4 Tableau de contingence, distance et test du khi-2

On a relevé sur un ensemble d'individus statistiques les valeurs prises par deux variables qualitatives, comportant un nombre réduit de modalités. On peut rassembler les résultats dans un tableau de contingence (tri croisé).

Exemple : origine sociale des étudiants de 1ère année et choix d'un secteur disciplinaire à l'université :

	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**

La première ligne et la première colonne donnent les modalités des deux variables X et Y étudiées. On se pose la question suivante : existe-t-il un lien entre l'origine sociale des étudiants (CSP des parents) et le choix de l'un ou l'autre des secteurs disciplinaires ?

Autrement dit : les variables X et Y sont-elles (statistiquement) dépendantes ?

Le tableau précédent constitue le tableau des effectifs observés O_{ij}. On forme alors le tableau des effectifs théoriques T_{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

Les valeurs contenues dans ce tableau sont calculées à partir de la formule :

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

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

On calcule alors un tableau des 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

Chaque contribution est calculée par :

$$\text{Ctr}_{ij} = \frac{(O_{ij} - T_{ij})^2}{T_{ij}} \; ;$$

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

La somme de toutes ces contributions est la distance du χ^2 séparant ces deux tableaux.

$$\chi^2_{Obs} = \sum_{i,j} \text{Ctr}_{ij} = 0,05 + ... + 72,12 = 320,2$$

Pour réaliser un test du χ^2 (ce qui suppose que les données observées constituent un échantillon tiré au hasard dans une population), on pose les hypothèses :

H_0 : Les variables X et Y sont indépendantes
H_1 : Les variables X et Y sont dépendantes

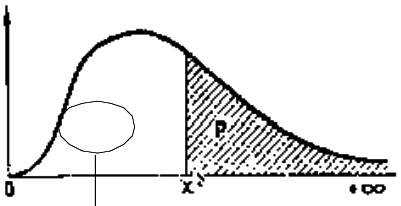
Sous l'hypothèse H_0, la distance entre les deux tableaux suit une loi du χ^2 à 12 degrés de liberté. Ce dernier nombre est défini par la formule :

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

On choisit un seuil (5% par exemple) et on lit dans une table la valeur critique correspondante :

DISTRIBUTION DE χ^2 (Loi de K. Pearson)

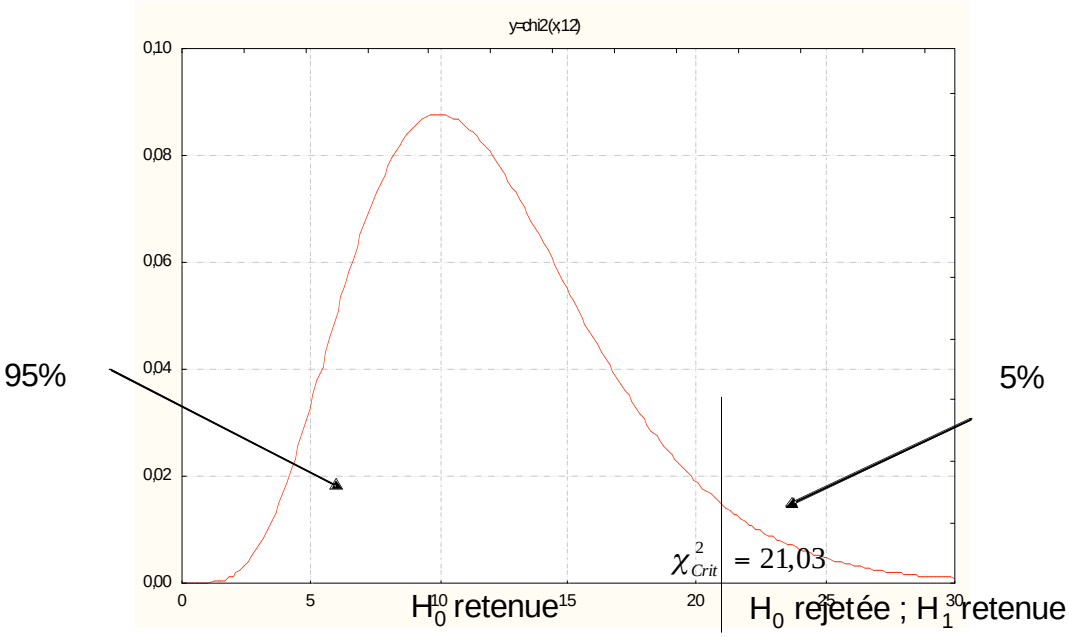
Valeur de χ^2 ayant la probabilité *P* dêtre dépassée.



v	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

On formule ensuite la règle de décision :

Loi du khi -2



Dans notre exemple, le χ^2 observé est très supérieur au χ^2 critique. On retient donc l'hypothèse H_1 : il existe une lien entre les deux variables étudiées.

5 Analyse Factorielle des Correspondances

5.1 Introduction

L'analyse factorielle des correspondances (AFC), ou analyse des correspondances simples, est une méthode exploratoire d'analyse des tableaux de contingence. Elle a été développée essentiellement par J.-P. Benzecri durant la période 1970-1990.

Soient deux variables nominales X et Y, comportant respectivement p et q modalités. On a observé les valeurs de ces variables sur une population et on dispose d'un tableau de contingence à p lignes et q colonnes donnant les effectifs conjoints c'est-à-dire les effectifs observés pour chaque combinaison d'une modalité i de X et d'une modalité j de Y.
Les valeurs de ce tableau seront notées n_{ij}, l'effectif total sera noté N.

L'AFC vise à analyser ce tableau en apportant des réponses à des questions telles que :

- Y a-t-il des lignes du tableau (modalités de X) qui se "ressemblent", c'est-à-dire telles que les distributions des modalités de Y soient analogues ?
- Y a-t-il des lignes du tableau (modalités de X) qui s'opposent, c'est-à-dire telles que les distributions des modalités de Y soient très différentes ?
- Mêmes questions pour les colonnes du tableau.
- Y a-t-il des associations modalité de X - modalité de Y qui s'attirent (effectif conjoint particulièrement élevé) ou qui se repoussent (effectif conjoint particulièrement faible) ?

La méthode se fixe également comme but de construire des représentations graphiques mettant en évidence ces propriétés des données.

5.2 Exemple

5.2.1 Enoncé

Réf. Résultats publiés dans "Le Monde" au lendemain du 22 avril 2007.

Les données qui suivent sont constituées par les résultats du premier tour des élections présidentielles de 2007. Pour chacune des 23 régions françaises (22 régions métropolitaines + 1 "région" Outremer), on donne les effectifs de suffrages pour chacun des 12 candidats (en colonnes). L'objectif est d'analyser la structure des votes ainsi que les liaisons entre candidats et régions.

Données : résultats du premier tour des présidentielles 2007.

	Sarkozy	Bayrou	Royal	Le Pen	Besanc.	Villiers	Voynet	Laguiller	Bove	Buffet	Nihous	Schivardi
Alsace	362391	214259	171282	135730	33310	22492	20382	13821	13758	6100	5142	2522
Aquitaine	532127	417546	557300	168664	78230	34028	28285	22046	27941	41791	35300	7572
Auvergne	238152	169395	225477	78704	41522	18730	12090	12936	13532	21920	12474	4207
Bourgogne	297544	175213	241094	119041	42246	24971	13690	14440	12296	18154	12079	3608
Bretagne	557507	451988	564100	143926	94205	41212	39026	25662	28484	31860	21207	5169
Centre	460425	278175	345352	168912	65347	45720	22655	22279	17395	30003	20567	5696
Cham-Ard	246680	122642	160280	114527	33424	20455	10727	13560	7414	12465	9016	2270
Corse	56819	18979	33493	23432	5941	1908	2119	1346	1659	5163	2260	450
Fr-Comte	212358	114148	165243	94212	30672	16361	12879	10880	10365	9204	7814	2446
Ile de Fr	1931429	1143081	1593033	430553	181247	89498	89885	52965	57453	110967	19890	12386
Lang-Rous	470017	234739	395509	214468	62597	28166	20787	17175	27412	38590	21356	11436
Limousin	123870	82445	142237	38525	24040	10789	6144	6566	6629	15695	8029	2284
Lorraine	403919	250195	315596	196696	70940	29183	21562	25574	16094	19229	10378	4211
Midi-Pyr	458093	341651	542038	154777	69177	30850	25267	18623	33687	34076	25280	8778
Nord-PdC	639390	340679	573071	335855	127881	40702	31388	52695	24591	74027	43595	6348
Basse-Nor	280914	184256	209308	88569	43997	25722	14389	14652	13180	11211	21449	3090
Haute-Nor	309924	184615	257664	126795	58312	25692	15707	20048	12563	26376	13717	3604
Pays-Loire	636934	457560	552280	158844	93685	107895	38952	28481	25811	26737	26674	6481
Picardie	331053	161236	251862	168699	57769	26731	14219	23982	11619	22334	20439	4189
Poit.-Cha	304493	194126	322212	88138	45638	38735	16333	13949	13934	15901	21571	3882
PACA	1010234	419161	579036	377831	88331	54851	38339	26467	36963	61968	26538	9935
Rhone-Alp	1121615	689984	807220	360646	123776	75959	63032	38114	51441	57654	32624	10969
Outremer	337711	103933	398110	36714	22159	5139	12383	10234	14904	14062	2698	1772

Y a-t-il des régions qui se ressemblent, c'est-à-dire dans lesquels les résultats (en pourcentages) des différents candidats sont voisins ? Y a-t-il au contraire des régions qui s'opposent (résultats très différents) ?

Y a-t-il des régions dont les résultats sont proches des résultats nationaux ? Y a-t-il des régions "à part" (dont les résultats s'écartent notablement des résultats nationaux) ?

Y a-t-il des candidats dont les résultats se ressemblent : ils n'obtiennent pas nécessairement les mêmes scores, mais les régions où ils obtiennent de bons scores sont les mêmes ? Y a-t-il des candidats dont les résultats s'opposent ?

Y a-t-il des candidats pour lesquels la répartition des votes est la même dans toutes les régions ? Y a-t-il des candidats pour lesquelles le vote est concentré dans certaines régions ?

Comment les régions "à part" et les candidats à "vote inégalement réparti" s'associent-ils ?

5.2.2 Etude descriptive du tableau de contingence

On fixe les notations suivantes :

n_{ij} : effectif de la cellule (i,j),

$n_{i\bullet}$: effectif total de la ligne i,

$n_{\bullet j}$: effectif total de la colonne j

$n_{\bullet\bullet}$: effectif total

5.2.2.1 Tableau des fréquences

Les fréquences sont calculées par : $f_{ij} = \frac{n_{ij}}{n_{\bullet\bullet}} = \frac{Effectif\ \ de\ \ la\ \ cellule\ (i,j)}{Effectif\ \ total}$

	Sarkozy	Bayrou	Royal	Le Pen	Besanc.	Villiers	Voynet	Laguiller	Bove	Buffet	Nihous	Schivardi	Total
Alsace	1,00	0,59	0,47	0,37	0,09	0,06	0,06	0,04	0,04	0,02	0,01	0,01	2,75
Aquitaine	1,46	1,15	1,53	0,46	0,21	0,09	0,08	0,06	0,08	0,11	0,10	0,02	5,36
Auvergne	0,65	0,47	0,62	0,22	0,11	0,05	0,03	0,04	0,04	0,06	0,03	0,01	2,33
Bourgogne	0,82	0,48	0,66	0,33	0,12	0,07	0,04	0,04	0,03	0,05	0,03	0,01	2,68
Bretagne	1,53	1,24	1,55	0,40	0,26	0,11	0,11	0,07	0,08	0,09	0,06	0,01	5,51
Centre	1,27	0,76	0,95	0,46	0,18	0,13	0,06	0,06	0,05	0,08	0,06	0,02	4,07
Champ Ard.	0,68	0,34	0,44	0,31	0,09	0,06	0,03	0,04	0,02	0,03	0,02	0,01	2,07
Corse	0,16	0,05	0,09	0,06	0,02	0,01	0,01	0,00	0,00	0,01	0,01	0,00	0,42
Fr-Comte	0,58	0,31	0,45	0,26	0,08	0,04	0,04	0,03	0,03	0,03	0,02	0,01	1,89
Ile-de-Fr.	5,31	3,14	4,38	1,18	0,50	0,25	0,25	0,15	0,16	0,30	0,05	0,03	15,70
Lang-Rous	1,29	0,64	1,09	0,59	0,17	0,08	0,06	0,05	0,08	0,11	0,06	0,03	4,24
Limousin	0,34	0,23	0,39	0,11	0,07	0,03	0,02	0,02	0,02	0,04	0,02	0,01	1,28
Lorraine	1,11	0,69	0,87	0,54	0,19	0,08	0,06	0,07	0,04	0,05	0,03	0,01	3,75
Midi-Pyr	1,26	0,94	1,49	0,43	0,19	0,08	0,07	0,05	0,09	0,09	0,07	0,02	4,79
Nord-PdeCa	1,76	0,94	1,57	0,92	0,35	0,11	0,09	0,14	0,07	0,20	0,12	0,02	6,29
Basse-Nor	0,77	0,51	0,58	0,24	0,12	0,07	0,04	0,04	0,04	0,03	0,06	0,01	2,50
Haute-Nor	0,85	0,51	0,71	0,35	0,16	0,07	0,04	0,06	0,03	0,07	0,04	0,01	2,90
Pays Loire	1,75	1,26	1,52	0,44	0,26	0,30	0,11	0,08	0,07	0,07	0,07	0,02	5,94
Picardie	0,91	0,44	0,69	0,46	0,16	0,07	0,04	0,07	0,03	0,06	0,06	0,01	3,01
Poitou-Char	0,84	0,53	0,89	0,24	0,13	0,11	0,04	0,04	0,04	0,04	0,06	0,01	2,96
PACA	2,78	1,15	1,59	1,04	0,24	0,15	0,11	0,07	0,10	0,17	0,07	0,03	7,50
Rhone-Alp.	3,08	1,90	2,22	0,99	0,34	0,21	0,17	0,10	0,14	0,16	0,09	0,03	9,43
Outremer	0,93	0,29	1,09	0,10	0,06	0,01	0,03	0,03	0,04	0,04	0,01	0,00	2,64
Total	31,11	18,55	25,83	10,51	4,11	2,24	1,57	1,34	1,32	1,94	1,15	0,34	100,00

5.2.2.2 Tableau des fréquences lignes

Les fréquences lignes (ou coordonnées des profils lignes) sont calculées par :

$$fl_{ij} = \frac{n_{ij}}{n_{i\bullet}} = \frac{f_{ij}}{f_{i\bullet}} = \frac{Effectif\ \ de\ \ la\ \ cellule\ (i,j)}{Effectif\ \ de\ \ la\ \ ligne\ \ i}$$

Les coordonnées du profil ligne moyen (dans le tableau des fréquences) sont calculées par : $f_{\bullet j} = \frac{n_{\bullet j}}{n_{\bullet\bullet}} = \frac{Effectif\ \ de\ \ la\ \ colonne\ \ j}{Effectif\ \ total}$

	Sarkozy	Bayrou	Royal	Le Pen	Besancenot	Villiers	Voynet	Laguiller	Bove	Buffet	Nihous	Schivardi	Total
Alsace	36,20	21,40	17,11	13,56	3,33	2,25	2,04	1,38	1,37	0,61	0,51	0,25	100,00
Aquitaine	27,28	21,40	28,57	8,65	4,01	1,74	1,45	1,13	1,43	2,14	1,81	0,39	100,00
Auvergne	28,05	19,95	26,55	9,27	4,89	2,21	1,42	1,52	1,59	2,58	1,47	0,50	100,00
Bourgogne	30,54	17,98	24,74	12,22	4,34	2,56	1,41	1,48	1,26	1,86	1,24	0,37	100,00
Bretagne	27,81	22,55	28,14	7,18	4,70	2,06	1,95	1,28	1,42	1,59	1,06	0,26	100,00
Centre	31,06	18,76	23,29	11,39	4,41	3,08	1,53	1,50	1,17	2,02	1,39	0,38	100,00
Champ-Ard.	32,74	16,28	21,27	15,20	4,44	2,71	1,42	1,80	0,98	1,65	1,20	0,30	100,00
Corse	37,00	12,36	21,81	15,26	3,87	1,24	1,38	0,88	1,08	3,36	1,47	0,29	100,00
Fr-Comte	30,93	16,63	24,07	13,72	4,47	2,38	1,88	1,58	1,51	1,34	1,14	0,36	100,00
Ile-de-France	33,81	20,01	27,89	7,54	3,17	1,57	1,57	0,93	1,01	1,94	0,35	0,22	100,00
Lang-Rous.	30,48	15,22	25,64	13,91	4,06	1,83	1,35	1,11	1,78	2,50	1,38	0,74	100,00
Limousin	26,51	17,64	30,44	8,24	5,14	2,31	1,31	1,41	1,42	3,36	1,72	0,49	100,00
Lorraine	29,62	18,35	23,14	14,43	5,20	2,14	1,58	1,88	1,18	1,41	0,76	0,31	100,00
Midi-Pyr	26,29	19,61	31,11	8,88	3,97	1,77	1,45	1,07	1,93	1,96	1,45	0,50	100,00
Nord-PdC	27,92	14,88	25,02	14,66	5,58	1,78	1,37	2,30	1,07	3,23	1,90	0,28	100,00
Basse-Norm	30,84	20,23	22,98	9,72	4,83	2,82	1,58	1,61	1,45	1,23	2,36	0,34	100,00

Haute-Norm	29,38	17,50	24,42	12,02	5,53	2,44	1,49	1,90	1,19	2,50	1,30	0,34	100,00
Pays Loire	29,48	21,18	25,56	7,35	4,34	4,99	1,80	1,32	1,19	1,24	1,23	0,30	100,00
Picardie	30,26	14,74	23,02	15,42	5,28	2,44	1,30	2,19	1,06	2,04	1,87	0,38	100,00
Poitou-Char	28,22	17,99	29,86	8,17	4,23	3,59	1,51	1,29	1,29	1,47	2,00	0,36	100,00
PACA	37,01	15,36	21,21	13,84	3,24	2,01	1,40	0,97	1,35	2,27	0,97	0,36	100,00
Rhone-Alpes	32,67	20,10	23,51	10,51	3,61	2,21	1,84	1,11	1,50	1,68	0,95	0,32	100,00
Outremer	35,18	10,83	41,48	3,83	2,31	0,54	1,29	1,07	1,55	1,47	0,28	0,18	100,00

5.2.2.3 Tableau des fréquences colonnes

Les fréquences colonnes (ou coordonnées des profils colonnes) sont calculées par :

$$fc_{ij} = \frac{n_{ij}}{n_{\bullet j}} = \frac{f_{ij}}{f_{\bullet j}} = \frac{Effectif\ \ de\ \ la\ \ cellule\ \ (i,j)}{Effectif\ \ de\ \ la\ \ colonne\ \ j}$$

Les coordonnées du profil colonne moyen (dans le tableau des fréquences) sont calculées par :

$$f_{i\bullet} = \frac{n_{i\bullet}}{n_{\bullet\bullet}} = \frac{Effectif\ \ de\ \ la\ \ ligne\ \ i}{Effectif\ \ total}$$

	Sarkozy	Bayrou	Royal	Le Pen	Besanc.	Villiers	Voynet	Laguiller	Bove	Buffet	Nihous	Schivardi
Alsace	3,20	3,17	1,82	3,55	2,23	2,76	3,57	2,84	2,87	0,86	1,22	2,05
Aquitaine	4,70	6,19	5,93	4,41	5,23	4,17	4,96	4,53	5,83	5,92	8,40	6,14
Auvergne	2,10	2,51	2,40	2,06	2,78	2,30	2,12	2,66	2,82	3,11	2,97	3,41
Bourgogne	2,63	2,60	2,56	3,11	2,83	3,06	2,40	2,97	2,57	2,57	2,88	2,93
Bretagne	4,92	6,70	6,00	3,76	6,30	5,05	6,84	5,27	5,95	4,52	5,05	4,19
Centre	4,07	4,12	3,67	4,42	4,37	5,60	3,97	4,58	3,63	4,25	4,90	4,62
Champ-Ard	2,18	1,82	1,70	2,99	2,24	2,51	1,88	2,79	1,55	1,77	2,15	1,84
Corse	0,50	0,28	0,36	0,61	0,40	0,23	0,37	0,28	0,35	0,73	0,54	0,36
Fr-Comte	1,88	1,69	1,76	2,46	2,05	2,01	2,26	2,24	2,16	1,30	1,86	1,98
Ile de France	17,06	16,93	16,94	11,26	12,13	10,97	15,76	10,89	11,99	15,73	4,73	10,05
Lang. Rous.	4,15	3,48	4,21	5,61	4,19	3,45	3,65	3,53	5,72	5,47	5,08	9,27
Limousin	1,09	1,22	1,51	1,01	1,61	1,32	1,08	1,35	1,38	2,22	1,91	1,85
Lorraine	3,57	3,71	3,36	5,14	4,75	3,58	3,78	5,26	3,36	2,73	2,47	3,42
Midi-Pyr	4,05	5,06	5,76	4,05	4,63	3,78	4,43	3,83	7,03	4,83	6,02	7,12
Nord-PdC	5,65	5,05	6,09	8,78	8,56	4,99	5,50	10,83	5,13	10,49	10,38	5,15
Basse-Norm	2,48	2,73	2,23	2,32	2,94	3,15	2,52	3,01	2,75	1,59	5,11	2,51
Haute-Norm	2,74	2,74	2,74	3,32	3,90	3,15	2,75	4,12	2,62	3,74	3,27	2,92
Pays Loire	5,62	6,78	5,87	4,15	6,27	13,23	6,83	5,85	5,39	3,79	6,35	5,26
Picardie	2,92	2,39	2,68	4,41	3,87	3,28	2,49	4,93	2,43	3,17	4,87	3,40
Poitou-Char	2,69	2,88	3,43	2,30	3,05	4,75	2,86	2,87	2,91	2,25	5,13	3,15
PACA	8,92	6,21	6,16	9,88	5,91	6,72	6,72	5,44	7,71	8,78	6,32	8,06
Rhone-Alpes	9,91	10,22	8,58	9,43	8,28	9,31	11,05	7,83	10,74	8,17	7,77	8,90
Outremer	2,98	1,54	4,23	0,96	1,48	0,63	2,17	2,10	3,11	1,99	0,64	1,44
Total	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00

5.2.2.4 Distances entre profils. Métrique du Φ^2

Chaque ligne du tableau des fréquences lignes peut être vue comme la liste des coordonnées d'un point dans un espace à q dimensions. On obtient ainsi le nuage des individus-lignes. On définit de même le nuage des individus-colonnes à partir du tableau des fréquences colonnes.

Comme en ACP, on s'intéresse alors aux directions de "plus grande dispersion" de chacun de ces nuages de points. Mais, pour mesurer la "distance" entre deux individus, on utilise la *métrique du* Φ^2 au lieu de la distance habituelle (dite *métrique euclidienne*). La distance du Φ^2 entre la ligne i et la ligne i' est ainsi définie par :

$$d^2_{\Phi^2}(L_i, L_{i'}) = \sum_j \frac{(fl_{ij} - fl_{i'j})^2}{f_{\bullet j}}$$

Pourquoi utiliser cette métrique plutôt que la métrique euclidienne ? Deux raisons fortes peuvent être avancées :

- Avec la métrique du Φ^2, la distance entre deux lignes ne dépend pas des poids respectifs des colonnes. Ainsi, sur notre exemple, les différents candidats obtiennent des scores très différents et l'usage de la métrique euclidienne aurait donné trop de poids aux candidats qui ont obtenu des scores élevés (Sarkozy, Royal, Bayrou).

- La métrique du Φ^2 possède la propriété d'*équivalence distributionnelle* : si on regroupe deux modalités lignes, les distances entre les profils-colonne, ou entre les autres profils-lignes restent inchangées.

Notons qu'en revanche, il n'existe pas d'outil mesurant une "distance" entre une ligne et une colonne.

5.2.2.5 Taux de liaison et Phi-2

Les taux de liaison sont définis par : $t_{ij} = \dfrac{f_{ij} - f_{i\bullet}.f_{\bullet j}}{f_{i\bullet}.f_{\bullet j}}$

	Sarkozy	Bayrou	Royal	Le Pen	Besanc.	Villiers	Voynet	Laguiller	Bove	Buffet	Nihous	Schivardi
Alsace	0,16	0,15	-0,34	0,29	-0,19	0,00	0,30	0,03	0,04	-0,69	-0,56	-0,26
Aquitaine	-0,12	0,15	0,11	-0,18	-0,02	-0,22	-0,07	-0,15	0,09	0,11	0,57	0,15
Auvergne	-0,10	0,08	0,03	-0,12	0,19	-0,02	-0,09	0,14	0,21	0,33	0,27	0,46
Bourgogne	-0,02	-0,03	-0,04	0,16	0,06	0,14	-0,10	0,11	-0,04	-0,04	0,07	0,09
Bretagne	-0,11	0,22	0,09	-0,32	0,14	-0,08	0,24	-0,04	0,08	-0,18	-0,08	-0,24
Centre	0,00	0,01	-0,10	0,08	0,07	0,38	-0,02	0,12	-0,11	0,04	0,20	0,13
Champ. Ard	0,05	-0,12	-0,18	0,45	0,08	0,21	-0,09	0,35	-0,25	-0,15	0,04	-0,11
Corse	0,19	-0,33	-0,16	0,45	-0,06	-0,45	-0,12	-0,34	-0,18	0,73	0,27	-0,14
Fr-Comte	-0,01	-0,10	-0,07	0,31	0,09	0,06	0,20	0,19	0,15	-0,31	-0,01	0,05
Ile-de-France	0,09	0,08	0,08	-0,28	-0,23	-0,30	0,00	-0,31	-0,24	0,00	-0,70	-0,36
Lang-Rous	-0,02	-0,18	-0,01	0,32	-0,01	-0,19	-0,14	-0,17	0,35	0,29	0,20	1,19
Limousin	-0,15	-0,05	0,18	-0,22	0,25	0,03	-0,16	0,05	0,08	0,73	0,49	0,44
Lorraine	-0,05	-0,01	-0,10	0,37	0,27	-0,05	0,01	0,40	-0,10	-0,27	-0,34	-0,09
Midi-Pyr	-0,15	0,06	0,20	-0,15	-0,03	-0,21	-0,07	-0,20	0,47	0,01	0,26	0,49
Nord-PdC	-0,10	-0,20	-0,03	0,40	0,36	-0,21	-0,13	0,72	-0,18	0,67	0,65	-0,18
Basse-Norma	-0,01	0,09	-0,11	-0,07	0,18	0,26	0,01	0,20	0,10	-0,36	1,04	0,00
Haute-Norma	-0,06	-0,06	-0,05	0,14	0,35	0,09	-0,05	0,42	-0,10	0,29	0,13	0,01
Pays Loire	-0,05	0,14	-0,01	-0,30	0,06	1,23	0,15	-0,01	-0,09	-0,36	0,07	-0,11
Picardie	-0,03	-0,21	-0,11	0,47	0,29	0,09	-0,17	0,64	-0,19	0,05	0,62	0,13
Poitou-Char	-0,09	-0,03	0,16	-0,22	0,03	0,60	-0,03	-0,03	-0,02	-0,24	0,73	0,06
PACA	0,19	-0,17	-0,18	0,32	-0,21	-0,10	-0,10	-0,27	0,03	0,17	-0,16	0,07
Rhone-Alpes	0,05	0,08	-0,09	0,00	-0,12	-0,01	0,17	-0,17	0,14	-0,13	-0,18	-0,06
Outremer	0,13	-0,42	0,61	-0,64	-0,44	-0,76	-0,18	-0,20	0,18	-0,24	-0,76	-0,46

Signification pratique du taux de liaison : le score de Sarkozy en Alsace est 16% plus élevé que le score théorique que l'on observerait si les votes étaient indépendants des régions. Au contraire, celui de Royal est 34% moins élevé que le score théorique.

Par construction, les valeurs prises par le taux de liaison sont :

- des nombres positifs quelconques (un score observé peut être 200% ou 300% supérieur au score théorique)
- des nombres négatifs compris entre -1 et 0 (le "déficit" le plus extrême d'un score observé est d'être 100% moins élevé que le score théorique).

Notez que le coefficient $f_{i\bullet}.f_{\bullet j}$ représente le "poids théorique" de chaque cellule dans le tableau. La somme de ces coefficients vaut 1.

La moyenne de la série des taux de liaison pondérée par les coefficients $f_{i\bullet}.f_{\bullet j}$ est nulle. La variance de cette série (avec les mêmes pondérations) est le coefficient Φ^2 :

$$\Phi^2 = \sum_{i,j} f_{i\bullet}.f_{\bullet j}.t_{ij}^2 = \sum_{i,j} \frac{(f_{ij} - f_{i\bullet}.f_{\bullet j})^2}{f_{i\bullet}.f_{\bullet j}} = \frac{\mathrm{X}^2}{n_{\bullet\bullet}}$$

Ici, on obtient : $\Phi^2 = 0{,}03341$.

La méthode d'analyse factorielle des correspondances peut être vue comme une décomposition pertinente du Φ^2 selon plusieurs axes factoriels.

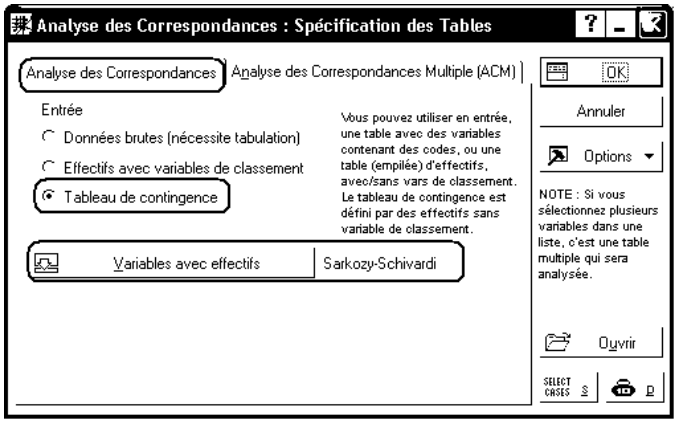
5.2.3 L'analyse factorielle des correspondances proprement dite

L'application de la méthode a deux effets :

- d'une part, on construit des images des nuages d'"individus-lignes" et d'"individus-colonnes" de départ, de façon que les distances entre images soient des distances euclidiennes et non plus des distances calculées selon la métrique du Φ^2;
- d'autre part, on recherche les directions de plus grande dispersion dans ces nuages de points images.

Ouvrez le classeur Statistica Presidentielles-2007.stw et rendez active la feuille "Par-région".

Utilisez ensuite le menu Statistiques - Techniques exploratoires multivariées - Analyse des correspondances, indiquez la forme des données d'entrée (tableau de contingence) et sélectionnez les variables qui vont être étudiées :



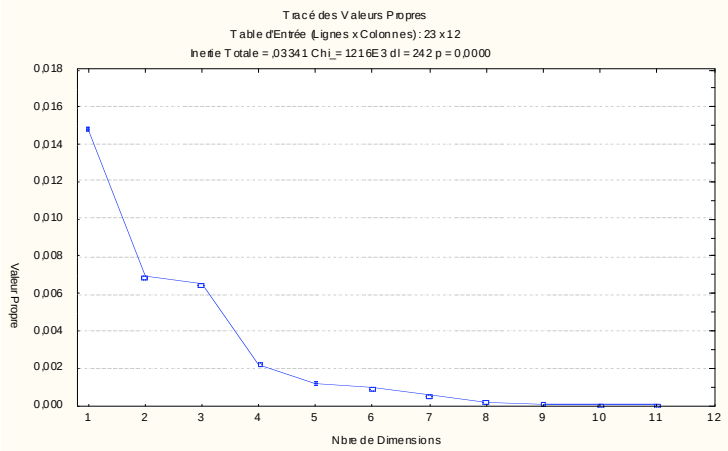
5.2.3.1 Valeurs propres

Le nombre de valeurs propres produites par la recherche des facteurs principaux est égal au minimum du nombre de lignes et du nombre de colonnes du tableau de contingence. Cependant, la première valeur propre est systématiquement égale à 1, et n'est pas mentionnée dans les résultats. Les autres valeurs propres sont des nombres positifs inférieurs à 1 et leur somme est égale à Φ^2.

Les boutons "Valeurs propres" de l'onglet "Avancé" permettent d'obtenir les éléments suivants.

Valeurs Propres et Inertie de toutes les Dimensions (Par-Region dans Presidentielles-2007.stw)
Inertie Totale = ,03341 Chi² = 1216E3 dl = 242 p = 0,0000

	ValSing.	ValProp.	%age Inertie	%age Cumulé	Chi²
1	0,12148	0,01476	44,17	44,17	537082
2	0,08322	0,00693	20,73	64,90	252072
3	0,08075	0,00652	19,52	84,42	237324
4	0,04709	0,00222	6,64	91,05	80692
5	0,03449	0,00119	3,56	94,61	43300
6	0,03083	0,00095	2,85	97,46	34602
7	0,02274	0,00052	1,55	99,01	18827
8	0,01333	0,00018	0,53	99,54	6465
9	0,00939	0,00009	0,26	99,80	3209
10	0,00667	0,00004	0,13	99,94	1619
11	0,00463	0,00002	0,06	100,00	781



Le choix du nombre d'axes factoriels à conserver se fait comme dans le cas de l'ACP. Ici, on observe une brusque décroissance des valeurs propres entre la 3è et la 4è valeur propre. On retient donc les 3 premiers axes factoriels.

5.2.3.2 Résultats relatifs aux individus-lignes

Après avoir fixé le nombre de valeurs propres retenues (onglet "Base") on obtient les résultats relatifs aux individus-lignes en cliquant sur le bouton "Coordonnées lignes et colonnes".

Coordonnées Ligne et Contributions à l'Inertie (Par-Region dans Presidentielles-2007-v2.stw)
Table d'Entrée (Lignes x Colonnes) : 23 x 12
Standardisation : Profils ligne et colonne

NomLigne	ligne numéi	Coord. Dim.1	Coord. Dim.2	Coord. Dim.3	Masse	Qualité	Inertie Relative	Inertie Dim.1	Cosinus² Dim.1	Inertie Dim.2	Cosinus² Dim.2	Inertie Dim.3	Cosinus
Alsace	1	-0,1219	0,1067	-0,1876	0,0275	0,9192	0,0550	0,0277	0,2222	0,0452	0,1704	0,1485	0,526
Aquitaine	2	0,0821	-0,0806	0,0276	0,0536	0,6685	0,0336	0,0245	0,3220	0,0503	0,3101	0,0063	0,036
Auvergne	3	0,0251	-0,0807	0,0272	0,0233	0,6607	0,0083	0,0010	0,0526	0,0220	0,5461	0,0026	0,062
Bourgogn	4	-0,0607	-0,0139	-0,0006	0,0268	0,8474	0,0037	0,0067	0,8052	0,0007	0,0421	0,0000	0,000
Bretagne	5	0,1259	-0,0753	-0,0354	0,0551	0,8223	0,0456	0,0591	0,5724	0,0450	0,2046	0,0106	0,045
Centre	6	-0,0593	-0,0350	-0,0397	0,0407	0,8317	0,0093	0,0097	0,4634	0,0072	0,1608	0,0099	0,207
Champag	7	-0,1799	0,0291	-0,0207	0,0207	0,9107	0,0229	0,0454	0,8761	0,0025	0,0229	0,0014	0,011
Corse	8	-0,1891	0,1485	0,0941	0,0042	0,8574	0,0098	0,0102	0,4598	0,0134	0,2838	0,0057	0,113
Franche-(	9	-0,1048	0,0093	0,0012	0,0189	0,6643	0,0094	0,0140	0,6590	0,0002	0,0052	0,0000	0,000
Ile-de-Fra	10	0,1263	0,0854	-0,0231	0,1570	0,9244	0,1209	0,1697	0,6201	0,1654	0,2836	0,0128	0,020
Languedc	11	-0,1029	0,0235	0,0902	0,0424	0,7125	0,0343	0,0304	0,3914	0,0034	0,0205	0,0529	0,300
Limousin	12	0,0673	-0,1103	0,1199	0,0128	0,8325	0,0143	0,0039	0,1214	0,0226	0,3259	0,0283	0,385
Lorraine	13	-0,1264	-0,0054	-0,0137	0,0375	0,6174	0,0294	0,0406	0,6090	0,0002	0,0011	0,0011	0,007
Midi-Pyre	14	0,1022	-0,0708	0,0738	0,0479	0,7606	0,0394	0,0339	0,3799	0,0347	0,1827	0,0399	0,198
Nord-Pas	15	-0,1688	-0,0608	0,1367	0,0629	0,9286	0,1032	0,1215	0,5201	0,0336	0,0675	0,1803	0,341
Basse-Nc	16	-0,0283	-0,0936	-0,0604	0,0250	0,5544	0,0179	0,0014	0,0336	0,0317	0,3676	0,0140	0,153
Haute-No	17	-0,0824	-0,0522	0,0329	0,0290	0,7599	0,0121	0,0133	0,4867	0,0114	0,1957	0,0048	0,077
Pays-de-l	18	0,0760	-0,1216	-0,1343	0,0594	0,7547	0,0909	0,0233	0,1131	0,1267	0,2890	0,1642	0,352
Picardie	19	-0,1995	-0,0360	0,0648	0,0301	0,9326	0,0437	0,0811	0,8195	0,0056	0,0267	0,0193	0,086
Poitou-Ch	20	0,0786	-0,1151	0,0190	0,0296	0,6629	0,0265	0,0124	0,2068	0,0567	0,4440	0,0016	0,012
Provence	21	-0,1308	0,1378	-0,0105	0,0750	0,9131	0,0890	0,0870	0,4315	0,2056	0,4788	0,0013	0,002
Rhone-Al	22	-0,0028	0,0338	-0,0659	0,0943	0,8336	0,0186	0,0001	0,0012	0,0156	0,1731	0,0629	0,659
Outremer	23	0,3201	0,1623	0,2393	0,0264	0,9058	0,1622	0,1831	0,4989	0,1003	0,1282	0,2316	0,278

Le tableau ci-dessus rassemble tous les résultats relatifs aux individus-lignes.

La colonne "Masse" rappelle les fréquences marginales des lignes c'est-à-dire le profil colonne moyen. Contrairement à l'ACP normée, dans laquelle chaque individu était affecté du même poids, les régions ont ici un "poids" dépendant de l'effectif total d'électeurs inscrits dans le département.

La colonne "Qualité" indique les qualités de représentation des individus ligne par les trois premiers axes principaux.

Pour chacun des trois axes factoriels, le tableau nous donne les coordonnées ou *scores factoriels* de l'individu-ligne selon cet axe.

Le tableau donne également la contribution de chaque individu à la formation de l'axe, ou inertie selon cet axe. Ces valeurs sont des contributions relatives (la somme de la colonne vaut 1). On peut donc utiliser des colonnes pour rechercher quels sont les individus-lignes qui ont eu une influence supérieure à la moyenne dans la formation de l'axe factoriel considéré.

Enfin, ce tableau nous donne les cosinus-carrés ou qualités de représentation des individus-lignes par chaque axe factoriel. L'interprétation géométrique de ces valeurs est analogue à celle développée pour l'ACP : c'est le carré du cosinus de l'angle entre le vecteur représentant l'Alsace dans l'espace à 11 dimensions et sa projection sur le premier axe factoriel.

5.2.3.3 Résultats relatifs aux individus-colonnes

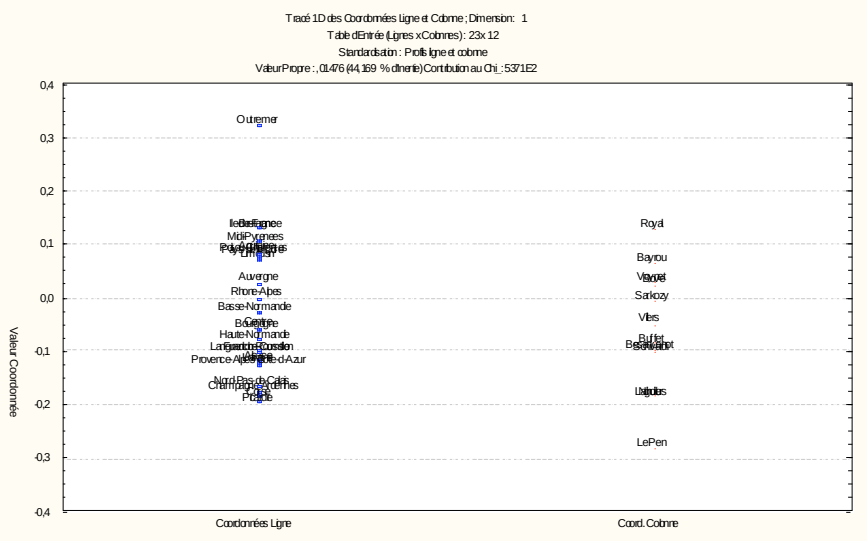
Dans une AFC, les individus-lignes et les individus-colonnes jouent des rôles symétriques. Les résultats relatifs aux individus-colonnes s'interprètent donc de la même façon que les résultats relatifs aux individus-lignes.

	Coordonnées Colonne et Contributions à l'Inertie (Par-Region dans Presidentielles-2007-v2.stw) Table d'Entrée (Lignes x Colonnes) : 23 x 12 Standardisation : Profils ligne et colonne												
Nom Col.	lonr umér	Coord. Dim.1	Coord. Dim.2	Coord. Dim.3	Masse	Qualité	Inertie Relative	Inertie Dim.1	Cosinus² Dim.1	Inertie Dim.2	Cosinus² Dim.2	Inertie Dim.3	Cosinus² Dim.3
Sarkozy	1	-0,0071	0,0911	-0,0274	0,3111	0,9194	0,0922	0,0011	0,0050	0,3731	0,8383	0,0359	0,0760
Bayrou	2	0,0610	-0,0500	-0,1032	0,1855	0,8085	0,1159	0,0468	0,1785	0,0670	0,1200	0,3028	0,5100
Royal	3	0,1244	-0,0146	0,0864	0,2583	0,9791	0,1829	0,2710	0,6547	0,0080	0,0091	0,2955	0,3154
Le Pen	4	-0,2839	0,0274	0,0297	0,1051	0,9824	0,2633	0,5739	0,9629	0,0114	0,0090	0,0142	0,0105
Besancer	5	-0,1026	-0,1497	0,0368	0,0411	0,8532	0,0494	0,0293	0,2619	0,1328	0,5576	0,0085	0,0337
Villiers	6	-0,0527	-0,2195	-0,2150	0,0224	0,6467	0,1008	0,0042	0,0185	0,1559	0,3205	0,1589	0,3077
Voynet	7	0,0269	-0,0095	-0,0991	0,0157	0,6394	0,0078	0,0008	0,0434	0,0002	0,0054	0,0236	0,5906
Laguiller	8	-0,1878	-0,1426	0,0812	0,0134	0,6823	0,0365	0,0319	0,3869	0,0392	0,2231	0,0135	0,0723
Bove	9	0,0222	-0,0298	0,0358	0,0132	0,0736	0,0143	0,0004	0,0136	0,0017	0,0245	0,0026	0,0355
Buffet	10	-0,0886	-0,0149	0,1921	0,0194	0,5316	0,0491	0,0103	0,0928	0,0006	0,0026	0,1097	0,4362
Nihous	11	-0,1880	-0,3514	0,1310	0,0115	0,7982	0,0762	0,0276	0,1602	0,2058	0,5602	0,0304	0,0778
Schivardi	12	-0,1060	-0,0930	0,0921	0,0034	0,2435	0,0118	0,0026	0,0964	0,0042	0,0743	0,0044	0,0728

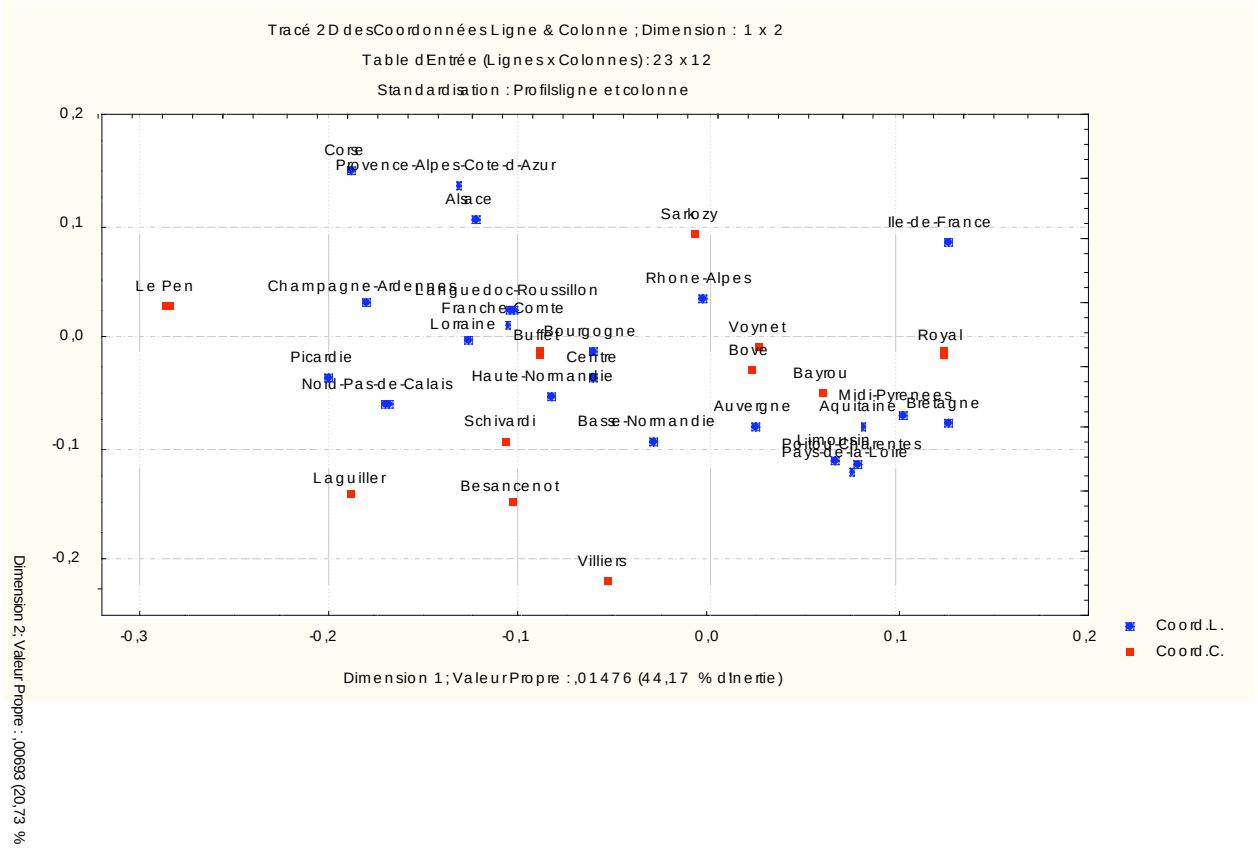
5.2.3.4 Résultats graphiques

Les transformations et les pondérations introduites rendent tout à fait comparables les valeurs obtenues pour les individus lignes et les individus colonnes. Contrairement à l'ACP, les graphiques factoriels pourront être construits en faisant figurer sur un même graphique les individus lignes et les individus colonnes.
On peut réaliser et essayer d'interpréter des graphiques :

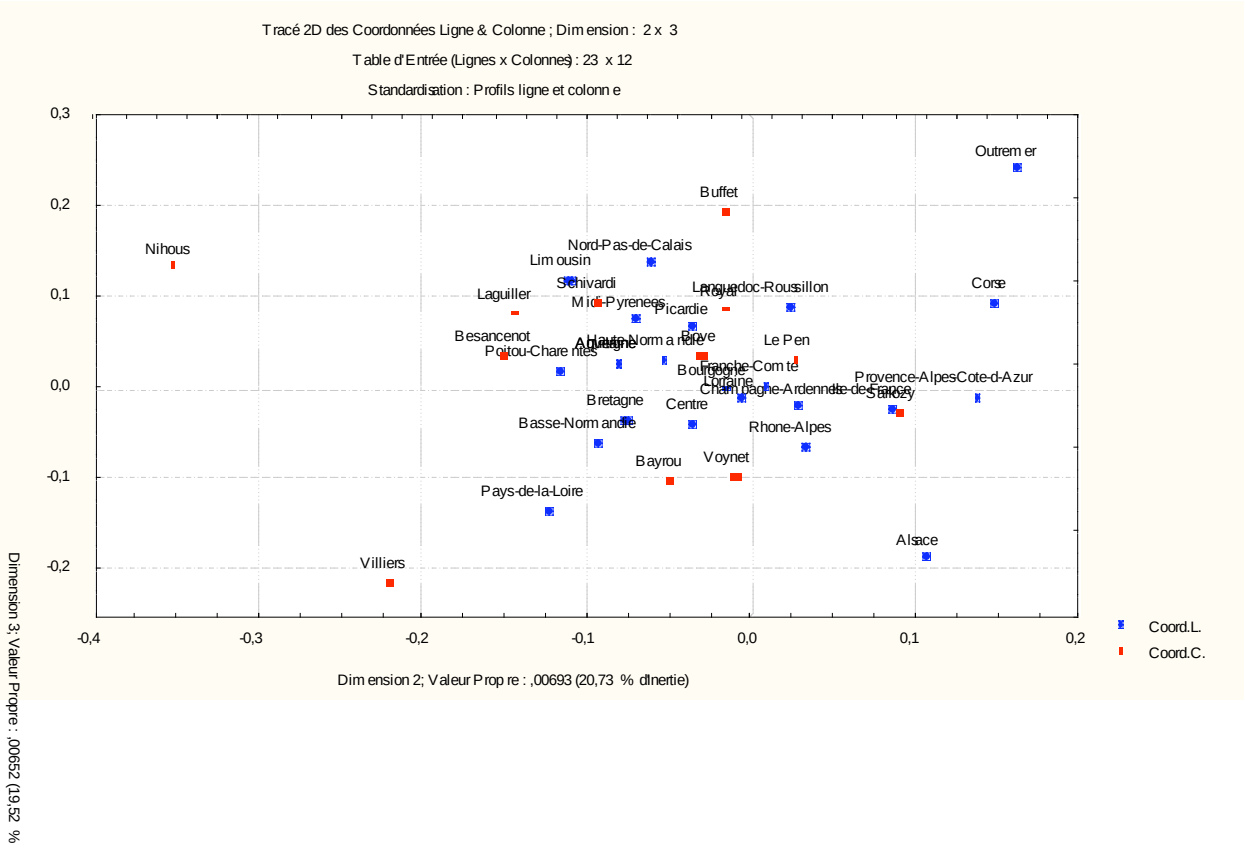
- en dimension 1 : on place les individus le long d'un axe factoriel,
- en dimension 2 : on place les individus dans un plan défini à partir de deux axes factoriels,
- éventuellement, en dimension 3 : on place les individus dans une représentation en perspective d'un espace à 3 dimensions.



Graphique selon les axes 1 et 2
N.B. L'individu ligne "Outremer" et l'individu colonne "Nihous" sont en dehors du dessin.



Graphique selon les axes 2 et 3



5.2.3.5 Interprétation géométrique

Les distances entre deux individus-lignes, ou entre un indivivu-ligne et l'origine des axes, peuvent être facilement interprétées. En effet : la distance euclidienne entre deux points-lignes, représentés par leurs coordonnées factorielles est égale à la distance du Φ^2 entre les profils-lignes initiaux.

La proximité entre un point-ligne L et un point-colonne C ne possède pas d'interprétation géométrique immédiate. En revanche, l'angle de sommet O dont les côtés passent par L et C a la propriété suivante :

- si l'angle (OL, OC) est aigu, la modalité-ligne L et la modalité colonne C s'attirent (taux de liaison positif)
- si l'angle (OL, OC) est obtus, la modalité-ligne L et la modalité colonne C se repoussent (taux de liaison négatif)
- si l'angle (OL, OC) est droit, la modalité-ligne L et la modalité colonne C n'interagissent pas (taux de liaison voisin de 0).

5.2.3.6 Reconstitution des données

Il est possible de reconstituer les données à partir des scores factoriels des lignes et des colonnes. En effet, on peut montrer la relation suivante entre les taux de liaison t_{ij}, les scores factoriels des lignes, les scores factoriels des colonnes et les valeurs propres :

$$t_{ij} = \sum_{Axes\ \ factoriels} \frac{(Score\ \ fact.\ \ ligne\ \ i\ \ selon\ \ axe\ \ \alpha)(Score\ \ fact.\ \ colonne\ \ j\ \ selon\ \ axe\ \ \alpha)}{\sqrt{Valeur\ \ propre\ \ associée\ \ à\ \ l'axe\ \ \alpha}}$$

Connaissant les profils moyens des lignes et des colonnes, et l'effectif total N, l'ensemble des données peut ainsi être retrouvé.

5.2.4 Interprétation des résultats de l'AFC

Au niveau global, on pourra noter que les inerties relatives les plus fortes sont observées sur l'Outremer, l'Ile de France, le Nord Pas de Calais et les Pays de la Loire pour les régions, et sur Le Pen, Royal, Bayrou, Villiers et Sarkozy pour les candidats. Ce sont donc essentiellement ces modalités lignes et modalités colonnes qui vont apparaître dans l'étude qui suit. On pourra noter que ces modalités correspondent soit à des modalités de poids important (Ile de France, Nord Pas de Calais, Royal, Sarkozy) soit à des modalités éloignées du profil moyen (Outremer, Pays de Loire, Villiers, Le Pen).

L'interprétation pourra être faite axe par axe, en étudiant d'abord séparément lignes et colonnes.
Pour chaque axe, on pourra dresser un tableau des individus qui ont apporté une contribution supérieure à la moyenne à la formation de cet axe.

5.2.4.1 Interprétation des axes

Pour le premier axe :

- Points lignes :

-	+
Nord - Pas de Calais (12%) Provence - Alpes Côte d'Azur (9%) Picardie (8%) Champagne Ardennes (5%)	Outremer (18%) Ile de France (17%) Bretagne (6%)

- Points colonnes :

-	+
Le Pen (57%)	Royal (27%)

Le premier axe oppose les régions du Nord et de l'Est, et la région PACA d'une part, à des régions telles que l'Outremer, l'Ile de France et la Bretagne.

Pour les modalités colonnes, cet axe est essentiellement unipolaire (la modalité Le Pen représente plus de la moitié de son inertie) et oppose les modalités Le Pen et Royal.

La synthèse entre l'analyse des lignes et des colonnes montre que cet axe oppose les régions où le vote pour le candidat Le Pen est supérieur à la moyenne nationale à celles où ce vote est inférieur à la moyenne (particulièrement l'Outremer, notamment). On constate que ces dernières sont également des régions de fort vote "Royal". Il ne faudrait pas pour autant en conclure que les deux candidats recrutent leurs voix dans le même électorat. C'est vraisemblablement plutôt l'ensemble du corps électoral qui possède une sensibilité plus "à gauche" dans certaines régions. Il faut également remarquer que le candidat Sarkozy intervient peu dans la formation de cet axe.

Pour le deuxième axe :

- Points lignes :

-	+
Pays de la Loire (13%) Poitou-Charentes (6%) Aquitaine (5%) Bretagne (5%)	Provence Alpes Côte d'Azur (21%) Ile de France (17%) Outremer (10%) Alsace (5%)

- Points colonnes :

-	+
Nihous (20%)	Sarkozy (37%)

Villiers (15%) Besancenot (13%)	

Cet axe oppose les régions de l'Ouest à des régions telles que PACA, l'Ile de France, l'Outremer et l'Alsace.

Pour les modalités colonnes, cet axe oppose certains "petits" candidats au candidat Sarkozy, mais il est, dans une certaine mesure unipolaire (Sarkozy représente 37% de son inertie)..

Cet axe est essentiellement organisé autour du vote pour le candidat Sarkozy : la partie positive de l'axe correspond aux régions où le vote Sarkozy est supérieur à la moyenne nationale, tandis que la partie négative correspond à des régions où ce vote est inférieur à la moyenne. Il semble par ailleurs qu'un faible vote pour Sarkozy soit lié à un vote plus significatif pour certains petits candidats.

Pour le troisième axe :

- Points lignes :

-	+
Pays de la Loire (16%) Alsace (14%) Rhône Alpes (6%)	Outremer (23%) Nord Pas de Calais (18%) Languedoc Roussillon (5%)

- Points colonnes :

-	+
Bayrou (30%)	Royal (29%)
Villiers (16%)	

Le troisième axe oppose nettement le vote pour Bayrou, bien représenté dans des régions telles que les Pays de la Loire et l'Alsace (partie négative de l'axe) au vote pour Royal, particulièrement élevé dans la "région" Outremer (partie positive de l'axe).

5.2.4.2 Remarques :

1. Etant donné le poids des suffrages obtenus par le candidat Sarkozy (31% de l'ensemble), on aurait pu s'attendre à ce que cette modalité colonne ait une grande influence dans la détermination du profil moyen et donc que le point représentant le candidat soit très proche de l'origine. On remarque malgré tout que ce point reste bien distinct de l'origine.
2. Il est tout à fait remarquable que l'étude ne montre pas d'opposition entre les votes pour les deux candidats arrivés en tête. Mais, dans des régions telles que l'Outremer ou l'Ile de France, ces candidats obtiennent tous les deux des scores supérieurs à leur moyenne nationale, alors qu'ils obtiennent simultanément des scores inférieurs à leur moyenne nationale dans le Nord Pas de Calais. Et, l'Ile de France et le Nord Pas de Calais sont très importantes numériquement.