Analyse multidimensionnelle des données
Master 2ème année - Psychologie Sociale des Représentations

Réf. (polycopié et fichiers de données utilisés) :
http://geai.univ-brest.fr/~carpenti/

1 Présentation

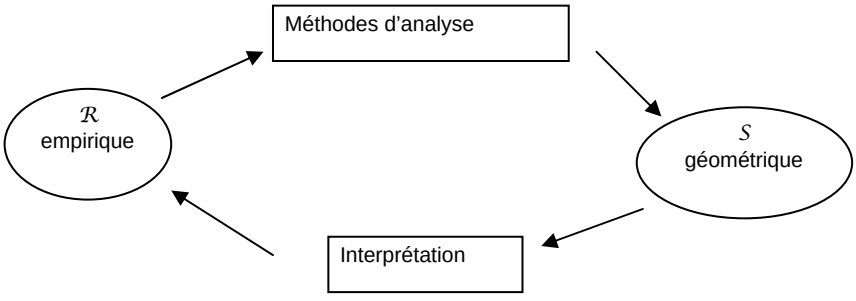
1.1 Introduction

Comment peut-on définir l'analyse multidimensionnelle des données ?

L'analyse statistique élémentaire s'applique à des situations dans lesquelles une ou deux variables ont été observées sur un ensemble d'individus statistiques (populations ou échantillons). L'extension de ces méthodes aux cas où le nombre de variables devient plus élevé est souvent appelé analyse *multivariée*. Cependant les conclusions ou résultats obtenus par ces méthodes restent de même nature, *unidimensionnelle*. Par exemple, la MANOVA (analyse de variance multivariée) permet d'étudier l'effet de facteurs de variation sur un "vecteur" de variables dépendantes, mais apporte une conclusion analogue à celle de l'ANOVA : les facteurs ont (ou n'ont pas) un effet sur le vecteur des VD.

L'analyse multidimensionnelle (ou plutôt, les méthodes qui en relèvent) étudie également des situations où un ensemble de variables doit être étudié simultanément sur un ensemble d'objets statistiques. Par nature, ces données se modélisent dans un espace à plusieurs dimensions. Mais, à la différence des méthodes précédentes, l'analyse multidimensionnelle des données s'attache à fournir des résultats en réduisant le nombre de dimensions, mais en ne se limitant pas à une seule. La plupart des méthodes d'analyse multidimensionnelle utilisent un modèle géométrique (une géométrie dans un espace de dimension supérieure à 3) et ses possibilités de projection sur des sous-espaces de dimension plus réduite, notamment sur des plans bien choisis. Les "écarts" entre objets y sont alors traduits par les distances habituelles.

G. Drouet d'Aubigny schématise ce traitement d'un tableau de données complexes, ou système relationnel empirique de la façon suivante :



Le plus souvent, les méthodes d'analyse multidimensionnelle s'appliquent à des tableaux de l'un des types suivants :

- Tableau protocole individus x variables numériques. Exemple :

On dispose des consommations annuelles de 8 types de denrées alimentaires pour 8 catégories socio-professionnelles (en 1972).

	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

Légende :

Variables :	Observations :
PAO Pain ordinaire	AGRI Exploitants agricoles
PAA Autre pain	SAAG Salariés agricoles
VIO Vin ordinaire	PRIN Professions indépendantes
VIA Autre vin	CSUP Cadres supérieurs
POT Pommes de terre	CMOY Cadres moyens
LEC Légumes secs	EMPL Employés
RAI Raisin de table	OUVR Ouvriers
PLP Plats préparés	INAC Inactifs

- Tableau de contingence. Exemple :

Répartition des étudiants selon la catégorie socio-professionnelle des parents et le type d'études suivi en 1975-1976 (simplifié) :

	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

- Tableau protocole pour des variables nominales

	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 individus x variables comportant des variables numériques et une variable dichotomique

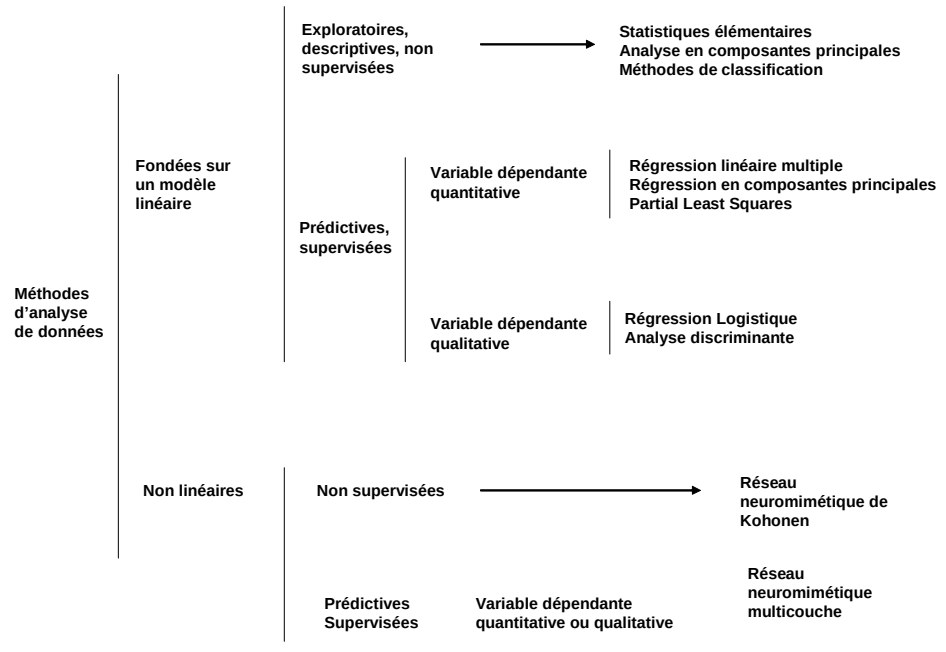
	Age	Etat-Civil	Feministe	Frequence	Agressivite	Harcelement
1	13	1	102	2	4	**0**
2	45	2	101	3	6	**0**
3	19	2	102	2	7	**1**
4	42	2	102	1	2	**1**
5	27	1	77	1	1	**0**
6	19	1	98	0	6	**1**
7	37	1	96	1	6	**0**

On cherche à analyser les résultats contenus dans ces tableaux, en explicitant plusieurs dimensions, si possible indépendantes l'une de l'autre.

1.2 Quelques méthodes utilisées

De nombreuses méthodes ont été proposées. Ces méthodes peuvent être regroupées d'une part selon les outils mathématiques utilisés (méthodes linéaires ou non linéaires), d'autre part selon la nature du résultat recherché (méthodes descriptives ou prédictives).

Méthodes descriptives : toutes les variables jouent des rôles analogues.
Méthodes prédictives : on cherche à "expliquer" ou "prévoir" une ou plusieurs variables (variables dépendantes ou VD) à l'aide des autres variables (variables indépendantes ou VI).



1.3 Concepts fondamentaux

Selon [Doise], toute distribution de réponses sur plusieurs variables peut être statistiquement décomposée en trois éléments : le niveau (la moyenne des réponses des individus), la dispersion (le degré d'éparpillement des réponses individuelles autour de la moyenne), et la corrélation (le lien entre les réponses individuelles pour deux variables). Ces composantes sont autant de points de vue sur les données.

Un tableau de données carré ou rectangulaire est appelé *matrice*. L'élément générique du tableau est désigné par une notation à double indice, par exemple x_{ij} . En général, le premier indice désigne le numéro de ligne, et le second indice le numéro de colonne. Un tableau comportant n lignes et p colonnes est dit *de dimension (n, p).*

Lorsque l'on traite un tableau Individus x Variables de dimension (n, p), les individus peuvent être représentés comme des points d'un espace à p dimensions, les variables comme des points d'un espace à n dimensions. L'ensemble des points représentant les individus est appelé *nuage des individus*.

La distance entre deux individus M_i, M_j est calculée par :

$$M_i M_j^2 = d^2(M_i, M_j) = (x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \ldots + (x_{ip} - x_{jp})^2 = \sum_{k=1}^{p} (x_{ik} - x_{jk})^2$$

L'inertie du nuage de points par rapport à un point donné O de l'espace est la somme des carrés des distances des points Mi à O.

$$I = \sum_{i=1}^{n} OM_i^2$$

L'inertie du nuage de points par rapport au point moyen du nuage est encore appelée somme des carrés ou variation totale.

Le "lien" entre deux variables X_k et X_l peut être mesuré par leur coefficient de corrélation $r(X_k,X_l)$. Lorsque les variables sont centrées et réduites, ce coefficient de corrélation est, à une division par n près, le produit scalaire des vecteurs représentant ces variables. C'est aussi le cosinus de l'angle entre ces deux vecteurs. Pour des variables centrées réduites :

$$r(X_k, X_l) = \frac{1}{n} \sum_{i=1}^{n} x_{ik} x_{il} = \cos\left(\overrightarrow{X_k}, \overrightarrow{X_l}\right)$$

2 Méthodes exploratoires, descriptives

2.1 Analyse en composantes principales ou ACP

2.1.1 Introduction

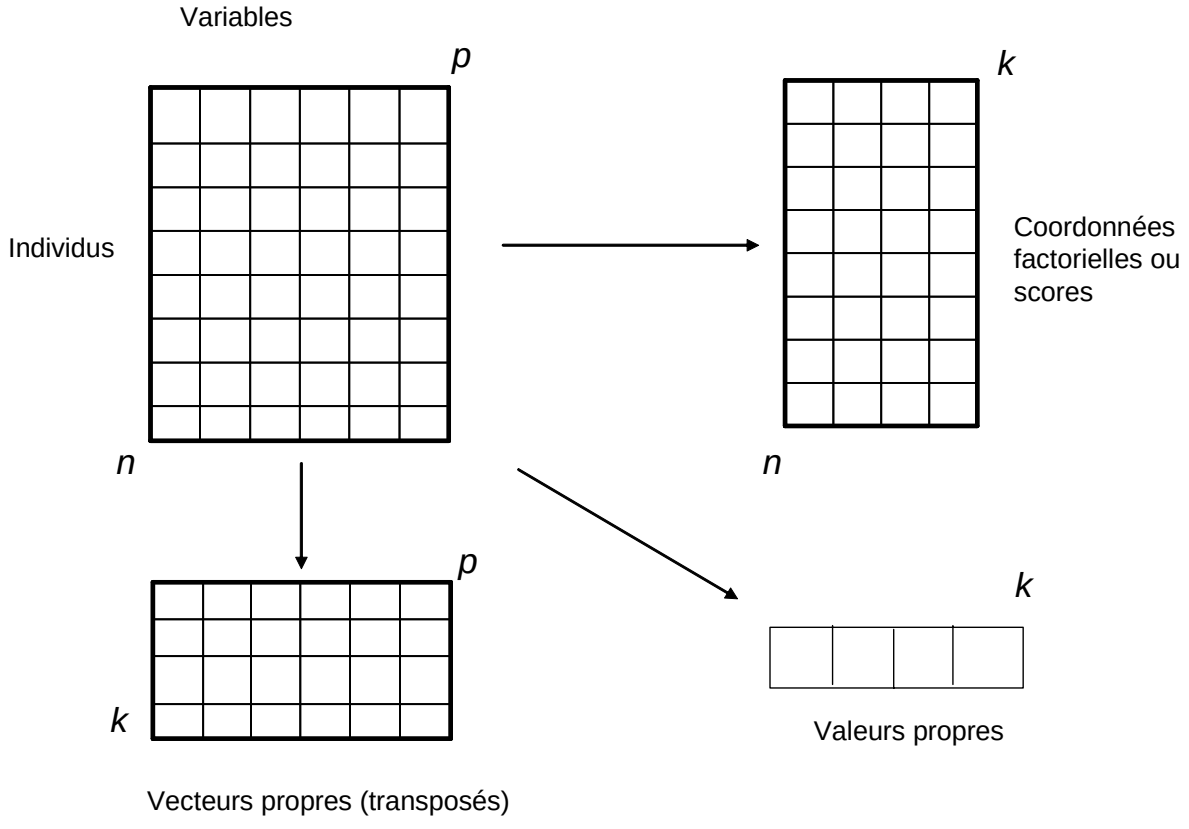
On a observé p variables sur n individus. On dit qu'il s'agit d'un protocole multivarié. Les données à traiter forment une matrice :

$$\begin{array}{cccccc} & X_1 & X_2 & \ldots & X_p \\ i_1 & x_{11} & x_{21} & \ldots & x_{1p} \\ i_2 & x_{21} & x_{22} & \ldots & x_{2p} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ i_n & x_{n1} & x_{n2} & \ldots & x_{np} \end{array}$$

On cherche à remplacer ces p variables par q nouvelles variables (composantes principales ou facteurs) résumant au mieux le protocole, avec $q \leq p$ et si possible q=2.
L'une des solutions à ce problème est l'ACP, méthode qui a l'avantage de résumer un ensemble de variables corrélées en un nombre réduit de facteurs non corrélés. Les principaux résultats d'une ACP sont donnés par :

- Les coordonnées des individus sur les composantes principales ou scores des individus ;
- Les coordonnées des variables sur les composantes principales, ou saturations des variables ; dans le cas d'une ACP normée, les saturations sont aussi les coefficients de corrélation entre les variables initiales et les composantes principales ;
- Les valeurs propres associées à chacune des composantes principales, qui représentent l'inertie du nuage prise en compte par la composante.

Principaux résultats d'une ACP



Principe de la méthode :

- Pour éliminer les effets dus aux choix d'unités des différentes variables, on fait un centrage-réduction des différentes variables.

- Les distances entre les individus sont mesurées par la distance euclidienne dans un espace de dimension p. Par exemple, pour les points représentant les individus 1 et 2 :

$$d^2(M_1, M_2) = (x_{11} - x_{21})^2 + (x_{12} - x_{22})^2 + ... + (x_{1p} - x_{2p})^2$$

- On recherche alors la direction dans laquelle le nuage de points est le plus dispersé : cette direction est le premier axe principal, et l'inertie (dispersion) le long de cet axe est la valeur propre associée à cet axe.

- On projette alors les points dans le sous-espace orthogonal au premier axe principal, et on cherche de nouveau la direction de plus grande dispersion du nuage projeté. On obtient ainsi le deuxième axe principal, et la seconde valeur propre.

- On poursuit la méthode, jusqu'à ce que l'essentiel de l'inertie du nuage de points ait été prise en compte.

2.1.2 Exemple

On reprend l'exemple donné en introduction : consommations annuelles de 8 types de denrées alimentaires pour 8 catégories socio-professionnelles (en 1972).

	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

Légende :

Variables :	Observations :
PAO Pain ordinaire	AGRI Exploitants agricoles
PAA Autre pain	SAAG Salariés agricoles
VIO Vin ordinaire	PRIN Professions indépendantes
VIA Autre vin	CSUP Cadres supérieurs
POT Pommes de terre	CMOY Cadres moyens
LEC Légumes secs	EMPL Employés
RAI Raisin de table	OUVR Ouvriers
PLP Plats préparés	INAC Inactifs

Données après centrage et réduction :

	PAO	PAA	VIO	VIA	POT	LEC	RAI	PLP
AGRI	1,43	-1,22	1,72	-1,15	0,30	0,49	-0,93	-1,50
SAAG	1,25	-0,90	1,16	-1,50	0,17	1,90	-1,38	-0,77
PRIN	-0,29	0,35	-0,70	-0,09	0,05	-0,58	0,65	1,36
CSUP	-1,44	1,92	-0,85	1,66	-1,48	-1,28	1,77	1,19
CMOY	-0,86	0,04	-0,73	0,58	-0,84	-0,93	0,20	0,46
EMPL	-0,58	-0,27	-0,62	0,23	-0,59	-0,22	-0,03	0,30
OUVR	0,10	-0,59	-0,52	-0,22	0,56	0,13	-0,70	-0,68
INAC	0,39	0,67	0,54	0,48	1,83	0,49	0,42	-0,36

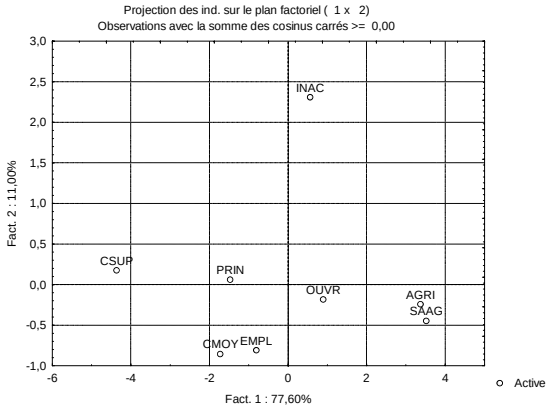
Corrélations entre variables :

	PAO	PAA	VIO	VIA	POT	LEC	RAI	PLP
PAO	1,00	-0,77	0,93	-0,91	0,66	0,89	-0,83	-0,86
PAA	-0,77	1,00	-0,60	0,90	-0,33	-0,67	0,96	0,77
VIO	0,93	-0,60	1,00	-0,75	0,52	0,79	-0,67	-0,83
VIA	-0,91	0,90	-0,75	1,00	-0,42	-0,84	0,92	0,72
POT	0,66	-0,33	0,52	-0,42	1,00	0,60	-0,41	-0,55
LEC	0,89	-0,67	0,79	-0,84	0,60	1,00	-0,82	-0,75
RAI	-0,83	0,96	-0,67	0,92	-0,41	-0,82	1,00	0,83
PLP	-0,86	0,77	-0,83	0,72	-0,55	-0,75	0,83	1,00

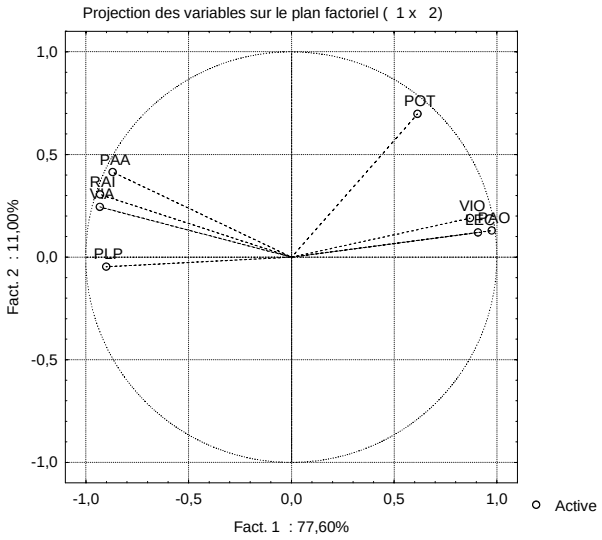
Valeurs propres de l'ACP

	Val Propre	Pourcentage	Cumul Inertie	Cumul %
1	6,2079	77,60	6,21	77,60
2	0,8797	11,00	7,09	88,60
3	0,4160	5,20	7,50	93,79
4	0,3065	3,83	7,81	97,63
5	0,1684	2,11	7,98	99,73
6	0,0181	0,23	8,00	99,96
7	0,0034	0,04	8,00	100,00

Représentation graphique des individus (plan 1-2)



Représentation graphique des variables (plan 1-2)

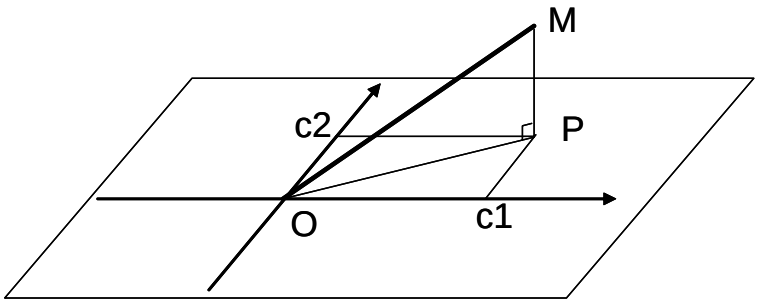


Aides à l'interprétation

Contributions ou inerties relatives des individus

	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
OUVR	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

Qualités de représentation



Cosinus carrés

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

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

$\overrightarrow{OM}$: 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

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

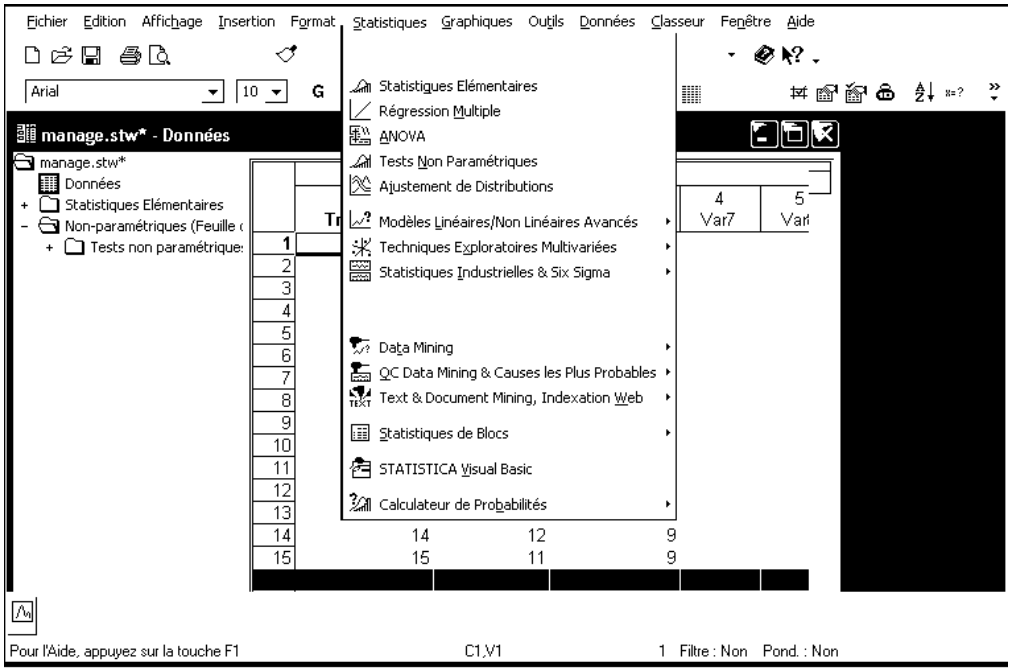
2.1.3 Analyse en composantes principales avec Statistica

2.1.3.1 Présentation de Statistica

. Statistica : l'interface utilisateur

L'écran de travail

Statistica 7.1 est un logiciel dédié aux traitements statistiques. C'est également la "brique" de base des logiciels proposés par Statsoft, et ses possibilités d'interaction avec d'autres logiciels (tableurs, systèmes de gestion de bases de données, traitements de textes, ...) sont nombreuses. En revanche, l'interface utilisateur pourra sembler un peu déconcertante au premier abord.



Les objets manipulés par Statistica

La **feuille de données** est organisée en variables et observations. Les colonnes sont les variables. Chaque ligne représente un individu statistique, appelé observation.

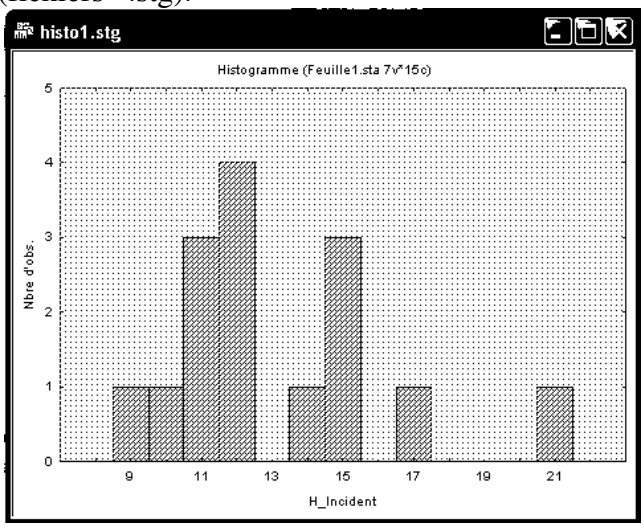
Données : Feuille1.sta (7v par 15c)

	1 Trimestre	2 H_Incident	3 D_Incident	4 Var7
1	1	11	8	
2	2	11	13	
3	3	14	12	
4	4	21	17	
5	5	12	14	
6	6	10	9	
7	7	15	10	
8	8	15	12	
9	9	17	13	
10	10	9	10	
11	11	12	8	
12	12	12	13	
13	13	15	12	
14	14	12	9	
15	15	11	9	

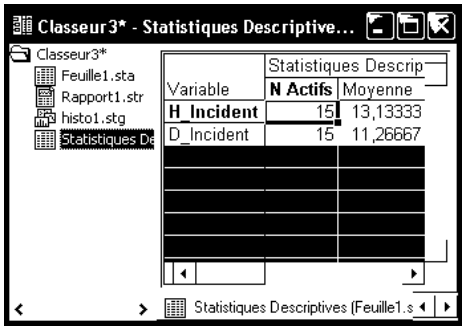
Les feuilles de données peuvent être enregistrées comme fichiers autonomes (fichiers *.sta). Elles contiennent les données d'entrée sur lesquelles s'effectuent les traitements statistiques. Les résultats de ces traitements s'affichent dans un document de sortie. Plusieurs possibilités sont offertes.

Fenêtre de rapport : C'est la méthode traditionnelle pour gérer les résultats produits par le logiciel. Un rapport se comporte plus ou moins comme un document produit par un traitement de textes. On peut insérer des commentaires, modifier la mise en forme, spécifier la mise en page, la numérotation des pages, l'en-tête et le pied de page en vue de l'impression. Les rapports peuvent être enregistrés comme fichiers autonomes (fichiers *.str).

Les résultats de sortie peuvent également être dirigés vers des fenêtres individuelles. Les résultats numériques sont alors affichés dans des fenêtres de données. Les graphiques sont affichés dans des **fenêtres de graphiques** (fichiers *.stg).



Les classeurs : les données d'entrée et de sortie peuvent également être stockées comme onglets dans un classeur. Un classeur est un "container" accueillant d'autres objets, organisés sous forme hiérarchique. Ils correspondent aux fichiers de type *.stw.



Traitements statistiques

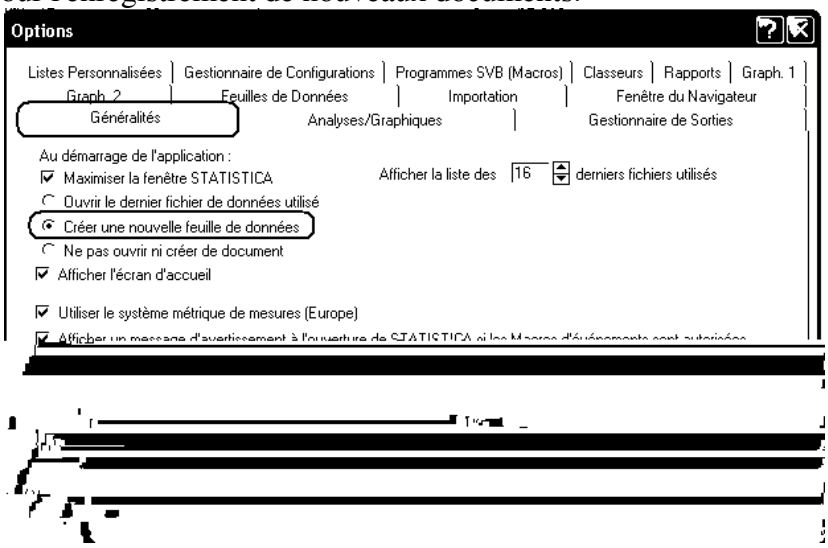
Statistica est organisé en modules, accessibles à partir du menu Statistiques. Chaque module contient un groupe de procédures statistiques reliées entre elles. Par exemple, le module "Statistiques élémentaires" se présente comme suit :

Gérer les sorties

Modifier le comportement de Statistica

Le comportement de Statistica peut être modifié en intervenant dans la fenêtre de dialogue affichée par le menu Outils - Options.
Par exemple, nous souhaitons :
- que Statistica n'ouvre plus systématiquement la dernière feuille de données utilisée lors du chargement du logiciel ;
- que Statistica nous propose par défaut le volume U: pour enregistrer nos documents, au lieu du répertoire "Mes Documents".

Exécutez le menu Outils - Options. Sous l'onglet Généralités, activez le bouton radio "Créer une nouvelle feuille de données".
Désactivez la boîte à cocher "mémoriser les répertoires pour l'ouverture ou la sauvegarde des fichiers". Complétez la zone d'édition "Répertoire par défaut" en indiquant U:\, puis réactivez la boîte à cocher (N.B. Bien que l'option soit en apparence désactivée, Statistica proposera par défaut le répertoire U:\ pour l'enregistrement de nouveaux documents.



Gérer les sorties

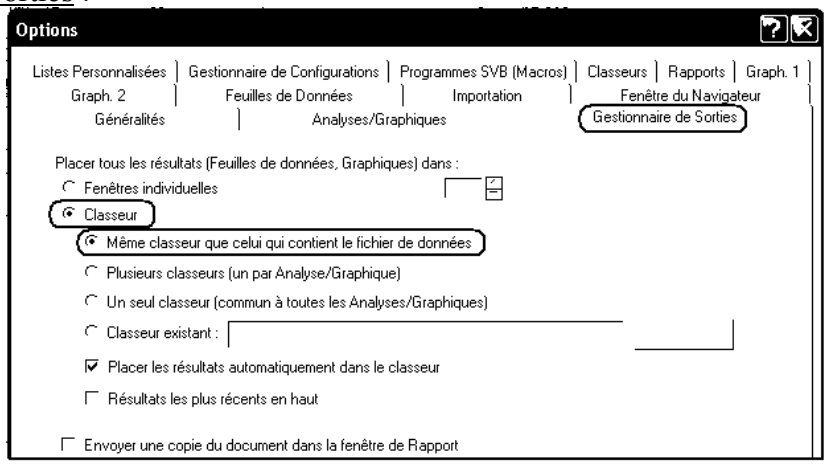
Lorsqu'on utilise Statistica sans se préoccuper des options de sortie des résultats, on se retrouve vite à la tête d'une quantité de fenêtres (classeurs, feuilles de données de résultats, fenêtres de graphiques...). Pour réaliser un travail que l'on souhaite conserver et reprendre au cours de plusieurs

séances de travail, il paraît indispensable d'organiser correctement son espace de travail et ses sauvegardes.

Enregistrer données et résultats dans un seul classeur

Cette méthode consiste à enregistrer les données, les résultats de traitements, et les commentaires éventuels comme objets d'un même classeur. Ainsi, un unique fichier du disque rassemble l'ensemble de notre travail sur un cas donné.

Ce comportement correspond aux réglages suivants dans le menu Outils - Options - Onglet Gestionnaire de Sorties :



Remarque : Le réglage ne sera actif que si la feuille de données se trouve effectivement dans un classeur. Or, ce ne sera pas le cas si la feuille de données a été ouverte à partir d'un fichier *.sta, ou importée à partir d'une feuille Excel. Dans ce cas, vous devez insérer la feuille de données dans le classeur comme il a été indiqué au paragraphe précédent.

Indiquer quelle est la feuille de données active

Lors des premières manipulations avec Statistica, nous n'avons pas eu besoin de nous préoccuper de la notion de "feuille de données active", les choix par défaut faits par Statistica nous convenant parfaitement. Cependant, cette notion permet de résoudre plusieurs problèmes :

- Ouvrir plusieurs fichiers .sta et effectuer un travail sur l'un d'eux (pas nécessairement le dernier ouvert)

- Utiliser une feuille de résultats comme feuille de données pour des traitements ultérieurs.

- Lorsque l'on travaille avec une feuille de données insérée dans un classeur, il arrive couramment que Statistica ne retrouve pas la feuille à partir de laquelle les traitements doivent être effectués. Mais on peut éviter ce comportement en spécifiant la propriété "feuille de données active" pour l'objet du classeur qui contient nos données.

Pour spécifier comme feuille de données active une feuille d'un classeur :

- Cliquez avec le bouton droit de la souris sur l'icône de la feuille de données dans le volet gauche du classeur.

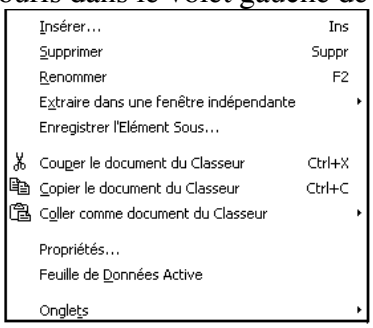
- Utilisez l'item Feuille de données active du menu local.

On peut également utiliser le menu Données - Feuille de données active.
Remarquez que le volet gauche d'un classeur indique si une feuille insérée dans le classeur est active ou non : l'icône d'un feuille active est encadrée en rouge :

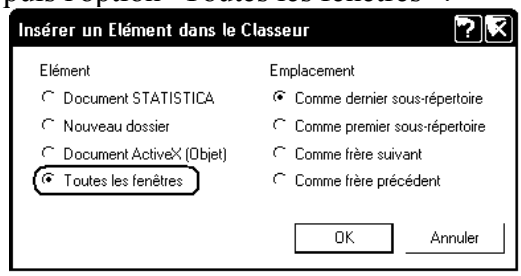
Enregistrer les données et l'ensemble des traitements réalisés dans un même classeur

Ouvrez un fichier de données (un fichier d'extension .sta) et réalisez un ou plusieurs traitements relatifs à ces données (par exemple, des statistiques descriptives et un graphique). Si vous avez gardé les options par défaut de Statistica, les résultats de tous ces traitements se trouvent dans un classeur.

Pour enregistrer données, traitements et rapport dans un seul classeur :
Affichez la fenêtre du classeur contenant les résultats.
Cliquez avec le bouton droit de la souris dans le volet gauche de la fenêtre du classeur.



Sélectionnez l'item Insérer..., puis l'option "Toutes les fenêtres" :



N'oubliez pas, ensuite, de spécifier la feuille contenant les données de base comme feuille active du classeur.

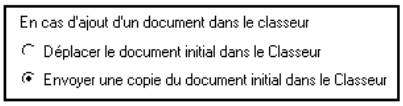
Manipuler les objets contenus dans un classeur

Copier - coller entre classeurs, entre un classeur et un objet Statistica

Pour déplacer un objet d'un classeur à un autre, il suffit de déplacer son icône depuis le volet gauche du premier classeur dans le volet gauche du second. On peut également utiliser les menus locaux Copier et Coller obtenus à l'aide d'un clic droit dans le volet gauche de chaque classeur.

Le menu local "Insérer" du volet gauche d'un classeur permet également d'insérer dans ce classeur un document contenu dans une fenêtre indépendante. Il suffit de choisir les options : Document Statistica - Créer à partir d'une fenêtre.

L'opération faite par Statistica est soit une copie (l'original de l'objet est conservé) soit un déplacement (l'original de l'objet n'est pas conservé) selon le paramétrage choisi dans le menu Outils - Options - Onglet Classeurs - Item "En cas d'ajout d'un document dans le classeur".



Supprimer un objet d'un classeur

Il est également possible de supprimer un objet d'un classeur, à l'aide d'un clic droit et de l'item de menu Supprimer. Cela permet notamment de ne garder, pour un traitement donné, que le résultat le plus abouti. Attention cependant : lorsque l'on supprime un objet qui n'est pas une feuille de la hiérarchie, on supprime en même temps tous les objets qui en dépendent.

2.1.3.2 Présentation de l'exemple

Source de l'exemple : Claude FLAMENT, Laurent MILLAND, Un effet Guttman en ACP, Mathématiques & Sciences humaines (43e année, n° 171, 2005, p. 25-49)

Cet exemple a trait à la représentation sociale de l'homosexualité. Le questionnaire, composé d'une liste de 31 traits plus ou moins sexués, a été administré à 70 hommes homosexuels et à 70 hommes hétérosexuels [Rallier, Ricou, 2000]. Tous les sujets devaient, dans un premier temps, se décrire à partir de cette liste de traits, en se positionnant à chaque fois sur une échelle allant de 1 (= négatif) à 7 (= positif). Après avoir réalisé cette auto-description, les sujets devaient répondre à ce même questionnaire « comme le feraient les X en général », la cible « X » pouvant être : les hommes, les femmes, ou les homosexuels. Nous disposons ainsi de 8 profils moyens, qui se définissent à partir de la combinaison entre les caractéristiques des répondants et les consignes données pour remplir les questionnaires. Nous travaillons ici sur un extrait des données complètes (15 traits), extrait qui respecte scrupuleusement le type de résultat obtenu sur l'ensemble des 31 traits de l'étude.
Pour faciliter le repérage des consignes, nous avons fait le choix de coder les 8 profils en repérant en premier les répondants, puis le type de consigne parmi les 4 possibles :

Ho : Soi = sujets Homosexuels répondant à la consigne d'auto-description Soi ;
Hé : Soi = sujets Hétérosexuels répondant à la consigne d'autodescription Soi ;
Ho : H = sujets Homosexuels répondant comme le feraient les Hommes ;
Hé : H = sujets Hétérosexuels répondant comme le feraient les Hommes;
Ho : F = sujets Homosexuels répondant comme le feraient les Femmes ;
Hé : F = sujets Hétérosexuels répondant comme le feraient les Femmes ;
Ho : Ho = sujets Homosexuels répondant comme le feraient les Homosexuels ;
Hé : Ho = sujets Hétérosexuels répondant comme le feraient les Homosexuels.

Nous partons ici d'un tableau de données comprenant, pour chacune des 8 conditions expérimentales, les moyennes de chaque trait calculées sur les 70 réponses obtenues dans chacune des conditions expérimentales. On retrouve, dans le tableau ci-dessous, le rang (solidarisation des variables) de chacun des 15 traits dans les 8 profils

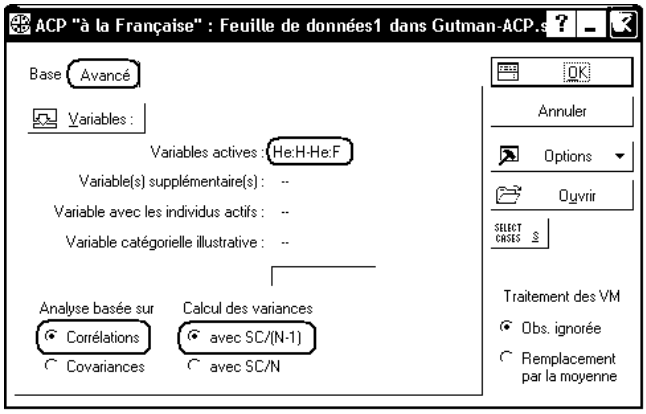
	He:H	Ho:H	He:Soi	Ho:Soi	Ho:Ho	Ho:F	He:Ho	He:F
Est meneur	5	6	12	13	15	13	14	13
Aime competition	3	3	13	14	11	14	13	14
FEMININ	15	15	15	15	13	2	4	1
A confiance en soi	4	8	6	11	14	12	12	12
Devoue	11	12	10	7	10	11	8	7
MASCULIN	1	1	1	12	12	15	15	15
Bienveillant	10	10	9	9	7	9	7	6
Attentif aux besoins des autres	12	13	11	4	9	8	5	5
Energique	8	4	5	8	6	10	11	11
Ambitieux	6	7	3	10	8	7	10	10

Sensible	14	14	14	2	1	1	1	2
Agreable	9	9	7	5	3	6	6	3
Affectueux	13	11	8	1	4	5	2	4
A du caractere	2	5	4	6	5	4	9	8
Defend ses opinions	7	2	2	3	2	3	3	9

Remarque. A l'examen du tableau précédent, on constate que les rangs ont été déterminés à l'inverse de ce qui est généralement fait en statistiques : les rangs élevés correspondent aux traits les moins typiques du stéréotype considéré, tandis que les rangs faibles correspondent aux traits les plus typiques. Cette remarque est importante pour l'interprétation des résultats de l'ACP.

Ouvrez le classeur Statistica Rep-Soc-Homo.stw.

Pour effectuer l'ACP, nous utilisons le menu Statistiques - Techniques exploratoires multivariées - ACP "à la française".



La fenêtre de dialogue permet de spécifier les variables qui participeront à l'analyse. Elle permet également d'indiquer les différentes options choisies pour le traitement.
Utilisez l'onglet "Avancé" de cette fenêtre.

- Comment seront traitées les valeurs manquantes ? Nous voyons que Statistica propose soit de neutraliser la ligne correspondante, soit de remplacer la valeur manquante par la moyenne observée sur la variable.

- L'analyse sera-t-elle basée sur les covariances ou sur les corrélations ?

- Utilise-t-on les variances et covariances non corrigées (SC/N) ou les variances et covariances corrigées (SC/(N-1)). Dans le cas d'une ACP normée, les deux méthodes fournissent des résultats presque identiques : seuls les scores des individus sont légèrement modifiés. En fait, l'ACP est une méthode descriptive et non une méthode inférentielle. Elle est effectuée dans un but exploratoire : on étudie les données pour elles-mêmes, et non en vue d'une généralisation à une population. C'est pourquoi l'utilisation des variances non corrigées est généralement justifiée.

Nous ferons ici une analyse basée sur les corrélations, en utilisant les variances et covariances corrigées (SC/(N-1)), de manière à retrouver les résultats publiés. Cliquez ensuite sur le bouton OK.

N.B. Ne fermez pas l'analyse en cours pendant la suite des manipulations. Ainsi, vous n'aurez pas à indiquer de nouveau les options ci-dessus, vos résultats seront cohérents entre eux et se rassembleront dans un même classeur.

2.1.3.3 Statistiques descriptives - Matrice des corrélations

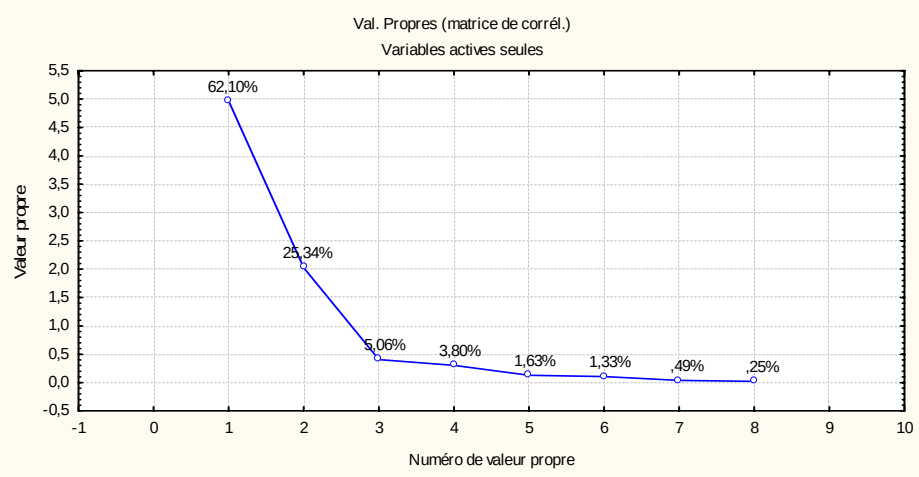
Ces résultats peuvent être obtenus à l'aide de l'onglet "Descriptives".

Variable	Corrélations (Repr-Soc-Homo dans Rep-Soc-Homo.stw)							
	He:H	Ho:H	He:Soi	Ho:Soi	Ho:Ho	Ho:F	He:Ho	He:F
He:H	1,0000	0,8679	0,5857	-0,4071	-0,3143	-0,6036	-0,8179	-0,8714
Ho:H	0,8679	1,0000	0,6786	-0,2393	-0,0607	-0,4821	-0,6429	-0,8357
He:Soi	0,5857	0,6786	1,0000	0,1679	0,2179	-0,1321	-0,2821	-0,4464
Ho:Soi	-0,4071	-0,2393	0,1679	1,0000	0,8429	0,5607	0,7143	0,5036
Ho:Ho	-0,3143	-0,0607	0,2179	0,8429	1,0000	0,6750	0,6821	0,4929
Ho:F	-0,6036	-0,4821	-0,1321	0,5607	0,6750	1,0000	0,8714	0,8071
He:Ho	-0,8179	-0,6429	-0,2821	0,7143	0,6821	0,8714	1,0000	0,8857
He:F	-0,8714	-0,8357	-0,4464	0,5036	0,4929	0,8071	0,8857	1,0000

2.1.3.4 Choix des valeurs propres

Affichez d'abord le tableau des valeurs propres et le diagramme correspondant.
Pour cela, cliquez sur les boutons "Valeurs propres" et "Tracé des valeurs propres" de l'onglet "Base".

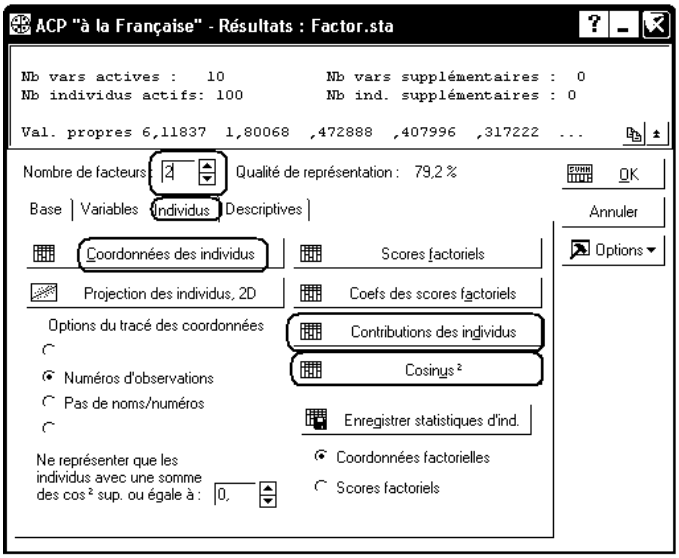
Valeur numéro	Val. Propres (matrice de corrél.) & stat. associées Variables actives seules			
	Val. propr	% Total variance	Cumul Val. propr	Cumul %
1	4,9682	62,1026	4,9682	62,10
2	2,0268	25,3355	6,9950	87,44
3	0,4045	5,0562	7,3995	92,49
4	0,3038	3,7979	7,7034	96,29
5	0,1308	1,6346	7,8341	97,93
6	0,1064	1,3301	7,9405	99,26
7	0,0391	0,4892	7,9797	99,75
8	0,0203	0,2541	8,0000	100,00



Dans notre cas, on peut choisir de retenir 2 composantes principales. Dans les manipulations qui suivent, on indiquera donc 2 dans la zone d'édition "nombre de facteurs".
Pour les résultats relatifs aux individus et aux variables, on utilisera de préférence les onglets correspondants.

2.1.3.5 Résultats relatifs aux individus

On pourra obtenir successivement les scores des individus, leurs contributions à la formation des composantes principales et leurs qualités de représentation en utilisant les boutons "Coordonnées des individus", "Contributions des individus", "Cosinus²".



Individus	Coordonnées factorielles des ind	
	Fact. 1	Fact. 2
Est meneur	-2,5273	1,6292
Aime competition	-2,7956	1,0317
FEMININ	2,2340	2,9293
A confiance en soi	-2,1315	0,6348
Devoue	0,4389	1,0207
MASCULIN	-3,9200	-0,7793
Bienveillant	0,5732	0,4503
Attentif aux besoins	1,6498	0,7581
Energique	-1,0549	-0,8752
Ambitieux	-0,8932	-0,6719
Sensible	3,9415	-0,0333
Agreable	1,4885	-0,8338
Affectueux	2,6229	-0,7360
A du caractere	-0,3598	-1,7357
Defend ses opinions	0,7335	-2,7890

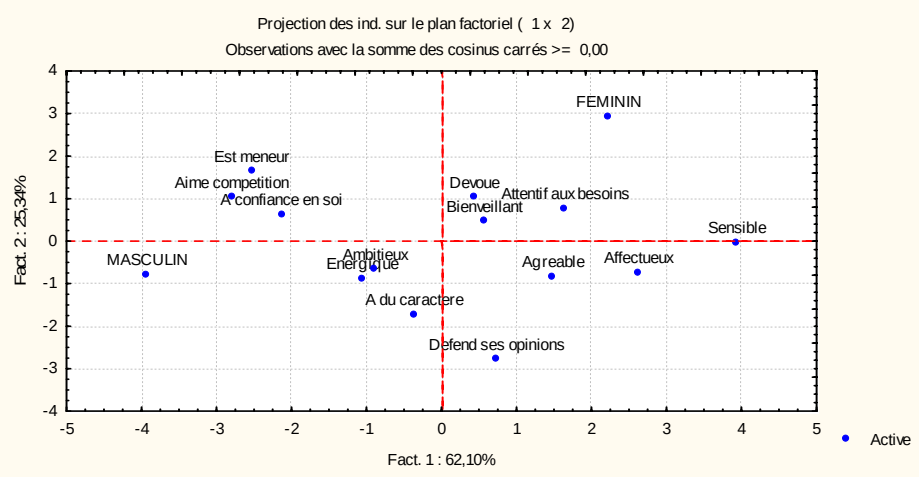
Individus	Contributions des ind	
	Fact. 1	Fact. 2
Est meneur	9,18	9,35
Aime competition	11,24	3,75
FEMININ	7,18	30,24
A confiance en soi	6,53	1,42
Devoue	0,28	3,67
MASCULIN	22,09	2,14
Bienveillant	0,47	0,71
Attentif aux besoins	3,91	2,03
Energique	1,60	2,70
Ambitieux	1,15	1,59
Sensible	22,34	0,00
Agreable	3,19	2,45
Affectueux	9,89	1,91
A du caractere	0,19	10,62
Defend ses opinions	0,77	27,41

Individus	Cosinus carrés,		
	Fact. 1	Fact. 2	Fact. 1 & 2 =v1+v2
Est meneur	0,6759	0,2809	0,9568
Aime competition	0,7203	0,0981	0,8184
FEMININ	0,3100	0,5330	0,8429
A confiance en soi	0,8041	0,0713	0,8755
Devoue	0,0875	0,4736	0,5611
MASCULIN	0,9427	0,0373	0,9800
Bienveillant	0,3866	0,2385	0,6251
Attentif aux besoins	0,6404	0,1352	0,7757
Energique	0,4364	0,3004	0,7368
Ambitieux	0,3711	0,2099	0,5810
Sensible	0,9502	0,0001	0,9503
Agreable	0,6330	0,1986	0,8317
Affectueux	0,8600	0,0677	0,9277
A du caractere	0,0284	0,6621	0,6906
Defend ses opinions	0,0582	0,8409	0,8991

Remarquez que les résultats ainsi obtenus sont présentés dans des feuilles de résultats sur lesquelles il est possible d'effectuer les mêmes transformations (tris, ajout ou suppression de colonne, etc) que sur les feuilles contenant les données de base. Ainsi, une colonne supplémentaire a été ajouté au tableau des cosinus-carrés pour indiquer la qualité de représentation des individus dans le premier plan factoriel.

On peut ensuite obtenir les projections du nuage des individus selon les premiers axes factoriels à l'aide du bouton "Projection de individus, 2D". Lorsque les individus ne sont pas anonymes (ce qui est le cas ici), il est utile d'étiqueter chaque point. Plusieurs méthodes sont possibles :

- Utiliser les identifiants d'individus figurant dans la première colonne du tableau de données
- Utiliser les numéros des observations
- Utiliser les étiquettes indiquées dans la variable "illustrative" : ces étiquettes peuvent être des identifiants des individus, mais peuvent également représenter un groupe d'appartenance, etc.



Dans certains cas, il pourra être utile de modifier les échelles sur les axes de manière à obtenir une représentation en axes orthonormés. L'importance de la part d'inertie expliquée par le premier axe principal apparaît ainsi plus clairement.

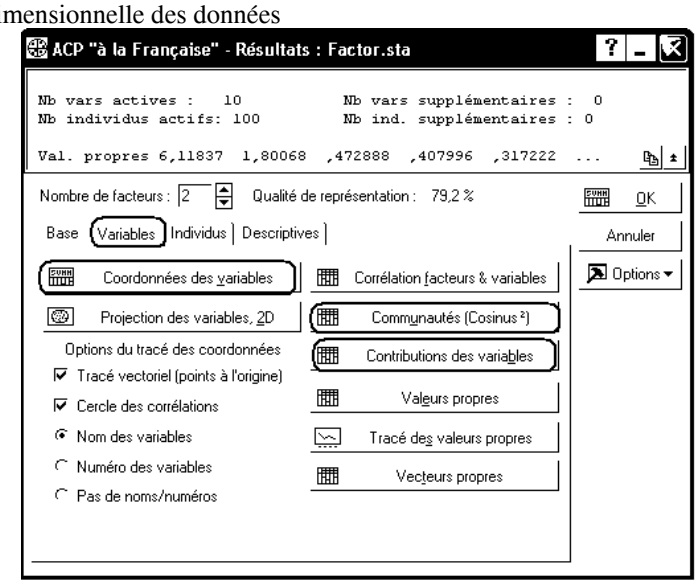
2.1.3.6 Résultats relatifs aux variables

Activons ensuite l'onglet "Variables".

On obtient les saturations des variables en cliquant sur le bouton "Coordonnées des variables" ou le bouton "Corrélation facteurs et variables" : dans le cas d'une ACP normée, ces deux traitements fournissent le même résultat.

On obtient leurs contributions à la formation des composantes principales en utilisant le bouton "Contributions des variables".

Les qualités de représentation sont calculées, de façon cumulative (qualité de la projection selon F1, puis selon le plan (F1,F2), puis selon l'espace (F1,F2,F3) en utilisant le bouton "Communautés (Cosinus²)".



Saturations des variables

Variable	Coord. factorielles des var	
	Fact. 1	Fact. 2
He:H	0,8863	0,3388
Ho:H	0,7743	0,5518
He:Soi	0,4047	0,8013
Ho:Soi	-0,6701	0,6053
Ho:Ho	-0,6317	0,7093
Ho:F	-0,8511	0,2387
He:Ho	-0,9663	0,1361
He:F	-0,9555	-0,1428

Contributions des variables

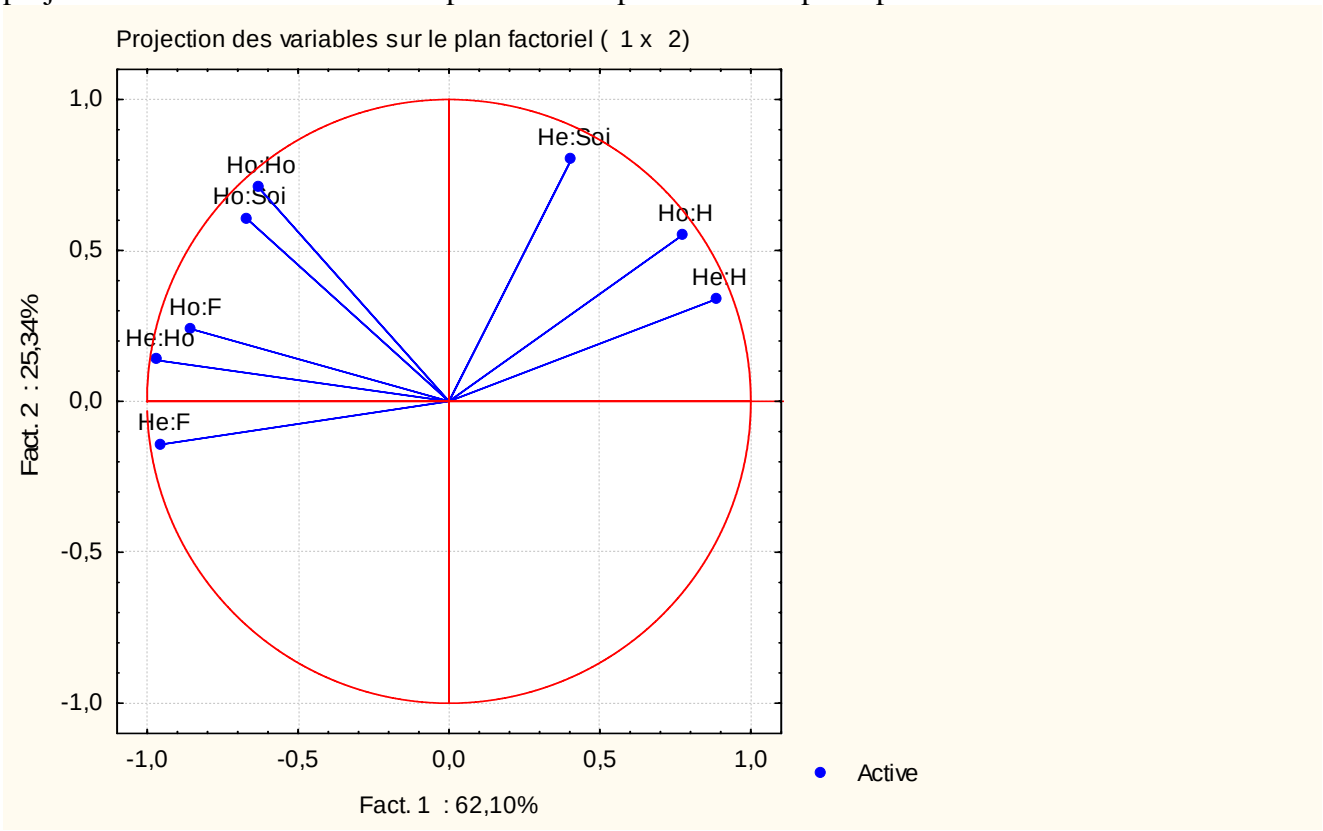
Variable	Contributions des var	
	Fact. 1	Fact. 2
He:H	0,1581	0,0566
Ho:H	0,1207	0,1502
He:Soi	0,0330	0,3168
Ho:Soi	0,0904	0,1808
Ho:Ho	0,0803	0,2482
Ho:F	0,1458	0,0281
He:Ho	0,1879	0,0091
He:F	0,1838	0,0101

Qualités des représentations des variables

Variable	Communautés,	
	Avec 1 facteur	Avec 2 facteurs
He:H	0,7856	0,9004
Ho:H	0,5996	0,9041
He:Soi	0,1638	0,8060
Ho:Soi	0,4491	0,8154
Ho:Ho	0,3991	0,9022
Ho:F	0,7243	0,7813
He:Ho	0,9337	0,9522
He:F	0,9131	0,9334

Représentation des variables

Le bouton "Projection des variables, 2D" permet d'obtenir les diagrammes représentant les projections des variables selon les plans définis par deux axes principaux.



On peut remarquer que toutes les variables se projettent dans un même demi-plan du premier plan factoriel. Autrement dit, une rotation des axes factoriels convenablement choisie permettrait de ramener toutes les variables dans le demi-plan correspondant aux valeurs positives du premier facteur.

2.1.3.7 Coefficients des variables

Les coefficients des variables (c'est-à-dire la matrice permettant de passer des variables centrées réduites aux composantes principales et vice-versa) sont obtenus à l'aide du bouton "Vecteurs propres" de l'onglet "Variables".

Variable	Vecteurs propres de la matrice de corrélation (Repr-Soc-Homo dans Rep-Soc-Homo.stw) Variables actives seules							
	Fact. 1	Fact. 2	Fact. 3	Fact. 4	Fact. 5	Fact. 6	Fact. 7	Fact. 8
He:H	0,398	0,238	0,172	-0,198	0,731	-0,039	-0,325	-0,272
Ho:H	0,347	0,388	0,135	-0,416	-0,440	-0,137	-0,329	0,466
He:Soi	0,182	0,563	0,217	0,750	-0,149	0,097	0,022	-0,092
Ho:Soi	-0,301	0,425	-0,617	0,082	0,318	-0,345	0,036	0,345
Ho:Ho	-0,283	0,498	-0,111	-0,411	-0,090	0,589	0,198	-0,309
Ho:F	-0,382	0,168	0,687	-0,142	0,196	-0,294	0,408	0,206
He:Ho	-0,434	0,096	0,065	-0,030	-0,261	-0,450	-0,496	-0,531
He:F	-0,429	-0,100	0,189	0,168	0,184	0,463	-0,577	0,401

2.1.4 Interprétation des résultats de l'ACP

2.1.4.1 Examen des valeurs propres. Choix du nombre d'axes

On examine les résultats relatifs aux valeurs propres.
Plusieurs critères peuvent nous guider :

- "méthode du coude" on examine la courbe de décroissance des valeurs propres pour déterminer les points où la pente diminue de façon brutale ; seuls les axes qui précèdent ce changement de pente seront retenus.
- si l'analyse porte sur p variables et n > p individus, la variation totale est répartie sur p axes. On peut alors choisir de conserver les axes dont la contribution relative est supérieure à $\frac{100\%}{p}$. Dans le cas d'une ACP normée, cela revient à conserver les axes correspondant aux valeurs propres supérieures à 1.

Sur le cas étudié, les différentes méthodes conduisent à ne garder que les deux premiers axes.

2.1.4.2 Interpréter les résultats relatifs aux individus

Très souvent, les individus pris en compte pour une ACP sont en nombre très élevé et sont considérés comme anonymes. Les éléments qui suivent concernent évidemment les cas où ils ne le sont pas.

Contributions des individus à la formation d'un axe

On relève, pour chaque axe, quels sont les individus qui ont la plus forte contribution à la formation de l'axe. Par exemple, on retient (pour l'analyse) les individus dont la contribution relative est supérieure à $\frac{100\%}{n}$. On note également si cette contribution intervient dans la partie positive ou dans la partie négative de l'axe.

On peut ainsi caractériser l'axe en termes d'opposition entre individus. Il peut également être intéressant d'étudier comment l'axe classe les individus.

Si un individu a une contribution très forte à la formation d'un axe, on peut choisir de recommencer l'analyse en retirant cet individu, puis de l'introduire en tant qu'individu supplémentaire.

Ainsi, pour le premier axe, on relève les traits qui ont contribué pour plus de 6,67% à sa formation et le signe de la coordonnée de chacun de ces traits. On obtient :

-	+
MASCULIN (22,09) Aime competition (11,24) Est meneur (9,18)	Sensible (22,34) Affectueux (9,89) FEMININ (7,18)

On voit que cet axe oppose le trait "masculin", et des traits qui sont souvent associés à ce sexe (meneur, aime compétition, a confiance en soi), sur la partie négative de l'axe, à des traits tels que "sensible", "affectueux", "attentif", et "féminin" sur la partie positive.

Pour le deuxième axe, la même démarche conduit au tableau suivant :

-	+
Defend ses opinions (27,41) A du caractere (10,62)	FEMININ (30,24) Est meneur (9,35)

Cet axe oppose deux traits pratiquement indépendants du premier axe (partie négative de l'axe) au trait "féminin" (partie positive de l'axe).

Projections des individus dans un plan factoriel

Même s'il s'agit du plan (F1, F2), les proximités entre individus doivent être interprétées avec prudence : deux points proches l'un de l'autre sur le graphique peuvent correspondrent à des individus éloignés l'un de l'autre. Pour interpréter ces proximités, il est nécessaire de tenir compte des qualités de représentation des individus.
Se méfier également des individus proches de l'origine : mal représentés, ou proches de la moyenne, ils ont, de toutes façons, peu contribué à la formation des axes étudiés.

2.1.4.3 Interpréter les résultats relatifs aux variables

Contributions des variables

L'examen du tableau des contributions des variables peut permettre d'identifier des variables qui ont un rôle dominant dans la formation d'un axe factoriel. Comme précédemment, on retient (par exemple) les variables dont la contribution relative est supérieure à $\frac{100\%}{p}$. On note également si cette contribution intervient dans la partie positive ou dans la partie négative de l'axe.

Ainsi, pour le premier axe, en fixant la "limite" à 12,5%, on obtient :

-	+
He:Ho (0,1879) He:F (0,1838) Ho:F (0,1458)	He:H (0,1581)

Ainsi, cet axe oppose les profils féminins et homosexuels vus par les hétérosexuels (partie négative de l'axe) au profil masculin vu par les hétérosexuels (partie positive de l'axe).

Remarque importante. L'analyse des individus (traits) avait associé la partie négative du premier axe aux traits masculins. L'analyse des variables semble a priori conduire à un résultat opposé. Mais la contradiction n'est qu'apparente : ici, le protocole des rangs accorde le rang le moins élevé au trait le plus caractéristique du profil. La variable He:H par exemple, est fortement corrélée positivement avec le facteur 1. Le trait "masculin" par exemple obtient un score faible aussi bien sur cette variable (rang 1) que sur le premier facteur (-3,92, minimum des coordonnées de points).

Pour le second axe factoriel, on obtient :

-	+
	He:Soi (0,3168) Ho:Ho (0,2482) Ho:Soi (0,1808) Ho:H (0,1502)

On remarque que les quatre variables retenues sont celles qui ne figuraient pas dans le tableau précédent. Ces quatre variables sont corrélées positivement avec le deuxième axe.

Analyse des projections des variables sur les plans factoriels

Les diagrammes représentant les projections des variables sur les axes factoriels nous fournissent plusieurs types d'informations :

- La longueur du vecteur représentant la variable est liée à la qualité de la représentation de la variable par sa projection dans ce plan factoriel

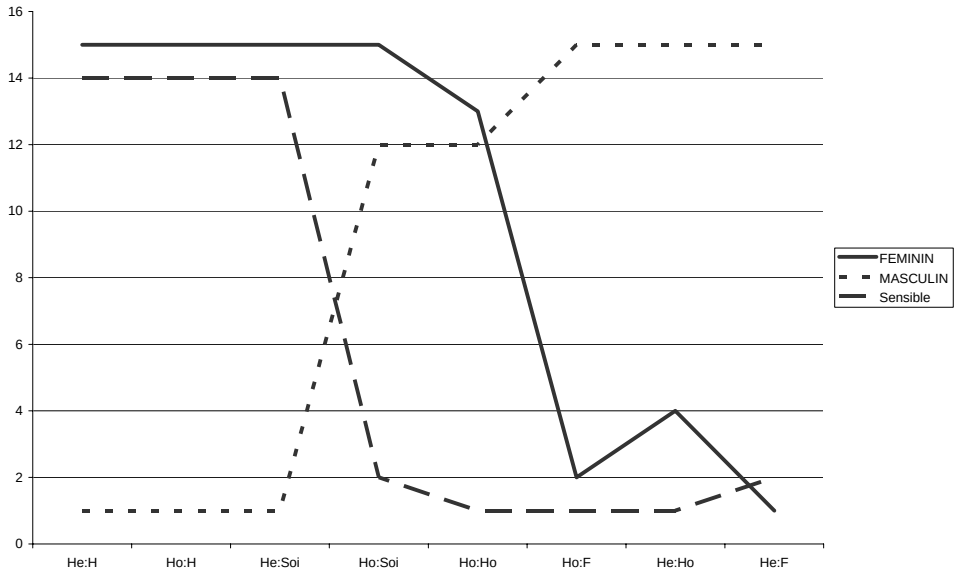
- Pour les variables bien représentées, l'angle entre deux variables est lié au coefficient de corrélation entre ces variables (si la représentation est exacte, le coefficient de corrélation est le cosinus de cet angle). Ceci permet de dégager des "groupes de variables" de significations voisines, des groupes de variables qui "s'opposent", des groupes de variables relativement indépendantes entre eux.

- De même, pour les variables bien représentées, l'angle que fait la projection de la variable avec un axe factoriel est lié au coefficient de corrélation de cette variable et de l'axe factoriel.

Ainsi, dans notre exemple, toutes les variables sont bien représentées dans le premier plan factoriel. Des variables telles que Ho:Soi et Ho:Ho par exemple, sont fortement corrélées positivement entre elles, alors que Ho:Ho et Ho:H sont pratiquement non corrélées. Les variables He:Ho et He:F par exemple, sont fortement anti-corrélées (corrélées négativement) avec le premier axe.

Synthèse des résultats obtenus

On voit que les sujets hétérosexuels ont tendance à estimer que les homosexuels se décrivent comme "féminin" plutôt que "masculin". L'étude des résultats de l'ACP pourrait nous conduire à associer la description que les homosexuels se font d'eux-mêmes à "féminin". Mais, cette conclusion est contredite par les données : les homosexuels ne se voient jamais comme "féminin", mais font appel à des items identifiés ici comme des caractéristiques féminines (sensible, affectueux, etc). Le graphique suivant, dans lequel on a représenté les scores des traits "féminin", "masculin" et "sensible" en fonction des profils convenablement ordonnés, le met en évidence :



Sur ce graphique, les profils sont ordonnés en fonction de leur ordre d'apparition sur le cercle des corrélations (graphique du paragraphe 2.1.3.6). Cet ordre peut également être schématisé de la manière suivante :

Répondants		
He	Ho	Cible
H Soi	H	Masculine
	Soi Ho	Homosexuelle
Ho Fe	Fe	Féminine

2.1.5 ACP avec individus et variables supplémentaires

Lorsqu'on réalise une ACP, il est possible de déclarer certains individus "inactifs" et/ou certaines variables "supplémentaires". Les données correspondantes n'interviennent plus dans le calcul de détermination des composantes principales. En revanche, on leur applique les mêmes transformations qu'aux autres données afin de les ré-introduire dans les tableaux et graphiques de résultats.

Cette méthode peut notamment être utilisée lorsque des individus ou des variables ont une influence trop importante sur les résultats d'une ACP. On recommence alors les calculs en les déclarant comme individus inactifs ou variables supplémentaires. Elle peut également être utilisée pour introduire des variables plus synthétiques, et des moyennes par groupe d'individus, comme c'est le cas dans l'exemple ci-dessous.

Avec Statistica, il est simple de déclarer une variable comme variable supplémentaire : le premier dialogue de l'ACP prévoit une zone d'édition pour cela. Pour déclarer des individus comme "inactifs", il est nécessaire de construire une variable supplémentaire, qui ne contiendra que deux modalités, et d'utiliser les zones d'édition "Variable avec individus actifs" et "Code des individus actifs".

Ouvrez le fichier Proteines-2008.stw.

Source : Exemple fourni avec le logiciel Statistica.
Cet exemple particulier est présenté par Greenacre (1984) dans le cadre d'une comparaison entre l'analyse en composantes principales (voir l'Analyse Factorielle) et l'analyse des correspondances.

Les données du fichier d'exemple Protein.sta représentent des estimations de la consommation protéique issue de 9 sources différentes, par habitant dans 25 pays (les données ont initialement été reportées par Weber, 1973, dans un polycopié publié à l'Université de Kiel, Institut für Agrarpolitik und Marktlehre, intitulé "Agrarpolitik im Spannungsfeld der Internationalen Ernährungspolitik").

Au fichier de données initial ont été ajoutées les 5 variables suivantes :

- Consommation en protéines animales (somme des variables v1 à v5)
- Consommation en protéines végétales (somme des variables v6 à v9)
- Un code du nom du pays sur 2 ou 3 lettres
- Le groupe auquel appartient le pays (4 groupes ont été définis : NW (Europe du Nord et de l'Ouest), NE (Europe de l'Est, pays du Nord), SW (Europe de l'Ouest, pays du Sud) et SE (Europe de l'Est, pays du Sud)).
- Une variable codant pour les individus actifs (1) et inactifs (0).

Quatre individus ont été ajoutés, correspondant aux moyennes observées dans les 4 groupes de pays définis précédemment

Extrait des données :

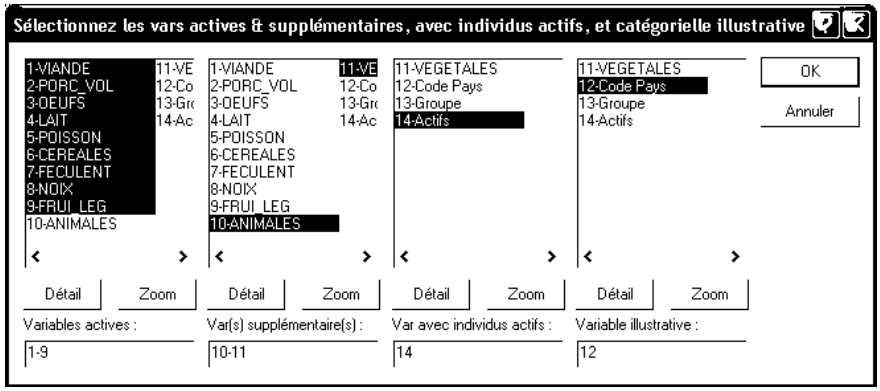
	Evaluation des consommations de protéïnes, en grammes/habitant/jour								
	1 VIANDE	2 ORC_VO	3 OEUFS	4 LAIT	5 'OISSON	6 EREALE	7 ECULEN	8 NOIX	9 RUI_LE
Belgique/Lux.	13,5	9,3	4,1	17,5	4,5	26,6	5,7	2,1	4,0
Bulgarie	7,8	6,0	1,6	8,3	1,2	56,7	1,1	3,7	4,2
Tchécoslovaquie	9,7	11,4	2,8	12,5	2,0	34,3	5,0	1,1	4,0
Danemark	10,6	10,8	3,7	25,0	9,9	21,9	4,8	0,7	2,4
R.D.A.	8,4	11,6	3,7	11,1	5,4	24,6	6,5	0,8	3,6
Finlande	9,5	4,9	2,7	33,7	5,8	26,3	5,1	1,0	1,4

Toutes les variables s'expriment ici avec la même unité (g.hab/jour). Pour réaliser une ACP, deux possibilités s'offrent à nous :

- Faire une ACP sur les valeurs non réduites. Ainsi, une information telle que "l'apport protéique des viandes, porc et volailles est, dans tous les cas, supérieur à celui des fruits et légumes" est prise en compte dans l'étude.

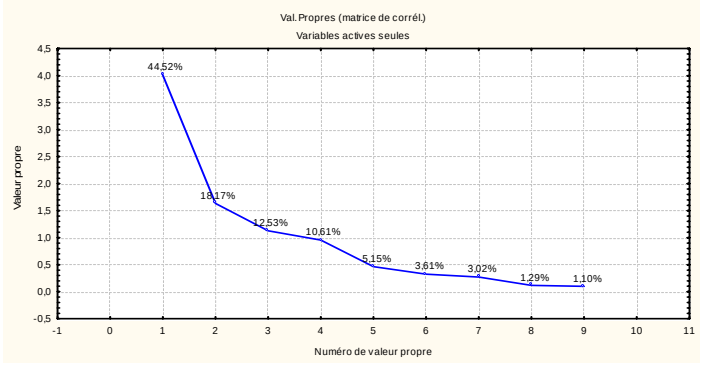
- Faire une ACP sur les valeurs réduites (ACP calculée à partir du tableau des corrélations). Dans ce cas, l'étude "gomme" les inégalités des apports protéiques des différentes sources.

Réalisons une ACP sur les corrélations en spécifiant individus actifs et variables supplémentaires comme suit :



Affichez les tableaux des covariances et des corrélations. On voit déjà apparaître une opposition entre protéines d'origine animale et protéines d'origine végétale.

Combien de valeurs propres faut-il ici retenir ? Seules 3 valeurs propres sont supérieures à 1, mais la règle du coude conduit à retenir soit 2, soit 4 axes factoriels. En fait, il faut conserver 4 axes pour mettre en évidence certaines spécificités des pays d'Europe Centrale (axe 3) ou de la France (axe 4).



Exercice : Calculez les résultats de l'ACP pour les 4 premiers axes à l'aide de Statistica, puis interprétez les résultats.

2.1.6 ACP avec rotation

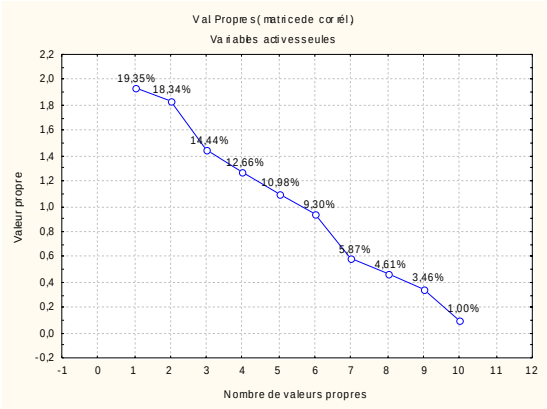
Par construction, les composantes principales sont des abstractions mathématiques et ne possèdent pas nécessairement de signification intuitive. Après avoir réalisé l'ACP, il peut parfois être intéressant de définir d'autres variables en effectuant une combinaison linéaire des composantes principales retenues, à l'aide d'une "rotation". L'objectif est généralement d'augmenter les saturations, c'est-à-dire les corrélations entre ces nouveaux "facteurs" et certaines variables de départ. Les nouveaux "facteurs" ainsi obtenus perdent les propriétés des facteurs principaux. Par exemple, le premier d'entre eux ne correspond plus à la direction de plus grande dispersion du nuage des individus. En revanche, la part de variance expliquée par les facteurs retenus reste identique. Il existe différents critères (varimax, quartimax, equamax, etc) permettant d'obtenir une rotation conduisant à des saturations proches de 1 ou -1, ou au contraire proches de 0.

Cette possibilité n'est pas disponible dans la méthode "ACP à la française" de Statistica. En revanche, on peut l'utiliser en utilisant le module "Analyse factorielle" convenablement paramétré.

2.1.7 Une ACP fournit-elle toujours des informations interprétables ?

Tout tableau de données peut être soumis à une ACP, et les méthodes d'analyse qui ont été développées permettent de "trouver des résultats". Mais ces résultats correspondent-ils à une réalité plus ou moins cachée ou ne constituent-ils qu'un artefact de la méthode ?
Pour étudier cet aspect, réalisons une ACP sur des données ... où il n'y a rien à dire (il s'agit de données produites à l'aide d'un générateur de nombres aléatoires).

Ouvrez le fichier aleatoire-20sujets.stw et réalisez une ACP normée sur ces données. La représentation graphique des valeurs propres nous indique déjà l'absence d'intérêt des données traitées :



2.2 Analyse Factorielle des Correspondances

Bibliographie :
Escofier, Pagès : Analyses factorielles simples et multiples
Lebart, Morineau, Piron, Statistique exploratoire multidimensionnelle
G. Saporta. Probabilité, Analyse des données et statistique. Editions Technip , 1990

2.2.1 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.

2.2.2 Traitement classique d'un tableau de contingence : exemple

	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 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

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

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

Effectifs observés O

	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 théoriques 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

Ecarts à l'indépendance : E = O - 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

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

	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

D'où : Khi-2 = 320,2.
L'hypothèse d'indépendance est largement rejetée.

2.2.3 Analyse factorielle des correspondances proprement dite

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 $n_{i.}$.
- La somme des éléments d'une colonne est notée $n_{.j}$.

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

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

Exemple : distance entre les lignes 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

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

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

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

Exemple : distance entre les colonnes 1 et 2

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

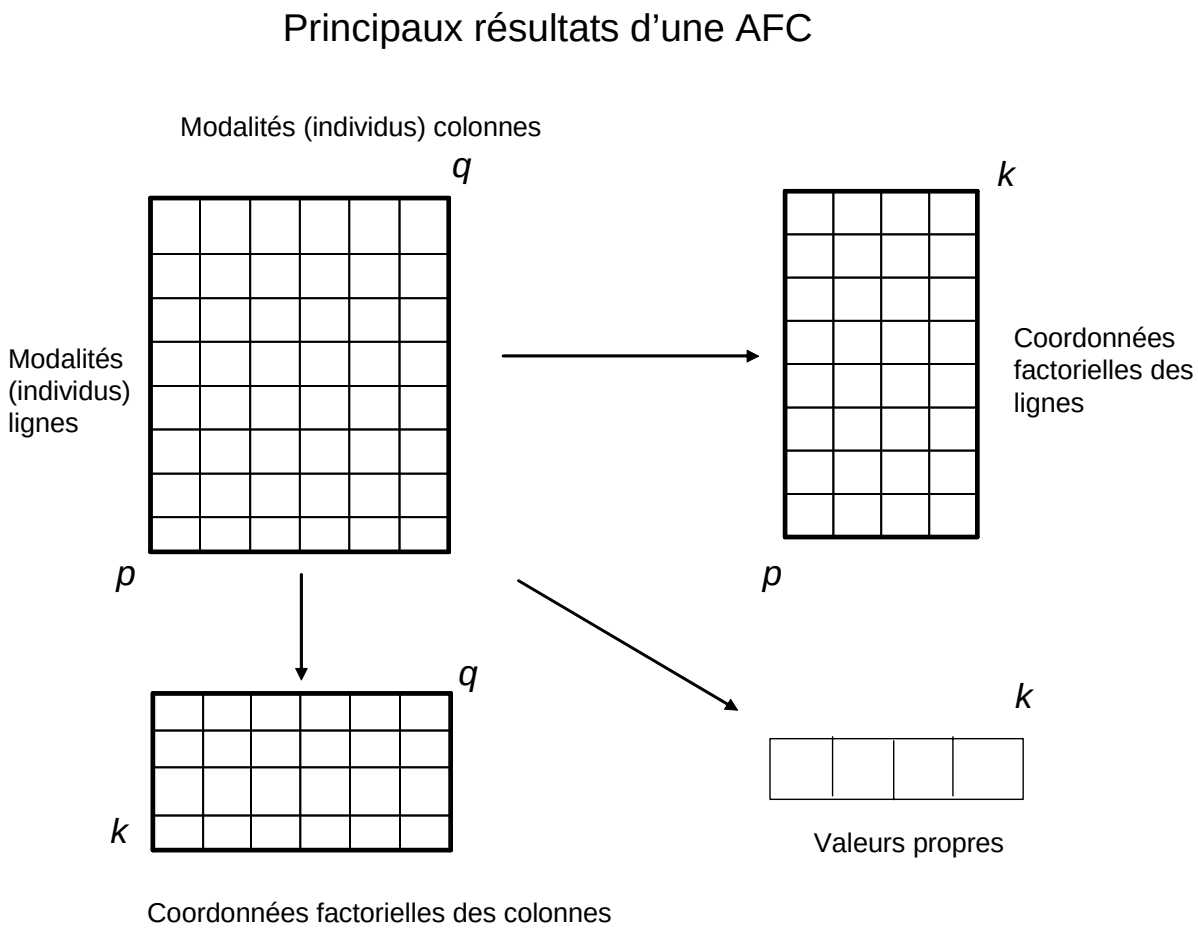
Propriété d'équivalence distributionnelle :

- 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.

L'analyse des correspondances détermine une représentation "optimale" de la distance du Phi-2 entre les individus lignes, et de même, une représentation optimale de la distance du Phi-2 entre les individus colonnes. Elle permet également de représenter les individus lignes et les individus colonnes sur une même carte factorielle.

Principaux résultats de l'AFC :



Valeurs propres

	ValProp.	%age inertie	%age cumulé	Chi²
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

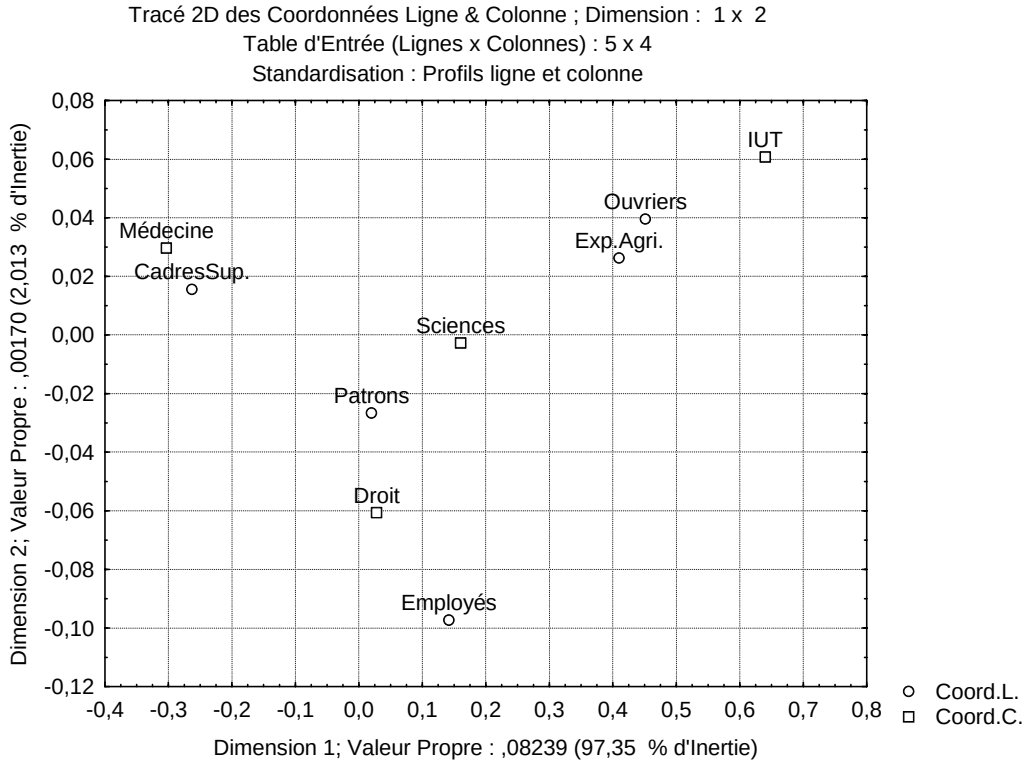
Résultats relatifs aux lignes

	Coord. Dim.1	Coord. Dim.2	Masse	Qualité	Inertie Relative	Inertie Dim.1	Cosinus² Dim.1	Inertie Dim.2	Cosinus² 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	Cosinus² Dim.1	Inertie Dim.2	Cosinus² 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

Carte factorielle lignes et colonnes



2.2.4 Analyse factorielle des correspondances avec Statistica

2.2.4.1 Traitement des données avec Statistica

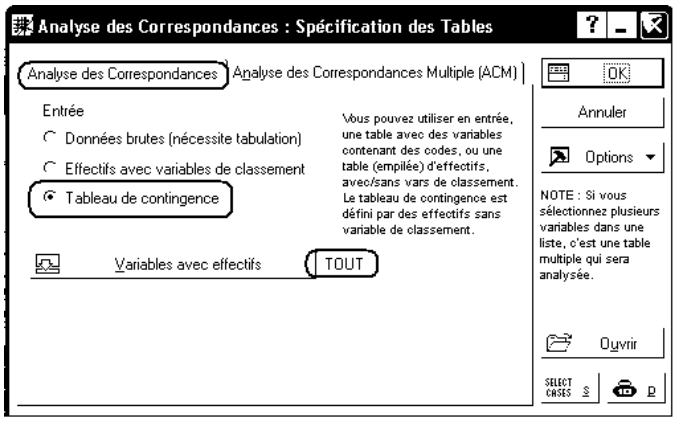
Source : Site Eurostat de l'Union Européenne.
http://epp.eurostat.ec.europa.eu/portal/

Ouvrez le classeur Regions-2001.stw

La feuille "Regions-Milliers-2001" rapporte des données relatives à la structure de la population : elle indique, pour chacune des 22 régions françaises (en lignes) le nombre d'habitants (en milliers) par âge (en colonnes) :
HF00 signifie Hommes et Femmes de 0 à 4 ans,
HF05 signifie Hommes et Femmes de 5 à 9 ans, ...
HF80 signifie Hommes et Femmes de plus de 80 ans

	1 HF00	2 HF05	3 HF10	4 HF15
ILEF	744	724	703	706
CHAM	82	86	93	95
PICA	120	128	138	134
HNOR	114	120	129	131

Pour effectuer l'AFC, nous utilisons le menu Statistiques - Techniques exploratoires multivariées - Analyse des correspondances.



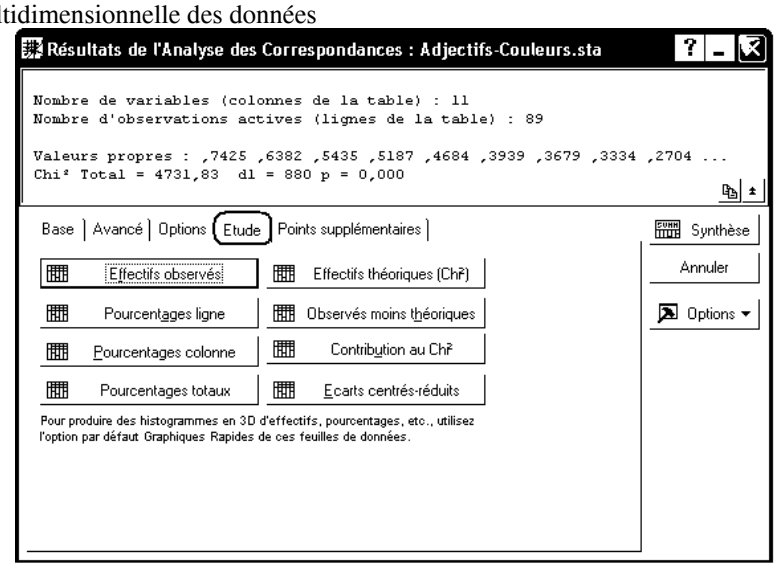
La fenêtre de dialogue permet d'indiquer la manière dont se présentent nos données. La situation la plus classique est celle d'un tableau de contingence : les modalités lignes sont indiquées dans une variable spécifiques, les modalités colonnes sont les autres variables du tableau, et la feuille de données contient les effectifs n_{ij}.

On indique également les variables qui participeront à l'analyse (ici toutes les variables). Notez que les zéros éventuels sont obligatoires, car une cellule laissée vide est interprétée comme une valeur manquante, et c'est alors l'ensemble de la ligne qui est éliminé de l'analyse.

N.B. Ne fermez pas l'analyse en cours pendant la suite des manipulations. Ainsi, vous n'aurez pas à indiquer de nouveau les options ci-dessus, vos résultats seront cohérents entre eux et se rassembleront dans un même classeur.

Statistiques descriptives

Les principaux résultats de statistiques descriptives pourront être obtenus à partir de l'onglet "Etude".
On peut ainsi obtenir les fréquences, les fréquences lignes, les fréquences colonnes et les profils moyens.



Par exemple, le tableau des fréquences et les profils ligne et colonne moyens sont :

Pourcentages Totaux (Regions-Milliers-2001 dans Regions-2001-Def.stw)
Table d'Entrée (Lignes x Colonnes) : 22 x 17
Inertie Totale = ,00882 Chi² = 515,83 dl = 336 p = 0,0000

	HF00	HF05	HF10	HF15	HF20	HF25	HF30	HF35	HF40	HF45	HF50	HF55	HF60	HF65	HF70	HF75	HF80	Total
ILEF	1,27	1,24	1,20	1,21	1,29	1,59	1,56	1,49	1,37	1,37	1,26	0,88	0,74	0,67	0,57	0,46	0,54	18,71
CHAM	0,14	0,15	0,16	0,16	0,15	0,17	0,16	0,17	0,17	0,17	0,15	0,10	0,11	0,10	0,09	0,08	0,08	2,29
PICA	0,21	0,22	0,24	0,23	0,20	0,23	0,23	0,24	0,23	0,24	0,20	0,13	0,14	0,14	0,12	0,10	0,09	3,17
HNOR	0,19	0,21	0,22	0,22	0,19	0,22	0,22	0,22	0,23	0,22	0,19	0,13	0,13	0,13	0,11	0,09	0,10	3,04
CENT	0,24	0,26	0,27	0,27	0,24	0,29	0,29	0,30	0,30	0,31	0,27	0,19	0,21	0,21	0,19	0,16	0,18	4,17
BNOR	0,15	0,15	0,17	0,17	0,15	0,16	0,17	0,17	0,18	0,17	0,15	0,10	0,12	0,12	0,11	0,09	0,09	2,43
BOUR	0,15	0,16	0,17	0,18	0,16	0,18	0,19	0,19	0,20	0,20	0,18	0,13	0,14	0,15	0,13	0,11	0,12	2,75
NORD	0,46	0,48	0,52	0,54	0,49	0,50	0,49	0,49	0,49	0,48	0,41	0,26	0,29	0,29	0,26	0,21	0,18	6,82
LORR	0,23	0,25	0,27	0,28	0,26	0,28	0,29	0,30	0,29	0,29	0,24	0,18	0,19	0,19	0,17	0,12	0,12	3,95
ALSA	0,19	0,19	0,19	0,19	0,19	0,23	0,24	0,23	0,23	0,22	0,17	0,14	0,14	0,13	0,11	0,09	0,08	2,96
FCOM	0,12	0,12	0,13	0,14	0,12	0,14	0,14	0,14	0,14	0,14	0,12	0,09	0,09	0,09	0,08	0,06	0,07	1,91
PAYS	0,34	0,35	0,37	0,40	0,36	0,38	0,38	0,39	0,39	0,39	0,34	0,24	0,26	0,26	0,24	0,19	0,20	5,51
BRET	0,29	0,30	0,32	0,34	0,32	0,34	0,34	0,36	0,35	0,35	0,31	0,22	0,26	0,26	0,24	0,19	0,19	4,97
POIT	0,15	0,16	0,17	0,18	0,16	0,18	0,19	0,19	0,20	0,20	0,18	0,14	0,15	0,16	0,14	0,12	0,13	2,80
AQUI	0,26	0,28	0,30	0,31	0,30	0,33	0,34	0,36	0,36	0,37	0,33	0,24	0,25	0,26	0,25	0,21	0,22	4,97
MIDI	0,23	0,24	0,25	0,27	0,27	0,29	0,31	0,32	0,31	0,31	0,28	0,21	0,22	0,23	0,22	0,18	0,20	4,36
LIMO	0,05	0,06	0,06	0,07	0,07	0,08	0,08	0,08	0,09	0,09	0,08	0,06	0,07	0,08	0,07	0,06	0,07	1,21
RHON	0,61	0,63	0,64	0,66	0,62	0,70	0,73	0,72	0,69	0,68	0,63	0,47	0,43	0,43	0,38	0,31	0,32	9,65
AUVE	0,11	0,12	0,13	0,14	0,14	0,15	0,15	0,16	0,16	0,17	0,15	0,11	0,12	0,12	0,11	0,10	0,10	2,24
LANG	0,22	0,23	0,24	0,25	0,24	0,26	0,26	0,28	0,27	0,28	0,25	0,19	0,20	0,22	0,20	0,17	0,17	3,92
PROV	0,44	0,47	0,48	0,48	0,45	0,50	0,54	0,55	0,54	0,54	0,51	0,41	0,39	0,40	0,36	0,31	0,34	7,70
CORS	0,02	0,03	0,03	0,03	0,02	0,03	0,03	0,03	0,03	0,03	0,03	0,03	0,02	0,02	0,02	0,02	0,02	0,44
Total	6,09	6,29	6,55	6,74	6,36	7,20	7,31	7,37	7,22	7,22	6,43	4,66	4,66	4,66	4,17	3,43	3,63	100,00

Remarque :

Statistica ne permet pas d'obtenir directement le tableau des taux de liaison, qui est pourtant un outil exploratoire intéressant. Mais on pourra utiliser les tableaux "Observés moins théoriques" et "Effectifs théoriques". On peut même recopier ces deux tableaux dans une feuille Excel et diviser chaque cellule du premier par la cellule correspondante du second pour obtenir le tableau des taux de liaison :

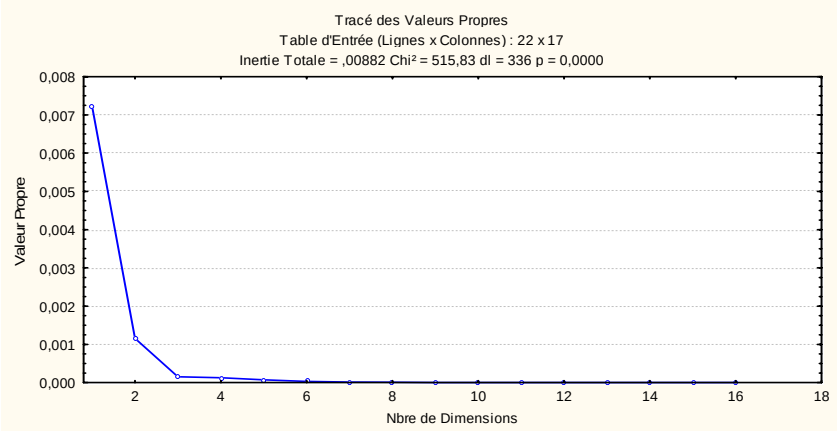
	HF00	HF05	HF10	HF15	HF20	HF25	HF30	HF35	HF40	HF45	HF50	HF55	HF60	HF65	HF70	HF75	HF80
ILEF	0,12	0,05	-0,02	-0,04	0,09	0,18	0,14	0,08	0,02	0,02	0,05	0,01	-0,15	-0,23	-0,27	-0,28	-0,21

CHAM	0,00	0,02	0,06	0,05	0,01	0,00	-0,02	-0,01	0,00	0,01	-0,00	-0,07	-0,01	-0,02	-0,02	-0,02	-0,04
PICA	0,06	0,10	0,14	0,07	-0,03	-0,01	0,01	0,01	0,02	0,04	-0,02	-0,10	-0,06	-0,09	-0,08	-0,12	-0,18
HNOR	0,05	0,07	0,11	0,09	-0,01	-0,01	-0,02	-0,00	0,04	0,02	-0,01	-0,08	-0,06	-0,07	-0,10	-0,10	-0,10
CENT	-0,04	-0,02	-0,01	-0,03	-0,09	-0,05	-0,05	-0,03	-0,01	0,01	0,01	-0,01	0,05	0,07	0,08	0,11	0,20
BNOR	0,00	0,01	0,05	0,05	-0,04	-0,07	-0,07	-0,04	0,00	-0,02	-0,05	-0,08	0,09	0,10	0,10	0,07	0,01
BOUR	-0,09	-0,06	-0,03	-0,03	-0,09	-0,09	-0,07	-0,05	-0,01	0,02	0,01	0,02	0,09	0,15	0,16	0,18	0,23
NORD	0,11	0,12	0,17	0,16	0,12	0,02	-0,03	-0,02	-0,01	-0,03	-0,07	-0,18	-0,10	-0,08	-0,08	-0,11	-0,28
LORR	-0,03	0,01	0,04	0,04	0,02	-0,02	-0,01	0,02	0,03	0,02	-0,04	-0,02	0,04	0,04	0,01	-0,08	-0,17
ALSA	0,03	0,04	0,01	-0,02	0,01	0,07	0,10	0,07	0,06	0,01	-0,09	0,04	-0,02	-0,06	-0,11	-0,14	-0,24
FCOM	-0,00	-0,00	0,04	0,06	-0,03	-0,02	-0,01	-0,03	0,00	-0,01	0,02	-0,00	0,02	0,02	-0,01	-0,03	-0,04
PAYS	0,02	0,01	0,03	0,09	0,03	-0,04	-0,05	-0,03	-0,01	-0,01	-0,05	-0,07	0,02	0,02	0,03	0,03	0,02
BRET	-0,04	-0,04	-0,02	0,02	0,00	-0,06	-0,08	-0,03	-0,01	-0,02	-0,03	-0,05	0,11	0,12	0,15	0,14	0,06
POIT	-0,14	-0,12	-0,07	-0,04	-0,08	-0,12	-0,09	-0,06	0,01	0,01	-0,01	0,04	0,14	0,19	0,23	0,25	0,31
AQUI	-0,13	-0,11	-0,09	-0,06	-0,07	-0,07	-0,06	-0,03	-0,00	0,02	0,02	0,04	0,10	0,14	0,19	0,22	0,24
MIDI	-0,12	-0,11	-0,12	-0,09	-0,03	-0,07	-0,03	-0,01	-0,01	-0,01	0,01	0,04	0,10	0,14	0,19	0,22	0,27
LIMO	-0,26	-0,22	-0,18	-0,14	-0,11	-0,12	-0,13	-0,08	-0,02	0,01	0,03	0,06	0,21	0,33	0,42	0,44	0,59
RHON	0,04	0,04	0,02	0,02	0,00	0,01	0,03	0,01	-0,01	-0,02	0,02	0,04	-0,04	-0,05	-0,06	-0,08	-0,08
AUVE	-0,17	-0,15	-0,12	-0,06	-0,05	-0,09	-0,07	-0,06	-0,00	0,04	0,06	0,08	0,13	0,18	0,23	0,25	0,22
LANG	-0,09	-0,07	-0,06	-0,05	-0,04	-0,09	-0,08	-0,05	-0,05	-0,02	0,00	0,07	0,09	0,18	0,20	0,24	0,22
PROV	-0,06	-0,03	-0,05	-0,07	-0,09	-0,10	-0,04	-0,03	-0,03	-0,03	0,02	0,13	0,08	0,11	0,13	0,17	0,21
CORS	-0,12	-0,08	-0,06	-0,09	-0,21	-0,09	-0,00	-0,01	0,01	0,01	0,02	0,24	0,15	0,16	0,11	0,12	0,17

Choix des valeurs propres

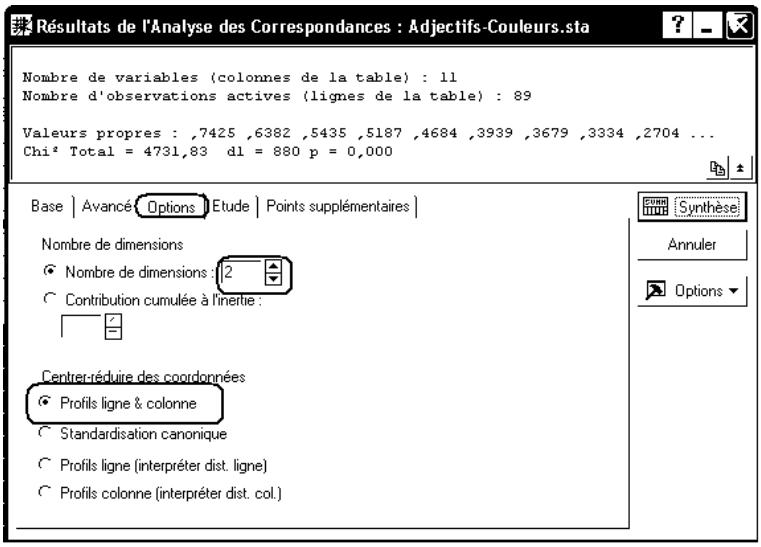
C'est ensuite l'onglet "Avancé" qui nous permettra d'afficher les valeurs propres, et donc de choisir le nombre d'axes à garder.

Nombre de Dims.	Valeurs Propres et Inertie de toutes les Dimensions) Table d'Entrée (Lignes x Colonnes) : 22 x 17 Inertie Totale = ,00882 Chi² = 515,83 dl = 336 p = 0,000				
	ValSing.	ValProp.	%age Inertie	%age Cumulé	Chi²
1	0,0850	0,0072	81,92	81,92	422,54
2	0,0340	0,0012	13,14	95,06	67,78
3	0,0123	0,0002	1,71	96,76	8,81
4	0,0113	0,0001	1,44	98,20	7,42
5	0,0086	0,0001	0,85	99,05	4,37
6	0,0058	0,0000	0,38	99,43	1,97
7	0,0042	0,0000	0,20	99,64	1,05
8	0,0035	0,0000	0,14	99,78	0,71
9	0,0024	0,0000	0,07	99,84	0,35
10	0,0023	0,0000	0,06	99,90	0,30
11	0,0018	0,0000	0,04	99,94	0,18
12	0,0015	0,0000	0,02	99,96	0,13
13	0,0012	0,0000	0,02	99,98	0,08
14	0,0010	0,0000	0,01	99,99	0,06
15	0,0008	0,0000	0,01	100,00	0,04
16	0,0006	0,0000	0,00	100,00	0,02



Résultats relatifs aux individus-lignes et aux individus-colonnes.

Pour les résultats qui suivent, on indique le nombre d'axes factoriels à conserver sous l'onglet "Base" ou sous l'onglet "Options". Ce dernier permet également de choisir plusieurs types d'échelles pour représenter lignes et colonnes. Le type de représentation le plus classique, qui fait jouer des rôles symétriques aux lignes et aux colonnes, correspond à la première option.



On retourne ensuite sous l'onglet "Avancé" pour afficher les coordonnées des individus-lignes et des individus-colonnes. On notera que Statistica produit deux tableaux de résultats, et on passera de l'un à l'autre à l'aide des onglets du classeur.

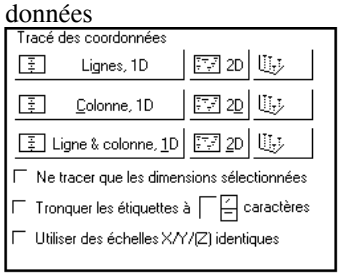
Coordonnées Ligne et Contributions à l'Inertie (Regions-Milliers-2001 dans Regions-2001-Def.stw)
Table d'Entrée (Lignes x Colonnes) : 22 x 17
Standardisation : Profils ligne et colonne

NomLigne	Ligne Numéro	Coord. Dim.1	Coord. Dim.2	Masse	Qualité	Inertie Relative	Inertie Dim.1	Cosinus² Dim.1	Inertie Dim.2	Cosinus² Dim.2	Cos² 1&2
ILEF	1	-0,1223	-0,0427	0,1871	0,9968	0,3574	0,3877	0,8885	0,2944	0,1082	0,9968
CHAM	2	-0,0129	0,0222	0,0229	0,7971	0,0022	0,0005	0,1997	0,0098	0,5974	0,7971
PICA	3	-0,0568	0,0396	0,0317	0,8540	0,0202	0,0142	0,5745	0,0430	0,2794	0,8540
HNOR	4	-0,0444	0,0356	0,0304	0,8471	0,0132	0,0083	0,5163	0,0332	0,3308	0,8471
CENT	5	0,0547	-0,0018	0,0417	0,8358	0,0170	0,0173	0,8349	0,0001	0,0009	0,8358
BNOR	6	0,0332	0,0420	0,0243	0,9231	0,0086	0,0037	0,3556	0,0369	0,5675	0,9231
BOUR	7	0,0900	0,0005	0,0275	0,9771	0,0258	0,0308	0,9771	0,0000	0,0000	0,9771
NORD	8	-0,0774	0,0788	0,0682	0,9772	0,0967	0,0566	0,4795	0,3662	0,4977	0,9772
LORR	9	-0,0151	0,0250	0,0395	0,4218	0,0090	0,0012	0,1126	0,0213	0,3092	0,4218
ALSA	10	-0,0618	-0,0090	0,0296	0,6520	0,0201	0,0157	0,6385	0,0021	0,0135	0,6520
FCOMTE	11	-0,0029	0,0160	0,0191	0,3961	0,0014	0,0000	0,0124	0,0042	0,3837	0,3961
PAYS	12	0,0088	0,0342	0,0551	0,8109	0,0096	0,0006	0,0498	0,0557	0,7611	0,8109
BRET	13	0,0569	0,0241	0,0497	0,8945	0,0241	0,0223	0,7586	0,0249	0,1359	0,8945
POIT	14	0,1223	-0,0047	0,0280	0,9891	0,0481	0,0580	0,9876	0,0005	0,0015	0,9891
AQUI	15	0,0985	-0,0184	0,0497	0,9830	0,0576	0,0668	0,9500	0,0145	0,0330	0,9830
MIDI	16	0,0994	-0,0293	0,0436	0,9659	0,0550	0,0597	0,8888	0,0323	0,0771	0,9659
LIMO	17	0,2133	-0,0393	0,0121	0,9774	0,0663	0,0765	0,9454	0,0162	0,0321	0,9774
RHON	18	-0,0312	-0,0007	0,0965	0,7453	0,0143	0,0130	0,7450	0,0000	0,0003	0,7453
AUVE	19	0,1155	-0,0238	0,0224	0,9426	0,0374	0,0413	0,9042	0,0110	0,0385	0,9426
LANG	20	0,1002	-0,0046	0,0392	0,9730	0,0460	0,0546	0,9709	0,0007	0,0021	0,9730
PROV	21	0,0791	-0,0202	0,0770	0,9122	0,0638	0,0667	0,8561	0,0272	0,0561	0,9122
CORS	22	0,0872	-0,0389	0,0044	0,7346	0,0063	0,0047	0,6129	0,0058	0,1216	0,7346

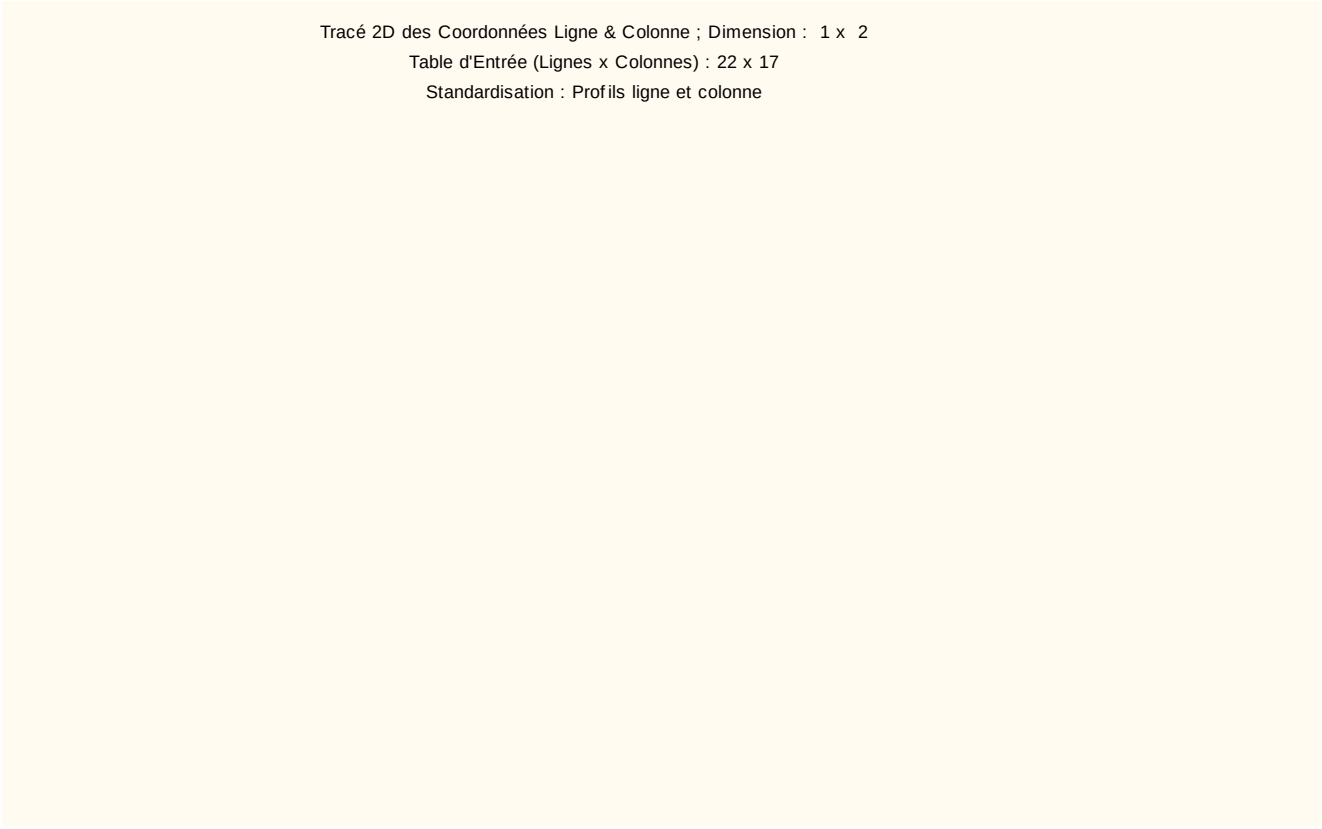
Coordonnées Colonne et Contributions à l'Inertie (Regions-Milliers-2001 dans Regions-2001-Def.stw)
Table d'Entrée (Lignes x Colonnes) : 22 x 17
Standardisation : Profils ligne et colonne

Nom Col.	Colonne Numéro	Coord. Dim.1	Coord. Dim.2	Masse	Qualité	Inertie Relative	Inertie Dim.1	Cosinus² Dim.1	Inertie Dim.2	Cosinus² Dim.2	Cos² 1&2
HF00	1	-0,0882	0,0106	0,0609	0,9564	0,0570	0,0656	0,9427	0,0059	0,0137	0,9564
HF05	2	-0,0637	0,0262	0,0629	0,9260	0,0365	0,0353	0,7916	0,0374	0,1343	0,9260
HF10	3	-0,0426	0,0602	0,0655	0,9515	0,0424	0,0164	0,3172	0,2047	0,6342	0,9515
HF15	4	-0,0262	0,0634	0,0674	0,9676	0,0372	0,0064	0,1407	0,2338	0,8269	0,9676
HF20	5	-0,0585	0,0094	0,0636	0,7551	0,0336	0,0302	0,7360	0,0049	0,0191	0,7551
HF25	6	-0,0881	-0,0335	0,0720	0,9645	0,0753	0,0774	0,8427	0,0698	0,1219	0,9645
HF30	7	-0,0661	-0,0373	0,0731	0,9848	0,0485	0,0443	0,7471	0,0878	0,2378	0,9848
HF35	8	-0,0384	-0,0207	0,0737	0,9260	0,0172	0,0150	0,7180	0,0271	0,2080	0,9260
HF40	9	-0,0121	-0,0029	0,0722	0,3271	0,0039	0,0015	0,3089	0,0005	0,0182	0,3271
HF45	10	-0,0041	-0,0067	0,0722	0,1446	0,0034	0,0002	0,0391	0,0028	0,1055	0,1446
HF50	11	-0,0018	-0,0329	0,0643	0,6808	0,0116	0,0000	0,0021	0,0601	0,6788	0,6808
HF55	12	0,0316	-0,0602	0,0466	0,7834	0,0312	0,0065	0,1696	0,1456	0,6138	0,7834
HF60	13	0,0982	0,0074	0,0466	0,9549	0,0537	0,0623	0,9495	0,0022	0,0054	0,9549
HF65	14	0,1405	0,0229	0,0466	0,9866	0,1086	0,1274	0,9612	0,0211	0,0255	0,9866
HF70	15	0,1672	0,0250	0,0417	0,9936	0,1361	0,1615	0,9719	0,0225	0,0217	0,9936
HF75	16	0,1851	0,0122	0,0343	0,9904	0,1351	0,1626	0,9861	0,0044	0,0043	0,9904
HF80	17	0,1932	-0,0471	0,0363	0,9643	0,1688	0,1876	0,9103	0,0694	0,0540	0,9643

On utilise ensuite les boutons du bloc "Tracé des coordonnées" pour obtenir des représentations graphiques des résultats de l'AFC.



Les graphiques "par axe" pourront être obtenus à l'aide du bouton "Ligne & colonne, 1D". Le graphique dans un plan, superposant les résultats des lignes et des colonnes, pourra être obtenu à l'aide du bouton "2D" de la même ligne. En revanche, il n'est pas évident d'éliminer certaines étiquettes pour améliorer la lisibilité du graphique. La seule méthode paraît être de faire un clic droit sur une étiquette, de sélectionner l'item de menu "Propriétés..." puis d'éditer manuellement le tableau des étiquettes qui s'affiche.



supérieure à la moyenne 0,0088/16 = 0,00055. Nous étudierons donc les deux premiers axes factoriels.

2.2.5.1 Interprétation du premier axe

Les individus lignes dont la contribution à l'inertie du premier axe est supérieure à la moyenne sont :

-	+
ILEF (39%) NORD (5,7%)	LIMO (7,7%) AQUI (6,7%) PROV (6,7%) MIDI (6%) POIT (5,8%) LANG (5,5%)

On voit que cet axe oppose des régions telles que l'Ile de France et le Nord Pas de Calais à un ensemble de régions "du sud" : Limousin, Aquitaine, Provence, etc. L'Ile de France représente à elle seule 39% de l'inertie de cet axe, et on peut s'étonner que cette région, malgré son poids démographique, soit représentée par un point aussi éloigné de l'origine des axes.

Pour les individus colonnes, les résultats sont :

-	+
HF25 (8%) HF00 (7%)	HF80 (19%) HF75 (16%) HF70 (16%) HF65 (13%) HF60 (6%)

Clairement, le premier axe oppose les classes d'âge élevées (partie positive de l'axe) aux autres classes, notamment la classe 25-29 ans et la classe 0-4 ans.

La synthèse des études menées sur les individus lignes et sur les individus colonnes en découle aussitôt : le premier axe oppose des régions où la population âgée est importante à des régions plus jeunes, ou dans lesquelles apparaît un déficit en personnes âgées (Ile de France et Nord Pas de Calais, mais aussi Alsace, Picardie, Haute Normandie, etc).

2.2.5.2 Etude du second axe factoriel

Les individus lignes dont la contribution à l'inertie du premier axe est supérieure à la moyenne sont :

-	+
ILEF (29,4%)	NORD (36,6%)

Les individus colonnes dont la contribution à l'inertie du premier axe est supérieure à la moyenne sont :

-	+

HF55 (14,6%)	HF15 (23,3%)
HF30 (8,8%)	HF10 (20,4%)
HF25 (7%)	
HF80 (7%)	
HF50 (6%)	

Le tableau des individus lignes semble montrer que cet axe oppose essentiellement deux régions "jeunes" : l'Ile de France et le Nord Pas de Calais. En fait, dans la partie négative de cet axe, on retrouve à la fois des régions "jeunes", telles que l'Ile de France et des régions "âgées" telles que le Limousin, pendant que la partie positive de l'axe rassemble des régions (Nord, mais aussi Picardie, Basse Normandie, Pays de la Loire, etc) où la population des adolescents (HF10, HF15) est bien représentée.

2.2.5.3 Quelques remarques sur les qualités de représentation

On voit que les âges correspondant aux adultes actifs (HF35, HF40, HF45) sont très peu intervenus dans l'étude. Les effectifs de ces classes d'âge diffèrent peu de l'indépendance : il y a peu de différences entre les régions du point de vue de la proportion de 35-49 ans dans la population. De faible inertie et donc intervenant peu dans la formation des premiers axes, ces individus colonnes peuvent être mal représentés (qualité de représentation égale à 0,14, par exemple, pour HF45 et à 0,32 pour HF40 : il faut donc s'abstenir d'interpréter, sans élément supplémentaire, leur proximité sur le graphique).

De même, la qualité de représentation de la Franche Comté (0,39) est assez faible, car cette région est peu importante numériquement et a un profil assez proche du profil moyen. Sur le schéma, elle apparaît proche de la Champagne, ce qui ne correspond pas vraiment à la réalité.

2.2.5.4 Synthèse

L'élément dominant que l'AFC fait apparaître est l'opposition entre d'une part les régions comportant beaucoup de personnes âgées (60 ans et plus), et par voie de conséquence, un déficit d'enfants et de jeunes adultes, et d'autre part, les régions comportant beaucoup de jeunes de moins de 35/40 ans et peu de personnes âgées. Une structure secondaire distingue, parmi les régions "jeunes" celles dont la population comporte de nombreux adultes (classes HF25, HF30 particulièrement nombreuses) à celles dont la population comporte beaucoup d'enfants (HF05, HF10, HF15).

On est ainsi tenté de définir quatre groupes de régions, sans pour autant pouvoir affecter objectivement chaque région à un groupe :
- Régions à population de personnes âgées importante : Limousin, Corse, Midi-Pyrénées, Auvergne, Provence, Aquitaine, Languedoc, Poitou, Bourgogne.
- Régions "intermédaires" : Centre, Bretagne, Basse Normandie, Pays de la Loire et peut-être Lorraine, Champagne, Franche Comté
- Régions à forte population de jeunes adultes : Ile de France, Alsace et peut-être Rhône-Alpes.
- Régions à forte population de jeunes enfants : Nord Pas-de-Calais, Picardie, Haute-Normandie.

2.2.6 Structures possibles pour les données d'entrée

Source : Exemple fourni avec le logiciel Statistica.

Supposons que vous ayez collecté des données sur les habitudes de différents salariés d'une entreprise concernant la cigarette. Les données suivantes sont présentées dans l'ouvrage de Greenacre (1984, p. 55).

Ouvrez le classeur Smoking.stw et observez les 3 feuilles de données saisies.

2.2.6.1 Données structurées sous forme d'un tableau de contingence

Commençons, par exemple, par rendre active la feuille de données Smoking1.sta (tableau de contingence).

	Analyse des correspondances simple.			
	1 NON_FUM	2 OCCAS	3 MOYEN	4 GROS_FUM
CADRE_EXPER	4	2	3	2
CADRE_DEBUT	4	3	7	4
EMPLOY_EXPER	25	10	12	4
EMPLOY_DEBUT	18	24	33	13
SECRETAIRES	10	6	7	2

Réalisez une AFC sur ce tableau de données.
N.B. On remarquera que le test du khi-2 sur ce tableau ne démontre pas l'existence d'une dépendance significative entre les habitudes concernant la cigarette et l'emploi occupé. L'analyse factorielle des correspondances est donc d'un intérêt très limité ici.

2.2.6.2 Données structurées sous forme de tableau d'effectifs

Statistica nous permet également de réaliser l'AFC à partir d'un tableau d'effectifs (feuille de données Smoking2.sta).

Refaites l'AFC précédente, d'abord en utilisant Smoking2.sta comme feuille de données active.

2.2.6.3 Données structurées sous forme de tableau protocole

On peut aussi réaliser l'AFC à partir d'un tableau protocole (données non recensées - feuille de données Smoking3.sta).

Refaites l'AFC précédente, d'abord en utilisant Smoking3.sta comme feuille de données active.

2.2.7 Ajout de lignes ou de colonnes supplémentaires : application à la comparaison de tableaux de fréquence binaire

On dispose des données suivantes relatives aux élèves scolarisés en 1972/73, sortis du système éducatif en 1973 et ayant trouvé un emploi :

Hommes	Sans diplôme	BEPC	BEP/CAP	BAC général	BAC technique	DEUG/ENT	DUT/BTS/ Santé	SUP	Total
Agriculteurs	15068	2701	5709	297	1242	0	322	0	25339
Ingénieurs	0	337	309	917	0	308	0	4383	6254
Techniciens	302	1697	2242	1969	1399	357	1943	381	10290
Ouvriers Qualifiés	10143	3702	30926	314	1861	0	0	337	47283
Ouvriers non qualifiés	59394	8087	17862	2887	1696	0	0	323	90249
Cadres supérieurs	596	298	892	1227	298	2362	318	6781	12772

Cadres Moyens	2142	2801	672	6495	924	2807	2301	4030	22172
Employés qualifiés	5445	7348	4719	4353	1280	614	982	0	24741
Employés non qualifiés	4879	4987	1514	3478	886	1326	0	661	17731
Total	97969	31958	64845	21937	9586	7774	5866	16896	256831

Femmes	Sans diplôme	BEPC	BEP/CAP	BAC général	BAC technique	DEUG/ENT	DUT/BTS/ Santé	SUP	Total
Agriculteurs	5089	1212	1166	0	0	0	0	0	7467
Ingénieurs	0	0	0	316	0	0	304	1033	1653
Techniciens	281	0	320	320	283	0	683	0	1887
Ouvriers Qualifiés	7470	1859	4017	1752	657	0	285	0	16040
Ouvriers non qualifiés	29997	4334	4538	1882	0	0	0	0	40751
Cadres supérieurs	0	0	0	2236	595	911	569	6788	11099
Cadres Moyens	1577	1806	4549	17063	875	4152	15731	3991	49744
Employés qualifiés	21616	19915	32452	16137	5865	1256	3332	1286	101859
Employés non qualifiés	19849	7325	6484	5111	898	294	635	0	40596
Total	85879	36451	53526	44817	9173	6613	21539	13098	271096

Source : B. Escoffier, J. Pagès, Analyses factorielles simples et multiples, 3è édition - Dunod 1998.

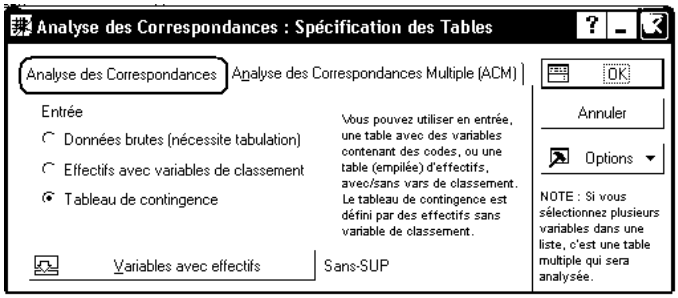
Ces tableaux croisent trois variables qualitatives : l'emploi, le diplôme et le sexe. Les buts de notre étude peuvent être multiples. D'une part, on peut s'intéresser à la liaison entre emploi et diplôme, indépendamment du sexe, et mettre ainsi en évidence une structure commune à ces deux tableaux. D'autre part, et de façon complémentaire, on peut s'intéresser aux écarts entre ces deux tableaux : les répartitions croisées des emplois et des diplômes sont-elles similaires selon le sexe, ou au contraire, sont-elles très différentes ?

2.2.7.1 Première analyse : les tableaux "par sexe" en éléments supplémentaires dans l'AFC de leur somme

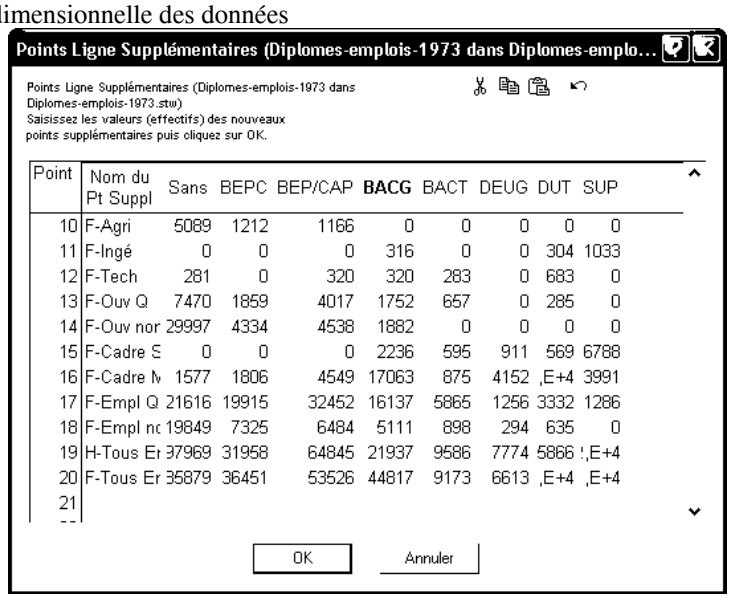
Ouvrez le classeur Diplomes-emploi-1973.stw et observez la manière dont les données y ont été saisies. Ouvrez également le classeur Excel Diplomes-emplois-1973.xls.

On va réaliser une AFC sur le tableau "somme", en plaçant en éléments supplémentaires les tableaux relatifs aux données par sexe.

Réalisez une AFC sur les variables 1 à 8 du tableau de données Statistica :

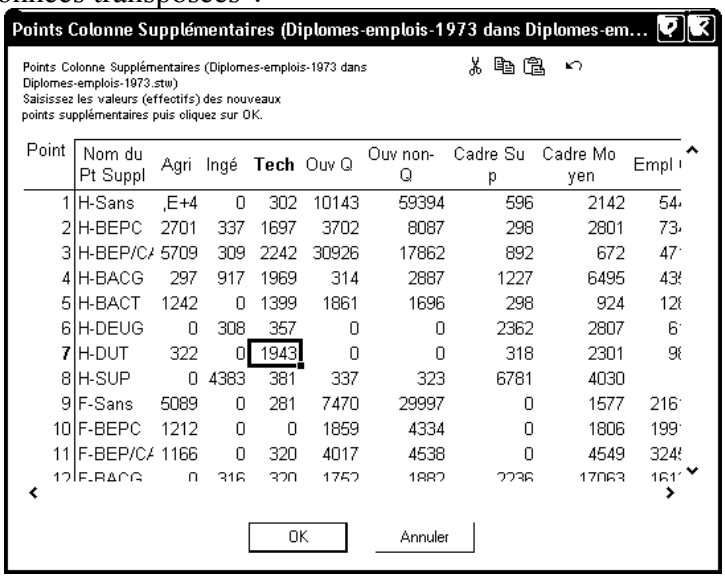


Activez ensuite l'onglet "Points supplémentaires" et cliquez sur le bouton "Ajouter des points lignes". Plutôt que de saisir ces données supplémentaires à la main, copiez, puis collez dans la fenêtre la plage A11:I30 de la feuille "Donnees" du classeur Excel.



Point	Nom du Pt Suppl	Sans	BEPC	BEP/CAP	**BACG**	BACT	DEUG	DUT	SUP
10	F-Agri	5089	1212	1166	0	0	0	0	0
11	F-Ingé	0	0	0	316	0	0	304	1033
12	F-Tech	281	0	320	320	283	0	683	0
13	F-Ouv Q	7470	1859	4017	1752	657	0	285	0
14	F-Ouv nor	29997	4334	4538	1882	0	0	0	0
15	F-Cadre S	0	0	0	2236	595	911	569	6788
16	F-Cadre M	1577	1806	4549	17063	875	4152	,E+4	3991
17	F-Empl Q	21616	19915	32452	16137	5865	1256	3332	1286
18	F-Empl no	19849	7325	6484	5111	898	294	635	0
19	H-Tous Er	37969	31958	64845	21937	9586	7774	5866	,E+4
20	F-Tous Er	35879	36451	53526	44817	9173	6613	,E+4	,E+4
21									

Après avoir validé, cliquez de même sur "Ajouter des points colonnes" et collez la plage A10:J27 de la feuille Excel "Donnees transposees".

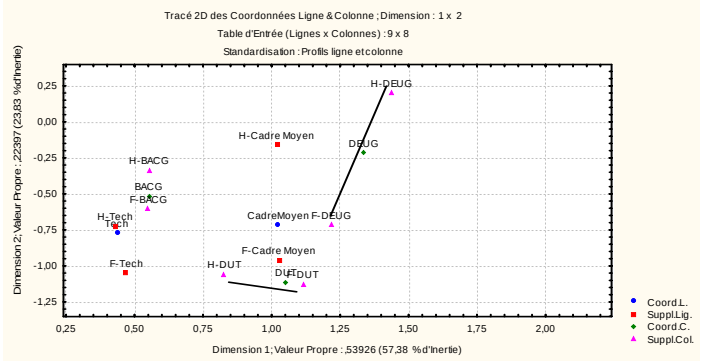


Point	Nom du Pt Suppl	Agri	Ingé	**Tech**	Ouv Q	Ouv non-Q	Cadre Sup	Cadre Moyen	Empl
1	H-Sans	,E+4	0	302	10143	59394	596	2142	54
2	H-BEPC	2701	337	1697	3702	8087	298	2801	73
3	H-BEP/CA	5709	309	2242	30926	17862	892	672	47
4	H-BACG	297	917	1969	314	2887	1227	6495	43
5	H-BACT	1242	0	1399	1861	1696	298	924	12
6	H-DEUG	0	308	357	0	0	2362	2807	6
7	H-DUT	322	0	1943	0	0	318	2301	9
8	H-SUP	0	4383	381	337	323	6781	4030	
9	F-Sans	5089	0	281	7470	29997	0	1577	216
10	F-BEPC	1212	0	0	1859	4334	0	1806	199
11	F-BEP/CA	1166	0	320	4017	4538	0	4549	324
12	F-BACG	0	316	320	1752	1882	2236	17063	161

Poursuivez ensuite l'exécution de l'ACP : valeurs propres, coordonnées lignes et colonnes, graphiques des points lignes, des points colonnes et graphique lignes et colonnes.

Pour interpréter les résultats trouvés, on commence par s'intéresser aux individus lignes et colonnes actifs. Ici, le premier axe classe les emplois et les diplômes en plaçant sur la partie gauche de l'axe "Sans diplôme" et les emplois peu qualifiés et sur la partie droite les diplômes "supérieurs" et les emplois d'ingénieurs et cadres supérieurs. Le second axe oppose les diplômes et emplois "moyens" (techniciens, cadres moyens, Bac, DEUG), qui occupent la partie négative de l'axe aux diplômes et emplois "extrêmes" (emplois non qualifiés, cadres supérieurs, sans diplôme, études supérieures) sur la partie positive de l'axe. Cette configuration est classique lorsque l'AFC s'applique à des variables ordinales, et porte le nom d'effet Guttman.

Pour étudier les points lignes et points colonnes supplémentaires, on compare leur position à celle du point correspondant du tableau "somme" :



Le point DEUG, par exemple, est situé à la moyenne pondérée des points H-DEUG et F-DEUG. Comme les effectifs masculins et féminins pour le DEUG sont sensiblement équivalents, ce point se trouve approximativement au milieu du segment (H-DEUG, F-DEUG).

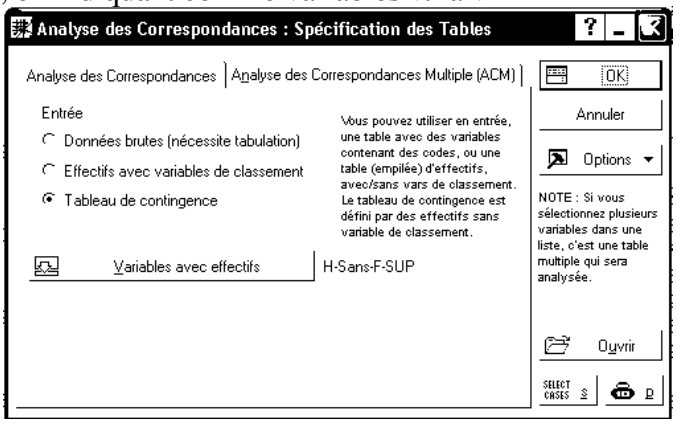
On constate que, sur le premier axe, pour tous les diplômes, les deux points représentant les hommes et les femmes sont presque confondus. En revanche, pour les points relatifs au DEUG par exemple, la différence des coordonnées sur le deuxième axe est très importante. D'une manière générale, on constate que, s'agissant des diplômes, les points relatifs aux femmes ont généralement une coordonnée sur l'axe 2 inférieure à celle du correspondant relatif aux hommes : les femmes occupent, plus que les hommes, les emplois "moyens". Inversement, les hommes sont plus nombreux à occuper des emplois "extrêmes".

Deux remarques méritent d'être faites

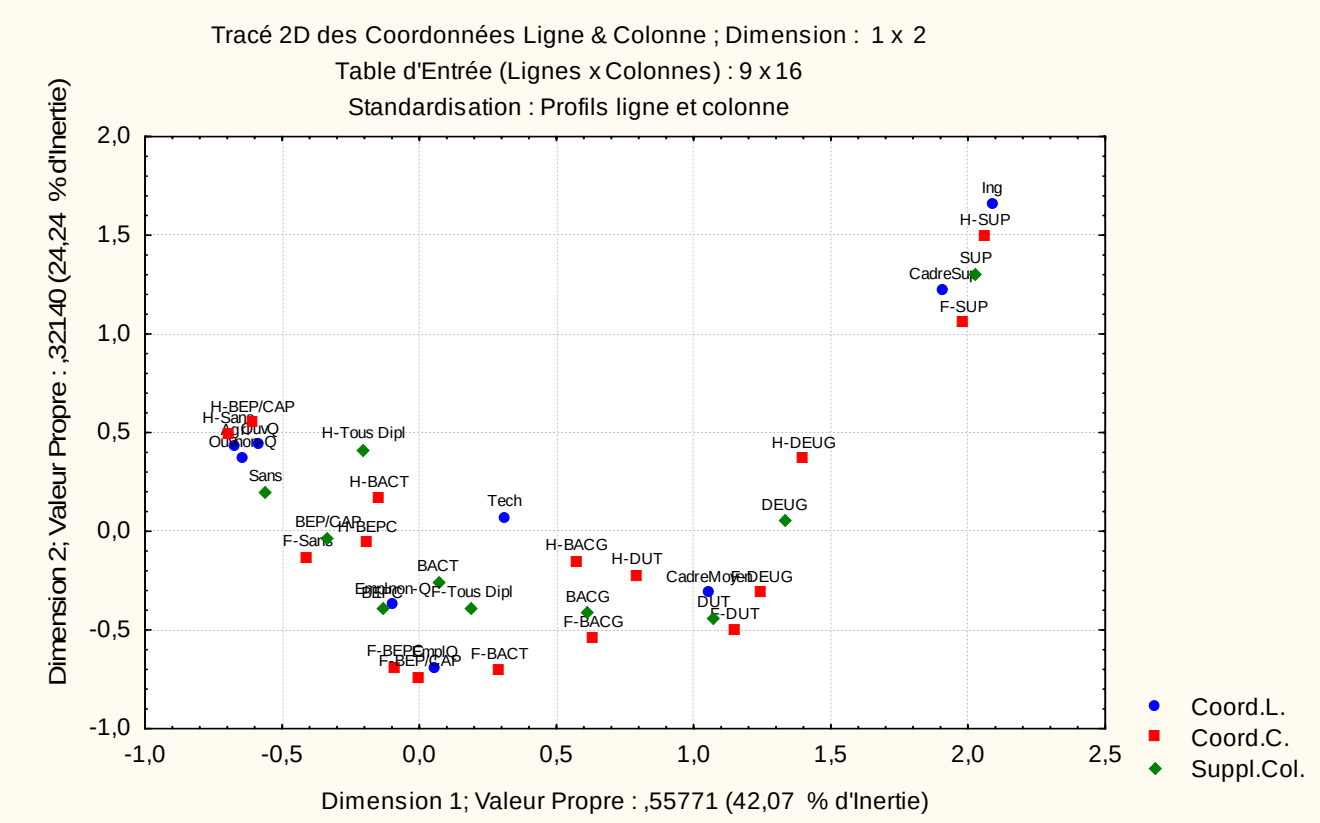
- Dans l'étude menée ici, l'inertie prise en compte (Phi-2 = 0,94) est celle du tableau "somme". On ne tient donc pas compte de la dispersion des données due aux discriminations liées au sexe.
- Deux points supplémentaires correspondant aux deux sexes peuvent être représentés proches l'un de l'autre sur le graphique, alors qu'il existe une forte disparité entre hommes et femmes pour cette modalité, et nous disposons de peu de moyens pour le mettre en évidence. Ce type de situation se produit lorsque la dispersion "entre sexes" est orthogonale à la dispersion due aux autres deux autres facteurs.

2.2.7.2 Deuxième analyse : tableaux "par sexe" juxtaposés et tableau "somme" en éléments supplémentaires.

Réalisez une autre AFC, en indiquant comme variables v9 à v24.



Ajoutez comme points colonnes supplémentaires des données relatives au tableau somme et à la synthèse des emplois par sexe, c'est-à-dire les plages A2:J9 et A26:J27 de la feuille Excel "Donnees transposees".



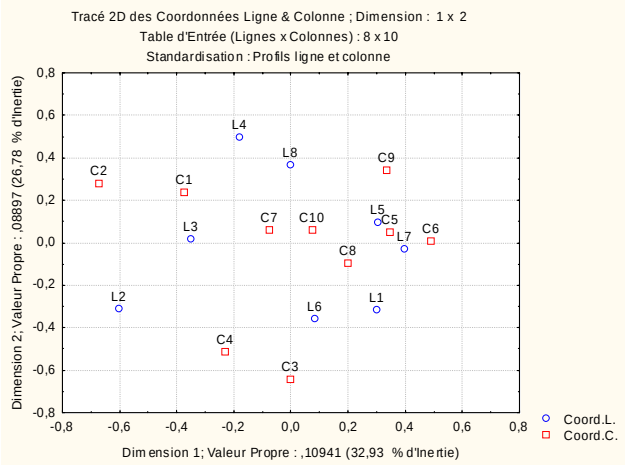
L'interprétation du graphique se fait comme précédemment. Cependant, l'inertie du nuage de points fait maintenant intervenir les points relatifs aux diplômes par sexe. On constate que le coefficient Phi-2 est plus élevé que dans l'étude précédente : 1,326 au lieu de 0,940. La différence entre les deux coefficients représente l'inertie "intra" (dispersion liée au sexe, pour chaque diplôme), qui représente ici presque 30% du total. Une étude plus poussée (dont les détails sortent du cadre de ce cours) permettrait de montrer que cette inertie intra est très faible sur le premier axe, mais représente presque la moitié de l'inertie du deuxième axe. Ainsi cette méthode permet, dans une certaine mesure, d'évaluer l'importance des écarts des colonnes homologues aux colonnes moyennes.

2.2.8 Quelques configurations remarquables dans les résultats produits par une AFC.

On pourra consulter le fichier Configurations-Types.stw qui rassemble quelques configurations classiques de nuages, générées à partir de données fictives.

2.2.8.1 Forme générale du nuage

L'inertie totale (le Phi-2) est un indicateur de la dispersion totale du nuage. La comparaison des inerties de chacun des axes (c'est-à-dire des valeurs propres associées aux axes) renseigne sur la forme du nuage de points. Si les premières valeurs propres sont proches les unes des autres, la dispersion est relativement homogène : il n'y a pas vraiment de direction privilégiée et le nuage de points est approximativement sphérique. Si au contraire, les valeurs propres sont nettement différentes, cela traduit un nuage de points fortement allongé selon une (ou plusieurs) direction.

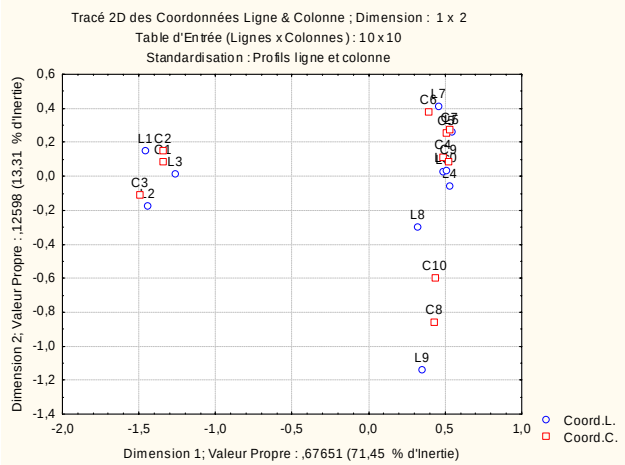


2.2.8.2 Deux paquets de points - Valeurs propres proches de 1

Les valeurs propres sont toutes inférieures à 1. Mais, une valeur propre proche de 1 indique une dichotomie des données, c'est-à-dire un tableau de contingence qui, après reclassement des modalités, aurait l'allure suivante :

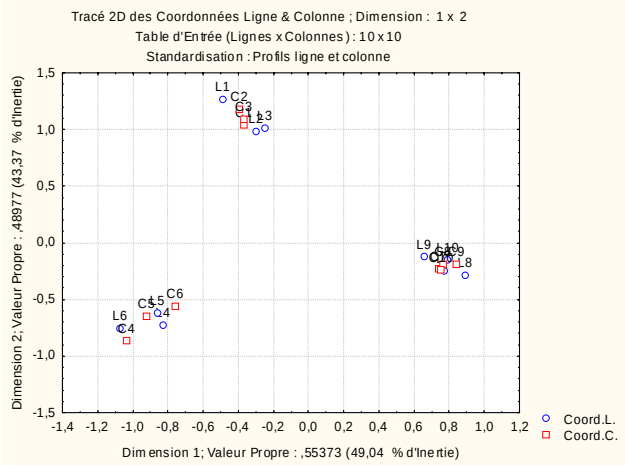
	0
0	

Le nuage est alors divisé en deux paquets de points. La feuille de données "Deux-paquets" fournit une illustration de cette situation.



2.2.8.3 Trois paquets de points

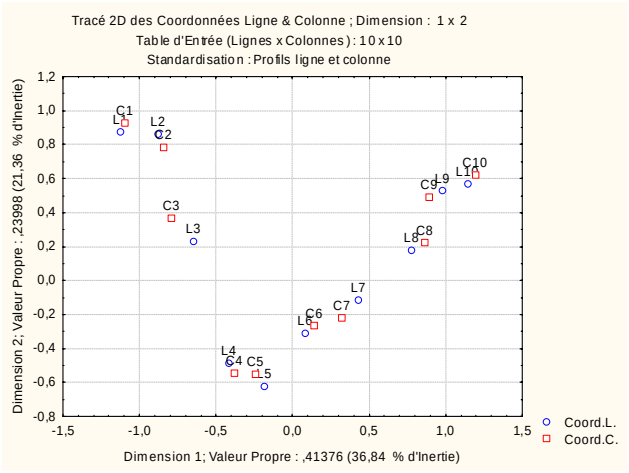
De même, l'existence de deux valeurs propres proches de 1 indique une partition des observations en 3 groupes. Si toutes les valeurs propres sont proches de 1, cela indique une correspondance entre chaque modalité ligne et une modalité colonne "associée". Avec une réorganisation convenable des modalités, les effectifs importants se trouvent alors le long de la diagonale.
La feuille de données "Trois-paquets" fournit une illustration de cette situation.

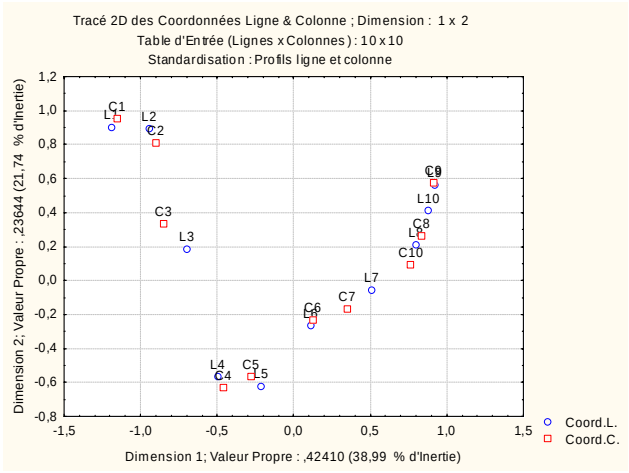


2.2.8.4 L'effet Guttman.

Un nuage de points de forme parabolique indique une redondance entre les deux variables étudiées : la connaissance de la ligne i donne pratiquement celle de la colonne j. Dans un tel cas, pratiquement toute l'information est contenue dans le premier facteur. Cette configuration se rencontre notamment lorsque les deux variables sont ordinales, et classent les sujets de la même façon. Dans ce cas, le premier axe oppose les valeurs extrêmes et classe les valeurs, tandis que le deuxième axe oppose les intermédiaires aux extrêmes.

La feuille de données "Effet-Guttman" fournit une illustration assez caractéristique de cette situation. Dans ce cas, on a intérêt à ne pas limiter l'étude au plan (1, 2). La configuration-type dans les trois plans de projection définis par les 3 premiers axes prend souvent les allures indiquées dans l'exemple.
Il pourra alors être intéressant d'examiner les accidents des courbes qui joignent les points, qui reflètent les particularités des situations étudiées. Voir par exemple la situation des modalités L10 et C10 dans l'exemple "Guttman-perturbé".





2.2.8.5 Nuage tétraédrique

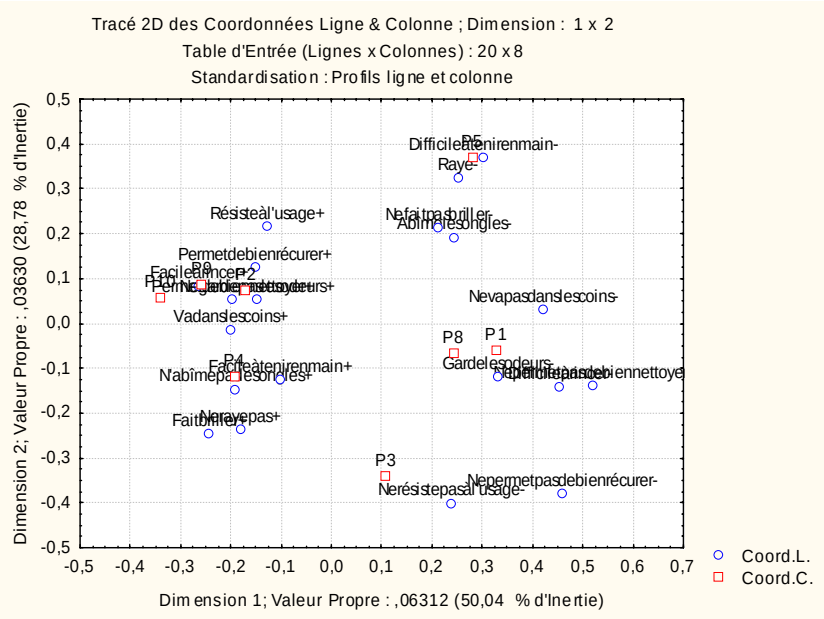
Le premier exemple ("Deux-paquets") est également caractéristique d'une forme classique de nuage : tétraédrique, ou en forme de "berlingot" comme on peut s'en rendre compte en construisant les projections du nuage sur les 3 premiers axes.

2.2.9 L'extension de la notion de tableau de contingence

En toute rigueur, l'analyse de correspondances ne s'applique qu'aux tableaux de contingence. Elle peut cependant être appliquée à des tableaux qui, a priori, ne sont pas des tableaux de contingence. Un critère essentiel pour décider si un tableau peut être assimilé à un tableau de contingence est le suivant : on doit pouvoir donner un sens à la somme des cases du tableau, qu'elle soit faite par ligne ou par colonne.

2.2.9.1 Tableaux juxtaposés

Considérons l'exemple fourni dans le classeur Echelles-Likert.stw. On obtient ainsi un point par produit et deux points par échelle bipolaire. On peut facilement montrer que le barycentre (pondéré) des deux points correspondant à une échelle donnée se trouve au centre de gravité du nuage. Si le point "+" se trouve plus près de l'origine que le point "-", cela signifie que l'intensité de la propriété positive est supérieure à celle de la propriété négative correspondante. Cet effet est connu sous le nom d'effet de levier.



Dans certains cas, on peut juxtaposer, par exemple, deux tableaux de contingence correspondant à des dates différentes, par exemple la ventilation de la population française par région et par CSP pour deux recensements différents. Il sera alors pertinent d'étudier comment chaque modalité s'est déplacée entre l'époque 1 et l'époque 2.

2.2.9.2 Juxtaposer plusieurs tableaux : vers l'ACM

Source : Hahn A., Eirmbter W. H., Jacob R., Le sida : savoir ordinaire et insécurité, traduction française de Herrmann M.

Il s'agit d'une enquête réalisée durant l'été 1990, auprès d'un échantillon représentatif des ménages de RFA.

Résumé du questionnaire :

Variable	Modalité	Codage
Sexe	masculin	m
	féminin	f

Confession	protestant	ev
	catholique	rk
	autre	an
	sans	ke
Liens avec l'église	forts	f1
	moyens	f2
	inexistants	f3
Catégorie Sociale	élèves/étud	s1
	classe sup.	s2
	cl. moy. sup.	s3
	cl. moyenne	s4
	cl. moy. inf.	s5
	cl. populaire	s6
	autres	s7
Taille du lieu de résidence	< 2	k1
	2 à < 5	k2
	5 à < 20	k3
	20 à < 50	k4
	50 à < 100	k5
	100 à < 500	k6
	> 500	k7
Classe d'âge	18 à < 30	a1
	30 à < 40	a2
	40 à < 50	a3
	50 à < 60	a4
	60 et plus	a5
Fidélité dans les rapports sexuels	très pour	t1
	plutôt pour	t2
	indécis	t3
	plutôt contre	t4
	très contre	t5
Plusieurs partenaires	oui	p1
	non	p2
Préférences politiques	CDU/CSU	cd
	SPD	sp
	FDP	fd
	Verts	gr
Nombre de situations jugées contaminantes		
	0	w0
	1	w1
	2	w2
	3	w3
	4	w4
	5	w5
	6	w6

7	w7
8	w8

Le sida est la conséquence d'une faute et d'une punition

très pour	c1
plutôt pour	c2
indécis	c3
plutôt contre	c4
très contre	c5

Dispositions d'évitement et d'expulsion des contaminés de la sphère personnelle

très pour	m1
plutôt pour	m2
indécis	m3
plutôt contre	m4
très contre	m5

Nombre de mesures obligatoires acceptées

0	z0
1	z1
2	z2
3	z3
4	z4
5	z5

Nombre de situations en public jugées dangereuses

0-1	o1
2	o2
3	o3
4	o4
5-6	o5

Le sida est un péril omniprésent

d'accord	g1
indécis	g2
pas d'accord	g3

Ouvrez le classeur Hahn.stw et observez la façon dont a été constitué le tableau de contingence : la variable "groupe" est croisée avec toutes les autres variables, et on juxtapose ainsi 14 tableaux de contingence portant sur des populations presque identiques (*presque*, car pour la plupart des questions, il y a quelques non-réponses).

Réalisez une analyse des correspondances sur ce tableau et retrouvez ainsi les résultats de l'auteur :

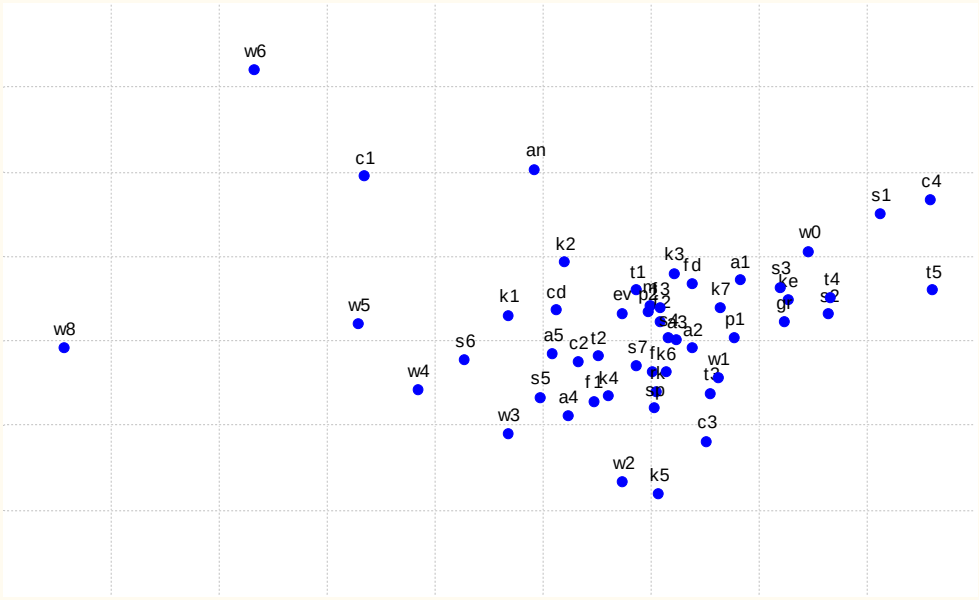
> "L'analyse des correspondances confirme l'existence de deux syndromes nettement distincts, attribuables, avec la prudence qui s'impose, à deux catégories ou milieux, qu'à la suite de Schulze on pourrait appeler "milieu harmoniste" et "milieu autodéterministe".
>
> Notre analyse utilise la dangerosité ressentie du sida comme la variable à décrire, les autres caractéristiques servant d'indices de cette appréciation. Etant donné les trois configurations de la variable à décrire, une solution bidimensionnelle serait théoriquement possible. Mais, puisque le premier axe d'inertie rend compte de 90,25% de la variation, nous négligerons ce deuxième axe.

Graphique et tableau numérique montrent que la vision du sida comme péril a été reportée sur l'ordonnée. On distingue nettement deux groupes, qui approuvent ou rejettent les termes de la question. Ceux qui ne se prononcent pas se situent entre les deux, mais sont enclins le cas échéant à considérer le sida comme une maladie omniprésente et très infectieuse.

À cela correspond la localisation des indicateurs de dispositions (perceptions, réactions) et des repères de morphologie sociale. Les enquêtés considérant le sida comme un péril le jugent très infectieux jusque dans la vie quotidienne (3 situations courantes ou plus jugées contaminantes par un taux supérieur à la moyenne). La maladie est ressentie comme conséquence et punition d'une faute morale; les dispositions d'exclusion se manifestent nettement, et les mesures obligatoires antisida - y compris la généralisation du test obligatoire - rencontrent un taux d'adhésion supérieur à la moyenne. Ceci vérifie nos hypothèses de départ : poussée à l'extrême, la conception du sida comme danger permanent de contamination fait considérer comme porteurs de virus potentiels non seulement les membres des principaux groupes à risque.(donc une minorité), mais tous les étrangers. Les mêmes enquêtés ressentent la sphère publique comme généralement inquiétante et hostile. Leurs opinions politiques plutôt conservatrices sont attestées par une préférence très nette pour les partis CDU/CSU. Ce groupe comprend une proportion importante de personnes âgées, de niveau social peu élevé, résidant plutôt dans des communes petites ou très petites.

A l'inverse, ceux pour qui le sida n'est pas un péril au sens indiqué ci-dessus, ont pour caractéristique commune de ne pas chercher un risque de contamination là où, en l'état actuel des connaissances, un tel risque n'existe pas. On n'envisage guère la maladie en termes de culpabilité, et on réclame rarement l'exclusion des contaminés ou l'adoption de mesures répressives. Or, ces, personnes sont objectivement plus exposées à la contamination.: la fidélité sexuelle est jugée relativement moins importante, le changement de partenaire est relativement fréquent. Les considérations éthico-religieuses passent à l'arrière-plan, la proportion des personnes sans confession est relativement élevée. Politiquement, ce segment se situe majoritairement à gauche du centre, avec une préférence marquée pour les Verts. Morphologiquement, il s'agit d'une population plutôt jeune, étudiante, de niveau social élevé et majoritairement citadine."

Tracé 2D des Coordonnées Ligne & Colonne ; Dimension : 1 x 2
Table d'Entrée (Lignes x Colonnes) : 69 x 3
Standardisation : Profils ligne et colonne



Le pays le plus "attiré" par la modalité "Or" est le Costa-Rica, qui n'a obtenu qu'une seule médaille, mais en or, alors que des pays tels que Cuba et l'Iran, avec des palmarès très différents, sont représentés proches l'un de l'autre, au voisinage de l'origine. En effet, les résultats de l'AFC ne concernent pas le nombre de médailles obtenues par les différents pays, mais l'écart entre les proportions de médailles de bronze, argent, or obtenues par le pays et la distribution totale (environ 1/3 de médailles de chaque type). Mais cet écart constitue-t-il vraiment un sujet d'étude ?