

ANALYSE DES DONNÉES

- **TEST DU KHI-DEUX**
- **ANALYSE FACTORIELLE DES
CORRESPONDANCES SIMPLES**

Pierre-Louis Gonzalez

Mesure de la liaison entre deux variables qualitatives

Khi deux

Enquête :

- Êtes-vous « pas du tout d'accord » (1)
 « pas tellement d'accord » (2)
 « peut-être d'accord » (3)
 « bien d'accord » (4)
 « entièrement d'accord » (5)

avec cette phrase ? :

« On en a assez de ceux qui bloquent la vie du pays par leurs revendications ».

Tendance politique	1	2	3	4	5	TOTAL
Extrême gauche	10	1	0	2	1	14
Gauche	134	102	94	82	60	472
Centre	22	27	58	85	62	254
Droite	5	27	49	85	148	314
Extrême droite	1	1	0	3	9	14
Indifférent	15	25	51	63	55	209
Non-réponse	17	24	52	55	45	193

Khi-deux = 319,48

d.d.l. = 24

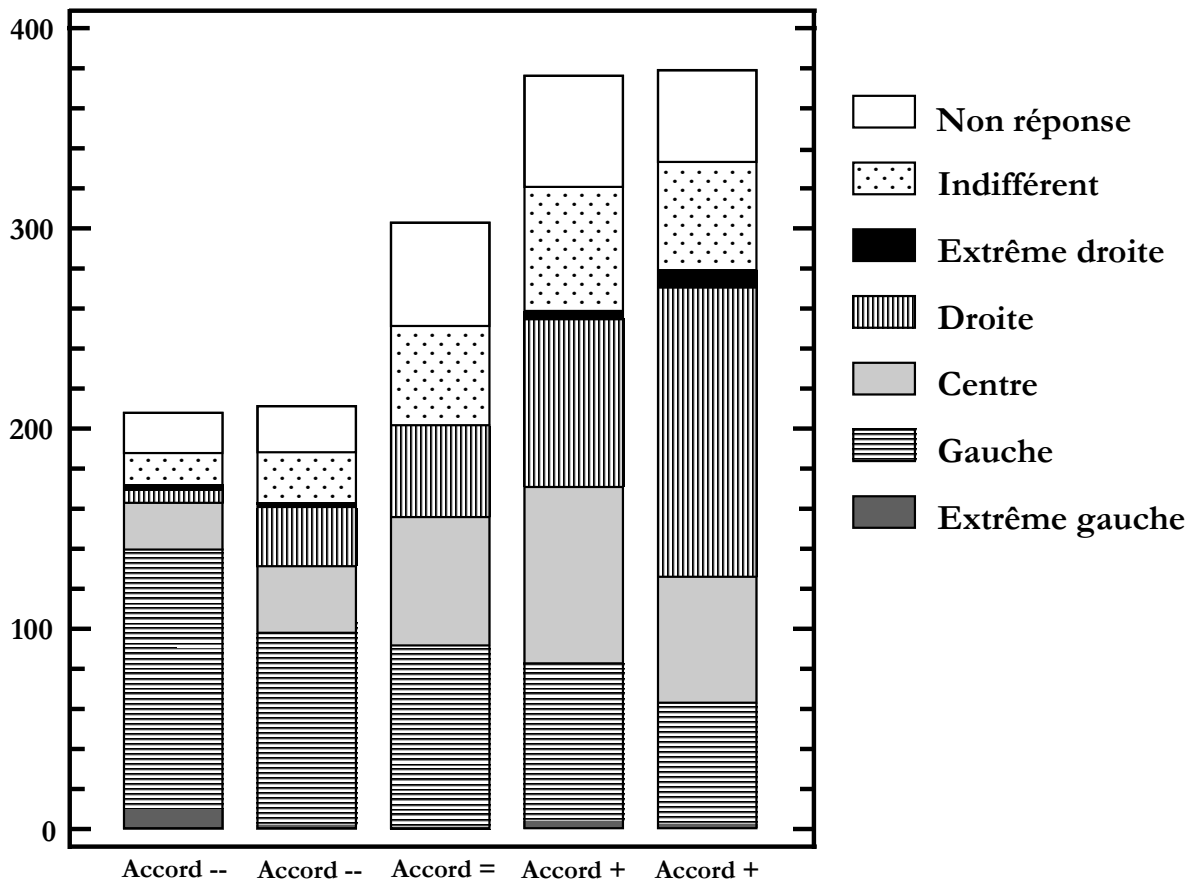
Proba = $2 \cdot 10^{-53}$

Existe-t-il un lien entre les réponses et la tendance politique ?

Tableau des profils lignes

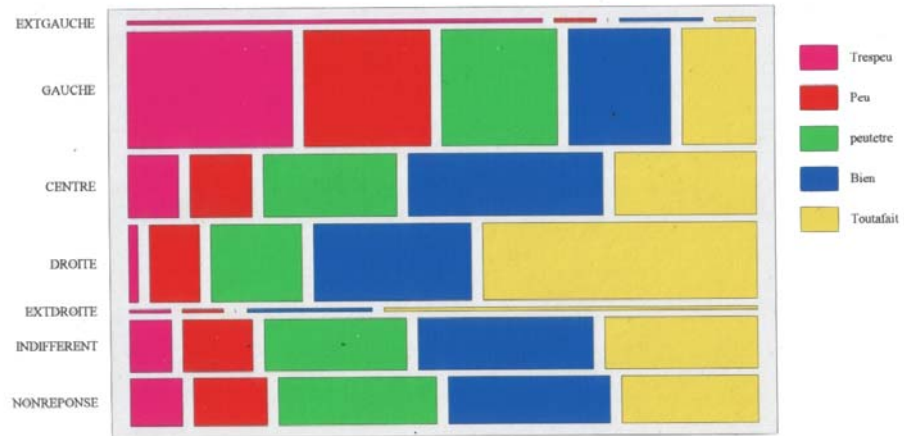
Tendance politique	1	2	3	4	5	TOTAL
Extrême gauche	714	71	0	143	71	1 000
Gauche	284	216	199	174	127	1 000
Centre	87	106	228	335	244	1 000
Droite	16	86	156	271	471	1 000
Extrême droite	71	71	0	214	643	1 000
Indifférent	82	120	244	301	263	1 000
Non-réponse	88	124	269	285	233	1 000

DIAGRAMME EN BATONS



REPRESENTATION GRAPHIQUE DES COLONNES

Mosaic Plot



TEST DU KHI-DEUX

Il s'agit de tester l'indépendance de deux variables qualitatives.

Y a-t-il indépendance entre :

la catégorie socioprofessionnelle et le vote à l'élection présidentielle ?

le niveau d'études et les journaux lus ?

1. Tableau de contingence

Croisement de deux variables qualitatives I et J à p et q modalités.

	1	2		j		q	
1							
2							
⋮							
i				n_{ij}			$n_{i\bullet}$
⋮							
p							
				$n_{\bullet j}$			n

$$n_{i\bullet} = \sum_{j=1}^q n_{ij} \quad (\text{total ligne})$$

$$n_{\bullet j} = \sum_{i=1}^p n_{ij} \quad (\text{total colonne})$$

$$n = \sum_{i=1}^p \sum_{j=1}^q n_{ij} \quad (\text{total})$$

2. Profils lignes - profils-colonnes - profils marginaux

► Profil ligne i :

On divise chaque terme de la ligne i par l'effectif de la ligne

$$\ell_i \quad \boxed{\frac{n_{i1}}{n_{i\bullet}} \quad \dots \quad \frac{n_{ij}}{n_{i\bullet}} \quad \dots \quad \frac{n_{iq}}{n_{i\bullet}}}$$

Profil marginal correspondant :

$$f_j \quad \boxed{\frac{n_{\bullet 1}}{n} \quad \dots \quad \frac{n_{\bullet j}}{n} \quad \dots \quad \frac{n_{\bullet q}}{n}}$$

Si les deux variables qualitatives I et J étaient indépendantes, les profils-lignes seraient tous identiques, et donc identiques au profil marginal correspondant.

$$\frac{n_{ij}}{n_{i\bullet}} = \frac{n_{\bullet j}}{n}$$



$$\boxed{n_{ij} = \frac{n_{i\bullet} \cdot n_{\bullet j}}{n}}$$

Remarques

① On pouvait établir la relation précédente en raisonnant sur les profils-colonnes.

② Souvent cette relation est exprimée

$$f_{ij} = f_{i\bullet} f_{\bullet j}$$

$$\text{avec } f_{ij} = \frac{n_{ij}}{n} \text{ et } f_{i\bullet} = \frac{n_{i\bullet}}{n}$$

$$f_{\bullet j} = \frac{n_{\bullet j}}{n}$$

Elle exprime clairement que dans le cas de l'indépendance le tableau de contingence est entièrement déterminé par ses marges.

3. Définition du Khi-deux

Pour chaque case, on peut donc calculer le nombre de cas attendus (sous hypothèse d'indépendance) $nf_{i\bullet}f_{\bullet j} = \frac{n_{i\bullet}n_{\bullet j}}{n}$

On peut comparer les nombres de cas attendus E_{ij} aux nombres observés.

$$\chi^2 = \sum_{i=1}^p \sum_{j=1}^q \frac{\left(n_{ij} - \frac{n_{i\bullet} n_{\bullet j}}{n} \right)^2}{\frac{n_{i\bullet} n_{\bullet j}}{n}}$$
$$\chi^2 = \sum_{i=1}^p \sum_{j=1}^q \frac{(n_{ij} - E_{ij})^2}{E_{ij}}$$

4. Test

Si les deux variables sont réellement indépendantes, cette expression suit une distribution du Khi-deux avec un nombre de degrés de liberté égal à $(p-1)(q-1)$

Dans une table on lit $\chi_{\alpha,k}^2$ valeur ayant une probabilité α d'être dépassée pour une distribution du khi-deux avec $k = (p-1)(q-1)$ degrés de liberté.

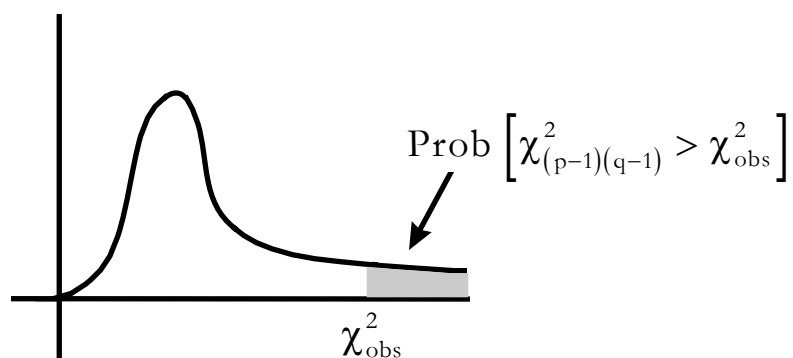
Si $\chi^2 \leq \chi_{\alpha,k}^2$ on accepte H_0 : indépendance

Si $\chi^2 > \chi_{\alpha,k}^2$ on rejette H_0 .

Remarque : test du khi-deux

Les logiciels statistiques proposent la présentation suivante :

- ▶ calcul du χ^2 associé au tableau de contingence noté χ_{obs}^2 .
- ▶ probabilité pour une v.a. suivant une loi du khi-deux à $(p-1)(q-1)$ d.d.l. de dépasser χ_{obs}^2 .



Si cette probabilité est faible (en général $< 5\%$), on rejette l'hypothèse d'indépendance entre les deux variables qualitatives.

5. Autres coefficients

Du χ^2 associé à un tableau de contingence, on déduit d'autres mesures d'association.

► **Coefficient de contingence** $\left(\frac{x^2}{n + x^2} \right)^{1/2}$

► ϕ^2 de Pearson $\phi^2 = \frac{x^2}{n}$

► **Coefficient de Tschuprow** $T = \frac{\phi^2}{\sqrt{p-1} \sqrt{q-1}}$

où p et q = nombre de modalités des 2 variables

Ce coefficient compris entre 0 et 1 est en fait l'analogue d'un coefficient de corrélation.

$T = 1 \Leftrightarrow$ liaison fonctionnelle réciproque

► **Coefficient de Cramer** $C = \left[\frac{\phi^2}{\text{Inf}(p-1, q-1)} \right]^{1/2}$

Thème n° 290 : Des gardiens armés dans supermarchés															
	1 PAS DU TOUT D'ACCORD	163	16.1			36.7	*****								
	2 PAS TELLEMENT D'ACCORD	208	20.6			21.8	*****								
	3 PEUT-ETRE D'ACCORD	220	21.8			41.5	*****								
	4 BIEN D'ACCORD	239	23.7				*****								
	5 ENTIEREMENT D'ACCORD	180	17.8				*****								
	1	2	3	4	5	TOTAL		1	2	3	4	5	TOTAL		MOY.
Région	chi-2 = 72.48					ddl = 28	Proba = 8.3E-006								
REGION PARISIENNE	28	45	43	57	17	190		147	237	226	300	89	1000		2.95
NORD	15	23	13	17	18	86		174	267	151	198	209	1000		3.00
EST	14	25	19	30	19	107		131	234	178	280	178	1000		3.14
BASSIN PARISIEN	25	34	49	45	40	193		130	176	254	233	207	1000		3.21
OUEST	31	30	31	31	14	137		226	219	226	226	102	1000		2.7
CENTRE-EST	32	26	23	19	22	122		262	213	189	156	180	1000		2.78
SUD-OUEST	6	7	25	22	19	79		76	89	316	278	241	1000		3.52
MEDITERRANEE	12	18	17	18	31	96		125	188	177	188	323	1000		3.40
Habitat (4 modalités)	chi-2 = 29.33					ddl = 12	Proba = 0.00352								
- DE 2000 H.	43	59	73	82	48	305		141	193	239	269	157	1000		3.11
-100000 H.	43	57	67	63	65	295		146	193	227	214	220	1000		3.17
+100000 H.	53	51	42	47	52	245		216	208	171	192	212	1000		2.98
PARIS	24	41	38	47	15	165		145	248	230	285	91	1000		2.93
Sexe	chi-2 = 13.38					ddl = 4	Proba = 0.00958								
HOMME	97	98	95	102	83	475		204	206	200	215	175	1000		2.95
FEMME	66	110	125	137	97	535		123	206	234	256	181	1000		3.17
Age (5 Modalités)	chi-2 = 26.17					ddl = 16	Proba = 0.05166								
18 A 24 ANS	18	39	31	47	26	161		112	242	193	292	161	1000		3.15
25 A 34 ANS	46	50	44	48	47	235		196	213	187	204	200	1000		3.00
35 A 49 ANS	45	56	62	51	37	251		179	223	247	203	147	1000		2.92
50 A 64 ANS	37	39	41	45	39	201		184	194	204	224	194	1000		3.05
65 ANS ET PLUS	17	24	42	48	31	162		105	148	259	296	191	1000		3.32
Religion	chi-2 = 51.39					ddl = 16	Proba = 0.00001								
SANS RELIGION	51	45	33	30	24	183		279	246	180	164	131	1000		2.62
CATHO.NON PRATIQUEANT	58	78	79	89	85	389		149	201	203	229	219	1000		3.17
PRAT.OCCASIONNEL	32	50	67	88	48	285		112	175	235	309	168	1000		3.25
PRAT.REGULIER	8	19	24	22	16	89		90	213	270	247	180	1000		3.21
AUTRES CAS	14	16	17	10	7	64		219	250	266	156	109	1000		2.69
Diplôme	chi-2 = 74.70					ddl = 20	Proba = 3.1E-008								
AUCUN DIPLOME	10	21	45	43	40	159		63	132	283	270	252	1000		3.52
C.E.P..C.A.P.	43	58	62	75	57	295		146	197	210	254	193	1000		3.15
BREVET	35	46	52	57	53	243		144	189	214	235	218	1000		3.19
BAC	23	34	28	33	14	132		174	258	212	250	106	1000		2.86
BTS.DUT	24	12	11	12	7	66		364	182	167	182	106	1000		2.48
SUPERIEUR	28	37	22	19	9	115		243	322	191	165	78	1000		2.51
CSP Interviewé (11 Mod.)	chi-2 = 67.02					ddl = 40	Proba = 0.00471								
AGRICULTEURS	4	7	13	14	10	48		83	146	271	292	208	1000		3.40
COMMERÇANTS ET PATRONS	11	12	9	9	10	51		216	235	176	176	196	1000		2.90
CAD.SUPERIEURS.PROF.LIBER	15	13	16	14	4	62		242	210	258	226	65	1000		2.66
ENSEIGNANTS ET ETUDIANTS	21	40	20	29	15	125		168	320	160	232	120	1000		2.82
PROFESSIONS INTERMEDIAIRE	17	20	26	13	8	84		202	238	310	155	95	1000		2.70
EMPLOYES	28	31	30	40	33	162		173	191	185	247	204	1000		3.12
OUVRIERS ET CONTREMAITRES	20	23	19	22	24	108		185	213	176	204	222	1000		3.06
OUVRIERS SPECIALISES	10	4	5	15	10	44		227	91	114	341	227	1000		3.25
RETRAITES OUVRIERS	3	4	4	12	7	30		100	133	133	400	233	1000		3.53
AUTRES RETRAITES	22	30	43	46	33	174		126	172	247	264	190	1000		3.22
SANS PROFESSION	12	24	35	25	26	122		98	197	287	205	213	1000		3.24
Politique	chi-2 = 101.54					ddl = 24	Proba = 1.6E-011								
EXTREME-GAUCHE	4	2	1	0	2	9		444	222	111	0	222	1000		2.33
GAUCHE	70	65	48	49	31	263		266	247	183	186	118	1000		2.64
CENTRE	29	47	33	57	26	192		151	245	172	297	135	1000		3.02
DROITE	15	31	57	45	43	191		79	162	298	236	225	1000		3.37
EXTREME-DROITE	1	0	0	3	9	13		77	0	0	231	692	1000		4.46
INDIFFERENT	32	47	53	48	39	219		146	215	242	219	178	1000		3.07
NON REPONSE	12	16	28	37	30	123		98	130	228	301	244	1000		3.46
Proximité politique	chi-2 = 77.44					ddl = 32	Proba = 0.00001								
EXTREME GAUCHE	1	4	1	3	1	10		100	400	100	300	100	1000		2.90
P.C.F.	7	8	6	11	6	38		184	211	158	289	158	1000		3.03
P.SOCIALISTE, MRO	51	57	44	53	29	234		218	244	188	226	124	1000		2.79
ECOLOGISTE	31	28	20	20	12	111		279	252	180	180	108	1000		2.59
U.D.F	8	17	24	23	10	82		98	207	293	280	122	1000		3.12
R.P.R	12	20	32	36	41	141		85	142	227	255	291	1000		3.52
FRONT NATIONAL	2	1	3	6	11	23		87	43	130	261	478	1000		4.00
AUTRE	11	14	13	13	9	60		183	233	217	217	150	1000		2.92
NON REPONSE	40	59	77	74	61	311		129	190	248	238	196	1000		3.18
TOTAL	163	208	220	239	180	1010		PROFILS LIGNE							

Thème n° 289 : Changer la société												
1 PAS DU TOUT D'ACCORD	47	4.7				28.1	*****					
2 PAS TELLEMENT D'ACCORD	237	23.5				36.9	*****					
3 PEUT-ETRE D'ACCORD	373	36.9				35.0	*****					
4 BIEN D'ACCORD	205	20.3					*****					
5 ENTIEREMENT D'ACCORD	148	14.7					*****					
	1	2	3	4	5	TOTAL	1	2	3	4	5	TOTAL MOY.
Région	chi-2 = 43.59 ddl = 28 Proba = 0.03048											
REGION PARISIENNE	11	47	75	41	16	190	58	247	395	216	84	1000 3.02
NORD	3	23	27	19	14	86	35	267	314	221	163	1000 3.21
EST	4	26	46	24	7	107	37	243	430	224	65	1000 3.04
BASSIN PARISIEN	11	48	65	33	38	193	57	238	337	171	197	1000 3.21
OUVEST	7	24	51	36	19	137	51	175	372	263	139	1000 3.26
CENTRE-EST	9	31	40	23	19	122	74	254	328	189	156	1000 3.10
SUD-OUEST	1	20	23	17	18	79	13	253	291	215	228	1000 3.39
MEDITERRANEE	1	20	46	12	17	96	10	208	479	125	177	1000 3.25
Habitat (4 modalités)	chi-2 = 30.56 ddl = 12 Proba = 0.00230											
- DE 2000 M.	10	78	119	62	36	305	33	256	390	203	118	1000 3.12
-100000 M.	13	58	93	67	64	295	44	197	315	227	217	1000 3.38
+100000 M.	16	60	94	39	36	245	65	245	384	159	147	1000 3.08
PARIS	8	41	67	37	12	165	48	248	406	224	73	1000 3.02
Sexe	chi-2 = 6.59 ddl = 4 Proba = 0.15941											
HOMME	27	119	175	83	71	475	57	251	368	175	149	1000 3.11
FEMME	20	118	198	122	77	535	37	221	370	228	144	1000 3.22
Age (5 Modalités)	chi-2 = 17.68 ddl = 16 Proba = 0.34318											
18 A 24 ANS	10	40	62	29	20	161	62	248	385	180	124	1000 3.06
25 A 34 ANS	14	48	90	49	34	235	60	204	383	209	145	1000 3.17
35 A 49 ANS	6	76	81	49	39	251	24	303	323	195	155	1000 3.16
50 A 64 ANS	9	41	73	48	30	201	45	204	363	239	149	1000 3.24
65 ANS ET PLUS	8	32	67	30	25	162	49	198	414	185	154	1000 3.20
Religion	chi-2 = 14.62 ddl = 16 Proba = 0.55259											
SANS RELIGION	10	40	66	37	30	183	55	219	361	202	164	1000 3.20
CATHO.NON PRATIQUEANT	16	92	144	75	62	389	41	237	370	193	159	1000 3.19
PRAT.OCCASIONNEL	13	69	109	64	30	285	46	242	382	225	105	1000 3.10
PRAT.REGULIER	4	23	29	22	11	89	45	258	326	247	124	1000 3.15
AUTRES CAS	4	13	25	7	15	64	62	203	391	109	234	1000 3.25
Diplôme	chi-2 = 61.75 ddl = 20 Proba = 3.8E-006											
AUCUN DIPLOME	4	27	58	44	26	159	25	170	365	277	164	1000 3.38
C.E.P.,C.A.P.	6	60	114	56	59	295	20	203	386	190	200	1000 3.35
BREVET	14	52	90	57	30	243	58	214	370	235	123	1000 3.15
BAC	7	31	58	19	17	132	53	235	439	144	129	1000 3.06
BTS,DUT	6	20	20	11	9	66	91	303	303	167	136	1000 2.95
SUPERIEUR	10	47	33	18	7	115	87	409	287	157	61	1000 2.70
CSP Interviewé (11 Mod.)	chi-2 = 78.52 ddl = 40 Proba = 0.00026											
AGRICULTEURS	2	10	22	9	5	48	42	208	458	188	104	1000 3.10
COMMERCANTS ET PATRONS	1	11	26	6	7	51	20	216	510	118	137	1000 3.14
CAD.SUPERIEURS,PROF.LIBER	7	20	20	8	7	62	113	323	323	129	113	1000 2.81
ENSEIGNANTS ET ETUDIANTS	9	42	47	16	11	125	72	336	376	128	88	1000 2.82
PROFESSIONS INTERMEDIAIRE	5	27	23	19	10	84	60	321	274	226	119	1000 3.02
EMPLOYES	3	44	53	34	28	162	19	272	327	210	173	1000 3.25
OUVRIERS ET CONTREMAITRES	5	17	45	18	23	108	46	157	417	167	213	1000 3.34
OUVRIERS SPECIALISES	1	4	16	17	6	44	23	91	364	386	136	1000 3.52
RETRAITES OUVRIERS	0	3	13	8	6	30	0	100	433	267	200	1000 3.57
AUTRES RETRAITES	9	40	68	32	25	174	52	230	391	184	144	1000 3.14
SANS PROFESSION	5	19	40	38	20	122	41	156	328	311	164	1000 3.40
Politique	chi-2 = 60.13 ddl = 24 Proba = 0.00006											
EXTREME-GAUCHE	2	0	2	1	4	9	222	0	222	111	444	1000 3.56
GAUCHE	7	76	92	55	33	263	27	289	350	209	125	1000 3.12
CENTRE	9	62	64	33	24	192	47	323	333	172	125	1000 3.01
DROITE	15	43	68	41	24	191	79	225	356	215	126	1000 3.08
EXTREME-DROITE	2	3	4	0	4	13	154	231	308	0	308	1000 3.08
INDIFFERENT	9	30	89	52	39	219	41	137	406	237	178	1000 3.37
NON REPONSE	3	23	54	23	20	123	24	187	439	187	163	1000 3.28
Proximité politique	chi-2 = 47.95 ddl = 32 Proba = 0.03480											
EXTREME GAUCHE	1	0	5	2	2	10	100	0	500	200	200	1000 3.40
P.C.F.	1	7	9	10	11	38	26	184	237	263	289	1000 3.61
P.SOCIALISTE, MRO	10	70	84	42	28	234	43	299	359	179	120	1000 3.03
ECOLOGISTE	5	25	40	29	12	111	45	225	360	261	108	1000 3.16
U.D.F	2	30	32	9	9	82	24	366	390	110	110	1000 2.91
R.P.R	9	34	43	34	21	141	64	241	305	241	149	1000 3.17
FRONT NATIONAL	2	3	8	6	4	23	87	130	348	261	174	1000 3.30
AUTRE	4	15	22	10	9	60	67	250	367	167	150	1000 3.08
NON REPONSE	13	53	130	63	52	311	42	170	418	203	167	1000 3.28
TOTAL	47	237	373	205	148	1010	PROFILS LIGNE					

THEMES	REGIO	HABIT	SEXE	AGE	RELI	DIPLO	CSP	POLIT	PRPOL	TOTAL
1 LIBERALISER L'AVORTEMENT	I	II		III	III	III	III	II	II	8
2 PUBLICITE INDISPENSABLE									I	1
4 CONSTRUIRE DES CENTRALES NUCLEAIRES	I	I	III					III		4
6 DIEU EXISTE				III	III		II	III	II	6
7 EGALISER LES REVENUS	I	II	I			I		II		6
8 EFFORTS POUR LES JEUX OLYMPIQUES					III	III	II			4
9 CONFIANCE EN LA JUSTICE	I	II						I	I	4
13 ETUDIANTS PARASITES						III	I			3
15 TROP DE TRAVAILLEURS IMMIGRES	I	III		II	III	III	III	III	III	8
16 LIMITER LES HERITAGES		II								1
18 CONTRE LA PORNOGRAPHIE	II	I	III	III	III	II	III			7
19 PRIS POUR ABRUTIS A LA TELEVISION					I					1
21 AIDER LES PAYS SOUS-DEVELOPPES				II	I	II	I	III	III	6
26 POUR LA PATRIE		III		III	III	III	III	I	II	7
28 ENCOURAGER LA NATALITE				I						1
29 PROGRES SCIENTIFIQUE NEFASTE	II							II		3
30 POUR LA FAMILLE	III			III		III	I			7
31 CONTRE LE TRAVAIL	I	II			I			II		4
33 DEFENSE DU CONSOMMATEUR	I	II								3
34 CONTRE LE MARIAGE					I				I	3
35 MAINTENIR LA CROISSANCE		I			II			II	II	5
36 REDUIRE LES DEPENSES MILITAIRES	I				II			III	III	4
41 CENSURER CERTAINS LIVRES	III			III	I	III	III	III	III	6
42 GOUVERNEMENT INEFFICACE		II				II	II	III	III	6
43 RESPECTER LES CONVENANCES	III	III		III	III	III	III	III	III	7
46 SENTIMENT D'INSECURITE	III	III	II	II	III	III	III	III	III	9
52 LE POUVOIR DE PARIS AUX REGIONS	III	III		I						3
57 CRISE DE L'ENERGIE PREOCCUPANTE	III	I	II			I				4
62 A LA MERCI DES PAYS PETROLIERS	I	III	II				I			6
65 AMENDES AUX GASPIS		I		III						4
66 SOUTENIR LES ECOLOGISTES			I					III	III	3
73 PLUS DE LIBERTES AUX MUNICIPALITES	II	III								2
77 POUR LES 35 HEURES	II	I		I	II		III	III	III	7
80 SUPPRIMONS LE SERVICE MILITAIRE	III	II		III	III	III	III	II	III	8
83 MONTEE DE LA VIOLENCE	III	III	III		II	III	II	III	III	7
84 SYNDICATS INDISPENSABLES		I					I			4
85 PRENDRE EN CHARGE LES CHOMEURS		II						II	I	3
90 ON N'APPREND PLUS RIEN A L'ECOLE				II		III	III		II	5
92 AGRICULTEURS TROP SUBVENTIONNES	I	III	I			II	III		I	6
93 ORDINATEURS MENACENT NOS LIBERTES	I	I		III		III	III			5
96 GUERRE MONDIALE INEVITABLE			I			III				2
99 RETRAITE PLUS JEUNE	I	II		II		II	III	III	III	7
101 TROP DE GREVES				II	I		II		III	5
103 HOMOSEXUELS COMME LES AUTRES	III	III	II	III	III	III	III	II	II	9
104 TROP DE FONCTIONNAIRES	I	II					II		II	5
106 HOMMES POLITIQUES INTEGRES								I		2
109 MARCHÉ COMMUN FAVORABLE A LA FRANCE	I	II				II				3
112 INDUSTRIE CHIMIQUE POLLUANTE	II	III								2
117 RETABLIR LA PEINE DE MORT	III	III		II	III	III	III	III	III	8
122 HASCHISCH EN VENTE LIBRE	III	III		III	III	III	III	III	I	7
128 CONTROLES D'IDENTITE INDISPENSABLES	III	III		III	III	III	III	III		8
131 TROP DE FAUX CHOMEURS	III	III			III	III	III	II	II	7
134 LIMITER LES IMPORTATIONS JAPONAISES			I	III		II	III			4
149 SOUTENIR LES PACIFISTES	I	II						II	III	4
157 NE PAS COUVRIR LES PETITES MALADIES	II	II					I			3
163 TROP DE CHARGES SUR LES ENTREPRISES	III	III			II	II	III	III	II	7
168 IL FALLAIT DES CENTRALES NUCLEAIRES		I	III		I		I	III	II	6
177 VOTE MUNICIPAL AUX ETRANGERS	II	II		I	III	III	II	III	III	8
180 EXPORTER LES ARMES FRANCAISES	I	II	II	I	II		I	III	III	8
182 MOINS DE ROBOTS	I	III	III	I	I	III	III	II		8
190 POUR DES FEMMES NUES A LA TELE	II	I	III	III	III		III			6

THEMES	REGION	HABIT	SEXE	AGE	RELI	DIPLO	CSP	POLIT	PRPOL	TOTAL
195 LIBERTE DES PRIX	I		II				II			3
200 GORBATCHEV INSPIRE CONFIANCE	I		II					II	III	4
202 VERITE CACHEE EN FRANCE SUR TCHERNOBYL	III	II				I	I			4
204 ISOLER MALADES DU SIDA	III	I		III	II	III		I	II	8
205 LA POLICE REMPLIT BIEN SA MISSION				II	II					2
211 ARMEE FRANCAISE MEILLEURE		II		I	I	II			I	5
212 PRISON POUR ALCOOL AU VOLANT		I	III							2
217 TRANSPORT DANGEREUX MAL CONTROLE	I	II	II				I		I	5
219 PLUS DE POUVOIR AU PARLEMENT		I	II							2
221 L'URSS MENACE LA FRANCE	I		III		II	III	III	III	II	7
222 POUVOIR DEVENIR FRANCAIS					III	III	II	II	III	5
228 FORTUNES RESPECTABLES	I	I		I	I	I	I			6
230 ARMES NUCLEAIRES POUR DEFENDRE L'EUROPE			III		II		II	III	II	5
233 GARANTIR A TOUS UN REVENU MINIMUM	I	III					I			3
240 TRANSFORMER L'ENSEIGNEMENT EN FRANCE										0
242 PLUS PERSONNE POUR PAYER NOS RETRAITES	II	III	II			II				4
244 CONSTRUIRE VOITURES MOINS RAPIDES	I	III	III	III	II	I	III			7
248 CRISE DE L'ENERGIE ECONOMIE D'ENERGIE				II						1
250 POUR LA FORCE DE DISSUASION		I	II		II			III	III	5
251 LA POLLUTION DETRUIRA L'HUMANITE	I	II	II			I	I			5
255 VERTS CAPABLES DE DIRIGER PAYS			III				II	III	III	4
267 L'ECOLOGIE EST UNE MODE		I		I		III	III	III	III	6
269 EPARGNER POUR LA RETRAITE	II			I	II			III	I	5
275 FRANCE EN AVANCE POUR L'ENVIRONNEMENT						II	I		III	3
276 UNE FEMME A L'ELYSEE		I	III				II	II	I	4
277 FAIRE PAYER LES COMMUNES RICHES	I	I						III	III	4
278 RENFORCER LE ROLE DE L'ONU	I	II	I		III	I				5
279 LE SPORT EST POURRI PAR L'ARGENT		I								1
280 VIVE LA REUNIFICATION DE L'ALLEMAGNE		II		II						2
281 LA FRANCE DOIT AIDER L'URSS	II	III	I		II	II	III	II	II	8
282 LIMITER LA VITESSE EN VILLE A 50 KM/H	I	III	II	I			I			5
283 UN SALAIRE POUR LA FEMME AU FOYER	I	I				I	III			4
284 LE TAG ET LE RAP C'EST DE L'ART	II	II	I	III	II	II	II	III	III	9
285 CONSTRUIRE L'EUROPE AVEC L'EST	I	I			I	II	I	III	I	7
286 POUR UN GRAND CLUB DE FOOT		III	II	II	I					5
287 LES INEGALITES AUGMENTENT	I	II				I				3
288 DETRUIRE L'ARMEE DE L'IRAK		II			II	II	I			4
289 CHANGER LA SOCIETE		I				III	II	II		4
290 DES GARDIENS ARMES DANS SUPERMARCHES	III	I	I		II	III	II	III	II	8
291 ACCUEILLIR LES REFUGIES POLITIQUES	III	III			III	III	III	III	III	7
292 TROP DE POLITIQUE DANS LA JUSTICE	II	II								2
293 POUR L'INTERVENTION DANS LE GOLFE	I	I						I		3
TOTAL	63	77	35	45	50	56	62	56	52	496

Significativité des liaisons :

I : à 10^{-2} II : à 10^{-3} III : à 10^{-5} IIII : à 10^{-10} IIIIII : à 10^{-20}

LISTE DES ABREVIATIONS :

REGIO : REGION

HABIT : HABITAT

RELI : RELIGION

DIPLO : DIPLOME

POLIT : TENDANCE POLITIQUE

PRPOL : PROXIMITE POLITIQUE

L'analyse des correspondances simples (A.F.C.)

Méthode proposée par J.P. Benzecri dans le but d'étudier la liaison dite encore **correspondance** entre deux variables qualitatives.

Exemple : répartition des habitants de Paris selon leur lieu d'habitation : un des 80 quartiers et leur C.S.P.

Questions :

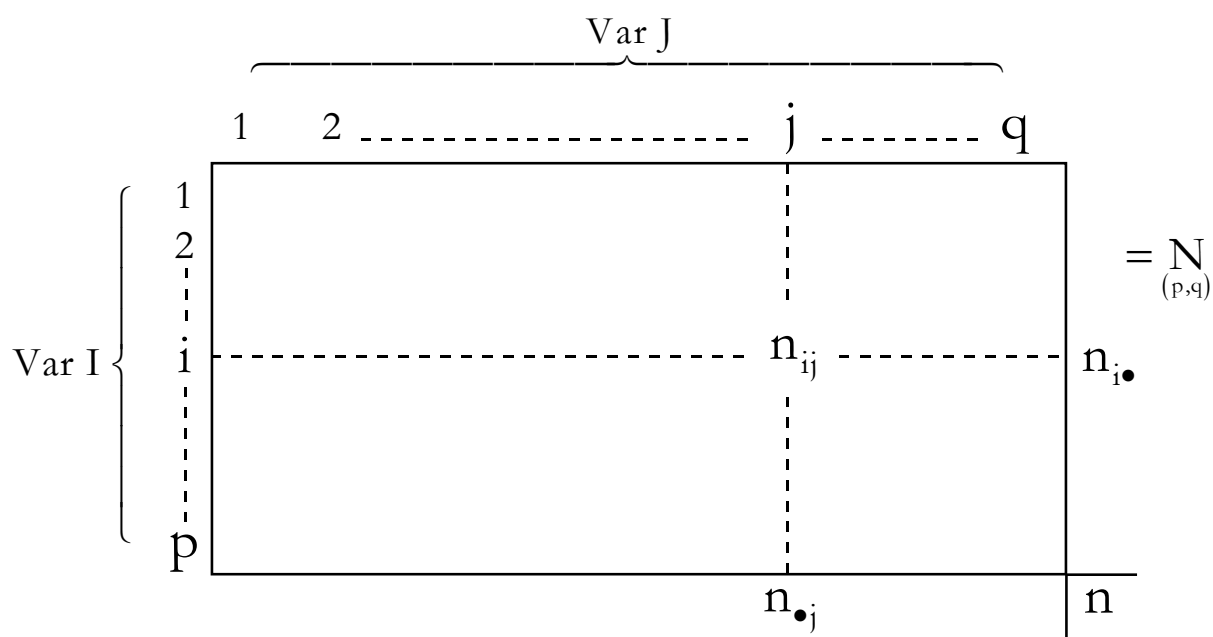
- ◆ Certains quartiers sont-ils proches ?
au sens même répartition des C.S.P. ?
- ◆ Certaines C.S.P. sont-elles proches ?
- ◆ Certaines C.S.P. sont-elles plus souvent associées à certains quartiers ?

L'analyse des correspondances traite des **tableaux de contingence**.

I. NOTATIONS ET PRÉSENTATION

1) Notations : tableau de contingence : N

Croisement de deux variables qualitatives à p et q modalités



► p Profils des lignes $\frac{n_{ij}}{n_{i\bullet}} = \frac{f_{ij}}{f_{i\bullet}}$

profil de la ligne i noté ℓ_i

$$\left(\frac{n_{i1}}{n_{i\bullet}} \quad \frac{n_{i2}}{n_{i\bullet}} \quad \dots \quad \frac{n_{iq}}{n_{i\bullet}} \right) \Leftrightarrow \left(\frac{f_{i1}}{f_{i\bullet}} \quad \frac{f_{i2}}{f_{i\bullet}} \quad \dots \quad \frac{f_{iq}}{f_{i\bullet}} \right)$$

► q Profils des colonnes

profil de la colonne j
noté c_j

$$\left(\frac{n_{1j}}{n_{\bullet j}} \quad \frac{n_{2j}}{n_{\bullet j}} \quad \dots \quad \frac{n_{pj}}{n_{\bullet j}} \right) \Leftrightarrow \left(\frac{f_{1j}}{f_{\bullet j}} \quad \frac{f_{2j}}{f_{\bullet j}} \quad \dots \quad \frac{f_{pj}}{f_{\bullet j}} \right)$$

2) Représentation des profils

a) Les profils-lignes

Les profils lignes sont considérés comme des individus.

Les p profils-lignes forment un nuage de p points dans R^q

A chaque profil-ligne est associé un poids égal à sa fréquence marginale $f_{i\bullet}$.

On note $N(I)$ le nuage de points formé des profils-lignes pondérés :

$$\begin{array}{cc} (\ell_i ; f_{i\bullet}) \\ \text{profil ligne} & \text{poids} \end{array}$$

Le centre de gravité g_ℓ est défini par :

$$g_\ell = \sum_{i=1}^p f_{i\bullet} \ell_i$$

La $j^{\text{ème}}$ coordonnée de g_ℓ vaut $f_{\bullet j}$

$$\text{En effet : } \sum_{i=1}^p f_{i\bullet} \left(\frac{f_{ij}}{f_{i\bullet}} \right) = \sum_{i=1}^p f_{ij} = f_{\bullet j}$$

Donc $g_\ell = \text{profil marginal de la variable J (à } q \text{ modalités)}$

$$g_\ell = f_j$$

b) Les profils-colonnes

$N(J)$ = nuage de points formé des q profils - colonnes pondérés $(c_j, f_{\bullet j})$

Le centre de gravité g_c est le profil marginal
de la variable I à p modalités.

$$g_c = f_I$$

Le problème qui se pose est **l'étude de la dépendance** entre les deux variables qualitatives.

Dans le cas où les deux variables sont **indépendantes**, on a identité des profils :

$$(1) \quad \frac{f_{ij}}{f_{i\bullet}} = f_{\bullet j} \quad \text{profil - ligne}$$

$$(2) \quad \frac{f_{ij}}{f_{\bullet j}} = f_{i\bullet} \quad \text{profil - colonne}$$

$$\boxed{f_{ij} = f_{i\bullet} f_{\bullet j}}$$

Dans le cas de l'indépendance, le nuage des profils-lignes se réduit à un point $\underline{g}_\ell$

De même, le nuage des profils-colonnes se réduit à un point g_c .

$\Rightarrow$ L'étude de la dépendance consiste à étudier la **forme des nuages**.

$\Rightarrow$ **Problème d'analyse en composantes principales.**

Quelle métrique ?

Remarque

Remarquons que les profils ayant pour somme 1, les p profils-lignes sont en réalité situés dans le sous-espace de dimension $q-1$ défini par

$$\sum_{j=1}^q x_j = 1 \quad (\text{avec } x_j \geq 0).$$

De même pour les q profils-colonnes.

Métrique du χ^2

a) Pour les profils lignes :

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

Justifications :

- ▶ Donne un poids important aux différences portant sur les petits pourcentages.
- ▶ Vérifie le principe d'équivalence distributionnelle : si deux colonnes ont le même profil, on les réunit en une seule d'effectif somme sans modifier les distances entre profils-lignes.

b) Pour les profils-colonnes :

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

Exemple de calcul de la distance entre deux profils-lignes

Distance du χ^2 :

$$d^2(\text{EXPA}, \text{LICS})$$

$$= \frac{1}{0.1592} (0.1566 - 0.1551)^2 + \frac{\dots\dots\dots}{\text{SCEC}} + \frac{\dots\dots\dots}{\text{LETT}} + \frac{\dots\dots\dots}{\text{SCIE}} + \\ + \frac{1}{0.2005} (0.1272 - 0.2890)^2 + \frac{\dots\dots\dots}{\text{MEDD}} + \frac{\dots\dots\dots}{\text{PHAR}} + \frac{\dots\dots\dots}{\text{PLUR}} + \frac{1}{0.0571} (0.1135 - 0.0261)^2$$

$$\left\{ \begin{array}{l} \text{contribution dans le calcul de la distance du } \chi^2 \\ \text{pour MEDD} : 0,130 \\ \text{pour IUT} : 0,133 \end{array} \right.$$

$$\left\{ \begin{array}{l} \text{avec la distance euclidienne usuelle :} \\ \text{pour MEDD} : 0,0261 \\ \text{pour IUT} : 0,0076 \end{array} \right.$$

la contribution d'IUT aurait été trop faible :

En effet : 11,35 % des fils d'EXPA vont à l'IUT au lieu de 5,71 %
pour l'ensemble de la population (soit le double)

2,61 % des enfants de LICS vont à l'IUT au lieu de 5,71 %
population entière

Inertie du nuage N(I)

$I_{N(I)}$ l'inertie du nuage $N(I)$ calculée par rapport
au centre de gravité f_J vaut $\frac{\chi^2}{n}$

où $\chi^2 = \text{Khi-deux}$ associé au tableau de contingence étudié.

En effet :

$$\begin{aligned} I_{N(I)} &= \sum_{i=1}^p f_{i\bullet} d_{\chi^2}^2(\ell_i, f_J) \\ &= \sum_{i=1}^p \frac{n_{i\bullet}}{n} \sum_{j=1}^q \frac{n}{n_{\bullet j}} \left(\frac{n_{ij}}{n_{i\bullet}} - \frac{n_{\bullet j}}{n} \right)^2 \\ &= \sum_{i=1}^p \sum_{j=1}^q \frac{n_{i\bullet}}{n_{\bullet j}} \left(\frac{n_{ij} n - n_{\bullet j} n_{i\bullet}}{n_{i\bullet} n} \right)^2 \\ &= \sum_{i=1}^p \sum_{j=1}^q \frac{n_{i\bullet}}{n_{\bullet j} n_{i\bullet}^2} \left(n_{ij} - \frac{n_{i\bullet} n_{\bullet j}}{n} \right)^2 \\ &= \sum_{i=1}^p \sum_{j=1}^q \frac{\left(n_{ij} - \frac{n_{i\bullet} n_{\bullet j}}{n} \right)^2}{n_{i\bullet} n_{\bullet j}} \\ &= \frac{\chi^2}{n} \end{aligned}$$

Remarque : on obtient le même résultat pour l'inertie du
nuage $N(J)$.

II. L'ANALYSE EN COMPOSANTES PRINCIPALES DU NUAGE DES PROFILS-LIGNES, DES PROFILS-COLONNES

1. On réalise l'A.C.P. du nuage des profils-lignes :

- ▶ Les profils-lignes jouent le rôle d'individus ; ils sont affectés des poids $f_{i\bullet}$.
- ▶ La métrique utilisée pour le calcul des distances entre individus est la métrique du khi-deux.

Le premier axe principal du nuage des profils-lignes est la droite passant le plus près possible de l'ensemble des points de $N(I)$.

Notons $\underline{a}^1$ la première composante principale

$$\underline{a}^1 = \begin{pmatrix} \vdots \\ \vdots \end{pmatrix} \leftarrow \begin{array}{l} \text{coordonnées des } p \\ \text{profils-lignes sur l'axe 1} \end{array}$$

Notons λ_1 la variance de $\underline{a}^1$ (égale à l'inertie portée par l'axe qui lui est associé).

$\underline{a}^2$ = deuxième composante principale de variance λ_2

$\underline{a}^3$ = troisième composante principale de variance λ_3

2. On peut également réaliser l'A.C.P. du nuage des profils-colonnes affectés des poids $f_{\bullet j}$ en utilisant la métrique du χ^2 .

Notons $\underline{b}^1$ la première composante principale

$$\underline{b}^1 = \begin{pmatrix} \vdots \end{pmatrix} \leftarrow \begin{array}{l} \text{coordonnées des } q \\ \text{profils-colonnes sur l'axe 1} \end{array}$$

$$\underline{b}^2 = \text{deuxième composante principale}$$

Les composantes principales de l'A.C.P. des profils-colonnes sont associées aux mêmes valeurs propres que les composantes principales de l'A.C.P. des profils-lignes.

$$\underline{b}^1 \text{ a pour variance } \lambda_1$$

$$\underline{b}^2 \text{ a pour variance } \lambda_2$$

.....

3.

Lien entre les deux analyses : formules de transition
(deuxième forme)

En notant b_j et a_i les $j^{\text{ème}}$ et $i^{\text{ème}}$ coordonnées des composantes principales $\underline{b}$ et $\underline{a}$ associées à la même valeur propre λ :

$$\sqrt{\lambda} b_j = \sum_{i=1}^p \frac{n_{ij}}{n_{\bullet j}} a_i$$

$$\sqrt{\lambda} a_i = \sum_{j=1}^q \frac{n_{ij}}{n_{i\bullet}} b_j$$

avec $\underline{a} = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_p \end{pmatrix}$ $\underline{b} = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_q \end{pmatrix}$

À $\sqrt{\lambda}$ près, la coordonnée d'une modalité i d'une variable est la moyenne des coordonnées des catégories de l'autre variable pondérées par les fréquences conditionnelles du profil de i .

III. REPRÉSENTATIONS GRAPHIQUES A.F.C.

Divers types de représentations sont possibles :

1. Les modalités de la variable I sont représentées en tant qu'individus (profils-lignes) de l'A.C.P. des profils-lignes.

La modalité i de la variable I a pour coordonnées dans un espace de dimension k :

$$(a_i^1, a_i^2, \dots, a_i^k)$$

avec $a_i^1 = i^{\text{ème}}$ coordonnée du vecteur $\underline{a}^1$

$a_i^2 = i^{\text{ème}}$ coordonnée du vecteur $\underline{a}^2$

.....

Pour les modalités de la variable J, la modalité j a pour coordonnées :

$$(\sqrt{\lambda_1} b_j^1, \sqrt{\lambda_2} b_j^2, \dots, \sqrt{\lambda_k} b_j^k)$$

$b_j^1 = j^{\text{ème}}$ coordonnée du vecteur $\underline{b}^1$

$b_j^2 = j^{\text{ème}}$ coordonnée du vecteur $\underline{b}^2$

Les modalités du deuxième groupe (J) sont les barycentres des modalités du premier groupe (variable I).

(voir formules de transition)

2. Même principe mais en inversant les rôles des deux groupes de modalités.

3. Abandon du principe barycentrique

Les modalités de chaque ensemble sont représentées par les :

$$\begin{cases} a_i^k & i=1 \dots p \\ b_j^k & j=1 \dots q \end{cases}$$

Cette dernière solution est la plus pratique (bien que les deux groupes de modalités se trouvent en réalité dans des espaces différents).

Cette représentation permet de déterminer les proximités entre certains éléments de I et certains éléments de J (compte tenu de la qualité de la représentation).

4. Aides à l'interprétation : identiques à celles de l'A.C.P.

a) Contributions

de la ligne i à l'axe k

$$\frac{f_{i\bullet} (a_i^k)^2}{\lambda_k} \quad \text{avec } f_{i\bullet} = \frac{n_{i\bullet}}{n}$$

de la colonne j à l'axe k

$$\frac{f_{\bullet j} (a_j^k)^2}{\lambda_k} \quad \text{avec } f_{\bullet j} = \frac{n_{\bullet j}}{n}$$

b) Cosinus carrés

Modalité i représentée sur l'axe k

$$\frac{(a_i^k)^2}{d^2(i, G)}$$

Modalité j représentée sur l'axe k

$$\frac{(b_j^k)^2}{d^2(j, G)}$$

c) Aspects pratiques de l'interprétation

- L'interprétation peut se faire à partir des représentations graphiques (en s'assurant de la qualité de représentation de chaque modalité à l'aide des $\cos^2$).
- Quand le nombre de modalités est élevé, il est conseillé d'éditer d'abord le graphique des profils-lignes, puis celui des profils-colonnes, enfin la représentation simultanée.

Les profils ayant des poids différents la lecture de leurs contributions à l'inertie de chaque axe s'avère très utile.

On peut repérer les profils dont la contribution est supérieure au poids