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

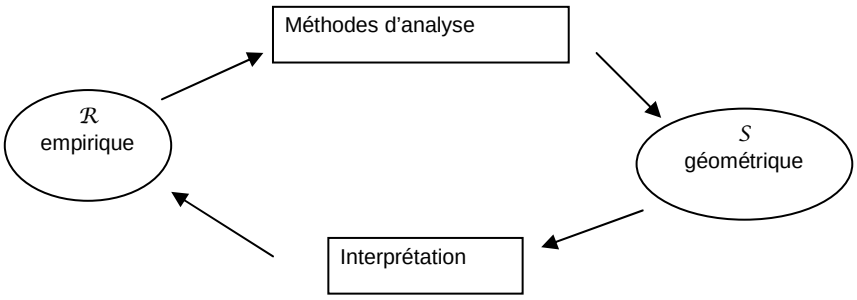
1.1 Introduction

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

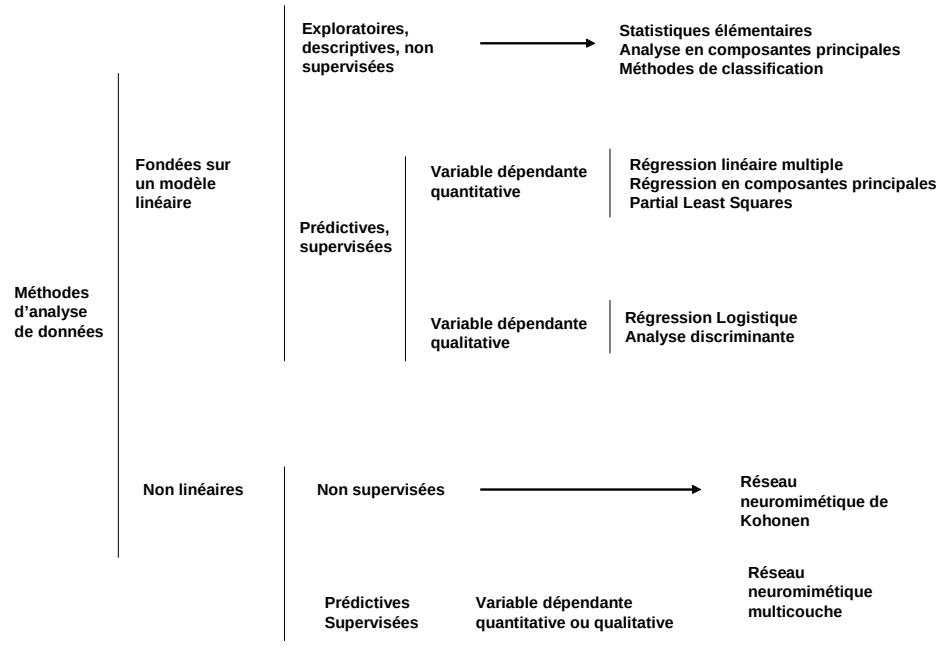
	Age	Etat-Civil	Feministe	Frequence	Agressivite	Harcelement
1	13	1	102	2	4	**0**
2	45	2	101	3	6	**0**
3	19	2	102	2	7	**1**
4	42	2	102	1	2	**1**
5	27	1	77	1	1	**0**
6	19	1	98	0	6	**1**
7	37	1	96	1	6	**0**

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 :

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

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

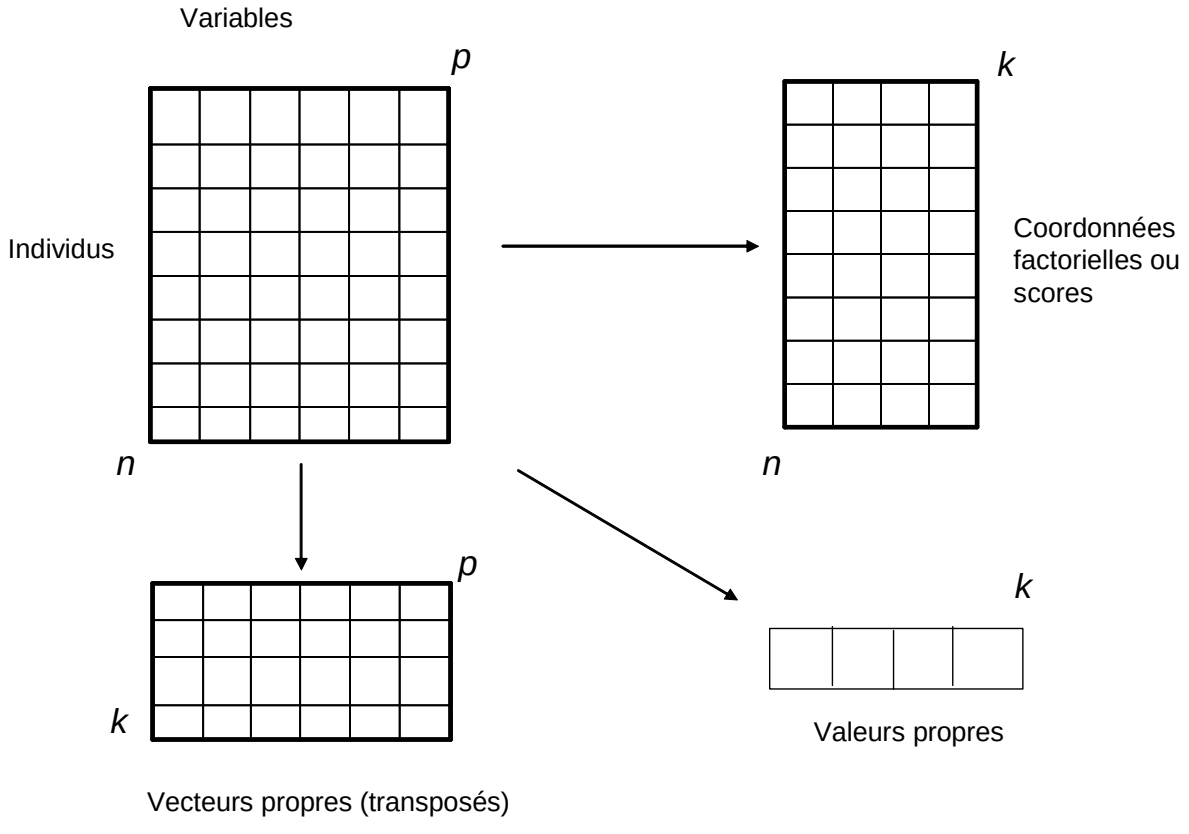
On a observé *p* variables sur *n* individus. On dit qu'il s'agit d'un protocole multivarié. Les données à traiter forment une matrice :

	X_1	X_2	...	X_p
i_1	x_{11}	x_{21}	...	x_{1p}
i_2	x_{21}	x_{22}	...	x_{2p}
...	...	...	...	...
i_n	x_{n1}	x_{n2}	...	x_{np}

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

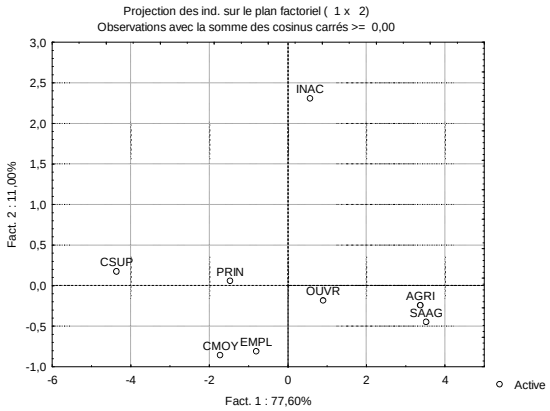
Corrélations entre variables :

	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

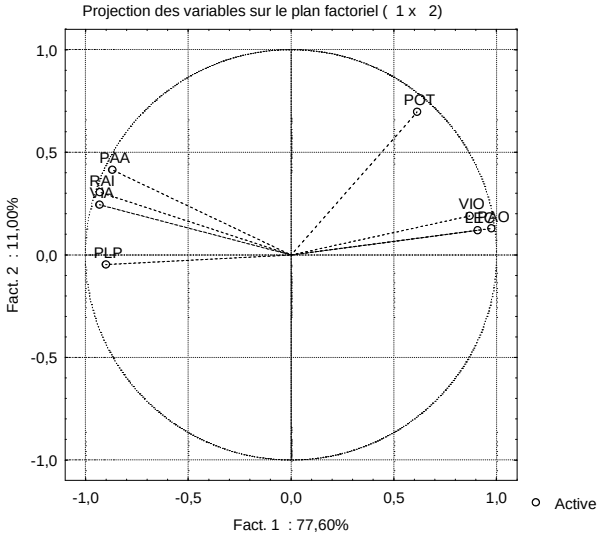
Valeurs propres de l'ACP

	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)

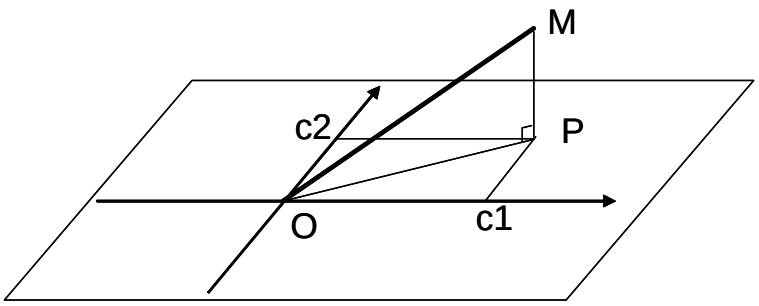


Aides à l'interprétation

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

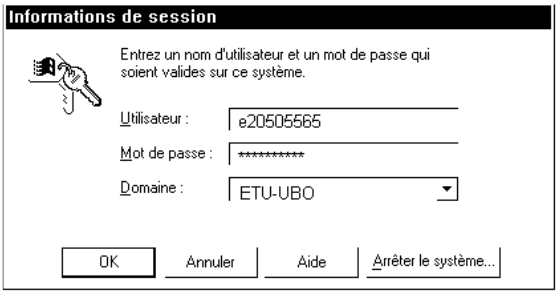
Qualité

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

2.1.3 Analyse en composantes principales avec Statistica

2.1.3.1 Organiser son espace de travail sous Statistica:

Affichez le dialogue d'ouverture de session en appuyant simultanément sur les trois touches Ctrl+Alt+Suppr. Complétez le dialogue en ouvrant la session à l'aide de vos identifiants ENT :



N.B. Pour des raisons de confidentialité, le mot de passe ne s'affiche pas "en clair".

Remarque 1. Si vous ne disposez pas encore de votre identifiant ENT, ou si votre mot de passe n'est pas reconnu, vous pouvez ouvrir une session en utilisant le compte :

Utilisateur : LETA20xPyy\etudiant
Mot de passe : ubo

Cependant, la configuration de nos appareils impose que la session soit ouverte par un utilisateur identifié, dans le domaine ETU-UBO pour que le logiciel Statistica soit disponible. Si vous avez ouvert la session avec l'identifiant etudiant, le mot de passe ubo, vous devez ensuite effectuer le montage du disque réseau contenant le logiciel Statistica (ainsi que, de préférence, de celui qui contient les fichiers utilisés en TD) à l'aide du menu Monter un volume réseau du poste de travail. Pour ces montages, vous devez utiliser les coordonnées ENT d'un utilisateur reconnu par l'ENT (l'enseignant, un collègue étudiant, etc.).

Paramètres pour le montage du volume réseau contenant le logiciel Statistica :

- Volume réseau : \\servsciences\statistica
- Lettre de lecteur : S:
- Utilisateur : login et mot de passe d'un utilisateur reconnu sur l'ENT.

Paramètres pour le montage du volume contenant les fichiers utilisés en TD :

- Volume réseau : \\serv-bu\tdlettres ou \\172.18.127.1\tdlettres
- Lettre de lecteur : W:
- Utilisateur : login et mot de passe d'un utilisateur reconnu sur l'ENT.

Chargez le logiciel Statistica en double-cliquant sur l'icône présente sur le bureau. La configuration par défaut du logiciel n'est pas vraiment satisfaisante. Nous allons donc commencer par adapter la configuration à nos besoins.

2.1.3.2 Au premier chargement du logiciel

Le logiciel propose deux options possibles pour l'affichage des menus. Vous pouvez choisir celle que vous préférez. Notez toutefois que les copies d'écran de ce polycopié utilisent l'option "Menus Classiques".

N.B. Il sera toujours possible de basculer d'une interface à l'autre à l'aide des menus Affichage - Ruban des menus classiques ou Affichage - Menus Classique de l'autre option;

Le logiciel demande ensuite si l'on souhaite installer les composants permettant au logiciel d'interagir avec R. Vous pouvez décliner la proposition et cocher l'option "Ne plus me proposer cette boîte de dialogue" car vous n'avez pas les droits d'administration permettant de faire cette installation.

2.1.3.3 Le menu Outils - Options

Le menu Outils - Options contient de nombreuses possibilités de paramétrage de Statistica. Heureusement, seules quelques-unes d'entre elles méritent d'être retouchées.
Ouvrez la fenêtre de dialogue accessible par le menu Outils-Options et explorez les différents onglets qui y sont rassemblés.

N.B. Les options ainsi choisies sont enregistrées dans le profil de l'utilisateur lorsque l'on quitte le logiciel. *Il n'y a aucun enregistrement si le compte est verrouillé ou si Statistica se plante en cours de travail.*

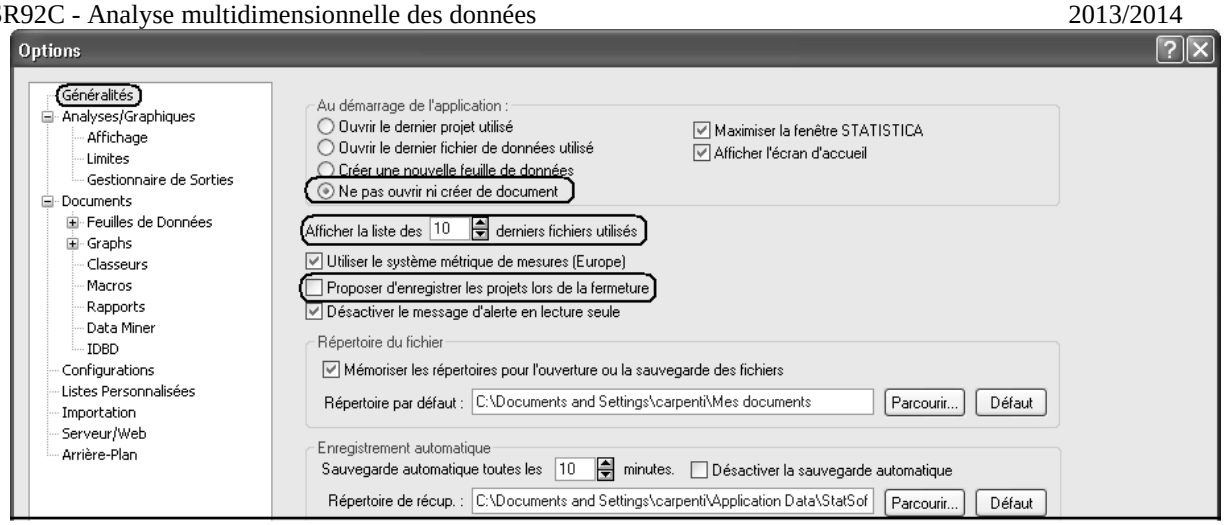
Spécifier le répertoire d'enregistrement par défaut

Affichez les options disponibles sous l'onglet Généralités.

- Choisissez de préférence l'option : *Au démarrage de l'application, ne pas ouvrir ni créer de document*., les autres options étant plutôt déroutantes.

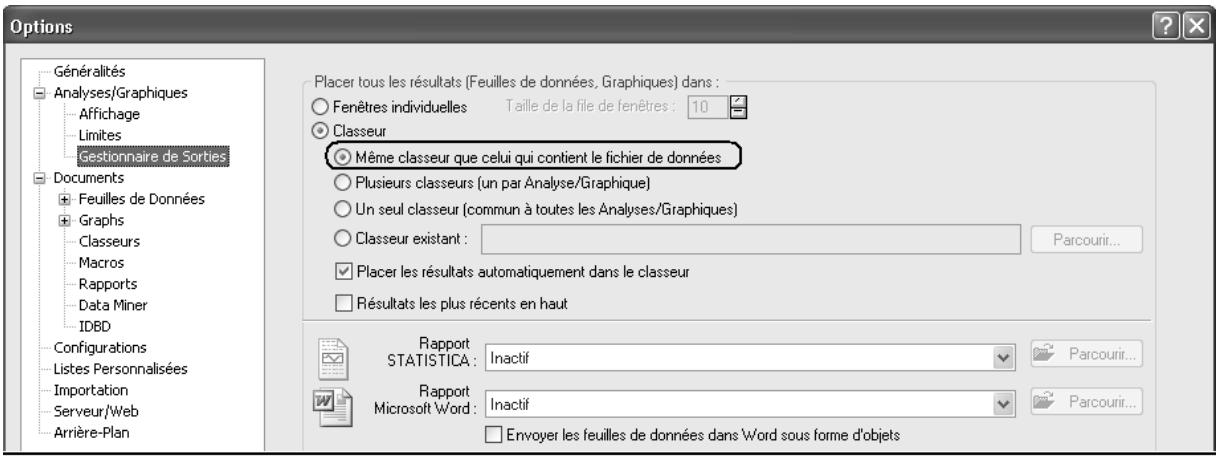
- Par défaut, Statistica affiche les noms des 16 derniers fichiers utilisés. On peut modifier ce comportement. Par exemple, on peut aussi, sans inconvénient, réduire la longueur de la liste à 10 au lieu de 16.

- L'option : *Proposer d'enregistrer les projets lors de la fermeture* peut également être déroutante si on ignore ce qu'est un fichier de projet Statistica (cf. § 1.3.3) et peut sans inconvénient être désactivée.

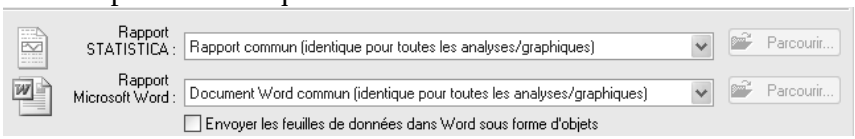


Gérer les sorties

La manière la plus commode de gérer nos documents avec Statistica consiste à rassembler dans un même classeur la ou les feuilles de données et les résultats de traitements concernant ces données. Ce comportement sera obtenu à l'aide du réglage suivant, sous l'onglet "Analyses/Graphiques - Gestionnaire de sorties" :



Il peut également être commode de demander à Statistica de placer une copie des résultats dans un rapport, en utilisant des options telles que :



En effet un rapport peut être enregistré au format .rtf pour être relu sur une autre machine par un logiciel de traitement de textes, même si Statistica n'est pas installé sur l'appareil. *Cependant, cette pratique présente plus d'inconvénients que d'avantages.* En effet :

- Les rapports produisent rapidement des fichiers très volumineux. Un rapport, ou un classeur contenant un ou des rapports devra être compressé (zippé) avant d'être envoyé par mail. Et par ailleurs, un rapport trop volumineux semble provoquer des plantages du logiciel dans certains cas.

- Si plusieurs séances de travail sont nécessaires pour réaliser le traitement, un nouveau rapport sera créé à chaque séance, ce qui est assez peu pratique.

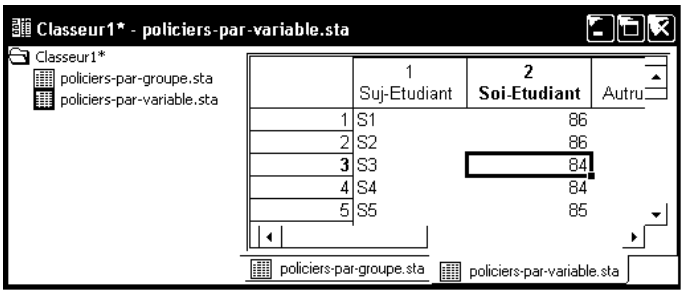
En revanche, on pourra utiliser un rapport pour y taper de courts commentaires textuels, l'interprétation du résultat d'un traitement par exemple.

Réglages concernant les graphiques

Sous l'onglet Documents > Graphs, veillez à désactiver l'option "Permettre un rendu avancé des graphiques". En effet, sur nos postes, lorsque cette option est active, Statistica se plante dès que l'on essaie de réaliser un graphique :

La feuille de données active

Les traitements demandés via les menus s'appliquent à la fenêtre de données **active**. Dans le cas de données rassemblées dans plusieurs fenêtres indépendantes, la feuille active est celle qui se trouve au premier plan sur l'écran. Dans le cas d'un classeur, la feuille active est repérée par un liseré rouge :



Dans le classeur ci-dessus, la feuille active est "policiers-par-variable.sta"

On peut rendre active une feuille, ou changer de feuille active :

- soit en cliquant sur l'icône de la feuille et en utilisant le menu : Classeur - Feuille de données active;

- soit en cliquant avec le bouton droit sur l'icône de la feuille et en utilisant l'item "Feuille de données active" du menu local.

N.B. Il faut parfois rendre inactive la feuille actuellement active (mêmes menus que ci-dessus) pour pouvoir en activer une autre.

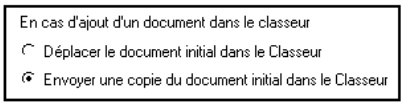
2.1.3.4 Manipulations de base sur un classeur

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".



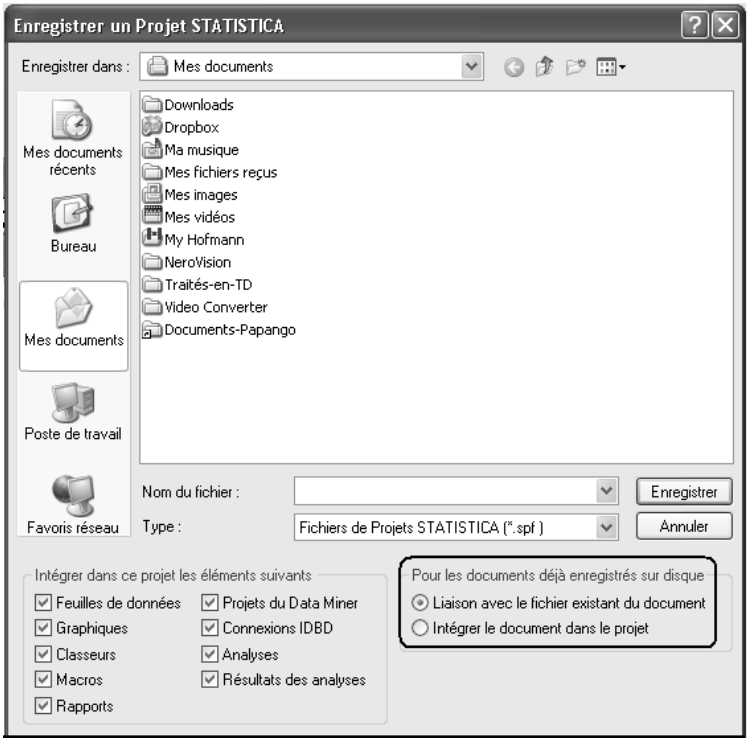
Supprimer un objet d'un classeur

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.

Qu'est ce qu'un fichier de "Projet"

On peut enregistrer un projet soit en réponse à une fenêtre de dialogue, si l'option : *Proposer d'enregistrer les projets lors de la fermeture* est active soit en utilisant le menu Fichier - Enregistrer le projet....

Un fichier de projet permet de mémoriser un "instantané" au cours d'une séance de travail : feuilles de données et de résultats, analyses actives, etc. Mais, avec les options par défaut, le fichier lui-même ne contient pas les données proprement dites, il contient seulement des liens vers les classeurs, feuilles de données, feuilles de graphiques, etc. comme le précise la fenêtre de dialogue d'enregistrement :



Avec l'option par défaut, ce format de fichier ne permet donc pas de recopier votre travail sur un autre compte, ou de transmettre votre travail à un autre utilisateur. Le logiciel enverra également des messages d'erreur si vos fichiers sont déplacés ou renommés après l'enregistrement du projet.

2.1.3.5 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.

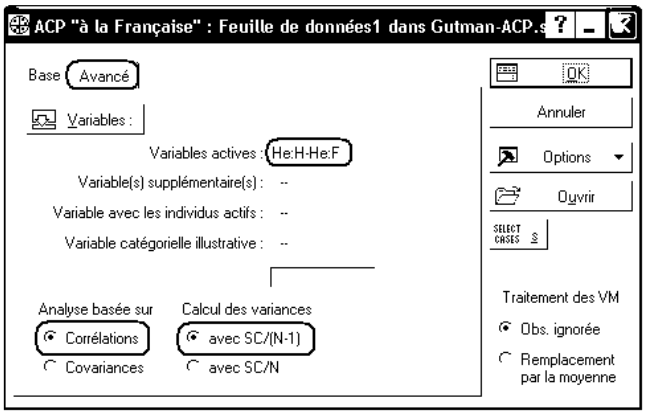
Nous partons ici d'un tableau de données comprenant, pour chacune des 8 conditions expérimentales, les moyennes de chaque trait calculées sur les 70 réponses obtenues dans chacune des conditions expérimentales. On retrouve, dans le tableau ci-dessous, le rang (solidarisation des variables) de chacun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

2.1.3.6 Statistiques descriptives - Matrice des corrélations

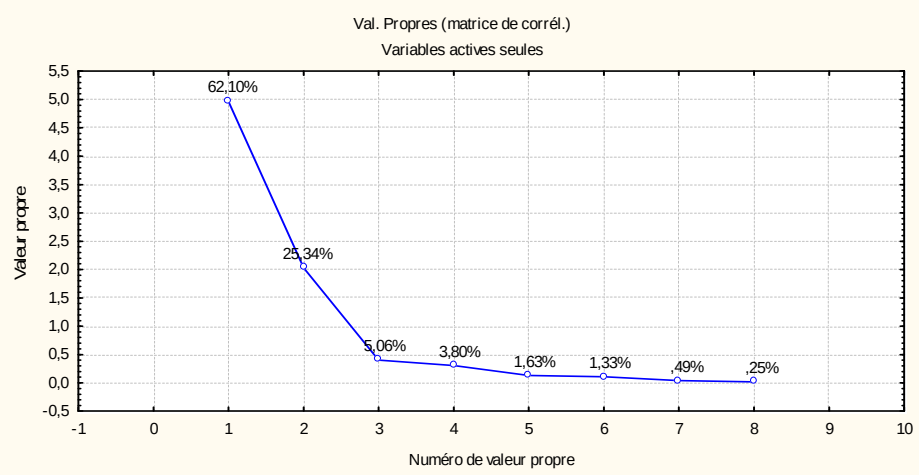
Ces résultats peuvent être obtenus à l'aide de l'onglet "Descriptives".

	Corrélations (Repr-Soc-Homo dans Rep-Soc-Homo.stw)							
Variable	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

2.1.3.7 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".

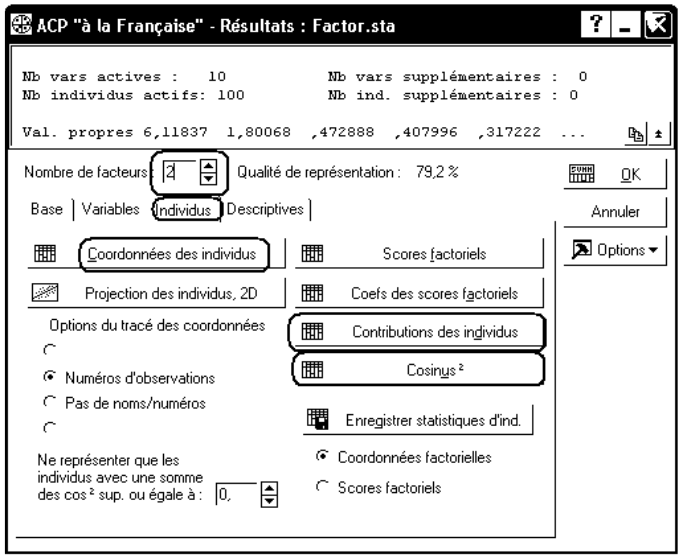
Valeur numéro	Val. Propres (matrice de corrél.) & stat. associées Variables actives seules			
	Val. propr	% Total variance	Cumul Val. propr	Cumul %
1	4,9682	62,1026	4,9682	62,10
2	2,0268	25,3355	6,9950	87,44
3	0,4045	5,0562	7,3995	92,49
4	0,3038	3,7979	7,7034	96,29
5	0,1308	1,6346	7,8341	97,93
6	0,1064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

2.1.3.8 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

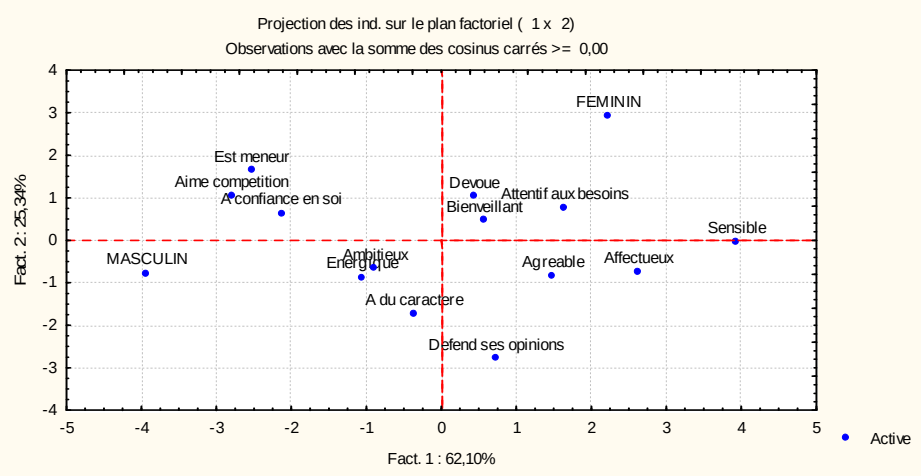
Individus	Contributions des ind	
	Fact. 1	Fact. 2
Est meneur	9,18	9,35
Aime competition	11,24	3,75
FEMININ	7,18	30,24
A confiance en soi	6,53	1,42
Devoue	0,28	3,67
MASCULIN	22,09	2,14
Bienveillant	0,47	0,71
Attentif aux besoins	3,91	2,03
Energique	1,60	2,70
Ambitieux	1,15	1,59
Sensible	22,34	0,00
Agreable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

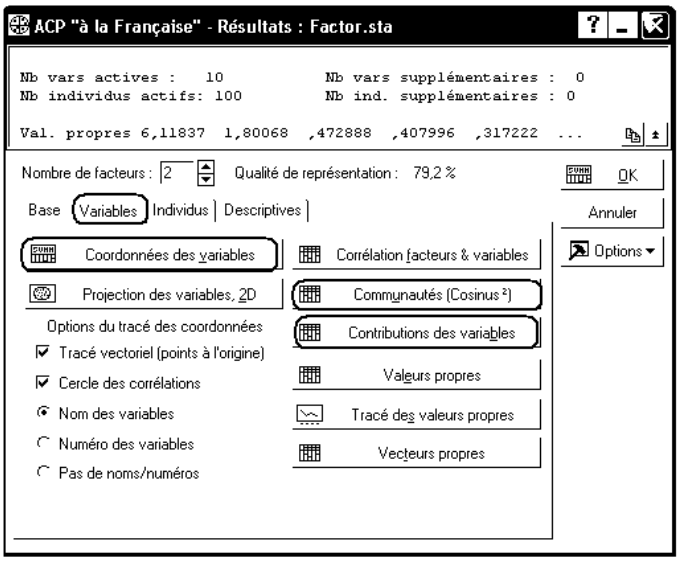
2.1.3.9 Résultats relatifs aux variables

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

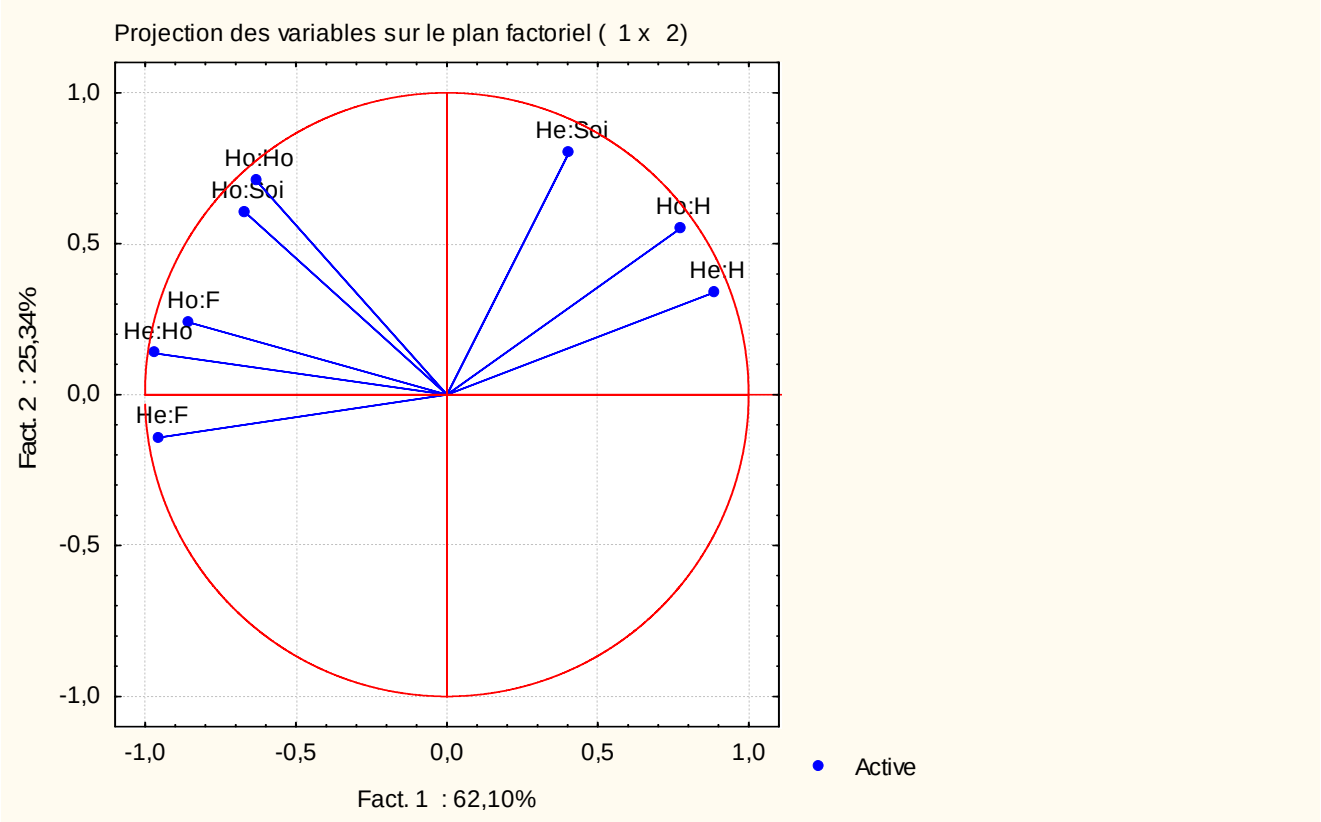
Variable	Contributions des var	
	Fact. 1	Fact. 2
He:H	0,1581	0,0566
Ho:H	0,1207	0,1502
He:Soi	0,0330	0,3168
Ho:Soi	0,0904	0,1808
Ho:Ho	0,0803	0,2482
Ho:F	0,1458	0,0281
He:Ho	0,1879	0,0091
He:F	0,1838	0,0101

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.

2.1.3.10 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".

	Vecteurs propres de la matrice de corrélation (Repr-Soc-Homo dans Rep-Soc-Homo.stw) Variables actives seules							
Variable	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

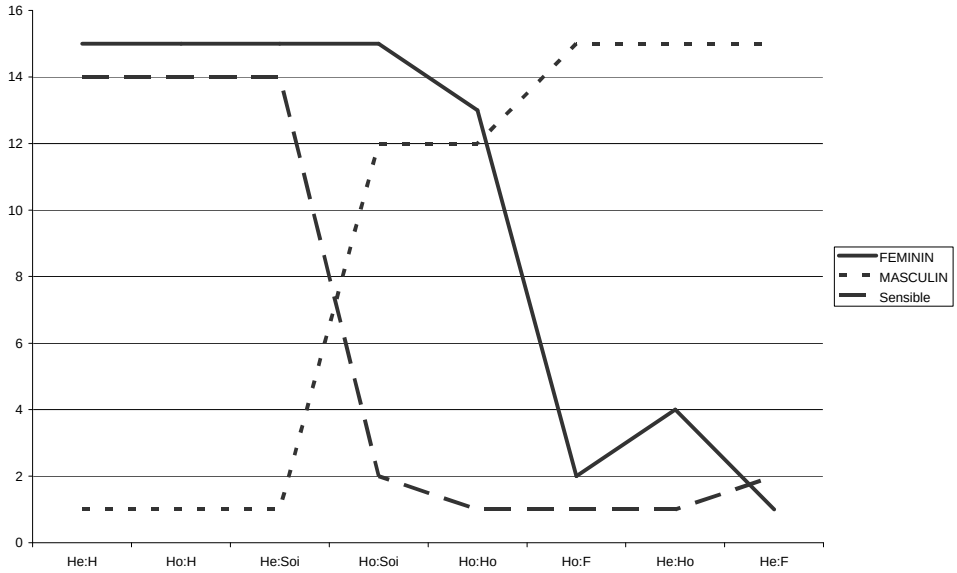
voisines, des groupes de variables qui "s'opposent", des groupes de variables relativement indépendantes entre eux.

- 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 :

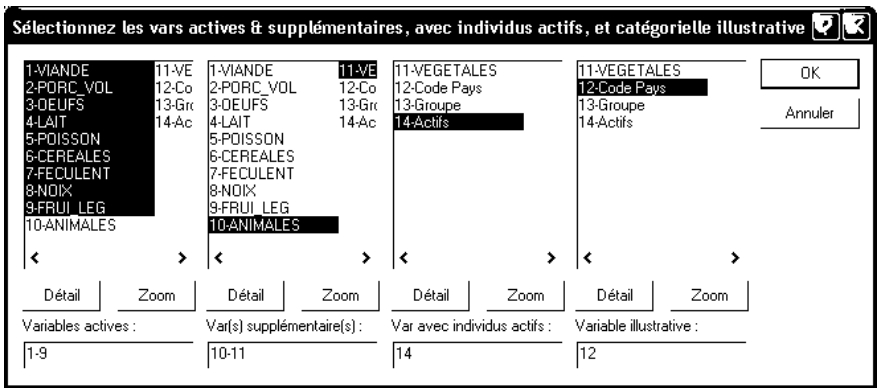
	Evaluation des consommations de protéïnes, en grammes/habitant/jour								
	1 VIANDE	2 ORC_VO	3 OEUFS	4 LAIT	5 'OISSON	6 EREALE	7 ECULEN	8 NOIX	9 RUI_LE
Belgique/Lux.	13,5	9,3	4,1	17,5	4,5	26,6	5,7	2,1	4,0
Bulgarie	7,8	6,0	1,6	8,3	1,2	56,7	1,1	3,7	4,2
Tchécoslovaquie	9,7	11,4	2,8	12,5	2,0	34,3	5,0	1,1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

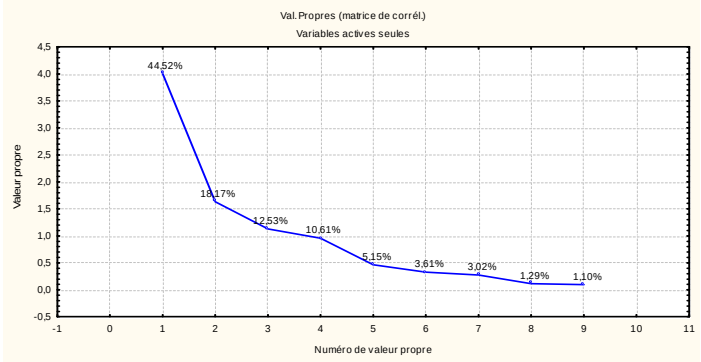
- Faire une ACP sur les valeurs réduites (ACP calculée à partir du tableau des corrélations). Dans ce cas, l'étude "gomme" les inégalités des apports protéiques des différentes sources.

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

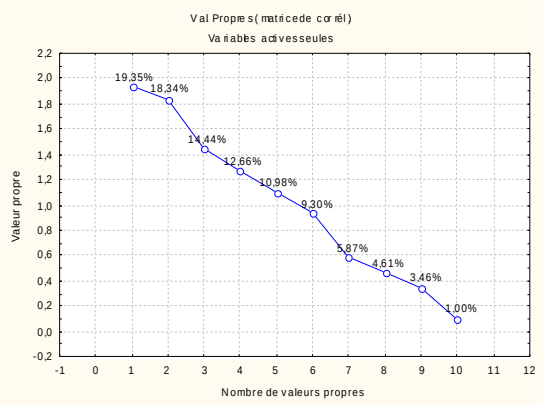
Par construction, les composantes principales sont des abstractions mathématiques et ne possèdent pas nécessairement de signification intuitive. Après avoir réalisé l'ACP, il peut parfois être intéressant de définir d'autres variables en effectuant une combinaison linéaire des composantes principales 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 Combiner description et prédiction : Analyse factorielle

2.2.1 Introduction

Le terme *analyse factorielle* (factor analysis ou FA) désigne un ensemble de techniques dont les origines peuvent être situées dans les travaux de Pearson (1901). Elle a été tout d'abord développée par des psychologues, sans que les justifications théoriques, au niveau statistique ne soient clairement établies et a donné lieu à diverses controverses entre psychologues. C'est pourquoi on a pu parler à son sujet de "mouton noir des statistiques". Ce n'est que plus tard, vers 1940 que les fondements théoriques, au niveau statistique, ont été établis pour certaines des variantes de l'analyse factorielle.
Quelques noms associés à ces méthodes : Spearman, Thomson, Thurstone, Burt, etc.

Comme l'ACP, l'analyse factorielle s'applique à des protocoles multivariés, c'est-à-dire des tableaux décrivant *n* sujets à l'aide de *p* variables numériques. Quelques remarques :

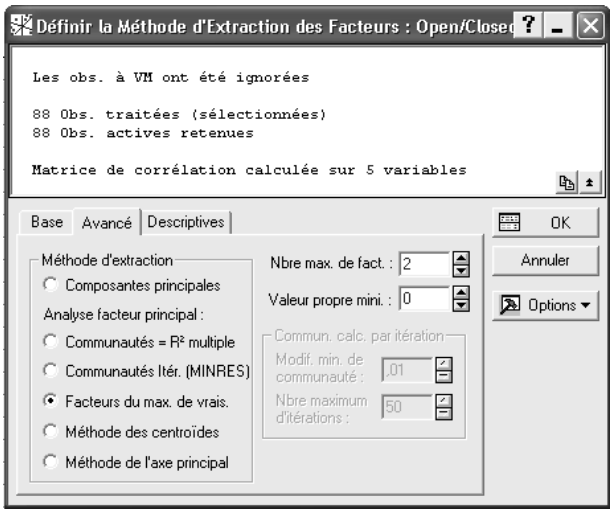
- l'intérêt porte ici sur les variables et non sur les individus statistiques ; il s'agit donc plus d'une méthode d'analyse multivariée que d'une méthode d'analyse multidimensionnelle.
- de nombreuses variantes existent : l'analyse factorielle est parfois désignée par le terme "analyse en facteurs communs et spécifiques", selon les variantes on parlera *d'analyse factorielle exploratoire* (exploratory factor analysis ou EFA) ou *d'analyse factorielle confirmatoire* (confirmatory factor analysis ou CFA). *L'analyse en facteurs principaux* (principal factor analysis ou PFA) est l'une des variantes de l'analyse factorielle.

2.2.2 Exemple introductif

Source : Mardia, K.V., Kent, J.T., Bibby, J.M., Multivariate Analysis, Academic Press, London 1979.

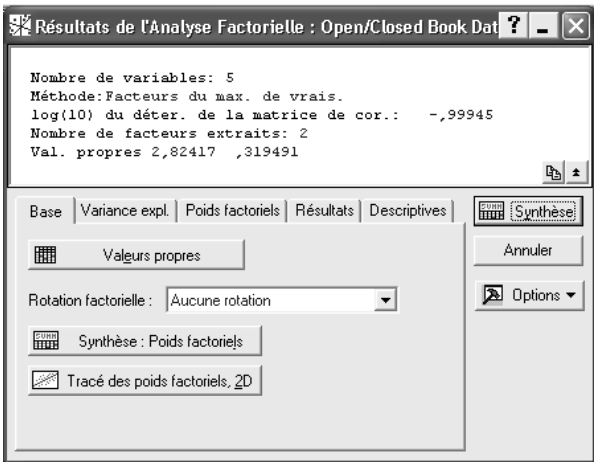
On dispose des notes obtenues par 88 sujets dans 5 matières : Mechanics(C), Vectors(C), Algebra(O), Analysis(O), Statistics(O). Pour deux matières, les étudiants n'avaient pas accès à leurs documents (closed book - C), pour les trois autres, les documents pouvaient être consultés (open book - O).

On utilise le menu Statistiques - Statistiques exploratoires multivariées - Analyse Factorielle de Statistica. Sous l'onglet "Avancé", on obtient le dialogue suivant :



Nous voyons que Statistica nous demande de fixer a priori le nombre de facteurs à extraire et nous propose plusieurs méthodes d'extraction des facteurs. Choisissons d'extraire deux facteurs par la méthode du maximum de vraisemblance.

Statistica fournit alors les résultats sous plusieurs onglets :



Sous l'onglet "Variance expliquée", on obtient notamment les 4 tableaux de résultats suivants :

- un tableau de "valeurs propres" :

Val. Propres (Open/Closed Book Data)				
Extraction : Facteurs du max. de vrais.				
	Val Propre	% Total variance	Cumul Val propre	Cumul %
1	2,824170	56,48341	2,824170	56,48341
2	0,319491	6,38983	3,143662	62,87323

- un tableau des "communautés" :

Communautés (Open/Closed Book) Rotation : Sans rot.			
	Pour 1 Facteur	Pour 2 Facteurs	R-deux Multiple
Mechanics(C)	0,394878	0,534103	0,376414
Vectors(C)	0,483548	0,580944	0,445122
Algebra(O)	0,808935	0,811431	0,671358
Analysis(O)	0,607779	0,648207	0,540864
Statistics(O)	0,529029	0,568977	0,479319

- un test d'adéquation du modèle aux données, utilisant une statistique du khi-2

Qualité d'ajust.,2 (Open/Closed Book Data) (Test de la nullité des éléments en dehors de la diagonale dans la matrice de corr.)				
	% expl.	Chi²	dl	p
Résultat	62,87323	0,074710	1	0,784601

- un tableau dit "de corrélation des résidus" :

Corrélations des Résidus (Open/Closed Book Data) (Résidus marqués sont > ,100000)					
	Mechanics(C)	Vectors(C)	Algebra(O)	Analysis(O)	Statistics(O)

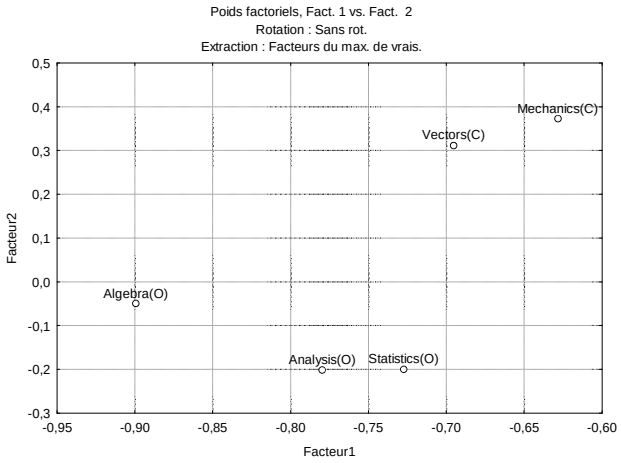
Mechanics(C)	0,47	-0,00	0,00	-0,01	0,01
Vectors(C)	-0,00	0,42	-0,00	0,01	-0,01
Algebra(O)	0,00	-0,00	0,19	-0,00	0,00
Analysis(O)	-0,01	0,01	-0,00	0,35	-0,00
Statistics(O)	0,01	-0,01	0,00	-0,00	0,43

L'onglet "Poids factoriels" nous offre la possibilité de transformer les facteurs par rotation. Il nous donne également les résultats suivants :

- les poids factoriels des variables selon chacun des facteurs :

Poids Factoriels(Sans rot.) (Open/Closed Book Data) (Poids marqués >,700000)		
	Facteur 1	Facteur 2
Mechanics(C)	-0,628393	0,373128
Vectors(C)	-0,695376	0,312083
Algebra(O)	-0,899408	-0,049958
Analysis(O)	-0,779602	-0,201066
Statistics(O)	-0,727344	-0,199869
Var. Expl.	2,824170	0,319491
Prp.Tot	0,564834	0,063898

- Le graphique correspondant :



Enfin, l'onglet "Résultats" nous fournit :

- les coefficients des scores factoriels :

Coefficients des Scores Factoriels (Open/Closed Book Data) Extraction : Facteurs du max. de vrais.		
	Facteur 1	Facteur 2
Mechanics(C)	-0,131635	0,457102
Vectors(C)	-0,161949	0,425053
Algebra(O)	-0,465496	-0,151209
Analysis(O)	-0,216280	-0,326209
Statistics(O)	-0,164691	-0,264662

- les scores factoriels des individus :

Scores Factoriels (Open/Closed Book Data) Extraction : Facteurs du max. de vrais.		
	Facteur 1	Facteur 2
1	-2,05705	0,73671
2	-2,51565	-0,00951
3	-2,09181	0,35850
4	-1,51263	0,02871
....	...	....

Comme on peut le voir, l'analyse factorielle, par certains aspects, semble ressembler à l'analyse en composantes principales. Mais qu'en est-il véritablement ?

2.2.3 Justification conceptuelle de l'analyse factorielle exploratoire

L'analyse en composantes principales est une méthode qui, à partir d'un ensemble X_1, X_2, ... X_p de variables observées corrélées entre elles permet d'obtenir un nouvel ensemble Y_1, Y_2, ... , Y_p de variables non corrélées tout en conservant la dispersion observée entre les individus. La méthode travaille sur les variances dans la mesure où Y_1 est la combinaison linéaire des X_i ayant la plus grande variance, Y_2 satisfait à la même condition tout en étant non corrélée avec Y_1, etc. L'analyse en composantes principales est essentiellement une transformation des données. C'est une méthode descriptive qui ne fait aucune hypothèse a priori sur les variables à traiter.

L'analyse factorielle est une méthode inférentielle qui vise à expliquer la matrice des covariances par un minimum, ou un petit nombre de variables hypothétiques (non observables) : les facteurs.

Par exemple, Spearman fait passer trois tests d'aptitude à un échantillon de sujets et les scores observés aux trois tests produisent la matrice de corrélation suivante :

$$\begin{bmatrix} 1 & 0,83 & 0,78 \\ 0,83 & 1 & 0,67 \\ 0,78 & 0,67 & 1 \end{bmatrix}$$

On souhaiterait étudier l'hypothèse suivante :
Les valeurs observées sont la somme de deux éléments :
- Une quantité proportionnelle à une variable ou facteur (non observable) mesurant l'intelligence du sujet
- Une quantité spécifique au test, à laquelle s'ajoute une erreur aléatoire.

Autrement dit :
- On a observé un ensemble X_1, X_2, ..., X_p de variables sur un échantillon
- On fait l'hypothèse que ces variables dépendent (linéairement) en partie de k variables non observables, ou variables latentes ou facteurs F_1, F_2, ..., F_k.

On cherche donc à décomposer les variables observées X_i (supposées centrées) de la façon suivante :

$$X_i = \sum_{r=1}^{k} l_{ir} F_r + E_i$$

ou, de façon moins formelle :

$$\text{Variable observée} = \sum \text{coeff.} \times \text{variable latente} + \text{erreur spécifique}$$

avec les conditions suivantes :

- Le nombre k de facteurs est fixé à l'avance.
- Les facteurs F_r sont centrés réduits, non corrélés entre eux
- Les termes d'erreur E_i sont non corrélés avec les facteurs
- Les termes d'erreur E_i sont non corrélés entre eux.

Remarque. Dans la formulation ci-dessus, on a choisi pour simplifier, de ne pas distinguer les paramètres observés sur l'échantillon des paramètres théoriques sur la population. Comme nous n'envisageons de développements théoriques à partir de ces équations, ce choix n'a guère d'importance.

Afin d'exploiter les conditions indiquées ci-dessus, le traitement mathématique porte sur les matrices de covariance (si les données ne sont pas réduites) ou de corrélation (si elles le sont). Notons c_{ij} la covariance des variables X_i et X_j et v_i la variance de la variable E_i.

On a les égalités :

$$c_{ii} = \sum_{r=1}^{k} l_{ir}^2 + v_i$$

$$c_{ij} = \sum_{r=1}^{k} l_{ir} l_{jr} \quad \text{si} \quad i \neq j$$

c'est-à-dire, matriciellement :

$$C = LL' + V .$$

Ce problème n'admet en général pas une solution unique. On ajoute alors une condition supplémentaire telle que :

$$J = L'V^{-1}L \quad \text{est diagonale}$$

Mais, toute rotation des facteurs ainsi déterminés fournit également aussi une solution.

Vocabulaire : les coefficients l_{ir} sont appelés *poids factoriels* (loadings) des variables sur les facteurs. La quantité $h_i^2 = \sum_{r=1}^{k} l_{ir}^2$ qui représente la partie de la variance de X_i due aux facteurs et donc "partagée" avec les autres variables est appelée communauté (*communality*).

Remarque. L'analyse factorielle n'exige pas que les données de départ soient centrées et réduites. Pour certaines méthodes insensibles aux échelles (scale free) les résultats ne dépendent pas d'une éventuelle réduction des données. Il importe par ailleurs de remarquer que, lorsque les données sont centrées réduites, les poids factoriels sont les coefficients de corrélation entre les facteurs et les variables, et la communauté d'une variable représente le carré du coefficient de corrélation multiple de cette variable par rapport aux facteurs.

2.2.4 Méthodes d'extraction des facteurs

Comme nous le montre Statistica, plusieurs méthodes d'extraction des facteurs ont été proposées et fournissent des résultats analogues, mais pas identiques.

2.2.4.1 Analyse en composantes principales

Une première méthode (souvent appelée PCA, *principal component analysis* dans les ouvrages anglo-saxons) utilise les valeurs propres et la diagonalisation des matrices. Les résultats sont alors identiques à ceux obtenus par ACP normée, se limitant à k axes. La différence la plus importante par rapport à l'ACP est la possibilité d'effectuer une rotation des facteurs.

2.2.4.2 Méthode de l'axe principal

La méthode de l'axe principal (PFA, principal factor analysis ou PAF, principal axis factoring) est une méthode itérative cherchant à maximiser les communautés. Les estimations initiales des communautés sont les coefficients de corrélation multiple de chaque variable par rapport à toutes les autres.

2.2.4.3 L'analyse factorielle du maximum de vraisemblance

Notion de vraisemblance d'une valeur d'un paramètre :

On cherche à répondre à des questions du type : "Etant donné des résultats observés sur un échantillon, est-il vraisemblable qu'un paramètre donné de la population ait telle valeur ?".

Exemple 1 : (variable discrète) Lors d'un référendum, on interroge trois personnes. Deux déclarent voter "oui", la troisième déclare voter "non".
Au vu de ces observations, laquelle de ces deux hypothèses est la plus vraisemblable :

- *Le résultat du référendum sera 40% de "oui"*
- *Le résultat du référendum sera 60% de "oui".*

Solution. Si le résultat du référendum est de 40% de "oui", la probabilité d'observer trois personnes votant respectivement "oui", "oui" et "non" est : P1 = 0,4x0,4x0,6 = 0,096. Si le résultat du référendum est de 60% de oui, la même probabilité est : P2 = 0,6x0,6x0,4 = 0,144. La seconde hypothèse est donc plus vraisemblable que la première.

Exemple 2 (variable continue) Lors d'un test effectué sur un échantillon de 5 sujets, on a observé les scores suivants :

90, 98, 103, 107, 112.

Deux modèles sont proposés pour représenter la distribution des scores dans la population parente :

- *La loi normale de moyenne 100 et d'écart type 15*
- *La loi normale de moyenne 102 et d'écart type 10.*

Quel est le modèle le plus vraisemblable ?

Dans le cas d'une variable continue, on utilise la valeur de la distribution de la loi théorique au lieu de la probabilité de la valeur observée. La vraisemblance associée à chaque hypothèse, calculée à l'aide d'Excel, est donc :

Obs	*Modèle 1*	*Modèle 2*
90	*0,02130*	*0,01942*
98	*0,02636*	*0,03683*
103	*0,02607*	*0,03970*
107	*0,02385*	*0,03521*
112	*0,01931*	*0,02420*
Vraisemblance	*6,74E-09*	*2,42E-08*

On voit que le modèle 2, dont la vraisemblance est de 2,42 10^{-8} est plus vraisemblable que le modèle 1.

Estimation du maximum de vraisemblance

L'estimation du maximum de vraisemblance (EMV, maximum likelihood estimation ou MLE dans les ouvrages anglo-saxons) est la valeur du paramètre pour laquelle la vraisemblance est maximum. Reprenons l'exemple du référendum.

Si le pourcentage de "oui" est p, la probabilité d'observer trois personnes votant respectivement "oui", "oui" et "non" est : $P = p^2(1-p)$. La dérivée de cette fonction est $P' = p(2 - 3p)$. Cette dérivée s'annule pour p=2/3=0,67, et cette valeur correspond à un maximum de P. Ainsi, au vu des observations, le résultat le plus vraisemblable est : 67% de "oui" ... ce qui n'est guère surprenant.

On notera que les calculs de vraisemblance sont souvent multiplicatifs et conduisent à des nombres très proches de 0. C'est pourquoi on utilise généralement la fonction L, opposée du logarithme de la vraisemblance. Dans le cas précédent on aurait ainsi :

$$L = - \ln P = - 2 \ln p - \ln(1 - p).$$

La recherche de l'estimation du maximum de vraisemblance revient alors à chercher le minimum de cette fonction.

Méthode du maximum de vraisemblance

La méthode du maximum de vraisemblance est la seule qui permette de calculer un test statistique d'adéquation du modèle.

Dans cette méthode, on fixe a priori un nombre k de facteurs à extraire. Les poids factoriels des variables sur les différents facteurs sont alors déterminés de manière à optimiser une fonction de vraisemblance.

Cette méthode utilise des concepts de statistique inférentielle classiques. Mais elle suppose que les données vérifient des propriétés de régularité convenables. La condition d'application est la multinormalité des variables X_i sur la population parente de l'échantillon observé. Certains auteurs expriment cette condition en termes d'asymétrie et d'aplatissement des distributions observées.

Un test statistique permet d'évaluer la validité du résultat. Selon Lawley et Maxwell, les hypothèses H_0 et H_1 du test sont :

H_0 : Il y a exactement *k* facteurs communs.
H_1 : Plus de *k* facteurs sont nécessaires.

La statistique utilisée dépend évidemment des covariances des X_i et des poids factoriels obtenus. Elle dépend également de la taille de l'échantillon tiré. Elle suit approximativement une loi du khi-2 avec $\frac{1}{2}\left[(p-k)^2-(p+k)\right]$ degrés de liberté (p : nombre de variables, k : nombre de facteurs extraits).
Selon Lawley et Maxwell, si le khi-2 trouvé excède la valeur critique correspondant au niveau de significativité choisi, H0 est rejetée, et il faut considérer au moins k+1 facteurs dans le modèle.

Remarques.

1. On doit avoir $(p+k)<(p-k)^2$ ce qui limite le nombre de facteurs.

2. Certains auteurs énoncent une règle en termes de taille des échantillons pour utiliser cette statistique. Par exemple, Mardia et Kent indiquent : $n \geq p + 50$.

3. Cette statistique peut être utilisée pour déterminer le nombre de facteurs à extraire. On calcule alors la statistique pour k=1, k=2, ... L'extraction d'un facteur supplémentaire se traduit par une diminution de la valeur de la statistique, mais également par une diminution du nombre de degrés de liberté. La p-value correspondante n'est donc pas nécessairement améliorée par l'augmentation du nombre de facteurs. On choisit ensuite le nombre de facteurs qui conduit à la meilleure p-value (celle qui est la plus proche de 1).

4. Cette statistique est malheureusement très sensible à la taille de l'échantillon.

2.2.5 Résultats obtenus - Scores des individus

2.2.5.1 Poids factoriels et communautés

Les résultats obtenus sont essentiellement constitués des poids factoriels des variables sur les différents facteurs et des communautés des différentes variables. Sur l'exemple donné en introduction, les poids factoriels sont donnés par :

Poids Factoriels(Sans rot.) (Open/Closed Book Data) (Poids marqués >,700000)		
	Facteur 1	Facteur 2
Mechanics(C)	-0,628393	0,373128
Vectors(C)	-0,695376	0,312083
Algebra(O)	-0,899408	-0,049958
Analysis(O)	-0,779602	-0,201066
Statistics(O)	-0,727344	-0,199869
Var. Expl.	2,824170	0,319491
Prp.Tot	0,564834	0,063898

On cherche alors à attribuer une signification à chacun des facteurs. Sur notre exemple, toutes les variables sont fortement corrélées (négativement) avec le premier facteur, qui peut ainsi apparaître comme une mesure "globale" relative à l'individu. Quant au deuxième facteur, il oppose les matières évaluées à livre fermé (poids factoriels positifs) à celles évaluées à livre ouvert (poids factoriels négatifs). On pourra parler de facteur *unipolaire* dans le premier cas, de facteur *bipolaire* dans le second.

Comme nous l'avons souligné plus haut, les facteurs ne sont pas déterminés de manière unique, et notamment, toute transformation des facteurs par rotation orthogonale conduit à une autre solution. Il peut intéressant d'effectuer une telle rotation pour obtenir des facteurs plus faciles à interpréter. C'est ce que nous ferons un peu plus loin.

Dans l'exemple traité en introduction les communautés sont les suivantes :

Communautés (Open/Closed Book) Rotation : Sans rot.			
	Pour 1 Facteur	Pour 2 Facteurs	R-deux Multiple
Mechanics(C)	0,394878	0,534103	0,376414
Vectors(C)	0,483548	0,580944	0,445122
Algebra(O)	0,808935	0,811431	0,671358
Analysis(O)	0,607779	0,648207	0,540864
Statistics(O)	0,529029	0,568977	0,479319

Ces quantités se calculent facilement à partir du tableau des poids factoriels. Par exemple, pour la variable Mechanics(C), la communauté se calcule de la manière suivante :

$$h_1^2 = (-0{,}628393)^2 + (0{,}373128)^2 = 0{,}534103$$

Pour une ACP, ces quantités sont interprétées en termes de qualité de représentation, ou de déformation due à la projection. Dans le cadre de l'analyse factorielle, elles nous indiquent quelle est la part de variabilité de chacune des variables observées qui participe à la variance "commune" et, par différence, quelle est la part qui est spécifique à chaque variable, et donc non prise en compte dans le modèle factoriel. Par exemple, pour la variable Algebra(O), la part "commune" est de 81% et la part spécifique, non prise en compte par les facteurs est de 19%.

2.2.5.2 Scores des individus

Les valeurs prises par les différents facteurs (qui sont des variables statistiques, même si elles ne sont pas observables directement) sur les individus statistiques composant l'échantillon sont appelées *scores des individus*. Contrairement à l'ACP, l'exploitation des résultats d'une analyse factorielle n'utilise généralement pas ces scores. En effet, les facteurs ne prennent pas en compte la totalité de la variation observée sur les données et celles-ci comportent une part de variation aléatoire due aux fluctuations d'échantillonnage. Les scores des individus ne peuvent donc pas être calculés de manière exacte mais seulement estimés à partir des autres résultats. Plusieurs méthodes ont été proposées, par exemple une méthode basée sur le maximum de vraisemblance a été proposée par Bartlett : le *Bartlett factor score*. La justification de ces méthodes approchées est particulièrement délicate lorsqu'on travaille sur les corrélations et non sur les covariances.

Dans l'exemple donné en introduction, Statistica nous donne d'une part l'expression des facteurs en fonction des variables :

Coefficients des Scores Factoriels (Open/Closed Book Data) Extraction : Facteurs du max. de vrais.		
	Facteur 1	Facteur 2
Mechanics(C)	-0,131635	0,457102
Vectors(C)	-0,161949	0,425053
Algebra(O)	-0,465496	-0,151209
Analysis(O)	-0,216280	-0,326209
Statistics(O)	-0,164691	-0,264662

Ainsi, par exemple :

$$\text{Facteur}\,1 = -0{,}132 \times \text{Mechanics} - 0{,}162 \times \text{Vectors} - 0{,}465 \times \text{Algebra} - 0{,}216 \times \text{Analysis} - 0{,}165 \times \text{Statistics}$$

D'autre part, il donne également les valeurs des facteurs sur les différentes observations, telles qu'elles peuvent être calculées à partir des formules précédentes et des valeurs centrées réduites associées aux valeurs observées. Par exemple pour le premier sujet, le logiciel indique :

	Facteur 1	Facteur 2
1	-2,05705	0,73671

Les valeurs centrées réduites des 5 variables sont :

	Mechanics(C)	Vectors(C)	Algebra(O)	Analysis(O)	Statistics(O)
1	2,17573873	2,38907869	1,54334732	1,36866891	2,24235647

Et on vérifie que :

$$\text{Facteur}\,1_{\text{Sujet}\,1} = -0{,}132 \times 2{,}176 - 0{,}162 \times 2{,}390 - 0{,}465 \times 1{,}543 - 0{,}216 \times 1{,}369 - 0{,}165 \times 2{,}242 = -2{,}057$$

Remarque. A l'exception des scores factoriels des individus, l'ensemble des résultats d'une analyse factorielle peut être obtenu à partir de la matrice des corrélations (ou des covariances) des variables, et de la taille de l'échantillon. C'est pourquoi Statistica propose de deux formats pour les données d'entrée : données brutes ou matrice de corrélations.

2.2.6 Rotation des facteurs : rotations orthogonales, rotations obliques

Les facteurs extraits par l'une ou l'autre des méthodes précédentes ne sont pas déterminés de manière unique et c'est généralement une condition arbitraire qui permet de choisir une solution dans l'ensemble des solutions possibles.
Il en résulte que les facteurs ainsi produits ne sont pas toujours simples à interpréter. Mais toute rotation sur les facteurs produit une autre solution et on peut être tenté de rechercher une solution qui "fasse sens", c'est-à-dire qui produise des facteurs plus simples à interpréter.

Il importe de noter que la transformation par rotation n'affecte pas l'adéquation du modèle aux données. Les communautés, notamment, restent les mêmes. Mais les solutions avant ou après rotation peuvent être interprétés de façon notablement différente.

Ainsi, sur notre exemple :

	Poids Factoriels (sans rotation)		Poids Factoriels (après rotation varimax normalisé)	
	Facteur 1	Facteur 2	Facteur 1	Facteur 2
Mechanics(C)	-0,628393	0,373128	0,270028	0,679108
Vectors(C)	-0,695376	0,312083	0,360346	0,671636
Algebra(O)	-0,899408	-0,049958	0,742939	0,509384
Analysis(O)	-0,779602	-0,201066	0,740267	0,316563
Statistics(O)	-0,727344	-0,199869	0,698141	0,285615
Var. Expl.	2,824170	0,319491	1,790119	1,353543
Prp.Tot	0,564834	0,063898	0,358024	0,270709

On examine les poids factoriels après rotation varimax. Les trois matières évaluées à livre ouvert sont alors fortement corrélées avec le premier facteur, alors que le second facteur correspond aux deux matières évaluées à livre fermé et dans une moindre mesure à l'algèbre.

La rotation la plus fréquemment utilisée est la rotation varimax (Kaiser 1958). L'effet produit par une telle rotation est généralement le suivant : pour chaque facteur, les poids factoriels élevés concernent un nombre réduit de variables et les autres poids factoriels sont proches de 0.

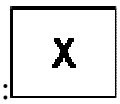
D'autres rotations ont également été proposées. Les rotations dites orthogonales produisent des facteurs non corrélés entre eux, tandis que les transformations par rotation oblique produisent de nouveaux facteurs qui peuvent être corrélés.

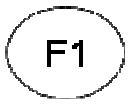
2.2.7 Analyse factorielle confirmatoire

L'analyse factorielle confirmatoire est apparentée à l'analyse factorielle exploratoire. Mais c'est aussi un cas particulier de modélisation d'équations structurelles (SEM : structural equation modelling). Différents algorithmes ont été développés dans ce cadre (par exemple : LISREL).

En analyse factorielle confirmatoire, le point de vue est différent de celui de l'analyse factorielle exploratoire : on se fixe a priori un modèle :

- nombre de facteurs
- corrélations éventuelles entre ces facteurs
- termes d'erreur attachés à chaque variable observée et corrélations éventuelles entre eux
- pour chaque facteur, variables avec lesquelles il sera significativement corrélé.

- Une variable observée est représentée dans un rectangle : 

- Une variable latente (un facteur) est représentée dans un ovale : 

- Un terme d'erreur, ou perturbation du modèle, est représenté par une variable sans cadre : 

- Une flèche entre deux variables signifie que les variations de la seconde sont dues, au moins en partie, aux variations de la première.

Exemple :

Source : pages en ligne de Michael Friendly à l'adresse :
http://www.psych.yorku.ca/lab/psy6140/fa/facfoils.htm

Calsyn et Kenny (1971) ont étudié la relation entre les aptitudes perçues et les aspirations scolaires de 556 élèves du 8è grade. Les variables observées étaient les suivantes :

Self : auto-évaluation des aptitudes
Parent : évaluation par les parents
Teacher : évaluation par l'enseignant
Friend : évaluation par les amis
Educ Asp : aspirations scolaires
Col Plan : projets d'études supérieures

Sur l'échantillon étudié, les corrélations observées entre ces six variables sont les suivantes :

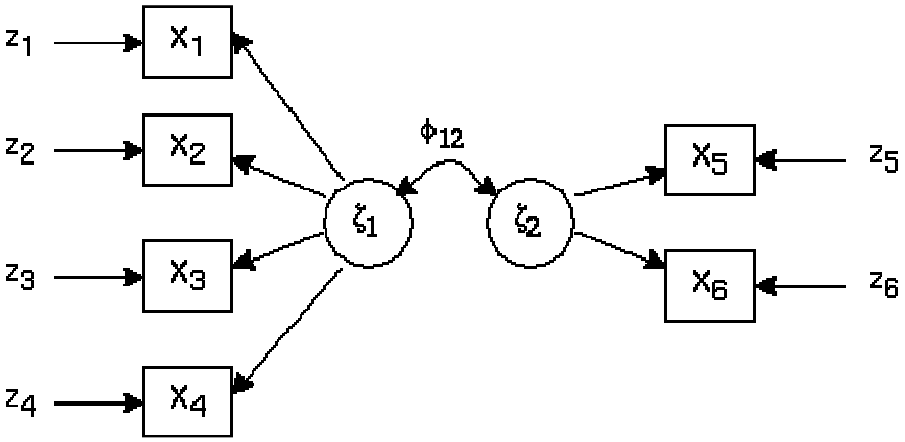
	Self	Parent	Teacher	Friend	Educ Asp	Col Plan
Self	1,00	0,73	0,70	0,58	0,46	0,56
Parent	0,73	1,00	0,68	0,61	0,43	0,52
Teacher	0,70	0,68	1,00	0,57	0,40	0,48
Friend	0,58	0,61	0,57	1,00	0,37	0,41
Educ Asp	0,46	0,43	0,40	0,37	1,00	0,72
Col Plan	0,56	0,52	0,48	0,41	0,72	1,00

Le modèle à tester fait les hypothèses suivantes :

- Les 4 premières variables mesurent la variable latente "aptitudes"
- Les deux dernières mesurent la variable latente "aspirations".

Ce modèle est-il valide ? Et, s'il en est bien ainsi, les deux variables latentes sont-elles corrélées ?

Le schéma correspondant à ce modèle peut être représenté ainsi (les variables sont renommées X_1 à X_6 et les facteurs sont désignés par la lettre grecque ζ dans ce schéma emprunté à Michael Friendly) :



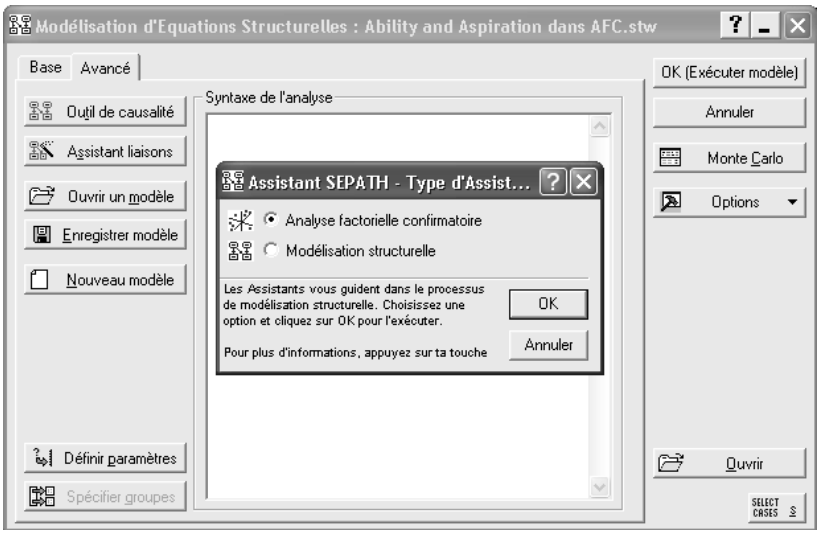
Traitement avec Statistica.

La matrice de corrélations précédente est saisie comme objet de type "matrice" de Statistica :

	Feuille de données3					
	1 Self	2 Parent	3 Teacher	4 Friend	5 Educ Asp	6 Col Plan
Self	1,00	0,73	0,70	0,58	0,46	0,56
Parent	0,73	1,00	0,68	0,61	0,43	0,52
Teacher	0,70	0,68	1,00	0,57	0,40	0,48
Friend	0,58	0,61	0,57	1,00	0,37	0,41
Educ Asp	0,46	0,43	0,40	0,37	1,00	0,72
Col Plan	0,56	0,52	0,48	0,41	0,72	1,00
Moyennes	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000
Ec-Types	1,00000	1,00000	1,00000	1,00000	1,00000	1,00000
Nb Obs.	556,00000					
Matrice	1,00000					

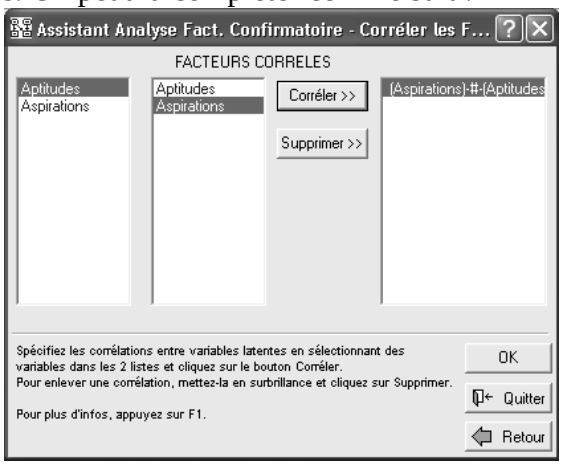
On choisit ensuite le menu Statistiques - Modèles linéaires / non linéaires avancés - Modélisation d'équations structurelles.

Sous l'onglet "Avancé", on clique sur le bouton "Assistant liaisons" et on choisit l'option "Analyse factorielle confirmatoire" :



On peut alors saisir le modèle sous la forme suivante :

Lorsqu'on clique sur le bouton OK, Statistica affiche une fenêtre permettant d'indiquer les corrélations entre les facteurs. On peut la compléter comme suit :



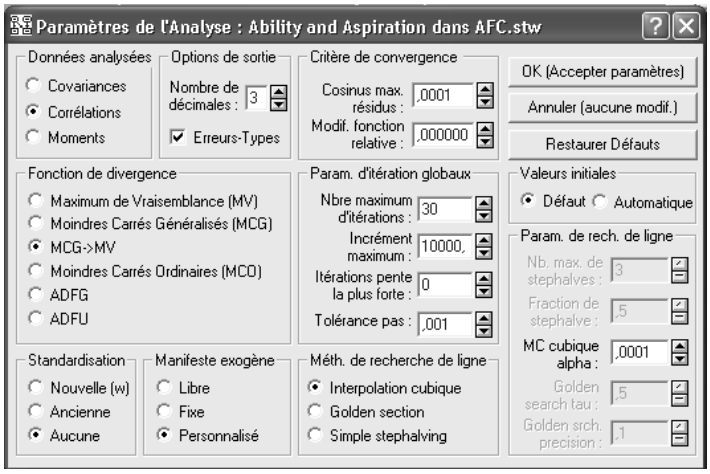
Lorsque la fenêtre suivante s'affiche, cliquer sur OK :

Le modèle spécifié est alors traduit en "langage" PATH1 sous la forme suivante :

```
(Aptitudes)-1->[Self]
(Aptitudes)-2->[Parent]
(Aptitudes)-3->[Teacher]
(Aptitudes)-4->[Friend]

(Aspirations)-5->[Educ Asp]
(Aspirations)-6->[Col Plan]

(DELTA1)-->[Self]
(DELTA2)-->[Parent]
(DELTA3)-->[Teacher]
```

```
(DELTA4)-->[Friend]
(DELTA5)-->[Educ Asp]
(DELTA6)-->[Col Plan]

(DELTA1)-7-(DELTA1)
(DELTA2)-8-(DELTA2)
(DELTA3)-9-(DELTA3)
(DELTA4)-10-(DELTA4)
(DELTA5)-11-(DELTA5)
(DELTA6)-12-(DELTA6)

(Aspirations)-13-(Aptitudes)
```

Ce "programme" peut éventuellement être enregistré dans un fichier autonome.
Cliquez ensuite sur le bouton "Paramètres de l'analyse". Le dialogue qui s'affiche est particulièrement abscons, mais nous nous contenterons d'y indiquer que les données analysées sont de type "corrélations", en laissant les autres paramètres à leurs valeurs par défaut :



Cliquez ensuite sur OK (Exécuter modèle), puis sur le bouton OK de la fenêtre suivante.

Le bouton "Synthèse du modèle" permet d'obtenir la feuille de résultats suivante :

Modèle Estimé (Ability and Aspiration dans AFC.stw)				
	Estimation Paramètre	Erreur Type	Stat. T	Niveau Proba
(Aptitudes)-1->[Self]	0,863	0,015	57,973	0,000
(Aptitudes)-2->[Parent]	0,849	0,016	54,296	0,000
(Aptitudes)-3->[Teacher]	0,805	0,018	44,287	0,000
(Aptitudes)-4->[Friend]	0,695	0,025	28,217	0,000
(Aspirations)-5->[Educ Asp]	0,775	0,026	30,279	0,000
(Aspirations)-6->[Col Plan]	0,929	0,024	39,165	0,000
(DELTA1)-->[Self]				
(DELTA2)-->[Parent]				
(DELTA3)-->[Teacher]				
(DELTA4)-->[Friend]				
(DELTA5)-->[Educ Asp]				
(DELTA6)-->[Col Plan]				
(DELTA1)-7-(DELTA1)	0,255	0,026	9,915	0,000
(DELTA2)-8-(DELTA2)	0,279	0,027	10,487	0,000
(DELTA3)-9-(DELTA3)	0,352	0,029	12,020	0,000
(DELTA4)-10-(DELTA4)	0,517	0,034	15,078	0,000
(DELTA5)-11-(DELTA5)	0,399	0,040	10,061	0,000
(DELTA6)-12-(DELTA6)	0,137	0,044	3,111	0,002

(Aspirations)-13-(Aptitudes)	0,666	0,031	21,528	0,000

On retrouve dans ce tableau le poids factoriel de chacune des variables sur le facteur spécifié par le modèle (sur une seule colonne - ce qui ne facilite pas la lecture du tableau). On y trouve également les variances des termes d'erreur DELTA1 à DELTA6 et enfin l'estimation de la corrélation entre les facteurs Aspirations et Aptitudes : 0,666.

Ces résultats seraient plus lisibles disposés de la façon (plus classique) suivante :

Modèle Estimé (Ability and Aspiration dans AFC.stw)				
	Aptitudes	Aspirations	Communauté	Spécificité
Self	0,863		0,745	0,255
Parent	0,849		0,721	0,279
Teacher	0,805		0,648	0,352
Friend	0,695		0,483	0,517
Educ Asp		0,775	0,601	0,399
Col Plan		0,929	0,863	0,137

Dans ce tableau, les communautés sont simplement les carrés des poids factoriels et les spécificités sont les compléments à 1 des communautés.

Le logiciel donne ensuite de nombreux indices évaluant la qualité du modèle.

En particulier, le bouton "Statistiques de synthèse" nous fournit la valeur d'une statistique du khi-2 du maximum de vraisemblance :

Statistiques de Synthèse (Ability and Aspiration dans AFC.stw)	
	Valeur
Chi-Deux MV	9,256
Degrés de Liberté	8,000
Niveau p	0,321

La valeur trouvée ici (p-value = 0,32) montre une bonne adéquation du modèle aux données.
D'autres indices de qualités

D'autres indices sont aussi couramment utilisés :

- AIC (Akaike Information Criterion ou Critère d'information de Akaike)
- BIC (Bayesian Information Criterion ou Critère Bayésien de Schwarz)
- TLI (Tucker-Lewis Index) : les modèles "acceptables" doivent vérifier TLI>0,90, les "bons" modèles, TLI>0,95
- RMSEA (root mean square error of approximation). les modèles "acceptables" doivent vérifier RMSEA<=0,08, les "bons" modèles, RMSEA <= 0,05
- CFI (Comparative Fit Index)

2.2.8 Bibliographie :

Ouvrages :

Lawley, D.N., Maxwell, A.E., Factor Analysis as a Statistical Method, Butterworths Mathematical Texts, England, 1963.
Mardia, K.V., Kent, J.T., Bibby, J.M., Multivariate Analysis, Academic Press, London 1979.

Articles :

Sites internet :
http://faculty.chass.ncsu.edu/garson/PA765/factor.html

Documents mis en ligne par Michael Friendly et notamment :
http://www.psych.yorku.ca/lab/psy6140/lectures/

Une discussion intéressante sur l'utilisation pratique de l'analyse factorielle :
http://core.ecu.edu/psyc/wuenschk/stathelp/EFA.htm

Pages mises en ligne par Peter Tryfos
http://www.yorku.ca/ptryfos/methods.htm

Site pour télécharger ce polycopié et les fichiers d'exemples :
http://geai.univ-brest.fr/~carpenti/

2.2.9 EFA avec Statistica sur le cas HSdata

Nous nous proposons d'utiliser Statistica pour explorer un jeu de données largement cité dans les publications, celui de Holzinger et Swineford (1939), en suivant la démarche de L. K. Muthen et B. Muthen (réf. http://www.ats.ucla.edu/stat/seminars/muthen_08)

Source : Holzinger, K. J. and Swineford, F. A. (1939). A study in factor analysis: The stability of a bi-factor solution. Supplementary Education Monographs, 48. University of Chicago.

Complete Data Set of Holzinger and Swineford's (1939) Study. Description:

A total number of 301 pupils from Paster School and Grant-White School participated in Holzinger and Swineford's (1939) study. This study consists of 26 tests, which are used to measure the subjects' spatial, verbal, mental speed, memory, and mathematical ability.

The spatial tests consist of visual, cubes, paper, flags, paperrev, and flagssub. The test 25, paper form board test (paperrev), can be used as a substitute for test 3, paper form board test (paper). The test 26, flags test (flagssub), is a possible substitute for test 4, lozenges test (flags).

The verbal tests consist of general, paragrap, sentence, wordc, and wordm.

The speed tests consist of addition, code, counting, and straight.

The memory tests consist of wordr, numberr, figurer, object, numberf, and figurew.

The mathematical-ability tests consis of deduct, numeric, problemr, series, and arithmet.

Data : a data frame with 301 observations on the following 32 variables.

id :	subject's ID number
Gender :	subject's gender
grade :	the grade the subject is on
agey :	the year part of the subject's age
agem :	the month part of the subject's age
school :	the school the subject is from
visual :	scores on visual perception test, test 1

cubes :	scores on cubes test, test 2
paper :	scores on paper form board test, test 3
flags :	scores on lozenges test, test 4
general :	scores on general information test, test 5
paragrap :	scores on paragraph comprehension test, test 6
sentence :	scores on sentence completion test, test 7
wordc :	scores on word classification test, test 8
wordm :	scores on word meaning test, test 9
addition :	scores on add test, test 10
code :	scores on code test, test 11
counting :	scores on counting groups of dots test, test 12
straight :	scores on straight and curved capitals test, test 13
wordr :	scores on word recognition test, test 14
numberr :	scores on number recognition test, test 15
figurer :	scores on figure recognition test, test 16
object :	scores on object-number test, test 17
numberf :	scores on number-figure test, test 18
figurew :	scores on figure-word test, test 19
deduct :	scores on deduction test, test 20
numeric :	scores on numerical puzzles test, test 21
problemr :	scores on problem reasoning test, test 22
series :	scores on series completion test, test 23
arithmet :	scores on Woody-McCall mixed fundamentals, form I test, test 24
paperrev :	scores on additional paper form board test, test 25
flagssub :	scores on flags test, test 26

Remarque : Holzinger and Swineford (1939) data is widely cited, but generally only the Grant-White School data is used. The present dataset contains the complete data of Holzinger and Swineford (1939).

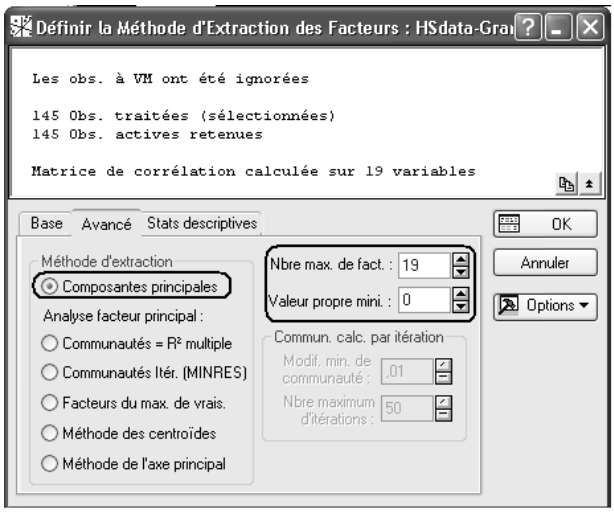
Chargez le classeur Statistica HSdata.stw et rendez active la feuille HSdata-Grant-White.

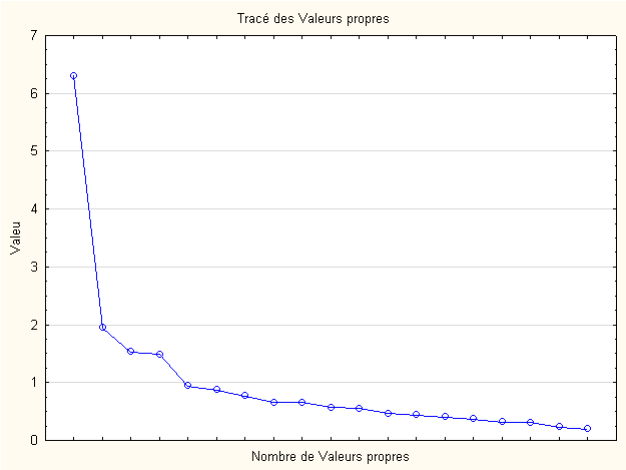
La première partie de l'étude sera menée sur les 19 variables portant les numéros de 7 à 25 (de visual à figurew). Les 19 variables correspondent à 19 tests dont le but est de mesurer la performance dans 4 domaines : les capacités au niveau spatial (v7 à v10), au niveau verbal (v11 à v15), la rapidité (v16 à v19) et la mémoire (v20 à v25).

Affichez le tableau des corrélations entre ces 19 variables.

On recherche un modèle pertinent d'analyse factorielle pour les données recueillies.

Utilisez le menu Statistiques > Techniques Exploratoires Multivariées > Analyse factorielle.
Utilisez le bouton Variables pour sélectionner les variables v7 à v25.
Calculez d'abord l'ensemble des valeurs propres en utilisant la méthode d'extraction "Composantes principales", et en ne posant pas de condition sur les valeurs propres. :





On cherche à déterminer le nombre de facteurs pertinents. De façon à pouvoir utiliser le test basé sur le khi-2, nous utilisons maintenant la méthode basée sur le maximum de vraisemblance.
Sélectionnez cette méthode et indiquez le nombre de facteurs désiré (1 puis 2, 3, 4). Le résultat du test du khi-2 est accessible sous l'onglet Variance Expliquée et le bouton "Test de la qualité d'ajustement".

	Qualité d'ajust	Nb facteurs	% expl.	Chi²	dl	p
	Résultat	1	28,99	441,19	152	1,08E-29
	Résultat	2	37,94	258,33	134	7,30E-10
	Résultat	3	43,77	175,51	117	0,00039
	Résultat	4	49,24	102,06	101	0,4518

Le choix de 4 facteurs semble pertinent. Sous l'onglet "Poids factoriels", indiquez une rotation de type "varimax normalisé" et affichez le tableau des poids en réglant la surbrillance à 0,32:

Variable	Poids Factoriels(Varimax normalisé) Extraction : Facteurs du max. de vrais. (Poids marqués >,320000)			
	Facteur 1	Facteur 2	Facteur 3	Facteur 4
visual	0,183	0,143	0,666	0,193
cubes	0,117	0,043	0,487	0,072
paper	0,191	0,126	0,455	0,170
flags	0,241	0,068	0,608	0,135
general	0,743	0,183	0,230	0,133
paragrap	0,772	0,038	0,195	0,244
sentence	0,808	0,146	0,158	0,119
wordc	0,589	0,242	0,267	0,174
wordm	0,806	0,021	0,180	0,219
addition	0,177	0,754	-0,062	0,189
code	0,197	0,486	0,190	0,367
counting	0,034	0,748	0,224	0,110
straight	0,206	0,545	0,489	0,103
wordr	0,184	0,064	0,077	0,522
numberr	0,103	0,054	0,144	0,506
figurer	0,081	0,021	0,398	0,524
object	0,155	0,228	-0,036	0,673
numberf	0,034	0,293	0,326	0,483
figurew	0,173	0,118	0,160	0,392
Var. Expl.	3,158	1,980	2,123	2,095
Prp.Tot	0,166	0,104	0,112	0,110

On peut alors décider d'éliminer les tests correspondant aux variables code, straight, figurer et numberf, qui ont des poids factoriels importants sur au moins 2 facteurs (variables ayant des "cross loadings"). On reprend alors l'analyse avec 15 variables.

Variable	Poids Factoriels(Varimax normalisé) Extraction : Facteurs du max. de vrais. (Poids marqués >,320000)			
	Facteur 1	Facteur 2	Facteur 3	Facteur 4
visual	0,204	0,127	0,583	0,164
cubes	0,114	0,031	0,521	0,018
paper	0,190	0,081	0,428	0,190
flags	0,211	0,039	0,698	0,126
general	0,730	0,140	0,283	0,150
paragrap	0,767	0,000	0,202	0,256
sentence	0,815	0,136	0,151	0,141
wordc	0,596	0,232	0,271	0,174
wordm	0,791	0,006	0,214	0,221
addition	0,197	0,696	-0,018	0,184
counting	0,027	0,808	0,269	0,121
wordr	0,161	0,040	0,087	0,574
numberr	0,071	0,050	0,185	0,549
object	0,162	0,239	-0,001	0,613
figurew	0,192	0,071	0,135	0,362
Var. Expl.	3,033	1,321	1,676	1,453
Prp.Tot	0,202	0,088	0,112	0,097

On peut alors décider d'éliminer le facteur 2, qui ne s'appuie que sur deux variables (addition et counting) et décider de faire une analyse factorielle à 3 facteurs sur 13 variables (v7 à v15, v20, v21, v23, v25) :

Variable	Poids Factoriels(Varimax normalisé) Extraction : Facteurs du max. de vrais. (Poids marqués >,320000)		
	Facteur 1	Facteur 2	Facteur 3
visual	0,219	0,557	0,186
cubes	0,112	0,533	0,009
paper	0,188	0,433	0,208
flags	0,207	0,706	0,128
general	0,728	0,305	0,161
paragrap	0,760	0,202	0,245
sentence	0,824	0,148	0,165
wordc	0,603	0,283	0,216
wordm	0,788	0,209	0,214
wordr	0,157	0,086	0,578
numberr	0,071	0,181	0,545
object	0,175	0,009	0,628
figurew	0,185	0,154	0,367
Var. Expl.	3,005	1,623	1,460
Prp.Tot	0,231	0,125	0,112

2.2.10 Modélisation d'équations structurelles avec Statistica sur le cas HSdata

On reprend l'étude de ces 13 variables (v7 à v15, v20, v21, v23, v25) à l'aide des outils de modélisation d'équations structurelles.

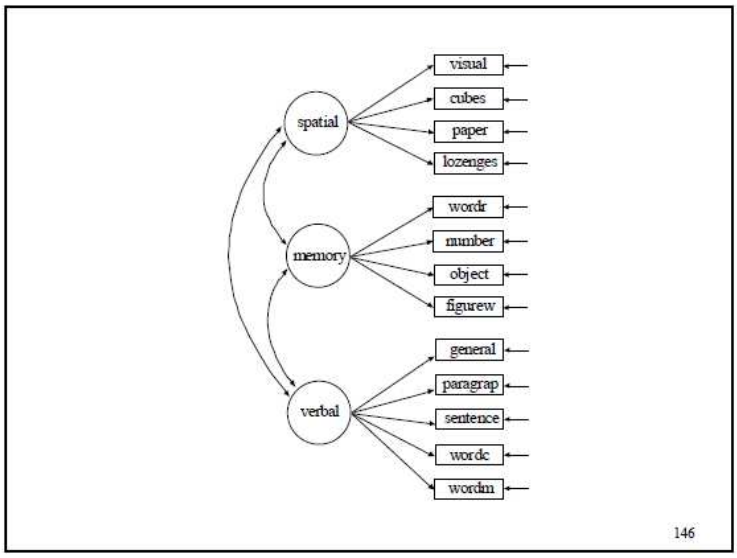
Utilisez le menu Statistiques > Modèles linéaires/non linéaires avancés > Modélisation d'équations structurelles.

Activez l'onglet Avancé puis Assistant liaisons et Analyse factorielle confirmatoire. On veut tester un modèle s'appuyant sur trois variables latentes :

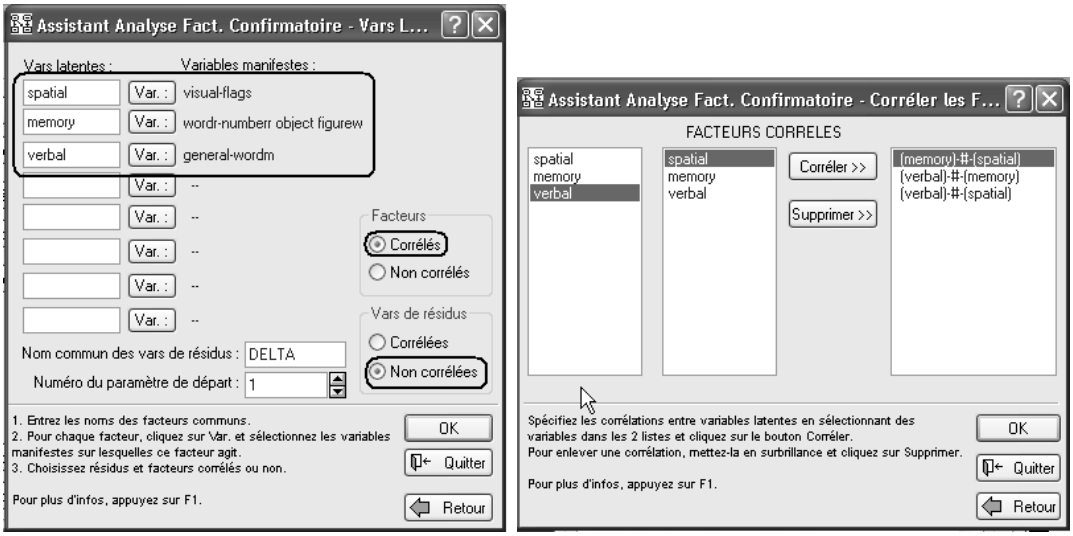
- une variable latente "spatial", mesurée par les 4 variables v7 à v10 (visual à flags)
- une variable latente "memory", mesurée par les 4 variables v20, v21, v23, v25
- une variable latente "verbal", mesurée par les 5 variables v11 à v15.

Nous supposerons que les trois variables latentes peuvent être corrélées entre elles, et que les résidus ne sont pas corrélés entre eux.

Ce modèle correspond au diagramme suivant (emprunté au site Web mentionné ci-dessus) :



Spécification du modèle sous Statistica :



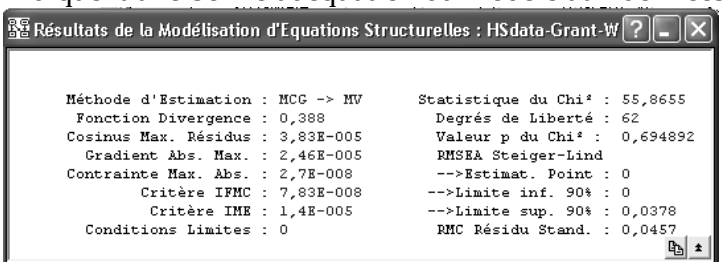
Cela conduit au programme suivant en langage PATH1 :

```
(spatial)-1->[visual]
(spatial)-2->[cubes]
(spatial)-3->[paper]
(spatial)-4->[flags]

(memory)-5->[wordr]
(memory)-6->[numberr]
(memory)-7->[object]
(memory)-8->[figurew]

(verbal)-9->[general]
(verbal)-10->[paragrap]
(verbal)-11->[sentence]
(verbal)-12->[wordc]
(verbal)-13->[wordm]

(DELTA1)-->[visual]
(DELTA2)-->[cubes]
(DELTA3)-->[paper]
(DELTA4)-->[flags]
(DELTA5)-->[wordr]
(DELTA6)-->[numberr]
```

```
(DELTA7)-->[object]
(DELTA8)-->[figurew]
(DELTA9)-->[general]
(DELTA10)-->[paragrap]
(DELTA11)-->[sentence]
(DELTA12)-->[wordc]
(DELTA13)-->[wordm]

(DELTA1)-14-(DELTA1)
(DELTA2)-15-(DELTA2)
(DELTA3)-16-(DELTA3)
(DELTA4)-17-(DELTA4)
(DELTA5)-18-(DELTA5)
(DELTA6)-19-(DELTA6)
(DELTA7)-20-(DELTA7)
(DELTA8)-21-(DELTA8)
(DELTA9)-22-(DELTA9)
(DELTA10)-23-(DELTA10)
(DELTA11)-24-(DELTA11)
(DELTA12)-25-(DELTA12)
(DELTA13)-26-(DELTA13)

(memory)-27-(spatial)
(verbal)-28-(memory)
(verbal)-29-(spatial)
```

Cliquez ensuite sur le bouton "Définir paramètres". Dans le bloc "Données analysées", cliquez sur "Corrélations" :

Cliquez ensuite sur le bouton "OK (Exécuter modèle)".
Les premiers résultats indiquent une bonne adéquation du modèle aux données :



En effet, la statistique du chi-2 indique une p-value de 69% et l'indice RMSEA est très proche de 0.
Le bouton "Synthèse du modèle" produit le tableau suivant (tableau dont on a éliminé les lignes vides) :

	Modèle Estimé			
	Estimation Paramètre	Erreur Type	Stat. T	Niveau Proba
(spatial)-1->[visual]	0,658	0,067	9,885	0,000
(spatial)-2->[cubes]	0,492	0,078	6,308	0,000
(spatial)-3->[paper]	0,530	0,075	7,026	0,000
(spatial)-4->[flags]	0,714	0,063	11,258	0,000
(memory)-5->[wordr]	0,604	0,077	7,815	0,000
(memory)-6->[numberr]	0,558	0,080	7,001	0,000
(memory)-7->[object]	0,624	0,076	8,161	0,000
(memory)-8->[figurew]	0,450	0,086	5,243	0,000
(verbal)-9->[general]	0,806	0,034	23,383	0,000
(verbal)-10->[paragrap]	0,822	0,032	25,300	0,000
(verbal)-11->[sentence]	0,834	0,031	26,910	0,000
(verbal)-12->[wordc]	0,691	0,048	14,356	0,000
(verbal)-13->[wordm]	0,847	0,030	28,701	0,000
(DELTA1)-14-(DELTA1)	0,567	0,088	6,464	0,000
(DELTA2)-15-(DELTA2)	0,757	0,077	9,851	0,000
(DELTA3)-16-(DELTA3)	0,719	0,080	8,990	0,000
(DELTA4)-17-(DELTA4)	0,491	0,091	5,420	0,000
(DELTA5)-18-(DELTA5)	0,635	0,093	6,797	0,000
(DELTA6)-19-(DELTA6)	0,689	0,089	7,752	0,000
(DELTA7)-20-(DELTA7)	0,611	0,095	6,407	0,000
(DELTA8)-21-(DELTA8)	0,798	0,077	10,325	0,000
(DELTA9)-22-(DELTA9)	0,350	0,056	6,302	0,000
(DELTA10)-23-(DELTA10	0,324	0,053	6,064	0,000
(DELTA11)-24-(DELTA11	0,304	0,052	5,872	0,000
(DELTA12)-25-(DELTA12	0,523	0,066	7,864	0,000
(DELTA13)-26-(DELTA13	0,283	0,050	5,663	0,000
(memory)-27-(spatial)	0,449	0,107	4,204	0,000
(verbal)-28-(memory)	0,520	0,087	5,970	0,000
(verbal)-29-(spatial)	0,591	0,076	7,758	0,000

Exemple de lecture des résultats :

En données centrées réduites (car on travaille sur les corrélations), on a, selon le modèle :

visual = 0,658 spatial + delta1
cubes = 0,492 spatial + delta2
Variance(delta1) = 0,567
Variance(delta2) = 0,757
Corrélation(delta1, delta2) = 0
Corrélation(spatial, delta1) = 0

On peut confronter ces résultats avec ceux figurant sous l'onglet "Résidus". Le bouton "Matrice d'entrée" donne la matrice des corrélations entre les variables observées, tandis que le bouton "Matrice reproduite" calcule les corrélations des variables produites par le modèle.

Ainsi, selon le modèle : Var(visual) = $0,658^2$ Var(spatial) + Var(delta1)=$0,658^2$+$0,567^2$ = 1
La matrice d'entrée indique : Corrélation(visual, cubes) = 0,326.
La matrice reproduite indique : Corrélation(visual, cubes) = 0,324.
On retrouve cette valeur à partir du modèle :
Corr(visual, cubes) = 0,658 x 0,492 Var(spatial) = 0,658 x 0,492 = 0,3237.

Le tableau des différences entre les corrélations observées et les corrélations produites par le modèle est fourni par le bouton "Résidus centrés-réduits".

2.3 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

2.3.1 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.

2.3.2 Traitement classique d'un tableau de contingence : test du khi-2 sur un 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

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

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.

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

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

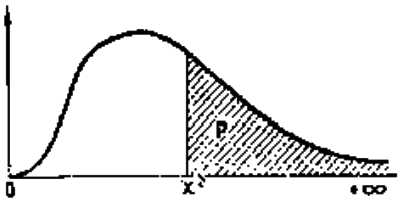
Sous l'hypothèse H_0, la distance entre les deux tableaux suit une loi du χ^2 à 12 degrés de liberté. Ce dernier nombre est défini par la formule :

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

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

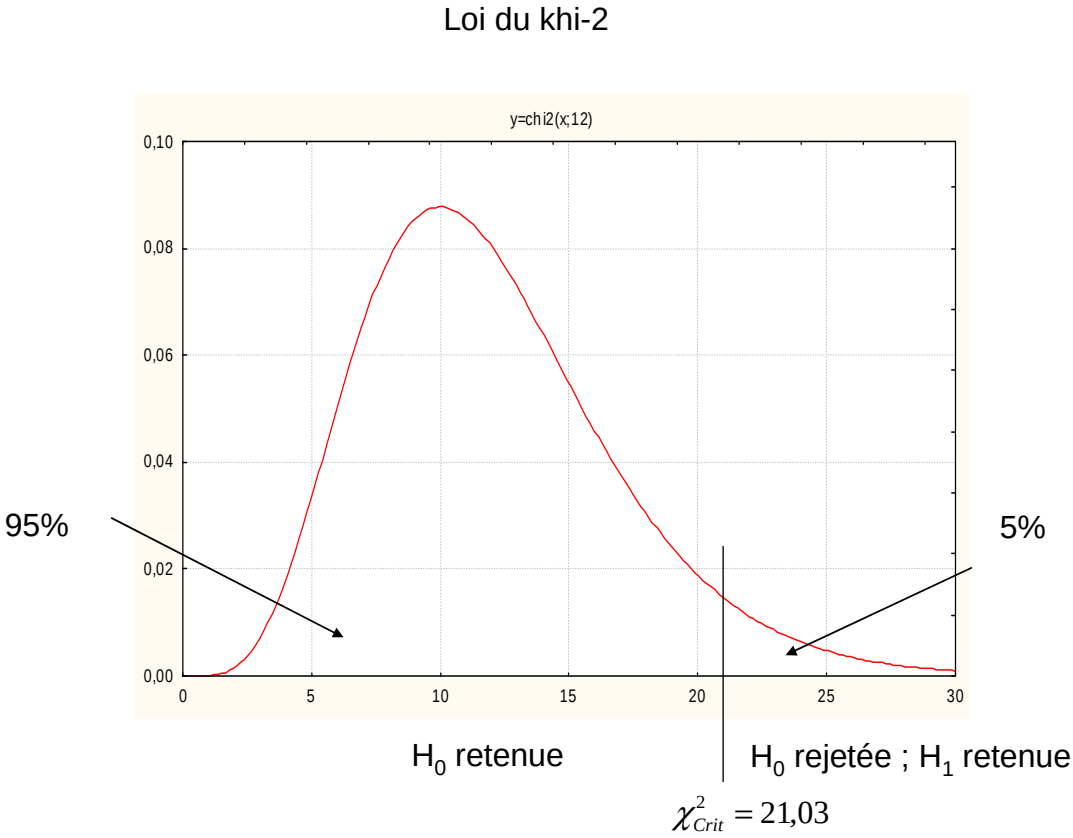
DISTRIBUTION DE χ^2 (Loi de K. Pearson)

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



v	0,9	0,8	0,7	0,5	0,3	0,2	0,1	0,05	0,02	0,01
1	0.0158	0.0642	0.148	0,455	1,074	1,642	2,706	3,841	5,412	6,635
2	0.211	0.446	0.713	1,386	2,408	3,219	4,605	5,991	7,824	9,210
3	0.584	1.005	1.424	2,366	3,665	4,642	6,251	7,815	9,837	11,345
4	1,064	1,649	2,195	3,357	4,878	5,989	7,779	9,488	11,668	13,277
5	1,610	2,343	3,000	4,351	6,064	7,289	9,236	11,070	13,388	15,086
6	2,204	3,070	3,828	5,348	7,231	8,558	10,645	12,592	15,033	16,812
7	2,833	3,822	4,671	6,346	8,383	9,803	12,017	14,067	16,622	18,475
8	3,490	4,594	5,527	7,344	9,524	11,030	13,362	15,507	18,168	20,090
9	4,168	5,380	6,393	8,343	10,656	12,242	14,684	16,919	19,679	21,666
10	4,865	6,179	7,267	9,342	11,781	13,442	15,987	18,307	21,161	23,209
11	5,578	6,989	8,148	10,341	12,899	14,631	17,275	19,675	22,618	24,725
12	6,304	7,807	9,034	11,340	14,011	15,812	18,549	21,026	24,054	26,217
13	7,041	8,634	9,926	12,340	15,119	16,985	19,812	22,362	25,471	27,688
14	7,790	9,467	10,821	13,339	16,222	18,151	21,064	23,685	26,873	29,141
15	8,547	10,307	11,721	14,339	17,322	19,311	22,307	24,996	28,259	30,578

On formule ensuite la règle de décision :



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

2.3.3 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

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

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

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

Exemple : distance entre les colonnes 1 et 2

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

Propriété d'équivalence distributionnelle :

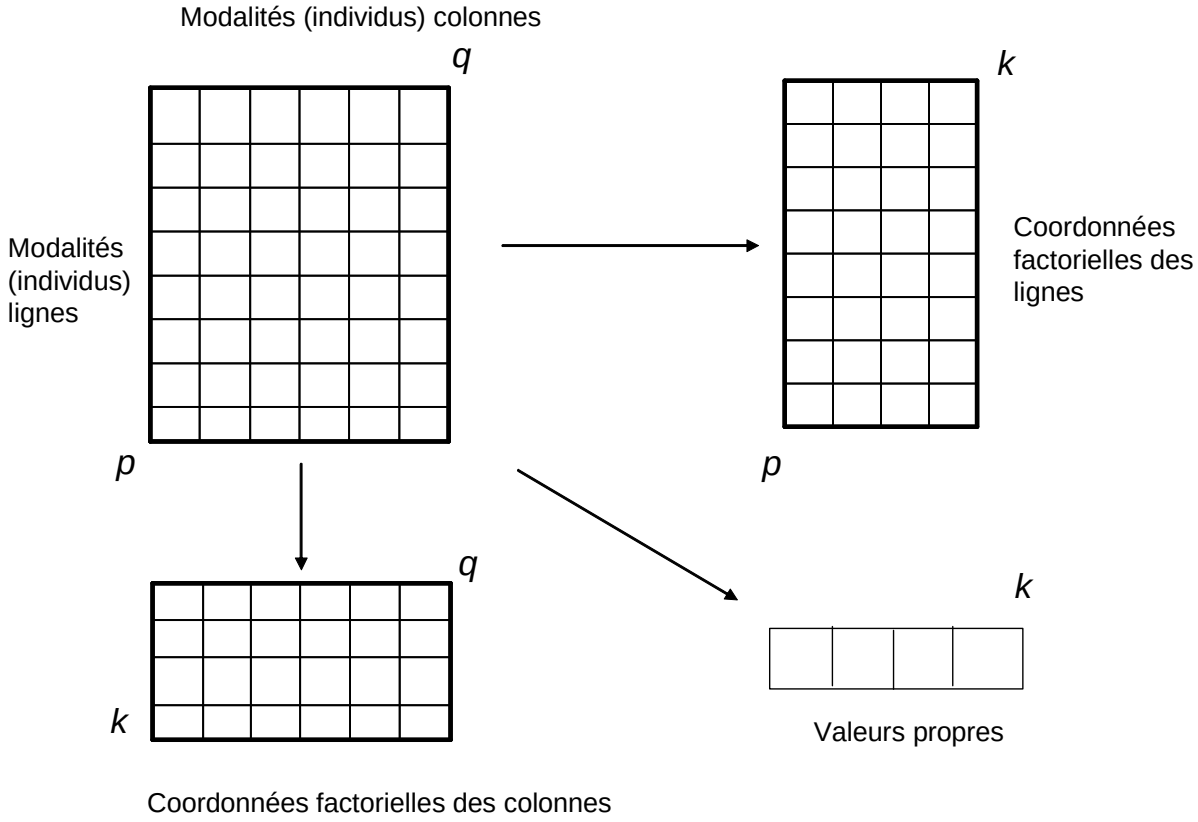
- Si on regroupe deux modalités lignes, les distances entre les profils-colonnes, ou entre les autres profils-lignes restent inchangées.

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

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 :

Principaux résultats d'une 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

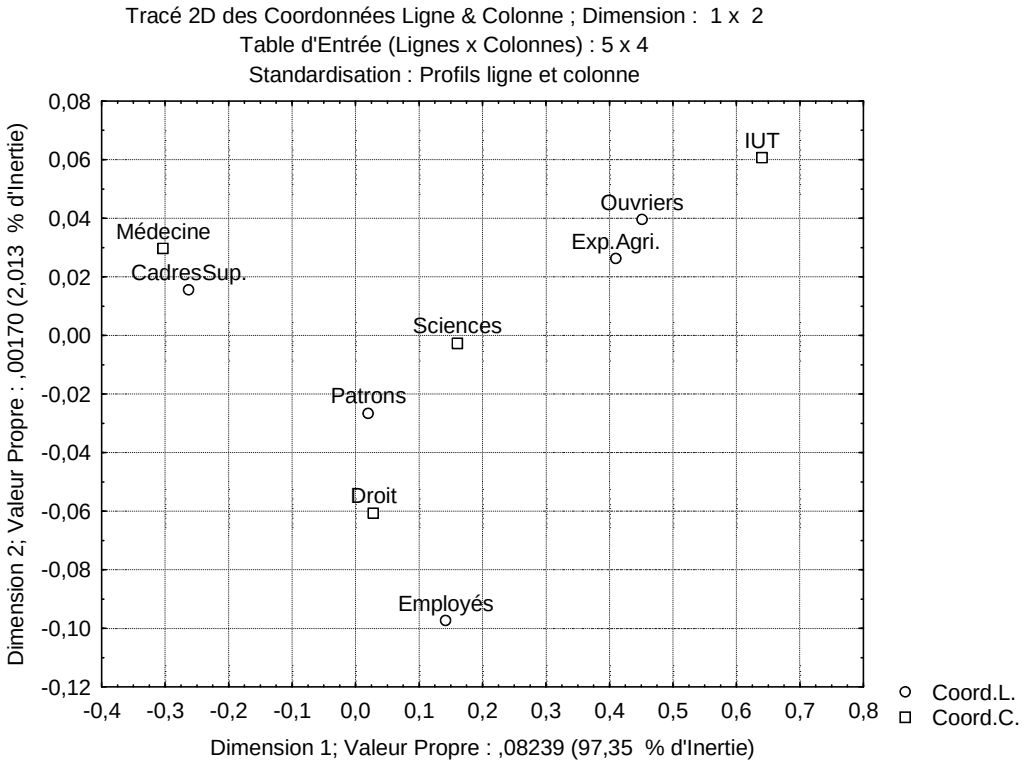
Résultats relatifs aux lignes

	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



2.3.4 Analyse factorielle des correspondances avec Statistica

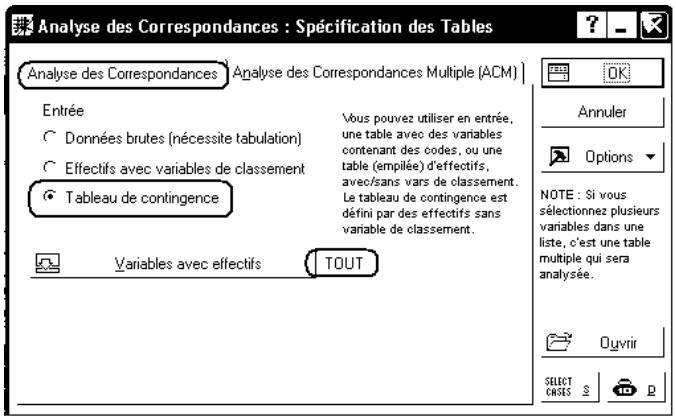
2.3.4.1 Traitement des données avec Statistica

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.



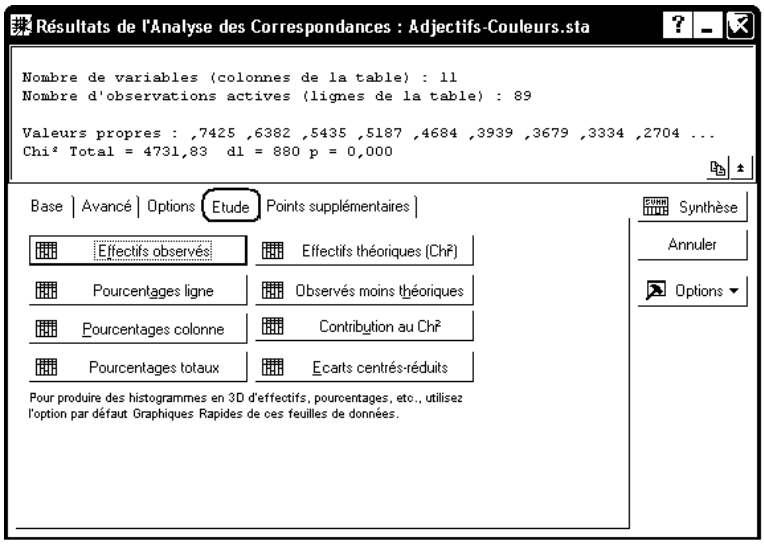
La fenêtre de dialogue permet d'indiquer la manière dont se présentent nos données. La situation la plus classique est celle d'un tableau de contingence : les modalités lignes sont indiquées dans une variable spécifiques, les modalités colonnes sont les autres variables du tableau, et la feuille de données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

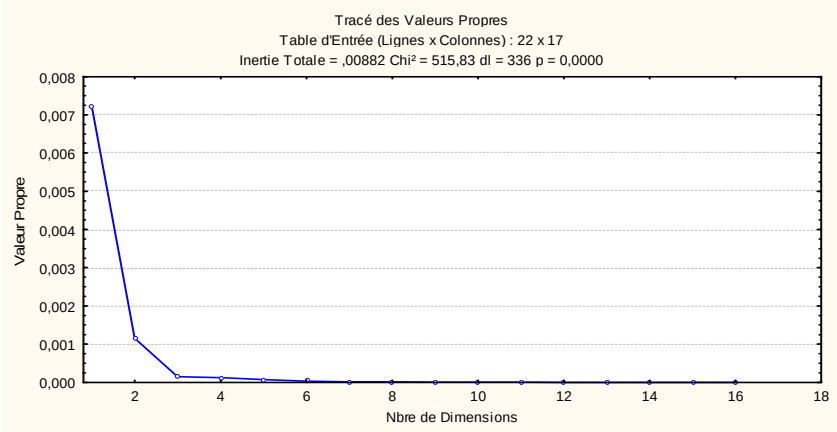
Statistica ne permet pas d'obtenir directement le tableau des taux de liaison, qui est pourtant un outil exploratoire intéressant. Mais on pourra utiliser les tableaux "Observés moins théoriques" et "Effectifs théoriques". On peut même recopier ces deux tableaux dans une feuille Excel et diviser chaque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

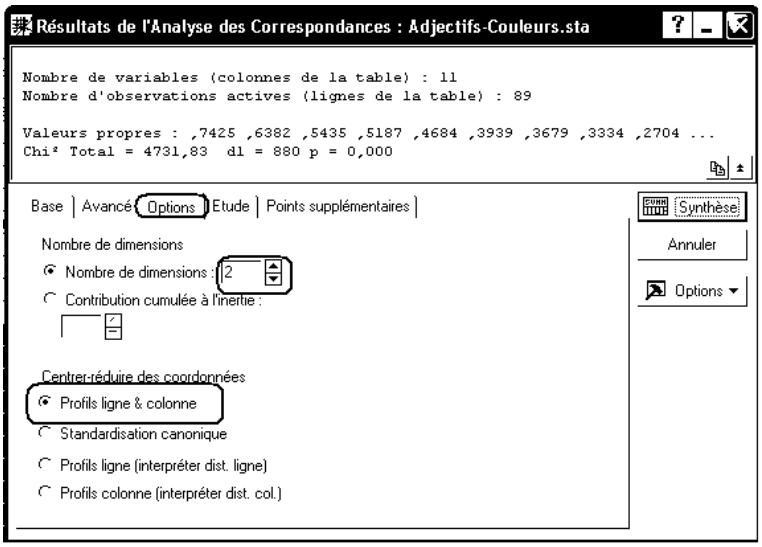
C'est ensuite l'onglet "Avancé" qui nous permettra d'afficher les valeurs propres, et donc de choisir le nombre d'axes à garder.

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.



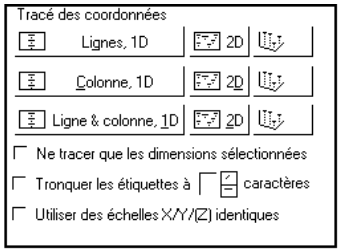
On retourne ensuite sous l'onglet "Avancé" pour afficher les coordonnées des individus-lignes et des individus-colonnes. On notera que Statistica produit deux tableaux de résultats, et on passera de l'un à l'autre à l'aide des onglets du classeur.

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

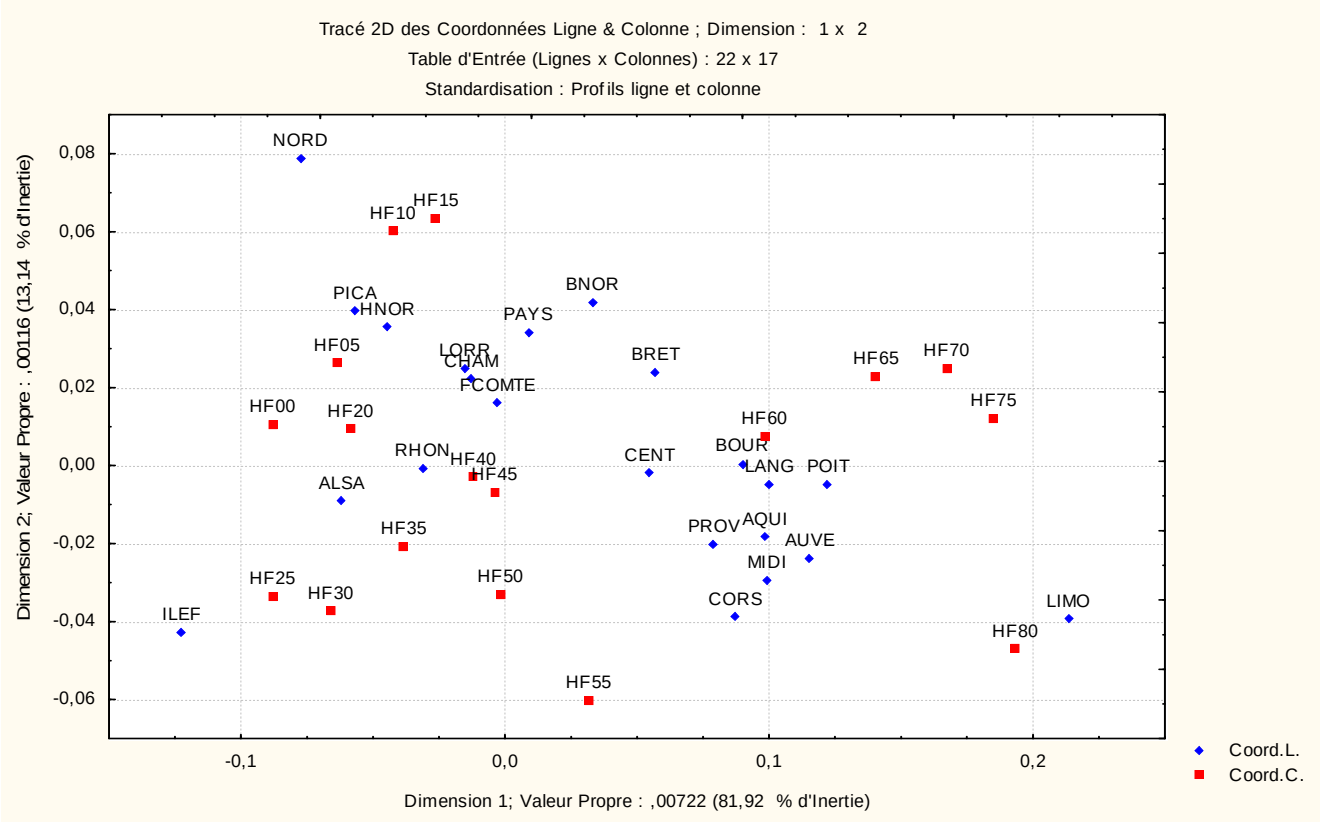
NomLigne	Ligne Numéro	Coord. Dim.1	Coord. Dim.2	Masse	Qualité	Inertie Relative	Inertie Dim.1	Cosinus² Dim.1	Inertie Dim.2	Cosinus² Dim.2	Cos² 1&2
ILEF	1	-0,1223	-0,0427	0,1871	0,9968	0,3574	0,3877	0,8885	0,2944	0,1082	0,9968
CHAM	2	-0,0129	0,0222	0,0229	0,7971	0,0022	0,0005	0,1997	0,0098	0,5974	0,7971
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

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
	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										
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.



2.3.5 Interprétation des résultats de l'AFC

On sait que la distance du khi-2 est sensible à l'importance de l'effectif observé. Sans surprise, même en exprimant les effectifs en milliers, nous obtenons ici un khi-2 de 515,83. En revanche, le coefficient Phi-2 est assez faible : 0,0088.

De même, on constate que les taux de liaison restent modérés, compris entre -0,28 et +0,59. Autrement dit le déficit d'une classe d'âge est au plus de 28% de l'effectif théorique que l'on obtiendrait si la structure par âge de la population française se retrouvait à l'identique dans toutes les régions, et l'excès d'une classe d'âge est d'au plus 59% de cet effectif théorique.

On sait que, dans une AFC, les valeurs propres sont toutes inférieures à 1, et que leur somme est égale au coefficient Phi-2. Ici, la décroissance des valeurs propres est très rapide, puisque la première représente plus de 80% de l'inertie. La deuxième, bien que très inférieure à la première, est supérieure à la moyenne 0,0088/16 = 0,00055. Nous étudierons donc les deux premiers axes factoriels.

2.3.5.1 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).

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

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.

2.3.5.3 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.

2.3.5.4 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.

2.3.6 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.

2.3.6.1 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.

2.3.6.2 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.

2.3.6.3 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.

2.3.7 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.

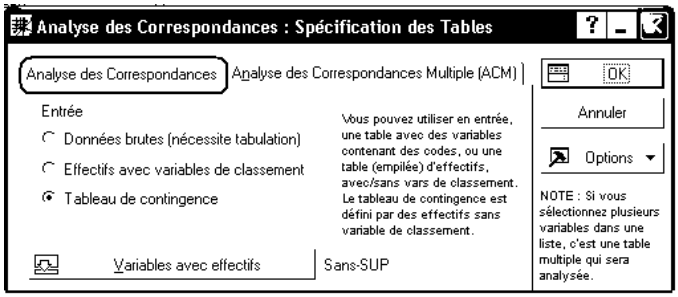
Ces tableaux croisent trois variables qualitatives : l'emploi, le diplôme et le sexe. Les buts de notre étude peuvent être multiples. D'une part, on peut s'intéresser à la liaison entre emploi et diplôme, indépendamment du sexe, et mettre ainsi en évidence une structure commune à ces deux tableaux. D'autre 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 ?

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

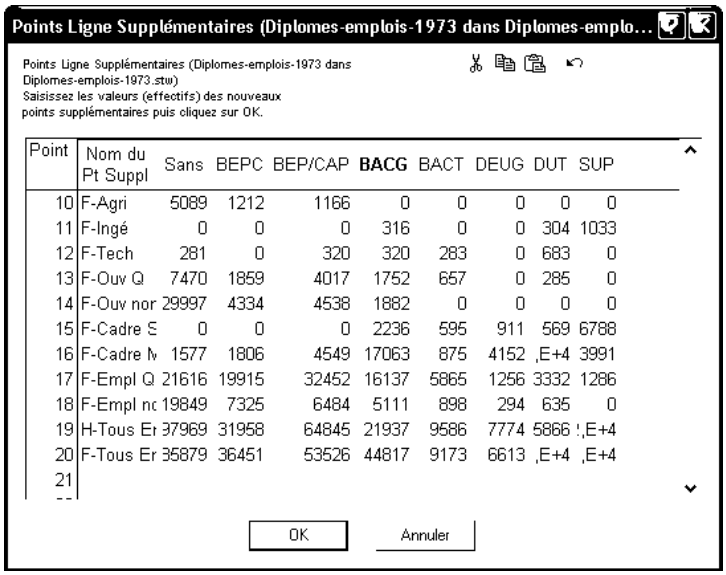
Ouvrez le classeur Diplomes-emploi-1973.stw et observez la manière dont les données y ont été saisies. Ouvrez également le classeur Excel Diplomes-emplois-1973.xls.

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 :

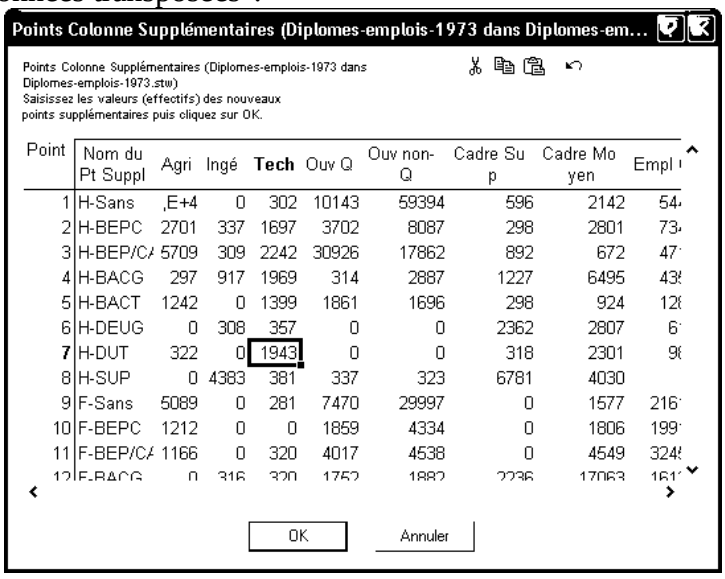


Activez ensuite l'onglet "Points supplémentaires" et cliquez sur le bouton "Ajouter des points lignes". Plutôt que de saisir ces données supplémentaires à la main, copiez, puis collez dans la fenêtre la plage A11:I30 de la feuille "Donnees" du classeur Excel.



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".

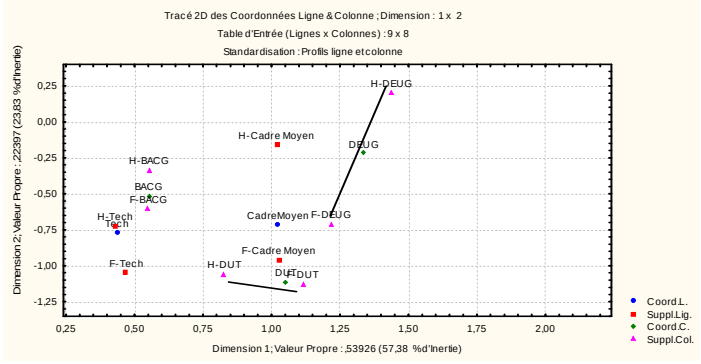


Point	Nom du Pt Suppl	Agri	Ingé	**Tech**	Ouv Q	Ouv non-Q	Cadre Sup	Cadre Moyen	Empl
1	H-Sans	,E+4	0	302	10143	59394	596	2142	54
2	H-BEPC	2701	337	1697	3702	8087	298	2801	73
3	H-BEP/CA	5709	309	2242	30926	17862	892	672	47
4	H-BACG	297	917	1969	314	2887	1227	6495	43
5	H-BACT	1242	0	1399	1861	1696	298	924	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).

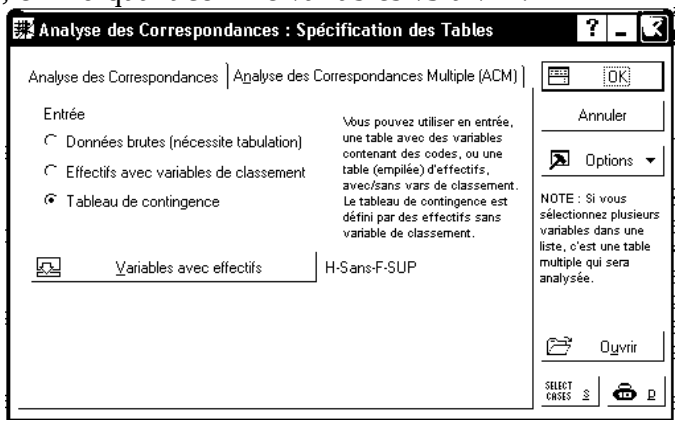
On constate que, sur le premier axe, pour tous les diplômes, les deux points représentant les hommes et les femmes sont presque confondus. En revanche, pour les points relatifs au DEUG par exemple, la différence des coordonnées sur le deuxième axe est très importante. D'une manière générale, on constate 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

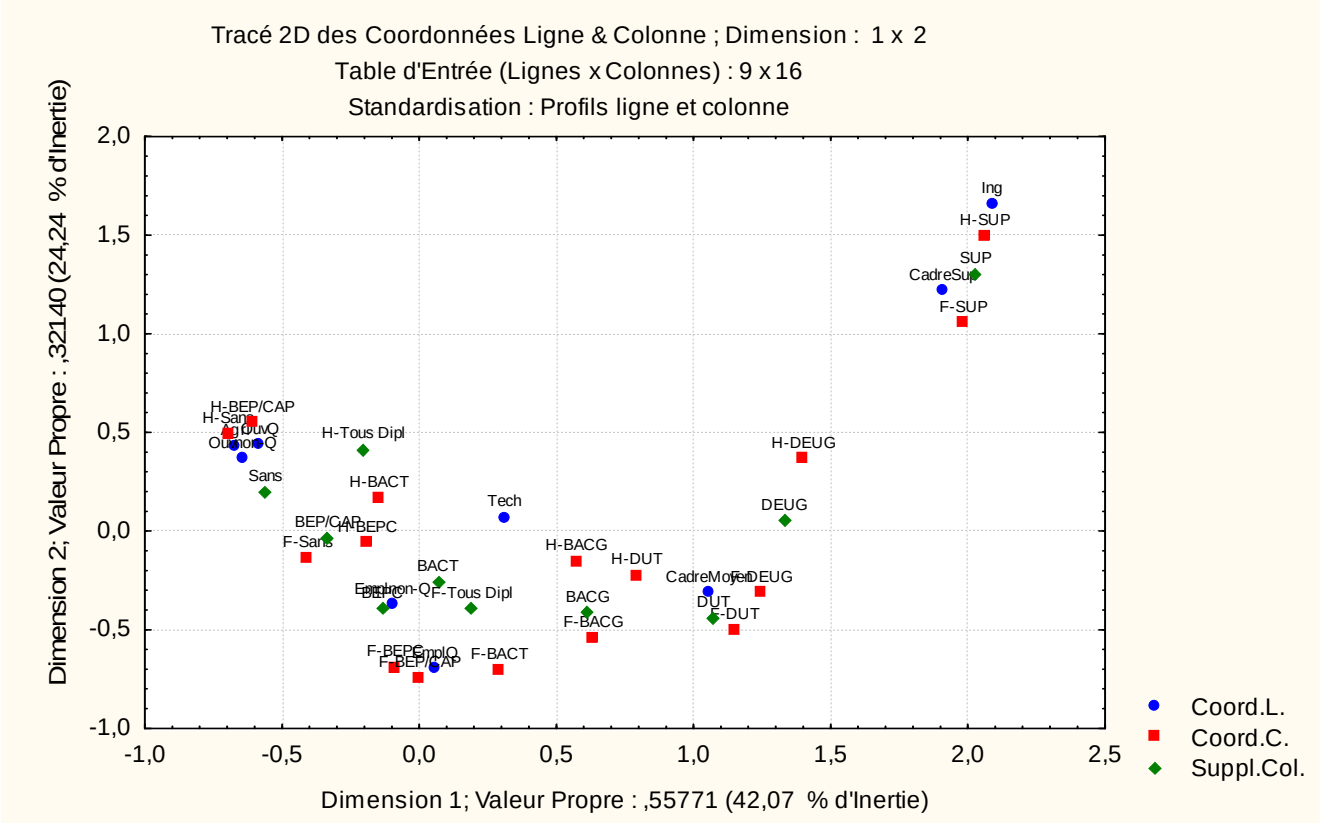
- Dans l'étude menée ici, l'inertie prise en compte (Phi-2 = 0,94) est celle du tableau "somme". On ne tient donc pas compte de la dispersion des données due aux discriminations liées au sexe.
- Deux points supplémentaires correspondant aux deux sexes peuvent être représentés proches l'un de l'autre 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.

2.3.7.2 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".



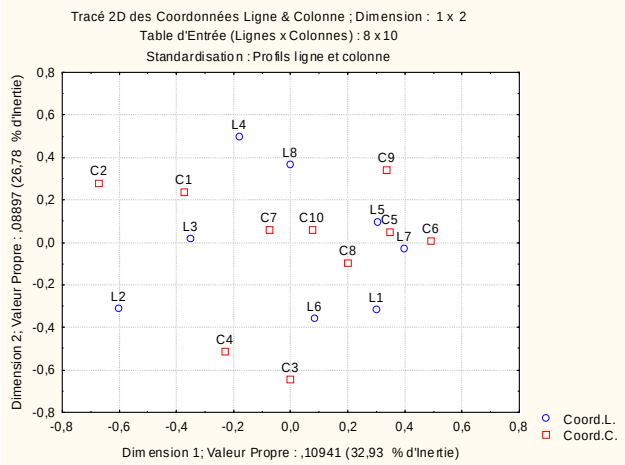
L'interprétation du graphique se fait comme précédemment. Cependant, l'inertie du nuage de points fait maintenant intervenir les points relatifs aux diplômes par sexe. On constate que le coefficient Phi-2 est plus élevé que dans l'étude précédente : 1,326 au lieu de 0,940. La différence entre les deux 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.

2.3.8 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.

2.3.8.1 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.

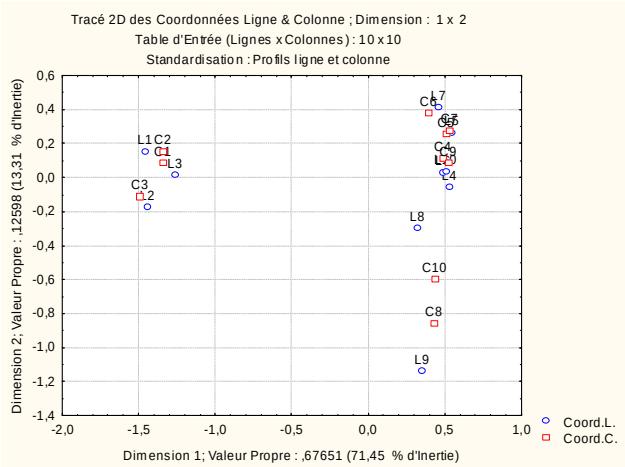


2.3.8.2 Deux paquets de points - Valeurs propres proches de 1

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 :

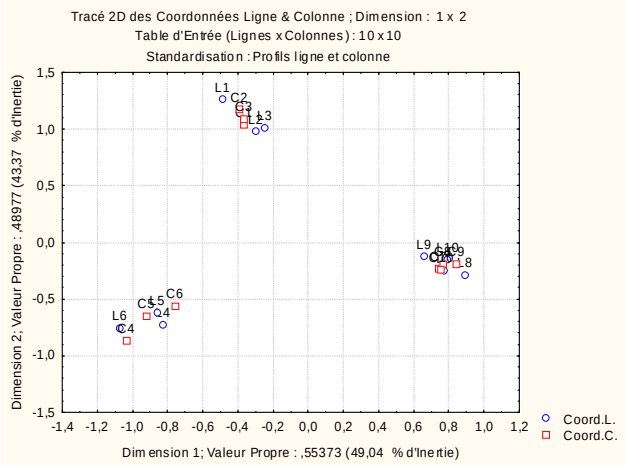
	0
0	

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



2.3.8.3 Trois paquets de points

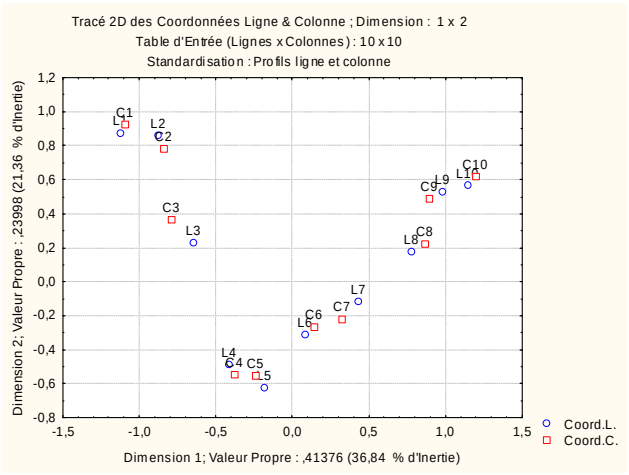
De même, l'existence de deux valeurs propres proches de 1 indique une partition des observations en 3 groupes. Si toutes les valeurs propres sont proches de 1, cela indique une correspondance entre chaque modalité ligne et une modalité colonne "associée". Avec une réorganisation convenable des modalités,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

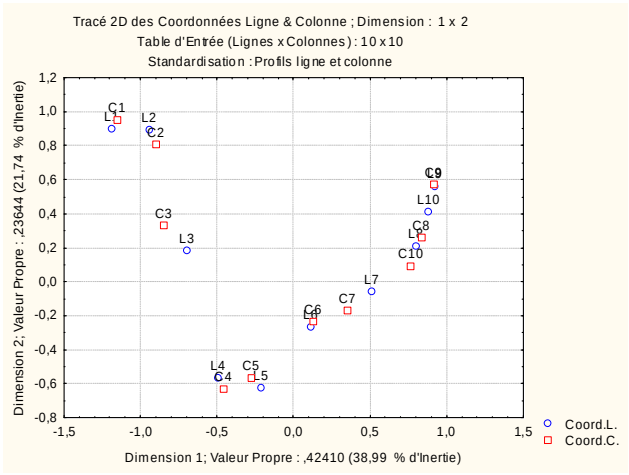


2.3.8.4 L'effet Guttman.

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".





2.3.8.5 Nuage tétraédrique

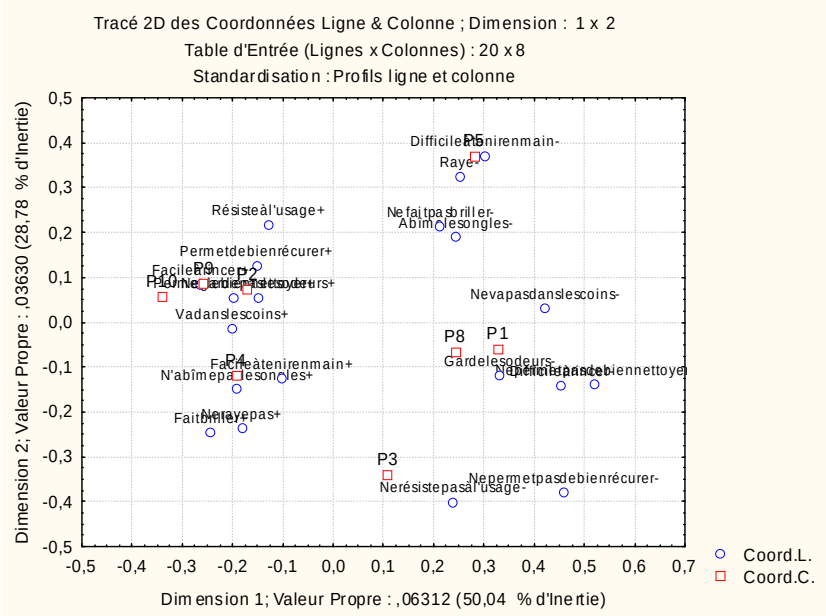
Le premier exemple ("Deux-paquets") est également caractéristique d'une forme classique de nuage : tétraédrique, ou en forme de "berlingot" comme on peut s'en rendre compte en construisant les projections du nuage sur les 3 premiers axes.

2.3.9 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.

2.3.9.1 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.

2.3.9.2 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).

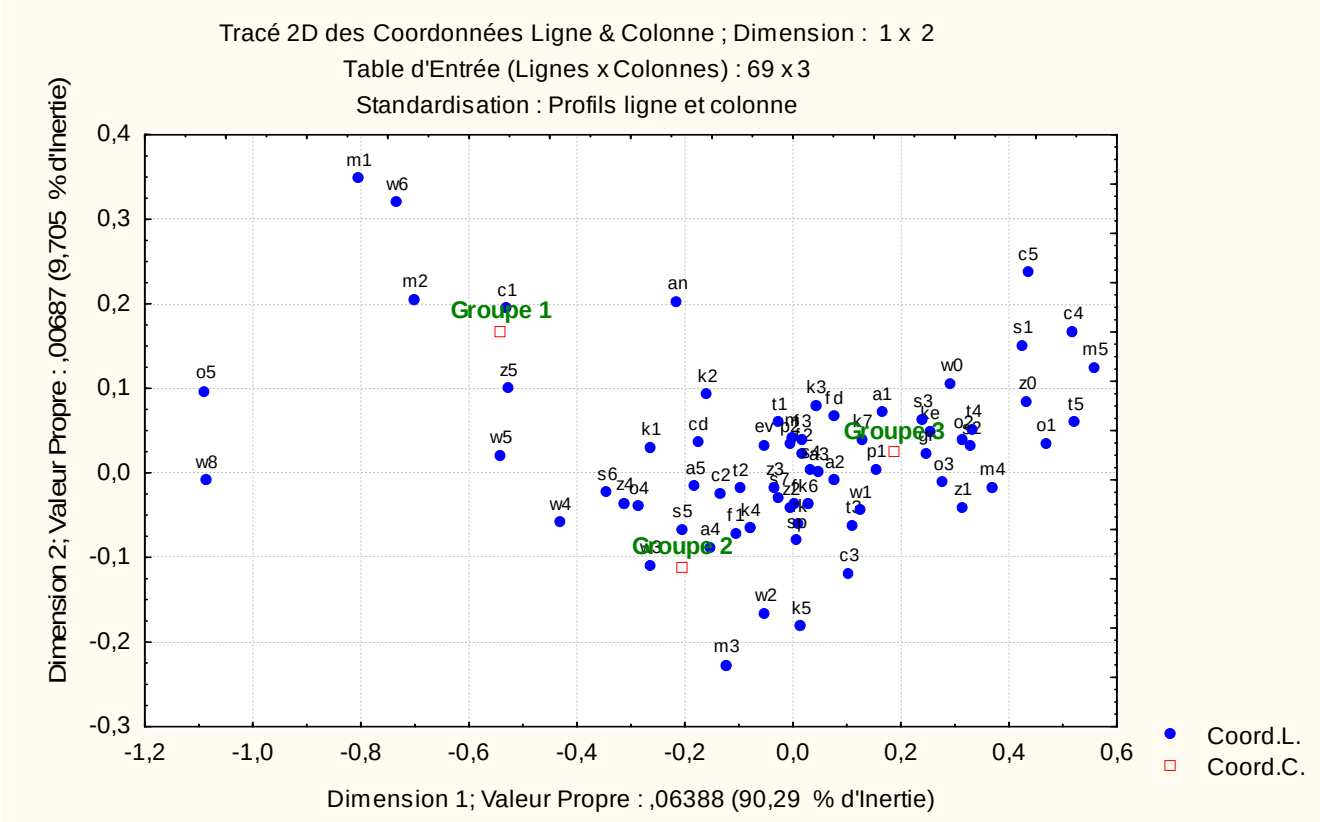
Réalisez une analyse des correspondances sur ce tableau et retrouvez ainsi les résultats de l'auteur :

> "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."



2.3.10 Conclusion

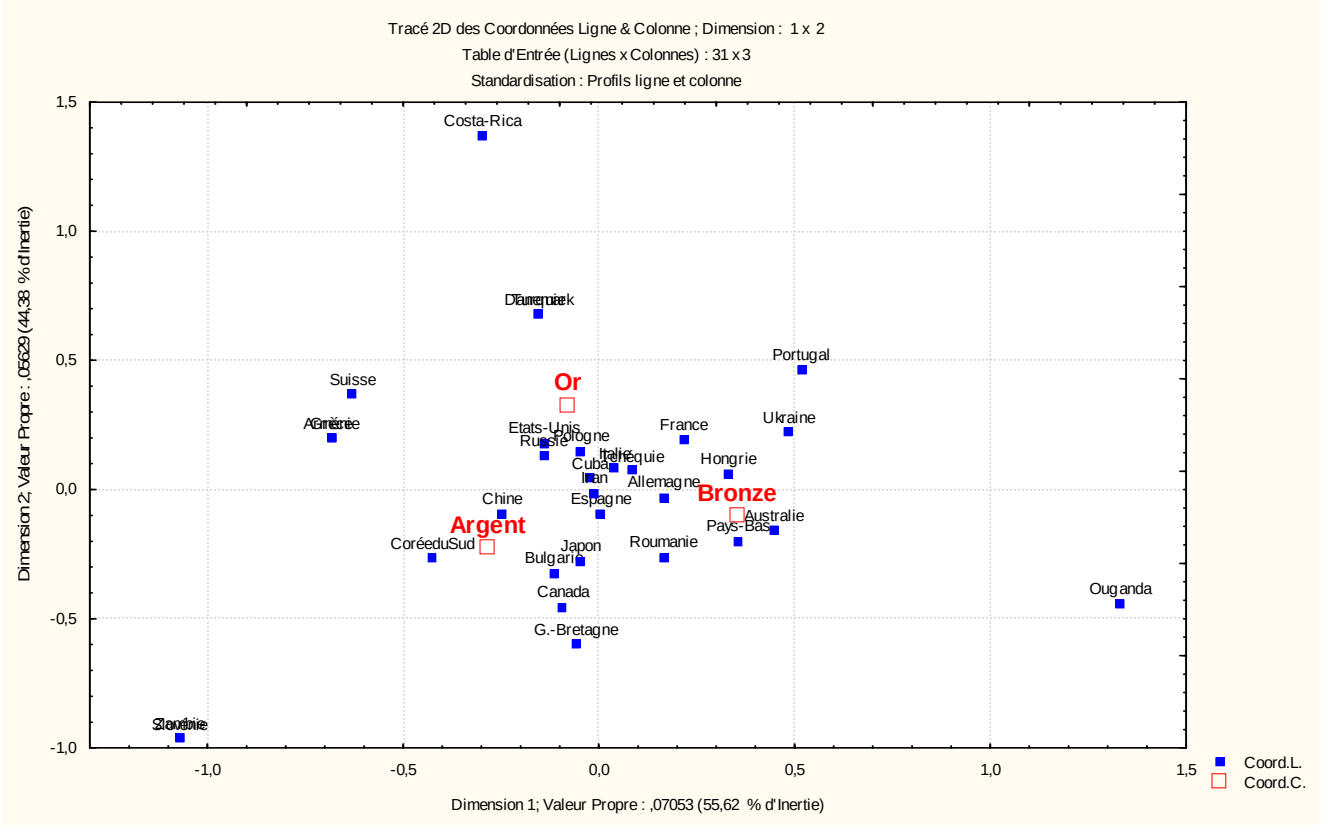
Au travers des exemples qui précédent, on voit que l'AFC peut être utilisée dans des situations variées. En revanche, comme l'indique Philippe Cibois dans son article "les pièges de l'AFC", il existe des situations où il vaut mieux s'abstenir d'utiliser cette méthode :

- L'AFC mettra toujours en évidence des attractions - répulsions entre modalités lignes et modalités colonne. Mais, lorsqu'on travaille sur un échantillon et que le khi-2 du tableau de contingence n'est pas significatif, l'effet mis en évidence n'est que le fruit du hasard.

- L'AFC n'a d'intérêt que si notre étude porte sur les liaisons existant entre lignes et colonnes. Comme exemple, caricatural, d'un mauvais usage de l'AFC, P. Cibois fournit le tableau suivant qui donne, pour 30 pays, le nombre de médailles d'or, d'argent et de bronze obtenues aux Jeux Olympiques d'Atlanta :

	Or	Argent	Bronze		Or	Argent	Bronze
Etats-Unis	44	32	25	Grèce	4	4	0
Russie	26	21	16	Tchéquie	4	3	4
Allemagne	20	18	27	Suisse	4	3	0
Chine	16	22	12	Danemark	4	1	1
France	15	7	15	Turquie	4	1	1
Italie	13	10	12	Canada	3	11	8
Australie	9	9	23	Bulgarie	3	7	5
Cuba	9	8	8	Japon	3	6	5
Ukraine	9	2	12	G.-Bretagne	1	8	6
Corée du Sud	7	15	5	Iran	1	1	1
Pologne	7	5	5	Arménie	1	1	0
Hongrie	7	4	10	Portugal	1	0	1
Espagne	5	6	6	Costa-Rica	1	0	0
Roumanie	4	7	9	Slovénie	0	2	0
Pays-Bas	4	5	10	Zambie	0	1	0

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 ?



2.4 Analyse des Correspondances Multiples

2.4.1 Introduction

L'analyse factorielle des correspondances, vue dans le paragraphe précédent, s'applique à des situations où les individus statistiques sont décrits par *deux* variables nominales. Mais il est fréquent que l'on dispose d'individus décrits par *plusieurs* (deux ou plus) variables nominales ou ordinales. C'est notamment le cas lorsque nos données sont les résultats d'une enquête basée sur des questions fermées. Une extension de l'AFC à ces situations a donc été proposée. Elle est généralement appelée Analyse des Correspondances Multiples[1] ou ACM.

Nous nous plaçons donc dans la situation où nous disposons de N individus statistiques, décrits par q variables nominales ou ordinales X_1, X_2, ..., X_q. L'ACM vise à mettre en évidence :

- les relations entre les modalités des différentes variables ;
- éventuellement, les relations entre individus statistiques ;
- les relations entre les variables, telles qu'elles apparaissent à partir des relations entre modalités.

2.4.2 Forme des données d'entrée

Selon leur origine, les données sur lesquelles nous nous proposons de faire une ACM peuvent se présenter sous différentes formes.
Imaginons, par exemple, une mini-enquête dans laquelle nous avons posé trois questions à 10 sujets : le sexe (F ou H), le niveau de revenus (M : modeste, E : élevé) et leur préférence sur un sujet donné (3 modalités : A, B ou C). Les données peuvent se présenter sous l'une des formes décrites ci-dessous. Le classeur Mini-ACM.stw contient 5 feuilles de données correspondant à ces 5 formes.

2.4.2.1 Tableau protocole

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

[1] Cette méthode est aussi parfois appelée *homogeneity analysis*.

2.4.2.2 Tableau d'effectifs

	1 Sexe	2 Revenu	3 reference	4 Effectif
1	F	M	A	2
2	F	E	B	1
3	F	E	C	2
4	H	E	C	1
5	H	E	B	1
6	H	M	B	2
7	H	M	A	1

2.4.2.3 Tableau disjonctif complet

Le tableau disjonctif complet ou TDC comporte une colonne pour chaque modalité des variables étudiées et une ligne pour chaque individu statistique. Les cellules du tableau contiennent 1 ou 0 selon que l'individu considéré présente la modalité ou non.

	1 Sexe:F	2 Sexe:H	3 Rev:M	4 Rev:E	5 Pref:A	6 Pref:B	7 Pref:C
s1	1	0	1	0	1	0	0
s2	1	0	1	0	1	0	0
s3	1	0	0	1	0	1	0
s4	1	0	0	1	0	0	1
s5	1	0	0	1	0	0	1
s6	0	1	0	1	0	0	1
s7	0	1	0	1	0	1	0
s8	0	1	1	0	0	1	0
s9	0	1	1	0	0	1	0
s10	0	1	1	0	1	0	0

2.4.2.4 Tableau disjonctif des patrons

En regroupant les lignes identiques du tableau disjonctif complet, on obtient le tableau disjonctif des patrons :

	1 Sexe:F	2 Sexe:H	3 Rev:M	4 Rev:E	5 Pref:A	6 Pref:B	7 Pref:C
FMA	2	0	2	0	2	0	0
FEB	1	0	0	1	0	1	0
FEC	2	0	0	2	0	0	2
HEC	0	1	0	1	0	0	1
HEB	0	1	0	1	0	1	0
HMB	0	2	2	0	0	2	0
HMA	0	1	1	0	1	0	0

2.4.2.5 Tableau de Burt

L'ACM peut également être réalisée à partir d'une structuration particulière des données, appelée tableau de Burt (TdB). Ce dernier tableau comporte une ligne et une colonne pour chaque modalité des variables étudiées. Chaque cellule du tableau indique le nombre d'individus statistiques qui possèdent en même temps la modalité ligne et la modalité colonne correspondantes. Le tableau de Burt apparaît ainsi comme une juxtaposition de tableaux de contingence des variables prises deux à deux.

	Table Observée (Effectifs) (Protocole dans Classeur2) Table d'Entrée (Lignes x Colonnes) : 7 x 7 (Table de Burt)						
	F	H	M	E	A	B	C
Sexe:F	5	0	2	3	2	1	2
Sexe:H	0	5	3	2	1	3	1
Revenu:M	2	3	5	0	3	2	0
Revenu:E	3	2	0	5	0	2	3
Preference:A	2	1	3	0	3	0	0
Preference:B	1	3	2	2	0	4	0
Preference:C	2	1	0	3	0	0	3

On peut noter qu'il est possible, sans grand problème de passer de l'une des 4 premières structures de données à une autre. De même, le TdB peut être obtenu facilement à partir du tableau disjonctif complet. En revanche, il n'existe pas de moyen simple pour recomposer l'une des 4 premières structures de données à partir du tableau de Burt.

2.4.3 Quelques règles d'interprétation

On cherchera d'une part à interpréter les oppositions entre modalités (ou entre groupes d'individus, si l'étude porte sur le TDC), et d'autre part à interpréter les proximités entre modalités.

L'interprétation des proximités entre les modalités devra tenir compte de la remarque suivante :

- Si deux modalités *d'une même variable* sont proches, cela signifie que les individus qui possèdent l'une des modalités et ceux qui possèdent l'autre sont globalement similaires *du point de vue des autres variables* ;
- Si deux modalités *de deux variables différentes* sont proches, cela peut signifier que ce sont globalement les mêmes individus qui possèdent l'une et l'autre.

Nous pouvons, comme en AFC, nous intéresser aux profils ligne et colonne, aux taux de liaison et au Φ^2 du tableau disjonctif complet, vu comme un tableau de contingence. Le nombre de lignes de ce tableau est égal au nombre d'individus statistiques étudiés. Cependant, nous avons vu que la métrique du Φ^2, utilisée pour l'AFC, possède la propriété d'équivalence distributionnelle : si on regroupe deux lignes correspondant au même patron de réponses, on ne change rien aux autres profils lignes, ni aux autres profils colonnes. Autrement dit, on retrouvera les mêmes résultats en effectuant une AFC sur le tableau disjonctif des patrons.

Comme en AFC, on peut calculer des fréquences, des fréquences lignes, des fréquences colonnes et des profils lignes et profils colonnes moyens.
L'élément le plus facile à interpréter est le profil colonne moyen : ce sont les fréquences des différents patrons de réponses dans la population étudiée.

L'élément le plus facile à interpréter est le profil colonne moyen : ce sont les fréquences des différents patrons de réponses dans la population étudiée.

Le profil ligne moyen est obtenu en calculant, pour chaque modalité, le quotient de sa fréquence par le nombre Q de questions. En notant respectivement n_k et f_k l'effectif et la fréquence de la modalité *k*, on a :

$$f_k = \frac{n_k}{N} = \frac{Nombre\ \ d'individus\ \ ayant\ \ choisi\ \ la\ \ modalité\ \ k}{Nombre\ \ total\ \ d'individus}$$

et le k-ième élément du profil-ligne moyen est :

$$f_{\bullet k} = \frac{f_k}{Q} = \frac{n_k}{QN} = \frac{Nombre\ \ d'individus\ \ ayant\ \ choisi\ \ la\ \ modalité\ \ k}{Nombre\ \ de\ \ questions \times Nombre\ \ total\ \ d'individus}$$

N.B. Ici, f_k et $f_{\bullet k}$ désignent des quantités différentes : f_k est la fréquence de la modalité k dans la population étudiée; $f_{\bullet k}$ est définie comme pour l'AFC, fréquence ligne marginale de la k-ième colonne du tableau disjonctif des patrons.

2.4.3.1 Taux de liaison et Phi-2

Pour le tableau disjonctif complet, ou le tableau disjonctif des patrons, considérés comme des tableaux de contingence, le coefficient Phi-2 vaut :

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

où K désigne le nombre de modalités et Q le nombre de questions

Dans notre exemple, on a : K=7, Q=3, et donc : $\Phi^2 = \frac{7}{3} - 1 = 1{,}33$.

Ce coefficient représente l'inertie totale du nuage de points des modalités colonnes. On montre que l'inertie absolue de chacune des questions est donnée par :

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

où K_q représente le nombre de modalités de la question q.

L'inertie relative de chacune des questions est donnée par :

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

Sur notre exemple, on a, pour l'inertie absolue :

$$I(\text{Sexe}) = I(\text{Revenu}) = \frac{2-1}{3} = 0{,}33$$

$$I(\text{Pref}) = \frac{3-1}{3} = 0{,}67$$

Quant aux inerties relatives :

$$Inr(\text{Sexe}) = Inr(\text{Revenu}) = \frac{2-1}{4} = 25\%$$

$$Inr(\text{Pref}) = \frac{3-1}{4} = 50\%$$

L'inertie d'une question est ainsi directement liée au nombre de ses modalités : on évitera donc d'utiliser la méthode lorsque les différentes questions présentent des nombres de modalités trop différents.

2.4.3.2 Distances entre profils lignes

En AFC, nous avions donné les formules permettant de calculer les distances entre deux profils lignes ou entre deux profils colonnes. La distance utilisée est la *métrique du* Φ^2. Ici, compte tenu de la structure particulière du tableau de contingence utilisé, les formules indiquées deviennent :

$$d^2_{\Phi^2}(L_i, L_{i'}) = \frac{1}{Q}\sum_k \frac{(\delta_{ik} - \delta_{i'k})^2}{f_k}$$

Notations utilisées : Li et Li' désignent deux patrons, Q est le nombre de questions. δ_{ik} prend la valeur 1 si la modalité k fait partie du patron i, et la valeur 0 sinon. Enfin, f_k est la fréquence de la modalité k dans la population.

Cette formule montre que deux individus (ou deux patrons) sont d'autant plus éloignés que leurs réponses diffèrent pour un plus grand nombre de questions et pour des modalités rares. Cette formule peut encore être écrite sous la forme :

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

où la somme est étendue à toutes les modalités faisant partie de l'un des deux patrons, sans faire partie des deux patrons.

Autrement dit, deux individus (ou deux patrons) sont d'autant plus éloignés que leurs réponses diffèrent pour un plus grand nombre de questions et pour des modalités rares.

Ainsi, sur notre exemple :

$$d^2_{\Phi^2}([\text{FMA}],[\text{HMA}]) = \frac{1}{3}\left(\frac{1}{0{,}5} + \frac{1}{0{,}5}\right) = 1{,}33$$

La distance d'un patron au profil ligne moyen est :

$$d^2_{\Phi^2}(O, L_i) = \left(\frac{1}{Q}\sum_k \frac{\delta_{ik}}{f_k}\right) - 1$$

Autrement dit, un patron sera d'autant plus loin de l'origine qu'il fait intervenir des modalités plus rares. On peut aussi écrire cette formule sous la forme :

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

où la somme est étendue à toutes les modalités faisant partie du patron i.

Par exemple :

$$d^2_{\Phi^2}(O, [FMA]) = \left(\frac{1}{3}\left(\frac{1}{0{,}5} + \frac{1}{0{,}5} + \frac{1}{0{,}3}\right)\right) - 1 = 1{,}44$$

La contribution (absolue) d'un patron à la variance du nuage est obtenue en multipliant la distance précédente par la fréquence du patron dans la population.

2.4.3.3 Distances entre profils colonnes

La distance entre les modalités k et k' est donnée par :

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

où $f_{kk'}$ est la fréquence de la combinaison de modalités k et k', ou encore :

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

Deux modalités sont d'autant plus éloignées qu'elles sont de fréquences faibles et rarement rencontrées simultanément.

Exemple :

$$d^2_{\Phi^2}(\text{Sexe}:\text{F}, \text{Revenu}:\text{M}) = \frac{1}{0,5} + \frac{1}{0,5} - 2\frac{0,2}{0,5\times 0,5} = \frac{5+5-2\times 2}{5\times 5/10} = 2,4$$

La distance d'une modalité au profil colonne moyen est donnée par :

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

Autrement dit, une modalité sera d'autant plus éloignée du profil moyen que sa fréquence est faible. Afin d'éviter que quelques modalités très rares ne prennent une importance excessive dans les résultats obtenus, il sera nécessaire de regrouper les modalités de fréquence trop faible (fréquence inférieure à 5% par exemple).

Exemple :

$$d^2_{\Phi^2}(O, \text{Pref}:\text{B}) = \frac{1}{0,4} - 1 = \frac{10}{4} - 1 = 1,5$$

La contribution absolue d'une modalité à la variance du nuage de points est :

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

La contribution relative d'une modalité à la variance du nuage de points est :

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

Exemples :

$$Ctr([\text{Sexe}:\text{F}]) = \frac{1-0,5}{4} = 12,5\%$$

$$Ctr([\text{Pref}:\text{A}]) = \frac{1-0,3}{4} = 17,5\%$$

2.4.4 Résultats de l'ACM sur l'exemple

Le tableau des valeurs propres est donné par :

Nombre de Dims.	Valeurs Propres et Inertie de toutes les Dimensions (Protocole dans Mini-ACM-v7.stw) Table d'Entrée (Lignes x Colonnes) : 7 x 7 (Table de Burt) Inertie Totale = 1,3333				
	ValSing.	ValProp.	%age Inertie	%age Cumulé	Chi²
1	0,7764	0,6028	45,2128	45,2128	25,3794
2	0,6810	0,4637	34,7781	79,9909	19,5221
3	0,4505	0,2030	15,2219	95,2128	8,5446
4	0,2526	0,0638	4,7872	100,0000	2,6872

Taux d'inertie modifiés

La décroissance des valeurs propres est en général très lente. Pour déterminer le nombre d'axes factoriels à conserver, Benzécri a proposé de calculer des taux d'inertie modifiés en utilisant la méthode suivante.

La somme des valeurs propres est égale à l'inertie totale, c'est-à-dire $\frac{K-Q}{Q}$ et la moyenne des valeurs propres est égale à $\lambda_m = \frac{1}{Q} = \frac{1}{Nb\ de\ questions}$. On ne conserve que les valeurs propres λ supérieures à λ_m et on calcule pour chacune d'entre elles : $\lambda' = (\lambda - \lambda_m)^2$. Le taux d'inertie modifié est alors calculé par : $\frac{\lambda'}{\sum \lambda'}$ et on conserve les valeurs propres dont le taux modifié est supérieur à la moyenne (des taux modifiés). Pour l'exemple traité, l'application de cette méthode donne les résultats suivants :

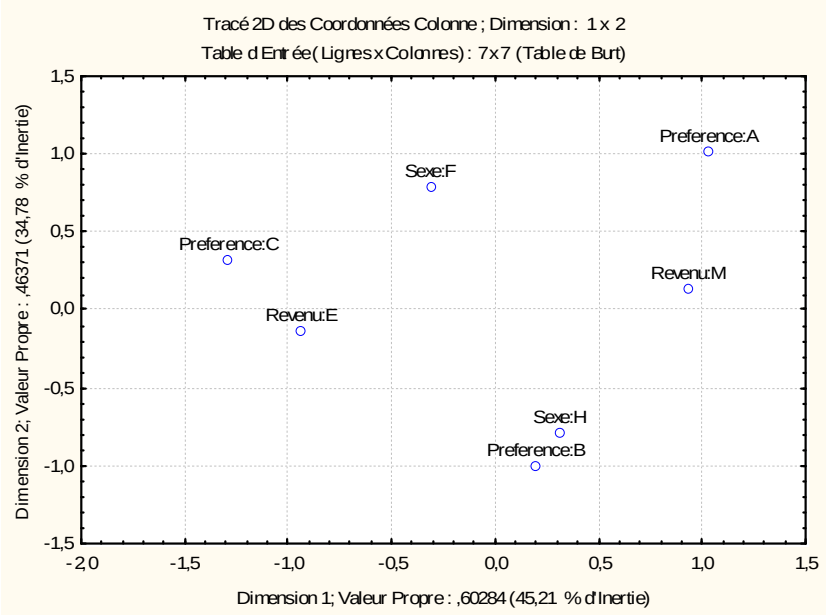
La moyenne des valeurs propres est : $\lambda_m = \frac{1}{3} = 0{,}33$, ce qui conduit à ne conserver que les 2 premières valeurs propres. La transformation précédente donne alors :

Nb de dim.	Val Prop.	$\lambda' = (\lambda - \lambda_m)^2$	Taux d'inertie modifié
1	0,6028	0,0726	81,04%
2	0,4637	0,0170	18,96%
3	0,2030		
4	0,0638		

Le taux d'inertie modifié moyen est de 100%/2= 50%. Seule la première valeur propre dépasse ce taux, mais une étude limitée seulement au premier axe principal présenterait peu d'intérêt. Nous étudierons donc les deux premiers.

Remarque : Selon Benzécri, les taux modifiés représentent l'écart du nuage de points par rapport au nuage parfaitement sphérique qui serait obtenu si aucun lien n'existait entre les modalités.

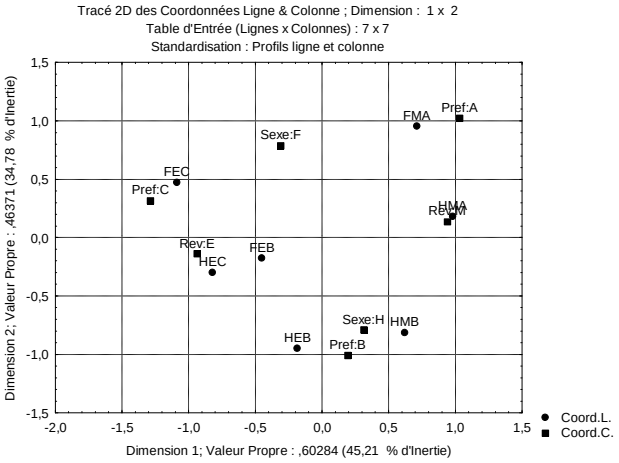
	Coordonnées Colonne et Contributions à l'Inertie (Protocole dans Mini-ACM-v7.stw Table d'Entrée (Lignes x Colonnes) : 7 x 7 (Table de Burt) Inertie Totale = 1,3333									
NomLigne	Ligne Numéro	Coord. Dim.1	Coord. Dim.2	Masse	Qualité	Inertie Relative	Inertie Dim.1	Cosinus Dim.1	Inertie Dim.2	Cosinus Dim.2
Sexe:F	1	-0,311	0,788	0,167	0,718	0,125	0,027	0,097	0,223	0,621
Sexe:H	2	0,311	-0,788	0,167	0,718	0,125	0,027	0,097	0,223	0,621
Revenu:M	3	0,938	0,138	0,167	0,899	0,125	0,243	0,880	0,007	0,019
Revenu:E	4	-0,938	-0,138	0,167	0,899	0,125	0,243	0,880	0,007	0,019
Preference:A	5	1,032	1,024	0,100	0,906	0,175	0,177	0,456	0,226	0,450
Preference:B	6	0,193	-1,007	0,133	0,701	0,150	0,008	0,025	0,292	0,677
Preference:C	7	-1,288	0,319	0,100	0,755	0,175	0,275	0,711	0,022	0,044



Bien que l'exemple ne comporte qu'un petit nombre d'observations, on remarque la proximité des modalités Préférence:B et Sexe:H, de même que l'opposition Préférence C, revenu E d'une part, Préférence A, Revenu M d'autre part selon le premier axe.

On note également que l'origine du repère est le milieu du segment joignant les deux modalités de la variable "Sexe", et aussi le milieu du segment joignant les deux modalités de la variable "Revenu". En effet, ces deux variables ont seulement deux modalités (d'où l'alignement de l'origine avec les modalités) et les deux modalités sont équiprobables (d'où la propriété du milieu).

La représentation du nuage de points représentant simultanément les modalités et les patrons de réponses est la suivante :



L'étude menée à partir du tableau de Burt mérite un commentaire particulier. En effet, dans un exposé théorique sur l'ACM, tels que ceux de [Crucianu] ou de [Rouanet, Le Roux], l'analyse du tableau de Burt est distinguée de celle du TDC ou du tableau disjonctif des patrons. Il est notamment indiqué que les valeurs propres produites par cette analyse sont les carrés des valeurs propres précédentes, et que le Phi-2 du tableau de Burt n'est pas celui du TDC. Cependant, les représentations graphiques produites (limitées aux seules modalités) peuvent être interprétées de façon analogue.

Lorsque l'on effectue une AFC en spécifiant le tableau de Burt comme tableau de contingence, on retrouve alors les résultats indiqués dans les exposés théoriques. Par exemple, le tableau des valeurs propres est alors donné par :

Nombre de Dims.	ValSing.	ValProp.	%age Inertie	%age Cumulé	Chi²
Valeurs Propres et Inertie de toutes les Dimensions (Tableau de Burt dans Mini-ACM-v7.stw) Table d'Entrée (Lignes x Colonnes) : 7 x 7 Inertie Totale = ,62370 Chi² = 56,133 dl = 36 p = ,01747					
1	0,6028	0,3634	58,2668	58,2668	32,7071
2	0,4637	0,2150	34,4755	92,7423	19,3523
3	0,2030	0,0412	6,6045	99,3468	3,7073
4	0,0638	0,0041	0,6532	100,0000	0,3667
5	0,0000	0,0000	0,0000	100,0000	0,0000
6	0,0000	0,0000	0,0000	100,0000	0,0000

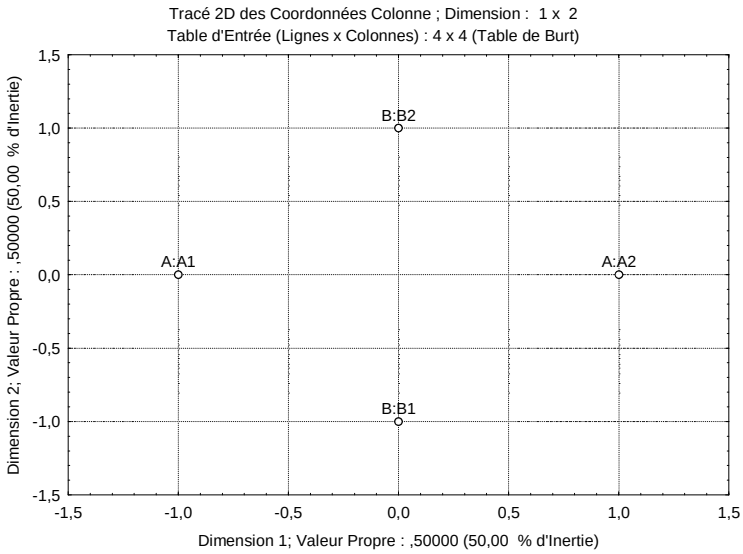
2.4.5 Exploration de l'ACM sur des mini-exemples

Etudions comment seront disposées les modalités colonnes lorsque la situation comporte 2 questions A et B à 2 modalités chacune (respectivement A1 et A2, B1 et B2). L'espace de représentation est alors de dimension 2, autrement dit, l'ACM produit une représentation non déformée dans un plan.

Cas 1 : les effectifs des modalités sont donnés par :

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

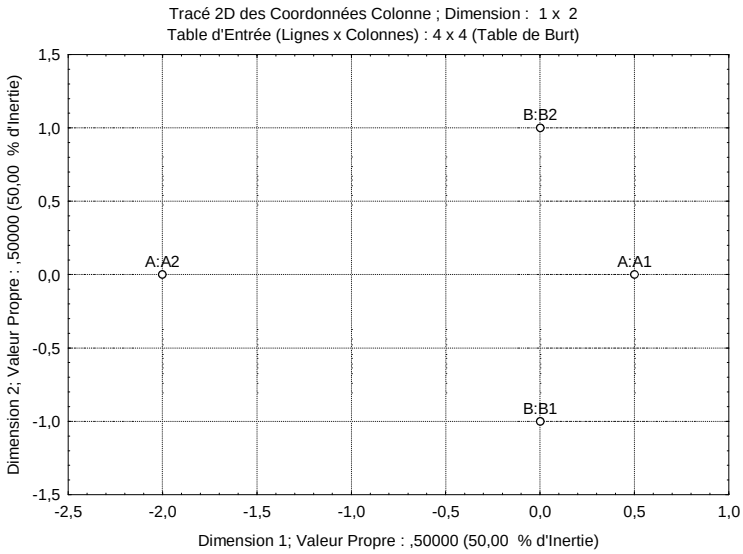
La représentation est alors :



Cas 2 : les effectifs des modalités sont donnés par :

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

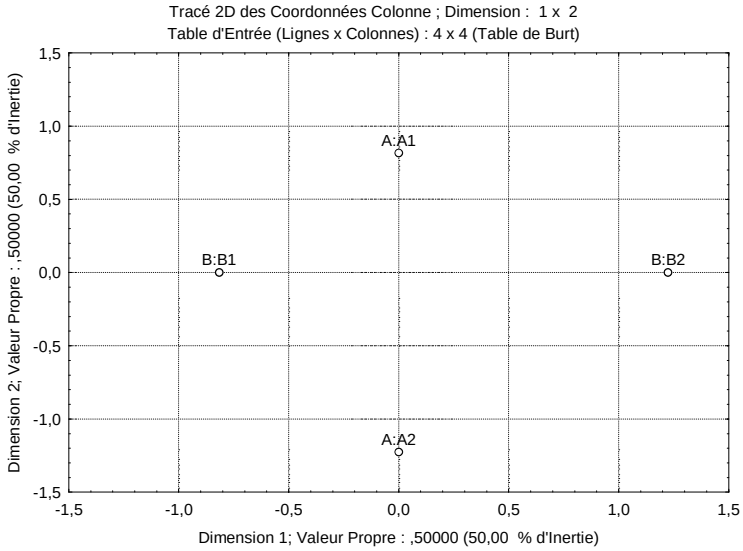
A2 est alors plus éloigné de O que A1. B1 et B2 sont à égale distance de O, et cette distance est intermédiaire entre celle de A1 et celle de A2. La représentation est alors :



Cas 3 : les effectifs des modalités sont donnés par :

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

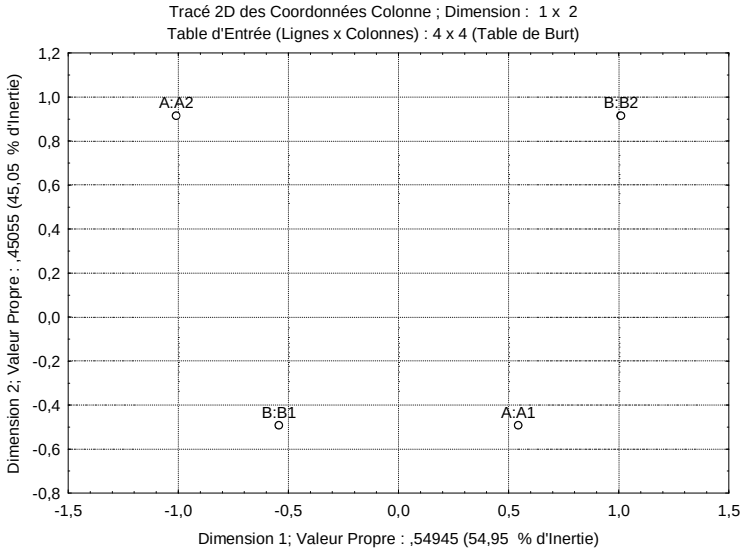
A2 est alors plus éloigné de O que A1. A1 et B1 sont à égale distance de O, et cette distance est intermédiaire entre celle de A1 et celle de A2. La représentation est alors :



Cas 4 : les effectifs des modalités sont donnés par :

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

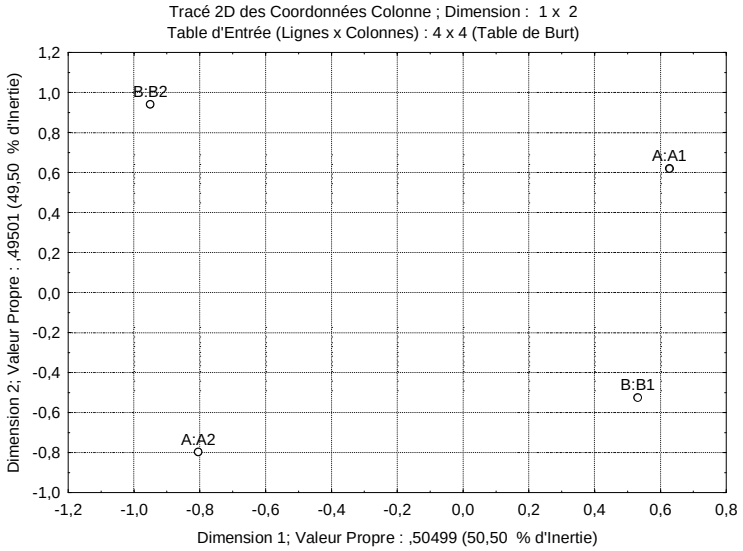
La situation est en apparence analogue à la précédente. En fait l'inertie due aux combinaisons de modalités l'emporte ici sur celle liée aux questions, et on obtient :



Cas 5 : les effectifs des modalités sont donnés par :

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

C'est la situation la plus générale. On obtient :



2.4.6 ACM avec Statistica

Comme l'indiquent Rouanet et Le Roux :

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

Quelle que soit la forme des données d'entrée, l'ACM sera réalisée à partir du menu Statistiques - Techniques exploratoires multivariées - Analyse des correspondances. Mais, selon la structure des

données, c'est l'onglet "Analyse de correspondances" ou l'onglet "Analyse des correspondances multiples (ACM)" qui sera utilisé, selon le tableau suivant :

Format des données	Onglet "Analyse des Correspondances"	Onglet "Analyse des Correspondances Multiple"	Observations
Tableau protocole	Non	Oui	AFC impossible si plus de 2 variables
Tableau d'effectifs	Non	Oui	AFC impossible si plus de 2 variables
Tableau Disjonctif Complet	Oui	Non	
Tableau Disjonctif des patrons	Oui	Non	
Tableau de Burt	Oui	Oui	Les deux analyses ne fournissent pas les mêmes résultats

Exemple.

Ref. L'exemple qui suit est accessible sur Internet à partir des adresses :
http://www.skeptron.uu.se/broady/sec/k-10-gda.htm
http://www.math-info.univ-paris5.fr/~lerb/livres/MCA/MCA_en.html
Le Roux, B., Rouanet, H., Savage, M., Warde, A., Class and Cultural Division in the UK, Sociology 2008, No 42, pp.1042-1071
http://soc.sagepub.com/content/42/6/1049

Il s'agit vraisemblablement de données recueillies dans le cadre de l'enquête "Cultural Capital and Social Exclusion" (CCSE) administrée en 2003 et 2004 au Royaume-Uni par le National Centre for Social Research.
Le questionnaire comportait notamment les questions suivantes :

Q1 : do you prefer leisure activities you ca
- Leisure:friends
- Leisure:family
- Leisure:alone
- Leisure:partner

Q2 : Would you say that during your free time
- lack time
- always sth to do
- Stimes nothing to do
- often do nothing

Q3 : If you had more time, 1st choice would b
- home DIY
- artistic activities
- to rest
- develop knowledge
- physical activities
- take care of family
- to take courses

Q4 : When you go out in the evening, do you u
- GoingOut:friends
- GoingOut:alone
- GoingOut:partner
- don't go out
- GoingOut:family

Q5 : Time watching TV (hours by week)
- TV:never
- TV:<10h
- TV:[10;19h[
- TV:[19;30h[
- TV:>=30h

Q6 : # of books or comic trips during last 12
- no book
- 1-4 books
- 5-12 books
- 13 -39 books
- 40 books or more

QS1 : Gender
- women
- men

QS2 : Education level
- no degree
- CEP
- CAP-BEP
- BEPC
- Bac
- Bac+2
- >Bac+2
- Students

QS3 : Age
- <=25
- 25-35
- 35-45
- 45-55
- 55-65
- >65

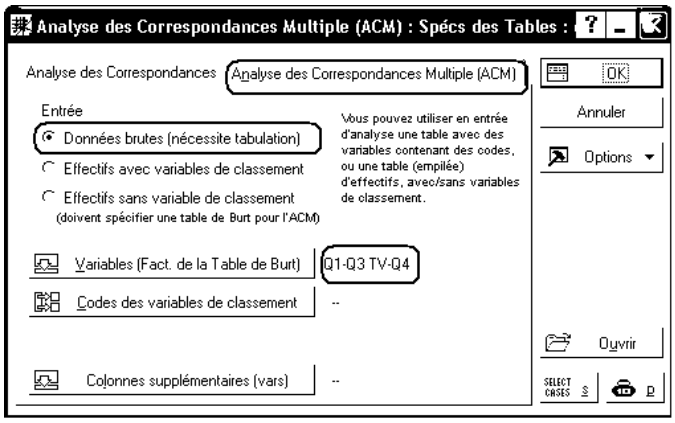
QS4 : PCS
- Femmes au foyer
- Retraités
- Etudiants, élèves
- Autres inactifs
- Cadres et profession
- Employés
- Ouvriers non qualifi
- Professions interméd
- Artisans, commerçant

Ouvriers qualifiés
Agriculteurs

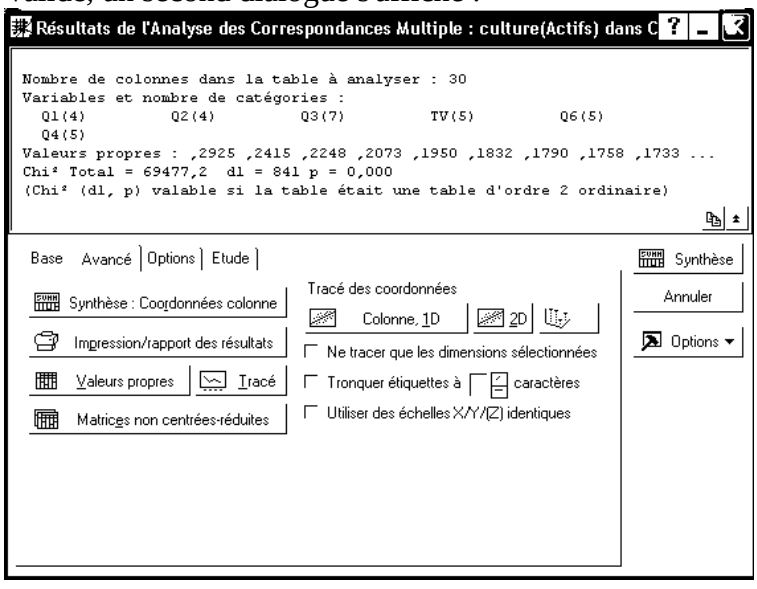
Ouvrez le fichier Culture.stw. Les données y sont saisies sous forme de tableau protocole. Deux jeux de données sont disponibles :

- la feuille de données Culture contient les résultats relatifs à 3002 observations, mais un certain nombre de réponses aux questions qui feront l'objet de l'ACM (Q1, Q2, Q3, Q4, TV et Q6) sont incomplètes, et la variable ISUP indique que les individus correspondants sont rendus "inactifs".
- la feuille de données Culture (Actifs) contient les réponses des 2720 individus actifs.

Réalisons, par exemple, une ACM sur les variables Q1, Q2, Q3, Q4, TV et Q6, à partir du tableau protocole. Après avoir déclaré cette feuille de données comme 'feuille active", on sélectionne l'onglet "Analyse des correspondances multiple" et on complète le premier dialogue comme suit :



Une fois ce dialogue validé, un second dialogue s'affiche :



Le bouton "Effectifs Observés de l'onglet "Etude" permet d'obtenir un tableau similaire au tableau de Burt. Les pourcentages ligne, pourcentages colonne, khi-2, etc utilisent ce dernier tableau.

L'onglet "Avancé" permet d'obtenir les autres résultats :

Valeurs Propres et Inertie de toutes les Dimensions (culture(Actifs) dans Culture.stw
Table d'Entrée (Lignes x Colonnes) : 30 x 30 (Table de Burt)
Inertie Totale = 4,0000

Nombre de Dims.	ValSing.	ValProp.	%age Inertie	%age Cumulé	Chi²
1	0,5409	0,2925	7,3131	7,3131	5080,9201
2	0,4914	0,2415	6,0374	13,3504	4194,5915
3	0,4742	0,2248	5,6210	18,9714	3905,2809
4	0,4553	0,2073	5,1814	24,1528	3599,8892
5	0,4416	0,1950	4,8745	29,0273	3386,6677
6	0,4281	0,1832	4,5808	33,6081	3182,5987
7	0,4230	0,1790	4,4738	38,0818	3108,2367
8	0,4193	0,1758	4,3960	42,4779	3054,2426
9	0,4162	0,1733	4,3313	46,8092	3009,2644
10	0,4108	0,1688	4,2197	51,0288	2931,6975
11	0,4068	0,1655	4,1363	55,1651	2873,7570
12	0,4007	0,1605	4,0132	59,1783	2788,2769
13	0,3991	0,1593	3,9821	63,1605	2766,6609
14	0,3952	0,1562	3,9042	67,0646	2712,5159
15	0,3905	0,1525	3,8123	70,8769	2648,6693
16	0,3881	0,1506	3,7658	74,6427	2616,3419
17	0,3801	0,1445	3,6113	78,2540	2509,0210
18	0,3777	0,1427	3,5665	81,8205	2477,8935
19	0,3691	0,1362	3,4062	85,2267	2366,5577
20	0,3618	0,1309	3,2734	88,5000	2274,2318
21	0,3535	0,1249	3,1235	91,6236	2170,1442
22	0,3469	0,1203	3,0083	94,6319	2090,0819
23	0,3417	0,1168	2,9198	97,5517	2028,5967
24	0,3129	0,0979	2,4483	100,0000	1701,0171

Taux d'inertie modifiés

On ne conserve que les valeurs propres λ supérieures à λ_m et on calcule pour chacune d'entre elles : $\lambda' = (\lambda - \lambda_m)^2$. Le taux d'inertie modifié est alors calculé par : $\frac{\lambda'}{\sum \lambda'}$ et on conserve les valeurs propres dont le taux modifié est supérieur à la moyenne (des taux modifiés). Pour l'exemple traité, l'application de cette méthode donne les résultats suivants :

	ValSing.	ValProp.	Val. Prop. Modifiées	%age modifié	Cumuls
1	0,5409	0,2925	0,0158	56,92%	56,92%
2	0,4914	0,2415	0,0056	20,12%	77,04%
3	0,4742	0,2248	0,0034	12,16%	89,20%
4	0,4553	0,2073	0,0016	5,92%	95,12%
5	0,4416	0,1950	0,0008	2,88%	98,00%
6	0,4281	0,1832	0,0003	0,99%	98,98%
7	0,4230	0,1790	0,0002	0,54%	99,53%
8	0,4193	0,1758	0,0001	0,30%	99,83%
9	0,4162	0,1733	0,0000	0,16%	99,98%
10	0,4108	0,1688	0,0000	0,02%	100,00%
11	0,4068	0,1655			
12	0,4007	0,1605			
13	0,3991	0,1593			
14	0,3952	0,1562			

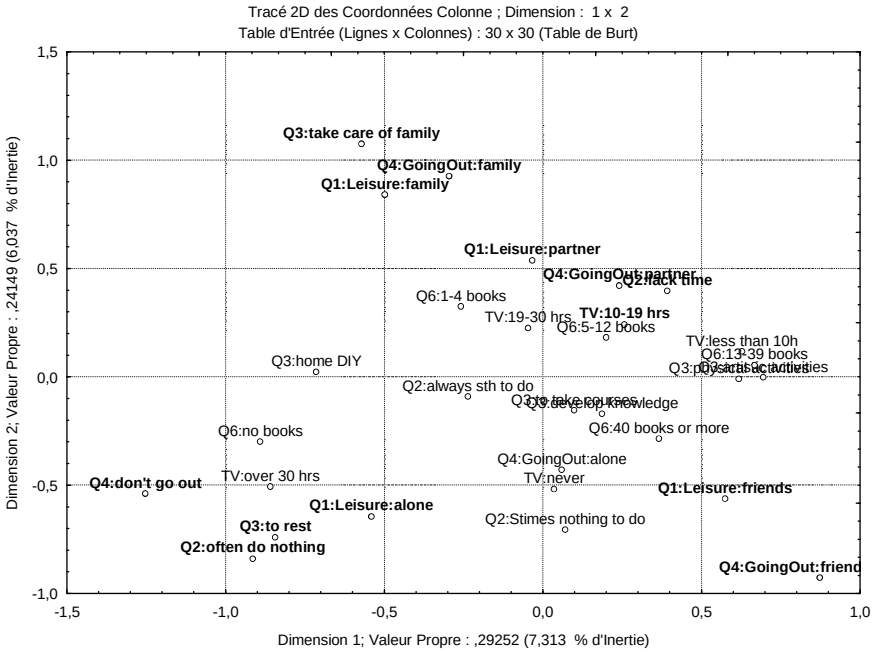
15	0,3905	0,1525			
16	0,3881	0,1506			
17	0,3801	0,1445			
18	0,3777	0,1427			
19	0,3691	0,1362			
20	0,3618	0,1309			
21	0,3535	0,1249			
22	0,3469	0,1203			
23	0,3417	0,1168			
24	0,3129	0,0979			
		0,1667	0,0278		

Le taux d'inertie modifié moyen est de 100%/10= 10%. Les trois premières valeurs propres modifiées dépassent ce taux. Nous étudierons ici les deux premiers axes.

Les coordonnées, contributions et qualités de représentation sont données dans le tableau ci-dessous.

	Coordonnées Colonne et Contributions à l'Inertie (culture(Actifs) dans Culture.stw) Table d'Entrée (Lignes x Colonnes) : 30 x 30 (Table de Burt) Inertie Totale = 4,0000									
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
Q1:Leisure:partner	1	-0,033	0,537	0,031	0,066	0,034	0,000	0,000	0,037	0,066
Q1:Leisure:friends	2	0,575	-0,563	0,065	0,413	0,025	0,073	0,211	0,085	0,202
Q1:Leisure:family	3	-0,498	0,841	0,044	0,343	0,031	0,037	0,089	0,129	0,254
Q1:Leisure:alone	4	-0,540	-0,645	0,027	0,134	0,035	0,027	0,055	0,046	0,079
Q2:always sth to do	5	-0,236	-0,091	0,071	0,047	0,024	0,013	0,041	0,002	0,006
Q2:lack time	6	0,392	0,396	0,069	0,218	0,024	0,036	0,108	0,045	0,110
Q2:Stimes nothing to d	7	0,071	-0,707	0,015	0,049	0,038	0,000	0,000	0,031	0,049
Q2:often do nothing	8	-0,914	-0,840	0,012	0,123	0,039	0,035	0,067	0,036	0,056
Q3:home DIY	9	-0,715	0,022	0,026	0,094	0,035	0,045	0,094	0,000	0,000
Q3:artistic activities	10	0,695	-0,003	0,024	0,082	0,036	0,040	0,082	0,000	0,000
Q3:to rest	11	-0,843	-0,741	0,019	0,159	0,037	0,045	0,089	0,042	0,069
Q3:develop knowledge	12	0,187	-0,171	0,028	0,013	0,035	0,003	0,007	0,003	0,006
Q3:physical activities	13	0,619	-0,010	0,035	0,102	0,033	0,046	0,102	0,000	0,000
Q3:take care of family	14	-0,571	1,076	0,019	0,195	0,037	0,022	0,043	0,093	0,152
Q3:to take courses	15	0,099	-0,155	0,016	0,004	0,038	0,001	0,001	0,002	0,003
TV:less than 10h	16	0,628	0,117	0,027	0,078	0,035	0,036	0,075	0,002	0,003
TV:19-30 hrs	17	-0,046	0,225	0,043	0,019	0,031	0,000	0,001	0,009	0,018
TV:10-19 hrs	18	0,258	0,241	0,049	0,051	0,030	0,011	0,027	0,012	0,024
TV:over 30 hrs	19	-0,859	-0,507	0,032	0,240	0,034	0,082	0,178	0,034	0,062
TV:never	20	0,035	-0,519	0,016	0,028	0,038	0,000	0,000	0,018	0,028
Q6:no books	21	-0,891	-0,299	0,037	0,251	0,032	0,100	0,226	0,014	0,025
Q6:5-12 books	22	0,199	0,182	0,039	0,022	0,032	0,005	0,012	0,005	0,010
Q6:1-4 books	23	-0,258	0,325	0,030	0,037	0,034	0,007	0,014	0,013	0,023
Q6:40 books or more	24	0,366	-0,286	0,026	0,041	0,035	0,012	0,025	0,009	0,015
Q6:13-39 books	25	0,667	0,053	0,034	0,117	0,033	0,053	0,116	0,000	0,001
Q4:don't go out	26	-1,253	-0,540	0,029	0,388	0,034	0,154	0,327	0,035	0,061
Q4:GoingOut:friends/o	27	0,873	-0,928	0,036	0,452	0,033	0,095	0,212	0,129	0,240
Q4:GoingOut:alone	28	0,060	-0,428	0,012	0,015	0,039	0,000	0,000	0,009	0,015
Q4:GoingOut:partner	29	0,242	0,421	0,056	0,119	0,028	0,011	0,029	0,041	0,089
Q4:GoingOut:family	30	-0,296	0,926	0,033	0,237	0,033	0,010	0,022	0,119	0,215

Dans le graphique suivant, les modalités qui ont une contribution supérieure à 3,3% à la formation du premier axe sont indiquées en caractères rouges, celles qui ont une contribution supérieure à la moyenne sur le second axe sont représentées en caractères gras.



2.4.7 Autres exemples d'ACM

Les autres exemples d'ACM que nous traiterons sont donnés à l'aide d'un tableau de Burt. En effet, c'est généralement sous cette forme que l'on trouve des données susceptibles de servir de base à un exercice.

2.4.7.1 Le cas "Aspirations des Français"

Ouvrez le classeur Aspi.stw. La présentation du cas, rappelée dans un rapport contenu dans le classeur est la suivante :

Source : Morineau A., Morin S., Pratique du traitement des enquêtes - Exemple d'utilisation du système SPAD, Cisia-Ceresta, Montreuil, 2000

On travaille sur des données extraites d'une enquête d'opinion réalisée en 1978, concernant les conditions de vie et les aspirations des Français.
Les questions prises en compte ici, et leurs modalités, sont les suivantes :

1- Sexe de la personne interrogée :
- masc : masculin
- femi : féminin

2- Possédez-vous des valeurs mobiliaires
- vmo1 : oui
- vmo2 : non

3- Taille d'agglomération
- agg1 : moins de 2000 h
- agg2 : de 2000 à 20000 h
- agg3 : de 20000 à 100000 h
- agg4 : plus de 100000h

agg5 : Paris

4- Diplome de l'enquêté :
- die1 : aucun
- die2 : CEP ou fin d'études
- die3 : BEPC - BE - BEPS
- die4 : bac - brevet sup.
- die5 : université, gde école

5- Statut du logement
- slo1 : en accession
- slo2 : propriétaire
- slo3 : locataire
- slo4 : logé gratuit, autre

6- Age de l'enquêté
- agc1 : moins de 25 ans
- agc2 : 25 à 34 ans
- agc3 : 35 à 49 ans
- agc4 : 50 à 64 ans
- agc5 : plus de 65 ans

7- Type d'emploi
- emp1 : ouvriers
- emp2 : employés
- emp3 : cadres
- emp4 : autres
- empNR : non réponse

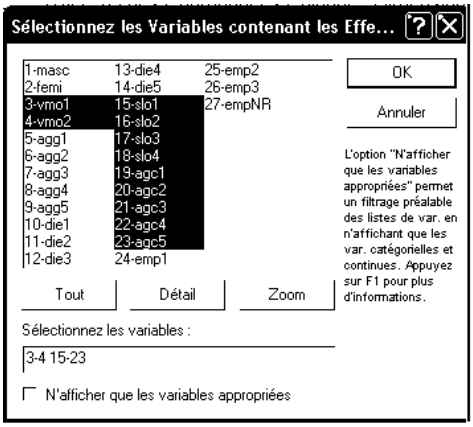
Remarque : pour une ACM sur la totalité des 27 modalités du TDB, les auteurs retiennent 5 axes principaux.

Faites tout d'abord une ACM sur la totalité du tableau de Burt (27 modalités - remarquez que seules 4 modalités de la variable "Type d'emploi" sont présentes.

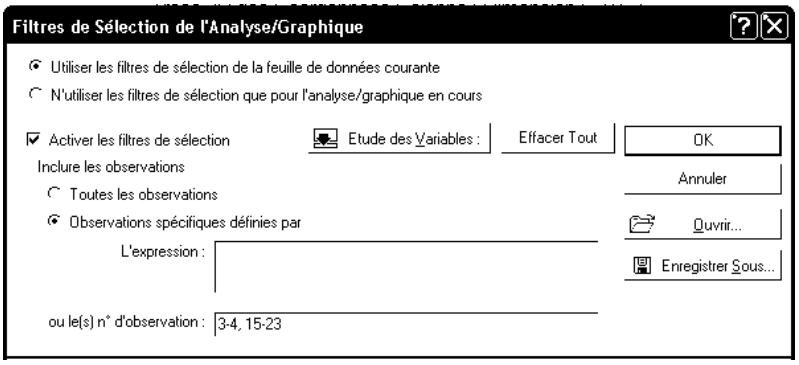
Remarque : le graphique ainsi obtenu est assez peu lisible. Il est cependant possible de l'améliorer en utilisant l'outil "Balayage/Habillage" :  . A l'aide de cet outil, il est par exemple possible de supprimer certains points qui se superposent au centre du graphique. Attention cependant à ce que le graphique conserve une certaine honnêteté intellectuelle !

Réalisez ensuite une ACM en ne prenant en compte que certaines variables, par exemple, la variable 2 (valeurs mobiliaires), la variable 5 (statut du logement) et la variable 6 (âge de l'enquêté). Pour cela :

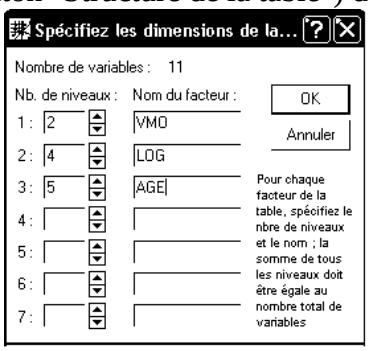
- Sélectionnez les variables comme suit :



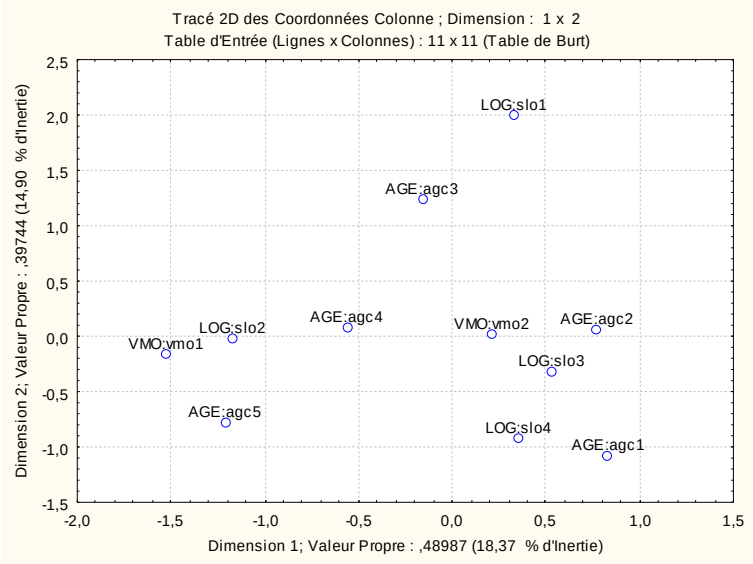
- Sélectionnez ensuite les observations correspondantes, par exemple en les désignant par leurs numéros. Pour cela, cliquez sur le bouton "Select Cases" et complétez le dialogue comme suit :



- Structurez enfin les variables (bouton "Structure de la table") de la façon suivante :



On obtient ainsi le graphique suivant :



La possession de valeurs mobiliaires est ainsi plutôt associée à l'occupation d'un logement en propriété, et à une personne relativement âgée (agc4, agc5), alors que la non-possession est plutôt le fait de personnes jeunes, locataires. L'âge agc3 est dans une certaine mesure associé à l'accession à la propriété alors que le dernier statut du logement est plutôt le fait des moins de 25 ans (qui, par ailleurs, ne possèdent généralement pas de valeurs mobiliaires).

2.4.7.2 Le cas "Avignon"

Source : Croutsche, J.-J., Pratiques statistiques en gestion et études de marchés, Editions ESKA, Paris, 1997

Une enquête sur la fréquentation du centre ville d'Avignon. On trouvera ci-dessous le texte d'une partie des questions posées, ainsi que le codage des modalités de réponse.

1- Combien de fois par mois allez-vous dans le centre ville pour faire des achats ?
- a1 : Plus de 3 fois par mois
- a2 : de 2 à 3 fois
- a3 : de 1 à 2 fois
- a4 : Autre

2- Votre fréquentation du centre ville est-elle plus ou moins importante qu'il y a 5 ans ?
- f1 : Beaucoup moins importante
- f2 : Un peu moins importante
- f3 : Identique
- f4 : Un peu plus importante
- f5 : Beaucoup plus importante

3-

4-

5- Etes-vous satisfait de la propreté du centre ville ?
- p1 : très satisfait
- p2 : satisfait
- p3 : moyennement satisfait
- p4 : peu satisfait
- p5 : très peu satisfait

6- Que pensez-vous de la sécurité dans le centre ville ?
- s1 : Très faible
- s2 : Faible
- s3 : Normale
- s4 : Importante
- s5 : Très importante

7- Si vous observez des problèmes de sécurité : vous arrive-t-il de ne pas vous rendre dans le centre ville à cause de ce problème ?
- r1 : oui
- r2 : non

8-

9-

10-

11- Où habitez-vous ?
- h1 : Avignon intra-muros
- h2 : Avignon extra-muros
- h3 : autre

12-

13- Dans quelle tranche d'âge vous situez-vous ?
- â1 : 15-19 ans
- â2 : 20-30 ans
- â3 : 31-40 ans
- â4 : 41-50 ans
- â5 : 51-60 ans
- â6 : Plus de 60 ans

14-

Dans le classeur Avignon.stw se trouvent diverses feuilles de données contenant les tableaux de Burt obtenus en sélectionnant 3 ou 4 des items du questionnaire. Analysez chacun des aspects ainsi définis à l'aide d'une ACM.

2.5 Méthodes de classification

Bibliographie : Lebart, L., Morineau, A., Piron M., Analyse exploratoire multidimensionnelle, Dunod, Paris, 2000.

2.5.1 Introduction

Classifier, c'est regrouper entre eux des objets similaires selon tel ou tel critère. Les diverses techniques de classification (ou d'"analyse typologique", de "taxonomie", ou "taxinomie" ou encore "analyse en clusters" (amas)) visent toutes à répartir *n* individus, caractérisés par *p* variables X_1, X_2, ..., X_p en un certain nombre m de sous-groupes aussi homogènes que possible.
On distingue deux grandes familles de techniques de classification :

- La classification non hiérarchique ou partitionnement, aboutissant à la décomposition de l'ensemble de tous les individus en m ensembles disjoints ou classes d'équivalence ; le nombre m de classes est fixé.
- La classification hiérarchique : pour un niveau de précision donné, deux individus peuvent être confondus dans un même groupe, alors qu'à un niveau de précision plus élevé, ils seront distingués et appartiendront à deux sous-groupes différents.

Remarques. Ces méthodes jouent un rôle un peu à part dans l'univers des méthodes statistiques. En effet :

- L'aspect inférentiel est ici inexistant ;
- Il existe un grand nombre de variantes de ces méthodes, et on peut être amené à appliquer plusieurs de ces méthodes sur un même jeu de données, jusqu'à obtenir une classification "qui fasse sens" ;
- Au contraire des méthodes factorielles, l'accent est souvent mis sur les n individus et non sur les p variables qui les décrivent.

2.5.2 Méthodes de type "centre mobile" : K-moyennes

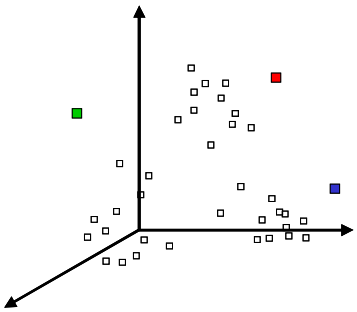
2.5.2.1 Principe de la méthode

On dispose d'un ensemble d'individus, ou observations, décrits par des variables numériques. On veut créer une partition de cet ensemble, en regroupant ces individus en un nombre déterminé K de classes : chaque individu devra appartenir à une classe et une seule. Pour cela :

On fixe de façon aléatoire K "centres de classes", ou "centres de gravité" et on exécute l'algorithme suivant :

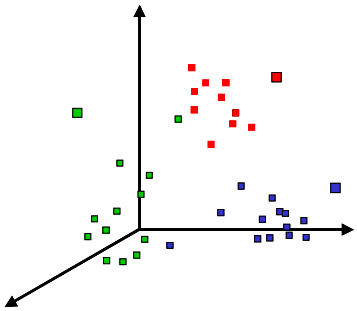
1) Chaque observation est classée en fonction de sa proximité au centre de gravité.
2) Chaque centre de gravité est déplacé de façon à être au centre du groupe correspondant.
3) On continue jusqu'à ce que les centres de gravité ne bougent plus

Méthodes de type « centres mobiles »



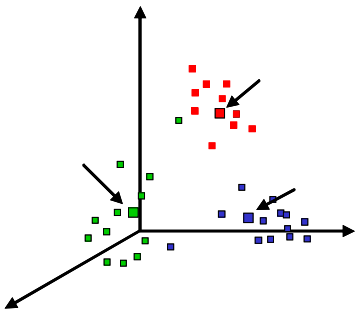
Au départ

Création aléatoire de centres de gravité.



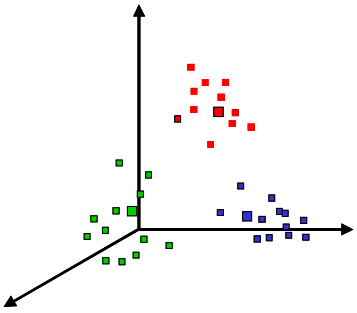
Etape 1

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



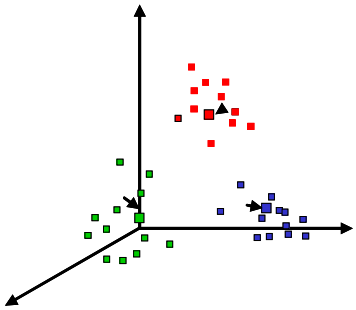
Etape 2

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



Etape 1’

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



Etape 2'

De nouveau, chaque centre de gravité est recalculé

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

Choix des variables représentant les individus

Les distances étant calculées sur les valeurs observées des variables, la classification n'aura pas de sens si les variables s'expriment avec des unités différentes, et ont des plages de variation très différentes. Si c'est le cas, il faut au préalable transformer les variables (par exemple en faisant un centrage-réduction) afin d'équilibrer les "poids" des différentes variables.

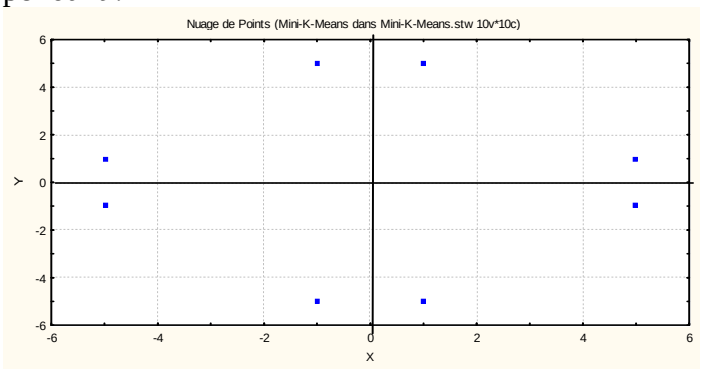
Dans le cas où les données observées sont les valeurs de p variables numériques sur n individus, on pourra choisir d'effectuer une classification des individus, ou une classification des variables. On peut choisir, par exemple, de retenir certains "traits" des individus (autrement dit certaines variables qui ont servi à les décrire) et réaliser la classification sur les individus décrits par ce choix de variables.

2.5.2.2 Mise en oeuvre avec Statistica sur un mini-exemple

On dispose de 8 individus décrits par 2 variables. Une troisième variable est constante sur l'ensemble des individus. Les données sont les suivantes :

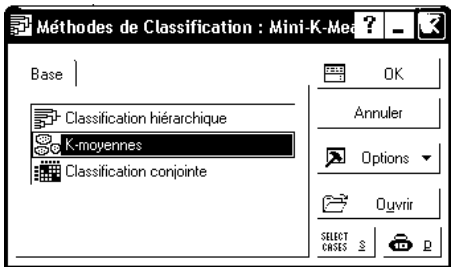
	X	Y	Z
1	5	1	10
2	5	-1	10
3	1	5	10
4	-1	5	10
5	-5	1	10
6	-5	-1	10
7	1	-5	10
8	-1	-5	10

Nuage de points correspondant :

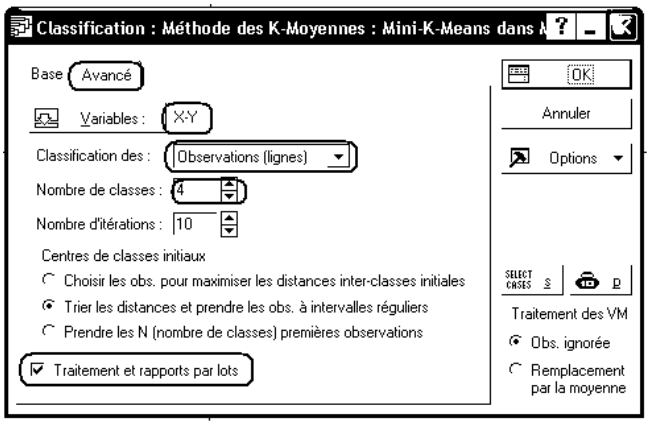


Nous souhaitons définir 4 classes à partir de ces 8 objets.

Ouvrez le classeur Mini-K-Means.stw.
Utilisez le menu Statistiques - Techniques exploratoires multivariées - Classifications et sélectionnez la méthode K-moyennes.



Sélectionnez X et Y comme variables d'analyse, et, sous l'onglet "Avancé", spécifiez une classification sur les observations, comportant 4 classes. Cochez également la case "traitements et rapports par lots", ce qui permettra de produire en une seule manipulation l'ensemble des résultats de la classification.



Comme prévu, les 4 classes formées par Statistica sont {O1, O2}, {O3, O4}, {O5, O6} et {O7, O8} (cf. les 4 feuilles de résultats "composition de la classe N° ...). Par exemple, pour la première classe :

	Composition de la Classe 1 et Distances au Centre de Classe Respecti Classe avec 2 obs.	
	Obs. # O_1	Obs. # O_2
Distance	0,707	0,707

Le centre C_1 de cette classe est évidemment le point de coordonnées (5, 0). On peut remarquer que la distance calculée par Statistica n'est pas tout à fait la distance euclidienne dans le plan, mais correspond à la formule suivante :

$$d^2(O_1, C_1) = \frac{(x_1 - \bar{x})^2 + (y_1 - \bar{y})^2}{2}$$

Le dénominateur introduit dans la formule représente le nombre de variables, comme on peut s'en rendre compte en introduisant la troisième variable (Z) dans la classification.

La même règle est appliquée pour le calcul des distances entre classes, autrement dit entre centres de classes :

Classe (Numéro)	Distances Euclidiennes Inter-Classes Dist. sous la diagonale (Dist.) ² au dessus de la diagonale			
	N° 1	N° 2	N° 3	N° 4
N° 1	0,00	25,00	50,00	25,00
N° 2	5,00	0,00	25,00	50,00
N° 3	7,07	5,00	0,00	25,00
N° 4	5,00	7,07	5,00	0,00

Les coordonnées des centres de classes sont disponibles dans la feuille de résultats "Moy. Classes" :

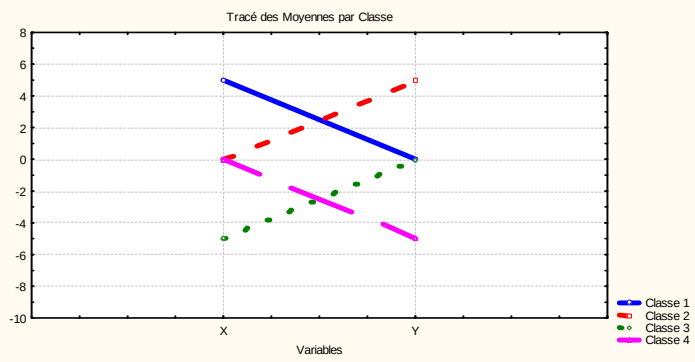
Variable	Moy. Classes (Mini-K-Means dans Mini-K-Means.stw)			
	Classe N° 1	Classe N° 2	Classe N° 3	Classe N° 4
X	5,00	0,00	-5,00	0,00
Y	0,00	5,00	0,00	-5,00

Statistica effectue également une analyse de variance à un facteur sur chacune des variables. Le facteur pris en compte ici est l'appartenance de l'observation à l'une des classes :

Variable	Analyse de Variance					
	SC Inter	dl	SC Intra	dl	F	signif. p
X	100,00	3	4,00	4	33,33	0,0027
Y	100,00	3	4,00	4	33,33	0,0027

Ces résultats peuvent être retrouvés à l'aide du menu ANOVA. On introduit une quatrième variable, nommée "Groupe", contenant le numéro de la classe à laquelle appartient l'observation. Puis, on effectue une analyse de variance à un facteur en indiquant X (par exemple) comme variable dépendante et Groupe comme variable de classement.

Le seul résultat qui n'est pas automatiquement produit par le traitement par lots est le graphique des moyennes. Pour l'obtenir, ré-affichez la fenêtre du traitement en cours, désactivez la case "traitement par lots" et cliquez sur OK. Dans la fenêtre de dialogue "Résultats de l'analyse par les k-moyennes", cliquez sur le bouton "Graphique" :



2.5.2.3 Mise en oeuvre sur les exemples traités dans les paragraphes ACP et AFC

Classification des variables du cas "Représentations sociales de l'homosexualité"

On reprend l'exemple "Représentations sociales de l'homosexualité" que nous avions traité par une ACP (classeur Statistica Rep-Soc-Homo.stw). Rappelons que les variables sont ici homogènes, puisque chaque variable est un protocole de rangs observés sur les 15 traits étudiés.

Une classification en 3 classes, portant sur les variables va-t-elle confirmer les résultats que nous avions obtenus en analysant les résultats de l'ACP ?

	Repr-Soc-Homo dans Rep-Soc-Homo-correction.st		
	1 VARIABL	2 CLASSE	3 DISTANC
He:H	1	3	1,79
Ho:H	2	3	1,43
He:Soi	3	3	2,36
Ho:Soi	4	1	1,21
Ho:Ho	5	1	1,21
Ho:F	6	2	1,48
He:Ho	7	2	1,10
He:F	8	2	1,42

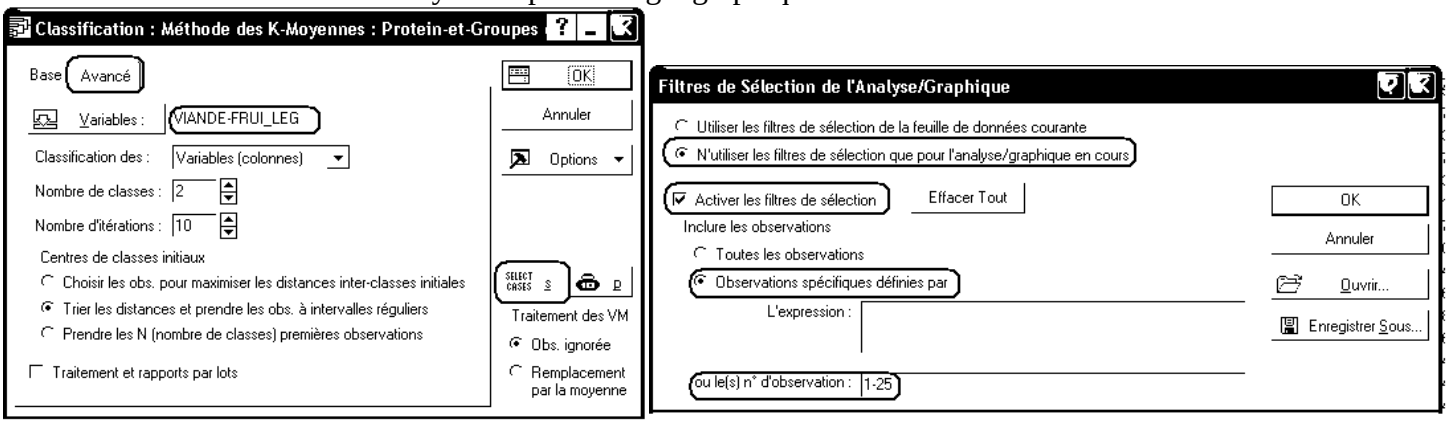
On constate que la classe 3 regroupe les variables correspondant à une cible masculine, la classe 1 regroupe les jugements portés par les homosexuels sur eux-mêmes et sur leur stéréotype, tandis que la classe 2 rassemble non seulement les variables correspondant à une cible féminine mais aussi He:Ho, c'est-à-dire la description de la cible "homosexuels" faite par les hétérosexuels.

Classifications sur le cas "Protéines"

On reprend le fichier Proteines-2008.stw.

La répartition en 2 groupes "protéines animales v/s protéines végétales" apparaît-elle naturellement dans les données étudiées ?

Effectuez une classification de type K-moyennes, portant sur les variables 1 à 9 de la feuille "Protein et Groupes" en indiquant deux classes. Faites une sélection des observations, de manière à éliminer de l'étude les moyennes par zone géographique :



On voit que l'une des classes est constituée de la seule variable "céréales" pendant que l'autre classe rassemble les 8 autres variables. En effet, l'étendue de la variable "Céréales" est très différente de celle des autres variables, et le résultat produit ne fait que l'illustrer.

On peut résoudre ce problème soit en travaillant sur des données centrées réduites, soit en utilisant les coordonnées des variables selon les axes factoriels produites par une ACP normée. Par exemple, activez la feuille "Proteines-Centre-Reduit". Reprenez une classification analogue, mais portant sur les variables centrées-réduites. Cette fois, la classification recouvre assez bien l'origine (animale v/s végétale) des protéines, mais les féculents restent regroupés avec les protéines animales :

	Proteines-Centre-Reduit dans Proteines-2008.st\		
	1 VARIABL	2 CLASSE	3 DISTANC
VIANDE	1	2	0,75
PORC_VOL	2	2	0,82
OEUFS	3	2	0,54
LAIT	4	2	0,69
POISSON	5	2	0,94
CEREALES	6	1	0,65
FECULENT	7	2	0,73
NOIX	8	1	0,46
FRUI_LEG	9	1	0,77

Classification des lignes dans le cas "Régions-2001"

On reprend le classeur Statistica Regions-2001.stw.

Une classification basée sur le tableau de contingence n'aurait pas grand sens. En revanche, on peut utiliser les résultats de l'AFC comme données de base pour essayer de faire une classification des régions en 3 ou 4 ensembles.

Refaites au besoin une AFC sur ce tableau de contingence et rendez active la feuille contenant les résultats relatifs aux individus lignes (les régions). Faites ensuite une classification de type "K-moyennes", en utilisant les variables "Coord." de cette feuille et en spécifiant 3 ou 4 classes. Vous devriez retrouver en grande partie la typologie que nous avions obtenue en analysant les résultats de l'AFC.

Remarque. Les résultats de la classification dépendent-ils du nombre d'axes factoriels représentés dans la feuille de résultats de l'AFC ? On pourra essayer de refaire la classification sur les coordonnées factorielles d'un plus grand nombre d'axes, et constater qu'il en résulte peu de modifications des résultats produits : l'essentiel de la variation est représenté par les premiers axes.

2.5.2.4 Remarques et conclusion

Cette méthode produit des résultats qui peuvent être facilement exploitables. On notera cependant que l'on doit indiquer a priori le nombre de classes, ce qui nuit à l'aspect véritablement "exploratoire" de la méthode. D'autre part, les variables traitées doivent être homogènes (s'exprimer avec la même unité, ou au moins avoir la même plage de variation) et c'est toujours la distance euclidienne qui est utilisée pour évaluer les distances entre objets.

2.5.3 Classification Ascendante Hiérarchique

2.5.3.1 Les 4 étapes de la méthode

Choix des variables représentant les individus

Les distances étant calculées sur les valeurs observées des variables, la classification n'aura pas de sens si les variables s'expriment avec des unités différentes, et ont des plages de variation très différentes. Si c'est le cas, il faut au préalable transformer les variables (par exemple en faisant un centrage-réduction) afin d'équilibrer les "poids" des différentes variables.

Dans le cas où les données observées sont les valeurs de p variables numériques sur n individus, on pourra choisir d'effectuer une classification des individus, ou une classification des variables. On peut choisir, par exemple, de retenir certains "traits" des individus (autrement dit certaines variables

qui ont servi à les décrire) et réaliser la classification sur les individus décrits par ce choix de variables.
On peut noter qu'il revient au même par exemple :

- de réaliser la CAH des individus à partir de p variables centrées réduites ;
- de réaliser la CAH des individus à partir des p facteurs obtenus à l'aide d'une ACP normée sur les variables précédentes.

Toutefois, il peut être intéressant de réaliser la CAH à partir des q premiers facteurs (q<p). Cela a pour effet d'éliminer une partie des variations entre individus, qui correspond en général à des fluctuations aléatoires, c'est-à-dire à un "bruit statistique".

Dans le cas où les données observées sont représentées par un tableau de contingence, c'est-à-dire sont les valeurs de 2 variables nominales sur n individus, on pourra effectuer une CAH des modalités-lignes par exemple, à partir des coordonnées lignes obtenues par une AFC. On pourra, de même, réaliser une CAH des modalités-colonnes.

Enfin, si les données observées sont les valeurs de p variables nominales sur n individus, on pourra effectuer une CAH des individus en partant du tableau disjonctif complet, ou en utilisant les coordonnées des individus obtenues par une ACM. On pourra également traiter les modalités comme dans le cas d'une AFC.

Choix d'un indice de dissimilarité

De nombreuses mesures de la "distance" entre individus ont été proposées. Le choix d'une (ou plusieurs) d'entre elles dépend des données étudiées. Statistica nous propose les mesures suivantes :

- Distance Euclidienne. C'est probablement le type de distance le plus couramment utilisé. Il s'agit simplement d'une distance géométrique dans un espace multidimensionnel.
$$d(I_i,I_j)=\sqrt{\sum_k (x_{ik}-x_{jk})^2}$$
- Distance Euclidienne au carré. On peut élever la distance euclidienne standard au carré afin de "sur-pondérer" les objets atypiques (éloignés).
$$d(I_i,I_j)=\sum_k (x_{ik}-x_{jk})^2$$
- Distance du City-block (Manhattan) :
$$d(I_i,I_j)=\sum_k \left|x_{ik}-x_{jk}\right|$$
- Distance de Tchebychev :
$$d(I_i,I_j)=Max\ \left|x_{ik}-x_{jk}\right|$$
- Distance à la puissance.
$$d(I_i,I_j)=\left(\sum_k \left|x_{ik}-x_{jk}\right|^p\right)^{1/r}$$
- Percent disagreement. Cette mesure est particulièrement utile si les données des dimensions utilisées dans l'analyse sont de nature catégorielle.
$$d(I_i,I_j)=\frac{Nombre\ \ de\ \ x_{ik}\neq x_{jk}}{K}$$
- 1- r de Pearson : calculée à partir du coefficient de corrélation, à l'aide de la formule :
$$d(I_i,I_j)=1-r_{ij}$$

Indices de dissimilarité et distances

On peut également utiliser d'autres indices de dissimilarité puisque Statistica permet d'effectuer la classification à partir du tableau des scores de dissimilarités entre individus. En fait, un indice de dissimilarité doit simplement satisfaire les conditions suivantes :

- non-négativité : $d(I_i,I_j) \geq 0$
- symétrie : $d(I_i,I_j) = d(I_j,I_i)$
- normalisation : $d(I_i,I_i) = 0$

Un indice de dissimilarité est une "vraie" distance, s'il vérifie également l'inégalité triangulaire :

$$d(I_i,I_j) \leq d(I_i,I_k) + d(I_k,I_j).$$

La plupart des "distances" proposées par Statistica sont de véritables distances.

De nombreux indices de dissimilarité (ou au contraire de similarité) ont été proposés dans le cas de variables qualitatives (à deux modalités, ou après codage disjonctif). Par exemple, si les individus sont décrits par K variables dichotomiques (oui/non), on peut introduire :

$a_{ij} =$ *Nombre co−occurrences entre les individus i et j*

$d_{ij} =$ *Nombre co−absences entre les individus i et j*

$b_{ij} =$ *Nombre d'attributs présents chez i et absents chez j*

$c_{ij} =$ *Nombre d'attributs absents chez i et présents chez j*

On peut proposer par exemple, comme indice de dissimilarité :

$$d(I_i,I_j) = \sqrt{b_{ij} + c_{ij}}$$

ou au contraire, comme indice de similarité :

$$s(I_i,I_j) = \frac{a_{ij}}{a_{ij} + b_{ij} + c_{ij}}$$

Un indice de similarité peut être converti en distance par la relation :

$$d(I_i,I_j) = s_{\max} - s(I_i,I_j)$$

Choix d'un indice d'agrégation

L'application de la méthode suppose également que nous fassions le choix d'une "distance" entre classes. Là encore, de nombreuses solutions existent. Il faut noter que ces solutions permettent toutes de calculer la distance entre deux classes quelconques sans avoir à recalculer celles qui existent entre les individus composant chaque classe.

Les choix proposés par Statistica sont les suivants :

- Saut minimum ou "single linkage" (distance minimum). C'est celle que nous avons utilisée ci-dessus.

- Diamètre ou "complete linkage" (distance maximum). Dans cette méthode, les distances entre classes sont déterminées par la plus grande distance existant entre deux objets de classes différentes (c'est-à-dire les "voisins les plus éloignés").

$$D(A,B) = \max_{I \in A} \max_{J \in B} d(I,J)$$

- Moyenne non pondérée des groupes associés. Ici, la distance entre deux classes est calculée comme la moyenne des distances entre tous les objets pris dans l'une et l'autre des deux classes différentes.

$$D(A,B) = \frac{1}{n_A n_B} \sum_{I \in A, J \in B} d(I,J)$$

- Moyenne pondérée des groupes associés. La moyenne précédente est étendue à l'ensemble des paires d'objets trouvées dans la réunion des deux classes.

$$D(A,B) = \frac{1}{(n_A + n_B)(n_A + n_B - 1)} \sum_{I,J \in A \cup B} d(I,J)$$

- Centroïde non pondéré des groupes associés. Le centroïde d'une classe est le point moyen d'un espace multidimensionnel, défini par les dimensions. Dans cette méthode, la distance entre deux classes est déterminée par la distance entre les centroïdes respectifs.

- Centroïde pondéré des groupes associés (médiane). Cette méthode est identique à la précédente, à la différence près qu'une pondération est introduite dans les calculs afin de prendre en compte les tailles des classes (c'est-à-dire le nombre d'objets contenu dans chacune).

- Méthode de Ward (méthode du moment d'ordre 2). Cette méthode se distingue de toutes les autres en ce sens qu'elle utilise une analyse de la variance approchée afin d'évaluer les distances entre classes. En résumé, cette méthode tente de minimiser la Somme des Carrés (SC) de tous les couples (hypothétiques) de classes pouvant être formés à chaque étape. Les indices d'agrégation sont recalculés à chaque étape à l'aide de la règle suivante : si une classe M est obtenue en regroupant les classes K et L, sa distance à la classe J est donnée par :

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

La méthode de Ward se justifie bien lorsque la "distance" entre les individus est le carré de la distance euclidienne. Choisir de regrouper les deux individus les plus proches revient alors à choisir la paire de points dont l'agrégation entraîne la diminution minimale de l'inertie du nuage. Le calcul des nouveaux indices entre la paire regroupée et les points restants revient alors à remplacer les deux points formant la paire par leur point moyen, affecté du poids 2.

Algorithme de classification et résultat produit

L'algorithme de classification

La classification proprement dite peut être décrite de la manière suivante : :

Étape 1 : il y a n éléments à classer (qui sont les n individus);

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

Étape 3 : on construit une nouvelle matrice des distances qui résultent de l'agrégation, en calculant les distances entre le nouvel élément et les éléments restants (les autres distances sont inchangées). On se trouve dans les mêmes conditions qu'à l'étape 1, mais avec seulement (n-1) éléments à classer et en ayant choisi un critère d'agrégation. On cherche de nouveau les deux éléments les plus proches, que l'on agrège. On obtient une deuxième partition avec n-2 classes et qui englobe la première;

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

Hiérarchie de classes et partition de l'ensemble des individus

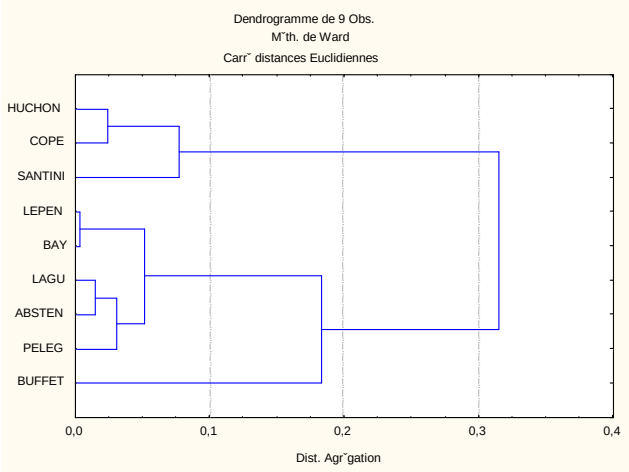
Opérer une classification, c'est définir une partition de l'ensemble des individus, c'est-à-dire, définir un ensemble de parties, ou classes de l'ensemble I des individus telles que :

- toute classe soit non vide
- deux classes distinctes sont disjointes
- tout individu appartient à une classe.

Le résultat d'une CAH n'est pas une partition de l'ensemble des individus. C'est une hiérarchie de classes telles que :

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

Ce résultat est souvent représenté sous forme de dendrogramme :



Sur la figure ci-dessus, l'axe vertical indique les individus statistiques qui ont été rassemblés pour former les classes, tandis que la graduation de l'axe horizontal indique la distance séparant les deux classes qui ont été rassemblées une étape donnée.

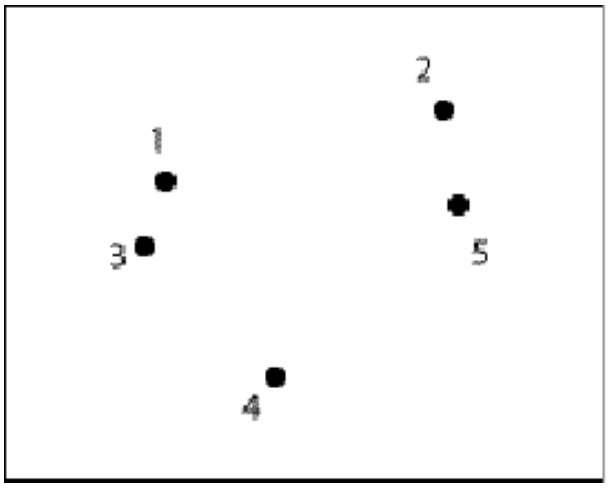
Choix d'une partition à partir de la hiérarchie des classes

Le dendrogramme nous indique l'ordre dans lequel les agrégations successives ont été opérées. Il nous indique également la valeur de l'indice d'agrégation à chaque niveau d'agrégation. Il est généralement pertinent d'effectuer la coupure après les agrégations correspondant à des valeurs peu élevées de l'indice et avant les agrégations correspondant à des valeurs élevées. En coupant l'arbre au niveau d'un saut important de cet indice, on peut espérer obtenir une partition de bonne qualité car les individus regroupés en-dessous de la coupure étaient proches, et ceux regroupés après la coupure sont éloignés.

2.5.3.2 CAH "à la main"

Le dessin suivant représente 5 objets "en vraie grandeur". La distance utilisée entre les objets est la distance euclidienne (mesurée au double-décimètre). L'indice d'agrégation est celui du "saut minimal".

Réalisez une CAH sur ces données :

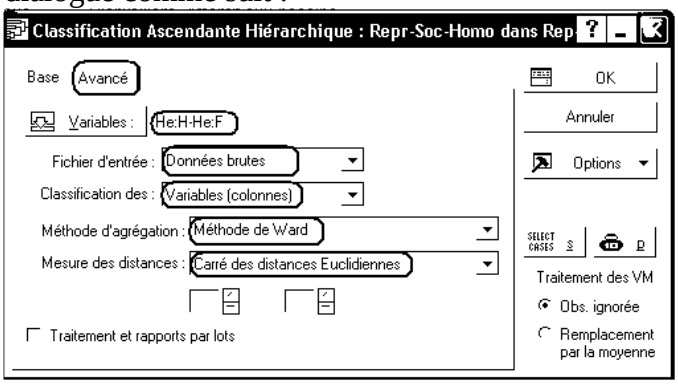


2.5.3.3 La CAH avec Statistica

CAH sur le cas "Représentations sociales de l'homosexualité

On reprend le classeur Rep-Soc-Homo.stw.

Faites une classification ascendante hiérarchique des variables, en utilisant le carré de la distance euclidienne et la méthode de Ward. Pour cela, utilisez le menu Statistiques - Techniques exploratoires multivariées - Classifications. Sélectionnez l'item "Classification hiérarchique" et complétez la fenêtre de dialogue comme suit :

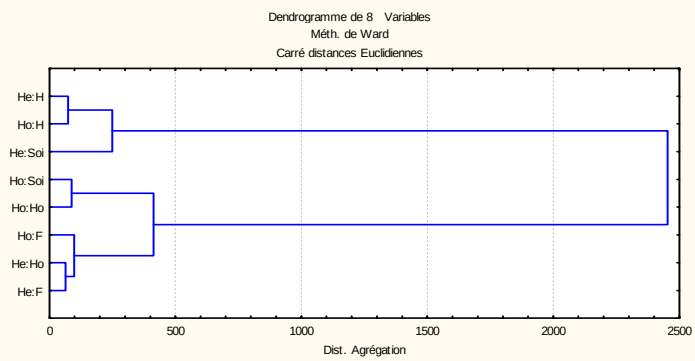


Pour l'essentiel, les résultats de la CAH rejoignent ceux de la classification précédente. Les principaux résultats fournis par Statistica sont les suivants :

La matrice des distances initiales entre les différents objets :

	Carré distances Euclidiennes							
Variable	He:H	Ho:H	He:Soi	Ho:Soi	Ho:Ho	Ho:F	He:Ho	He:F
He:H	0	74	232	788	736	898	1018	1048
Ho:H	74	0	180	694	594	830	920	1028
He:Soi	232	180	0	466	438	634	718	810
Ho:Soi	788	694	466	0	88	246	160	278
Ho:Ho	736	594	438	88	0	182	178	284
Ho:F	898	830	634	246	182	0	72	108
He:Ho	1018	920	718	160	178	72	0	64
He:F	1048	1028	810	278	284	108	64	0

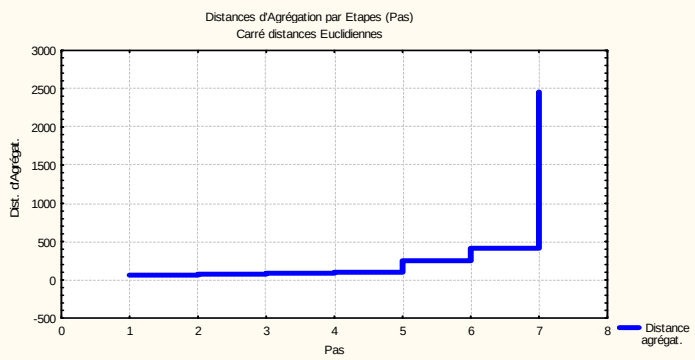
Le dendrogramme de la classification :



Le tableau donnant les différentes étapes de la classification :

distance agrégat.	Agrégation Finale (Repr-Soc-Homo dans Rep-Soc-Homo-correction.stw Méth. de Ward Carré distances Euclidiennes Objet # 1	Objet # 2	Objet # 3	Objet # 4	Objet # 5	Objet # 6	Objet # 7	Objet # 8
64,00	He:Ho	He:F						
74,00	He:H	Ho:H						
88,00	Ho:Soi	Ho:Ho						
98,66	Ho:F	He:Ho	He:F					
250,00	He:H	Ho:H	He:Soi					
413,33	Ho:Soi	Ho:Ho	Ho:F	He:Ho	He:F			
2453,5	He:H	Ho:H	He:Soi	Ho:Soi	Ho:Ho	Ho:F	He:Ho	He:F

Le graphique de l'agrégation finale, donnant à chaque étape l'indice d'agrégation des deux classes que l'on réunit :



Remarques.

1. Contrairement à d'autres logiciels de traitement statistique, Statistica ne propose pas de faire un centrage réduction des variables avant de faire la CAH. Si les variables retenues pour décrire les individus s'expriment avec des unités différentes, ou ont des plages de variation très différentes, une telle transformation des variables est pourtant indispensable.

2. Statistica ne permet pas de choisir a priori un nombre déterminer de classes et, en conséquence ne fournit pas non plus de table d'appartenance du type suivant (produit par Statgraphics) :

```
Table d'appartenance
--------------------
Méthode de classification: Ward
Distance:  Euclidienne au carré

Variable              Classe
```

```
---------------------- -------
He:H                        1
Ho:H                        1
He:Soi                      1
Ho:Soi                      2
Ho:Ho                       2
Ho:F                        3
He:Ho                       3
He:F                        3
---------------------- -------
```

Ces limitations ne sont guère gênantes sur l'exemple traité ici (8 variables et 15 individus) mais le deviennent lorsque le nombre d'objets à classer est important.

Un exemple de CAH effectué à partir d'un tableau de contingence

Source : Lebart L., Morineau A., Piron M. Statistique Exploratoire Multidimensionnelle.

L'exemple concerne l'analyse d'un tableau de contingence qui croise 8 professions et catégories socioprofessionnelles (PCS) et 6 types de médias pour un échantillon de 12 388 "contacts média" relatifs à 4433 personnes interrogées. L'individu statistique sera pour nous le "contact média" et non la personne interrogée dans l'enquête. Les données sont extraites de l'Enquête Budget-temps Multimédia 1991-1992 du CESP.

Afin d'interpréter plus efficacement les représentations obtenues, on projettera en éléments supplémentaires certaines autres caractéristiques de la population enquêtée telles que le sexe, l'âge, le niveau d'instruction.

Tables de contingence croisant les types de contacts-média (colonnes) avec professions, sexe, âge, niveau d'éducation (lignes).

	Radio	Tél.	Quot.N.	Quot R.	P.Mag.	P.TV
Professions						
Agriculteur	96	118	2	71	50	17
Petit patron	122	136	11	76	49	41
Prof. Cad. S.	193	184	74	63	103	79
Prof. interm	360	365	63	145	141	184
Employé	511	593	57	217	172	306
Ouvrier qual	385	457	42	174	104	220
Ouvrier n-q	156	185	8	69	42	85
Inactif	1474	1931	181	852	642	782

Nous disposons des tables de contingence suivantes (cf. tableau) : On trouve, à l'intersection de la ligne i et de la colonne j le nombre kij d'individus appartenant à la catégorie i et ayant eu la veille (un jour de semaine) au moins un contact avec le type de média j. Une personne interrogée pouvant avoir des contacts avec plusieurs médias, les valeurs en ligne représentent des "nombres de contacts".

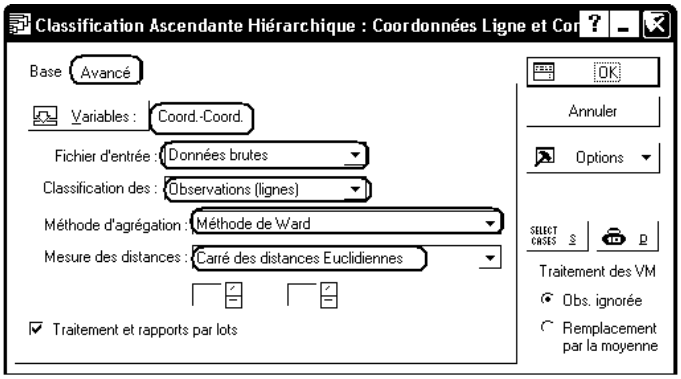
Nous nous proposons de réaliser une CAH sur les professions à partir de ce tableau. Comme nous l'avons vu dans le paragraphe sur l'AFC, la "distance" pertinente entre deux lignes du tableau est la distance du khi-2, ou, ce qui revient au même, le carré de la distance euclidienne entre les images des modalités lignes obtenues par AFC.

Dans un premier temps, ouvrez le classeur Contacts-Medias.stw et réalisez une AFC en calculant les coordonnées lignes et colonnes sur tous les facteurs.

Rendez ensuite active la feuille de données contenant les résultats relatifs aux lignes.

Utilisez ensuite le menu Statistiques - Techniques Exploratoires Multivariées - Classifications .

On choisit ici comme mesure des distances, le carré des distances euclidiennes. Cela revient à mesurer la distance entre deux lignes à l'aide de la distance du khi-2 (propriété de l'AFC). L'indice d'agrégation choisi est celui calculé par la méthode de Ward.



On obtient ainsi les résultats suivants :

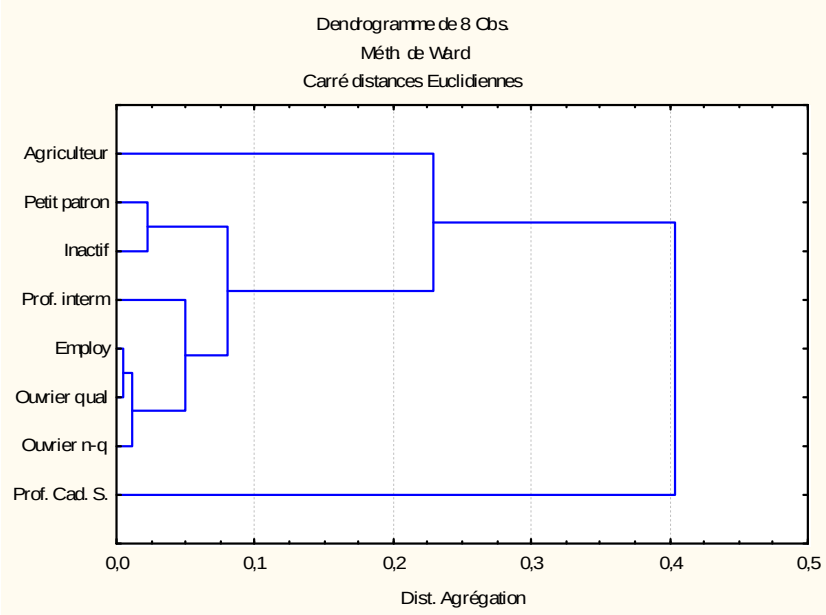
- un tableau donnant les étapes de la classification :

distance agrégat.	Agrégation Finale (Coordonnées Ligne et Contributions à l'Inertie (Contacts-medias.sta dans Classeur4) dans Classeur Méth. de Ward Carré distances Euclidiennes							
	Objet # 1	Objet # 2	Objet # 3	Objet # 4	Objet # 5	Objet # 6	Objet # 7	Objet # 8
,0041508	Employé	Ouvrier qual						
,0120153	Employé	Ouvrier qual	Ouvrier n-q					
,0225031	Petit patron	Inactif						
,0496726	Prof. interm	Employé	Ouvrier qual	Ouvrier n-q				
,0805143	Petit patron	Inactif	Prof. interm	Employé	Ouvrier qual	Ouvrier n-q		
,2296203	Agriculteur	Petit patron	Inactif	Prof. interm	Employé	Ouvrier qual	Ouvrier n-q	
,4046085	Agriculteur	Petit patron	Inactif	Prof. interm	Employé	Ouvrier qual	Ouvrier n-q	Prof. Cad. S.

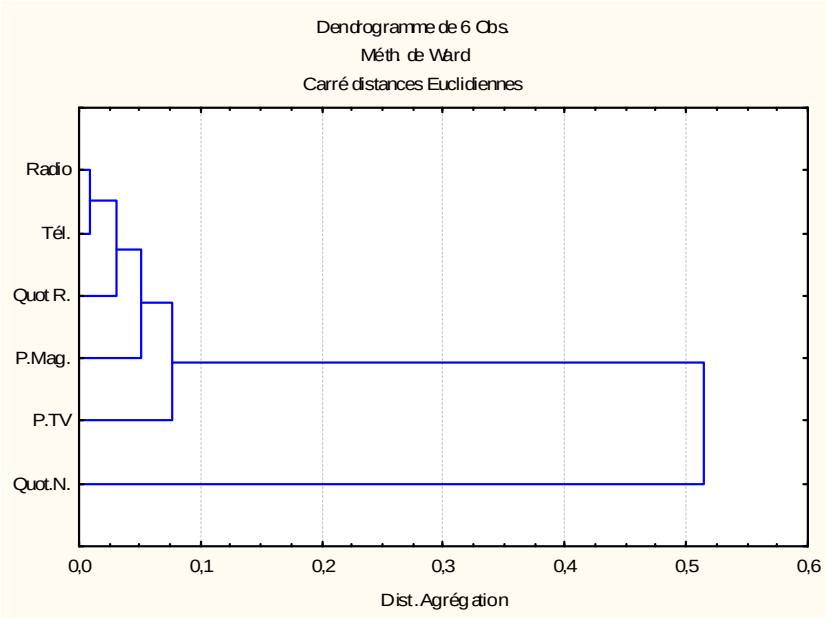
- Le tableau des distances entre individus :

N° Obs.	griculteu	Petit patron	Prof. Cad. S.	Prof. interm	mploye	Ouvrier qual	Ouvrier n-q	Inactif
Agriculteur	0,00	0,04	0,42	0,19	0,19	0,19	0,17	0,10
Petit patron	0,04	0,00	0,26	0,06	0,07	0,06	0,06	0,02
Prof. Cad. S.	0,42	0,26	0,00	0,12	0,22	0,25	0,33	0,22
Prof. interm	0,19	0,06	0,12	0,00	0,02	0,03	0,06	0,03
Employé	0,19	0,07	0,22	0,02	0,00	0,00	0,01	0,02
Ouvrier qual	0,19	0,06	0,25	0,03	0,00	0,00	0,01	0,02
Ouvrier n-q	0,17	0,06	0,33	0,06	0,01	0,01	0,00	0,03
Inactif	0,10	0,02	0,22	0,03	0,02	0,02	0,03	0,00

- Le dendrogramme correspondant à la CAH :



Une CAH analogue, réalisée à partir des individus colonnes, conduit au dendrogramme suivant :

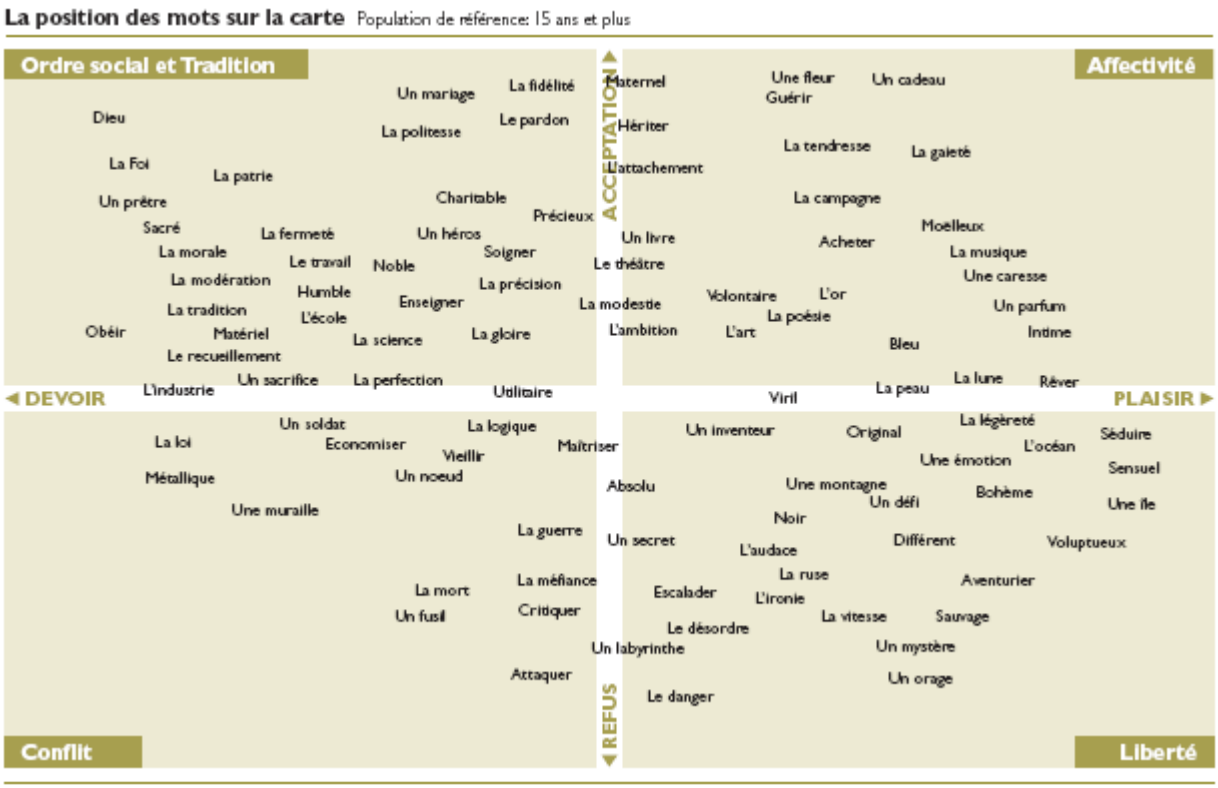


Une telle classification, dans laquelle chaque nouvelle classe est obtenue en agrégeant un unique individu à la classe formée à l'étape précédente revient en fait à définir une relation d'ordre sur les individus, et ne présente qu'un intérêt fort limité.

Classification à partir d'un tableau Individus x Variables Numériques

Réf. Lebart L., Piron M., Steiner J.-F., La Sémiométrie, Dunod, Paris, 2003.

Dans l'ouvrage cité en référence, les auteurs ont fait le choix de 210 mots. Il est ensuite demandé aux personnes interviewées de noter les mots en fonction de la sensation, agréable ou désagréable, que provoque leur lecture. L'échelle de notation comporte 7 modalités variant de -3 à 3. Pour les traitements statistiques ultérieurs, cette échelle est ramenée à une échelle variant de 1 à 7. L'échantillon interrogé entre 1990 et 2002 s'élève à 11055 personnes. Une enquête analogue, menée pour la Belgique, a conduit au résultat suivant (deux premiers axes d'une ACP) :



On mesure la proximité entre deux mots à l'aide du coefficient de corrélation des séries statistiques obtenues pour les deux mots. Plus précisément, le carré de la distance entre deux mots a et b est égal à (1-r(a, b)), où r(a,b) désigne le coefficient de corrélation des deux séries. Pour chaque mot, les autres mots qui lui sont le mieux corrélés constituent son champ sémantique interne. Cependant, un même mot peut être corrélé avec des mots non corrélés entre eux.

Une classification ascendante hiérarchique est effectuée à partir de la distance définie précédemment. Il n'est pas évident a priori que des notes fondées seulement sur l'agrément ou le désagrément engendrent des proximités sémantiques. On constate cependant que les classes obtenues regroupent des mots qui ne sont pas de vrais synonymes (la liste de mots excluait a priori la présence de synonymes) mais appartiennent au même halo sémantique. Dans une partition en 12 classes, par exemple, on trouvera rassemblés des mots ayant trait au concept de "sublimation" tels que :

absolu, immense, infini, admirer, adorer, éternel, précieux, secret, sublime.

Exercice :
A partir de la liste de 7 mots suivants :

efficace, courage, sensuel, montagne, magie, douceur, campagne

imaginez les réponses fournies par dix interviewés et traitez-les à l'aide d'une CAH en utilisant, évidemment , la "distance" 1-r de Pearson.

Représentation des similitudes par l'arbre de longueur minimale

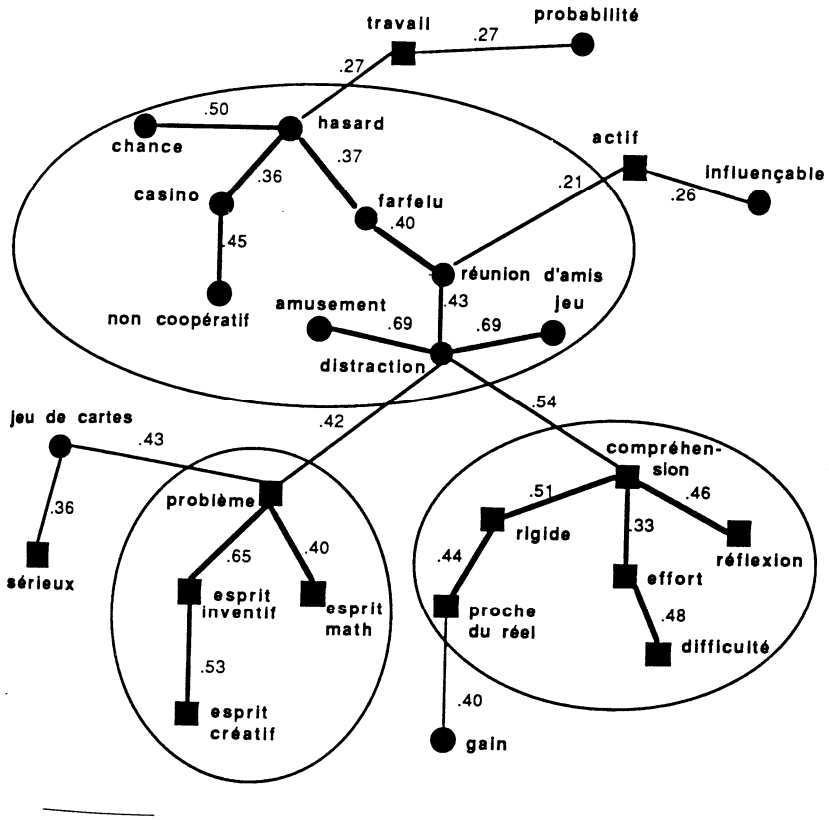
L'ensemble des n objets à classer peut être considéré comme un ensemble de points d'un espace. Si l'on ne dispose que des valeurs d'un indice de dissimilarité , on peut représenter les objets par des points (d'un plan par exemple), chaque couple d'objets étant joint par une ligne continue, à laquelle est attachée la valeur de l'indice de dissimilarité. On représente ainsi l'ensemble des objets et des valeurs de l'indice par un graphe complet valué. On cherchera ensuite à extraire de ce graphe un graphe partiel (ayant les mêmes sommets, mais moins d'arêtes) plus aisé à représenter, et

permettant néanmoins de bien résumer les valeurs de l'indice. Parmi tous les graphes partiels, ceux qui ont une structure d'arbre sont particulièrement intéressants, car ils peuvent faire l'objet d'une représentation plane. La longueur d'un arbre sera la somme des "longueurs" (valeurs de l'indice) de ses arêtes. Parmi tous les graphes partiels qui sont des arbres, l'arbre de longueur minimale a retenu depuis longtemps l'attention des statisticiens en raison de ses bonnes qualités descriptives, qui ne sont pas étrangères à sa parenté avec les classifications hiérarchiques. On peut, par exemple, montrer l'équivalence avec la classification selon le saut minimal.

Dans la procédure de Kruskal, par exemple, on range les n(n - 1)/2 arêtes dans l'ordre des valeurs croissantes de l'indice. On part des deux premières arêtes, puis on sélectionne successivement toutes les arêtes qui ne font pas de cycle avec les arêtes déjà choisies. On interrompt la procédure dès que l'on a n-1 arêtes. De cette façon, on est sûr d'avoir obtenu un arbre (graphe sans cycle ayant n-1 arêtes).

Exemple : Dans l'ouvrage "Représentations sociales et analyse des données", Doise et al. donnent l'exemple suivant :

Donnons un exemple que Flament emprunte à Abric et Vacherot (1976). Il s'agit d'une recherche effectuée sur la représentation d'une tâche de type «dilemme du prisonnier», tâche qui peut être perçue comme une situation de jeu ou une situation de résolution de problèmes. Les auteurs retiennent 26 termes d'une pré-enquête permettant de traduire l'une ou l'autre de ces situations. Ils demandent ensuite à des sujets ayant effectué une tâche de type dilemme du prisonnier de choisir parmi les 26 termes ceux qui évoquent la situation dans laquelle ils se trouvaient. L'arbre maximum du système de similitude (comprenant 325 corrélations) est de la forme suivante :



Dans cette figure, chaque terme représente un sommet. Le long des liaisons entre sommets (ou arêtes) sont indiqués les indices de similitude. Pour construire un tel arbre, la procédure est la suivante. Il s'agit d'abord d'ordonner les arêtes selon la valeur décroissante de l'indice de similitude qui leur est associé. On retient ensuite les deux premières arêtes qui appartiendront forcément à l'arbre maximum du fait qu'elles ne peuvent être les plus petites dans aucun cycle. Enfin, on ajoute à

ces deux premières arêtes, toute arête qui ne forme pas de cycle avec celles déjà retenues. Les arêtes qui sont donc retenues dans l'arbre maximum sont celles qui ne sont minimum dans aucun cycle (voir Degenne et Verges, 1973). Pour illustrer ce propos, prenons l'exemple des éléments Chance, Hasard et Casino qui figurent dans l'arbre maximum ci-dessus. Les arêtes (Chance, Hasard) et (Casino, Hasard) sont inscrites sur le graphe et valent respectivement .50 et .36. On en déduit par conséquent que l'arête (Chance, Casino) est inférieure à .36 ; si tel n'était pas le cas, l'arête (Casino, Hasard) serait supprimée au profit de l'arête (Chance, Casino). En termes de similitude, on peut dire que Chance et Hasard, d'une part, et Hasard et Casino, d'autre part, sont plus proches l'un de l'autre que Chance et Casino.

Sur la base de l'arbre maximum, il est possible de répondre à la question posée par Abric et Vacherot qui est d'identifier les termes associés à jeu ou à résolution de problème comme représentation de la tâche. Flament (1986, 144) en propose la lecture suivante: «Supprimons de l'arbre maximum les arêtes se trouvant entre items de catégories initiales différentes (voir figure) ; les sous-graphes ainsi obtenus sont alors de composition homogène (soit tout jeu, soit tout problème) ; on observe des items isolés (Travail, Probabilité, Actif, etc.), dont la signification initiale est fortement remise en cause (puisque chacun ressemble plus à des items de catégorie opposée qu'aux items de sa propre catégorie). Restent trois sous-graphes importants (indiqués dans la figure) - un pour jeu, deux pour problème -, dont les items voient leur signification initiale confirmée dans la représentation par le voisinage d'items de même catégorie.»

3 Méthodes prédictives

3.1 Régression linéaire

Bibliographie :
Bry, X., Analyses factorielles multiples, Economica, Paris, 1996.

3.1.1 Régression linéaire multiple

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

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

Exemple (source : fichiers d'exemples fournis avec Statistica) :

On dispose, pour 30 comtés américains, des données suivantes :

VARI_POP	Variation de la Population (1960-1970)
N_AGRIC	Nb. de personnes travaillant dans le secteur primaire (agriculture)
TX_IMPOS	Taux d'imposition des propriétés résidentielles et fermes
PT_PHONE	Pourcentage de résidences équipés d'une installation téléphonique
PT_RURAL	Pourcentage de la population vivant en milieu rural
AGE	Age médian
PT_PAUVR	Pourcentage de familles en dessous du seuil de pauvreté

L'objectif est d'identifier les facteurs liés au pourcentage de familles en deçà du seuil de pauvreté dans ces comtés (Pt_Pauvr), et de construire un modèle prédictif pour cette variable. Nous allons donc traiter la variable Pt_Pauvr comme la variable dépendante (réponse), et les 6 autres variables comme des prédicteurs continus.

Les données sont les suivantes :

	VARI_POP	N_AGRIC	PT_PAUVR	TX_IMPOS	PT_PHONE	PT_RURAL	AGE
Benton	13,7	400	19,0	1,09	82	74,8	33,5
Cannon	-0,8	710	26,2	1,01	66	100,0	32,8
Carrol	9,6	1610	18,1	0,40	80	69,7	33,4
Cheatheam	40,0	500	15,4	0,93	74	100,0	27,8
Cumberland	8,4	640	29,0	0,92	65	74,0	27,9
DeKalb	3,5	920	21,6	0,59	64	73,1	33,2
Dyer	3,0	1890	21,9	0,63	82	52,3	30,8
Gibson	7,1	3040	18,9	0,49	85	49,6	32,4
Greene	13,0	2730	21,1	0,71	78	71,2	29,2
Hawkins	10,7	1850	23,8	0,93	74	70,6	28,7
Haywood	-16,2	2920	40,5	0,51	69	64,2	25,1
Henry	6,6	1070	21,6	0,80	85	58,3	35,9
Houston	21,9	160	25,4	0,74	69	100,0	31,4
Humphreys	17,8	380	19,7	0,44	83	72,0	30,1
Jackson	-11,8	1140	38,0	0,81	54	100,0	34,1
Johnson	7,5	690	30,1	1,05	65	100,0	30,5
Lawrence	3,7	1170	24,8	0,73	76	69,5	30,0
McNairy	1,6	1280	30,3	0,65	67	81,0	32,4
Madison	8,4	2270	19,5	0,48	85	39,1	28,7

Marshall	2,7	960	15,6	0,72	84	58,4	33,4
Maury	5,6	1710	17,2	0,62	84	42,4	29,9
Montgomery	12,7	1410	18,4	0,84	86	36,4	23,3
Morgan	-4,8	200	27,3	0,73	66	99,8	27,5
Sevier	16,5	960	19,2	0,45	74	90,6	29,5
Shelby	15,2	11500	16,8	1,00	87	5,9	25,4
Sullivan	11,6	1380	13,2	0,63	85	44,2	28,8
Trousdale	4,9	530	29,7	0,54	70	100,0	33,1
Unicoi	1,1	370	19,8	0,98	75	52,6	30,8
Wayne	3,8	440	27,7	0,46	48	100,0	28,4
Weakley	19,0	1630	20,5	0,68	83	72,1	30,4

3.1.1.1 Formulation explicative : le modèle linéaire

On cherche à exprimer Y sous la forme :

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

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

La solution à ce problème est obtenue en prenant pour b_0 :

$$b_0 = \overline{Y} - b_1 \overline{X_1} - b_2 \overline{X_2} - ... - b_p \overline{X_p}$$

et pour les autres coefficients b_i, les solutions du système d'équations linéaires :

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

Pour les données citées en introduction, on obtient :

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

Interprétation des coefficients b_i : d'une manière générale, chaque coefficient b_i représente la variation relative de la variable Y rapportée à celle de la variable X_i, toutes les autres variables restant constantes.

Problème : les coefficients b_i dépendent des unités choisies pour mesurer les X_i. C'est pourquoi, on donne aussi les coefficients β_i, liés aux b_i par la relation :

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

Dans l'exemple, les coefficients β_i sont donnés par :

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

Même ainsi "normés", les coefficients de la régression restent d'interprétation délicate. En effet, il est impossible de faire varier l'une des X_i en laissant les autres constantes, car ces variables sont elles-mêmes corrélées entre elles.

Sur notre exemple, on constate que le coefficient correspondant à la variable N_AGRIC est positif (PT_PAUVR et N_AGRIC semblent varier dans le même sens), alors que le coefficient de corrélation entre les deux variables PT_PAUVR et N_AGRIC est négatif (r=-0,17), ce qui indiquerait plutôt une variation en sens contraires. Les corrélations entre les variables sont en effet données par :

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

Les coefficients b_i et β_i étant des valeurs "théoriques" estimées à partir des valeurs prises par les variables X_i sur l'échantillon de n individus statistiques, il est possible :
- de donner pour chaque b_i un intervalle de confiance à un degré de confiance donné ;
- de tester si chacun des coefficients est significativement différent de 0.

Dans l'exemple traité, on obtient pour les b_i :

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

N.B. : Tableau obtenu sous Statistica, à l'aide du menu Statistiques - Modèles généraux de régression - Régression Multiple puis l'onglet Synthèse et le bouton Coefficients.

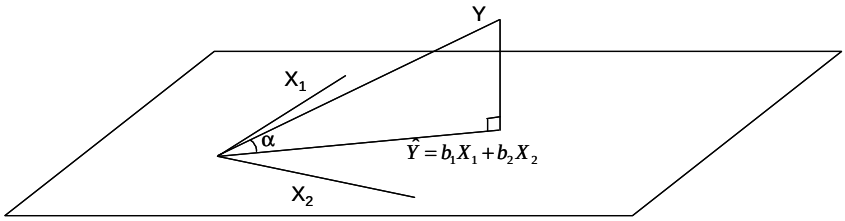
On voit que seuls b_0, b_1 et b_5 sont significativement différents de 0.

En raison des difficultés d'interprétation des résultats d'une régression multiple, différentes alternatives à la régression linéaire "ordinaire" ont été proposées. En particulier, l'interprétation d'une régression est nettement plus simple lorsque les prédicteurs sont non corrélés entre eux. C'est pourquoi il peut être intéressant de réaliser une ACP sur les prédicteurs, puis une régression de Y sur les facteurs de l'ACP. Cette méthode est appelée : *régression sur les composantes principales*.

3.1.1.2 Approche factorielle de la régression

Après centrage des données, le problème de la régression linéaire se ramène au suivant :

On cherche à expliquer la variabilité de Y à partir de celle des X_j : on cherche une combinaison linéaire des X_j qui reproduit "au mieux" la variabilité des individus selon Y. On prend donc la combinaison linéaire la plus corrélée avec Y. La solution est fournie par la combinaison linéaire des X_j qui fait avec Y un angle minimum.



A chaque valeur observée y_i de la variable Y correspond une valeur $\hat{y}_i$ estimée à l'aide de l'équation de régression. La variabilité de Y se décompose comme suit :

Variance de Y = Variance expliquée + Variance résiduelle

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

L'analyse de variance permet de tester globalement si la variable régressée dépend significativement des régresseurs qui ont été considérés :

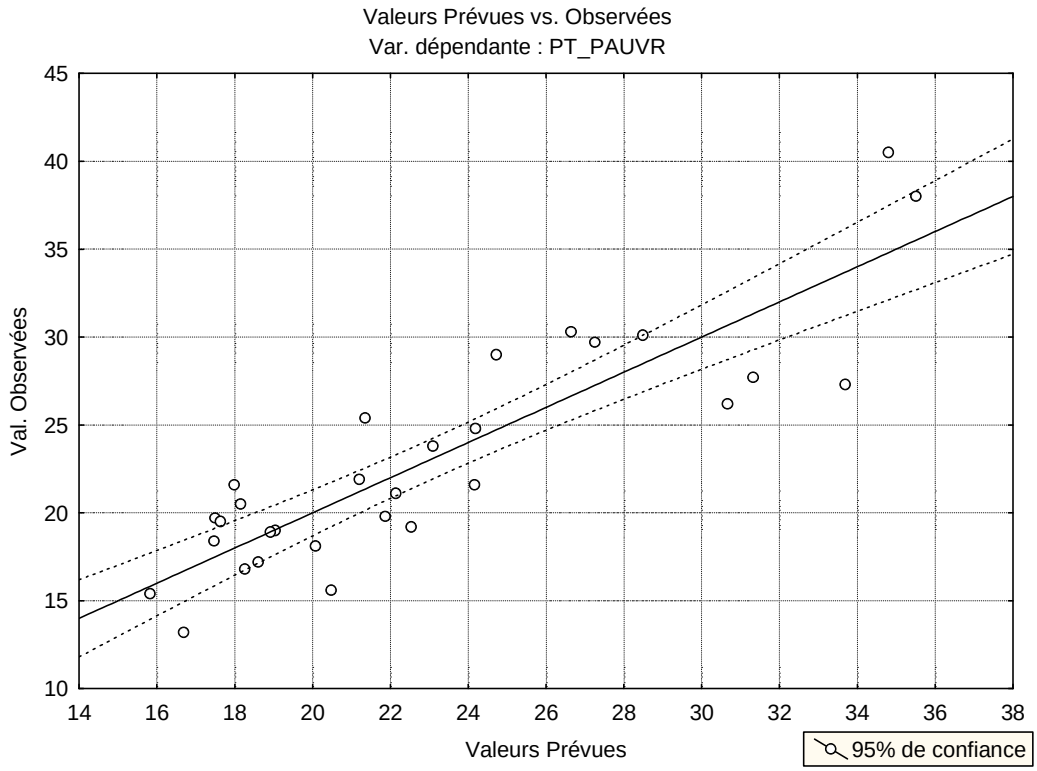
	Sommes Carrés	dl	Moyennes Carrés	F	niveau p
Régress.	932,065	6	155,3441	13,44909	0,000002
Résidus	265,662	23	11,5505		
Total	1197,727				

Le *coefficient de détermination* est le rapport : $R^2 = \dfrac{Var(\hat{Y})}{Var(Y)}$.

Cette valeur est aussi le carré du coefficient de corrélation $r(Y,\hat{Y})$, appelé *coefficient de corrélation multiple*. Sur l'exemple traité, on obtient :

$R = 0{,}8822$; $R^2 = 0{,}7782$

Le graphique suivant compare les valeurs observées de Y (les yi) avec les valeurs estimées par la régression (les y_i) :



3.1.2 Une application de la régression linéaire : analyse de médiation

Une référence fréquemment citée en ce qui concerne l'analyse de médiation et l'analyse de modération :
Baron, R. M., Kenny D.A., The moderator-Mediator Variable Distinction in Social Psychological Research: Conceptual, Strategic, and Statistical Considerations, Journal of Personality and Social Psychology, 1986, V. 51 N° 6, pp. 1173-1182.

Voir aussi : http://www.psychologie-sociale.org/reps2.php?article=7

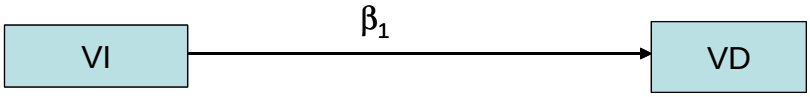
L'analyse de médiation est une technique statistique très utile pour identifier les processus responsables de l'effet d'une variable indépendante sur une variable dépendante. Ainsi, la médiation permet de distinguer, dans l'effet à expliquer, ce qui est directement imputable à la variable indépendante (effet direct de la VI sur la VD) et ce qui relève plutôt de l'intervention d'un facteur intermédiaire (effet indirect de la VI sur la VD via une variable de médiation M).

3.1.2.1 Principe de la méthode :

On effectue la régression linéaire de la VD sur la VI. On obtient l'équation de régression :

$VD = b0 + b1\ VI$

et un coefficient de régression standardisé : β_1 :



On effectue ensuite la régression linéaire de la variable de médiation M sur la VI. On obtient l'équation de régression :

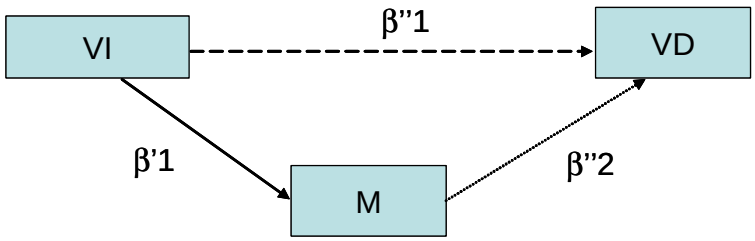
$M = b'0 + b'1\ VI$

et le coefficient de régression standardisé : β'_1

Enfin, on effectue la régression linéaire multiple de la VD sur les deux variables M et VI. On obtient l'équation de régression :

$VD = b''0 + b''1\ VI + b''2\ M$

et les coefficients de régression standardisés β''_1 et β''_2 :

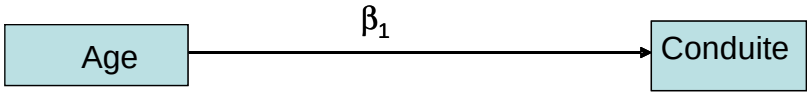


Interprétation :

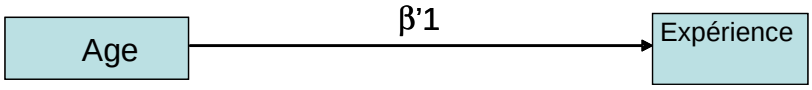
Si β''_2 est significativement différent de 0 et que β''_1 est nettement plus proche de 0 que β_1, en particulier si β''_1 n'est pas significativement différent de 0 alors que β_1 l'était, il y a médiation (partielle ou totale).

3.1.2.2 Exemple "de laboratoire" :

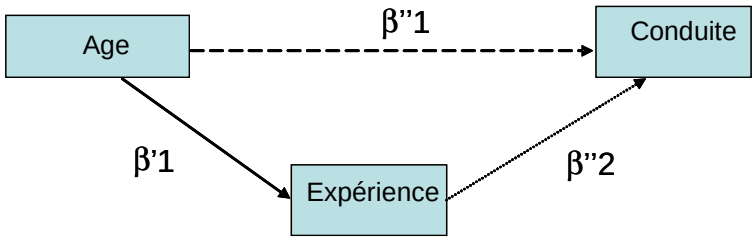
L'âge a un effet significatif sur la performance de conduite d'un véhicule :



Mais l'âge a également un effet sur l'expérience de conduite d'un véhicule :



Lorsque l'âge et l'expérience de conduite sont invoquées toutes deux comme prédicteurs, seule l'expérience a un effet significatif :



3.1.2.3 Exemple 2

Ref. Costarelli, S., Callà, R.-M.. Self-directed negative affect: The distinct roles of ingroup identification and outgroup derogation, Current research in Social Psychology, Volume 10 No 2, 2004.

Le Sud-Tyrol est une région de l'Italie du Nord dans laquelle coexistent une population de langue italienne et une population de langue allemande. La population de langue allemande a fait l'objet d'une discrimination négative durant le régime fasciste, puis a bénéficié de dispositions favorables ensuite. De ces événements résulte un fort sentiment d'appartenance à un groupe pour les membres de chacun de ces deux groupes ethniques.

Une enquête par questionnaire a été menée en 2002 auprès d'un échantillon de 71 lycéens italophones. En particulier, les sujets devaient se positionner sur des échelles unipolaires à 6 points (cotées de 0 à 5, 0=pas du tout, 5=extrêmement), selon leur opinion relativement aux deux communautés. Pour moitié, les adjectifs utilisés étaient à connotation positive (par exemple : les germanophones : ne sont pas du tout/sont extrêmement sympathiques), et pour moitié, les adjectifs utilisés étaient à connotation négative (par exemple : antipathique, repoussant, méprisable). En calculant un score moyen par sujet pour les échelles de même connotation, appliquées à la même cible ethnique, on obtient ainsi pour chaque sujet quatre mesures comprises dans l'intervalle de 0 à 5:

- l'évaluation positive de l'endogroupe, notée ici ENDOP
- l'évaluation positive de l'exogroupe, notée ici EXOP
- l'évaluation négative de l'endogroupe, notée ici ENDON
- l'évaluation négative de l'exogroupe, notée ici EXON.

Par ailleurs, le questionnaire comportait également des questions permettant d'évaluer deux autres variables, également dans l'intervalle de mesure de 0 à 5 :

- l'intensité de l'identification à l'endogroupe, notée ici IDENT;
- l'estime négative de soi (self-directed negative affect), notée ici SDNA.

Les paramètres descriptifs des variables observées sont donnés par :

Description	Notation	Moyenne	Ecart type
Identification à l'endogroupe	IDENT	3.61	0.57
Estime négative de soi	SDNA	3.07	0.63
Evaluation positive de l'endogroupe	ENDOP	4.39	0.71
Evaluation positive de l'exogroupe	EXOP	3.66	0.66
Evaluation négative de l'endogroupe	ENDON	0.56	0.50
Evaluation négative de l'exogroupe	EXON	1.58	0.59

On cherche à expliquer les variations de la variable "estime négative de soi" (SDNA) par celles des autres variables.

On définit une variable notée DEROG (outgroup derogation, ou partialité envers l'exogroupe) en formant la différence EXON - ENDON.

a) La régression linéaire de la variable SDNA sur la variable IDENT fournit les résultats suivants :

	Synthèse de la Régression; Variable Dép. : SDNA F(1.69)=4.0587 p<.04784 Err-Type de l'Estim.: .62106					
	Béta	Err-Type de Béta	B	Err-Type de B	t(69)	niveau p
OrdOrig.			2.129557	0.472590	4.506145	0.000026
IDENT	0.235700	0.116994	0.260511	0.129309	2.014632	0.047842

L'effet de IDENT sur SDNA est donc significatif au seuil de 5%.

La régression linéaire de IDENT sur DEROG fournit les résultats suivants :

	Synthèse de la Régression; Variable Dép. : DEROG F(1.69)=8.4320 p<.00495 Err-Type de l'Estim.: .52667					
	Béta	Err-Type de Béta	B	Err-Type de B	t(69)	niveau p
OrdOrig.			-0.129500	0.400765	-0.323132	0.747572
IDENT	0.329994	0.113642	0.318421	0.109657	2.903799	0.004948

L'effet de IDENT sur DEROG est donc significatif au seuil de 5% ?

Enfin, on réalise une régression linéaire multiple de SDNA sur les variables DEROG et IDENT. Les résultats sont alors les suivants :

	Synthèse de la Régression; Variable Dép. : SDNA F(2.68)=5.1029 p<.00861 Err-Type de l'Estim.: .60028					
	Béta	Err-Type de Béta	B	Err-Type de B	t(68)	niveau p
OrdOrig.			2.172575	0.457119	4.752756	0.000011
IDENT	0.140000	0.119789	0.154737	0.132398	1.168723	0.246596
DEROG	0.290005	0.119789	0.332182	0.137210	2.420970	0.018154

Dans cette dernière régression, l'effet de IDENT sur SDNA n'est plus significatif. L'effet constaté dans la première régression s'explique donc par un effet de médiation joué par la variable DEROG.

Remarque 1. Dans l'article cité supra, les auteurs définissaient également la variable FAVO (favoritisme pour l'endogroupe) comme la différence ENDOP-ENDON et réalisaient une analyse de médiation analogue. Mais, au contraire de la variable DEROG, la variable FAVO ne joue pas de rôle de médiation significatif.

Remarque 2. Ces résultats, obtenus sur des données analogues à celles utilisées par les auteurs peuvent être retrouvés dans le classeur Statistica Analyse-mediation1.stw.

3.1.2.4 Test de Sobel

Plusieurs tests statistiques ont été proposés pour évaluer la significativité des résultats d'une analyse de médiation. Ces tests ne sont pas disponibles dans le logiciel Statistica. En revanche, trois de ces tests pourront être réalisés à l'aide de la commande mediation.test() du package bstats de R. Sur l'exemple précédent, on obtient :

```
> with(mediat,mediation.test(DEROG,IDENT,SDNA))
             Sobel     Aroian    Goodman
z.value 1.86070289 1.79895801 1.92927563
p.value 0.06278615 0.07202532 0.05369665
```

Autrement dit, sur l'exemple précédent, l'effet médiateur tend à être significatif, mais n'atteint pas la p-value de 5% (en revanche, le test unilatéral est significatif).

3.1.3 Modèles de régression plus généraux : aperçu sur l'analyse de modération

Moins utilisée que l'analyse de médiation dans les publications de Psychologie, l'analyse de modération est également plus délicate à mettre en oeuvre.

3.1.3.1 Principe de la méthode

On souhaite étudier si l'effet du prédicteur (VI) sur la variable à prédire (VD) est influencé par les valeurs prises par une troisième variable M (variable modératrice, pas nécessairement corrélée à la VI). Autrement dit, l'effet de la VI sur la VD ne serait pas le même selon que le modérateur prend des valeurs faibles ou élevées : il existerait donc un effet d'interaction entre la VI et la variable M. Une telle situation est ainsi assez analogue à certaines de celles qui sont étudiées à l'aide d'une ANOVA factorielle.

3.1.3.2 Exemple

Dans le package QuantPsyc du logiciel R, les auteurs donnent l'exemple fictif suivant : on a simulé pour 1000 observations les valeurs de 4 variables : beliefs, values, attitudes, intentions, cohérence avec la Théorie de l'Action Raisonnée. La matrice des corrélations est la suivante :

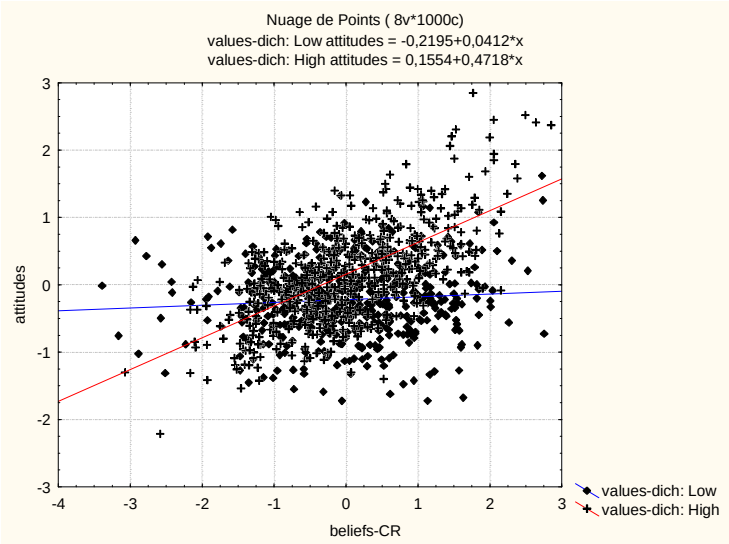
	Corrélations (tra.sta dans tra.stw) Corrélations significatives marquées à p < ,0500 N=1000 (Observations à VM ignorées)			
Variable	beliefs	values	attitudes	intentions
beliefs	1,00	0,03	0,38	0,07
values	0,03	1,00	0,36	0,09
attitudes	0,38	0,36	1,00	0,20
intentions	0,07	0,09	0,20	1,00

On souhaite étudier l'effet modérateur de values lorsque la VI est beliefs et la VD attitudes.
Pour cela, on effectue une régression linéaire multiple en utilisant comme prédicteurs les variables centrées (en fait centrées et réduites) beliefs et values ainsi que le produit beliefs-CR X values-CR de ces deux variables centrées réduites. On obtient comme résultats :

Effet	Paramètres Estimés (tra.sta dans tra.stw) Paramétrisation sigma-restreint			
	attitudes Param.	attitudes Err-Type	attitudes t	attitudes p
Ord.Orig.	-0,0383	0,0164	-2,3378	0,0196
beliefs-CR	0,2428	0,0164	14,8032	0,0000
values-CR	0,2136	0,0165	12,9837	0,0000
beliefs-CR*values-CR	0,2434	0,0161	15,0996	0,0000

Effet	Tests Univariés de Significativité de attitudes (tra.sta dans tra.stw Paramétrisation sigma-restreint Décomposition efficace de l'hypothèse				
	SC	Degré de Liberté	MC	F	p
Ord.Orig.	1,4676	1	1,46759	5,4652	0,019596
beliefs-CR	58,8452	1	58,84516	219,1343	0,000000
values-CR	45,2687	1	45,26872	168,5768	0,000000
beliefs-CR*values-CR	61,2251	1	61,22513	227,9971	0,000000
Erreur	267,4605	996	0,26853		

On observe que les 3 prédicteurs ont un effet sur la VD, ce qui démontre l'effet de modération. On peut illustrer l'effet d'interaction en définissant deux groupes selon les valeurs de values : groupe "Low" pour values-CR < 0 et groupe "High" pour values-CR > 0. Les nuages de points représentant les observations des deux groupes selon les valeurs de beliefs-CR et attitudes sont alors les suivants:



On voit sur le graphique que l'effet de beliefs-CR est nettement plus important lorsque attitudes a une valeur élevée que lorsque attitudes a une valeur faible.

3.1.3.3 Remarque

L'utilisation de cette méthode est délicate, notamment parce que les valeurs trouvées (importance relative des prédicteurs, R2, etc) dépendent des moyennes des prédicteurs. Pour que le terme "produit des prédicteurs" puisse être interprété comme une interaction, il est pratiquement indispensable de travailler avec des prédicteurs centrés (en revanche, les résultats ne sont pas changés de façon substantielle lorsqu' on effectue une réduction des variables, ce que nous avons fait ici).

3.1.4 Régression linéaire avec Statistica

Exemple

Source : A study on significant sources of the burnout syndrome in workers at occupational centres for mentally disabled, Pedro R. Gil-Monte and José Ma Peiró, Psychology in Spain, 1997, Vol. 2. No 1, 116-123.
Page Web : http://www.psychologyinspain.com/content/full/1997/6bis.htm

Subjects
Subjects were 95 employees in occupational institutions for mentally retarded people in the Valencia Autonomous Community (...).

Description des variables.

Self-confidence levels were measured by using five items of an adaptation of the Trait Sport-Confidence Inventory" (TSCI) (Vealey, 1986), in which the word "athlete" was replaced by "workmate". Cronbach's alpha coefficient for the present study was .84.
Social support at work was estimated using 6 items of the "Organisational Stress Questionnaire" (OSQ) (Caplan, Cobb, French, Van Harrison and Pinneau, 1975). These items reflect some aspects of social support coming from *workmates* (3 items) and *supervisors* (3 items). Reliability coefficient in this study was a=.86 for the supervisors' social support scale, and a=.76 for the workmates' social support scale.
Perceived *role conflict* and *role ambiguity* levels were measured by 3 items, for each of the variables, taken from their respective OSQ scales. Reliability values were a=.69 for role ambiguity and .68 for the role conflict scale.
The burnout syndrome was estimated by MBI (Maslach and Jackson, 1986). This instrument is comprised of 22 items measuring the three dimensions in the syndrome: *personal accomplishment* (8 items), *emotional exhaustion* (9 items), and *depersonalisation* (5 items). Reliability coefficients obtained in the study were: a=.76 for the personal accomplishment subscale, a=.87 for emotional exhaustion, and a=.52 for depersonalisation.

Ouvrez le classeur Valencia-Burnout.stw.

N.B. Les données figurant dans ce classeur ont été générées à partir des indications (moyennes, écarts-types, coefficients de corrélation) figurant dans l'article. Cela explique qu'il ne s'agisse pas de valeurs entières, comme on aurait pu le penser à la lecture de la description des variables.

Affichez les statistiques descriptives concernant ces variables. Vous devriez obtenir :

	Statistiques Descriptives (Valencia-Burnout dans Valencia-Burnout.stw				
Variable	N Actifs	Moyenne	Minimum	Maximum	Ecart-type
Self-Confidence	95	6,4800	4,1632	9,4430	1,0656
Workmates Social Support	95	3,2500	1,5810	5,0078	0,6635
Supervisor Social Support	95	2,9000	0,5435	5,2818	0,8545
Role Conflict	95	2,7300	0,9488	4,7904	0,7942
Role Ambiguity	95	2,1100	0,1915	4,3977	0,7640
Personal Accomplishment	95	36,4300	23,2596	57,2632	6,9266
Emotional Exhaustion	95	17,5600	-5,4779	42,1888	10,1737
Depersonalisation	95	4,6900	-4,6758	15,3578	4,4636

Affichez de même la matrice des corrélations :

Variable	Corrélations (Valencia-Burnout dans Valencia-Burnout.stw) Corrélations significatives marquées à p < ,05000 N=95 (Observations à VM ignorées)							
	f-Confiden	Workmates Social Support	Supervisor Social Support	Role Conflict	Role Ambiguity	Personal omplishme	Emotional Exhaustion	ersonalisat
Self-Confidence	1,00	0,08	0,16	-0,11	-0,33	0,35	-0,14	0,00
Workmates Soci	0,08	1,00	0,50	-0,41	-0,40	0,33	-0,45	-0,22
Supervisor Socia	0,16	0,50	1,00	-0,37	-0,43	0,22	-0,40	-0,12
Role Conflict	-0,11	-0,41	-0,37	1,00	0,38	-0,32	0,69	0,32
Role Ambiguity	-0,33	-0,40	-0,43	0,38	1,00	-0,48	0,40	0,28
Personal Accom	0,35	0,33	0,22	-0,32	-0,48	1,00	-0,40	-0,28
Emotional Exhau	-0,14	-0,45	-0,40	0,69	0,40	-0,40	1,00	0,40
Depersonalisatic	0,00	-0,22	-0,12	0,32	0,28	-0,28	0,40	1,00

Comparez avec les valeurs indiquées dans l'article :

Table 1
Descriptive Statistics, reliability values and variables correlations[a]

	M	SD	Range	1	2	3	4	5	6	7	8
1. Self-confidence	6.48	1.06	1-9	(.84)							
2. Workmates Social Support	3.25	.66	1-4	.08	(.76)						
3. Supervisor Social Support	2.90	.85	1-4	.16	.50	(.86)					
4. Role Conflict	2.73	.79	1-5	-.11	-.41	-.37	(.68)				
5. Role Ambiguity	2.11	.76	1-5	-.33	-.40	-.43	.38	(.69)			
6. Personal Accomplishment	36.43	6.89	0-48	.35	.33	.22	-.32	-.48	(.76)		
7. Emotional Exhaustion	17.56	10.12	0-54	-.14	-.45	-.40	.69	.40	-.40	(.87)	
8. Depersonalisation	4.69	4.44	0-30	-.00	-.22	-.12	.32	.28	-.28	.40	(.52)

3.1.4.1 La régression linéaire ordinaire

Effectuez ensuite une régression multiple ordinaire des 3 dernières variables sur les 5 premières :

Pour la variable Personal Accomplishment :

Le bouton "Synthèse de la régression" (onglet "Avancé") affiche les résultats suivants :

N=95	Synthèse de la Régression; Variable Dép. : Personal Accomplishment R= ,56173332 R²= ,31554432 R² Ajusté = ,27709175 F(5,89)=8,2061 p<,00000 Err-Type de l'Estim.: 5,8892					
	Bêta	Err-Type de Bêta	B	Err-Type de B	t(89)	niveau p
OrdOrig.			32,4726	7,3143	4,4396	0,0000
Self-Confidence	0,2293	0,0932	1,4906	0,6057	2,4610	0,0158
Workmates Social Support	0,1730	0,1074	1,8063	1,1213	1,6109	0,1108
Supervisor Social Support	-0,0923	0,1071	-0,7479	0,8678	-0,8618	0,3911
Role Conflict	-0,1351	0,1006	-1,1779	0,8773	-1,3426	0,1828
Role Ambiguity	-0,3235	0,1071	-2,9325	0,9710	-3,0201	0,0033

La colonne "B" donne les coefficients de l'équation de régression linéaire. Le modèle fourni par la régression linéaire est le suivant :
Personal Accomplishment = 32,47 + 1,49 *Self-Confidence +1,81 * Workmates Social Support - 0,75 * Supervisor Social Support - 1,18 * Role Conflict - 2,93 * Role Ambiguity

La valeur de R^2 est de 0,315 : 31,5% de la variance de la variable Personal Accomplishment est expliquée par le modèle.

Les coefficients de la colonne "Bêta" sont les coefficients standardisés, c'est-à-dire les coefficients que l'on observerait si on utilisait des variables centrées réduites au lieu des variables observées. On peut également les interpréter comme suit : lorsque "Self-Confidence" augmente d'un écart type, la variable "Personal Accomplishment" estimée augmente de 0,23 écart type, lorsque la variable "Role Conflict" augmente d'un écart type, "Personal Accomplishment" diminue de 0,135 écart type.
Par exemple, on pourra vérifier que

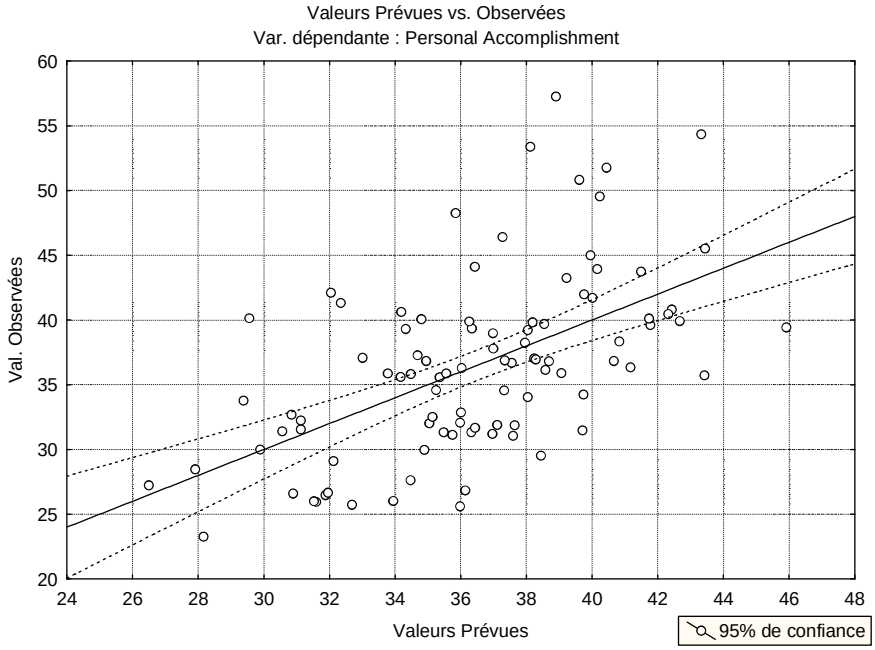
$$Beta(Self-Confidence) = \frac{Ecart\ type(Self-Confidence)}{Ecart\ type(Personal\ Accomplishment)} \times B(Self-Confidence) = \frac{1,0656}{6,9266} \times 1,4906 = 0,2293$$

Les valeurs de t sont obtenues en divisant la valeur correspondante de B par son erreur type. Autrement dit, on teste si le coefficient B est significativement différent de 0.

On peut afficher les résultats de l'ANOVA (bouton ANOVA) montrant qu'ici, le coefficient de régression multiple est significativement différent de 0, ou encore qu'il existe un lien linéaire significatif entre la variable dépendante et les autres variables :

	Analyse de Variance (Valencia-Burnout dans Valencia-Burnout.stw				
Effet	Sommes Carrés	dl	Moyennes Carrés	F	niveau p
Régress.	1423,058	5	284,6115	8,2061	0,0000
1Résidus	3086,793	89	34,6831		
Total	4509,850				

Sous l'onglet "Nuage", on pourra obtenir différentes représentations graphiques dont, par exemple, le graphique illustrant l'adéquation entre les valeurs observées et les valeurs théoriques :



3.1.4.2 La régression linéaire pas à pas

Dans l'article, les auteurs indiquent qu'ils ont fait une régression linéaire pas à pas des dimensions du MBI sur les 5 premières variables.

Principe de la méthode

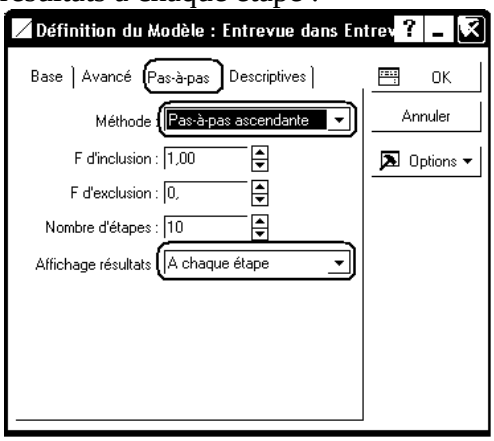
Les données sont formées par une VD Y et plusieurs variables explicatives X1, X2, ..., Xp.
On choisit, parmi les variables explicatives, celle qui est le mieux corrélée à Y. Pour simplifier les notations, nous supposerons qu'il s'agit de la variable X1.
On calcule l'équation de régression linéaire de Y sur X1 : Y = b1 X1 + b0.

On calcule alors les résidus : R1 = Y - b1 X1.- b0
On choisit, parmi les variables explicatives restantes, celle qui est le mieux corrélée à R1. Nous supposons ici qu'il s'agit de la variable X2.
On calcule l'équation de régression linéaire de Y sur X1 et X2 : Y = b'1 X1 + b2 X2 + b'0.
On calcule les nouveaux résidus : R2 = Y - (b'1 X1 + b2 X2 + b'0) et on poursuit la méthode jusqu'à ce que les variables explicatives restantes ne soient plus significativement corrélées aux résidus.

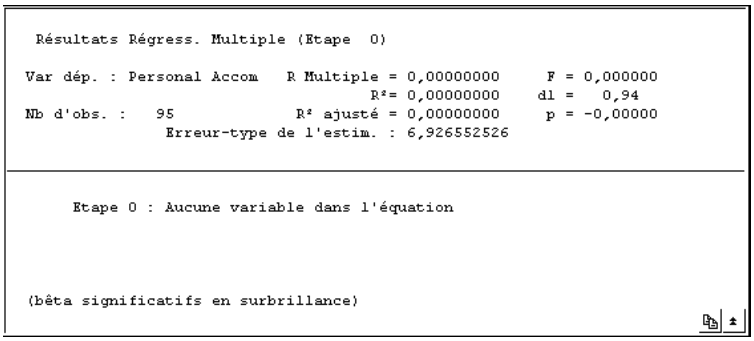
La régression linéaire pas à pas pour la variable Personal Accomplisment

Utilisez de nouveau le menu Statistiques - Régression Multiple
Sous l'onglet "Avancé", spécifiez "Personal Accomplishment" comme variable dépendante, les 5 premières variables comme variables indépendantes. Cochez l'option "régression ridge ou pas-à-pas".

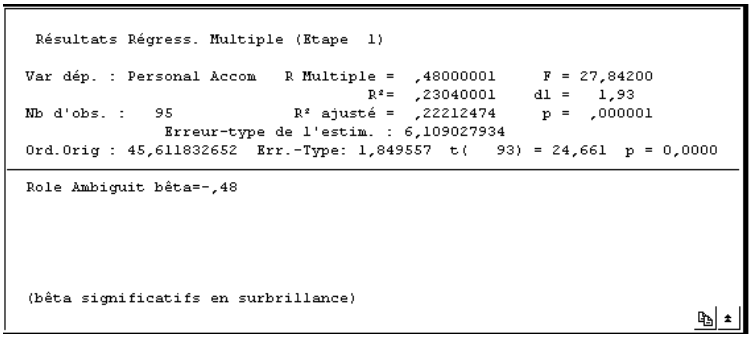
Dans le dialogue suivant, activez l'onglet "pas-à-pas" et sélectionnez la méthode "pas à pas ascendante", et l'affichage des résultats à chaque étape :



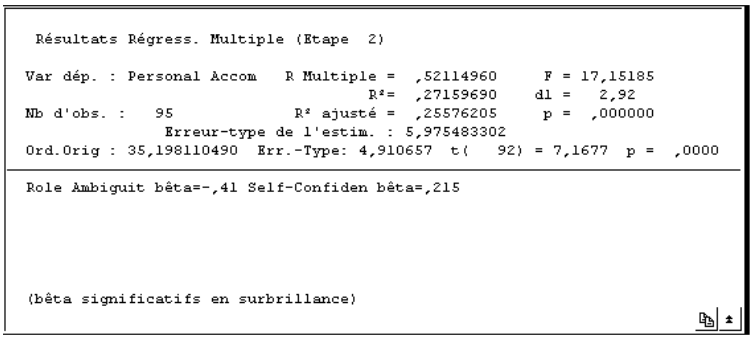
A la première étape, Statistica affiche les résultats suivants :



Cliquez sur "suivant". On obtient :



Puis :



Statistica accepte encore de faire rentrer deux autres variables dans la régression. Cependant, en affichant les résultats disponibles sous le bouton "Synthèse de la régression", on se rend compte que seules ces deux premières variables sont significativement corrélées aux résidus :

N=95	Synthèse de la Régression; Variable Dép. : Personal Accomplishmen R= ,55662608 R²= ,30983259 R² Ajusté = ,27915848 F(4,90)=10,101 p<,00000 Err-Type de l'Estim.: 5,8808					
	Bêta	Err-Type de Bêta	B	Err-Type de B	t(90)	niveau p
OrdOrig.			30,8772	7,0660	4,3698	0,0000
Role Ambiguity	-0,3029	0,1043	-2,7456	0,9451	-2,9050	0,0046
Self-Confidence	0,2253	0,0929	1,4647	0,6041	2,4247	0,0173
Workmates Social Support	0,1406	0,1005	1,4680	1,0489	1,3996	0,1651
Role Conflict	-0,1225	0,0994	-1,0682	0,8668	-1,2323	0,2210

On peut alors reprendre la méthode en ne spécifiant que deux étapes et retrouver les résultats indiqués par les auteurs :

N=95	Synthèse de la Régression; Variable Dép. : Personal Accomplishment R= ,52114960 R²= ,27159690 R² Ajusté = ,25576205 F(2,92)=17,152 p<,00000 Err-Type de l'Estim.: 5,9755					
	Bêta	Err-Type de Bêta	B	Err-Type de B	t(92)	niveau p
OrdOrig.			35,1981	4,9107	7,1677	0,0000
Role Ambiguity	-0,4090	0,0943	-3,7083	0,8545	-4,3395	0,0000
Self-Confidence	0,2150	0,0943	1,3976	0,6127	2,2811	0,0249

Résultats indiqués dans l'article :

Table 2
Stepwise regression analysis for MBI dimensions

Variable Step	R2 increase	Beta	F for equation
	personal Accomplishment		
1 Role ambiguity	.23	-.41	
2 Self-confidence	.04	.22	17.27***
	Emotional Exhaustion		
1 Role conflict	.47	.60	
2 Workmates' social support	.03	-.20	47.27***
	Depersonalisation		
1 Role conflict	.10	.32	10.45***

*** p < 001

3.1.5 Un exemple d'analyse de médiation avec Statistica

Source : Congruence de valeurs et engagement envers l'organisation et le groupe de travail. Stinglhamber, F., Bentein, K., Vandenberghe, C., Psychologie du Travail et des Organisations, Vol. 10, pp. 165-187, 2004.

L'étude citée supra avait pour objectif d'examiner le rôle des valeurs individuelles, organisationnelles et groupales ainsi que celui de la congruence de valeurs dans la prédiction de l'engagement des salariés envers l'organisation et le groupe de travail.

On lit notamment dans la partie consacrée aux objectifs de l'étude :

"La plupart des études ayant examiné les effets des valeurs culturelles sur les attitudes et les comportements des employés se sont concentrés sur le niveau organisationnel. Pourtant, il est de plus en plus reconnu dans la littérature que l'organisation est composée d'entités multiples. Celles-ci peuvent produire leurs propres valeurs culturelles (...)

Nous faisons l'hypothèse que le niveau d'ancrage des valeurs culturelles aura une importance décisive dans la prédiction des attitudes du personnel. Plus spécifiquement, les valeurs émanant de l'organisation en tant que telle devraient jouer un rôle primordial dans le développement d'attitudes dirigées vers l'organisation, alors que les valeurs caractérisant le groupe de travail devraient influencer en priorité les attitudes envers le groupe (Hypothèse 1). (...)

Par ailleurs, il est vraisemblable que cette influence des valeurs véhiculées au sein d'une entité particulière (organisation ou groupe de travail) sur les attitudes du personnel envers cette même entité se fasse par l'intermédiaire d'un ou de plusieurs des mécanismes évoqués précédemment, à savoir des effets directs de ces valeurs, un effet de la congruence (objective ou subjective) de valeurs, ou des effets interactifs. En outre, les résultats des travaux de Judge et Cable (1997 ; Cable et Judge, 1996) laissent à penser que la congruence subjective pourrait être un médiateur de la

relation postulée entre valeurs ou congruence objective de valeurs liée à une entité particulière et les attitudes dirigées vers cette entité (Hypothèse 2). (...)"

Les auteurs définissent 18 variables mesurées à partir d'items évalués sur des échelles de Likert ou de facteurs issus d'une analyse factorielle exploratoire sur un ensemble d'items. Le classeur Statistica Stinglhamber.stw contient un fichier de données créées artificiellement à partir du tableau des moyennes, variances et corrélations publié dans l'article cité et respectant le nombre d'observations faites (200 questionnaires).

N.B. Les auteurs ont utilisé une méthode de régression polynomiale différente de l'analyse de régression utilisée ici. C'est pourquoi les résultats publiés, tout en étant proches des résultats obtenus sous Statistica, ne sont pas identiques à ces derniers.

On s'intéresse ici aux 5 variables prédictives :
- valeurs orientées vers le support interpersonnel au niveau individuel (Support P)
- valeurs orientées vers le support interpersonnel au niveau organisationnel (Support O)
- valeurs orientées vers le support interpersonnel au niveau du groupe de travail (Support G)
- la congruence entre les valeurs personnelles et organisationnelles (par exemple "mes valeurs de travail correspondent à celles qui sont en vigueur dans mon organisation") (Congruence subj P-O)
- la congruence entre les valeurs personnelles et groupales (mes valeurs sont en accord avec celles des autres membres de mon groupe de travail) (Congruence subj P-G).

Les variables dépendantes étudiées sont :
- l'engagement affectif envers l'organisation (EA-Organisation)
- l'engagement normatif envers l'organisation (EN-Organisation)
- l'engagement affectif envers le groupe de travail (EA-Groupe)
- l'engagement normatif envers le groupe de travail (EN-Groupe).

Effectuez une régression linéaire multiple de EA-Organisation sur les trois variables support. Vous devriez obtenir :

N=200	Synthèse de la Régression; Variable Dép. : EA-Organisation R= ,35553188 R²= ,12640292 R² Ajusté = ,11303153 F(3,196)=9,4532 p<,00001 Err-Type de l'Estim.: ,84974					
	b*	Err-Type de b*	b	Err-Type de b	t(196)	valeur p
OrdOrig.			2,03	0,52	3,88	0,00
Support P	-0,03	0,08	-0,05	0,12	-0,44	0,66
Support O	0,26	0,07	0,21	0,06	3,46	0,00
Support G	0,18	0,08	0,16	0,07	2,28	0,02

Effectuez également une régression linéaire multiple de EA-Organisation sur les 5 variables prédictives. Vous devriez obtenir :

N=200	Synthèse de la Régression; Variable Dép. : EA-Organisation R= ,54964813 R²= ,30211306 R² Ajusté = ,28412629 F(5,194)=16,796 p<,00000 Err-Type de l'Estim.: ,76339					
	b*	Err-Type de b*	b	Err-Type de b	t(194)	valeur p
OrdOrig.			1,42	0,48	2,95	0,00
Support P	-0,04	0,07	-0,06	0,10	-0,55	0,58
Support O	0,10	0,07	0,08	0,06	1,40	0,16
Support G	0,09	0,08	0,08	0,08	1,11	0,27
Congruence subj P-O	0,46	0,07	0,45	0,07	6,49	0,00
Congruence subj P-G	0,03	0,08	0,02	0,08	0,32	0,75

Quels sont les éléments de la conclusion suivante qui peuvent être énoncés à partir des tableaux ci-dessus ?
"Les valeurs de support de l'organisation sont associées positivement à l'engagement affectif envers l'organisation (...) Par ailleurs, les valeurs de support du groupe de travail sont positivement liées à l'engagement tant affectif que normatif envers le groupe. En outre, la congruence de valeurs subjective P–O a un impact significatif sur l'engagement affectif envers l'organisation et la congruence subjective P–G a un effet significatif sur l'engagement affectif envers le groupe."

Réalisez les autres régressions linéaires multiples permettant de justifier les autres éléments indiqués.

Réalisez également des régressions linéaires multiples montrant que les variables support n'ont pas d'effet sur l'engagement normatif envers l'organisation mais qu'en revanche, un effet apparaît lorsqu'on introduit les 5 variables prédictives dans le modèle.

De même, quels sont les éléments de ces tableaux qui permettent d'énoncer la conclusion suivante :

"Les résultats indiquent que la congruence subjective P–O agit comme médiateur total dans la relation entre les valeurs organisationnelles de support et l'engagement affectif envers l'organisation. En effet :
• ces valeurs organisationnelles de support ont un effet principal sur l'engagement affectif organisationnel ;
• ces valeurs organisationnelles de support sont positivement liées à la congruence subjective P–O ($r = 0,41$, $p < 0,01$) ;
• cette dernière a un effet indépendant sur l'engagement affectif envers l'organisation ;
• et l'effet principal des valeurs organisationnelles de support sur l'engagement affectif envers l'organisation disparaît lorsque la congruence subjective P–O est introduite dans l'équation de régression."

Procédez de façon analogue pour obtenir les autres résultats indiqués par les auteurs :

- "La congruence subjective P–G est un médiateur de l'effet des valeurs de support du groupe sur l'engagement affectif envers le groupe. Étant donné que les valeurs de support du groupe exercent encore un effet significatif sur l'engagement affectif envers ce groupe lorsque la congruence subjective P–G est introduite dans l'équation, nous ne pouvons cependant conclure qu'à une médiation partielle et non totale."

- Les variables de congruence subjective n'ont pas d'effet de médiateur sur l'engagement normatif envers le groupe.

3.1.6 L'exemple d'analyse de modération traité avec Statistica

Les données nécessaires sont rassemblées dans le classeur Statistica tra.stw.

Calculez la matrice des corrélations pour les 4 variables beliefs, values, attitudes, intentions. Vous devriez retrouver le résultat donné au paragraphe 3.1.3.

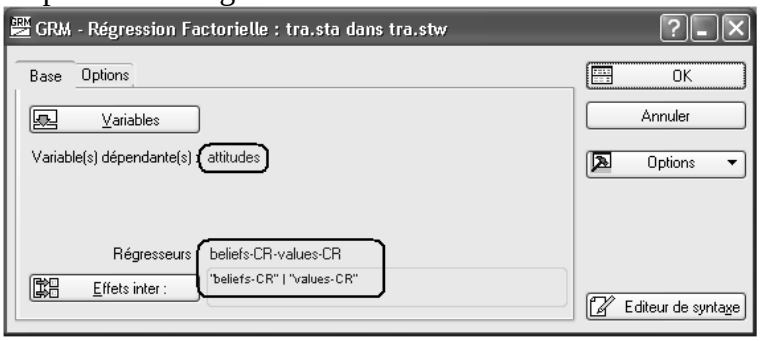
Réalisez la régression linéaire de attitudes sur beliefs :

N=1000	Synthèse de la Régression; Variable Dép. : attitudes (tra.sta dans tra.stw) R= ,37665219 R²= ,14186687 R² Ajusté = ,14100702 F(1,998)=164,99 p<0,0000 Err-Type de l'Estim.: ,61952					
	Bêta	Err-Type de Bêta	B	Err-Type de B	t(998)	niveau p
OrdOrig.			-0,0249	0,0196	-1,2696	0,2045
beliefs	0,3767	0,0293	0,2535	0,0197	12,8448	0,0000

Réalisez ensuite la régression linéaire multiple de attitudes sur beliefs et values :

N=1000	Synthèse de la Régression; Variable Dép. : attitudes (tra.sta dans tra.stw) R= ,51345125 R²= ,26363219 R² Ajusté = ,26215502 F(2,997)=178,47 p<0,0000 Err-Type de l'Estim.: ,57417					
	Bêta	Err-Type de Bêta	B	Err-Type de B	t(997)	niveau p
OrdOrig.			-0,0218	0,0182	-1,2010	0,2300
beliefs	0,3674	0,0272	0,2472	0,0183	13,5152	0,0000
values	0,3491	0,0272	0,2362	0,0184	12,8399	0,0000

Enfin, utilisez le menu Statistiques -> Modèles Linéaires/non linéaires avancés -> Modèles généraux de régression puis l'item Régression Factorielle :



Vous devriez retrouver le résultat donné au paragraphe 3.1.3 :

Effet	Paramètres Estimés (tra.sta dans tra.stw) Paramétrisation sigma-restreint						
	attitudes Param.	attitudes Err-Type	attitudes t	attitudes p	-95,00% Lim.Conf	+95,00% Lim.Conf	attitudes Bêta (ß)
Ord.Orig.	-0,0383	0,0164	-2,3378	0,0196	-0,0705	-0,0062	
beliefs-CR	0,2428	0,0164	14,8032	0,0000	0,2106	0,2750	0,3632
values-CR	0,2136	0,0165	12,9837	0,0000	0,1813	0,2459	0,3196
beliefs-CR*values-CR	0,2434	0,0161	15,0996	0,0000	0,2118	0,2751	0,3716

Réalisez ensuite le graphique donné à la fin du paragraphe 3.1.3.

3.2 Régression logistique

Bibliographie :

Howell, D.C., Méthodes Statistiques en Sciences Humaines, De Boeck, Paris Bruxelles, 1998.
Lebart, L., Morineau, A., Piron M., Analyse exploratoire multidimensionnelle, Dunod, Paris, 2000.

3.2.1 La régression logistique

La régression logistique peut être vue comme une extension de la régression linéaire au cas où la variable dépendante est dichotomique. Plus précisément, sur un échantillon de n individus statistiques, on a observé :

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

Dans le cas le plus simple, on cherche à expliquer une variable dichotomique Y par une variable numérique X. On dispose donc d'un tableau de données sous la forme :

	s1	s2		sn
Y	1	0	...	0
X	x1	x2		xn

Exemple : On considère un échantillon de 30 sujets pour lesquels on a relevé :

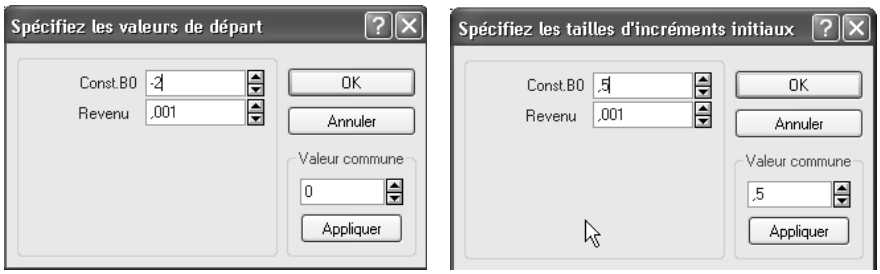
- d'une part le niveau des revenus (variable numérique)
- d'autre part la possession ou non d'un nouvel équipement électro-ménager.

On a obtenu les données suivantes :

Revenu	1085	1304	1331	1434	1541	1612	1729	1759	1863	2121	2395	2681	3390	4237	1241
Possède	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1

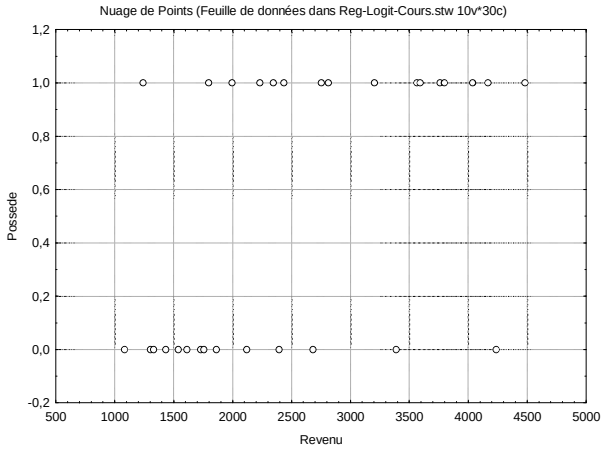
Revenu	1798	1997	2234	2346	2436	2753	2813	3204	3564	3592	3762	3799	4037	4168	4484
Possède	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

N.B. Cet exemple peut être traité sous Statistica à l'aide du menu Statistiques > Modèles linéaires/non linéaires avancés > Estimation non-linéaire > Régression Logit. Mais les valeurs initiales par défaut des paramètres ne conviennent pas. Il faut indiquer par exemple, sous l'onglet Avancé de la fenêtre "Estimation du modèle" :



3.2.1.1 Principe de la méthode

Ces données peuvent être représentées à l'aide d'un nuage de points, qui a l'allure suivante :



On cherche un modèle permettant d'estimer Y ("Possède") connaissant X ("Revenu"). Plutôt que de rechercher un modèle mathématique donnant pour une valeur donnée X exactement la valeur 0 ou la valeur 1, il peut sembler pertinent de rechercher un modèle produisant des valeurs comprises entre 0 et 1 qui seront interprétées comme des probabilités. Par exemple :

$\hat{Y} = 0{,}1$ signifie que : il y a 10% de chances que Y=1

Cependant, la droite de régression de la variable Y par rapport à la variable X ne constitue pas un bon modèle car les valeurs estimées ne seront pas limitées à 0 et 1.

Pour passer d'une variable prenant ses valeurs dans [0, 1] à une variable prenant ses valeurs dans [0, +∞[, on introduit le rapport de chances ou cote :

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

Ainsi, si P(Y=1)=0,9, le rapport de chances vaut p1 = 0,9/0,1=9 : on a 9 fois plus de chances d'observer Y=1 que Y=0.
De même, si P(Y=1)=0,2, le rapport de chances vaut p1 = 0,2/0,8=1/4 : on a 4 fois plus de chances d'observer Y=0 que Y=1.

Pour passer d'une quantité (le rapport de chances) variant dans [0, +∞[à une quantité prenant n'importe quelle valeur réelle, on applique une nouvelle transformation, en prenant le logarithme népérien du rapport. On obtient ainsi la transformation logit :

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

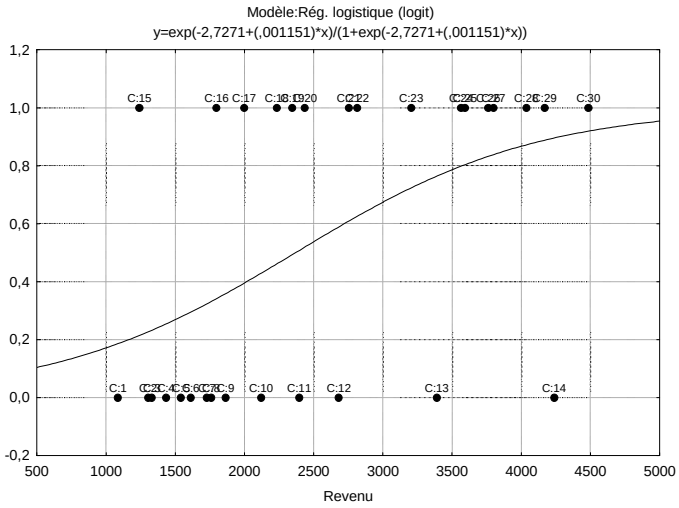
Ainsi,

- si P = 0,9, logit(P) = ln 9 = 2,1972
- si P = 0,5, logit(P) = ln 1 = 0
- si P = 0,2, logit(P) = ln(1/4) = -1,3863.

A partir d'une "valeur logit" y, on peut facilement revenir à la probabilité P correspondante en appliquant la transformation :

$$P = \frac{e^y}{1+e^y}$$

On ajuste alors logit(P) par une fonction affine, ce qui revient à déterminer une "sigmoïde" qui passe au mieux par les points expérimentaux :



L'équation correspondant à cet ajustement est :

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

Exemple d'utilisation de cette équation : à partir de quel revenu a-t-on 90% de chances de tirer un sujet possédant l'équipement envisagé ?

P = 0,9 correspond à P/(1-P) = 0,9/0,1 = 9 d'où logit(P) = 2,1972.
Or : 2,1972 = -2,7271 + 0,001151 X donne X = (2,1972 + 2,7271)/0,001151, c'est-à-dire : X=4278.

Remarque : Cette équation n'est pas obtenue par une "simple" régression linéaire, mais par des méthodes itératives. D'une part, il n'est pas envisageable de faire les calculs manuellement, d'autre part, il faudra, dans certains cas, "aider" les logiciels en indiquant des valeurs initiales plausibles pour les coefficients.

3.2.1.2 Aides à l'interprétation. Evaluation de la qualité du modèle obtenu.

La qualité du modèle peut être évaluée en comparant les résultats obtenus avec ceux du modèle "constant" qui attribuerait la probabilité 14/30 à la valeur 0 et 16/30 à la valeur 1. Une fonction de vraisemblance est évaluée dans les deux cas, et la différence des deux fonctions suit une loi du khi-2 à 1 degré de liberté lorsqu'il n'y a qu'une seule variable indépendante. Autrement dit, les hypothèses du test sont ici :

H_0 : le modèle n'est pas significativement différent du modèle constant ;
H_1 : le modèle est significativement différent du modèle constant.

Sur notre exemple, on obtient :

Chi-deux = 7,636181 ; dl = 1 ; p = ,0057242

Le revenu est donc un prédicteur significatif de la variable Y.

Une autre aide à l'interprétation courante est le rapport de cotes ou odds-ratio (OR). En particulier, la contribution de la variable X à la variation de Y est calculée par :

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

Ainsi, sur notre exemple, l'odds-ratio correspondant au coefficient 0,001151 est : $e^{0,001151}$=1,0012. Autrement dit, une augmentation du revenu de 1 unité se traduit par une multiplication de la probabilité par 1,0012.

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

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

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

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

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

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

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

On évalue également un Odds-ratio comparant valeurs observées et valeurs prévues. Pour cela, on définit deux classes dans les valeurs prévues : celles inférieures à 0,5 et celles supérieures à 0,5 et on forme le tableau de contingence croisant les valeurs observées (0 ou 1) avec les classes ainsi définies. Sur notre exemple, on obtient :

Obs	Prév. < 0,5	Prév. > 0,5
0	10	4
1	5	11

Le rapport est alors obtenu en formant le rapport ad/bc (produit des effectifs des cases d'accord divisé par le produit des effectifs des cases de désaccord).
On obtient ainsi :

$$\text{OR} = \frac{10 \times 11}{5 \times 4} = 5{,}50$$

3.2.2 La régression logistique avec Statistica

Source : Howell. p. 633, ex. 15.31 a 15.33

La feuille de données Harass contient des données légèrement modifiées relatives à 343 cas créés pour répliquer les résultats d'une étude sur le harcèlement sexuel (Brooke et Perot 1991). Les variables sont :
- l'âge
- l'état-civil (1 = marié(e), 2 = célibataire) (NB étonnant, n'est-ce pas l'inverse? cf données)
- l'idéologie féministe
- la fréquence du comportement
- le caractère agressif du comportement
- le fait qu'il ait été ou non signale (0 = non, 1 = oui).

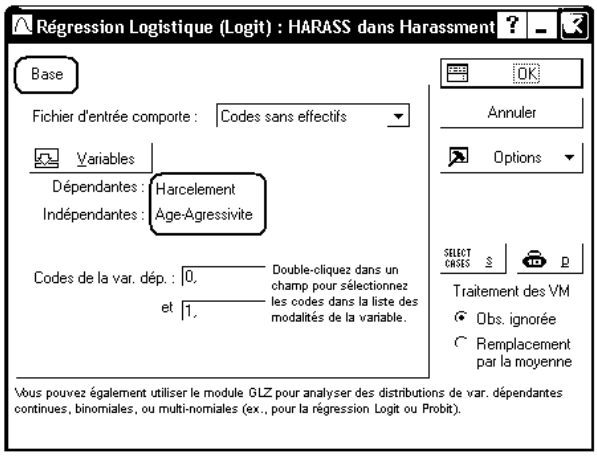
1) Utiliser un programme de régression logistique et examiner la probabilité qu'un sujet signale un cas de harcèlement sexuel sur la base des VI.

2) Même question, mais en n'utilisant que le prédicteur dichotomique relatif à l'état civil. Faire une table de contingence, calculer les rapports de chances et comparer ces résultats à ceux de la régression logistique. (résultats son significatifs, mais cela importe peu, selon Howell).

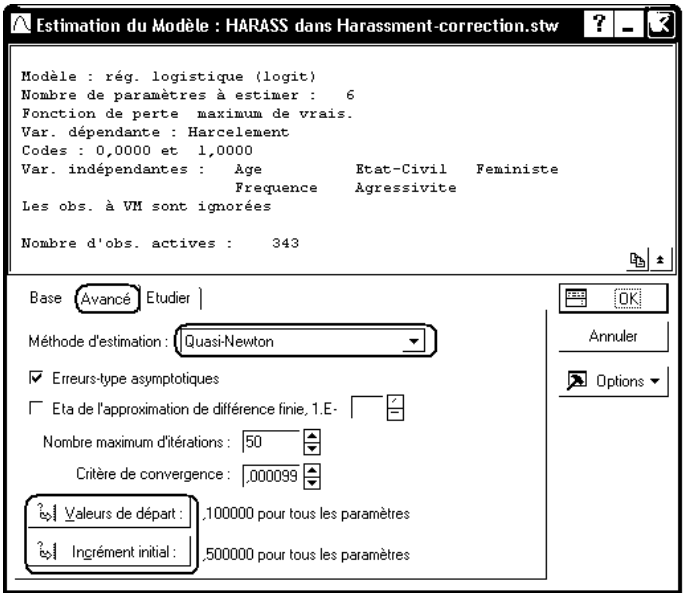
3) Apparemment, la fréquence du comportement n'est pas liée à la probabilité de voir la victime signaler le cas de harcèlement. Peut-on en imaginer les raisons ?

Ouvrez le classeur Harassment.stw.

On peut utiliser le menu Statistiques, Modèles linéaires/non-linéaires avancés, Estimation non linéaire, Régression Logit: On indique la variable dépendante et les variables indépendantes :



On peut ensuite choisir un algorithme d'estimation et éventuellement indiquer manuellement les valeurs initiales des coefficients bi, ce qui est souvent utile, si les plages de variations des VI sont très différentes de l'intervalle [0, 1] (et n'est pas prévu par le menu précédent). Pour obtenir le tableau de résultats indiqué ci-dessous, il faut également cocher la boîte "Erreurs-types asymptotiques".



Le tableau de résultats est alors accessible par le bouton "Synthèse : paramètres et erreurs-types" du dialogue des résultats.

N=343	Modèle: Rég. logistique (logit) Nbre de 0 : 174 1 : 169 Var dép. : Harcelement Perte : Max vraisemblance (MC-er. posit. à Perte finale= 219,99193498 Chi²(5)=35,442 p=,00000					
	Const.B0	Age	Etat-Civil	Feministe	Frequence	Agressivite
Estimat.	-1,7317	-0,0137	-0,0723	0,0070	-0,0464	0,4878
Erreur-type	1,4296	0,0129	0,2338	0,0146	0,1525	0,0949
t(337)	-1,2113	-1,0614	-0,3091	0,4771	-0,3043	5,1409
niveau p	0,2266	0,2893	0,7575	0,6336	0,7611	0,0000
-95%CL	-4,5439	-0,0391	-0,5321	-0,0218	-0,3464	0,3011
+95%CL	1,0804	0,0117	0,3876	0,0358	0,2536	0,6744
Chi² de Wald	1,4672	1,1265	0,0955	0,2277	0,0926	26,4292
niveau p	0,2258	0,2885	0,7573	0,6333	0,7609	0,0000
Odds ratio (unité)	0,1770	0,9864	0,9303	1,0070	0,9547	1,6287
-95%CL	0,0106	0,9617	0,5874	0,9784	0,7072	1,3514
+95%CL	2,9460	1,0118	1,4734	1,0364	1,2887	1,9629
Odds r. (étendue)		0,4644	0,9303	1,3985	0,8306	80,6394
-95%CL		0,1121	0,5874	0,3509	0,2501	15,0336
+95%CL		1,9244	1,4734	5,5731	2,7579	432,5462

L'équation de la courbe de régression est :

$$\text{logit}\, P = -1{,}7317 - 0{,}013698\,\text{Age} - 0{,}072251\,\text{EtatCivil} + 0{,}0069870\,\text{Feministe} - 0{,}046408\,\text{Frequence} + 0{,}4878\,\text{Agressivite}$$

Le khi-2 correspondant au modèle vaut 35,442, et il est significatif au seuil de 1%. En revanche, seule la variable Agressivite semble avoir un rôle explicatif supérieur à celui que le hasard est susceptible de produire.

Les odds-ratio unitaires correspondant aux différentes variables sont :

N=343	Modèle: Rég. logistique (logit) Nbre de 0 : 174 1 : 169 (HARASS) Var dép. : Harcelement Perte : Max vraisemblance (MC-er. posit. à Perte finale= 219,99193498 Chi²(5)=35,442 p=,00000					
	Const.B0	Age	Etat-Civil	Feministe	Frequence	Agressivite
Odds ratio (unité)	0,1770	0,9864	0,9303	1,0070	0,9547	1,6287

On voit que seules les variables Feministe et Agressivite possèdent des odds-ratio unitaires supérieurs à 1 et que seul celui de Agressivite est nettement différent de l'unité. Lorsqu'on affiche les résultats complets (boîte à cocher "erreurs asymptotiques" activée), on peut également observer les intervalles de confiance de ces odds ratio. On constate alors qu'Agressivite est la seule variable pour laquelle les deux bornes de l'intervalle de confiance sont d'un même côté de la valeur 1.

On peut également afficher le tableau des valeurs observées et des valeurs prévues de la variable dépendante :

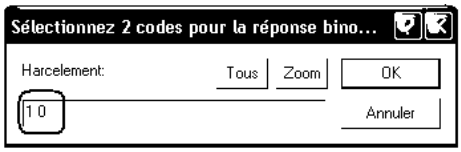
	Modèle : (HARASS dans Harassment-correction.stw Var. Dép. : Harcelement		
	Observée	Prév.	Résidus
1	0,0000	0,6431	-0,6431
2	0,0000	0,7312	-0,7312
3	1,0000	0,8696	0,1304
4	1,0000	0,3080	0,6920

Sous l'onglet Résidus, on peut obtenir le calcul de l'odds-ratio pour le modèle :

	Classification d'obs. (HARASS) Odds ratio : 2,1051 Pourc. corrigé : 59,18%		
Observée	Prév. 0,000000	Prév. 1,000000	% Corrigé
0,000000	111	63	63,79310
1,000000	77	92	54,43787

On peut également utiliser le menu : Statistiques, Modèles linéaires/non-linéaires avancés, Modèles linéaires/non linéaires généralisés, puis l'item Modèle logit dans l'onglet Base ou les items : Régression simple (ou multiple), Distribution: Binomiale et Fonction de liaison : logit de l'onglet Avancé.

Lorsqu'on indique les variables et leur rôle, il est important de préciser que c'est le code "1" de la variable Harcelement qui doit être assimilé à la modalité "succès" de la variable binomiale, faute de quoi les résultats seraient inversés :



On retrouve ainsi les résultats obtenus par les deux autres méthodes, mais avec une présentation différente. On peut également obtenir des résultats supplémentaires, tels que l'évolution des valeurs des coefficients à chaque itération de l'algorithme :

	Harcelement - Historique itérations (HARASS) Distribution : BINOMIALE Fonction de Liaison : LOGIT					
Effet	Niveau Effet	Colonne	Itérat. 0	Itérat. 1	Itérat. 2	Itérat. 3
Ord.Orig		1	0,000	-1,556	-1,727	-1,732
Age		2	0,000	-0,012	-0,014	-0,014
Etat-Civil		3	0,000	-0,066	-0,072	-0,072
Feministe		4	0,000	0,006	0,007	0,007
Frequence		5	0,000	-0,044	-0,046	-0,046
Agressivite		6	0,000	0,438	0,486	0,488
Vraisembl.			-237,749	-220,156	-219,992	-219,992

On peut également noter que l'on obtient des résultats légèrement différents lorsque l'on indique "Etat-Civil" comme variable catégorielle.

3.2.3 Un exemple de régression logistique issu d'un article.

Réf. : Factors Influencing Adolescents Engagement in Risky Internet Behavior, ALBERT KIENFIE LIAU, Ph.D., ANGELINE KHOO, Ph.D., and PENG HWAANG, Ph.D., CYBERPSYCHOLOGY & BEHAVIOR, Volume 8, Number 6, 2005, pp 513-520.

Dans l'article cité supra les auteurs se sont intéressés aux facteurs liés à la prise de risques dans le comportement sur Internet pour des adolescents de Singapour. Ils identifient notamment comme conduite à risques le fait de rencontrer physiquement une personne qu'ils ont d'abord connu "online".

Dans les résultats de leur étude, les auteurs indiquent notamment :

1045 (93.0% of the total sample) adolescents reported having used the Internet, and 827 (73.6%) adolescents reported having chatted on the Internet. The study focused on this group of 827 adolescents who have experienced chatting on the Internet. These adolescents have a mean age = 14.42 (SD = 1.33) and are 51.4% girls. (...)
A total of 169 adolescents (16.2% of Internet users, or 20.4% of those who chat) reported having met someone in real life that they first encountered online.

A series of multiple logistic regression analyses was used to examine the factors that influence adolescents' engagement in risky internet behavior, in particular, meeting in person with someone encountered online. Odds ratios (OR) were calculated to approximate relative risk and are presented with 99% confidence intervals. Age was a significant predictor of the risky behavior (OR = 1.26, 99% CI (1.06, 1.48), $p < 0.0001$) but gender was not a significant predictor; 80 out of the 169 (47.3%) adolescents were girls. For ease of interpretation, the frequency of use of the Internet variable was dichotomized so that 1 = "at least once a day" and 0 = "less than once a day." Controlling for age, frequency of use of the Internet was a significant predictor of the risky behavior (OR = 1.68, 99% CI (1.07, 2.65), $p < 0.01$). Parents' educational background and whether parents lived together were not significant predictors of the risky behavior. All subsequent analyses include age and frequency of use as covariates in order to control for the influence of these factors. The following factors were examined as predictors of the risky behavior: frequency of chatting and gaming behavior, parental supervision, communication with parents, type of personal information given out, amount of inappropriate messages received, whether inappropriate websites have been visited, and type of internet advice heard. Significant and marginally significant predictors of the risky behavior are reported in Table 2.

TABLE 2. SIGNIFICANT AND MARGINALLY SIGNIFICANT PREDICTORS OF THE RISKY INTERNET BEHAVIOR—MEETING IN PERSON SOMEONE ENCOUNTERED ONLINE

Predictor	*OR*	*99% CI*
Frequency of Internet activities	3.13**	1.75, 5.55
Frequency of chatting	1.77*	1.07, 2.91
Frequency of gaming		
Parental supervision		
Rules for Internet use		
Not allowed to meet in person someone encountered online	0.49**	0.30, 0.81
Not allowed to talk to strangers in chatrooms	0.46*	0.23, 0.93
Not allowed to give out personal information	0.62†	0.39, 1.01
People usually at home when arrive from school	1.56†	1.06, 1.48
Communication with parents		
Tell parents about receiving pornographic junk mail	0.49†	0.22, 1.06
Giving out personal information		
Phone number	2.17*	1.15, 4.09
Photograph	2.68*	1.16, 6.18
Favorite band, music	1.67*	1.03, 2.90
Receiving inappropriate message		
Met someone on the Internet who asked for personal information	4.16**	2.42, 6.67
Sent pornography from someone met only on the Internet	1.80†	0.97, 3.34
Received unwanted sexual comments on the Internet	2.59**	1.58, 4.23
Received pornographic junk mail in e-mail or Instant Messaging	1.90**	1.19, 3.04
Visiting Inappropriate websites		
Accidentally ended up in a pornographic website	1.68*	1.04, 2.73
Purposely visited a pornographic website	2.39**	1.33, 4.28
Accidentally ended up in a website with violent/gruesome images	1.60*	1.01, 2.54
Accidentally ended up in a hate website	1.44†	0.90, 2.33
Heard of the following Internet safety advice		
Never arrange to meet anyone	0.55*	0.33, 0.90
Do not download anything	1.88*	1.06, 3.17

**$p < 0.0001$.
*$p < 0.01$.
†$p < 0.05$.

3.3 Introduction à l'analyse discriminante

3.3.1 Présentation de la méthode

3.3.1.1 Position du problème

On dispose de n observations sur lesquelles on a relevé :

- les valeurs d'une variable catégorielle comportant quelques modalités (2, 3, ...) : c'est le groupe ou diagnostic.
- les valeurs de p variables numériques : X_1, X_2, ..., X_p : ce sont les prédicteurs.

On se pose des questions telles que :

- dans quelle mesure la valeur de Y est-elle liée aux valeurs de X_1, X_2, ..., X_p ?
- Etant donné d'autres observations, pour lesquelles X_1, X_2, ..., X_p sont connues, mais Y ne l'est pas, est-il possible de prévoir Y (le groupe), et avec quel degré de certitude ?

Exemples de situations où une telle méthode peut être intéressante :

Exemple 1. On étudie les différentes espèces de poissons peuplant un lac, mais la détermination exacte de l'espèce suppose que l'on sacrifie l'animal. Peut-on se contenter de relever différents paramètres concernant les poissons prélevés, et déduire l'espèce à partir de ces paramètres avec un degré de certitude raisonnable ?

Exemple 2. Pour déterminer le type d'utilisation de parcelles agricoles, on peut évidemment faire des relevés sur le terrain. Mais pourrait-on utiliser les informations données par des images satellites ?

La méthode est également utilisée sans que l'on ait un objectif de prédiction; on souhaite seulement déterminer les prédicteurs les plus liés au groupe d'appartenance. De ce point de vue, l'analyse discriminante est alors un complément à l'analyse de variance multivariée ou MANOVA.

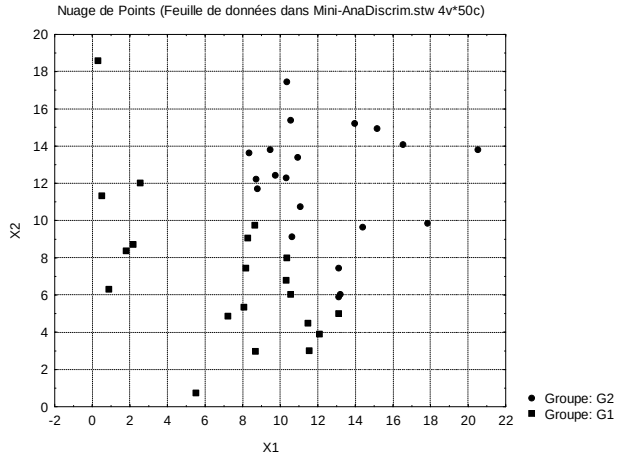
3.3.1.2 Précautions et limites de la méthode

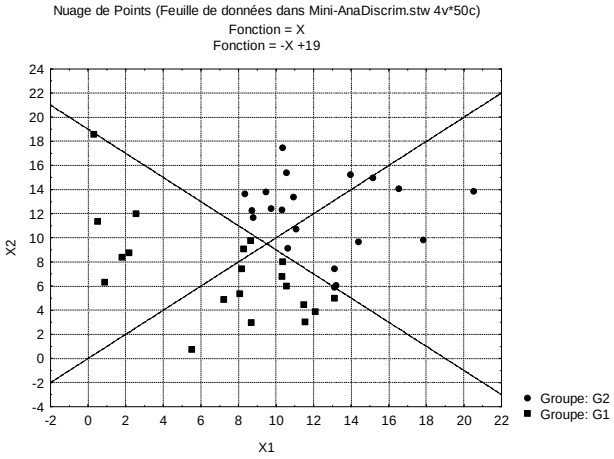
Comme dans le cas de la régression linéaire, l'emploi de cette méthode suppose que les variables prédictrices possèdent des propriétés de régularité satisfaisantes : distribution normale (voire multinormale) des variables X_i dans les différentes populations.
Par ailleurs (comme pour la régression linéaire), l'analyse discriminante peut conduire à des résultats incorrects si les variables X_i sont trop fortement corrélées entre elles.

3.3.2 Analyse discriminante sur un mini-exemple

3.3.2.1 Présentation de l'exemple

On a relevé les valeurs de deux variables X_1 et X_2 sur 40 individus statistiques répartis en deux groupes. Le nuage de points représentant ces observations est le suivant :





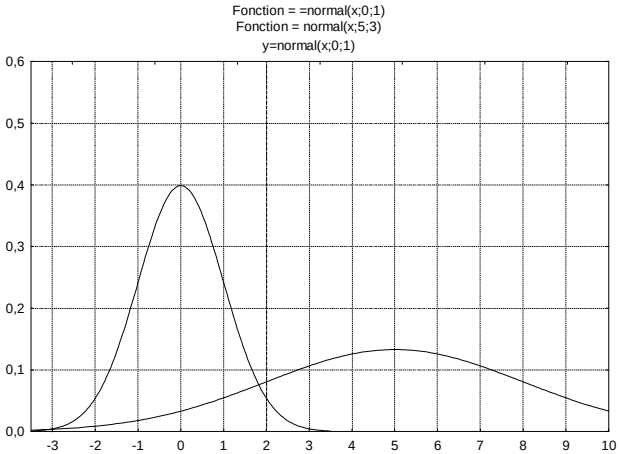
Prise isolément, aucune des deux variables X_1 et X_2 ne permet de différencier les deux groupes G_1 et G_2. Cependant, on voit bien que les deux groupes occupent des régions du plan bien spécifiques. On voit intuitivement que notre problème pourrait être résolu en considérant une variable abstraite, combinaison linéaire de X_1 et X_2 (approximativement $X_1 + X_2$) définie de façon que :

- la variance (dispersion) intra-groupes soit la plus petite possible
- la variance inter-groupes (variance calculée à partir des points moyens pondérés des groupes) soit la plus grande possible.

Ainsi, sur notre exemple, la droite d'équation $X_2 = - X_1 + 19$ semble séparer correctement les deux groupes et il semblerait que c'est en projetant les points sur la droite $X_2=X_1$ que l'on obtiendra une dispersion minimale dans les groupes et maximale entre les groupes.

Remarque : *distance de Mahalanobis*. Dans notre exemple, les deux groupes présentent à peu près la même dispersion de valeurs. Cependant, dans d'autres situations, l'un des groupes peut être nettement plus dispersé que l'autre.

Considérons la situation suivante, où l'on a représenté la distribution des valeurs issues de deux groupes sur un "facteur discriminant". Dans le premier groupe, cette distribution est normale, de moyenne 0 et d'écart type 1. Dans le second groupe, elle est normale, de moyenne 5 et d'écart type 3. On souhaite, par exemple, affecter la valeur x=2 à l'un des deux groupes. Pour la distance "habituelle" (euclidienne), cette valeur est plus près du centre du premier groupe (valeur $\bar{x}_1 = 0$) que du centre du second groupe (valeur $\bar{x}_2 = 5$). Cependant, x=2 a plus de chances d'être une observation provenant du second groupe qu'une observation provenant du premier groupe.



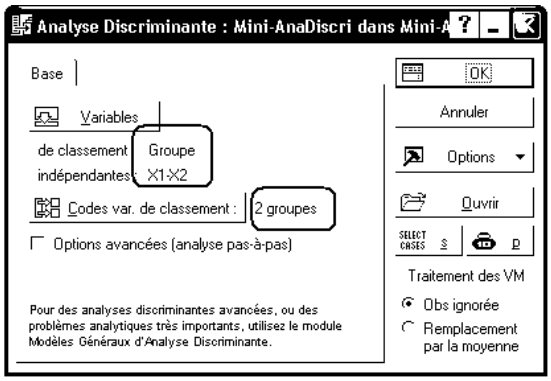
Pour résoudre ce problème, on introduit une distance particulière : la **distance de Mahalanobis** pour évaluer la distance entre un point et le centre d'un groupe. Pour calculer cette distance, on fait intervenir les écarts réduits entre x et les centres de groupes. On aura ainsi :

$$d_1^2(\bar{x}_1,2)=\left(\frac{2-0}{1}\right)^2=4 \quad ; \quad d_2^2(\bar{x}_2,2)=\left(\frac{2-5}{3}\right)^2=1$$

La définition de la distance de Mahalanobis est nettement plus compliquée lorsqu'il y a plusieurs variables à prendre en compte, car elle fait intervenir les covariances des variables prises deux à deux.

3.3.2.2 Traitement de l'exemple précédent avec Statistica

Ouvrez le fichier Mini-AnaDiscrim.stw
Faites une analyse discriminante (menu Statistiques - Techniques exploratoires multivariées - Analyse discriminante) *en indiquant les codes G2 et G1 comme codes pour la variable catégorielle* "Groupe", X1 et X2 comme variables indépendantes.
N.B. Un troisième code existe : G?. Mais nous voulons utiliser ces dernières observations pour tester les capacités de prédiction du modèle.



L'onglet Avancé nous donne accès aux boutons suivants :

Synthèse (variables dans le modèle) :

N=40	Synthèse de l'Analyse Discriminante (Mini-AnaDiscri dans Mini-AnaDiscrim.stw Vars dans le modèle : 2; Classmt : Groupe (2 grps) Lambda Wilk : ,38021 F approx. (2,37)=30,158 p< ,0000					
	Wilk (Lambda)	Partiel (Lambda)	F d'exc. (1,37)	niveau p	Tolér.	1-Tolér. (R^2)
X1	0,676419	0,562090	28,82580	0,000004	0,838237	0,161763
X2	0,668372	0,568857	28,04266	0,000006	0,838237	0,161763

Cette feuille donne les résultats de plusieurs tests.

Dans la boîte de synthèse, on lit la valeur du test lambda de Wilk pour le modèle formé par l'ensemble des prédicteurs : $\Lambda = 0{,}3821$. Il s'agit en fait de la statistique d'une MANOVA à un facteur : globalement, les centres de gravité des différents groupes sont-ils discernables (H1) ou non (H0) à partir des prédicteurs choisis ? La significativité de cette statistique de test est évaluée à partir d'une approximation par un F de Fisher (F=30,158, $p<10^{-4}$). On retrouve la valeur de cette statistique en réalisant une analyse de variance à un facteur dans laquelle on indique X1 et X2 comme variables dépendantes (menu Statistiques - ANOVA). On notera que les valeurs significatives de lambda sont les valeurs proches de 0.

Sur notre exemple, on constate que les deux groupes sont très significativement discernables ($p <10^{-4}$).

La colonne Wilk (Lambda) indique la valeur du lambda de Wilk que l'on obtiendrait en supprimant du modèle, la variable concernée. Par exemple, si on reprend l'analyse discriminante en n'indiquant seulement X2 comme prédicteur, on obtient lambda=0,676419.

La colonne Partiel (Lambda) donne la valeur du lambda de Wilk associée à la variable correspondante. Les deux colonnes suivantes testent la significativité de ce lambda partiel.

La colonne 1-Tolér. indique la valeur de R^2, où R est un coefficient de corrélation calculé à partir de la régression de cette variable sur toutes les autres, avec neutralisation de l'effet dû aux groupes. La tolérance est une mesure de la redondance de la variable correspondante. Par exemple, une valeur de tolérance de 0,10 signifie que la variable est redondante à 90% avec toutes les autres variables du modèle.

La valeur indiquée pour notre exemple peut être retrouvée de la manière suivante :

- On calcule les écarts à la moyenne des différentes observations, pour chaque variable dans chaque groupe.
- On calcule le carré du coefficient de corrélation entre les deux variables ainsi obtenues.

Distances inter-groupes fournit trois feuilles de résultats :

1. Distance entre les centroïdes des deux groupes

	Dist. de Mahalanobis au Carré (Mini-AnaDiscri dans Mini-AnaDiscrim.stw	
Groupe	G2	G1
G2	0,000000	6,194525
G1	6,194525	0,000000

Il s'agit du carré de la distance entre les points moyens des deux groupes, mesurée à l'aide de la distance de Mahalanobis.

2. Un test statistique concernant la séparation des deux groupes

	Valeurs F ; dl 2,37 (Mini-AnaDiscri dans Mini-AnaDiscrim.stw	
Groupe	G2	G1
G2		30,15756
G1	30,15756	

Il s'agit d'un test du même type que le lambda de Wilk global, mené en considérant les groupes deux à deux. La feuille suivante indique les niveaux de significativité des valeurs trouvées :

	niveau p (Mini-AnaDiscri dans Mini-AnaDiscrim.stw	
Groupe	G2	G1
G2		1,6998E-8
G1	1,6998E-8	

On retrouve ainsi le résultat déjà observé dans la synthèse : les centres de gravité des deux groupes sont très significativement distincts.

L'aspect "classification et prédiction" de la méthode est accessible sous l'onglet Classification qui donne accès aux résultats suivants :

N.B. Les résultats qui suivent dépendent des probabilités a priori spécifiées pour les groupes. Ici, nous pouvons prendre p=0,5 dans chacun des groupes. Mais dans d'autres situations, on peut savoir qu'a priori, un groupe est nettement plus fréquent que l'autre, même si les échantillons sont équilibrés.

Fonctions de classification

	Fonctions de classif. ; classement: Groupe	
Variable	G2 p=,50000	G1 p=,50000
X1	1,4380	0,83765
X2	1,5504	0,91594
Constte	-18,8231	-6,93504

La fonction de classification du groupe G2 est :

F2 = 1,4380 X1 + 1,5504 X2 - 18,8231

Celle du groupe G1 est :

F1 = 0,83765 X1 + 0,91594 X2 - 6,93504

La méthode classe un élément dans le groupe G1 si F1 > F2 et dans G2 dans le cas contraire.

Matrice de classification

Ce tableau est encore appelé *Matrice de confusion*. Il croise la classification observée avec la classification calculée par la méthode, à l'aide des fonctions de classification précédentes.

	Matrice de Classification Lignes : classifications observées Colonnes : classifications prévues		
Groupe	% Correct	G2 p=,50000	G1 p=,50000
G2	90,00000	18	2
G1	95,00000	1	19
Total	92,50000	19	21

Sur notre exemple, 18 des 20 observations du groupe 1 et 19 des 20 observations du groupe 2 sont correctement classées par la méthode.

Classification d'observations

	Classification d'observations Classif. incorrectes indiquées par *		
Observation	Classif. Observée	1 p=,50000	2 p=,50000
1	G2	G2	G1
2	G1	G1	G2
3	G1	G1	G2
4	G2	G2	G1
5	G2	G2	G1
* 6	G2	G1	G2
7	G2	G2	G1

Ce tableau donne pour chaque observation, le groupe le plus probable (selon le calcul), ainsi que le second candidat. Il indique également le classement calculé des valeurs qui n'étaient pas classées a priori :

Observation	Classification d'observations Classif. incorrectes indiquées par ' Classif. Observée	1 p=,50000	2 p=,50000
40	G2	G2	G1
41	---	G2	G1
42	---	G1	G2
43	---	G1	G2
44	---	G2	G1
45	---	G1	G2
46	---	G1	G2

Distances de Mahalanobis au carré

Observation	Dist. Mahalanobis Carrées aux Centroïdes de Group Classif. incorrectes indiquées par * Classif. Observée	G2 p=,50000	G1 p=,50000
1	G2	1,02516	3,21197
2	G1	9,43294	0,46701
3	G1	4,30475	0,81343
4	G2	1,71089	3,08567
5	G2	0,20940	6,53698
* 6	G2	2,95094	2,71562
7	G2	0,48679	4,17369

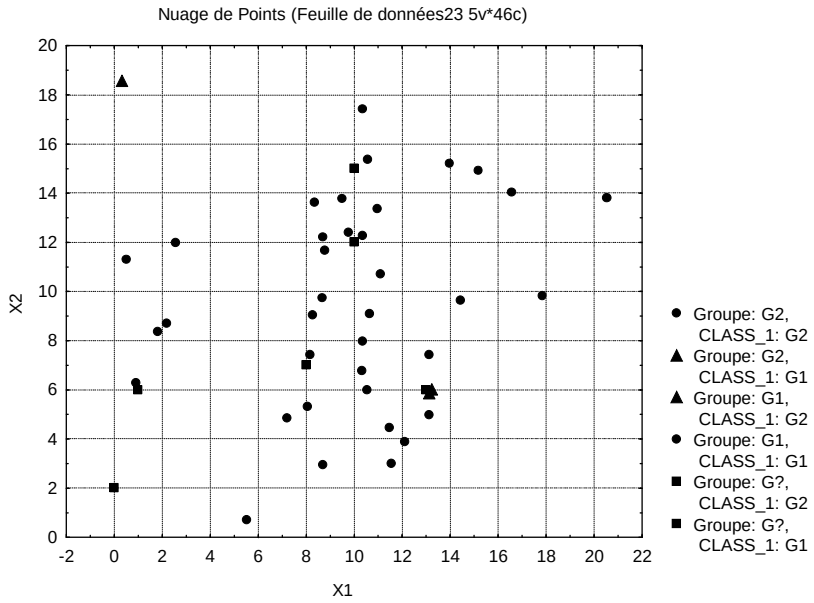
Probabilités a posteriori

Observation	Probabilités a posteriori Classif. incorrectes indiquées par ' Classif. Observée	G2 p=,50000	G1 p=,50000
1	G2	0,749022	0,250978
2	G1	0,011174	0,988826
3	G1	0,148595	0,851405
4	G2	0,665386	0,334614
5	G2	0,959449	0,040551
* 6	G2	0,470618	0,529382
7	G2	0,863356	0,136644

En fait pour chaque observation la méthode calcule une probabilité d'appartenance à chacun des deux groupes et affecte l'observation au groupe le plus probable.

Enregistrer les scores

Ce bouton permet de générer une feuille de données avec la classification produite par la méthode, et éventuellement, les variables et la classification initialement observées. Cette feuille de données peut être utilisée pour produire un nuage de points tel que le suivant :



Dans ce graphique, les points bien classés sont représentés par des cercles, les points mal classés par des triangles, et les points supplémentaires par des carrés. La couleur (rouge ou noir) correspond au groupe calculé.

La méthode correspondant aux résultats qui viennent d'être décrits est appelée **analyse discriminante décisionnelle** (*predictive discriminant analysis*). Comme le montrent ces résultats, son but est de définir une règle d'affectation permettant de classer un individu donné dans un groupe donné, parmi plusieurs groupes définis préalablement. Mais, la fenêtre de dialogue de Statistica comporte également un bouton "Réaliser une analyse canonique". Cette dernière méthode, appelée **analyse canonique discriminante** ou **analyse factorielle discriminante** (*canonical discriminant analysis, descriptive discriminant analysis*) a pour objectif de décrire les différences liées aux facteurs étudiés.

Le bouton *Réaliser une analyse canonique* donne accès aux résultats suivants :

Onglet avancé :

Le bouton "*Coefficients des variables canoniques*" produit deux feuilles de résultats, dont la définition de la première variable canonique.

	Coefficients bruts des Variables Canoniques
Variable	Comp_1
X1	-0,241220
X2	-0,254916
Constte	4,776475
V.Propre	1,630138
Prop.Cum	1,000000

Ici, la première (et seule) variable canonique est C1 = -0,24122 X1 - 0,25916 X2 + 4,776475.

Cette variable est (définie à un facteur et une constante près) la combinaison linéaire des variables X1 et X2 qui discrimine le mieux les deux groupes, au sens suivant : pour chaque combinaison linéaire Y des variables X1 et X2, on peut comparer les deux groupes G1 et G2 du point de vue de cette variable à l'aide d'une ANOVA à un facteur et calculer le rapport F correspondant. La variable canonique est alors celle pour laquelle le rapport F est maximum.

Dans le tableau ci-dessous, on a effectué l'ANOVA à un facteur pour différentes combinaisons linéaires : la variable canonique, la combinaison linéaire X1+X2, la combinaison linéaire X1+0,5X2, la combinaison linéaire X1-X2 :

Variable	Analyse de la Variance (Mini-AnaDiscri dans Mini-AnaDiscrim.stw) Effets significatifs marqués à p < ,05000							
	SC Effet	dl Effet	MC Effet	SC Erreur	dl Erreur	MC Erreur	F	p
Canonique	61,95	1	61,95	38,00	38	1,00	61,95	0,00
CombLin 1	1008,88	1	1008,88	620,00	38	16,32	61,83	0,00
CombLin 2	582,97	1	582,97	463,91	38	12,21	47,75	0,00
CombLin 3	1,66	1	1,66	1451,50	38	38,20	0,04	0,84

La combinaison linéaire X1+X2 produit un F proche de celui correspondant à la variable canonique, les deux autres des valeurs de F nettement inférieures. On peut également remarquer que les coefficients de la variable canonique sont choisis de façon que le carré moyen de l'erreur soit 1; la constante est alors choisie de manière que cette variable soit centrée (variable de moyenne nulle).

Le bouton "Test du Chi² avec suppr. des Comp. Successives" fournit des informations relatives au nombre d'axes factoriels à conserver.

Exclusion Compos.	Test du Chi² avec suprr. des Comp. Successives (Mini-AnaDiscri dans Mini-AnaDiscrim.stv					
	Valeur propre	R canoniq.	Lambda de Wilk	Chi²	dl	valeur p
0	1,630138	0,787269	0,380208	35,78034	2	0,000000

Les valeurs propres mesurent la dispersion des valeurs sur chacun des axes factoriels. La colonne "R canoniq." indique les *corrélations canoniques* successives. Leurs carrés sont calculés à partir de quotients du type (Somme des carrés intergroupes)/(Somme des carrés totale).
La première ligne correspond au modèle complet, la seconde ligne au modèle privé de la première composante, etc. Les tests du khi-2 permettent d'évaluer quels sont les rapports de corrélation qui sont significativement différents de 0.

Le bouton "Structure factorielle" affiche les corrélations des variables avec les fonctions discriminantes respectives, une fois neutralisé l'effet lié aux classes. Ces corrélations sont analogues aux poids factoriels dans l'analyse factorielle.

Enfin le dernier bouton donne les moyennes des composantes canoniques sur les différentes classes.

Groupe	Moyenne des Vars Canoniques
	Comp_1
G2	-1,24444
G1	1,24444

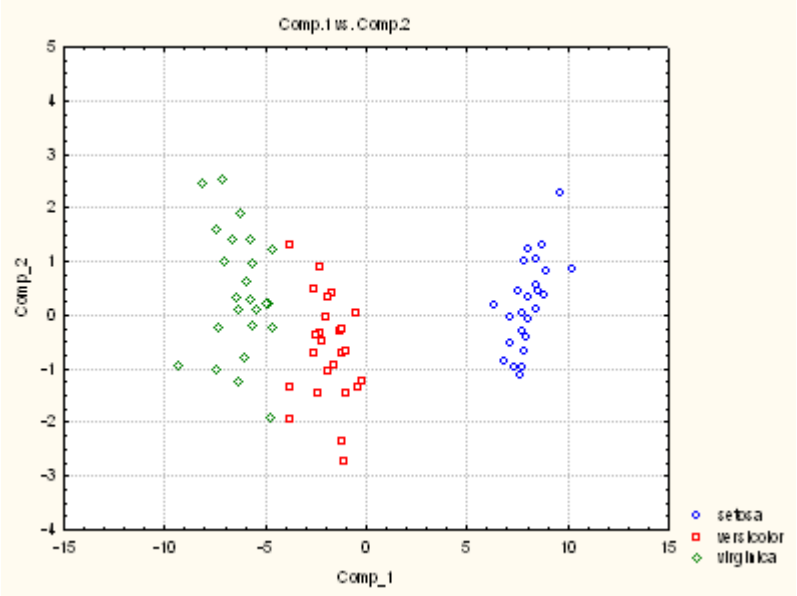
Onglet Scores canoniques

Le bouton "Scores canoniques de chaque observation" donne la valeur de la variable canonique sur chaque observation. On voit ainsi que, sauf exception, les observations classées dans le groupe G2 ont des scores négatifs pendant que celles classées dans le groupe G1 ont des scores positifs.

En cliquant sur le bouton Annuler, on revient aux résultats de l'analyse discriminante proprement dite.

3.3.3 Les iris de Fisher

Ouvrez le classeur Iris.stw. Il s'agit d'un exemple, initialement proposé par Fisher, et utilisé comme données de référence par la plupart des logiciels de statistiques.
On a noté, pour 150 iris, l'espèce (setosa, versicolor, virginica) et 4 variables numériques : la longueur et la largeur des sépales, la longueur et la largeur des pétales. Pour chaque espèce, on dispose de 50 observations. Les 25 premières observations de chaque espèce vont contituer l'ensemble d'apprentissage, tandis que les 25 observations restantes seront classifiées à l'aide des résultats de l'analyse discriminante. La classification ainsi obtenue pourra ainsi être comparée aux données réelles.
Procédez de même à une analyse discriminante sur ces données. Comme nous avons ici 4 variables numériques et 3 groupes, nous aurons deux facteurs discriminants, et Statistica nous permet de construire un graphique représentant les observations selon les valeurs de leurs scores canoniques :



3.3.4 Un exemple d'interprétation des résultats d'une analyse discriminante

Référence : Chrea, C., Valentin, D., Sulmont-Rossé, C., Hoang Nguyen, D., Abdi, H.. Semantic, Typicality and Odor Representation: A Cross-cultural Study, Chemical Senses, 2005, Vol. 30 No 1, pp. 37-49.

Dans l'étude citée en référence, une équipe de chercheurs s'est intéressée à l'effet de la culture sur la catégorisation des odeurs . Dans une première expérience, trois groupes de sujets ont été recrutés : 30 étudiants de l'Université de Bourgogne (Dijon - France), 30 étudiants de l'Université de Dallas (Texas - USA) et 30 étudiants de l'Institut Polytechnique de Danang (Vietnam). Les sujets devaient accomplir une tâche de classification sur 40 odorants fournis par une société spécialisée. Cette première expérience a permis de répartir les 40 odorants en 5 classes (floral, médical, doux, mauvais, naturel) pour la culture française et en 4 classes (floral, doux, mauvais, naturel) pour chacune des deux autres cultures.

Dans une deuxième expérience, on recrute de nouveau trois groupes de sujets correspondant aux trois cultures. Chacun des 40 odorants était présenté aux sujets, qui devaient évaluer la typicalité de son odeur par rapport à 11 catégories (animal, pâtisserie, sucrerie, nettoyant, cosmétique, fleur, fruit, pharmacie, moisissure, nature, épice). Les sujets donnaient leurs réponses sur des échelles de Likert en 7 points (1 = non typique, 7 = tout à fait typique).

On considère alors, pour chaque groupe et chaque odorant, la moyenne du score de typicalité observée sur l'ensemble des participants. Pour évaluer si les classes produites par les classifications de la première expérience s'organisent en fonction de la typicalité, les auteurs ont réalisé une série

d'analyses discriminantes. Ces analyses utilisent les 11 scores de typicalité pour prédire l'affectation des 40 odorants dans les classes définies par la première expérience.

On se propose d'utiliser Statistica pour retrouver les résultats indiqués par les auteurs.

Chargez le classeur Statistica Odors-AnaDiscrim.stw.

Dans ce classeur, les trois feuilles Odors-FR, Odors-US et Odors-VN donnent les scores de typicalité moyen pour chaque culture et chaque catégorie. La dernière variable de chaque feuille (variable "Classe") indique la classe de référence déterminée par la première expérience. La feuille Odors-Moyennes pourra être utilisée pour faciliter la correspondance entre les dénominations en français et en anglais.

Pour chacune des 3 feuilles réalisez une analyse discriminante permettant de retrouver les résultats suivants :

- Utilisez le bouton "Synthèse - Variables dans le modèle" pour vérifier que :

> *L'analyse discriminante produit des résultats significatifs pour chacune des trois cultures [F(44,77)=5,63, P<0,0001 pour la France, F(33,77)=9,10, P<0,0001 pour les USA, F(33,77)=3,28, P<0,0001 pour le Vietnam)].*

- Utilisez le bouton "Réaliser une analyse canonique" puis les boutons "Synthèse: Test du Chi^2, composantes successives" et "Coefficients des variables canoniques" pour vérifier que :

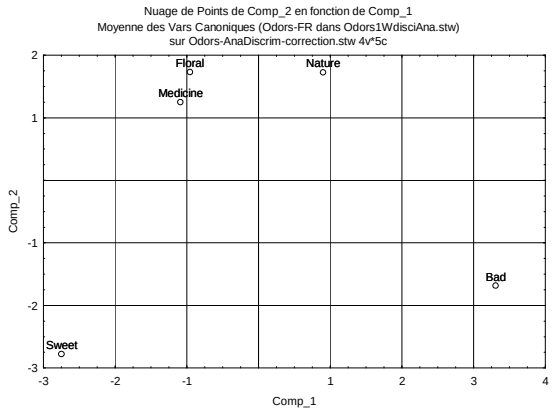
> *Trois fonctions discriminantes pour la France et les USA et deux fonctions discriminantes pour le Vietnam maximisent la discrimination des 40 odorants. Ces fonctions discriminantes linéaires prises ensemble représentent 97% de la variance pour la France, 99% pour les USA et 91% pour le Vietnam. Ainsi, dans chacune des cultures, les scores de typicalité permettent de prédire l'affectation des odorants dans les classes.*

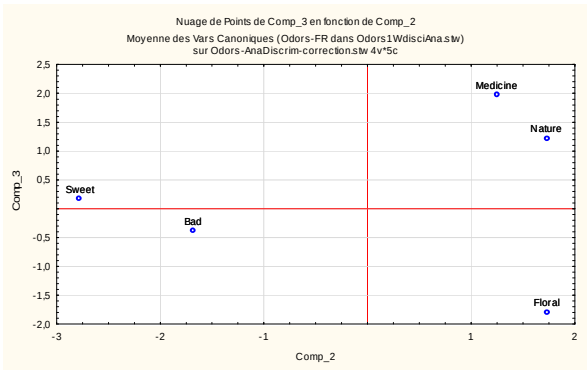
- Utilisez le bouton "Structure factorielle" pour déterminer les corrélations entre les variables de typicalité et les fonctions linéaires discriminantes et vérifier que :

> *C'est le score de typicalité "candy" qui a le poids le plus fort dans la première fonction, pour toutes les cultures. Cela signifie que la première fonction sépare les classes essentiellement selon le gradient de la typicalité "candy". Pour la seconde fonction, "musty" et "animal" ont un poids fort pour les USA et le Vietnam, pendant que "cleaner", "candy" et "fruit" ont les poids les plus forts pour la France. Enfin, "cosmetic" a un poids élevé dans la troisième fonction pour la France et les USA.*

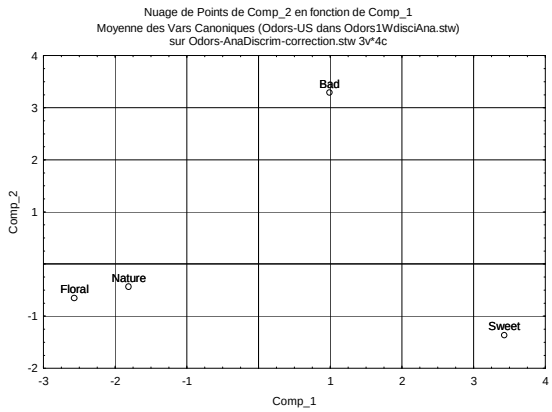
- Enfin, utilisez le bouton "Moyennes des variables canoniques" puis réalisez des graphiques de type "nuage de points" basés sur les feuilles de résultats produites pour réaliser les graphiques suivants :

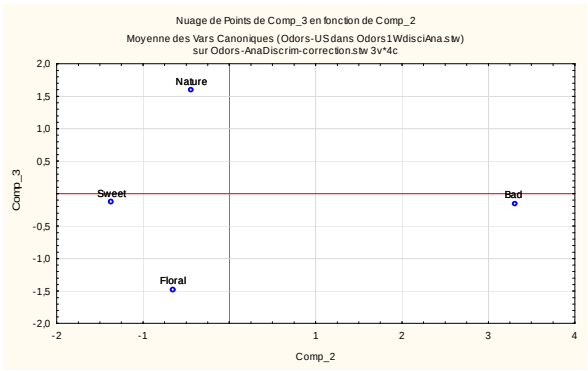
France :



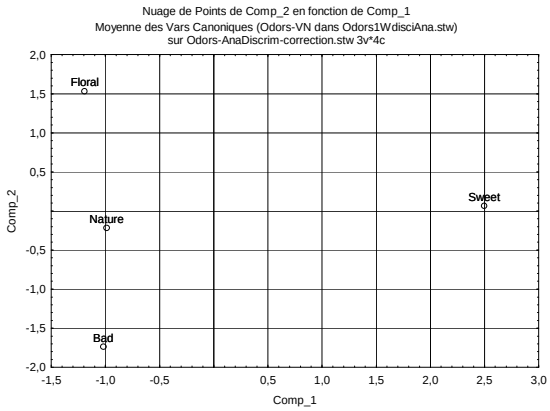


USA :





Vietnam :



La discrimination entre les classes n'est pas identique dans les trois cultures. Par exemple, la première dimension oppose les classes "bad" et "nature" aux classes "sweet", "medecine" et "floral" en France alors qu'elle oppose la classe "sweet" aux classes "floral" et "nature" pour les USA et le Vietnam.

Nous avons trouvé que certains odorants ont été évalués comme plus typiques d'une catégorie donnée que d'autres. Ces résultats suggèrent que les odeurs à l'intérieur d'une catégorie ne sont pas équivalentes. De plus, les résultats de l'analyse discriminante ont montré que les odorants étaient discriminés dans les trois cultures par quatre scores de typicalité (candy, animal, musty et cosmetic).(...) Nos résultats révèlent qu'une certaine variabilité dans les scores de typicalité a aussi contribué à des différences culturelles dans la discrimination des classes. Cette variabilité peut être due à des différences culturelles dans la consommation en ce qui concerne la nourriture et les produits cosmétiques et aussi à des différences culturelles dans la familiarité avec certaines odeurs spécifiques.

3.4 Analyse et régression PLS

3.4.1 Position du problème

On a observé sur un échantillon de n individus statistiques :
- d'une part, p variables indépendantes ou explicatives : X_1, X_2, ..., X_p
- d'autre part, q variables dépendantes, ou "à expliquer" : Y_1, Y_2, ..., Y_q.

On souhaite établir entre les variables indépendantes et les variables explicatives une relation linéaire du type :

$$Y_1 = b_{10} + b_{11}X_1 + ... + b_{1p}X_p + \varepsilon_1$$
$$Y_2 = b_{20} + b_{21}X_1 + ... + b_{2p}X_p + \varepsilon_2$$
$$...$$
$$Y_q = b_{q0} + b_{q1}X_1 + ... + b_{qp}X_p + \varepsilon_q$$

On dispose déjà d'un outil s'appliquant à ce type de problème : la régression linéaire multiple. Cependant, la régression linéaire classique présente les inconvénients suivants :
- Elle "met en compétition" les différentes variables X_i, et elle est très sensible aux collinéarités entre les Xi, et même inutilisable si l'une des variables X_i est combinaison linéaire des autres variables.
- Elle ne peut pas être utilisée si le nombre d'observations (n) est inférieur au nombre de prédicteurs (p).

Une façon de contourner ces problèmes consiste à faire d'abord une ACP sur les prédicteurs, puis de réaliser la régression des variables dépendantes sur les variables principales ainsi définies. Mais le résultat n'est pas facilement interprétable par l'utilisateur.

L'idée de la régression PLS est de procéder de façon analogue à la régression sur composantes principales, mais en formant des composantes ou *variables latentes* tenant compte des variables à expliquer.

3.4.2 Le principe de la régression PLS sur un mini-exemple

Considérons les données suivantes (1 variable dépendante Y, 4 variables explicatives X_j, 3 sujets observés) :

	Y	X_1	X_2	X_3	X_4
s1	12	8	2	7	6
s2	10	2	12	5	7
s3	5	15	6	5	5

Afin d'éliminer les effets dus aux unités avec lesquelles sont mesurés les X_j, on introduit les variables Z_j, variables centrées réduites associées aux X_j.

Ainsi, les variables Z_j sont ici données par :

Y	Z_1	Z_2	Z_3	Z_4
0,8321	-0,0512	-0,9272	1,1547	0,0000
0,2774	-0,9734	1,0596	-0,5774	1,0000
-1,1094	1,0246	-0,1325	-0,5774	-1,0000

La première composante, ou variable latente P_1 est obtenue en pondérant les Z_j proportionnellement aux coefficients de corrélation w_j=r(Y, X_j).

Sur notre exemple, les coefficients de corrélation valent :

	Y
X_1	-0,7247
X_2	-0,1653
X_3	0,7206
X_4	0,6934

On divise ces coefficients par un même nombre, de manière que la somme des carrés des poids soit égale à 1. On obtient ainsi les poids suivants :

$$w_1=-0,582 ; w_2 = -0,133 ; w_3 = 0,578 ; w_4 = 0,556$$

La variable latente P1 a donc pour valeur :

$$P_1 = - 0,582 * Z_1 - 0,133 * Z_2 + 0,578 * Z_3 + 0,556 * Z_4.$$

Sur les 3 observations, elle prend les valeurs suivantes :

	P_1
s1	0,8206
s2	0,6481
s3	-1,4687

La régression de Y par rapport à P_1 conduit à l'équation :

$$Y = 2,7640 P_1 + 9$$

et les valeurs estimées de Y sont :

	Y	Y estimé	Résidus
s1	12	11,2682	0,7318
s2	10	10,7915	-0,7915
s3	5	4,9404	0,0596

D'où un coefficient de détermination :

$$R^2(Y, Y \text{ estimé}) = 0,955$$

Il serait ensuite possible de recommencer la même méthode à partir des résidus de Y, pour produire une deuxième variable latente, et améliorer la qualité de l'estimation.

3.4.3 Un exemple de régression PLS avec Statistica

Dans l'ouvrage : M. Lewis-Beck, A. Bryman, T. Futing (Eds): Encyclopedia for research methods for the social sciences. Thousand Oaks (CA): Sage. pp. 792-795, Hervé Abdi donne l'exemple suivant, que l'on trouve également sur son site, à partir de la page

http://www.utdallas.edu/~herve/#Articles.

On veut prévoir l'évaluation subjective d'un ensemble de 5 vins. Les variables dépendantes que nous voulons prédire sont son appréciation générale et la façon dont il s'accorde avec la viande et les desserts. Les prédicteurs sont le prix, le taux de sucre, le taux d'alcool, et l'acidité.

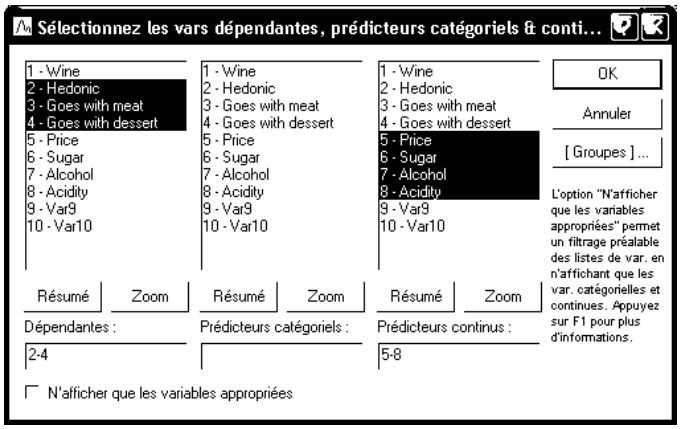
Les données sont les suivantes :

	1 Wine	2 Hedonic	3 Goes with meat	4 Goes with dessert	5 Price	6 Sugar	7 Alcohol	8 Acidity
1	1	14	7	8	7	7	13	7
2	2	10	7	6	4	3	14	7
3	3	8	5	5	10	5	12	5
4	4	2	4	7	16	7	11	3
5	5	6	2	4	13	3	10	3

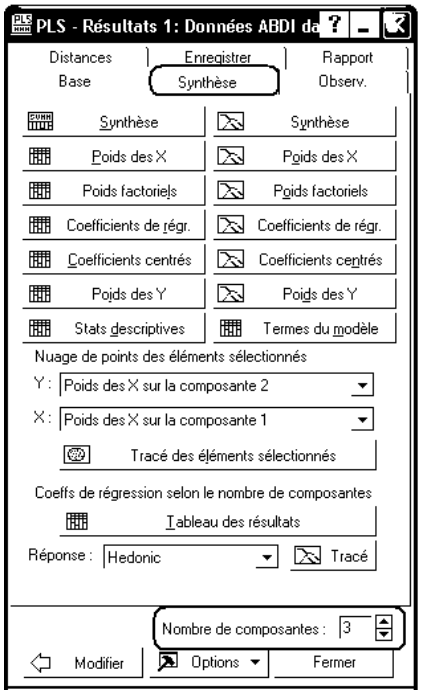
Ouvrez le fichier PLS-Abdi.stw.

La régression PLS est accessible à partir du menu : Statistiques - Modèles linéaires/non-linéaires avancés - Modèles généraux PLS - Modèles linéaires généraux.

Sélectionnez les variables comme suit :



La fenêtre de dialogue "Résultats" permet d'indiquer le nombre de variables latentes souhaité et comporte différents onglets :



L'onglet "Base" est entièrement repris dans l'onglet "Synthèse".
Le bouton "Synthèse" produit le résultat suivant :

	Synthèse de la PLS (Données ABDI dans PLS-Abdi.stw) Réponses : Hedonic Goes with meat Goes with dessert Options : NO-INTERCEPT AUTOSCALE						
	Augmente R^2 de Y	Moyenne R^2 de Y	Augmente R^2 de X	Moyenne R^2 de X	R^2 de Hedonic	R^2 de Goes with meat	R^2 de Goes with dessert
Comp 1	0,6333	0,6333	0,7045	0,7045	0,7053	0,9374	0,2572
Comp 2	0,2206	0,8540	0,2790	0,9835	0,7071	0,9851	0,8697
Comp 3	0,1044	0,9583	0,0165	1,0000	1,0000	1,0000	0,8750

Ce tableau nous donne le pourcentage de variance de chacune des variables dépendantes expliqué, pris en compte par le modèle, en séparant l'apport de chacune des composantes (colonnes R2 de Hedonic, R2 de Goes with meat, R2 de Goes with dessert). Il donne également le pourcentage global pour l'ensemble des 3 variables dépendantes (R2 de Y), obtenu simplement comme moyenne des 3 pourcentages précédents. Il indique également le pourcentage de variance des prédicteurs pris en compte par les composantes.

Le bouton "Poids des X" conduit au tableau suivant, qui donne l'expression des composantes en fonction des prédicteurs :

Poids des prédicteurs (Données ABDI dans PLS-Abdi.stw
Réponses : Hedonic Goes with meat Goes with dessert
Options : NO-INTERCEPT AUTOSCALE

	Price	Sugar	Alcohol	Acidity
Compo 1	-0,5137	0,2010	0,5705	0,6085
Compo 2	0,2343	0,9611	0,1267	0,0734
Compo 3	-0,3747	0,1291	-0,8069	0,4380

Ainsi, on a, sur les données centrées réduites :

Compo 1 = -0,51 * Price + 0,20 * Sugar + 0,57 * Alcohol + 0,61 * Acidity

Le bouton "Poids Factoriels" donne l'expression des prédicteurs en fonction des composantes :

Pds Fac. X (Données ABDI dans PLS-Abdi.stw)
Réponses : Hedonic Goes with meat Goes with desser
Options : NO-INTERCEPT AUTOSCALE

	Price	Sugar	Alcohol	Acidity
Comp 1	-0,5678	0,0142	0,5933	0,6032
Comp 2	0,3302	0,9638	-0,0136	-0,0268
Comp 3	-0,3496	0,1613	-0,8220	0,4222

Ainsi, en données centrées réduites :

Price = -0,57 * Compo1 + 0,33 * Compo2 - 0,350 * Compo3

Les boutons Coefficients de régression et Coefficients de régression centrés donnent les résultats de la régression (utilisant le modèle PLS). Les variables dépendantes estimées y sont exprimées en fonction des variables de départ.

Coefficient de régression PLS (Données ABDI dans PLS-Abdi.stw
Réponses : Hedonic Goes with meat Goes with dessert
Options : NO-INTERCEPT AUTOSCALE

	Ord.Ori	Price	Sugar	Alcohol	Acidity
Hedonic	48,5000	-1,0000	0,7500	-4,0000	2,7500
Goes with meat	-8,9167	-0,0333	0,2750	1,0000	0,1750
Goes with dessert	-3,8542	0,0417	0,5937	0,5000	0,0937

PLS coefficients de régression centrés (Données ABDI dans PLS-Abdi.stw
Réponses : Hedonic Goes with meat Goes with dessert
Options : NO-INTERCEPT AUTOSCALE

	Price	Sugar	Alcohol	Acidity
Hedonic	-1,0607	0,3354	-1,4142	1,2298
Goes with meat	-0,0745	0,2593	0,7454	0,1650
Goes with dessert	0,1250	0,7510	0,5000	0,1186

Ainsi, par exemple, en données non centrées réduites, on a :

Hedonic estimé = 48,5 - Price + 0,75 * Sugar - 4 * Alcohol + 2,75 * Acidity

(et il s'agit d'une valeur exacte, puisque R2=1 pour cette variable).

Le bouton "Poids des Y" donne l'expression des variables dépendantes (centrées réduites) en fonction des composantes :

Poids des réponses (Données ABDI dans PLS-Abdi.stw
Réponses : Hedonic Goes with meat Goes with dessert
Options : NO-INTERCEPT AUTOSCALE

	Hedonic	Goes with meat	Goes with dessert
Comp 1	0,6093	0,7024	0,3680
Comp 2	-0,0518	0,2684	0,9619
Comp 3	0,9672	-0,2181	-0,1301

L'onglet "Observ." donne quant à lui des tableaux des valeurs observées, valeurs prévues et résidus des variables dépendantes sur les différents individus statistiques observés :

Valeurs prévues (Données ABDI dans PLS-Abdi.stw)
Réponses : Hedonic Goes with meat Goes with desser
Options : NO-INTERCEPT AUTOSCALE

	Hedonic	Goes with meat	Goes with dessert
1	14,0000	7,0000	7,7500
2	10,0000	7,0000	5,7500
3	8,0000	5,0000	6,0000
4	2,0000	4,0000	6,7500
5	6,0000	2,0000	3,7500

Il donne également les scores des individus sur les composantes, calculés soit à partir des variables prédictives, soit à partir des variables dépendantes :

Valeurs des prédicteurs et réponses (Données ABDI dans PLS-Abdi.stw
Réponses : Hedonic Goes with meat Goes with dessert
Options : NO-INTERCEPT AUTOSCALE

	Comp. X 1	Comp. X 2	Comp. X 3	Comp. Y 1	Comp. Y 2	Comp. Y 3
1	1,4952	0,9663	0,2937	1,9451	0,7611	0,6191
2	1,7789	-1,0239	-0,2380	0,9347	-0,5305	-0,5388
3	0,0000	0,0000	0,0000	-0,2327	-0,6084	0,0823
4	-1,4181	1,1040	-0,2724	-0,9158	1,1575	-0,6139
5	-1,8560	-1,0464	0,2167	-1,7313	-0,7797	0,4513

3.5 Analyse de segmentation

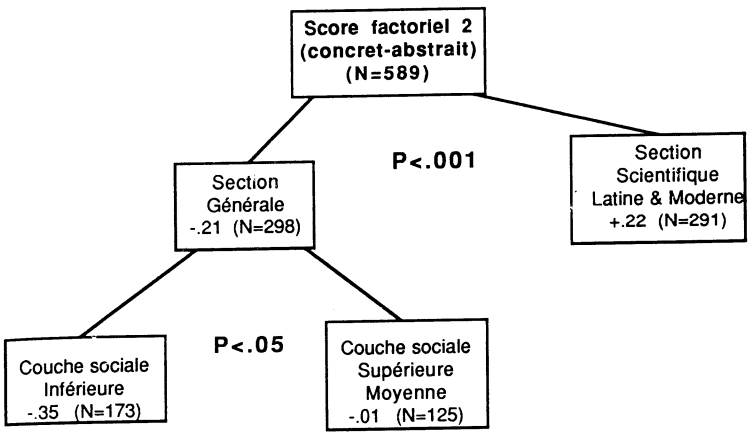
Bibliographie:
Lebart, L., Morineau, A., Piron M., Analyse exploratoire multidimensionnelle, Dunod, Paris, 2000.
Doise, W., Clémence A., Lorenzi-Cioldi, F., Représentations sociales et analyses de données, Presses Universitaires de Grenoble, Grenoble, 1992
Idams : http://ead.univ-angers.fr/~statidams/

3.5.1 But de la méthode

On a observé sur un échantillon d'individus statistiques une variable dépendante numérique ou qualitative Y et plusieurs variables numériques ou catégorielles X1, X2, ..., Xp.

La segmentation vise à expliquer la variable Y à l'aide d'une ou plusieurs variables quantitatives ou qualitatives. Elle permet également de créer des groupes d'individus ou d'observations homogènes.

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



3.5.2 Rappel : décomposition de l'inertie

Du point de vue de la variable dépendante, l'inertie totale est la somme des carrés des écarts à la moyenne générale :

$$I = \sum_{i=1}^{n} \left(y_i - \bar{y}\right)^2$$

On suppose les observations réparties en g groupes (j=1, 2, ..., g). Pour chacun des groupes, on a une moyenne du groupe : $\bar{y}_j$, un effectif nj et une inertie intra-groupe : $I_j = \sum_{i'=1}^{n_j} \left(y_{i'} - \bar{y}_j\right)^2$.

Une relation fondamentale est donnée par le théorème de Huygens : l'inertie totale est la somme des inerties intra-groupes et de l'inertie des points moyens des groupes, pondérés par l'effectif des groupes.

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

$$\underset{\text{totale}}{\text{Inertie}} = \sum_{\text{les groupes}} \underset{\text{dans}}{\text{Inertie}} + \underset{\text{pondérés par}}{\text{Inertie des points moyens}}$$
$$\text{les effectifs des groupes}$$

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

Groupe	A	B	A	B
Y	1	2	3	4

La moyenne générale est donnée par : $\bar{y} = 2{,}5$. L'inertie totale vaut :

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

Les inerties des deux groupes A et B (inerties intra-groupes) sont données par :

$$I_A = (1-2)^2 + (3-2)^2 = 2 \qquad I_B = (2-3)^2 + (4-3)^2 = 2$$

L'inertie des points moyens pondérés, ou inertie inter-groupes vaut :

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

On vérifie bien que :

$$\text{Inertie totale} = I_A + I_B + I_{Inter} = 2 + 2 + 1 = 5$$

3.5.3 Principe de la méthode :

L'inertie inter-groupes mesure l'hétérogénéité entre les différents groupes alors que l'inertie intra-groupes mesure l'homogénéité à l'intérieur des groupes. Pour obtenir des groupes les plus distincts possibles, il faut une inertie inter-groupes le plus élevée possible. L'inertie intra-groupe sera alors faible et donc les individus d'un même groupe seront homogènes.

1) Au départ, on dispose d'un seul segment contenant l'ensemble des individus.
2) A la première étape, la procédure de construction de l'arbre examine une par une toutes les variables explicatives. Pour chaque variable, elle passe en revue toutes les divisions possibles (de la forme Xj< A et Xj > A si Xj est numérique, regroupement des modalités en deux sous-ensembles si Xj est catégorielle). Pour chaque division, l'inertie inter-groupes est calculée.
3) La division choisie est celle qui maximise l'inertie inter-groupes.
4) On recommence la procédure dans chacun des deux groupes ainsi définis.

Critères d'arrêt :

On peut utiliser comme critères d'arrêt de l'algorithme de segmentation :
- La taille des groupes (classes) à découper
- Le rapport entre l'inertie intra et la variance totale
- Des tests statistiques (tests de Student de comparaison de moyennes, tests du Khi deux)

3.5.4 Exemple d'analyse de segmentation

Source : http://lib.stat.cmu.edu/datasets/

Determinants of Wages from the 1985 Current Population Survey

Summary:

The Current Population Survey (CPS) is used to supplement census information between census years. These data consist of a random sample of 534 persons from the CPS, with information on wages and other characteristics of the workers, including sex, number of years of education, years of work experience, occupational status, region of residence and union membership. We wish to determine (i) whether wages are related to these characteristics and (ii) whether there is a gender gap in wages.

Based on residual plots, wages were log-transformed to stabilize the variance. Age and work experience were almost perfectly correlated (r=.98). Multiple regression of log wages against sex, age, years of education, work experience, union membership, southern residence, and occupational status showed that these covariates were related to wages (pooled F test, $p < .0001$). The effect of age was not significant after controlling for experience. Standardized residual plots showed no patterns, except for one large outlier with lower wages than expected. This was a male, with 22 years of experience and 12 years of education, in a management position, who lived in the north and was not a union member. Removing this person from the analysis did not substantially change the results, so that the final model included the entire sample.

Adjusting for all other variables in the model, females earned 81% (75%, 88%) the wages of males ($p < .0001$). Wages increased 41% (28%, 56%) for every 5 additional years of education ($p < .0001$). They increased by 11% (7%, 14%) for every additional 10 years of experience ($p < .0001$). Union members were paid 23% (12%, 36%) more than non-union members ($p < .0001$). Northerns were paid 11% (2%, 20%) more than southerns ($p =.016$). Management and professional positions were paid most, and service and clerical positions were paid least (pooled F-test, $p < .0001$). Overall variance explained was $R2 = .35$.

In summary, many factors describe the variations in wages: occupational status, years of experience, years of education, sex, union membership and region of residence. However, despite adjustment for all factors that were available, there still appeared to be a gender gap in wages. There is no readily available explanation for this gender gap.

Authorization: Public Domain

Reference: Berndt, ER. The Practice of Econometrics. 1991. NY: Addison-Wesley.

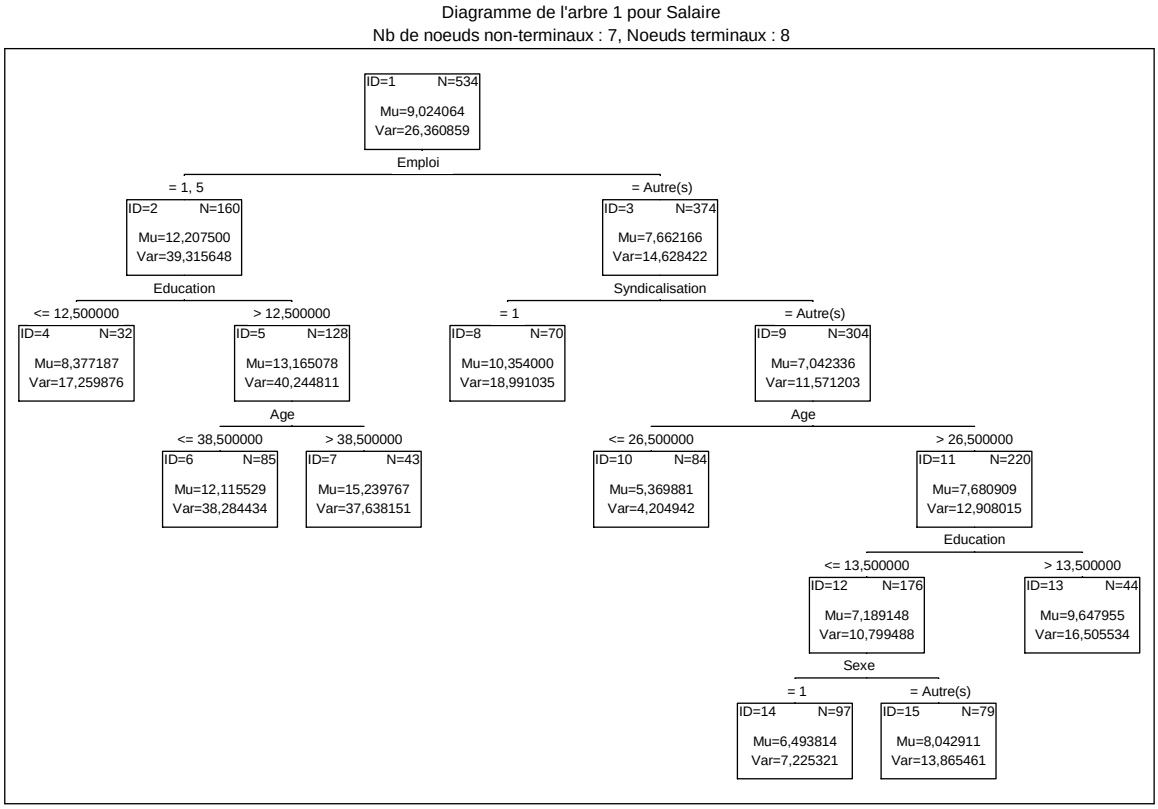
Description: The datafile contains 534 observations on 11 variables sampled from the Current Population Survey of 1985. This data set demonstrates multiple regression, confounding, transformations, multicollinearity, categorical variables, ANOVA, pooled tests of significance, interactions and model building strategies.

Variable names in order from left to right:
EDUCATION: Number of years of education.
SOUTH: Indicator variable for Southern Region (1=Person lives in South, 0=Person lives elsewhere).
SEX: Indicator variable for sex (1=Female, 0=Male).
EXPERIENCE: Number of years of work experience.
UNION: Indicator variable for union membership (1=Union member, 0=Not union member).
WAGE: Wage (dollars per hour).
AGE: Age (years).
RACE: Race (1=Other, 2=Hispanic, 3=White).
OCCUPATION: Occupational category (1=Management, 2=Sales, 3=Clerical, 4=Service, 5=Professional, 6=Other).
SECTOR: Sector (0=Other, 1=Manufacturing, 2=Construction).
MARR: Marital Status (0=Unmarried, 1=Married)

Ces données (avec des noms de variables francisés) se trouvent dans la feuille de données du classeur CPS_85_Wages.stw.

La variable à expliquer est évidemment la variable WAGE.
On constate que les variables SOUTH, SEX, UNION, RACE, OCCUPATION, SECTOR, MARR sont des variables catégorielles, alors que les variables EDUCATION, EXPERIENCE, AGE peuvent être considérées comme numériques.

On présente ci-dessous l'arbre obtenu en indiquant "Salaire" comme variable dépendante, Education, Expérience et Age comme variables numériques, Localisation, Sexe, Syndicalisation, Origine Ethnique, Emploi, Secteur et Situation de famille comme variables catégorielles et en adoptant la règle d'arrêt : on ne segmente pas les groupes d'effectifs inférieurs à 100 :

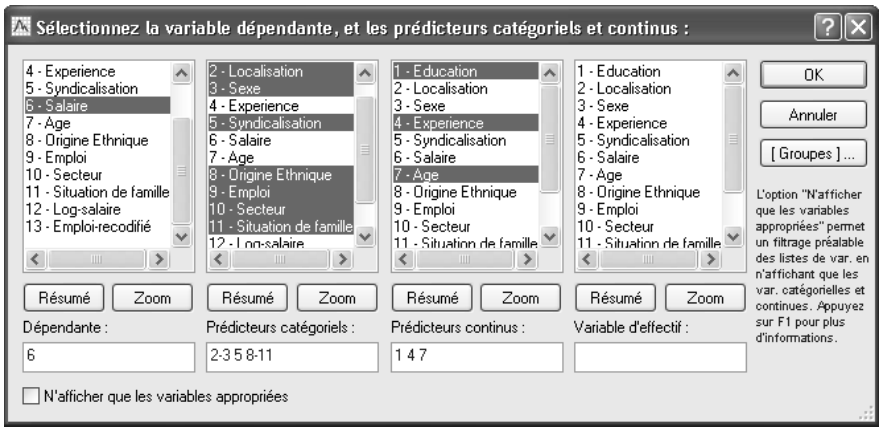


3.5.5 Traitements sous Statistica

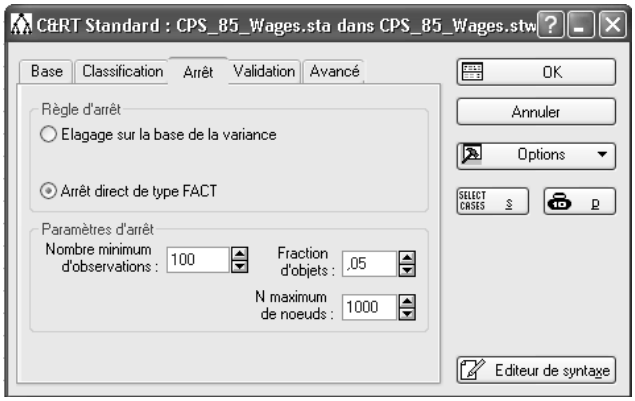
Ouvrir le classeur CPS_85_Wages.stw.

Utiliser le menu Statistiques - Data Mining - Modèles d'arbres de classification et de régression - C&RT Standard.
N.B. "C&RT" signifie : classification and regression trees. L'aide de Statistica indique : "Modèles d'Arbres de Classification et de Régression (GC&RT)",

Sous l'onglet "Standard", indiquer "Salaire" comme variable dépendante, Education, Expérience et Age comme variables numériques (prédicteurs continus), Localisation, Sexe, Syndicalisation, Origine Ethnique, Emploi, Secteur et Situation de famille comme variables catégorielles (prédicteurs catégoriels) :



Sous l'onglet "Arrêt", indiquer 100 comme nombre minimum d'observations.



Cliquer sur le bouton OK puis cliquer sur le bouton "Diagramme de l'arbre" du dialogue "Résultat". Vous devriez obtenir le graphique représenté ci-dessus.

3.5.6 Retrouver avec Statistica les résultats indiqués dans la présentation de l'exemple

Observez la colonne 12 : les logarithmes des salaires y ont été calculés.
Observez également la colonne 13 : la variable Emploi y a été recodifiée afin que les codes numériques des types d'emploi soient classés dans le même ordre que les salaires moyens des différents types.

Calculez le coefficient de corrélation entre Age et Expérience : on trouve effectivement r=0,977.

Faites la première régression indiquée : Log-Salaire (v12) est la variable dépendante, Sexe, Age, Education, Expérience, Syndicalisation, Localisation et Emploi-recodifié (v1 à v5, v7 et v13) sont les prédicteurs. On trouve effectivement F(7, 526)=39,56 et donc un effet significatif de l'ensemble de ces variables.

Calcul de l'effet de l'âge après contrôle de l'expérience : on peut, par exemple, calculer les résidus de la régression de Log-Salaire par rapport à Expérience, puis coller ces résidus dans une colonne supplémentaire de la feuille de données et tester le résultat de la régression de ces résidus par rapport à Age. Le coefficient b* vaut alors 0,09, significatif à 3% seulement.

La valeur atypique citée dans le texte pourra être mise en évidence sur l'un des graphiques de la première régression multiple (par exemple en demandant le tracé des bandes de prévision). Il s'agit de l'observation n° 200.

Les indications de l'avant-dernier paragraphe pourront être retrouvées en effectuant une nouvelle régression linéaire multiple en prenant Log-Salaire (v13) comme variable dépendante, Education, Localisation, Sexe, Expérience, Syndicalisation, Emploi-recodifié (v1 à v5 et v13) comme variables prédictrices.

On trouve bien un coefficient de détermination voisin de 0,35. On remarque que les coefficients des variables prédictrices sont alors tous significativement différents de 0. Pour retrouver les pourcentages indiqués, il faut reconvertir les effets (linéaires, additifs) des coefficients bi sur Log-Salaire en effets multiplicatifs sur Salaire (qui est l'exponentielle de Log-Salaire).

Par exemple, le coefficient b_i pour Expérience est b_4 = 0,011. Pour 10 ans d'expérience supplémentaires (toutes choses égales par ailleurs), Log-Salaire augmente de 0,1103. L'effet multiplicatif sur le salaire est obtenu en calculant l'exponentielle de cette valeur : exp(0,1103)=1,1166. L'effet se traduit donc par une augmentation du salaire de 11,66%.

Pour l'ensemble des prédicteurs, les calculs menés sous Excel conduisent au tableau suivant :

	b*	Err-Type	b	Err-Type	t(527)	valeur p	b* Coef	% Variation
OrdOrig.			0,6457	0,1193	5,4100	0,0000%		
Education	0,3788	0,0409	0,0764	0,0083	9,2611	0,0000%	0,3822	46,55%
Localisation	-0,0923	0,0358	-0,1070	0,0415	-2,5751	1,0294%	-0,1070	-10,15%
Sexe	-0,1791	0,0365	-0,1895	0,0387	-4,9009	0,0001%	-0,1895	-17,26%
Experience	0,2587	0,0382	0,0110	0,0016	6,7643	0,0000%	0,1103	11,66%
Syndicalisation	0,1407	0,0362	0,1932	0,0497	3,8904	0,0113%	0,1932	21,32%
Emploi-recodifié	0,2288	0,0386	0,0870	0,0147	5,9227	0,0000%		

Enfin, concernant l'effet du sexe, on pourra faire une troisième régression multiple en prenant l'ensemble des prédicteurs de la régression précédente, à l'exception du sexe. On calcule les résidus de cette régression et on les place dans une nouvelle colonne de la feuille de données. Enfin, on réalise une régression linéaire en utilisant ces résidus comme variable dépendante et le sexe comme prédicteur. On obtient alors (en ajoutant le traitement sous Excel) :

	b*	Err-Type	b	Err-Type	t(532)	valeur p	b* Coef	% Variation
OrdOrig.			0,0810	0,0252	3,2135	0,14%		
Sexe	-0,2015	0,0425	-0,1765	0,0372	-4,7442	0,00%	-0,1765	-16,18%

Il est assez remarquable que cet écart entre les sexes se retrouve, avec des valeurs presque identiques, dans l'ensemble des régressions qui ont été faites.

Remarque. Les résultats ainsi obtenus sont proches de ceux annoncés dans la présentation de l'exemple, sans être strictement identiques. Les différences proviennent probablement de la façon de traiter les variables catégorisées, notamment la variable Emploi.

4 Bibliographie :

Lebart, L., Morineau, A., Piron M., Analyse exploratoire multidimensionnelle, Dunod, Paris, 2000.
Doise, W., Clémence A., Lorenzi-Cioldi, F., Représentations sociales et analyses de données, Presses Universitaires de Grenoble, Grenoble, 1992
Escofier, B., Pagès J., Analyses factorielles simples et multiples, Dunod, Paris, 1998.
Bry, X., Analyses factorielles simples, Economica, Paris, 1995.
Bry, X., Analyses factorielles multiples, Economica, Paris, 1996.
Morineau A., Morin S., Pratique du traitement des enquêtes - Exemple d'utilisation du système SPAD, Cisia-Ceresta, Montreuil, 2000
Croutsche, J.-J., Pratiques statistiques en gestion et études de marchés, Editions ESKA, Paris, 1997
Howell, D.C., Méthodes Statistiques en Sciences Humaines, De Boeck, Paris Bruxelles, 1998.
Saporta, G., Probabilité Analyse des données et statistique. Editions Technip , 1990

Articles :
Flament C., Milland L., Un effet Guttman en ACP, Mathématiques & Sciences humaines (43e année, n° 171, 2005, p. 25-49)
Hahn A., Eirmbter W. H., Jacob R., Le sida : savoir ordinaire et insécurité, traduction française de Herrmann M.
Baron, R. M., Kenny D.A., The moderator-Mediator Variable Distinction in Social Psychological Research: Conceptual, Strategic, and Statistical Considerations, Journal of Personality and Social Psychology, 1986, V. 51 N° 6, pp. 1173-1182.
Costarelli, S., Callà, R.-M.. Self-directed negative affect: The distinct roles of ingroup identification and outgroup derogation, Current research in Social Psychology, Volume 10 No 2, 2004.
Gil-Monte P. R., Ma Peiró J., A study on significant sources of the burnout syndrome in workers at occupational centres for mentally disabled, , Psychology in Spain, 1997, Vol. 2. No 1, 116-123.
Page Web : http://www.psychologyinspain.com/content/full/1997/6bis.htm
Stinglhamber, F., Bentein, K., Vandenberghe, C. Congruence de valeurs et engagement envers l'organisation et le groupe de travail, Psychologie du Travail et des Organisations, Vol. 10, pp. 165-187, 2004.

Sites internet :
Site Eurostat de l'Union Européenne : http://epp.eurostat.ec.europa.eu/portal/
Site d'Hervé Abdi : http://www.utdallas.edu/~herve/#Articles.
Idams : http://ead.univ-angers.fr/~statidams/
Statlib - datasets archive : http://lib.stat.cmu.edu/datasets/
Psychologie Sociale : http://www.psychologie-sociale.org/reps2.php?article=7
Site pour télécharger ce polycopié et les fichiers d'exemples : http://geai.univ-brest.fr/~carpenti/

5 Table des matières