

Source : Saporta, 1990

Variables :

PAO	Pain ordinaire
PAA	Autre pain
VIO	Vin ordinaire
VIA	Autre vin
POT	Pommes de terre
LEC	Légumes secs
RAI	Raisin de table
PLP	Plats préparés

Observations :

AGRI	Exploitants agricoles
SAAG	Salariés agricoles
PRIN	Professions indépendantes
CSUP	Cadres supérieurs
CMOY	Cadres moyens
EMPL	Employés
OUVR	Ouvriers
INAC	Inactifs

Exemples de données relevant de l'analyse multidimensionnelle

Tableau de contingence : répartition d'étudiants en 1975-1976

	Droit	Sciences	Médecine	IUT
Exp. agri.	80	99	65	58
Patron	168	137	208	62
Cadre sup.	470	400	876	79
Employé	145	133	135	54
Ouvrier	166	193	127	129

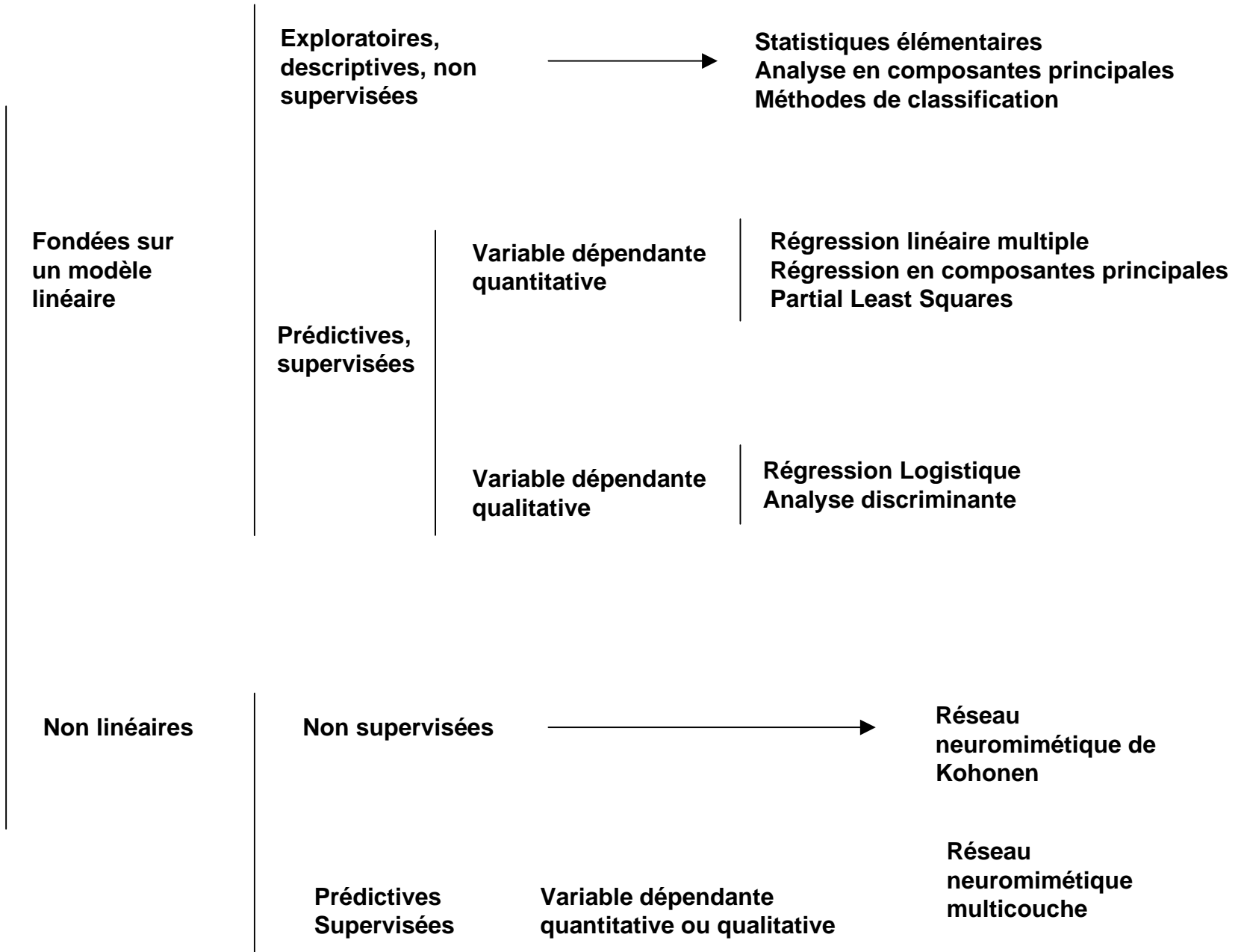
Cité par Saporta (1990)

Exemples de données relevant de l'analyse multidimensionnelle

Questions à réponses fermées : sexe (2 modalités), niveau de revenu (2 modalités), préférence (3 modalités)

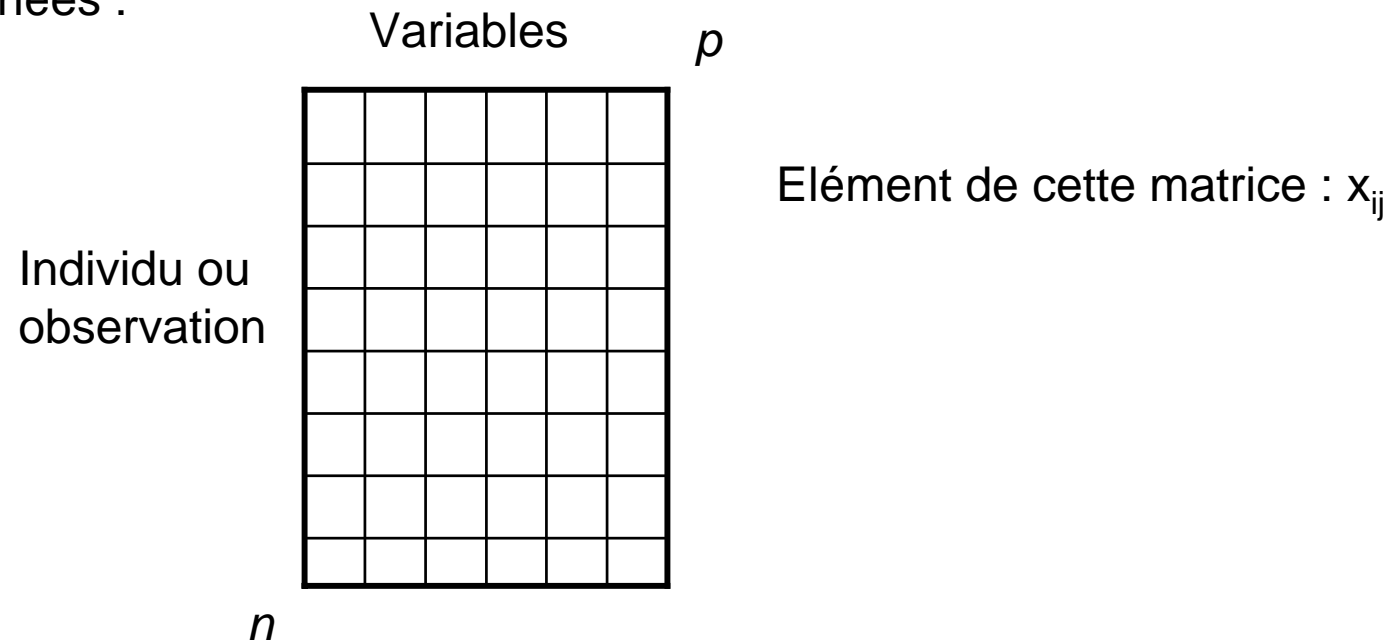
	1 Sexe	2 Revenu	3 Preference
s1	F	M	A
s2	F	M	A
s3	F	E	B
s4	F	E	C
s5	F	E	C
s6	H	E	C
s7	H	E	B
s8	H	M	B
s9	H	M	B
s10	H	M	A

**Méthodes
d'analyse
de données**

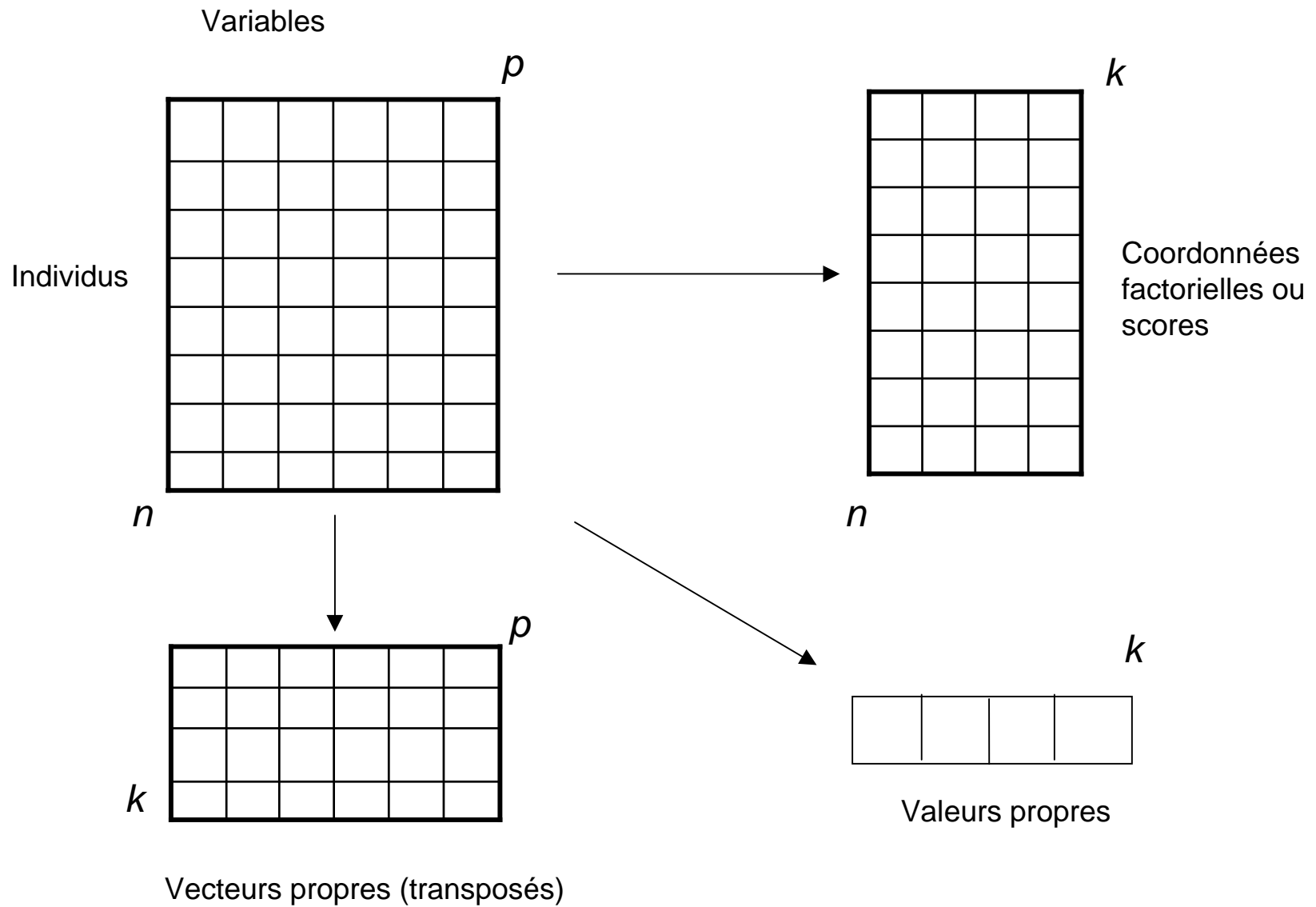


Analyse en composantes principales

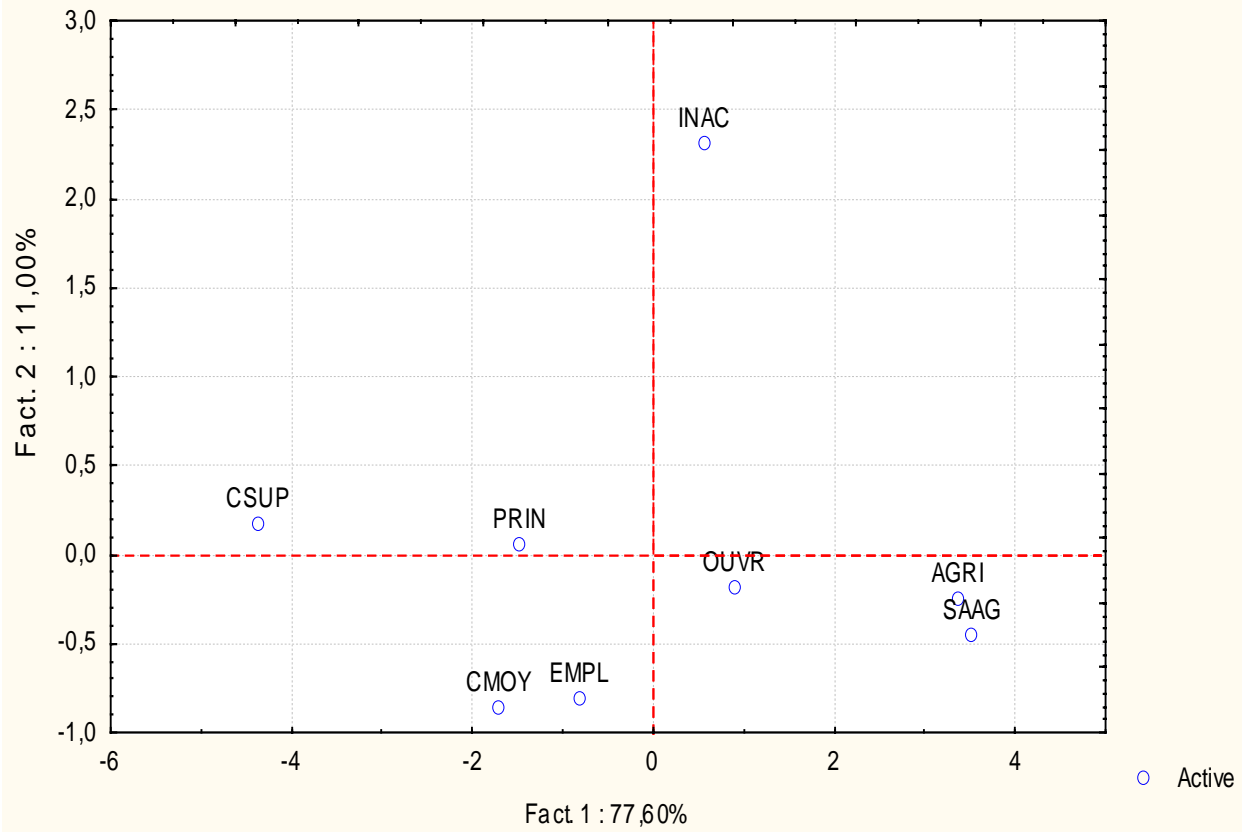
Données :



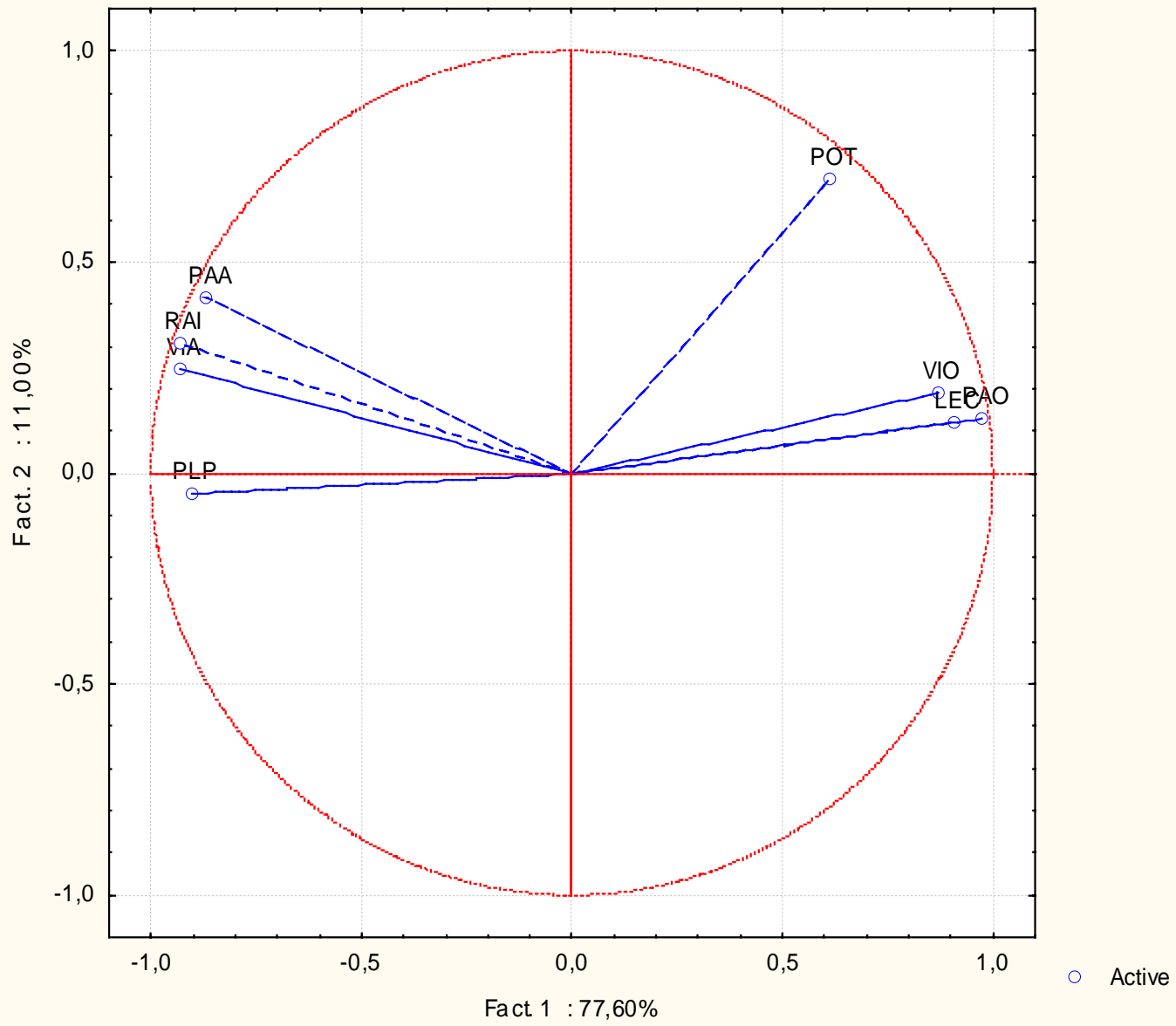
Principaux résultats d'une ACP

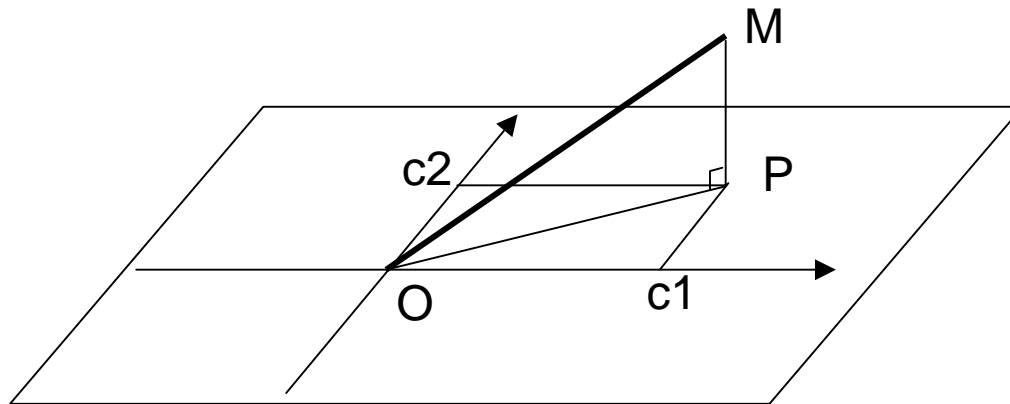


Projection des ind. sur le plan factoriel (1 x 2)
Observations avec la somme des cosinus carrés >= 0,00



Projection des variables sur le plan factoriel (1 x 2)





Cosinus carrés

$$\text{Cos}^2(\overrightarrow{OM}, CP_1) = \frac{Oc_1^2}{OM^2}$$

$$\text{Cos}^2(\overrightarrow{OM}, CP_2) = \frac{Oc_2^2}{OM^2}$$

\overrightarrow{OM} : vecteur de l'observation

\overrightarrow{OP} : vecteur de la projection sur le plan factoriel

$\overrightarrow{Oc_1}$: projection sur l'axe 1

$\overrightarrow{Oc_2}$: projection sur l'axe 2

Qualité

$$QUAL = \text{Cos}^2(\overrightarrow{OM}, \overrightarrow{OP}) = \frac{OP^2}{OM^2}$$

	QLT	Coord. 1	Cos2	Ctr	Coord. 2	Cos2	Ctr
AGRI	0,889	1,35	0,884	22,89	-0,26	0,005	0,86
SAAG	0,913	1,41	0,898	24,97	-0,48	0,014	2,84
PRIN	0,576	-0,59	0,575	4,36	0,06	0,001	0,05
CSUP	0,943	-1,75	0,942	38,26	0,19	0,002	0,44
CMOY	0,940	-0,69	0,753	5,94	-0,91	0,187	10,43
EMPL	0,858	-0,32	0,428	1,31	-0,86	0,430	9,29
OVR	0,376	0,36	0,361	1,63	-0,20	0,015	0,48
INAC	0,987	0,23	0,056	0,64	2,46	0,932	75,61
				100			100

Contributions des individus

Analyse factorielle des correspondances

Tableau de contingence : répartition d'étudiants en 1975-1976

	Droit	Sciences	Médecine	IUT
Exp. agri.	80	99	65	58
Patron	168	137	208	62
Cadre sup.	470	400	876	79
Employé	145	133	135	54
Ouvrier	166	193	127	129

Cité par Saporta (1990)

Test du khi-2 sur un tableau de contingence

Modalités lignes : variable X

Modalités colonnes : variable Y

Hypothèses du test :

H_0 : Les variables X et Y sont indépendantes

H_1 : Les variables X et Y sont dépendantes

	Droit	Sciences	Médecine	IUT
Exp. agri.	80	99	65	58
Patron	168	137	208	62
Cadre sup.	470	400	876	79
Employé	145	133	135	54
Ouvrier	166	193	127	129

Effectifs observés O

Construction de la statistique de test

	Droit	Sciences	Médecine	IUT	Total
Exp. agri.	80	99	65	58	302
Patron	168	137	208	62	575
Cadre sup.	470	400	876	79	1825
Employé	145	133	135	54	467
Ouvrier	166	193	127	129	615
Total	1029	962	1411	382	3784

Effectifs observés O_{ij}

	Droit	Sciences	Médecine	IUT
Exp. agri.	82,12	76,78	112,61	30,49
Patron	156,36	146,18	214,41	58,05
Cadre sup.	496,28	463,97	680,52	184,24
Employé	126,99	118,72	174,14	47,14
Ouvrier	167,24	156,35	229,32	62,09

Effectifs théoriques T_{ij}

$$T_{ij} = \frac{\text{Total ligne } i \times \text{Total colonne } j}{\text{Total Général}}$$

$$\text{Exemple : } 82,12 = \frac{302 \times 1029}{3784}$$

Contributions au khi-2

	Droit	Sciences	Médecine	IUT
Exp. agri.	0,05	6,43	20,13	24,83
Patron	0,87	0,58	0,19	0,27
Cadre sup.	1,39	8,82	56,15	60,11
Employé	2,55	1,72	8,80	1,00
Ouvrier	0,01	8,59	45,66	72,12

Contributions au khi-2 : $(O - T)^2/T$

$$Ctr_{ij} = \frac{(O_{ij} - T_{ij})^2}{T_{ij}} ;$$

$$\text{Exemple : } 0,05 = \frac{(80 - 82,12)^2}{82,12}$$

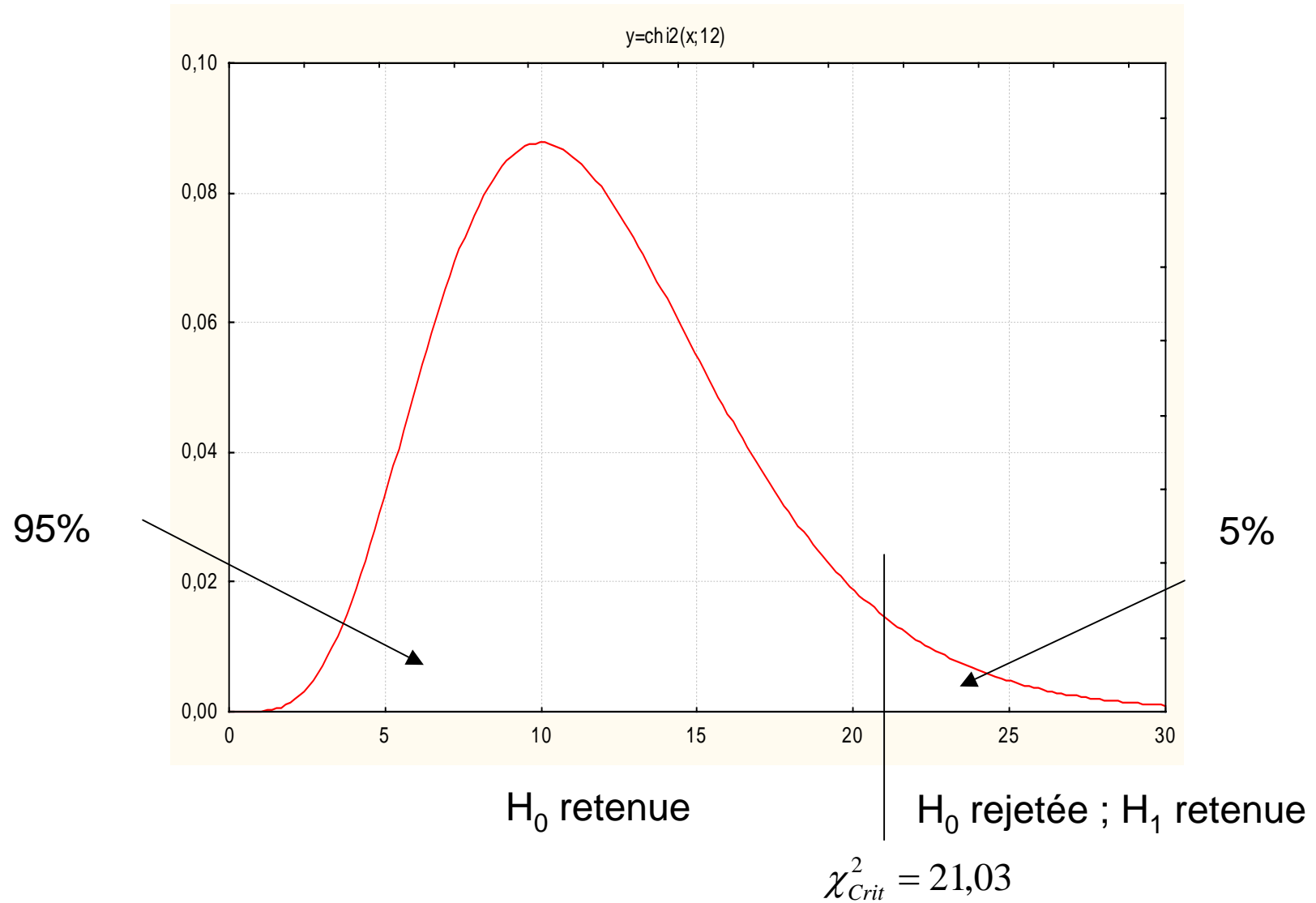
Calcul du khi-2

$$\chi_{Obs}^2 = \sum_{i,j} Ctr_{ij} = 0,05 + \dots + 72,12 = 320,2$$

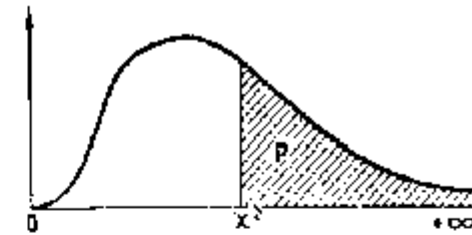
Nombre de degrés de liberté :

$$ddl = (\text{Nb Modalités lignes} - 1)(\text{Nb Modalités colonnes} - 1) = 12$$

Loi du khi-2



DISTRIBUTION DE χ^2 (Loi de K. Pearson)
Valeur de χ^2 ayant la probabilité P d'être dépassée.

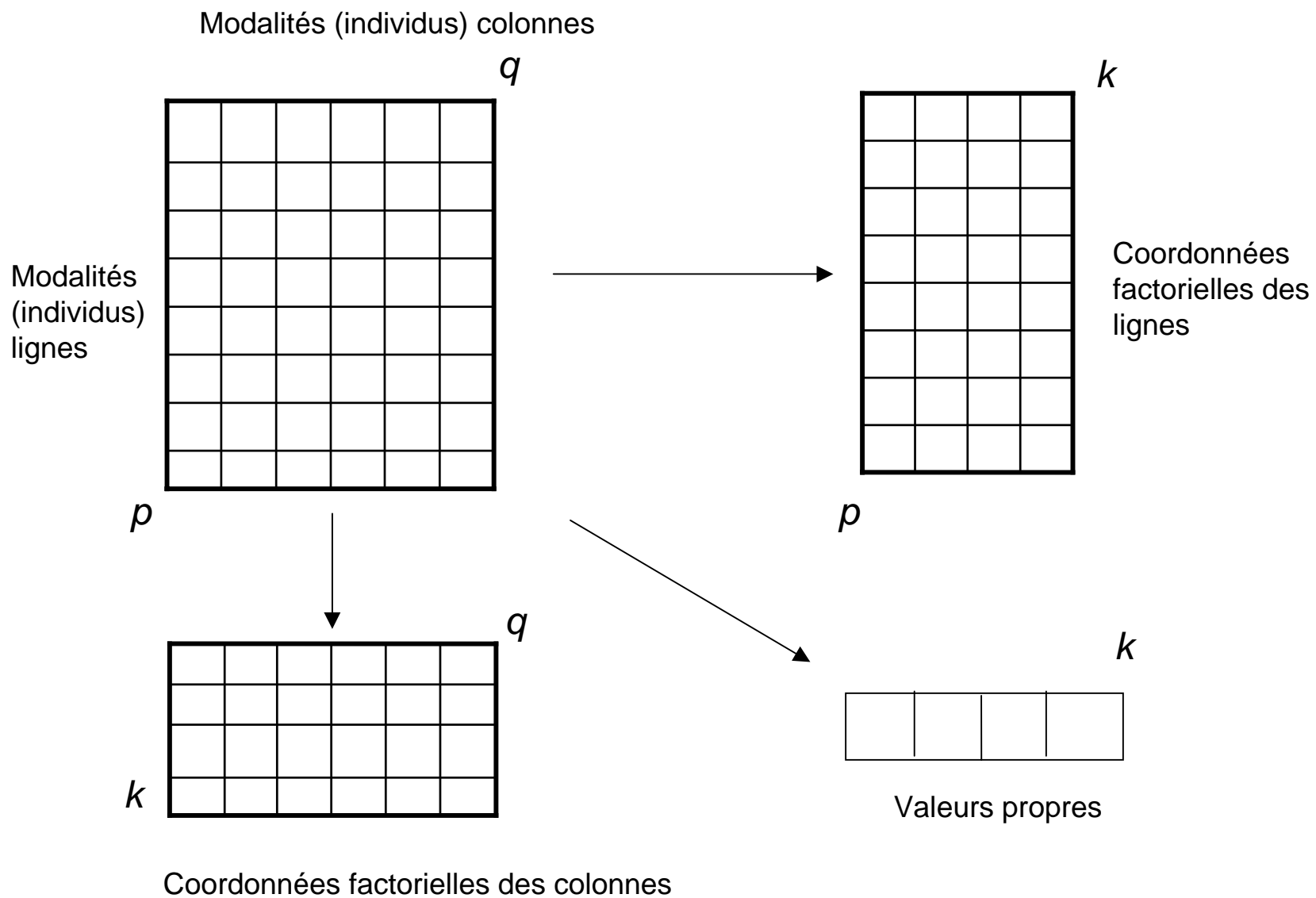


ν	0,9	0,8	0,7	0,5	0,3	0,2	0,1	0,05	0,02	0,01
1	0,0158	0,0642	0,148	0,455	1,074	1,642	2,706	3,841	5,412	6,635
2	0,211	0,446	0,713	1,386	2,408	3,219	4,605	5,991	7,824	9,210
3	0,584	1,005	1,424	2,366	3,665	4,642	6,251	7,815	9,837	11,345
4	1,064	1,649	2,195	3,357	4,878	5,989	7,779	9,488	11,668	13,277
5	1,610	2,343	3,000	4,351	6,064	7,289	9,236	11,070	13,388	15,086
6	2,204	3,070	3,828	5,348	7,231	8,558	10,645	12,592	15,033	16,812
7	2,833	3,822	4,671	6,346	8,383	9,803	12,017	14,067	16,622	18,475
8	3,490	4,594	5,527	7,344	9,524	11,030	13,362	15,507	18,168	20,090
9	4,168	5,380	6,393	8,343	10,656	12,242	14,684	16,919	19,679	21,666
10	4,865	6,179	7,267	9,342	11,781	13,442	15,987	18,307	21,161	23,209
11	5,578	6,989	8,148	10,341	12,899	14,631	17,275	19,675	22,618	24,725
12	6,304	7,807	9,034	11,340	14,011	15,812	18,549	21,026	24,054	26,217
13	7,041	8,634	9,926	12,340	15,119	16,985	19,812	22,362	25,471	27,688
14	7,790	9,467	10,821	13,339	16,222	18,151	21,064	23,685	26,873	29,141
15	8,547	10,307	11,721	14,339	17,322	19,311	22,307	24,996	28,259	30,578

$$\chi_{Obs}^2 > \chi_{Crit}^2 \quad : \text{ on conclut donc sur } H_1$$

Les deux variables étudiées dépendent l'une de l'autre

Principaux résultats d'une AFC



Effectifs et fréquences marginaux

	Droit	Sciences	Médecine	IUT	Effectifs marginaux lignes	Fréquence
Exp. agri.	80	99	65	58	302	0,0798
Patron	168	137	208	62	575	0,1520
Cadre sup.	470	400	876	79	1825	0,4823
Employé	145	133	135	54	467	0,1234
Ouvrier	166	193	127	129	615	0,1625
Effectifs marginaux colonnes	1029	962	1411	382	3784	
Fréquence	0,2719	0,2542	0,3729	0,1010		

Fréquences théoriques dans l'hypothèse d'indépendance

X	0,2719	0,2542	0,3729	0,1010					
0,0798						0,0217	0,0203	0,0298	0,0081
0,1520						0,0413	0,0386	0,0567	0,0153
0,4823					=	0,1312	0,1226	0,1798	0,0487
0,1234						0,0336	0,0314	0,0460	0,0125
0,1625						0,0442	0,0413	0,0606	0,0164

Fréquences théoriques dans l'hypothèse d'indépendance

$$\begin{bmatrix} 0,0798 \\ 0,1520 \\ 0,4823 \\ 0,1234 \\ 0,1625 \end{bmatrix} \times \begin{bmatrix} 0,2719 & 0,2542 & 0,3729 & 0,1010 \end{bmatrix} = \begin{bmatrix} 0,0217 & 0,0203 & 0,0298 & 0,081 \\ 0,0413 & 0,0386 & 0,0567 & 0,0153 \\ 0,1312 & 0,1226 & 0,1798 & 0,0487 \\ 0,0336 & 0,0314 & 0,0460 & 0,0125 \\ 0,0442 & 0,0413 & 0,0606 & 0,0164 \end{bmatrix}$$

Effectifs théoriques dans le cas d'indépendance

0,0217	0,0203	0,0298	0,0081		82,12	76,78	112,61	30,49
0,0413	0,0386	0,0567	0,0153		156,36	146,18	214,41	58,05
0,1312	0,1226	0,1798	0,0487		496,28	463,97	680,52	184,24
0,0336	0,0314	0,0460	0,0125		126,99	118,72	174,14	47,14
0,0442	0,0413	0,0606	0,0164	x 3784 =	167,24	156,35	229,32	62,09

	Droit	Sciences	Médecine	IUT
Exp. agri.	80	99	65	58
Patron	168	137	208	62
Cadre sup.	470	400	876	79
Employé	145	133	135	54
Ouvrier	166	193	127	129

Effectifs observés O

	Droit	Sciences	Médecine	IUT
Exp. agri.	82,12	76,78	112,61	30,49
Patron	156,36	146,18	214,41	58,05
Cadre sup.	496,28	463,97	680,52	184,24
Employé	126,99	118,72	174,14	47,14
Ouvrier	167,24	156,35	229,32	62,09

Effectifs théoriques T

	Droit	Sciences	Médecine	IUT
Exp. agri.	-2,12	22,22	-47,61	27,51
Patron	11,64	-9,18	-6,41	3,95
Cadre sup.	-26,28	-63,97	195,48	-105,24
Employé	18,01	14,28	-39,14	6,86
Ouvrier	-1,24	36,65	-102,32	66,91

Ecart à l'indépendance : $E = O - T$

	Droit	Sciences	Médecine	IUT
Exp. agri.	82,12	76,78	112,61	30,49
Patron	156,36	146,18	214,41	58,05
Cadre sup.	496,28	463,97	680,52	184,24
Employé	126,99	118,72	174,14	47,14
Ouvrier	167,24	156,35	229,32	62,09

Effectifs théoriques T

	Droit	Sciences	Médecine	IUT
Exp. agri.	-2,12	22,22	-47,61	27,51
Patron	11,64	-9,18	-6,41	3,95
Cadre sup.	-26,28	-63,97	195,48	-105,24
Employé	18,01	14,28	-39,14	6,86
Ouvrier	-1,24	36,65	-102,32	66,91

Ecart à l'indépendance : $E = O - T$

	Droit	Sciences	Médecine	IUT
Exp. agri.	0,05	6,43	20,13	24,83
Patron	0,87	0,58	0,19	0,27
Cadre sup.	1,39	8,82	56,15	60,11
Employé	2,55	1,72	8,80	1,00
Ouvrier	0,01	8,59	45,66	72,12

Contributions au khi-2 : $(O - T)^2/T$

D'où : $Khi-2 = 320,2$.

	Droit	Sciences	Médecine	IUT
Exp. agri.	82,12	76,78	112,61	30,49
Patron	156,36	146,18	214,41	58,05
Cadre sup.	496,28	463,97	680,52	184,24
Employé	126,99	118,72	174,14	47,14
Ouvrier	167,24	156,35	229,32	62,09

Effectifs théoriques T

	Droit	Sciences	Médecine	IUT
Exp. agri.	-2,12	22,22	-47,61	27,51
Patron	11,64	-9,18	-6,41	3,95
Cadre sup.	-26,28	-63,97	195,48	-105,24
Employé	18,01	14,28	-39,14	6,86
Ouvrier	-1,24	36,65	-102,32	66,91

Ecart à l'indépendance : $E = O - T$

	Droit	Sciences	Médecine	IUT
Exp. agri.	-0,03	0,29	-0,42	0,90
Patron	0,07	-0,06	-0,03	0,07
Cadre sup.	-0,05	-0,14	0,29	-0,57
Employé	0,14	0,12	-0,22	0,15
Ouvrier	-0,01	0,23	-0,45	1,08

Taux de liaison : $(O - T)/T$: valeurs dans l'intervalle $[-1, +\infty [$

-0,42 : l'effectif observé est inférieur de 42% à l'effectif théorique

1,08 : l'effectif observé est supérieur de 108% à l'effectif théorique

Analyse des correspondances

Les questions auxquelles on cherche à répondre :

- Quelles sont les modalités lignes qui sont « proches » du profil ligne moyen ? Quelles sont celles qui s'en écartent le plus ?
- Quelles sont les modalités colonnes qui sont « proches » du profil colonne moyen ? Quelles sont celles qui s'en écartent le plus ?
- Quelles sont les modalités lignes et les modalités colonnes qui « s'attirent » ? Quelles sont celles qui « se repoussent » ?

Analyse des correspondances

Notations :

Soit un tableau de contingence comportant p lignes et q colonnes.

- L'élément du tableau situé à l'intersection de la ligne i et de la colonne j est noté n_{ij} .
- La somme des éléments d'une ligne est notée $n_{i\bullet}$.
- La somme des éléments d'une colonne est notée $n_{\bullet j}$.

Distance (du Phi-2) entre deux profils lignes :

$$d_{ii'}^2 = \sum_{j=1}^q \frac{n}{n_{\bullet j}} \left(\frac{n_{ij}}{n_{i\bullet}} - \frac{n_{i'j}}{n_{i'\bullet}} \right)^2$$

Exemple :

	Droit	Sciences	Médecine	IUT	Effectifs marginaux lignes
Exp. agri.	80	99	65	58	302
Patron	168	137	208	62	575
Cadre sup.	470	400	876	79	1825
Employé	145	133	135	54	467
Ouvrier	166	193	127	129	615
Effectifs marginaux colonnes	1029	962	1411	382	3784

$$d_{12}^2 = \frac{3784}{1029} \left(\frac{80}{302} - \frac{168}{575} \right)^2 + \frac{3784}{962} \left(\frac{99}{302} - \frac{137}{575} \right)^2 + \frac{3784}{1411} \left(\frac{65}{302} - \frac{208}{575} \right)^2 + \frac{3784}{382} \left(\frac{58}{302} - \frac{62}{575} \right)^2$$

Distance (du Phi-2) entre deux profils colonnes :

$$d_{jj'}^2 = \sum_{i=1}^p \frac{n}{n_{i\bullet}} \left(\frac{n_{ij}}{n_{\bullet j}} - \frac{n_{ij'}}{n_{\bullet j'}} \right)^2$$

Exemple : distance entre les colonnes 1 et 2

	Droit	Sciences	Médecine	IUT	Effectifs marginaux lignes
Exp. agri.	80	99	65	58	302
Patron	168	137	208	62	575
Cadre sup.	470	400	876	79	1825
Employé	145	133	135	54	467
Ouvrier	166	193	127	129	615
Effectifs marginaux colonnes	1029	962	1411	382	3784

$$d_{12}^2 = \frac{3784}{302} \left(\frac{80}{1029} - \frac{99}{962} \right)^2 + \frac{3784}{575} \left(\frac{168}{1029} - \frac{137}{962} \right)^2 + \frac{3784}{1825} \left(\frac{470}{1029} - \frac{400}{962} \right)^2 + \frac{3784}{467} \left(\frac{145}{1029} - \frac{133}{962} \right)^2 + \frac{3784}{615} \left(\frac{166}{1029} - \frac{193}{962} \right)^2$$

Propriété d'équivalence distributionnelle :

- Si on regroupe deux modalités lignes, les distances entre les profils-colonnes, ou entre les autres profils-lignes restent inchangées.
- Si on regroupe deux modalités colonnes, les distances entre les profils-lignes, ou entre les autres profils-colonnes restent inchangées.

Valeurs propres

	ValProp.	%age inertie	%age cumulé	Chi ²
1	0,082	97,35	97,35	311,78
2	0,002	2,01	99,36	6,45
3	0,001	0,64	100,00	2,04

Inertie totale du nuage de points :

$$\Phi^2 = \frac{\chi^2}{N} = \sum \text{Valeurs Propres} = \sum GM_i^2$$

Résultats relatifs aux lignes

	Coord. Dim.1	Coord. Dim.2	Masse	Qualité	Inertie Relative	Inertie Dim.1	Cosinus ² Dim.1	Inertie Dim.2	Cosinus ² Dim.2
Exp. Agri.	0,410	0,026	0,080	0,991	0,161	0,163	0,987	0,032	0,004
Patrons	0,020	-0,027	0,152	0,336	0,006	0,001	0,123	0,063	0,213
Cadres Sup.	-0,263	0,016	0,482	0,999	0,395	0,404	0,996	0,069	0,004
Employés	0,142	-0,097	0,123	0,985	0,044	0,030	0,670	0,686	0,315
Ouvriers	0,451	0,040	0,163	1,000	0,395	0,402	0,992	0,150	0,008

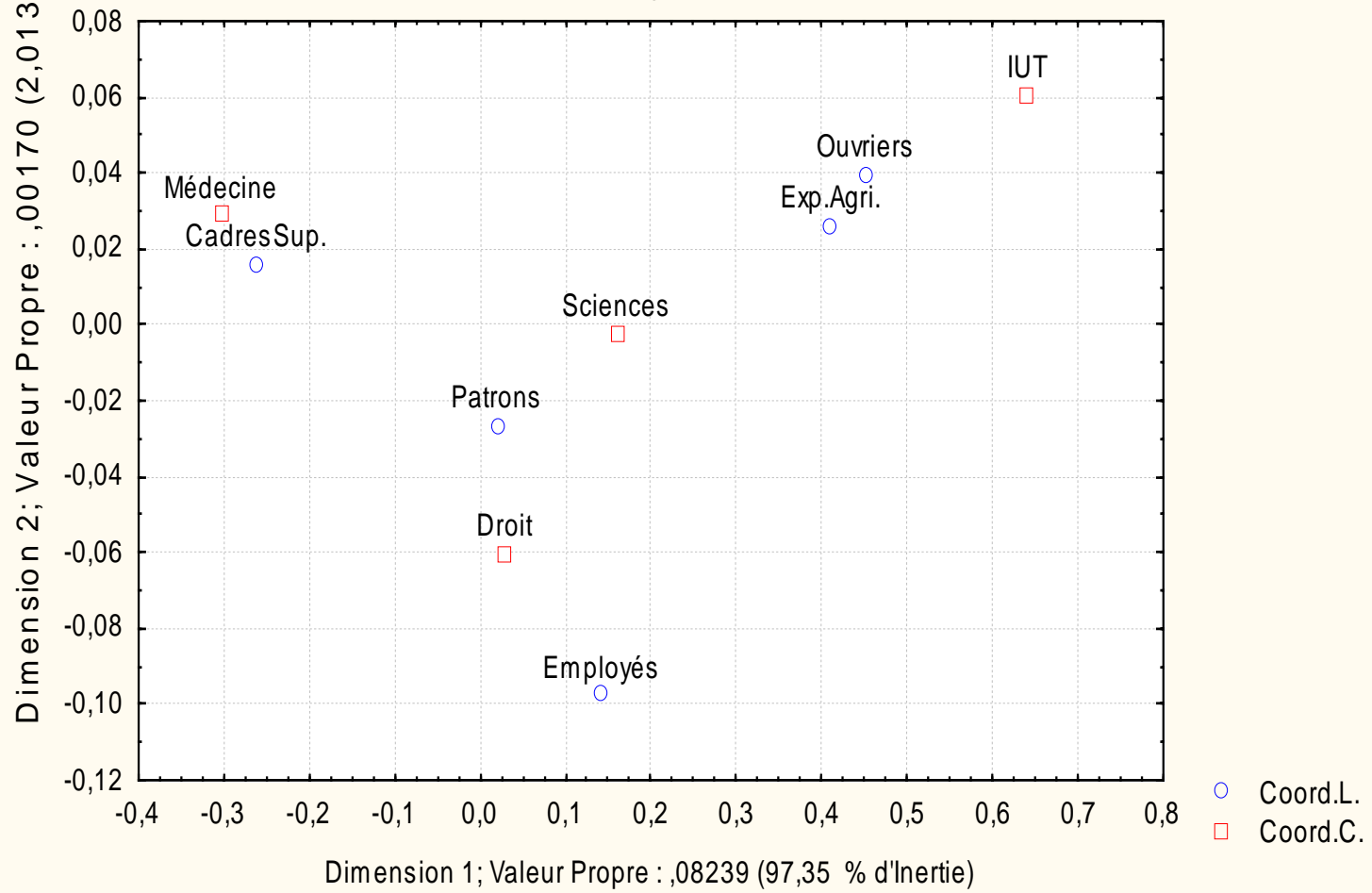
Résultats relatifs aux colonnes

	Coord. Dim.1	Coord. Dim.2	Masse	Qualité	Inertie Relative	Inertie Dim.1	Cosinus ² Dim.1	Inertie Dim.2	Cosinus ² Dim.2
Droit	0,028	-0,061	0,272	0,942	0,015	0,003	0,165	0,588	0,777
Sciences	0,160	-0,003	0,254	0,948	0,082	0,079	0,948	0,001	0,000
Médecine	-0,303	0,030	0,373	1,000	0,409	0,416	0,990	0,193	0,009
IUT	0,640	0,061	0,101	0,998	0,494	0,502	0,989	0,219	0,009

Tracé 2D des Coordonnées Ligne & Colonne ; Dimension : 1 x 2

Table d'Entrée (Lignes x Colonnes) : 5 x 4

Standardisation : Profils ligne et colonne



Analyse des correspondances multiples

Tableau protocole : 3 questions, 7 modalités

	Sexe	Revenu	Preference
s1	F	M	A
s2	F	M	A
s3	F	E	B
s4	F	E	C
s5	F	E	C
s6	H	E	C
s7	H	E	B
s8	H	M	B
s9	H	M	B
s10	H	M	A

Tableau disjonctif complet

	Sexe: F	Sexe: H	Rev: M	Rev:E	Pref:A	Pref:B	Pref:C
s1	1	0	1	0	1	0	0
s2	1	0	1	0	1	0	0
s3	1	0	0	1	0	1	0
s4	1	0	0	1	0	0	1
s5	1	0	0	1	0	0	1
s6	0	1	0	1	0	0	1
s7	0	1	0	1	0	1	0
s8	0	1	1	0	0	1	0
s9	0	1	1	0	0	1	0
s10	0	1	1	0	1	0	0

La disjonction complète

DEPARTEMENTS	BLE	VIN	LAIT
DEP 1	NON	ROUGE	PEU
DEP 2	OUI	ROSE	MOYEN
DEP 3	OUI	BLANC	MOYEN

LA DISJONCTION EST UNE CODIFICATION EN DONNEES BINAIRES

CREATION D'UNE VARIABLE POUR CHAQUE MODALITE

DEPARTEMENTS	BLE		VIN			LAIT		
	OUI	NON	ROUGE	ROSE	BLANC	PEU	MOYEN	BCP
DEP 1	0	1	1	0	0	1	0	0
DEP 2	1	0	0	1	0	0	1	0
DEP 3	1	0	0	0	1	0	1	0

Tableau d'effectifs ou tableau des patrons de réponses

Sexe	Revenu	Preference	Effectif
F	M	A	2
F	E	B	1
F	E	C	2
H	E	C	1
H	E	B	1
H	M	B	2
H	M	A	1

Tableau disjonctif des patrons de réponses

	Sexe: F	Sexe: H	Rev: M	Rev:E	Pref:A	Pref:B	Pref:C
FMA	2	0	2	0	2	0	0
FEB	1	0	0	1	0	1	0
FEC	2	0	0	2	0	0	2
HEC	0	1	0	1	0	0	1
HEB	0	1	0	1	0	1	0
HMB	0	2	2	0	0	2	0
HMA	0	1	1	0	1	0	0

Tableau de Burt

	F	H	M	E	A	B	C
Sexe:F	5	0	2	3	2	1	2
Sexe:H	0	5	3	2	1	3	1
Revenu:M	2	3	5	0	3	2	0
Revenu:E	3	2	0	5	0	2	3
Preference:A	2	1	3	0	3	0	0
Preference:B	1	3	2	2	0	4	0
Preference:C	2	1	0	3	0	0	3

Le tableau de BURT

Si X est une matrice disjonctive complète
La Matrice de BURT est tXX

				BLE		VIN			LAIT		
				OUI	NON	Rouge	Rosé	Blanc	Peu	Moyen	Bcp
tX				0	1	1	0	0	1	0	0
X				1	0	0	1	0	0	1	0
				1	0	0	0	1	0	1	0
OUI	0	1	1	2	0	0	1	1	0	2	0
NON	1	0	0	0	1	1	0	0	1	0	0
Rouge	1	0	0	0	1	1	0	0	1	0	0
Rosé	0	1	0	1	0	0	1	0	0	1	0
Blanc	0	0	1	1	0	0	0	1	0	1	0
Pau	1	0	0	0	1	1	0	0	1	0	0
Moyen	0	1	1	2	0	0	1	1	0	2	0
Bcp	0	0	0	0	0	0	0	0	0	0	0

MATRICE DE BURT

tXX

Tous les tris simples
Tous les tris croisés

Propriété de l'analyse des correspondances (simple)

Lorsqu'il y a deux variables qualitatives réunies dans un tableau disjonctif $X = [X_1 | X_2]$, l'analyse factorielle des correspondances du tableau disjonctif est équivalente à l'analyse des correspondances du tableau de contingence $N = {}^T X_1 \ X_2$

Analyse des correspondances multiples

Effectuer l'analyse des correspondances multiples, c'est effectuer l'analyse factorielle des correspondances du tableau disjonctif complet, muni des relations $K < Q >$ (modalités emboîtées dans les questions) et $I < K < q >$ (individus emboîtés dans les modalités de chaque question).

Rouanet et Le Roux

Résultats produits par l'ACM sur le tableau suivant :

	Sexe	Revenu	Preference
s1	F	M	A
s2	F	M	A
s3	F	E	B
s4	F	E	C
s5	F	E	C
s6	H	E	C
s7	H	E	B
s8	H	M	B
s9	H	M	B
s10	H	M	A

Valeurs propres

Valeurs Propres et Inertie de toutes les Dimensions (Protocole dans Mini-ACM.stw) Table d'Entrée (Lignes x Colonnes) : 7 x 7 (Table de Burt) Inertie Totale = 1,3333

	ValSing.	ValProp.	%age	%age	Chi²
1	0,776426	0,602837	45,21275	45,2128	25,37943
2	0,680961	0,463708	34,77810	79,9909	19,52211
3	0,450509	0,202959	15,22190	95,2128	8,54456
4	0,252646	0,063830	4,78725	100,0000	2,68724

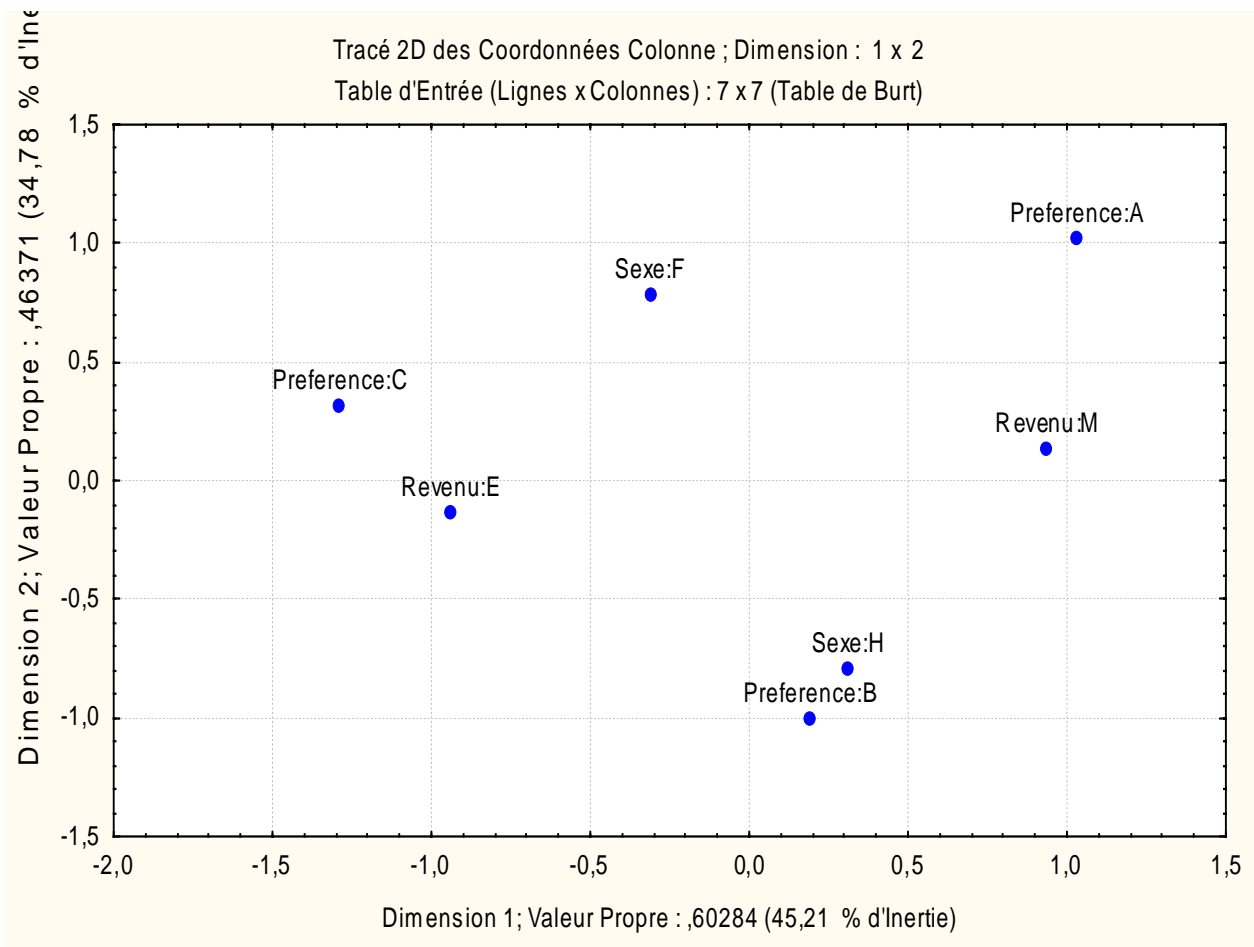
Valeurs propres : décroissance lente -> taux d'inertie modifiés de Benzécri

Calcul des taux modifiés :

	ValProp.	1/Q	(VP-1/Q)^2	%age
1	0,6028	0,3333	0,0726	81,04%
2	0,4637	0,3333	0,0170	18,96%
3	0,2030			
4	0,0638			
Somme	1,3333		0,089630	

Coordonnées, inertie et cosinus carrés

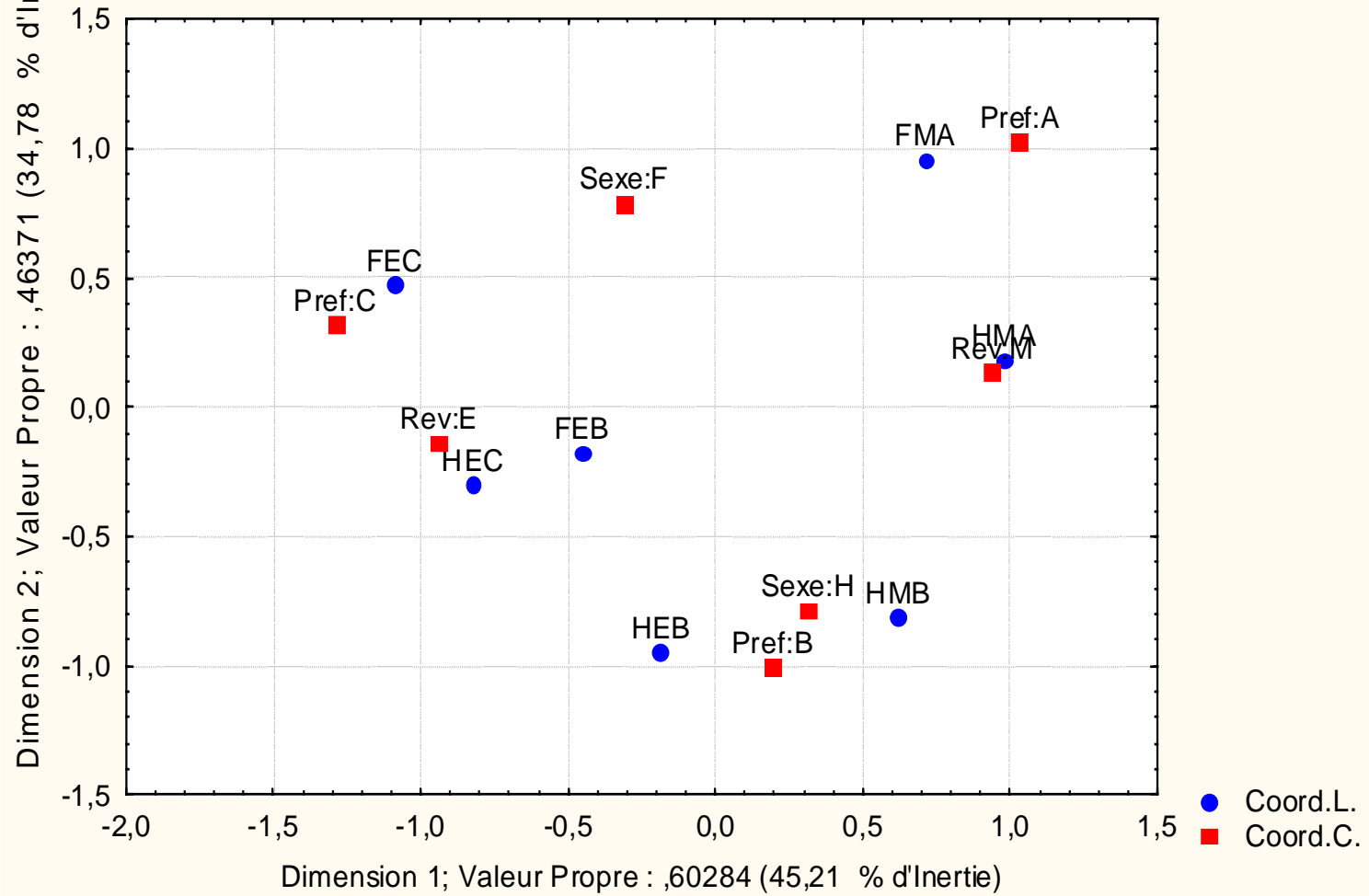
Coordonnées Colonne et Contributions à l'Inertie (Protocole dans Mini-ACM.stw)										
Table d'Entrée (Lignes x Colonnes) : 7 x 7 (Table de Burt)										
Inertie Totale = 1,3333										
NomLigne	Ligne Numéro	Coord. Dim.1	Coord. Dim.2	Masse	Qualité	Inertie Relative	Inertie Dim.1	Cosinus ² Dim.1	Inertie Dim.2	Cosinus ² Dim.2
Sexe:F	1	-0,31	0,79	0,17	0,72	0,12	0,03	0,10	0,22	0,62
Sexe:H	2	0,31	-0,79	0,17	0,72	0,12	0,03	0,10	0,22	0,62
Revenu:M	3	0,94	0,14	0,17	0,90	0,13	0,24	0,88	0,01	0,02
Revenu:E	4	-0,94	-0,14	0,17	0,90	0,13	0,24	0,88	0,01	0,02
Preference:A	5	1,03	1,02	0,10	0,91	0,18	0,18	0,46	0,23	0,45
Preference:B	6	0,19	-1,01	0,13	0,70	0,15	0,01	0,02	0,29	0,68
Preference:C	7	-1,29	0,32	0,10	0,75	0,18	0,28	0,71	0,02	0,04



Tracé 2D des Coordonnées Ligne & Colonne ; Dimension : 1 x 2

Table d'Entrée (Lignes x Colonnes) : 7 x 7

Standardisation : Profils ligne et colonne



Propriétés algébriques et géométriques de l'ACM

Valeur du Phi-2 :

$$\Phi^2 = \frac{K - Q}{Q} = \frac{\text{Nombre de modalités} - \text{Nombre de questions}}{\text{Nombre de questions}}$$

Contributions absolues et relatives des modalités colonnes à l'inertie :

$$Cta(M_k) = \frac{1 - f_k}{Q} \qquad Ctr(M_k) = \frac{1 - f_k}{K - Q}$$

Inerties absolue et relative d'une question :

$$I(X_q) = \frac{K_q - 1}{Q}$$

K_q : nombre de modalités de la question q

$$Inr(X_q) = \frac{K_q - 1}{K - Q} = \frac{\text{Nb de modalités de la question} - 1}{\text{Nb total de modalités} - \text{Nb de questions}}$$

Distances entre profils lignes :

$$d_{\Phi^2}^2(\text{Patron } i, \text{Patron } i') = \frac{1}{\text{Nb de Questions}} \sum \frac{1}{\text{fréquence de la modalité } k}$$

Somme étendue à toutes les modalités faisant partie de l'un des deux patrons, sans faire partie des deux patrons

Distance d'une ligne au profil moyen

$$d_{\Phi^2}^2(O, \text{Patron } i) = \left(\frac{1}{\text{Nombre de Questions}} \sum \frac{1}{\text{fréquence de la modalité } k} \right)^{-1}$$

Somme étendue à toutes les modalités faisant partie du patron i

Distances entre profils colonnes :

$$d_{\Phi^2}^2(M_k, M_{k'}) = \frac{1}{f_k} + \frac{1}{f_{k'}} - 2 \frac{f_{kk'}}{f_k f_{k'}} = \frac{n_k + n_{k'} - 2n_{kk'}}{n_k n_{k'} / n}$$

$$d_{\Phi^2}^2(M_k, M_{k'}) = \frac{\text{Effectif de } k + \text{Effectif de } k' - 2 \times \text{Effectif de la combinaison } k \text{ \& } k'}{\text{Effectif de } k \times \text{Effectif de } k' / \text{Effectif total}}$$

Distance d'une colonne au profil moyen :

$$d_{\Phi^2}^2(O, M_k) = \frac{1}{f_k} - 1 = \frac{n}{n_k} - 1 = \frac{\text{Effectif total}}{\text{Effectif de } k} - 1$$

1) Indépendance des modalités M_k et $M_{k'}$:

$$d^2(M_k, M_{k'}) = d^2(O, M_k) + d^2(O, M_{k'})$$

Autrement dit, dans l'espace multidimensionnel, le triangle $OM_kM_{k'}$ est alors un triangle rectangle en O .

2) Si les modalités M_k et $M_{k'}$ s'attirent, l'angle $(OM_k, OM_{k'})$ est un angle aigu.

3) Si les modalités M_k et $M_{k'}$ se repoussent, l'angle $(OM_k, OM_{k'})$ est un angle obtus.

4) Si l'effectif conjoint $n_{kk'}$ des modalités M_k et $M_{k'}$ est nul (en particulier si M_k et $M_{k'}$ sont deux modalités d'une même question) :

$$d^2(M_k, M_{k'}) = d^2(O, M_k) + d^2(O, M_{k'}) + 2$$

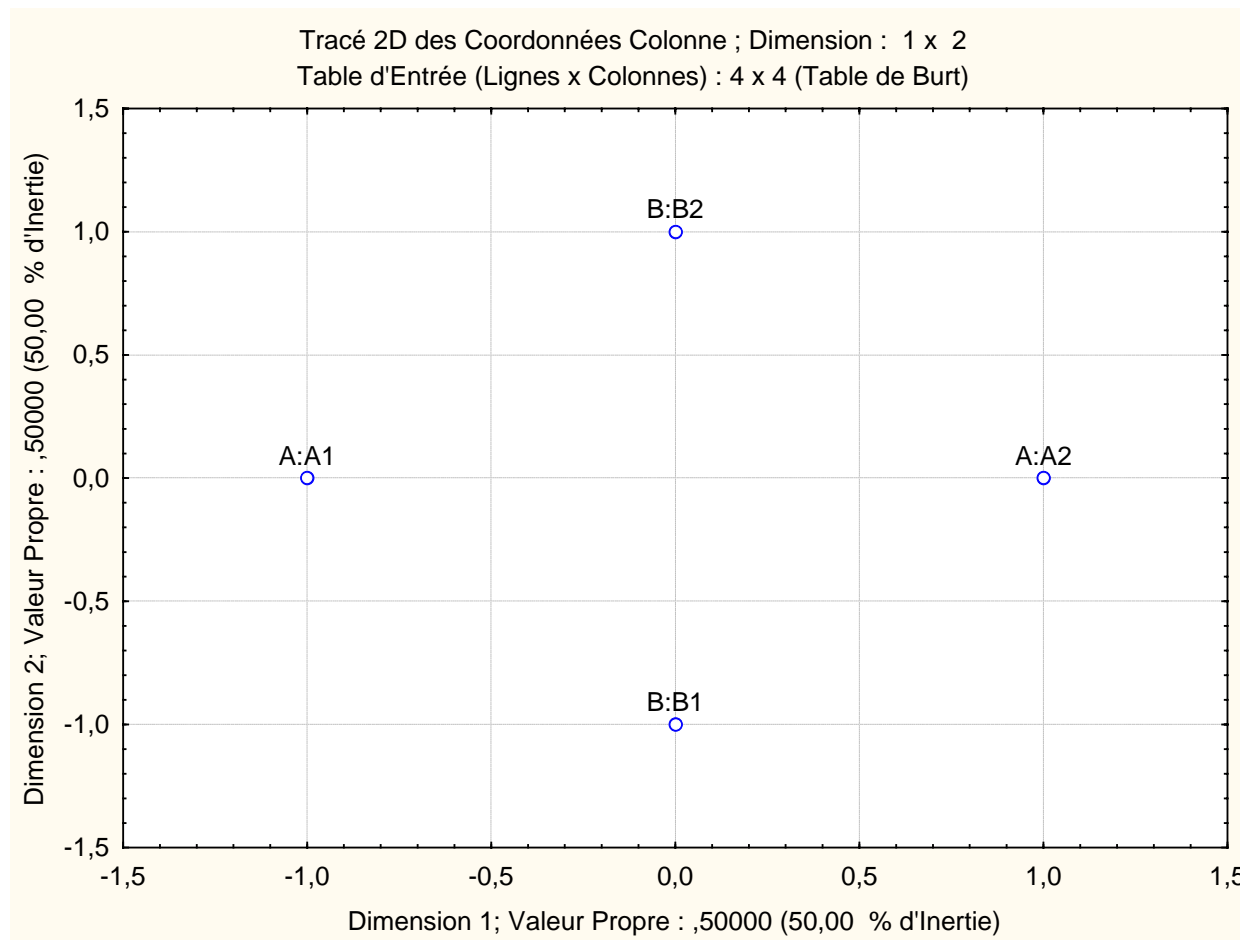
Deux questions à deux modalités chacune.

Cas 1 : les effectifs des modalités sont donnés par :

	A1	A2	Total
B1	50	50	100
B2	50	50	100
Total	100	100	200

Prévoir la forme de la représentation par rapport au premier plan factoriel.

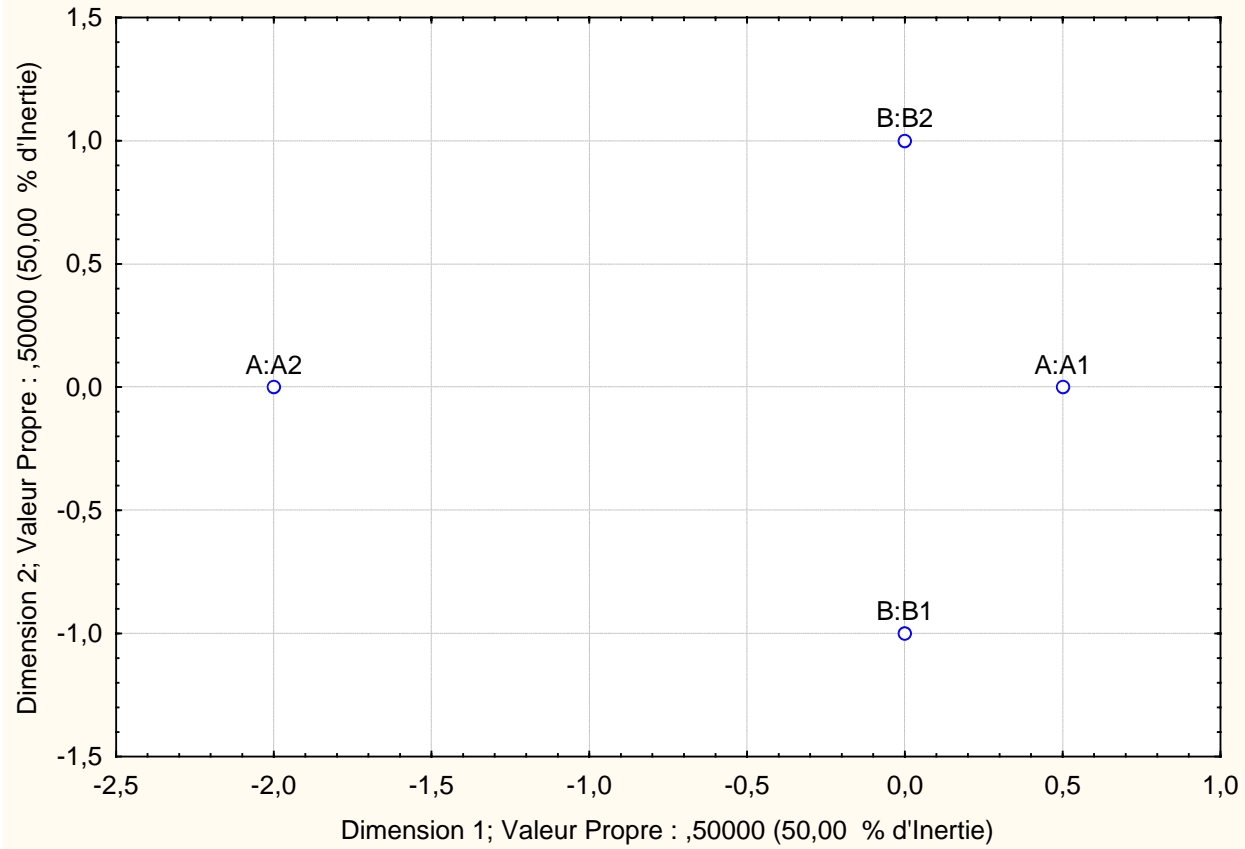
Réponse :



Cas 2 : les effectifs des modalités sont donnés par :

	A1	A2	Total
B1	80	20	100
B2	80	20	100
Total	160	40	200

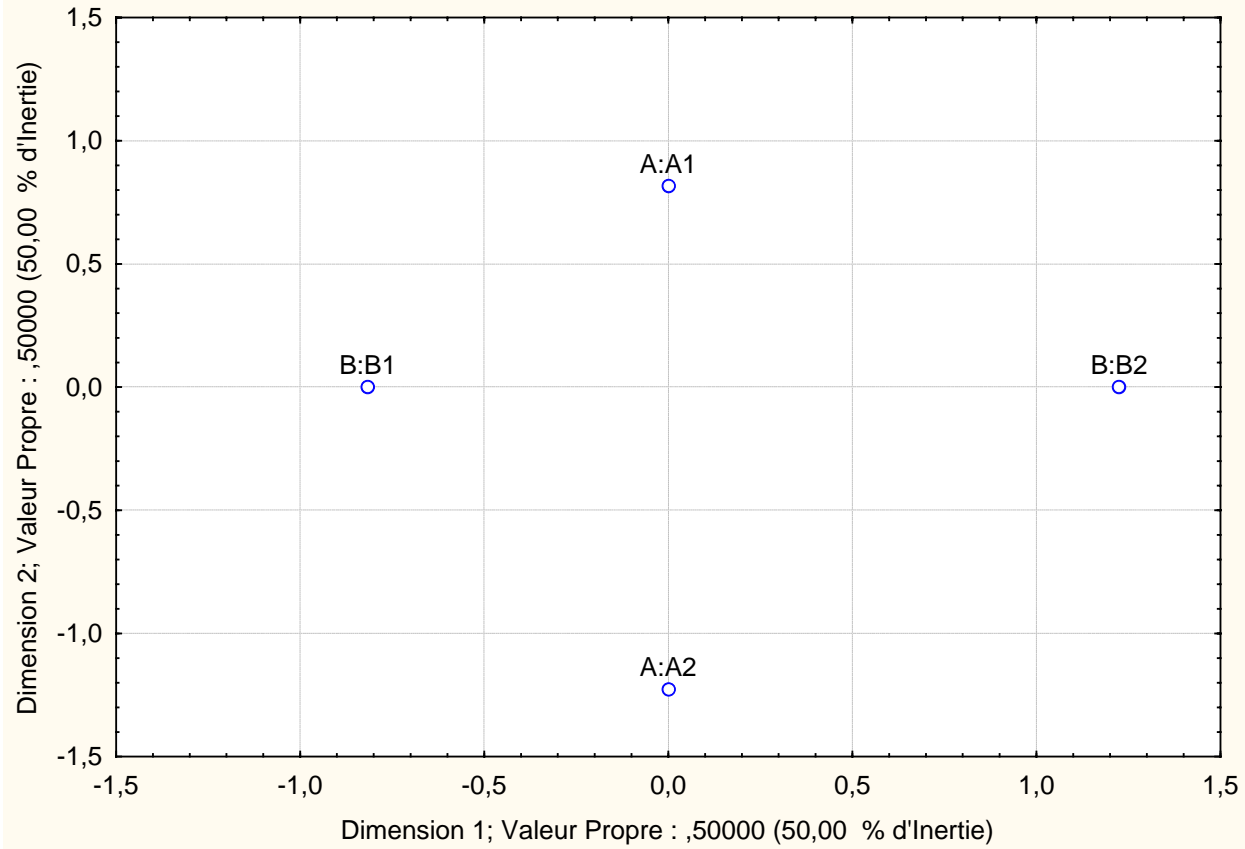
Tracé 2D des Coordonnées Colonne ; Dimension : 1 x 2
Table d'Entrée (Lignes x Colonnes) : 4 x 4 (Table de Burt)



Cas 3 : les effectifs des modalités sont donnés par :

	A1	A2	Total
B1	72	48	120
B2	48	32	80
Total	120	80	200

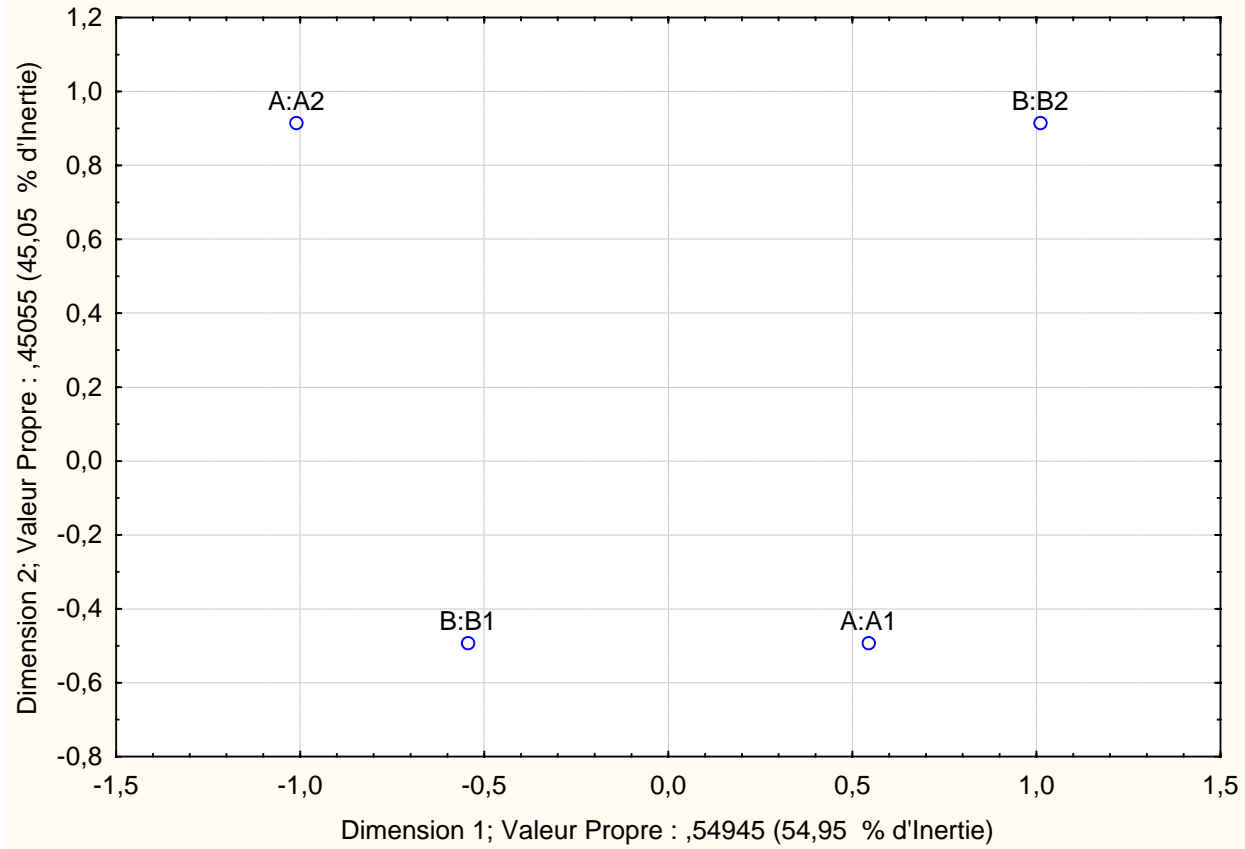
Tracé 2D des Coordonnées Colonne ; Dimension : 1 x 2
Table d'Entrée (Lignes x Colonnes) : 4 x 4 (Table de Burt)



Cas 4 : les effectifs des modalités sont donnés par :

	A1	A2	Total
B1	80	50	130
B2	50	20	70
Total	130	70	200

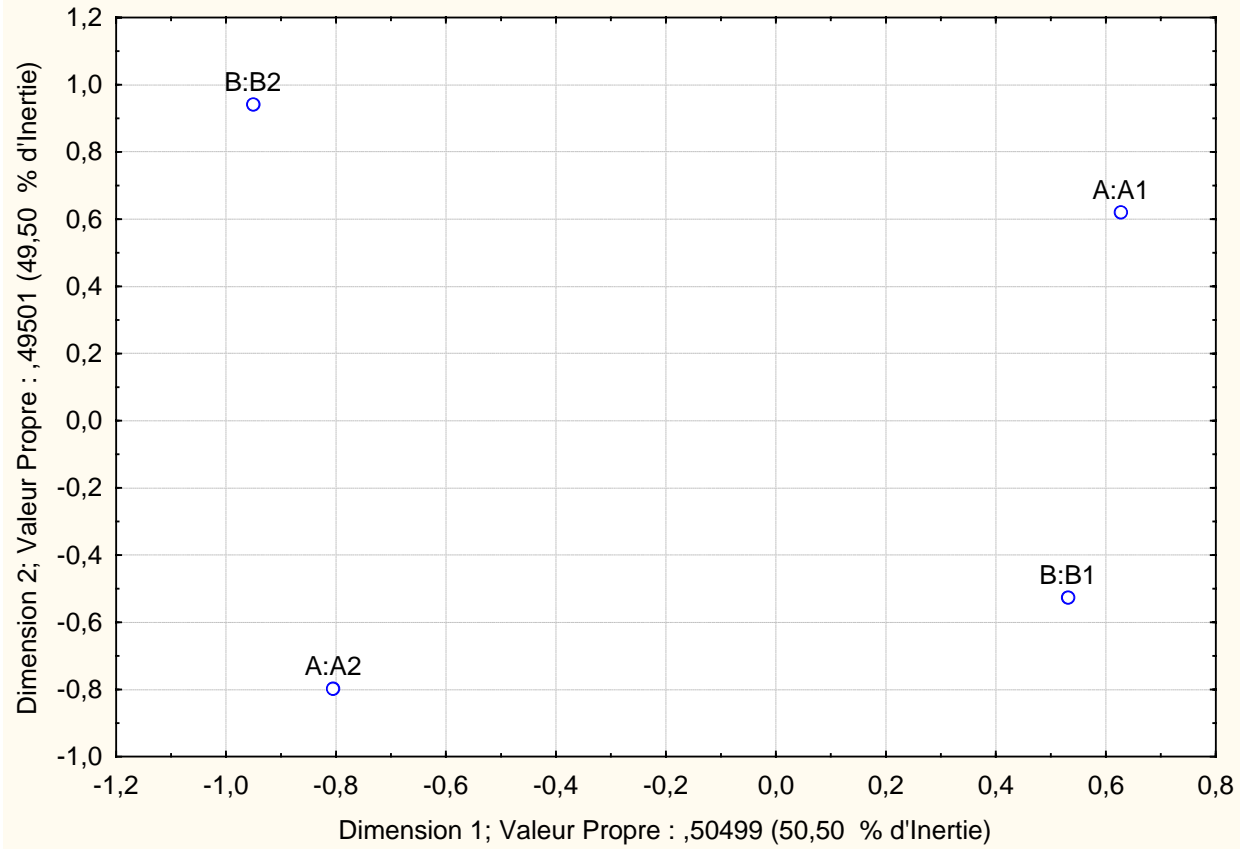
Tracé 2D des Coordonnées Colonne ; Dimension : 1 x 2
Table d'Entrée (Lignes x Colonnes) : 4 x 4 (Table de Burt)



Cas 5 : les effectifs des modalités sont donnés par :

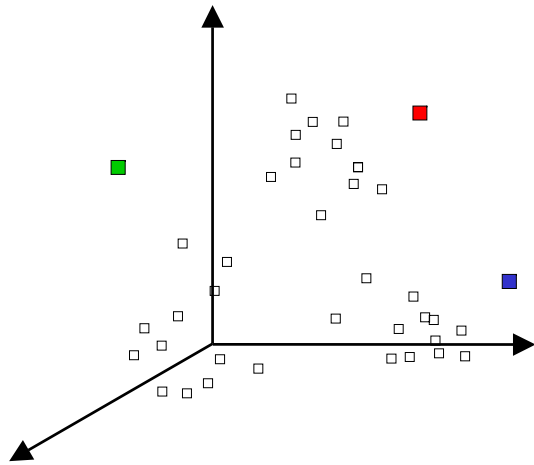
	A1	A2	Total
B1	73	56	129
B2	40	32	72
Total	113	88	201

Tracé 2D des Coordonnées Colonne ; Dimension : 1 x 2
Table d'Entrée (Lignes x Colonnes) : 4 x 4 (Table de Burt)



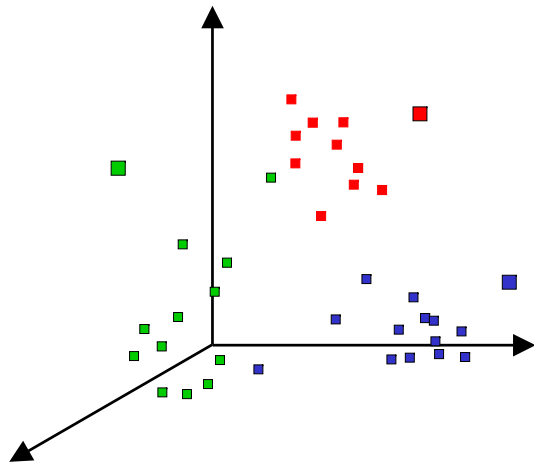
Méthodes de classification

Méthodes de type « centres mobiles »



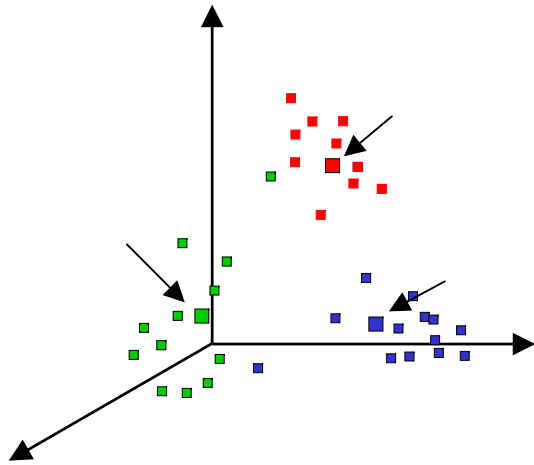
Au départ

Création aléatoire de centres de gravité.



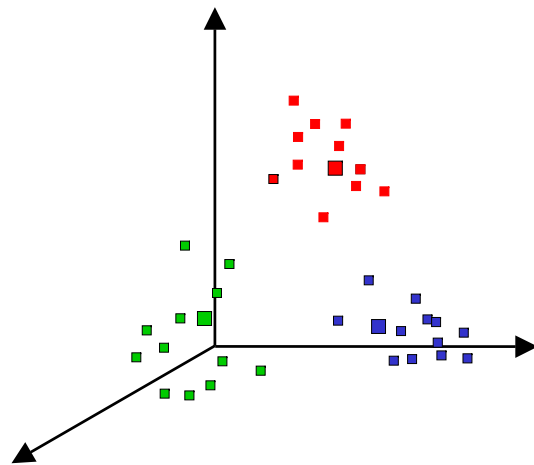
Etape 1

Chaque observation est classée en fonction de sa proximité aux centres de gravités.



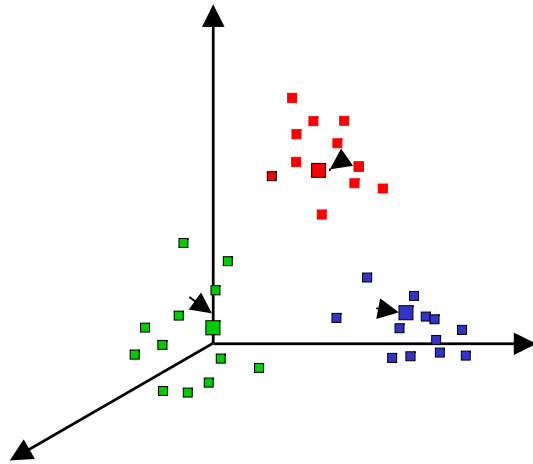
Etape 2

Chaque centre de gravité est déplacé de manière à être au centre du groupe correspondant



Etape 1'

On répète l'étape 1 avec les nouveaux centres de gravité.



Etape 2'

De nouveau, chaque centre de gravité est recalculé.

On continue jusqu'à ce que les centres de gravité ne bougent plus.

Exemple : typicalité des odeurs dans 3 cultures : FR, US, VN

Extrait des données

	1	2	3	4	5	6	7	8	9
	animal-FR	animal-US	animal-VN	bakery-FR	bakery-US	bakery-VN	candy-FR	candy-US	candy-VN
amber	1,17	1,05	1,87	1,83	1,16	2,53	2,27	1,58	2,73
anise	1,07	1	1,63	2,2	2,16	2,67	6,03	6,05	3,53
apricot	1,07	1,26	1,63	2,3	1,95	3,4	5,23	3,84	5,03
blackcurrant	1,03	1,37	1,43	2,2	1,74	3,27	6,63	5,42	4,43
butter	2,97	2,26	2,03	1,37	3,16	3,43	1,3	1,32	3,4
cat pee	3,5	5,21	2,07	1	1,05	1,73	1,03	1,37	1,7
cinnamon	1,33	1,11	1,57	2,67	4,26	2,87	2,47	5,11	3,2
civet	4,8	4,79	3,97	1	1,05	1,57	1,07	1,26	1,4
clove	1,23	1,37	2,27	1,43	2,89	1,83	1,37	2,63	1,8
cookies	1,03	1	1,33	4,37	4,33	5,4	5,27	4,33	4,83
detergent	1,37	1,37	2,37	1,13	1,32	1,73	1,2	1,16	1,9
eucalyptus	1,03	1,32	1,43	1	1,05	1,73	4,1	1,63	2,33
ginger	1,27	1,53	1,73	1,47	2,05	2,97	2,43	1,63	3,77
hazelnut	1,77	2,22	2,7	4,23	2,94	2,63	3,27	3,22	2,33
honey	2,03	2,33	2,67	2,1	2,28	2,27	2,8	2,17	2,8
jasmine	1,7	1,63	2,57	1,13	1,47	2,1	1,23	2,05	2,33
lavender	1,07	1,05	1,8	1	1,16	1,83	1,27	1,47	2,47

Exemple : typicalité des odeurs dans 3 cultures : FR, US, VN

Classe 1

FR	Composition de la Classe 1 (Odors dans Odors.stw) et Distances au Centre de Classe Respectif Classe avec 8 obs.
	Distance
anise	0,853238
apricot	0,556198
blackcurrant	0,475439
cookies	0,885102
melon	0,854534
milk	0,538125
pineapple	0,616581
strawberry	0,397054

VN	Composition de la Classe 1 (Odors dans Odors.stw) et Distances au Centre de Classe Respectif Classe avec 6 obs.
	Distance
apricot	0,221989
blackcurrant	0,356261
cookies	0,589538
melon	0,385943
pineapple	0,337747
strawberry	0,291606

US	Composition de la Classe 5 (Odors dans Odors.stw) et Distances au Centre de Classe Respectif Classe avec 5 obs.
	Distance
apricot	0,597914
blackcurrant	0,234945
melon	0,419215
pineapple	0,312601
strawberry	0,331206

Exemple : typicalité des odeurs dans 3 cultures : FR, US, VN

Classe 2

FR	Composition de la Classe 2 (Odors dans Odors.stw) et Distances au Centre de Classe Respectif Classe avec 8 obs.
	Distance
amber	0,557463
jasmine	0,585252
lavender	0,868097
mango	0,378329
orange blossom	0,588759
rose	0,451861
vanilla	0,874072
violet	0,546793

VN	Composition de la Classe 3 (Odors dans Odors.stw) et Distances au Centre de Classe Respectif Classe avec 8 obs.
	Distance
anise	0,366928
butter	0,344654
cinnamon	0,506660
ginger	0,282658
mango	0,411693
milk	0,314841
rose	0,444952
vanilla	0,402814

US	Composition de la Classe 4 (Odors dans Odors.stw) et Distances au Centre de Classe Respectif Classe avec 9 obs.
	Distance
anise	0,905972
cinnamon	0,884537
clove	0,830185
cookies	0,579065
hazelnut	0,739649
milk	0,474403
peanut	0,495004
vanilla	0,457066
walnut	0,600739

Exemple : typicalité des odeurs dans 3 cultures : FR, US, VN

Classe 3

FR	Composition de la Classe 3 (Odors dans Odors.stw) et Distances au Centre de Classe Respectif Classe avec 16 obs.
	Distance
butter	0,673286
cat pee	0,752818
cinnamon	0,810755
civet	1,056562
clove	0,818287
ginger	0,655657
hazelnut	1,093172
honey	0,605759
moldy	1,245204
mushroom	1,222151
nutmeg	0,669163
peanut	0,677682
tea	0,332057
truffle	0,551662
walnut	0,943539
woody	0,533684

VN	Composition de la Classe 2 (Odors dans Odors.stw) et Distances au Centre de Classe Respectif Classe avec 10 obs.
	Distance
cat pee	0,348711
civet	0,571433
hazelnut	0,354241
leather	0,348657
moldy	0,850417
mushroom	0,439731
nutmeg	0,302967
peanut	0,348855
truffle	0,380382
woody	0,252038

US	Composition de la Classe 2 (Odors dans Odors.stw) et Distances au Centre de Classe Respectif Classe avec 10 obs.
	Distance
butter	0,686537
cat pee	0,830411
civet	0,712920
honey	0,502343
leather	0,551503
moldy	0,694570
mushroom	0,602322
tea	0,417894
truffle	0,512148
woody	0,487014

Exemple : typicalité des odeurs dans 3 cultures : FR, US, VN

Classe 4

FR	Composition de la Classe 4 (Odors dans Odors.stw) et Distances au Centre de Classe Respectif Classe avec 6 obs.	
	Distance	
detergent	0,433169	
leather	0,445424	
moth ball	0,362147	
musk	0,483415	
pine	0,471897	
soap	0,583448	

VN	Composition de la Classe 4 (Odors dans Odors.stw) et Distances au Centre de Classe Respectif Classe avec 12 obs.	
	Distance	
clove	0,431683	
detergent	0,299191	
honey	0,441126	
jasmine	0,396467	
lavender	0,326566	
moth ball	0,371675	
musk	0,280186	
orange blossom	0,531775	
pine	0,395162	
soap	0,573349	
tea	0,357944	
violet	0,174551	

US	Composition de la Classe 3 (Odors dans Odors.stw) et Distances au Centre de Classe Respectif Classe avec 12 obs.	
	Distance	
amber	0,687928	
detergent	0,424100	
jasmine	0,652539	
lavender	0,461172	
mango	0,434236	
moth ball	0,814012	
musk	0,584099	
orange blossom	0,541148	
pine	0,579675	
rose	0,542605	
soap	0,883078	
violet	0,560355	

Exemple : typicalité des odeurs dans 3 cultures : FR, US, VN

Classe 5

FR	Composition de la Classe 5 (Odors dans Odors.stw) et Distances au Centre de Classe Respectif Classe avec 2 obs.	
	Distance	
eucalyptus	0,462265	
wintergreen	0,462265	

VN	Composition de la Classe 5 (Odors dans Odors.stw) et Distances au Centre de Classe Respectif Classe avec 4 obs.	
	Distance	
amber	0,435829	
eucalyptus	0,291722	
walnut	0,313654	
wintergreen	0,278881	

US	Composition de la Classe 1 (Odors dans Odors.stw) et Distances au Centre de Classe Respectif Classe avec 4 obs.	
	Distance	
eucalyptus	0,794737	
ginger	0,586028	
nutmeg	0,709252	
wintergreen	0,780444	

Classification Ascendante Hiérarchique

Les quatre étapes de la méthode :

- Choix des variables représentant les individus
- Choix d'un indice de dissimilarité
- Choix d'un indice d'agrégation
- Algorithme de classification et résultat produit

Quelques distances ou indices de dissimilarité

- Distance Euclidienne.
$$d(I_i, I_j) = \sqrt{\sum_k (x_{ik} - x_{jk})^2}$$
- Distance Euclidienne au carré.
$$d(I_i, I_j) = \sum_k (x_{ik} - x_{jk})^2$$
- Distance du City-block (Manhattan) :
$$d(I_i, I_j) = \sum_k |x_{ik} - x_{jk}|$$
- Distance de Tchebychev :
$$d(I_i, I_j) = \text{Max} |x_{ik} - x_{jk}|$$
- Distance à la puissance.
$$d(I_i, I_j) = \left(\sum_k |x_{ik} - x_{jk}|^p \right)^{1/p}$$
- Percent disagreement.
$$d(I_i, I_j) = \frac{\text{Nombre de } x_{ik} \neq x_{jk}}{K}$$
- 1- r de Pearson :
$$d(I_i, I_j) = 1 - r_{ij}$$

Quelques indices d'agrégation

- Diamètre ou « complete linkage » : $D(A,B) = \max_{I \in A} \max_{J \in B} d(I,J)$

- Moyenne non pondérée des groupes associés: $D(A,B) = \frac{1}{n_A n_B} \sum_{I \in A, J \in B} d(I,J)$

- Moyenne pondérée des groupes associés : $D(A,B) = \frac{1}{(n_A + n_B)(n_A + n_B - 1)} \sum_{I, J \in A \cup B} d(I,J)$

- Centroïde non pondéré des groupes associés.

- Centroïde pondéré des groupes associés (médiane).

- Méthode de Ward (méthode du moment d'ordre 2). Si une classe M est obtenue en regroupant les classes K et L, sa distance à la classe J est donnée par :

$$D(M,J) = \frac{(N_J + N_K)D(K,J) + (N_J + N_L)D(L,J) - N_J D(K,L)}{N_J + N_K + N_L}$$

L'algorithme de classification

Étape 1 : n éléments à classer ;

Étape 2 : Construction de la matrice de distances entre les n éléments et recherche les deux plus proches, que l'on agrège en un nouvel élément. On obtient une première partition à $n-1$ classes;

Étape 3 : Construction d'une nouvelle matrice des distances qui résultent de l'agrégation, en calculant les distances entre le nouvel élément et les éléments restants (les autres distances sont inchangées). Recherche des deux éléments les plus proches, que l'on agrège. On obtient une deuxième partition avec $n-2$ classes et qui englobe la première;

...

Étape m : on calcule les nouvelles distances, et l'on réitère le processus jusqu'à n'avoir plus qu'un seul élément regroupant tous les objets et qui constitue la dernière partition.

Distance Euclidienne au carré et méthode de Ward

Inertie totale = Inertie « intra » + Inertie « inter »

A chaque étape, on réunit les deux classes de façon à augmenter le moins possible l'inertie « intra »

$$I = \sum_{j=1}^g \sum_{i=1}^{n_j} G_j M_{ij}^2 + \sum_{j=1}^g n_j G G_j^2$$

Inertie totale = \sum Inertie dans les classes + Inertie des points moyens pondérés par les effectifs des classes

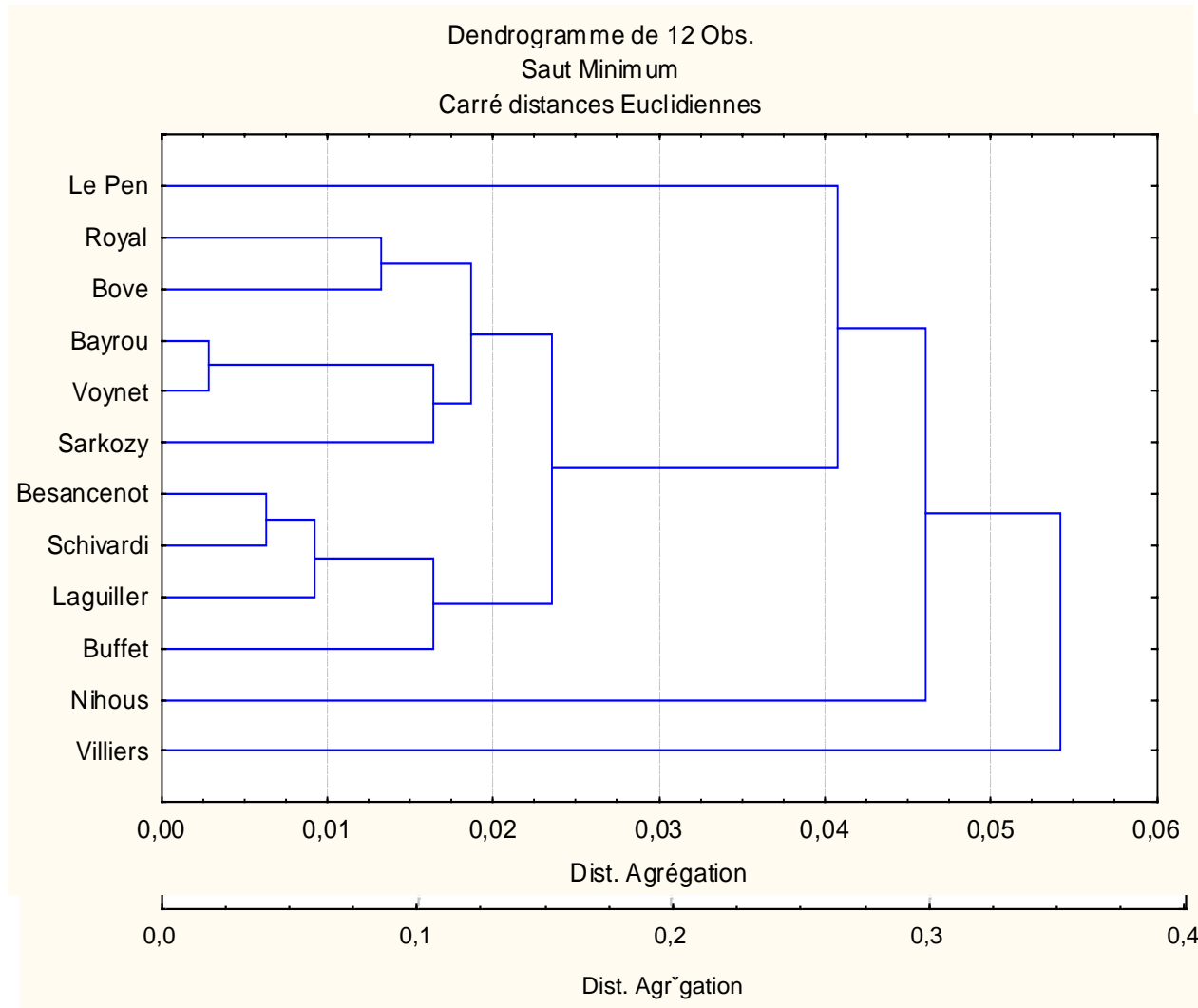
Résultat obtenu :

Une hiérarchie de classes telles que :

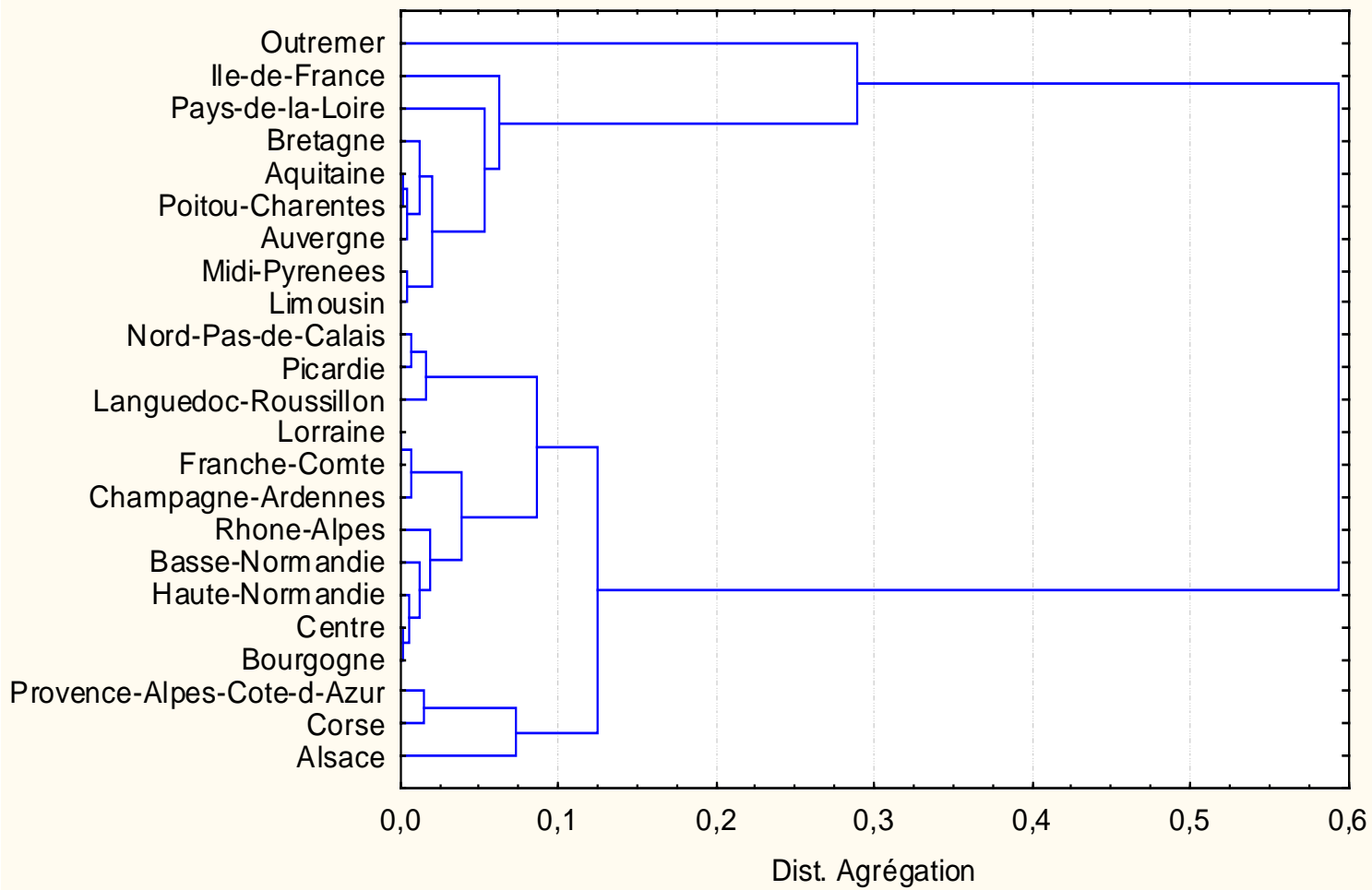
- toute classe est non vide
- tout individu appartient à une (et même plusieurs) classes
- deux classes distinctes sont disjointes, ou vérifient une relation d'inclusion (l'une d'elles est incluse dans l'autre)
- toute classe est la réunion des classes qui sont incluses dans elle.

Ce résultat est fréquemment représenté à l'aide d'un dendrogramme

Exemples de dendrogrammes



Dendrogramme de 23 Obs.
Méth. de Ward
Carré distances Euclidiennes



Régression linéaire Multiple

Echantillon de n individus statistiques :

- p variables numériques X_1, X_2, \dots, X_p (variables indépendantes ou explicatives)
- une variable numérique Y (variable dépendante, ou "à expliquer").

Exemple (30 comtés américains) :

VARI_POP : Variation de la Population (1960-1970)

N_AGRIC : Nb. de personnes travaillant dans le secteur primaire

TX_IMPOS : Taux d'imposition des propriétés

PT_PHONE : Pourcentage d'installations téléphoniques

PT_RURAL : Pourcentage de la population vivant en milieu rural

AGE : Age médian

PT_PAUVR : Pourcentage de familles en dessous du seuil de pauvreté

Matrice des corrélations

	VARI_POP	N_AGRIC	PT_PAUVR	TX_IMPOS	PT_PHONE	PT_RURAL	AGE
VARI_POP	1,00	0,04	-0,65	0,13	0,38	-0,02	-0,15
N_AGRIC	0,04	1,00	-0,17	0,10	0,36	-0,66	-0,36
PT_PAUVR	-0,65	-0,17	1,00	0,01	-0,73	0,51	0,02
TX_IMPOS	0,13	0,10	0,01	1,00	-0,04	0,02	-0,05
PT_PHONE	0,38	0,36	-0,73	-0,04	1,00	-0,75	-0,08
PT_RURAL	-0,02	-0,66	0,51	0,02	-0,75	1,00	0,31
AGE	-0,15	-0,36	0,02	-0,05	-0,08	0,31	1,00

Le modèle linéaire :

On cherche à exprimer Y sous la forme :

$$Y = b_0 + b_1 X_1 + b_2 X_2 + \dots + b_p X_p + E$$

où E (erreur commise en remplaçant Y par la valeur estimée) est nulle en moyenne, et de variance minimale.

Solution au problème :

Les coefficients b_i ($1 \leq i \leq p$) sont les solutions du système d'équations :

$$\begin{cases} Cov(X_1, X_1)b_1 + Cov(X_1, X_2)b_2 + \dots + Cov(X_1, X_p)b_p = Cov(X_1, Y) \\ Cov(X_2, X_1)b_1 + Cov(X_2, X_2)b_2 + \dots + Cov(X_2, X_p)b_p = Cov(X_2, Y) \\ \dots \\ Cov(X_p, X_1)b_1 + Cov(X_p, X_2)b_2 + \dots + Cov(X_p, X_p)b_p = Cov(X_p, Y) \end{cases}$$

et

$$b_0 = \bar{Y} - b_1 \bar{X}_1 - b_2 \bar{X}_2 - \dots - b_p \bar{X}_p$$

Sur l'exemple proposé :

$$\text{PT_PAUVR} = 31,2660 - 0,3923 \text{ VARI_POP} + 0,0008 \text{ N_AGRIC} + 1,2301 \text{ TX_IMPOS} - 0,0832 \text{ PT_PHONE} + 0,1655 \text{ PT_RURAL} - 0,4193 \text{ AGE}$$

Coefficients standardisés :

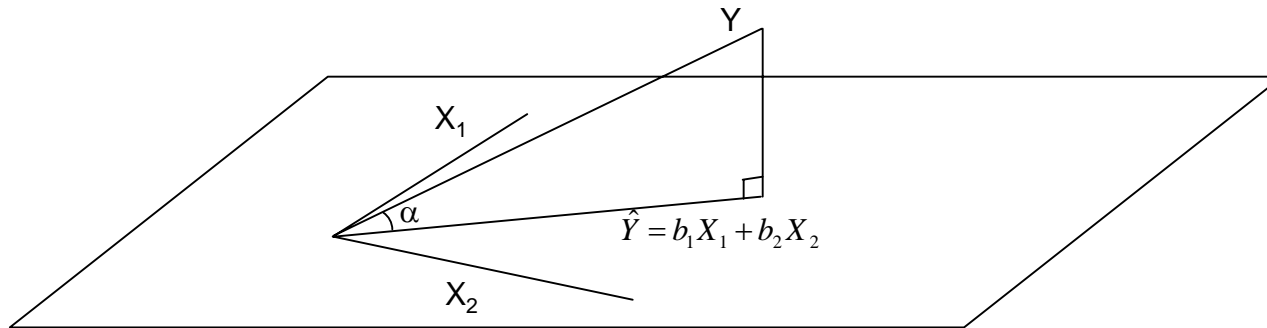
$$\beta_i = \frac{\sigma(X_i)}{\sigma(Y)} b_i$$

VARI_POP	N_AGRIC	TX_IMPOS	PT_PHONE	PT_RURAL	AGE
-0,630788	0,238314	0,038799	-0,129627	0,618746	-0,188205

Test des coefficients de la régression

	PT_PAU VR	PT_PAU VR	PT_PAU VR	PT_PAU VR	-95,00%	+95,00%
	(param.)	Err-Type	t	p	Lim.Conf	Lim.Conf
Ord.Orig.	31,2660	13,2651	2,3570	0,0273	3,8251	58,7070
VARI_POP	-0,3923	0,0805	-4,8742	0,0001	-0,5589	-0,2258
N_AGRIC	0,0008	0,0004	1,6903	0,1045	-0,0002	0,0017
TX_IMPOS	1,2301	3,1899	0,3856	0,7033	-5,3686	7,8288
PT_PHONE	-0,0832	0,1306	-0,6376	0,5300	-0,3533	0,1868
PT_RURAL	0,1655	0,0618	2,6766	0,0135	0,0376	0,2935
AGE	-0,4193	0,2554	-1,6415	0,1143	-0,9476	0,1091

Approche factorielle de la régression



Expliquer la variabilité de Y à partir de celle des X_j :

Combinaison linéaire des X_j qui reproduit « au mieux » la variabilité des individus selon Y : combinaison linéaire la plus corrélée avec Y .

Solution : combinaison linéaire des X_j qui fait avec Y un angle minimum.

Test de la régression :

Variance de Y = Variance expliquée + Variance résiduelle

$$Var(Y) = Var(\hat{Y}) + Var(Y - \hat{Y})$$

Analyse de variance

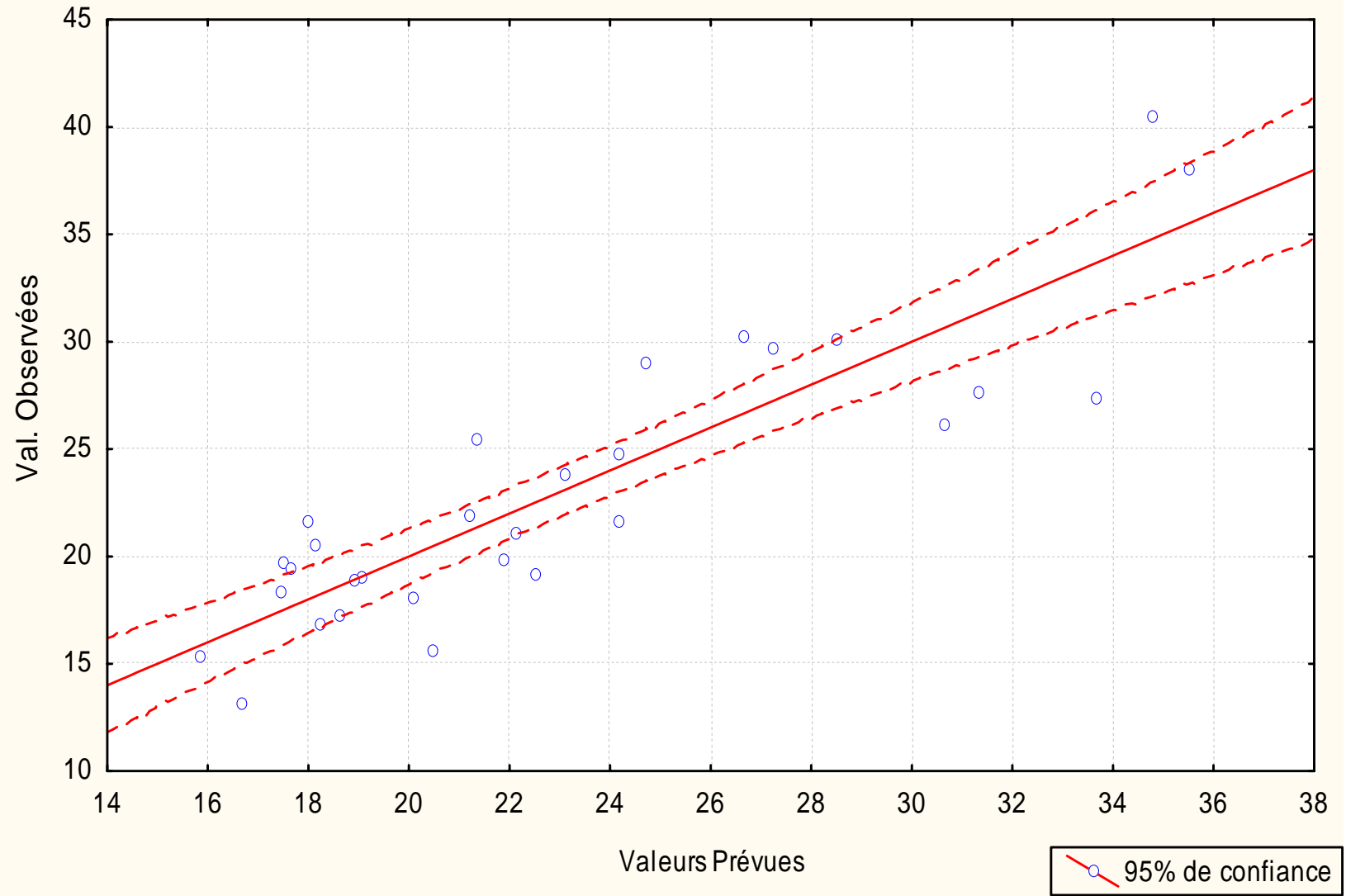
	Sommes	dl	Moyennes	F	niveau p
	Carrés		Carrés		
Régress.	932,065	6	155,3441	13,44909	0,000002
Résidus	265,662	23	11,5505		
Total	1197,727				

Coefficient de détermination :

$$R^2 = \frac{Var(\hat{Y})}{Var(Y)} = 0,7782$$

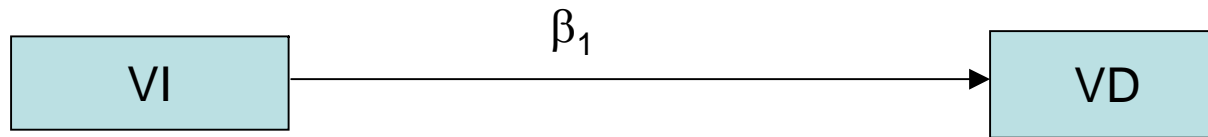
Valeurs Prévues vs. Observées

Var. dépendante : PT_PAUVR



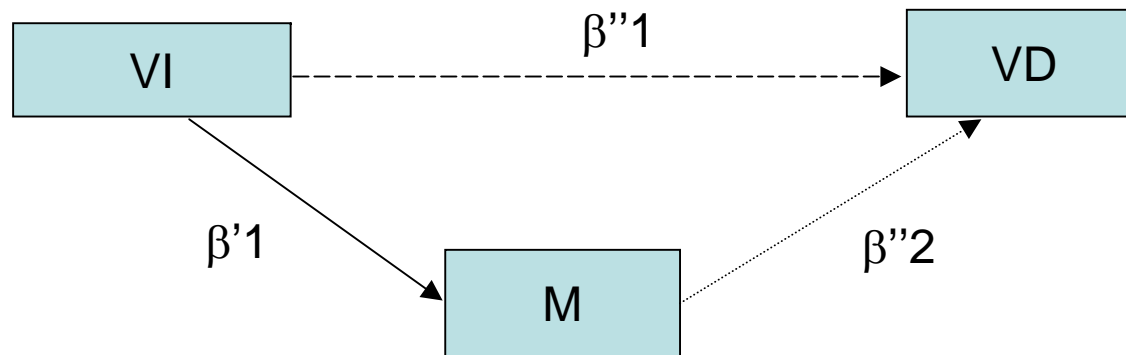
Analyse de médiation

1) Régression de la VD sur la VI : $VD = b_0 + b_1 VI$
Coefficient de régression standardisé : β_1



2) Régression de la médiation sur la VI : $M = b'_0 + b'_1 VI$
Coefficient de régression standardisé : β'_1

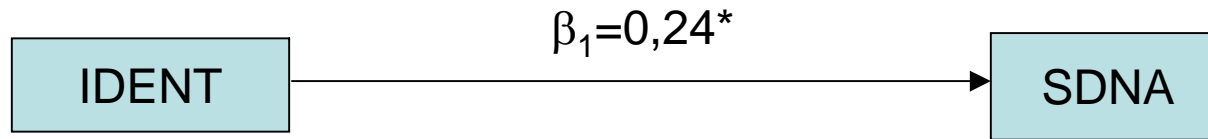
3) Régression multiple de la VD sur VI et M : $VD = b''_0 + b''_1 VI + b''_2 M$
Coefficients de régression standardisés : β''_1, β''_2



Interprétation :

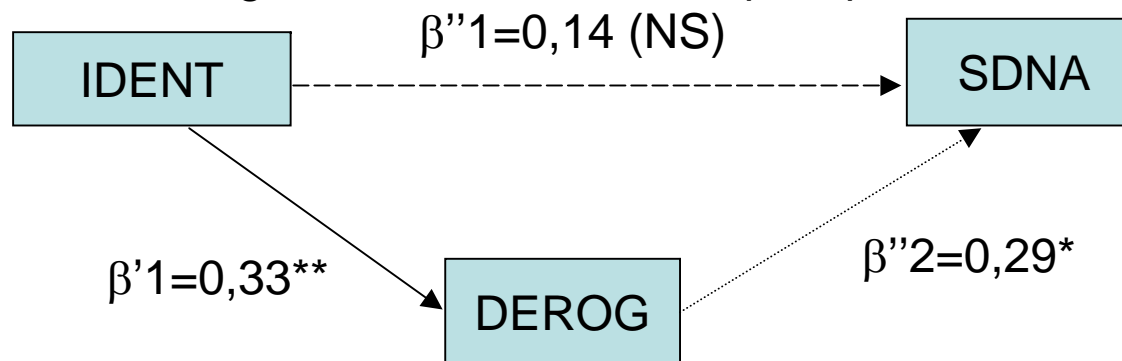
Si b''_1 est nettement plus proche de 0 que b_1 , en particulier si b''_1 n'est pas significativement différent de 0 alors que b_1 l'était, il y a médiation (partielle ou totale)

1) Régression de la VD sur la VI : $SDNA = b_0 + b_1 IDENT$
Coefficient de régression standardisé : β_1

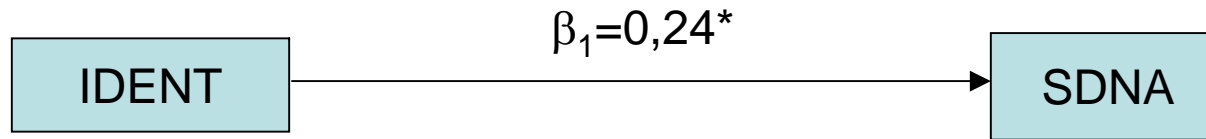


2) Régression de la médiation sur la VI : $DEROG = b'_0 + b'_1 IDENT$
Coefficient de régression standardisé : β'_1

3) Régression multiple de la VD sur VI et M :
 $SDNA = b''_0 + b''_1 IDENT + b''_2 DEROG$
Coefficients de régression standardisés : β''_1, β''_2

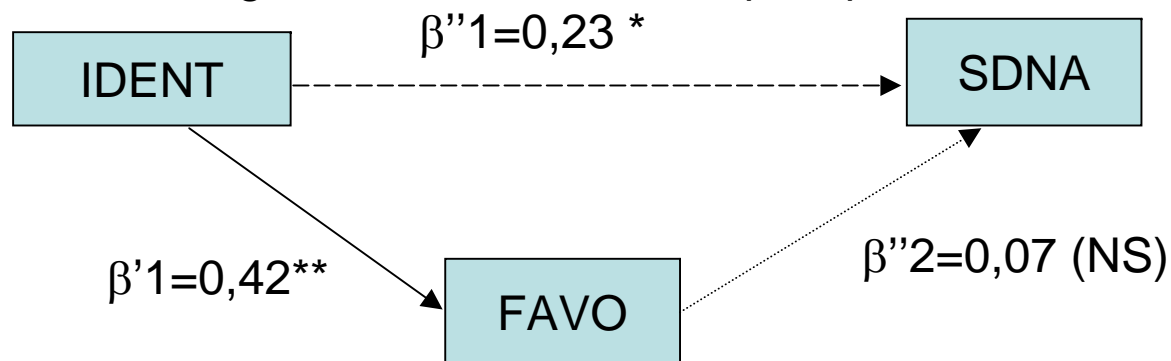


1) Régression de la VD sur la VI : $SDNA = b_0 + b_1 \text{ IDENT}$
Coefficient de régression standardisé : β_1



2) Régression de la médiation sur la VI : $DEROG = b'_0 + b'_1 \text{ IDENT}$
Coefficient de régression standardisé : β'_1

3) Régression multiple de la VD sur VI et M :
 $SDNA = b''_0 + b''_1 \text{ IDENT} + b''_2 \text{ DEROG}$
Coefficients de régression standardisés : β''_1, β''_2



Pas d'effet de médiation

Régression Logistique

Sur un échantillon de n individus statistiques, on a observé :

- p variables numériques ou dichotomiques X_1, X_2, \dots, X_p (variables indépendantes ou explicatives)
- une variable dichotomique Y (variable dépendante, ou "à expliquer").

Exemple :

Echantillon de 30 sujets pour lesquels on a relevé :

- d'une part le niveau des revenus (variable numérique)
- d'autre part la possession ou non d'un nouvel équipement électroménager.

Exemple

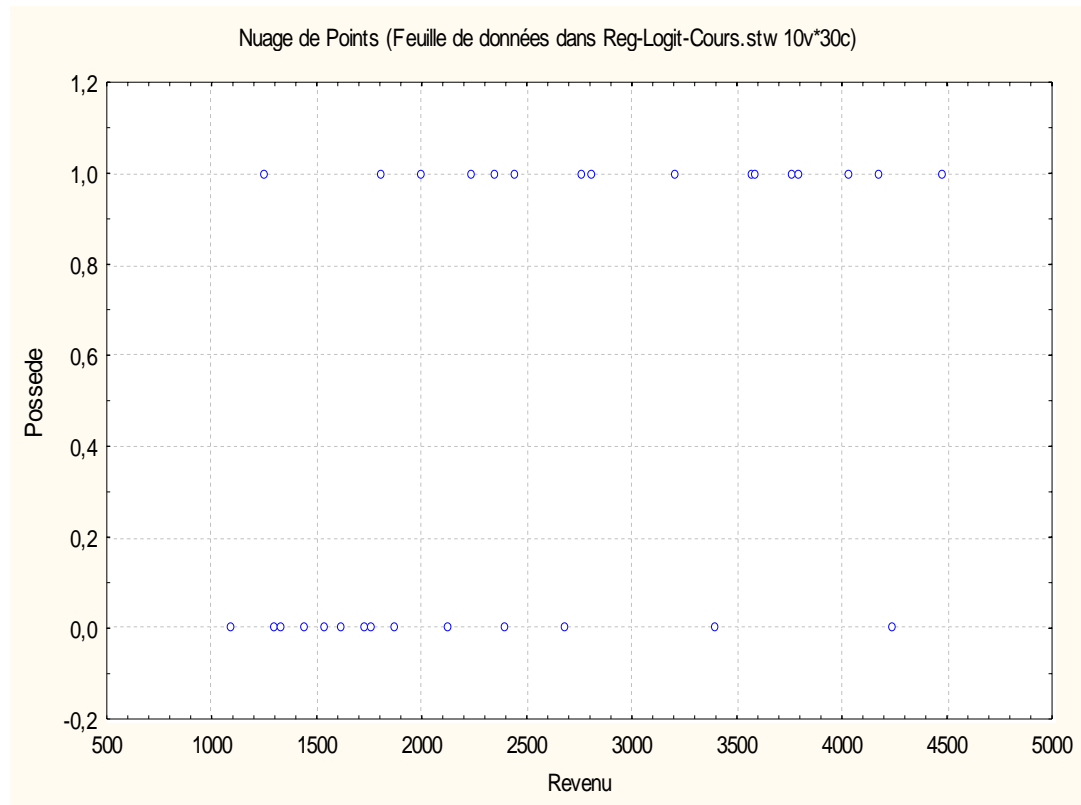
Revenu	1085	1304	1331	1434	1541	1612	1729	1759
Possède	0	0	0	0	0	0	0	0

Revenu	1863	2121	2395	2681	3390	4237	1241
Possède	0	0	0	0	0	0	1

Revenu	1798	1997	2234	2346	2436	2753	2813	3204
Possède	1	1	1	1	1	1	1	1

Revenu	3564	3592	3762	3799	4037	4168	4484
Possède	1	1	1	1	1	1	1

Nuage de points



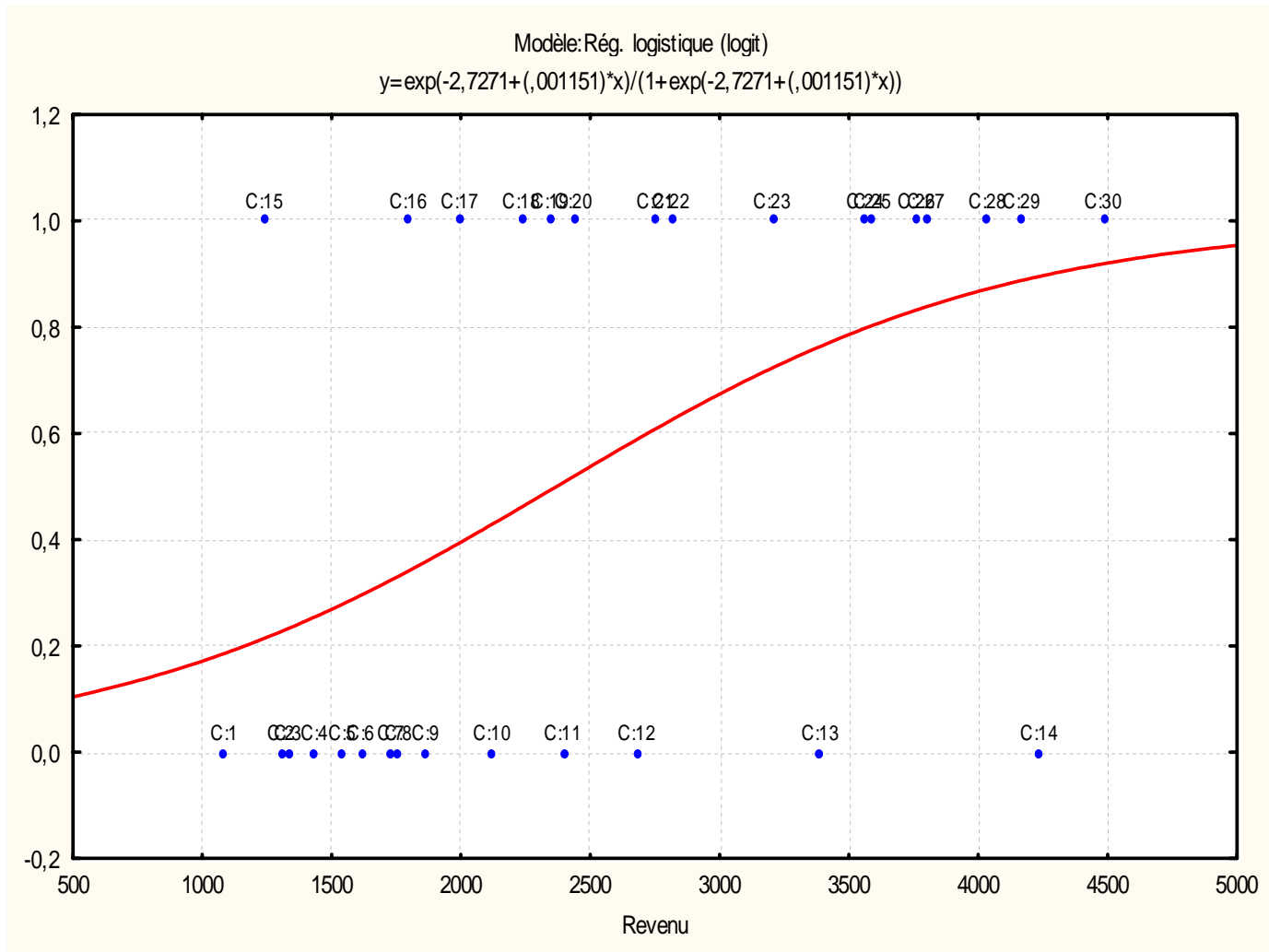
Rapport de chances et transformation logit

Rapport de chances ou cote :

$$p_1 = \frac{P(Y = 1)}{1 - P(Y = 1)}$$

Transformation logit

$$\text{logit}(P) = \ln\left(\frac{P}{1 - P}\right)$$



$$\text{logit}(Y) = -2,7271 + 0,001151 X$$

Aide à l'interprétation : odds-ratio ou rapport de cotes

La contribution de la variable X à la variation de Y est calculée par :

$$\text{OR} = \exp(\text{Coefficient de X dans le modèle})$$

L'odds-ratio correspondant au coefficient 0,001151 est :

$$e^{0,001151} = 1,0012.$$

Autrement dit, une augmentation du revenu de 1 unité se traduit par une multiplication de la probabilité par 1,0012.

L'odds-ratio est défini comme le rapport de deux rapports de chances. Ainsi, l'odds-ratio relatif à l'étendue des valeurs observées est défini de la manière suivante :

- On calcule le rapport de chances relatif à la plus grande valeur observée du revenu :

$$\text{Pour } X = 4484, P_1=0,919325 \text{ et } \frac{P_1}{1 - P_1} = 11,3954$$

- On calcule le rapport de chances relatif à la plus petite valeur observée du revenu :

$$\text{Pour } X = 1085, P_2=0,185658 \text{ et } \frac{P_2}{1 - P_2} = 0,2280$$

- L'odds-ratio est obtenu comme quotient des deux rapports précédents :

$$\text{OR} = \frac{\frac{P_1}{1 - P_1}}{\frac{P_2}{1 - P_2}} = \frac{11,3954}{0,2280} = 49,98$$

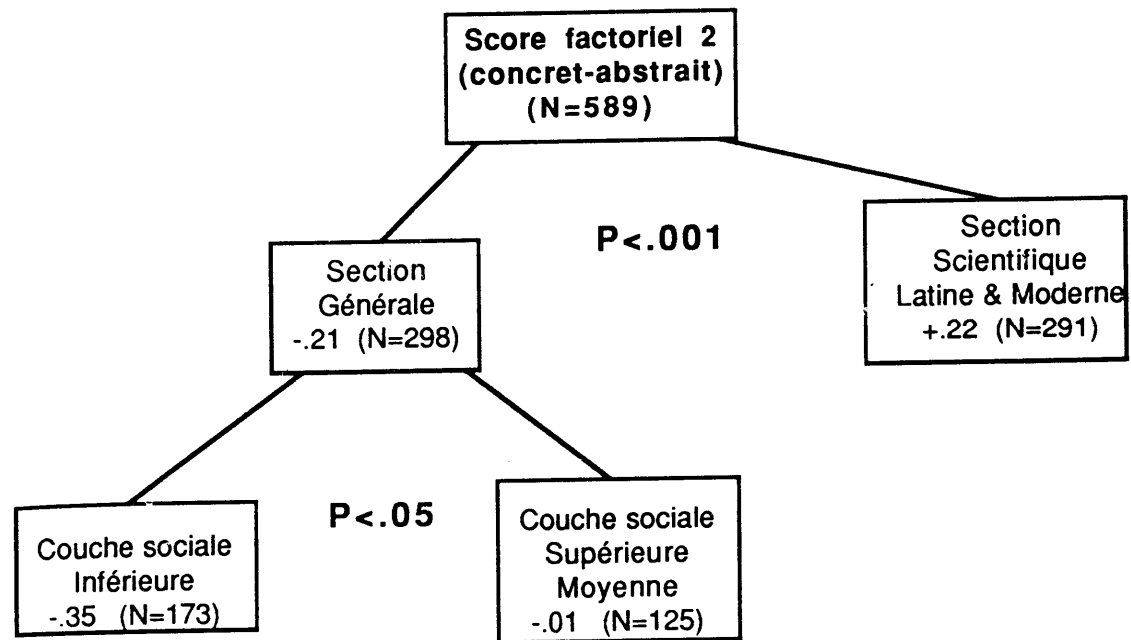
Analyse de segmentation

- Echantillon de n individus statistiques
- une variable dépendante numérique ou qualitative Y
- plusieurs variables numériques ou catégorielles X_1, X_2, \dots, X_p .

Expliquer la variable Y à l'aide d'une ou plusieurs variables quantitatives ou qualitatives.

Créer des groupes d'individus ou d'observations homogènes.

Résultat est fourni sous la forme d'un arbre de décision binaire du type suivant :



Rappel : théorème de Huygens

L'inertie totale est la somme des inerties intra-groupes et de l'inertie des points moyens des groupes, pondérés par l'effectif des groupes.

$$I = \sum_{j=1}^g I_j + \sum_{j=1}^g n_j (\bar{y}_j - \bar{y})^2$$

Inertie totale = \sum Inertie dans les groupes + Inertie des points moyens pondérés par les effectifs des groupes

Exemple : 4 observations suivantes, réparties en deux groupes A et B :

Groupe	A	B	A	B
Y	1	2	3	4

$$\bar{y} = 2,5$$

$$\text{Inertie totale} = (1 - 2,5)^2 + (2 - 2,5)^2 + (3 - 2,5)^2 + (4 - 2,5)^2 = 5$$

$$I_A = (1 - 2)^2 + (3 - 2)^2 = 2$$

$$I_B = (2 - 3)^2 + (4 - 3)^2 = 2$$

$$I_{Inter} = 2 \times (2 - 2,5)^2 + 2 \times (3 - 2,5)^2 = 1$$

Algorithme de segmentation

1) Au départ : un seul segment contenant l'ensemble des individus.

2) Examen de toutes les variables explicatives et de toutes les divisions possibles (de la forme $X_j < A$ et $X_j > A$ si X_j est numérique, regroupement des modalités en deux sous-ensembles si X_j est catégorielle).

Pour chaque division, l'inertie inter-groupes est calculée.

3) La division choisie est celle qui maximise l'inertie inter-groupes.

4) On recommence la procédure dans chacun des deux groupes ainsi définis.

Critères d'arrêt :

On peut utiliser comme critères d'arrêt de l'algorithme de segmentation :

- La taille des groupes (classes) à découper
- Le rapport entre l'inertie intra et la variance totale
- Des tests statistiques (tests de Student de comparaison de moyennes, tests du Khi deux)

Determinants of Wages from the 1985 Current Population Survey

Variable names in order from left to right:

EDUCATION: Number of years of education.

SOUTH: Indicator variable for Southern Region (1=Person lives in South, 0=Person lives elsewhere).

SEX: Indicator variable for sex (1=Female, 0=Male).

EXPERIENCE: Number of years of work experience.

UNION: Indicator variable for union membership (1=Union member, 0=Not union member).

WAGE: Wage (dollars per hour).

AGE: Age (years).

RACE: Race (1=Other, 2=Hispanic, 3=White).

OCCUPATION: Occupational category (1=Management, 2=Sales, 3=Clerical, 4=Service, 5=Professional, 6=Other).

SECTOR: Sector (0=Other, 1=Manufacturing, 2=Construction).

MARR: Marital Status (0=Unmarried, 1=Married)

Diagramme de l'arbre 1 pour Salaire
 Nb de noeuds non-terminaux : 7, Noeuds terminaux : 8

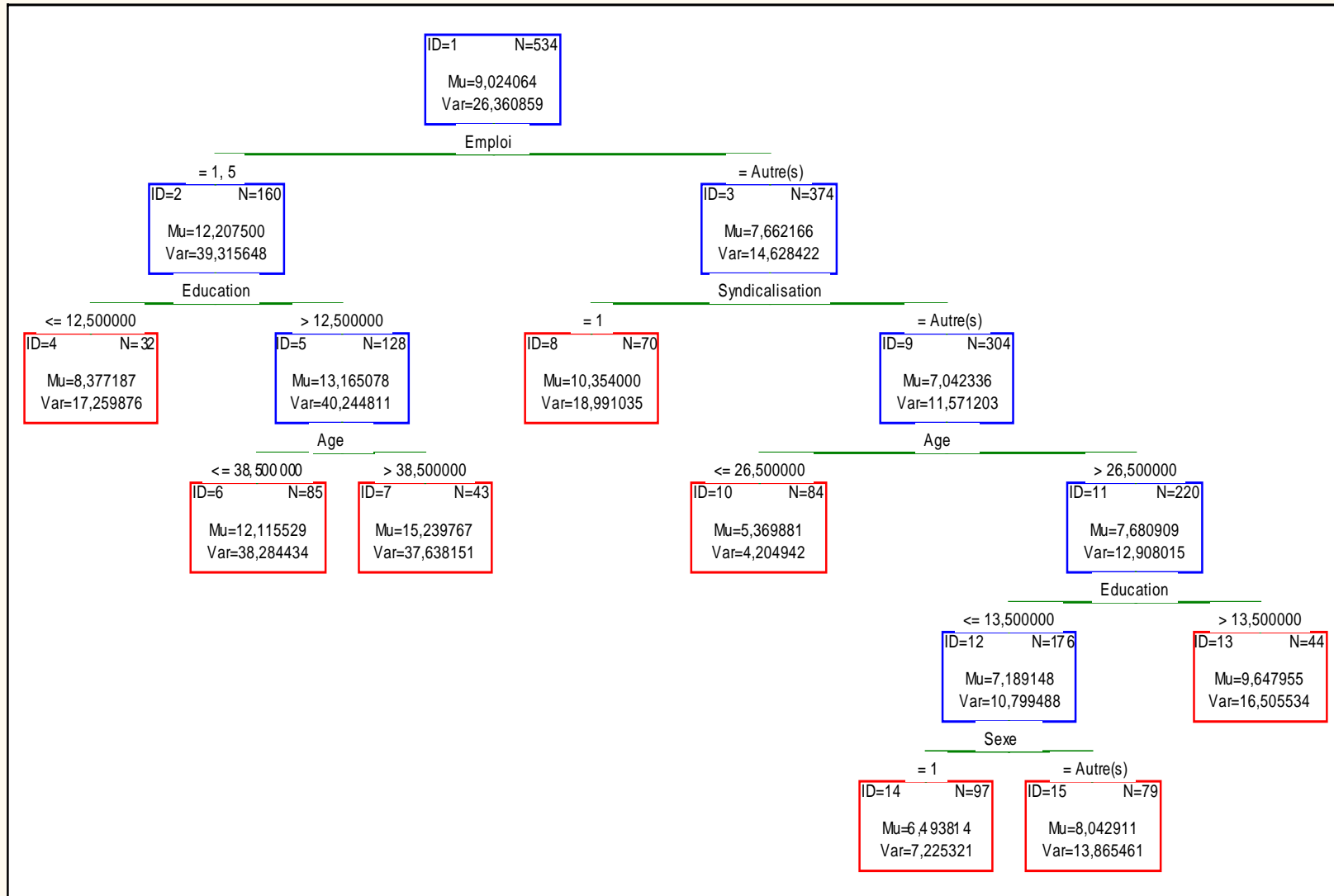


Diagramme de l'arbre 1 pour Log-salaire
 Nb de noeuds non-terminaux : 8, Noeuds terminaux : 9

