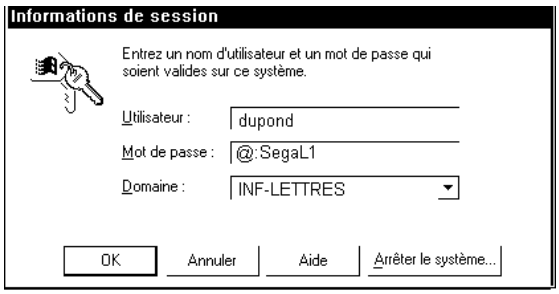
Analyse des données multidimensionnelles

1 Présentation des salles informatiques

Vous disposez d'un "compte" ouvert à votre nom sur le serveur des salles de travaux dirigés. Vous pouvez donc utiliser votre login et le mot de passe **@:SegaL1** pour ouvrir la session. Votre volume personnel sera alors monté comme volume U:.

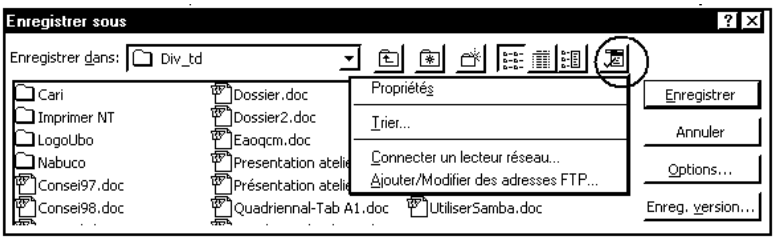


Remarque. Contrairement au compte "etudiant", la configuration de votre compte personnel n'est pas verrouillée, et il vous appartiendra de choisir les paramètres qui vous conviennent pour l'Explorer (affichage ou non des extensions), Word, Excel, etc. Si vous désirez utiliser les logiciels "libres" OpenOffice et Gimp, vous devrez également les "installer".

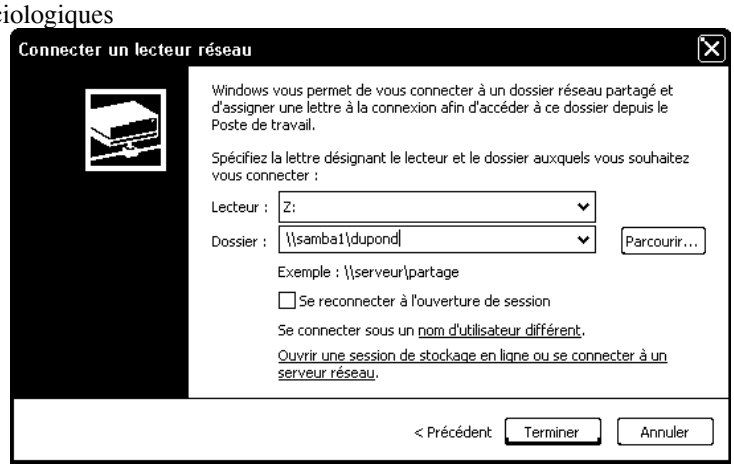
Si vous travaillez en binôme sur un appareil, le deuxième étudiant pourra se connecter à son compte de la manière suivante :

- Utiliser le menu Connecter un lecteur réseau, auquel on a accès d'au moins trois manières différentes :

  1. Par le menu Outils de l'explorateur Windows XP
  2. Par le menu local obtenu en utilisant le bouton droit de la souris sur l'icône Poste de travail
  3. A l'aide de l'icône ci-dessous de la fenêtre Enregistrer sous... de certains logiciels.

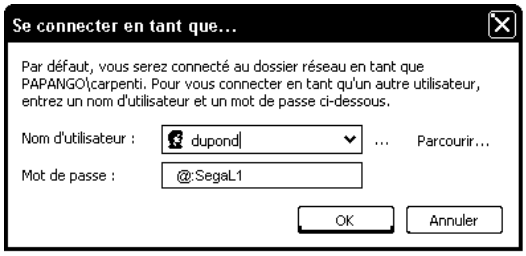


- Compléter les fenêtres de connexion comme suit (vous indiquerez évidemment votre login à la place de "dupond") :



- Désactiver l'item "Se reconnecter à l'ouverture de session"

- Cliquer sur l'item : "Se connecter sous un nom d'utilisateur différent" et compléter le dialogue qui s'affiche comme suit :



Si l'on veut connecter un troisième compte, on peut procéder de la même façon, mais le troisième se verra refuser la connexion au même serveur SAMBA1. Le serveur possède donc un alias, SAMBA2, qu'il faut utiliser pour tromper Windows XP.

Les différents noms letsamba, samba1, samba2 désignent le même serveur physique ; le contenu du répertoire personnel sera donc le même dans chacun des cas.

1.1.1 Changer son mot de passe sur le serveur Letsamba

Vous pouvez si vous le souhaitez, changer votre mot de passe sur le serveur Samba. Vous devez pour cela procéder de la façon suivante :

1.1.1.1 Qu'est-ce qu'un bon mot de passe ?

Un "bon" mot de passe doit être facile à mémoriser pour vous, et difficile à trouver pour les autres, et pour les programmes de piratage... Comme il devra être utilisé avec plusieurs systèmes d'exploitation, possédant des systèmes d'encodage de caractères différents (Macintosh, Windows, Linux, encodage particulier à Internet), il faut se limiter aux caractères dont le code ASCII[1] est inférieur à 127. Il convient, en général, de respecter les règles suivantes :

- de 6 à 8 caractères ;

- Obligatoirement, des majuscules, des minuscules, des chiffres et des caractères non alphabétiques (ponctuation, #, @, $, *, %)

- Evitez les caractères accentués du Français, le ç, et autres caractères bizarres.

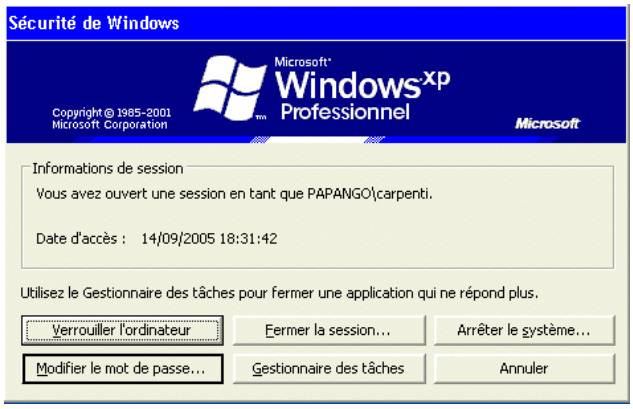
N.B. : Les mots de passe trop simples (11111111 ou zazazazaza par exemple) ne seront pas acceptés.

[1] Revoyez votre cours de 2ème année...

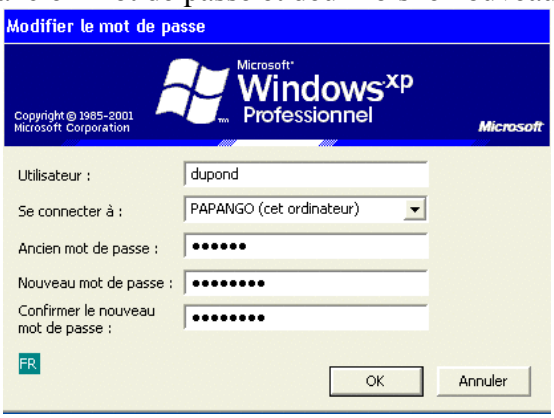
Mémorisez ce mot de passe. Notez-le sur un document personnel. Si vous l'oubliez, l'intervention du technicien sera nécessaire.

1.1.1.2 Changement du mot de passe

- Ouvrez une session sous votre login.
- Appuyez simultanément sur les trois touches Ctrl+Alt+Suppr.
- Dans le dialogue qui s'affiche, choisissez le bouton "modifier le mot de passe":



- Vous devez ensuite taper l'ancien mot de passe et deux fois le nouveau mot de passe :

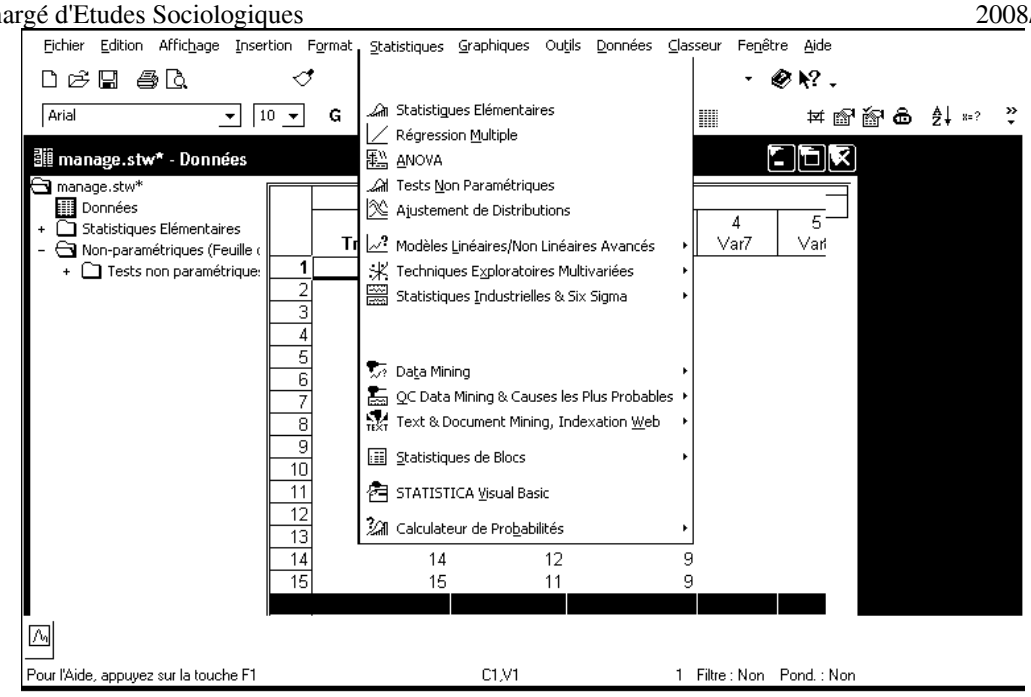


2 Présentation de Statistica

. Statistica : l'interface utilisateur

L'écran de travail

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



Les objets manipulés par Statistica

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

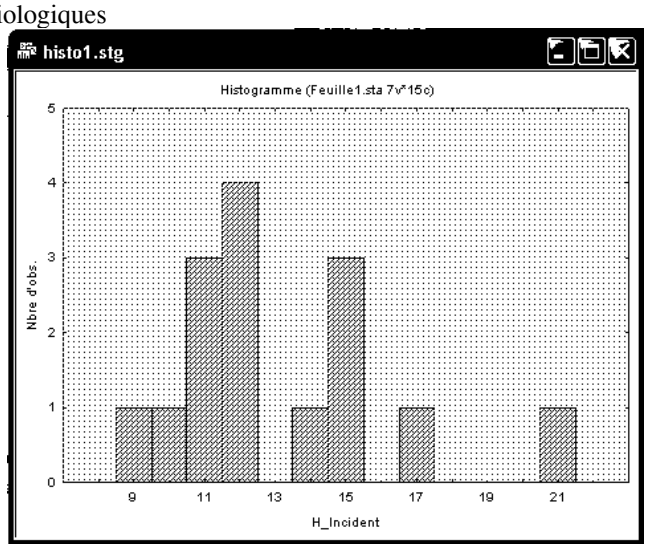
Données : Feuille1.sta (7v par 15c)

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

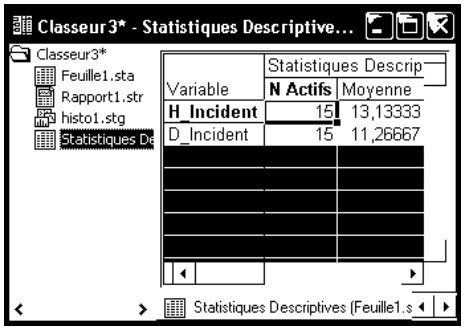
Les feuilles de données peuvent être enregistrées comme fichiers autonomes (fichiers *.sta). Elles contiennent les données d'entrée sur lesquelles s'effectuent les traitements statistiques. Les résultats de ces traitements s'affichent dans un document de sortie. Plusieurs possibilités sont offertes.

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

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



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



Traitements statistiques

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

Gérer les sorties

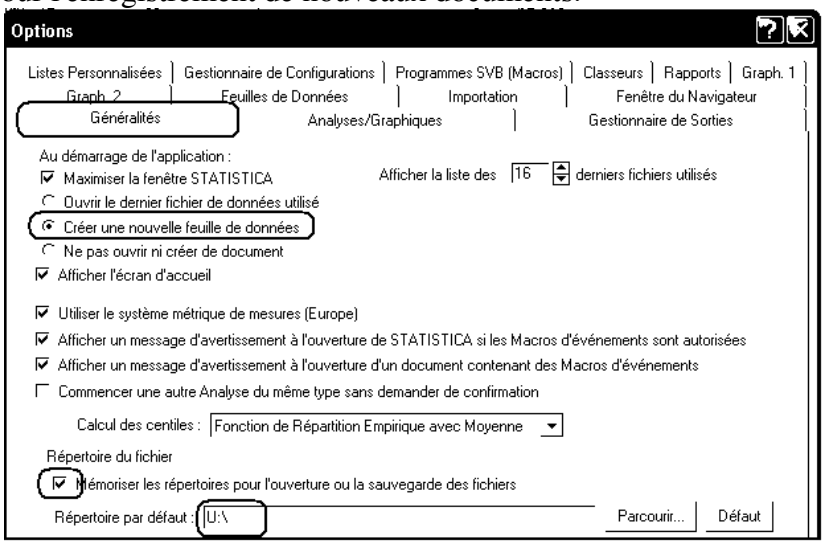
Modifier le comportement de Statistica

Le comportement de Statistica peut être modifié en intervenant dans la fenêtre de dialogue affichée par le menu Outils - Options.

Par exemple, nous souhaitons :

- que Statistica n'ouvre plus systématiquement la dernière feuille de données utilisée lors du chargement du logiciel ;
- que Statistica nous propose par défaut le volume U: pour enregistrer nos documents, au lieu du répertoire "Mes Documents".

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



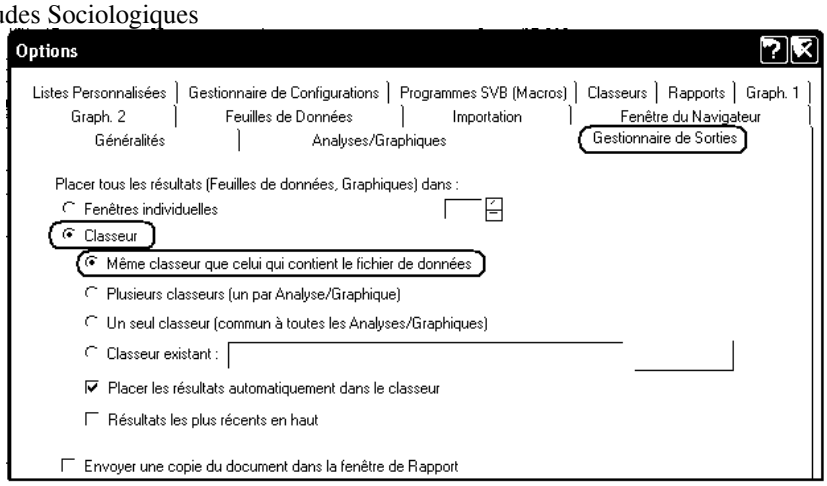
Gérer les sorties

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

Enregistrer données et résultats dans un seul classeur

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

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



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

Indiquer quelle est la feuille de données active

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

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

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

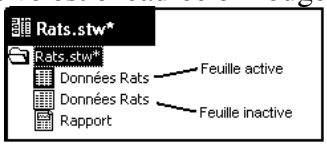
- Lorsque l'on travaille avec une feuille de données insérée dans un classeur, il arrive couramment que Statistica ne retrouve pas la feuille à partir de laquelle les traitements doivent être effectués. Mais on peut éviter ce comportement en spécifiant la propriété "feuille de données active" pour l'objet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

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

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

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

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

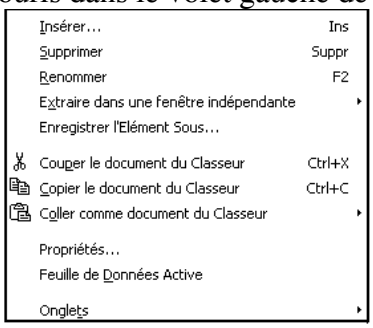


Enregistrer les données et l'ensemble des traitements réalisés dans un même classeur

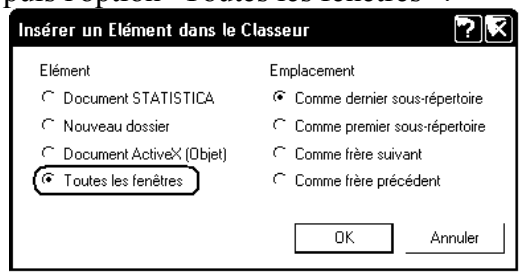
Ouvrez un fichier de données (un fichier d'extension .sta) et réalisez un ou plusieurs traitements relatifs à ces données (par exemple, des statistiques descriptives et un graphique). Si vous avez gardé les options par défaut de Statistica, les résultats de tous ces traitements se trouvent dans un classeur.

Pour enregistrer données, traitements et rapport dans un seul classeur :

Affichez la fenêtre du classeur contenant les résultats.
Cliquez avec le bouton droit de la souris dans le volet gauche de la fenêtre du classeur.



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



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

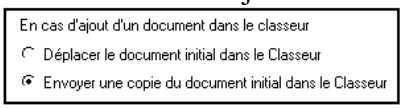
Manipuler les objets contenus dans un classeur

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

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

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

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



Supprimer un objet d'un classeur

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

3 Traitement d'un tableau de données individus x variables : l'analyse en composantes principales ou ACP

3.1 Introduction

On a observé un ensemble de variables numériques sur un ensemble d'individus statistiques. On dispose donc d'un tableau de données individus x variables tel que le suivant :

Ind.	V1	V2	V3
1	20	40	60
2	40	60	80
3	20	40	60
4	40	60	80

Selon [Doise], on peut définir trois notions fondamentales dans l'approche multivariée des différences individuelles: le niveau, la dispersion et la corrélation. Ces trois composantes sont autant de points de vue sur les données.

Le niveau des réponses peut être évalué en calculant la **moyenne** pour chaque variable.

La dispersion (le degré d'éparpillement des réponses individuelles autour de la moyenne) peut être évaluée en calculant la **variance** et **l'écart type** de chacune des variables.

La corrélation (le lien entre les réponses individuelles pour deux variables) est évaluée en calculant le **coefficient de corrélation** entre ces variables.

	CAS											
	1			2			3			4		
Ind.	V1	V2	V3	V1	V2	V3	V1	V2	V3	V1	V2	V3
1	20	40	60	60	40	60	40	35	30	60	35	10
2	40	60	80	40	60	60	60	65	70	80	65	50
3	20	40	60	40	60	40	40	35	30	60	35	10
4	40	60	80	60	40	40	60	65	70	80	65	50
Moy	30	50	70	50	50	50	50	50	50	70	50	30
s	10	10	10	10	10	10	10	15	20	10	15	20
CORRÉLATIONS												
	V1	V2	V3	V1	V2	V3	V1	V2	V3	V1	V2	V3
V1	1			1			1			1		
V2	1	1		-1	1		1	1		1	1	
V3	1	1	1	0	0	1	1	1	1	1	1	1

NB: Moy = moyenne ou niveau s = écart type ou dispersion

Dans le cas numéro (1), comme dans le cas numéro (4), les moyennes des trois variables sont différentes. Pourtant, les corrélations entre ces variables sont identiques. Dans le cas (4), la dispersion varie du simple au double, selon les variables, sans que cela affecte les corrélations. En outre, comme l'indique le cas (3), les dispersions ne sont pas non plus liées aux moyennes. Dans le cas (2) au contraire, il n'y a aucune variation de niveau, ni même de dispersion, alors que les corrélations diffèrent fortement. Arrêtons nous un instant sur les variables V1 et V2 du cas (2). On voit que ces variables comportent les mêmes chiffres compte non tenu de l'ordre successif des individus (des valeurs de 40 et 60) ; en tenant compte de l'ordre successif, ce que fait la corrélation comme mesure des liens entre profils, il apparaît qu'en progressant de l'individu 1 à l'individu 2, la

variable V1 diminue (de 60 à 40), alors que la variable V2 augmente (de 40 à 60). On s'aperçoit ainsi que des valeurs identiques (ou très similaires), mais arrangées différemment, affectent uniquement les corrélations. Ces exemples illustrent la propriété d'indépendance des trois éléments des distributions. Cette propriété nous permet de prédire que la corrélation de 1 dans le cas (2) n'aurait pas été affectée à la suite d'une transformation linéaire quelconque d'une des deux variables (par exemple, rajout à toutes les valeurs d'une variable d'une constante ; une telle transformation ne modifierait que le niveau de cette variable). Cette propriété des distributions sera largement exploitée par l'analyse multidimensionnelle.

L'analyse en composantes principales sert généralement à résumer les variations d'un champ de représentations dans une population donnée. Cette technique comporte un examen fort détaillé des liens entre profils des réponses individuelles dans cette population, au détriment toutefois de celui du niveau et de la dispersion de ces réponses.

3.2 L'ACP sur un mini-exemple

On a observé p variables sur n individus. On dit qu'il s'agit d'un protocole multivarié. On s'intéresse à l'étude de la *variabilité* observée sur l'ensemble des individus ou l'ensemble des variables, avec l'idée suivante :

> *trouver des variables abstraites, en petit nombre, reproduisant de la façon la moins déformée possible la variabilité observée.*

Du point de vue des variables : on cherche à remplacer les p variables par q nouvelles variables résumant au mieux le protocole, avec $q \leq p$ et si possible q=2. Nous verrons que l'ACP permet de résumer un ensemble de variables corrélées en un nombre réduit de variables (appelées facteurs) non corrélées.

Du point de vue des individus : chaque individu est représenté par un point dans un espace de dimension p. On peut calculer les distances (euclidiennes) entre deux individus, entre un individu et le point moyen du nuage, etc. On cherche alors à trouver une projection des individus dans un espace de dimension $q \leq p$, respectant au mieux les distances entre les individus (une "carte", la moins déformée possible).

3.2.1 Mini-exemple

Ci-dessous, un tableau de notes attribuées à 9 sujets dans 5 matières.

Sujet	Math	Sciences	Français	Latin	Musique
Jean	6	6	5	5,5	8
Aline	8	8	8	8	9
Annie	6	7	11	9,5	11
Monique	14,5	14,5	15,5	15	8
Didier	14	14	12	12	10
André	11	10	5,5	7	13
Pierre	5,5	7	14	11,5	10
Brigitte	13	12,5	8,5	9,5	12
Evelyne	9	9,5	12,5	12	18

Données centrées réduites :

En général, les variables retenues pour décrire les individus sont exprimées avec des unités différentes, et ne sont pas directement comparables entre elles. Dans la plupart des cas, on procède donc à un centrage-réduction des variables de départ.

Sujet	Math	Sciences	Français	Latin	Musique
Jean	-1,0865	-1,2817	-1,5037	-1,6252	-1,0190
Aline	-0,4939	-0,6130	-0,6399	-0,7223	-0,6794
Annie	-1,0865	-0,9474	0,2239	-0,1806	0,0000
Monique	1,4322	1,5604	1,5197	1,8058	-1,0190
Didier	1,2840	1,3932	0,5119	0,7223	-0,3397
André	0,3951	0,0557	-1,3597	-1,0835	0,6794
Pierre	-1,2347	-0,9474	1,0878	0,5417	-0,3397
Brigitte	0,9877	0,8916	-0,4959	-0,1806	0,3397
Evelyne	-0,1975	-0,1115	0,6559	0,7223	2,3778

On définit ainsi p variables $Z_1, Z_2, ..., Z_p$.

Nuage des individus - Inertie du nuage

Le nuage des individus est l'ensemble des 9 points correspondant aux 9 sujets, pris dans un espace de dimension 5 (le nombre de variables). La variabilité observée entre les 9 sujets est mesurée par *l'inertie* du nuage de points (vocabulaire issu de la mécanique).

L'inertie totale du nuage est $\sum OM_i^2 = \sum\sum z_{ij}^2 = n \times p = 9 \times 5 = 45$.

Inertie (absolue) de l'individu i : OM_i^2.

Inertie relative de l'individu i : $Inr_i = \dfrac{OM_i^2}{\sum_j OM_j^2}$

L'inertie relative d'un individu est d'autant plus grande que les valeurs des variables observées sur cet individu sont "loin de la moyenne".

L'inertie (absolue) de l'individu i le long d'un axe D est OH_i^2, où H_i est la projection orthogonale du point M_i sur l'axe D. L'inertie relative correspondante est $\dfrac{OH_i^2}{\sum_j OH_j^2}$

Nuage des variables

De façon duale, on peut considérer les 5 points correspondant aux 5 variables, dans un espace de dimension 9 (le nombre des individus).

L'inertie absolue de chaque variable est n, son inertie relative est $\dfrac{1}{p}$.

Corrélations des variables prises deux à deux :

	Math	Sciences	Français	Latin	Musique
Math	1,0000	0,9825	0,2267	0,4905	0,0112
Sciences	0,9825	1,0000	0,3967	0,6340	0,0063
Français	0,2267	0,3967	1,0000	0,9561	0,0380

Latin	0,4905	0,6340	0,9561	1,0000	0,0886
Musique	0,0112	0,0063	0,0380	0,0886	1,0000

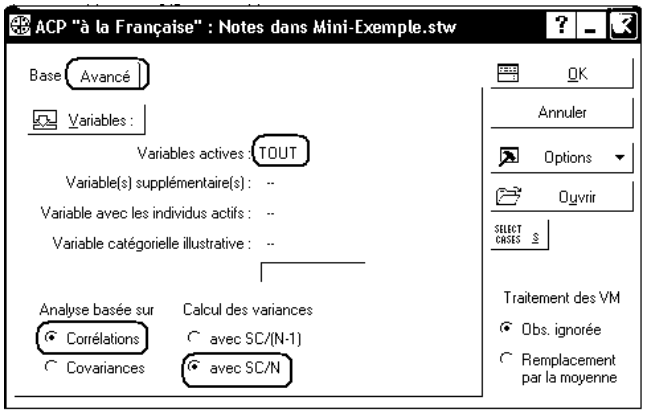
Dans notre exemple, toutes les variables sont corrélées positivement. La corrélation est forte entre les 2 premières, et entre la 3è et la 4è. La cinquième est faiblement corrélée aux autres variables.

3.2.2 L'ACP avec Statistica

Ouvrez un nouveau classeur Statistica, comportant une feuille de données et saisissez les données ci-dessous :

	1 Math	2 Sciences	3 Français	4 Latin	5 Musique
Jean	6	6	5	5,5	8
Aline	8	8	8	8	9
Annie	6	7	11	9,5	11
Monique	14,5	14,5	15,5	15	8
Didier	14	14	12	12	10
André	11	10	5,5	7	13
Pierre	5,5	7	14	11,5	10
Brigitte	13	12,5	8,5	9,5	12
Evelyne	9	9,5	12,5	12	18

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



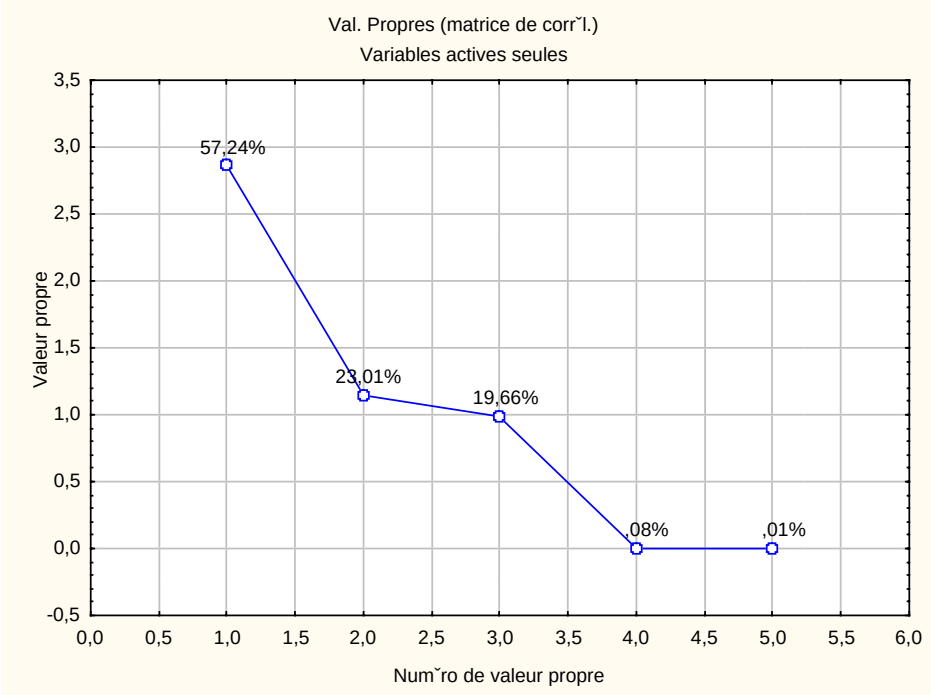
3.2.3 Valeurs propres et vecteurs propres. Composantes principales

Les composantes principales CP_1, CP_2, ..., CP_p sont des variables obtenues comme combinaisons linéaires des variables de départ, et qui vérifient les propriétés suivantes :

- CP_1 représente la direction de plus grande dispersion du nuage de points.
- CP_2 représente la direction de plus grande dispersion des résidus, une fois l'effet de CP_1 pris en compte
- même chose pour CP_3, CP_4, etc
- Les variables CP_k sont indépendantes : si $k \neq l$, alors $Cov(CP_k, CP_l) = 0$
- Les variables CP_k ne sont en général pas réduites : la variance de la composante principale CP_k est égale à la k-ième valeur propre.

Le terme de "valeur propre" (en anglais : eigenvalue) appartient au domaine de l'algèbre linéaire. Il s'agit en fait des valeurs propres de la matrice des corrélations.

	Val. propr	% Total variance	Cumul Val. propr	Cumul %
1	2,8618	57,24	2,86	57,24
2	1,1507	23,01	4,01	80,25
3	0,9831	19,66	5,00	99,91
4	0,0039	0,08	5,00	99,99
5	0,0004	0,01	5,00	100,00



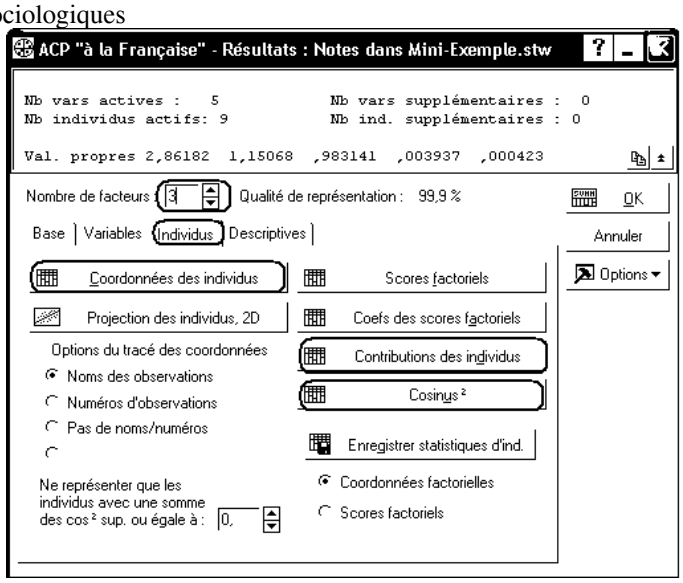
La variation totale (100%) est répartie selon 5 valeurs propres. D'où l'idée de ne garder que les valeurs propres (et directions propres) qui représentent au moins 20% de variation. Dans le cas d'une ACP normée, cela revient à conserver les valeurs propres supérieures à 1.

Variante : on observe une brusque décroissance des valeurs propres entre la 3è et la 4è valeur propre.

Au final, on décide de ne garder que trois valeurs propres.

3.2.4 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", "Cosinus^2".



Scores des individus

Les scores des individus (bouton "Coordonnées des individus") sont les valeurs des composantes principales sur les individus.

Coordonnées factorielles des ind., basées sur les corrélations
Var. illustrative : Sujet

	Fact. 1	Fact. 2	Fact. 3	Sujet
1	-2,7857	0,6765	0,7368	Jean
2	-1,2625	0,3303	0,5549	Aline
3	-1,0167	-1,0198	0,2881	Annie
4	3,1222	0,1659	1,1442	Monique
5	1,9551	0,7879	0,1892	Didier
6	-0,9477	1,2014	-1,1401	André
7	-0,3250	-1,7548	0,9095	Pierre
8	0,6374	1,1298	-0,6919	Brigitte
9	0,6231	-1,5173	-1,9909	Evelyne

Contributions des individus

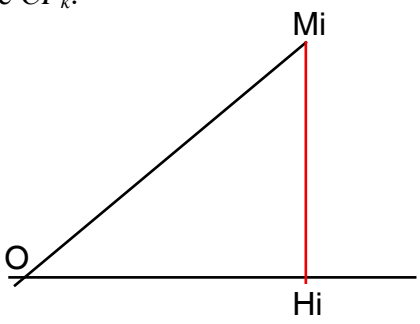
La contribution relative d'un individu i à la formation de la composante principale k est l'inertie relative de cet individu sur l 'axe factoriel k.

Contributions des ind., basées sur les corrélations
Var. illustrative : Sujet

	Fact. 1	Fact. 2	Fact. 3	Sujet
1	30,13	4,42	6,14	Jean
2	6,19	1,05	3,48	Aline
3	4,01	10,04	0,94	Annie
4	37,85	0,27	14,80	Monique
5	14,84	5,99	0,40	Didier
6	3,49	13,94	14,69	André
7	0,41	29,73	9,35	Pierre
8	1,58	12,33	5,41	Brigitte
9	1,51	22,23	44,79	Evelyne

Qualités de la représentation des individus

Géométriquement, la qualité de la représentation d'un individu i par la composante principale k est égale à $\cos^2\theta$, où θ est l'angle $(\overrightarrow{OM_i},\overrightarrow{CP_k})$. Elle mesure la "déformation" due à la projection sur la composante principale CP_k.



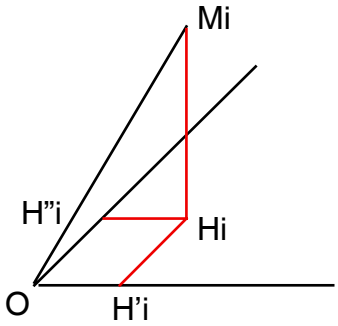
$$QLT(S_i,CP_k)=\frac{OH_i^2}{OM_i^2}$$

Cosinus carrés, basées sur les corrélations (crucianu-1-1.sta)
Var. illustrative : Sujet

	Fact. 1	Fact. 2	Fact. 3	Sujet
1	0,8855	0,0522	0,0619	Jean
2	0,7920	0,0542	0,1530	Aline
3	0,4784	0,4813	0,0384	Annie
4	0,8786	0,0025	0,1180	Monique
5	0,8515	0,1383	0,0080	Didier
6	0,2465	0,3962	0,3568	André
7	0,0263	0,7671	0,2061	Pierre
8	0,1877	0,5898	0,2211	Brigitte
9	0,0583	0,3458	0,5954	Evelyne

Les qualités de représentation sont additives. Par exemple, la qualité de représentation d'un individu i par le plan (CP1, CP2) est donnée par :

$$QLT(S_i,CP_1;CP_2)=\frac{(Score\ de\ S_i\ selon\ CP_1)^2+(Score\ de\ S_i\ selon\ CP_2)^2}{\sum_l (Score\ de\ S_i\ selon\ CP_l)^2}$$

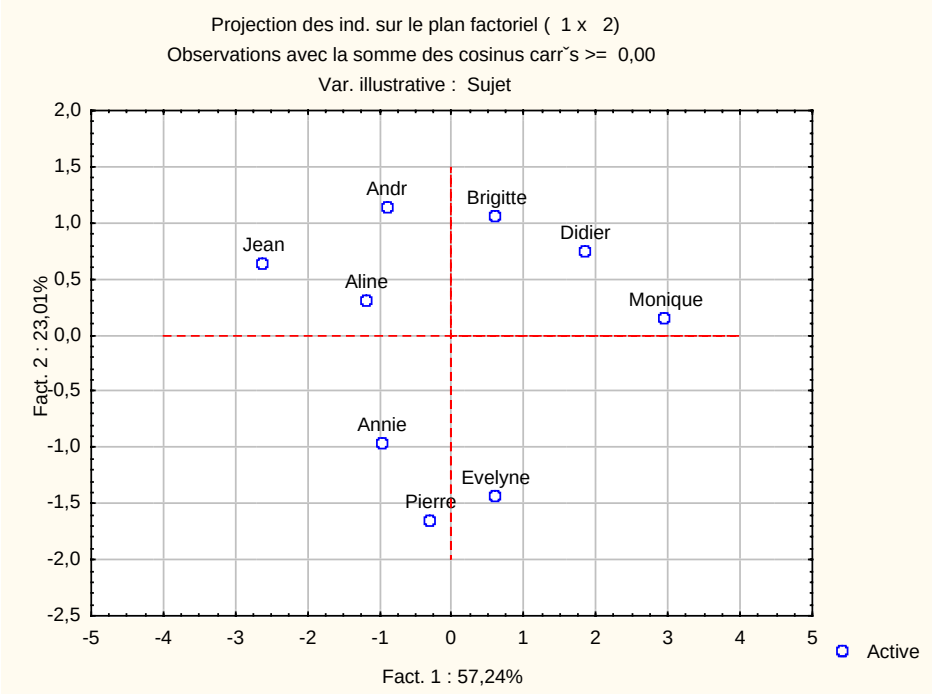


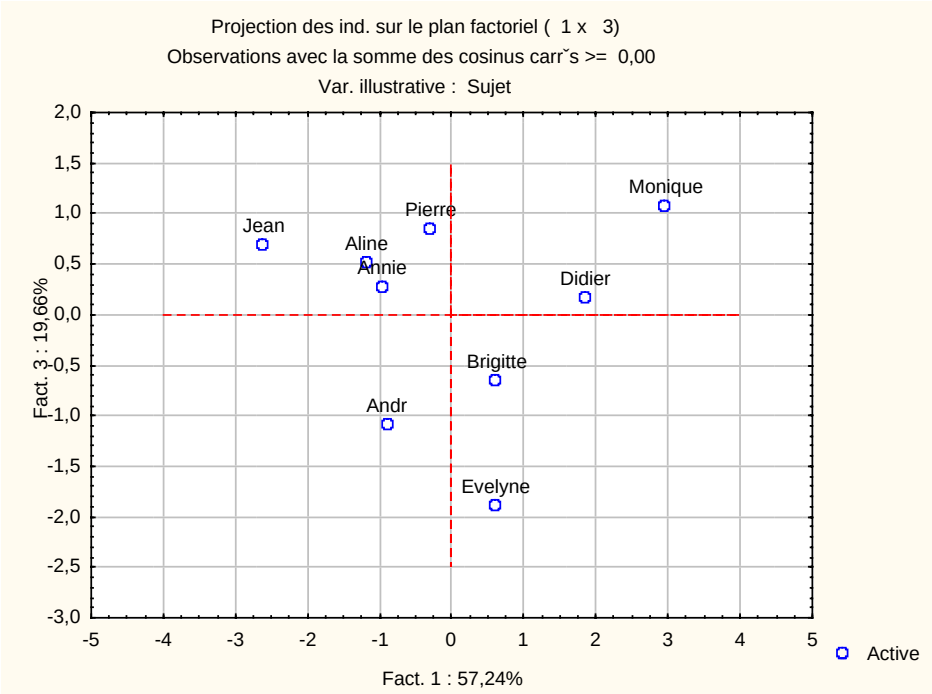
Pour le sujet 1 (Jean), la qualité de représentation par le plan factoriel 1x2 est : 0,8855+0,0522=0,9377.
Cette valeur représente le carré du cosinus de l'angle que fait $\overrightarrow{OM_1}$ avec le plan (CP1, CP2).

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 (c'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

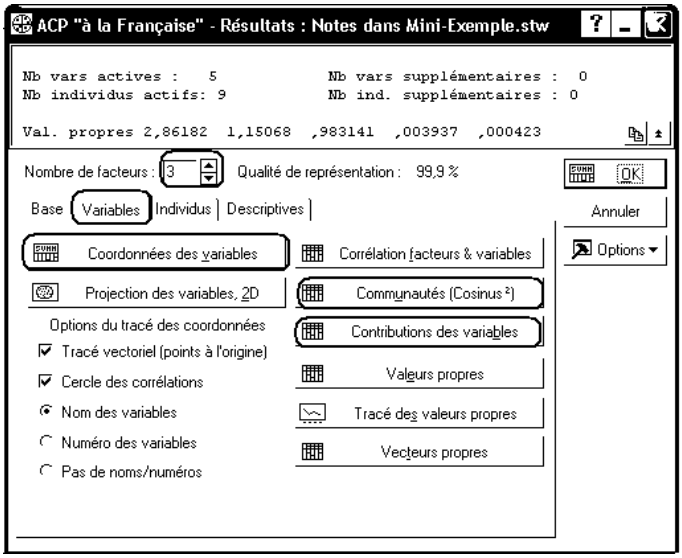
- Utiliser les numéros des observations
- Utiliser les étiquettes indiquées dans la variable "illustrative" : ces étiquettes peuvent être des identifiants des individus, mais peuvent également représenter un groupe d'appartenance, etc.





3.2.5 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 CP1, puis selon le plan (CP1,CP2), puis selon l'espace (CP1,CP2,CP3) en utilisant le bouton "Communautés (Cosinus^2)".

Saturations des variables

Les saturations des variables sont les coefficients de corrélation entre les variables (centrées réduites) de départ et les variables factorielles.

$$SAT(Z_j, CP_k) = \rho(Z_j, CP_k)$$

Coord. factorielles des var., basées sur les corrélations (crucianu-1-1.sta)

	Fact. 1	Fact. 2	Fact. 3
Math	0,8059	0,5714	-0,1534
Sciences	0,8970	0,4308	-0,0929
Français	0,7581	-0,6110	0,2257
Latin	0,9103	-0,3975	0,1084
Musique	0,0667	-0,3275	-0,9425

Contributions des variables

Les contributions des variables à la formation des composantes principales sont définies de la même façon que celles des individus.

Contributions des var., basées sur les corrélations (crucianu-1-1.sta)

	Fact. 1	Fact. 2	Fact. 3
Math	0,2269	0,2837	0,0239
Sciences	0,2812	0,1613	0,0088

Français	0,2008	0,3245	0,0518
Latin	0,2895	0,1373	0,0120
Musique	0,0016	0,0932	0,9035

Qualités de la représentation des variables

La qualité de la représentation d'une variable par une composante principale est définie de la même façon que pour les individus.

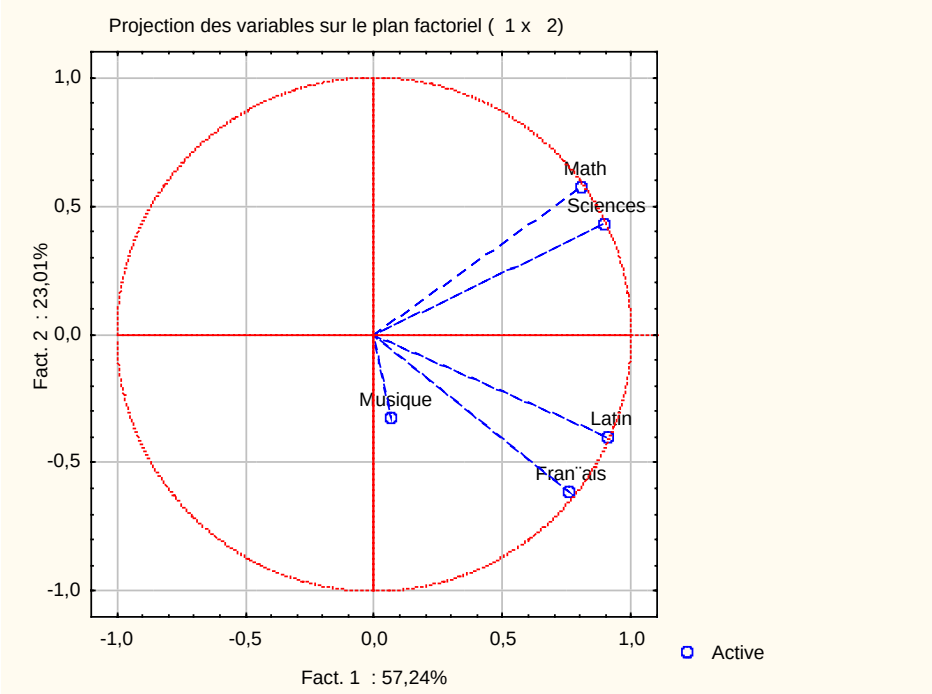
Comme dans le cas des individus, les qualités des représentations d'une variable selon les composantes principales s'additionnent. Le tableau ci-dessous donne les qualités de représentation selon la première composante principale, selon le plan des deux premières composantes et dans l'espace défini par les trois premières composantes.

Communautés, basées sur les corrélations (crucianu-1-1.sta)

	Avec 1 facteur	Avec 2 facteurs	Avec 3 facteurs
Math	0,6495	0,9759	0,9995
Sciences	0,8046	0,9902	0,9988
Français	0,5747	0,9481	0,9990
Latin	0,8286	0,9866	0,9983
Musique	0,0044	0,1117	1,0000

Graphiquement, la qualité de la représentation d'une variable dans le plan (CP_1, CP_2) est le carré de la norme (longueur) du vecteur représentant cette variable (projection de cette variable dans le plan).

Représentation des variables :



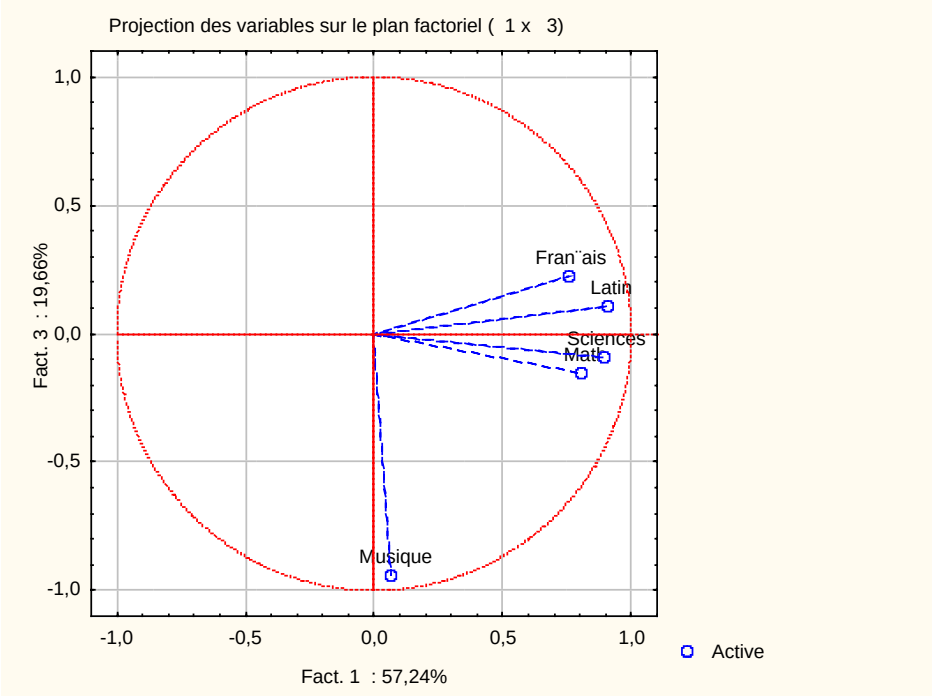
La proximité de l'extrémité du vecteur représentant une variable avec le cercle des corrélations renseigne sur la qualité de représentation de cette variable. Ici, les 4 premières matières sont bien représentées dans le premier plan factoriel. Seule la musique y est mal représentée.

Pour des variables bien représentées, l'angle entre deux vecteurs renseigne sur la corrélation entre ces variables. Ainsi :

- Math et Sciences sont fortement et positivement corrélées (angle proche de 0°)

- Math et Français sont faiblement corrélées (angle proche de 90°).

Enfin, un angle obtus proche de l'angle plat traduirait une forte corrélation négative.



3.2.6 Résultats relatifs à l'analyse elle-même :

Coefficients des variables :

Le tableau des coefficients des variables ("loadings" en anglais) peut être lu de deux façons :

- il permet de calculer les valeurs des composantes principales à partir des variables centrées réduites de départ
- il permet de retrouver les valeurs des variables centrées réduites de départ à partir des valeurs des composantes principales.

Vecteurs propres de la matrice de corrélation
Variables actives seules

	Fact. 1	Fact. 2	Fact. 3	Fact. 4	Fact. 5
Math	0,4764	0,5326	-0,1548	-0,3030	0,6112
Sciences	0,5302	0,4016	-0,0936	0,5168	-0,5308
Français	0,4481	-0,5696	0,2276	0,4775	0,4414
Latin	0,5381	-0,3706	0,1093	-0,6416	-0,3868
Musique	0,0394	-0,3053	-0,9505	0,0390	0,0140

3.3 Interprétation des résultats de l'ACP

3.3.1 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.

3.3.2 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 11% à sa formation et le signe de la coordonnée de chacun de ces traits. On obtient :

-	+
Jean (30%)	Monique (38%) Didier (15%)

On voit que cet axe oppose un sujet tel que Jean, dont les résultats sont globalement modestes, sur la partie négative de l'axe, à des sujets tels que Monique et Didier, dont les résultats sont globalement plus élevés, sur la partie positive.

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

-	+
Pierre (30%) Evelyne (22%)	Brigitte (12%)

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.

3.3.3 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 20%, on obtient :

-	+
	Math (23%) Sciences (28%) Français (20%) Latin (29%)

On retrouve d'un même côté du premier axe, l'ensemble des matières, à l'exception de la musique. Il s'agit là d'une configuration classique lorsque les variables de départ sont toutes corrélées positivement entre elles : **l'effet de taille**.