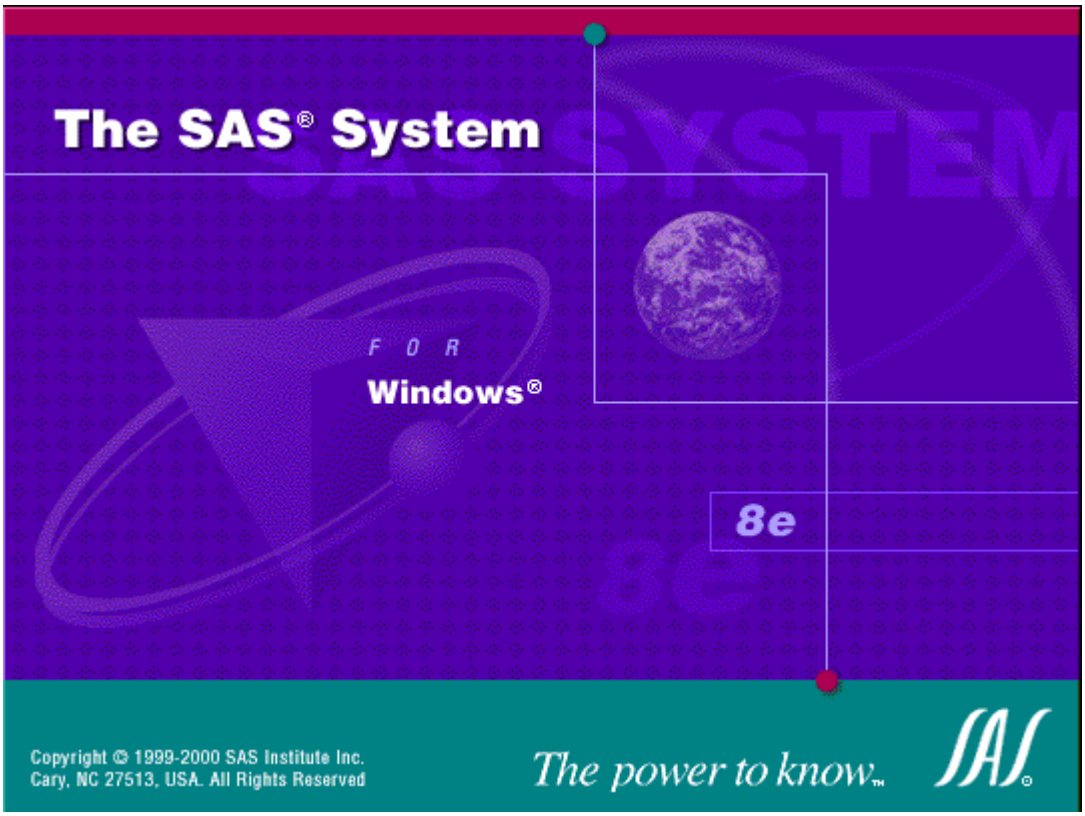


COURS, TP

Modules utilisés : Base, Stat, Insight, Ets, Graph, Assist

Introduction

Le « système SAS™ » est LE logiciel de traitement de données[1]. Les procédures SAS sont très complètes et dépassent largement le cadre du DUT STID.

SAS™ est très répandu. Il a acquis depuis son lancement en 1960 une situation dominante dans beaucoup de secteurs d'activités. En France l'INSEE, ELF, EDF et depuis peu les banques, les assurances, les CAF l'ont adopté.

SAS™ peut s'utiliser, dans le cadre de votre formation, en assimilant le langage SAS[2] ou en utilisant les modules « cliquer-résultat » comme SAS/ *ASSIST* ou SAS *Enterprise Guide*. Nous privilégierons la première approche car elle permet une utilisation plus approfondie de SAS™ même si c'est au prix d'un certain temps d'apprentissage.

Ajoutons que SAS™ ne peut être utilisé « convenablement » que par des personnes ayant les connaissances requises en statistiques et en programmation. Il est devenu incontournable dans la majorité des stages et offres d'emplois qui nous arrivent.

Le Système SAS™ fonctionne sur plusieurs systèmes (MAC, PC Dos et WINDOWS, UNIX...).

La version WINDOWS nécessite:	16 M0 de mémoire vive RAM (minimum) 500 M0 sur le disque dur. (minimum)

Ce logiciel est commercialisé par :

SAS INSTITUTE B.P.5 77166 GREGY-SUR-YERRES ☎: 0160621111 Fax:0160621199 Contact : Ariane Ligier – Bellair

[1] SAS commercialise un autre logiciel statistique (très convivial) SAS JMP3.2. Il n'occupe que quelques mégas sur le disque dur. Il est très convivial (menus...) et assez complet (Plans d'expériences, Surfaces de réponse, Régression logistique...) mais malheureusement limité dans certains domaines (importation de données, paramètrage des sorties, des plans d'expériences fractionnés...). Néanmoins, il peut être une alternative intéressante pour ceux qui n'ont pas besoin de toute la puissance du système SAS ou qui sont allergiques au langage SAS !

[2] SAS possède en fait 3 langages. Le langage SAS et son module Macros bien sûr, mais aussi le langage SQL bien connu dans l'univers des SGBD et le langage le SCL pour créer des applications type Visual Basic (SAS/AF, SAS/FSP)

SOMMAIRE

Pour bien utiliser ce cours...

"J'entends, j'oublie
J'apprends, je retiens
Je fais, je comprends"

Proverbe chinois[3]

Le but de ce cours est de vous présenter une petite palette d'outils SAS afin de mettre en pratique vos connaissances en statistiques et en informatique.

On ne peut faire le tour des possibilités de SAS en 50H de TD! Il en faudrait au moins 5 fois plus ! Nous avons donc omis une grande partie des nombreuses options des procédures, commandes et instructions[4]. L'aide en ligne du logiciel ou l'excellente documentation papier[5] sont là pour vous permettre d'approfondir les notions vues en cours.

Pour que ce cours soit profitable, il faut le travailler régulièrement. Il ne faut pas hésiter à y revenir hors séance.

N'hésitez pas à me faire part de vos commentaires sur ce document afin de l'améliorer pour les candidats futurs.

Bon courage !

[3] Cité par M. Tennenhaus dans une conférence de l'ASU sur l'emploi des logiciels en Statistique

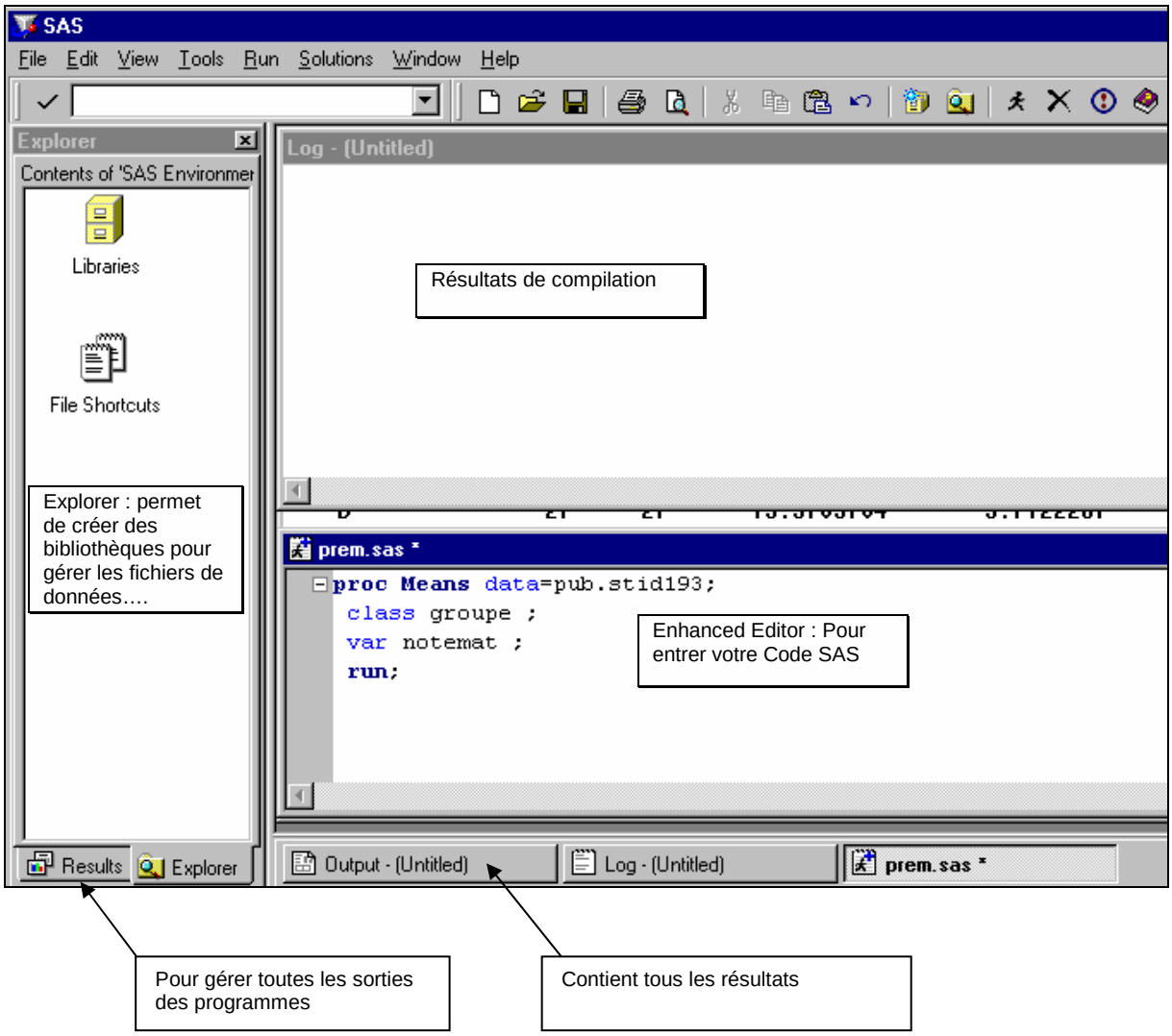
[4] Ce document totalise 400 pages environ contre plus de 10000 pour la documentation papier officielle SAS…

[5] Vous devrez impérativement la consulter si vous voulez être spécialiste SAS plus tard…

I. Premier contact avec SAS

Lancez le programme SAS, vous allez voir apparaître le DMS (Display Manager System) de SAS qui contient cinq fenêtres essentielles[6]:

A. Cinq fenêtres essentielles



[6] Si tel n'est pas le cas, allez dans le menu Windows/Cascade ou si une des fenêtres est absente, faites View/'nom de la fenêtre'

La fenêtre Explorer

Permet de gérer les bibliothèques et les fichiers de données. On peut créer, visualiser, modifier une fichier de données.

La fenêtre Enhanced EDITOR [7]

Elle contient, comme son nom l'indique, les instructions SAS à exécuter. Grâce aux menus attachés à CETTE fenêtre, vous pouvez saisir un programme, le sauvegarder, le rappeler, le modifier...

La fenêtre LOG (Touche F6)

Après exécution d'un programme, cette fenêtre contient chaque instruction exécutée et éventuellement les erreurs rencontrées. **Il est indispensable de consulter cette fenêtre avant de lire les résultats contenus dans la fenêtre OUTPUT**.

La fenêtre OUTPUT (Touche F7)

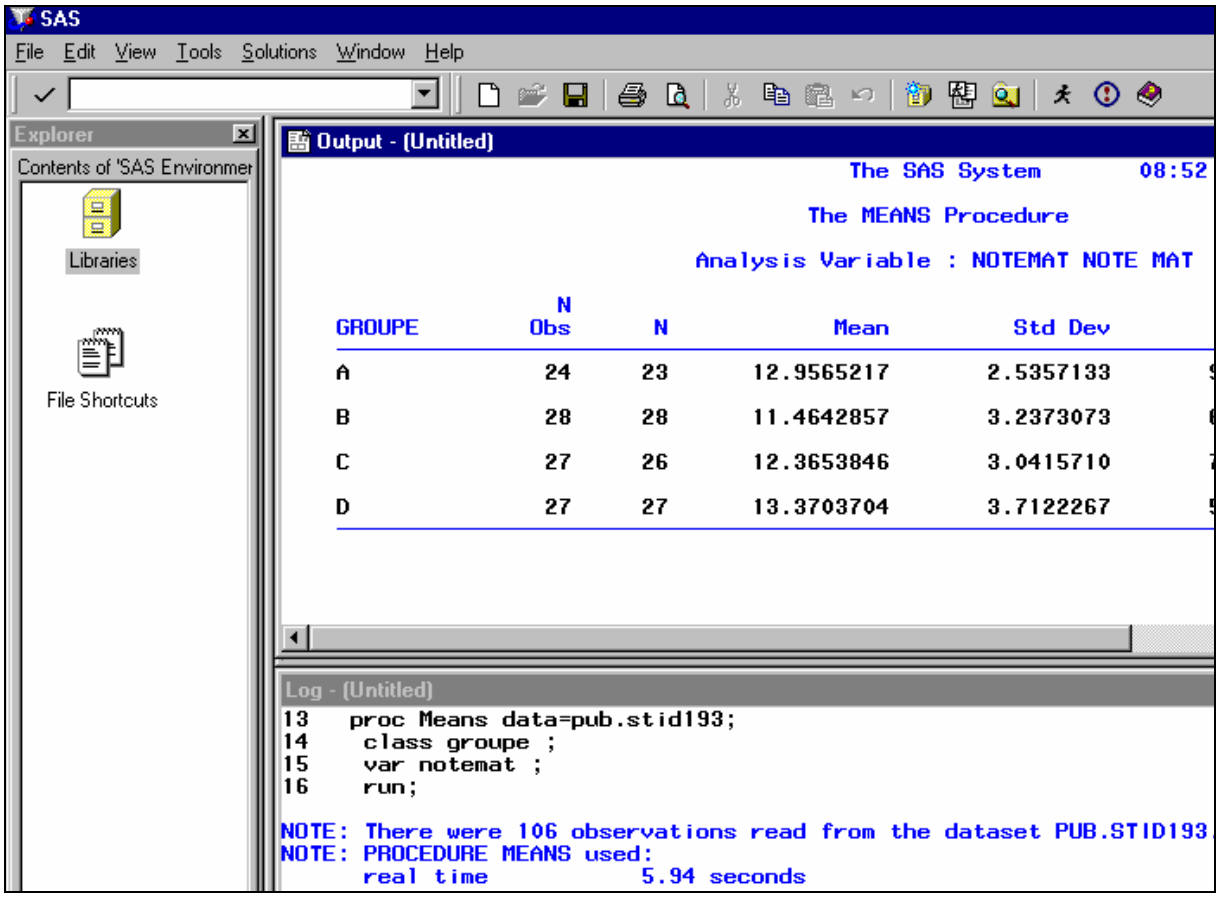
Cette fenêtre contient tous les résultats des instructions exécutées par SAS: tests, tableaux de valeurs ... **Elle ne doit être consultée qu'après la LOG**.
Il vous est possible de sauvegarder tout ou partie du contenu de cette fenêtre et de récupérer le contenu sous WORD.

La fenêtre RESULTS

Permet de gérer toutes les sorties produites par les programmes SAS exécutés préalablement. Elles permet d'accéder rapidement à la sortie qui vous interesse.

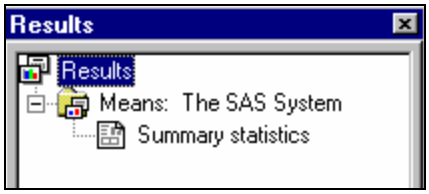
[7] En fait, il y en a deux. Il y a l'ancienne (Program Editor V6.12) et celle-ci beaucoup plus agréable à utiliser (indentation automatique, reconnaissance des caractères par des couleurs etc.)

Ainsi, après exécution du programme contenu dans l' *Enhanced* `EDITOR` précédent, on observe le résultat suivant:



Vous constatez que les instructions et les commentaires d'exécution sont "passés" dans la fenêtre "LOG", quant aux résultats, ils figurent dans la fenêtre "OUTPUT".

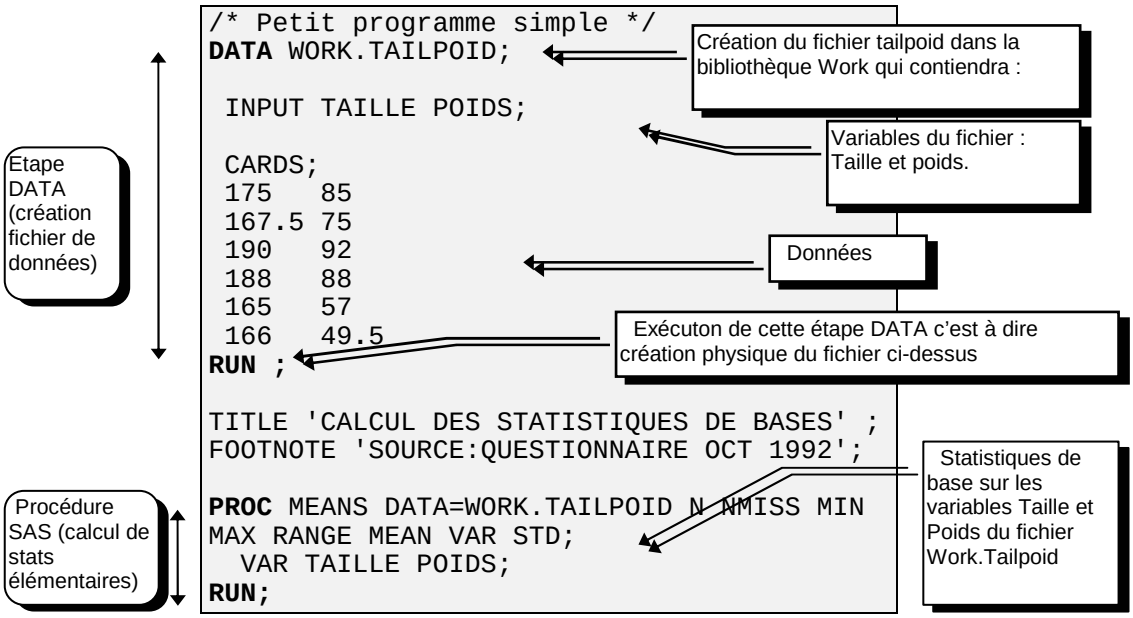
Quant à la fenêtre Results, elle contient le « plan » des résultats disponibles.
Nous voyons qu'une procédure MEANS a été exécutée et que nous disposons des « Summary Statistics » :



B. Mon premier programme SAS

1. Saisie du programme

Activez la fenêtre *Enhanced EDITOR* et cliquez sur FILE/NEW pour vider son contenu. Vous allez taper (à la lettre !) dans la fenêtre le programme suivant :



- Ce programme crée un fichier de données TAILPOID dans la bibliothèque WORK contenant deux variables numériques TAILLE et POIDS puis calcule quelques statistiques de base.

- L'instruction CARDS ; indique à SAS l'emplacement des données. Notez les RUN ; qui terminent les procédures et les étapes DATA et aussi le « . » qui sert de séparateur décimal (comme pour MINITAB)

Un mot sur PROC MEANS[8]

(à passer en première lecture)

« PROC MEANS etc. » demande à SAS d'effectuer quelques statistiques élémentaires sur les variables taille et poids dont voici la signification (nous en ajoutons d'autres qui peuvent être demandées en option) :

Terme	Signification	Terme	Signification
N	Nombre d'observations	RANGE	L'étendue (MAX-MIN)
NMISS	Nombre d'observations manquantes	MEAN	La moyenne
MIN	Le minimum	VAR	La variance
MEDIAN	Médiane	Q3	3eme quartile
Q1	Premier quartile	QRANGE	Q3-Q1
MAX	Le maximum	STD	L'écart-type

Sa syntaxe peut être très complexe :

L'instruction CLASS indique quelles variables utiliser pour définir les sous populations.

```
proc means data=moi.stid193 mean median q1 q3;
var taille;
class groupe sexe bac;
run;
```

Extrait de la sortie :

GROUPE	SEXE	BAC	N Obs	Mean	Median	Lower Quartile	Upper Quartile
A	1	B	2	174.0000000	174.0000000	168.0000000	180.0000000
		C	5	182.2000000	181.0000000	179.0000000	183.0000000
		D	2	173.0000000	173.0000000	168.0000000	178.0000000
		PRO	1	186.0000000	186.0000000	186.0000000	186.0000000
	2	A	2	168.5000000	168.5000000	167.0000000	170.0000000
		B	5	163.0000000	162.0000000	162.0000000	164.0000000

Pour limiter le croisement des variables précédentes, vous pouvez utiliser WAYS :

Complément sur Proc Means : WAYS, TYPE

```
proc means data=moi.stid193 mean median q1 q3;
var taille;
class groupe sexe bac;
ways 1 2;
run;
```

Ne va combiner les variables CLASS que 1 à 1 ou 2 à 2 et ce, grâce à WAYS… pour n'avoir qu'une seule combinaison il suffit de mettre 1. L'instruction TYPE GROUPE*SEXE ; permet de n'avoir que la variable groupe croisée avec la variable SEXE…

[8] Pour en savoir plus, allez dans l'aide, puis dans « Help on SAS Software PRODUCTS » puis dans Search entrez MEANS, vous pouvez ensuite accéder à la procédure PROC MEANS. Nous verrons plus loin une procédure plus complète pour traiter les données quantitatives :PROC UNIVARIATE

2. Sauvegarde des instructions du programme[9]

Toujours dans la fenêtre " EDITOR" faites FILE/SAVE AS et enregistrez ce programme dans votre répertoire sous le nom «PREMIER.SAS » par exemple.[10]

3. Exécution du programme (F8)

Exécution de tout le contenu de la fenêtre

Nous allons exécuter le programme précédent. Assurez-vous que la fenêtre LOG est visible. Dans la fenêtre "EDITOR" allez dans LOCALS/SUBMIT, ou cliquez sur le bouton  ou encore tapez sur F8.

Une fois l'exécution achevée, allons dans la fenêtre LOG (F6) pour voir les commentaires d'exécution[11]. Il est fondamental d'y aller AVANT d'interpréter les résultats car elle contient les éventuels messages d'erreurs.

Exécution partielle

Pour exécuter une partie du programme figurant dans la fenêtre Program Editor, sélectionnez la avec la souris (mettez la en surbrillance) puis faites un submit. Seule la partie sélectionnée a été exécutée.

---

9 Le fichier de programme est un fichier texte DOS banal qui peut ensuite être édité sous Word (en police courier new pour conserver l'alignement)

10 L'extension « .SAS » est réservée au fichiers de programme SAS. Ce sont des fichiers ASCII standard.

11 Il vous est possible aussi de sauvegarder le contenu de cette fenêtre LOG (menu File/...) dans un fichier que vous nommerez PREMIER.LOG par exemple. Vous pourrez ainsi regarder le contenu à tête reposée !

4. Visualisation des résultats et personnalisation de la fenêtre OUTPUT (complément)

Si tout s'est bien passé, la fenêtre OUTPUT (F8) apparaît avec les résultats.

Par défaut SAS affiche dans la fenêtre OUTPUT, la date, le titre SAS, votre titre, le numéro de page… Vous pouvez changer cela par un

```
Options nodate nonumber ;
```

à mettre au début de votre premier programme avant de l'exécuter à nouveau.

Vous pouvez aussi changer la taille du contenu de la fenêtre OUTPUT avec les options LINESIZE= nb de caractères par ligne PAGESIZE=nb de lignes par page :

```
Options linesize=70 pagesize=35 ;
```

Quant aux titres, vous pouvez les gérer par l'instruction TITLE ou par un CTRL T et les notes de bas de page par un FOOTNOTE ou par un CTRL F.

Correction de votre programme

En cas de problème, vous pouvez corriger votre programme.

Quand tout est correct, retournez dans la fenêtre OUTPUT. [12]

5. Sauvegarde des résultats contenus dans OUTPUT

a) Dans un document WORD

Avec la souris, sélectionnez le tableau des résultats (la zone change de couleur), faites `EDIT/COPY TO PASTE BUFFER` (la couleur d'origine revient)[13]. Basculez vers *WORD*, sélectionnez la police *COURIER NEW*, et collez le résultat !

b) Dans un fichier Texte depuis SAS

En effet, vous pouvez aussi sauvegarder directement les résultats de la fenêtre OUTPUT dans un fichier texte, nommé PREMIER.LST par exemple, (File/SAVE As) que vous pourrez rappeler sous WORD.

6. Sauvegarde du fichier des données

Elle est automatique ! Nous verrons plus loin que les données sont automatiquement placée en C :\SASWORK\TAILPOID.SD2 [14]

Vos données sur la taille et le poids sont dans le fichier SAS temporaire WORK.TAILPOID. Vous pouvez y accéder sans recréer ce fichier

Ainsi, si vous souhaitez faire un nuage de points avec ces données, tapez simplement dans la fenêtre PROGRAM EDITOR (à la suite du programme précédent)

```
PROC GPLOT DATA=WORK.TAILPOID ;
 PLOT TAILLE*POIDS ;
RUN ;
QUIT ;
```

[12] Si vous trouvez que la fenêtre OUTPUT est peu lisible (trop de lignes par page ou pas assez, trop de caractères par ligne ou pas assez, vous pouvez modifier cela en insérant au début de votre programme un « `OPTIONS LINESIZE=70 PAGESIZE=35 ;` » allez voir l'annexe pour plus d'informations sur les options.

[13] Si tel n'est pas le cas, vous avez probablement sélectionné une zone interdite. Vous ne devez pas faire descendre le curseur au delà de la dernière ligne de résultats.

[14] Néanmoins, ce fichier ne peut être lu QUE par un programme SAS (ou par le SAS viewer qui est un petit programme SAS libre de droits permettant de lire tous les fichiers de données SAS)

7. Ne confondez pas...

Vous venez de voir 4 sortes de fichiers qu'il ne faut pas confondre :

1)Le fichier de programme édité dans la fenêtre PROGRAM EDITOR et qui contient vos instructions SAS (PREMIER.SAS)

2)Le fichier de données WORK.TAILPOID qui a été créé par votre programme et qui contient les données de votre étude statistique.

3) et 4) Les fichiers PREMIER.LOG et PREMIER.LST qui contiennent, si vous les avez créés, les erreurs de compilation de la fenêtre LOG et les résultats contenus dans la fenêtre OUTPUT.

II. Fichiers de données SAS

Méthode
SAS ne peut effectuer de calculs que sur des fichiers de données type - SAS. Vous devez donc, avant toutes choses saisir vos données dans un programme SAS ou, ce qui est le plus courant, convertir votre fichier EXCEL, DBASE, Lotus 1-2-3 , ASCII au format SAS.

SAS sait convertir directement (grâce à FILE/IMPORT) les fichiers *EXCEL*, *DBASE*, ASCII[15](texte) etc.[16]

A. Préliminaires sur les fichiers de données SAS

1. Nom logique d'un fichier de données SAS

SAS utilise son propre système pour nommer les fichiers de données[17]. Tout fichier SAS a un nom du type `LIBREF.FILE` où `LIBREF` *e*st le nom de la bibliothèque (8 caractères maximum) et `FILE` le nom du fichier (32 caractères maximum). La bibliothèque est l'endroit où se trouve le fichier de données SAS. Elle se substitue au chemin du DOS.

Comparaison entre les noms des fichiers de données usuels et ceux de SAS

Sous EXCEL, WORD, etc.		**Sous SAS**	
« Chemin et Nom de fichier »	***« Extension »*** *caractérise l'application (XLS pour EXCEL, DOC pour WORD)*	***« Nom de bibliothèque »*** *identifie le « répertoire DOS »où SAS va chercher le fichier dont le nom figure à droite.*	***« Nom du fichier »***
C:\WORD\COURS .	**DOC**	**WORK .**	**STID193**
G:\MONREP\TOT .	**XLS**	**WORK .**	**DONNEES**
I:\MART\STID193 .	**XLS**	**SASUSER .**	**STID193**

Comment SAS s'y retrouve-t-il ?

Prenons par exemple le fichier de données WORK.BIDULE. Il désigne le fichier SAS Bidule dans la bibliothèque WORK [18]

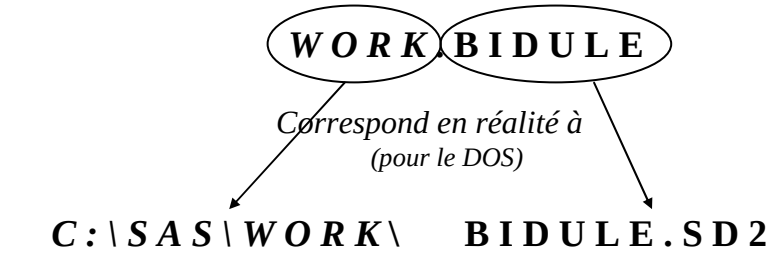
---

[15] S'ils ne sont pas trop compliqués : Pas de ligne de titre, le séparateur de variables est un espace. Si FILE/IMPORT ne fonctionne pas avec votre fichier, il faudra utiliser l'importation classique utilisant un programme SAS. Cf. « Importation d'un fichier ASCII » de ce document. Grâce à son langage puissant, SAS peut en effet importer n'importe quel fichier texte aussi compliqué soit il !

[16] Théoriquement SAS reconnaît les fichiers SPSS et BMDP. La procédure IMPORT permet d'effectuer le transfert.

[17] Ceci peut sembler bizarre à première vue. Cela dit SAS est multi plate-forme : il fonctionne également sous UNIX, NT etc. Ce système de noms particuliers permet aux programmes SAS de fonctionner sur n'importe quelle plate-forme après quelques modifications mineures des programmes (Libname par ex.)

[18] Work est une bibliothèque prédéfinie par SAS. Elle est située physiquement en C:\SAS\WORK. C'est à dire que le fichier Bidule s'y trouve physiquement



Remarques:

Tout fichier de données SAS a donc une adresse de stockage. Il est inutile de le 'sauvegarder' (contrairement à MINITAB ou EXCEL) car ceci est fait automatiquement fait par SAS. Par contre vous pouvez recopier le fichier à un autre endroit pour plus de sûreté... (cf. PROC DATASETS)[19]

Si vous voulez mettre vos fichiers de données sur disquette (A:) ou dans votre répertoire réseau (Z:\TOTO) il vous faudra créer une bibliothèque dont l'adresse sera l'endroit où vous voulez mettre vos fichiers.[20]

2. Les 2 bibliothèques prédéfinies WORK et SASUSER

Il y a d'origine deux bibliothèques sous SAS. Une temporaire (WORK) et une permanente (SASUSER). La bibliothèque WORK détruit les fichiers qu'elle contient dès que vous quittez SAS. A l'inverse de SASUSER.

Elles correspondent respectivement aux répertoires DOS physiques C:\Windows\Temp\SAS temporary Files et C:\Mes Documents\SASV8[21].

Remarques:

Vous pouvez vérifier que le fichier `TAILPOID` de votre « premier programme SAS » se trouve bien en `C:\WINDOWS\TEMP` sous le nom `tailpoid.sd2`. Toutefois vous ne pourrez visualiser ce fichier qu'avec un programme SAS ou le SAS VIEWER[22].

Les fichiers de ces bibliothèques sont donc sauvegardés physiquement sur C:, le disque dur de l'ordinateur dont vous vous servez. Ceci est dangereux si vous n'êtes pas le seul utilisateur de ce micro... Il est donc conseillé de créer votre propre bibliothèque et d'y mettre vos fichiers de données SAS.

---

[19] On peut être tenté de faire les copies de fichiers de données SAS en utilisant le gestionnaire de programmes ou l'explorateur de Windows puisque nous connaissons le nom DOS du fichier. Cela dit, si la copie d'un fichier devient systématique, il est préférable de l'effectuer via la procédure DATASETS pour respecter la compatibilité multi plateforme dont nous parlions dans la note précédente.

[20] (cf. instruction Libname plus loin dans ce document)

[21] Ceci peut changer d'un système à un autre.

[22] Application fournie « gracieusement » par SAS permettant de consulter les fichiers de données.

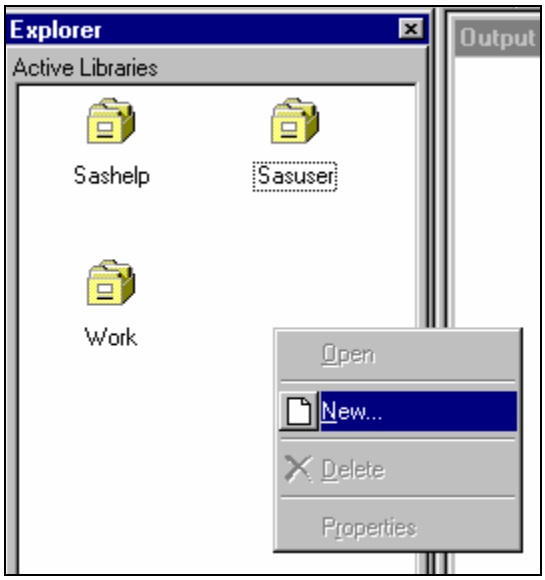
3. Comment créer VOTRE bibliothèque ?

Nous allons maintenant créer une nouvelle bibliothèque appelée MOI qui pointe sur le répertoire D:\DATA (ce répertoire n'existe pas chez vous, c'est simplement un exemple de démonstration).

a) Avec l'explorer

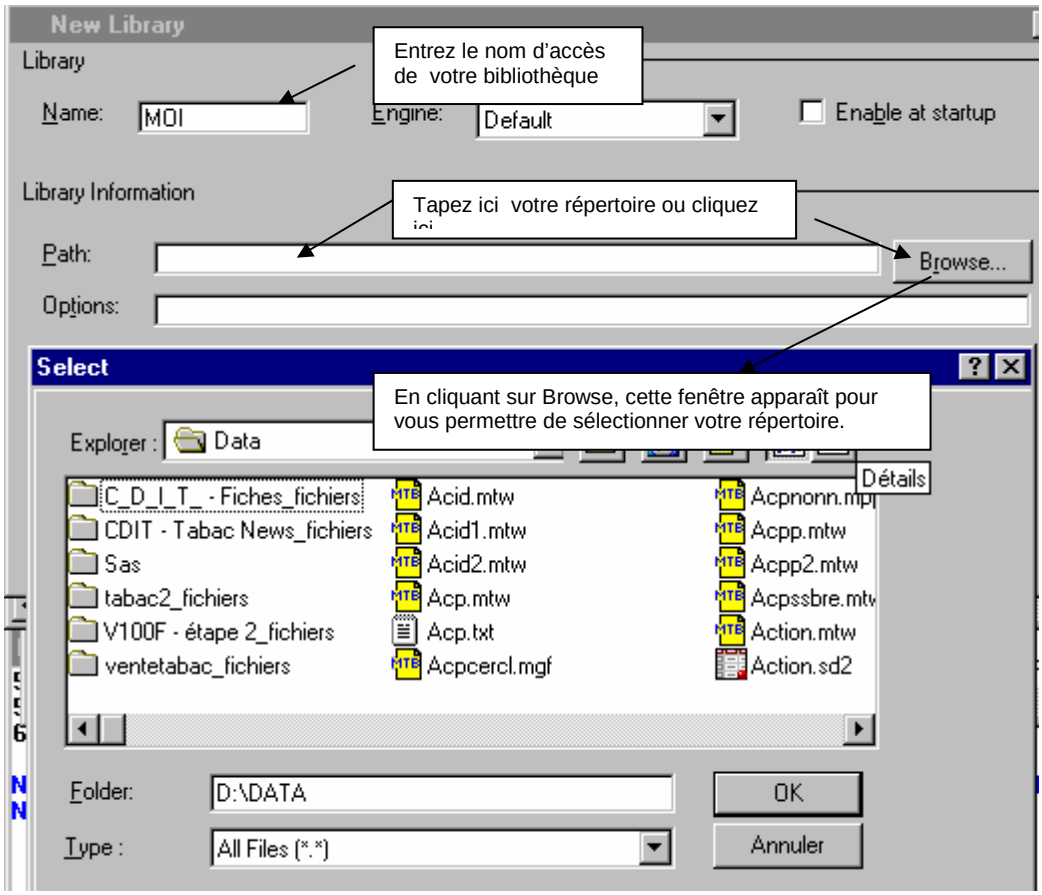
Activez la fenêtre Explorer.[23] Au moins trois bibliothèques par défaut sont actives sous SAS :

SASHELP, SASUSER, WORK

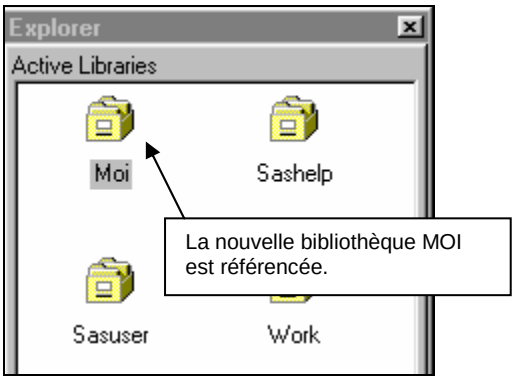


Pour ajouter une nouvelle bibliothèque, cliquez sur le bouton droit et choisissez NEW, ou cliquez sur  de la barre d'outils, ou entrez la commande LIBASSIGN dans la ligne de commande.

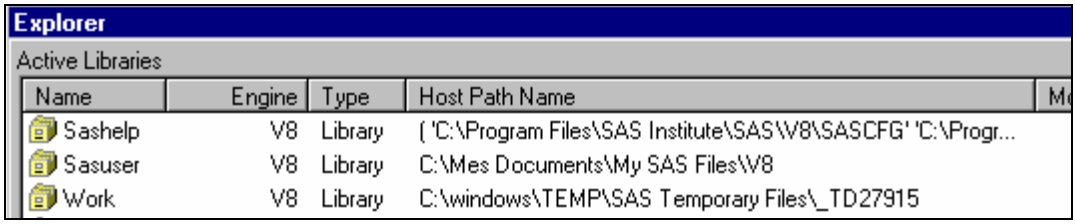
[23] Allez dans View/Explorer si vous ne la voyez pas.



Validez. Si tout a bien fonctionné, vous devriez avoir dans la fenêtre Explorer :



Pour avoir des détails sur ces bibliothèques, allez dans View/ Details :



b) Création d'une bibliothèque dans un programme SAS

Pour faire la même chose en utilisant un programme, il suffirait de taper : **`LIBNAME MOI 'D:\DATA';`** et de le compiler.

Remarque :

La bibliothèque MOI est référencée (elle a une adresse, SAS peut désormais y accéder), son adresse physique est D:\DATA . Si je crée sous SAS le fichier de données MOI.STID193, il sera physiquement stocké en D:\DATA\STID193.SD2

Exercice:

Créez-vous une bibliothèque (8 caractères maximum) avec comme adresse physique votre répertoire serveur(et éventuellement un sous - répertoire). Modifiez « mon premier programme » pour que le fichier de données TAILPOID soit directement créé dans votre répertoire. Vérifiez dans la "LOG" que tout s'est bien passé.

Remarques:

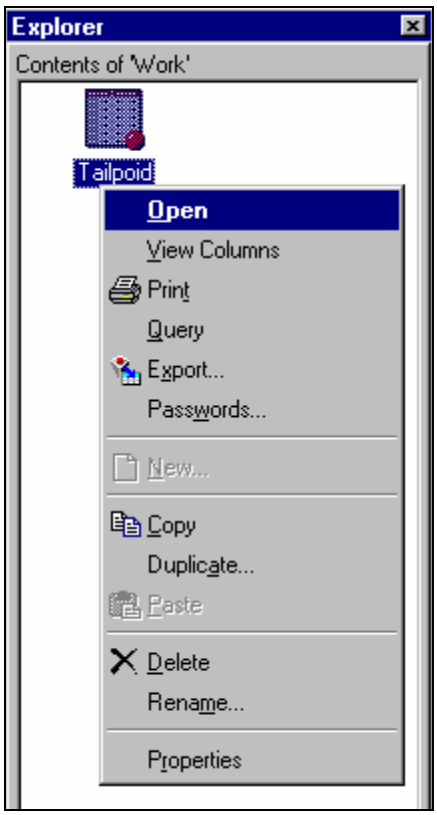
SAS « oublie » les noms de vos bibliothèques – mais pas le contenu !- dès que vous le quittez[24]. Pensez à les redéclarer au début de chaque session. D'autre part, le nom '**`MOI`**' n'a aucune importance ; on peut mettre n'importe quel nom (<=8 caractères) à condition de s'en rappeler !

[24] Sauf si vous cochez la case ENABLE AT STARTUP de la fenêtre de création de bibliothèques.

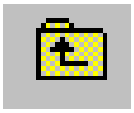
4. Visualisation du contenu d'un fichier, modifications...

Il suffit de cliquer sur la bibliothèque dans laquelle il se trouve, puis de cliquer sur le fichier concerné. En suite, en cliquant sur le bouton droit de la souris, vous faites apparaître un menu contextuel qui vous permet de visualiser le fichier :

Ici, nous avons ouvert la bibliothèque WORK, nous y avons trouvé notre fichier de données :



Remarque Importante :

Pour revenir en arrière (Fermer la fenêtre donnant le contenu d'une bibliothèque et afficher la liste des bibliothèques… cliquez sur  de la barre d'outils.

Exercice :

- Recherchez le fichier TAILPOID que vous avez créé Cf. « Mon premier programme ». Et visualisez-le avec les commandes précédentes.

- Pour modifier les données de façon intéractive, passez en `Edit/Edit Mode`[25] et `Edit/Table LEVEL EDIT ACCESS.` Ajoutez une nouvelle ligne de donnée (`Edit/Add Row`, faites un `Edit/Commit New Row` pour valider la saisie d'une nouvelle ligne).

- Un File/Close permet de terminer la modification.

- Vous pouvez maintenant réexécuter la fin[26] de votre petit programme pour obtenir des statistiques à jour. Pour cela mettez en surbrillance la portion du programme a exécuter et faites un Local/Submit.

```
PROC MEANS DATA=MOI.TAILPOID N NMISS MIN
MAX RANGE MEAN VAR STD;
  VAR TAILLE POIDS;
RUN;
```

[25] Le Browse Mode est le mode de lecture seule.

[26] Si vous exécutez tout le programme, le fichier de données sera remis à sa forme d'origine à cause de l'instruction DATA…

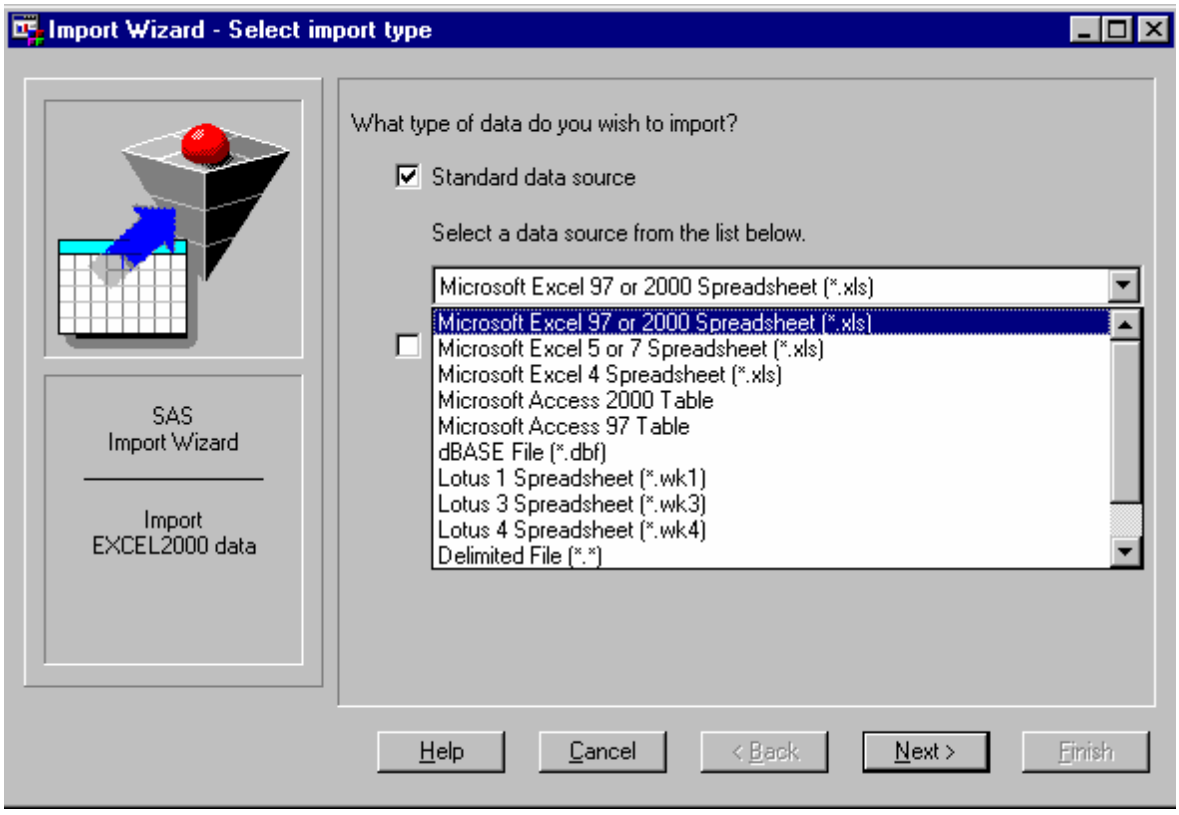
B. Conversion automatique d'un fichier EXCEL[27] (File/Import)

Prenons le fichier EXCEL : STID193.XLS qui contient des données sur les STID1ere année 1993 : (Les variable sont : Groupe, ordre (dans le groupe), Sexe, Série du Bac, Date (de naissance), Nombre de frères et soeurs(NBFS), Notfr, Nothis, Notmat (les notes en français, histoire géo et maths au bac), la façon dont ils ont connu l'IUT (IUT ?), leur taille et leur poids et le code postal de leur lycée.

Nous allons le transformer en un fichier de données SAS pour pouvoir travailler dessus sous SAS. Vous allez être guidé pas à pas par un assistant pour effectuer le travail.

1°) Déclarez, si ce n'est pas déjà fait votre bibliothèque sous SAS.
2°) Sous SAS, allez dans FILE/IMPORT DATA, vous obtenez :

1. Choix du type de fichier



Choisissez « *EXCEL 97 2000* » comme format de fichier à importer. Cliquez ensuite sur « NEXT ».

[27] Ceci suppose que le module ACCESS to PC FILE FORMAT est installé. Si tel n'est pas le cas, il faut convertir votre fichier Excel en CSV et importer ce type avec SAS. C'est un peu plus lourd mais cela fonctionne. Pour l'importation de fichiers ASCII voir en fin de ce document.

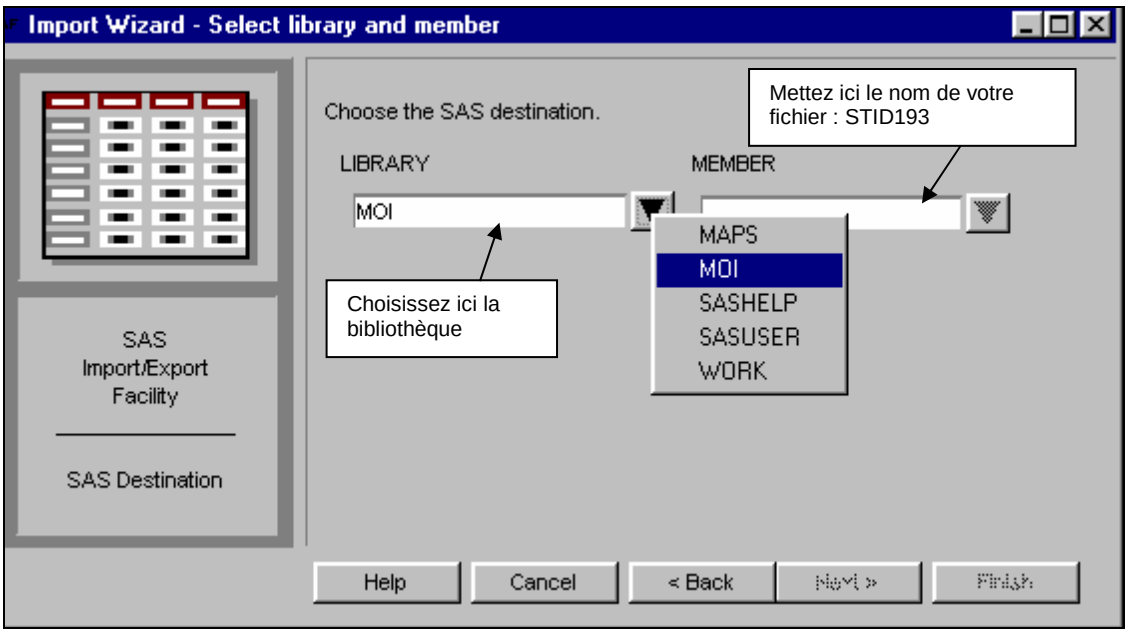
2. Emplacement du fichier à convertir

SAS vous demande ensuite « `WHERE IS THE FILE LOCATED` ? » C'est à dire : où se trouve le fichier à importer ?, Vous pouvez alors taper le chemin et le nom du fichier ou, si vous ne vous en rappelez plus, effectuer un « Browse » pour parcourir les différents répertoires.
En ce qui nous concerne, le fichier est en P:\LOGICIEL.

Il vous suffit donc de taper P:\LOGICIEL\STID193.XLS ou d'aller le chercher dans les répertoires (BROWSE)
et de cliquer sur Next.

3. Nom du fichier SAS obtenu

SAS demande ensuite la bibliothèque et le nom du fichier SAS résultat[28].



- Nous choisissons ici « MOI » comme bibliothèque[29] et STID193 comme nom de fichier.
- SAS demande ensuite si vous souhaitez récupérer le programme ayant permis de faire cette importation. Nous n'en avons pas besoin ici.[30]
- Cliquez sur Finish. Si tous s'est bien passé, dans la fenêtre LOG, vous devez avoir le message suivant :

```
NOTE: MOI.STID193 WAS SUCCESSFULLY CREATED.
```

[28] Comme vous le savez tout fichier de données SAS possède un nom accolé à son nom de bibliothèque qui n'est autre que le chemin du DOS

[29] (qui a été précédemment déclarée)

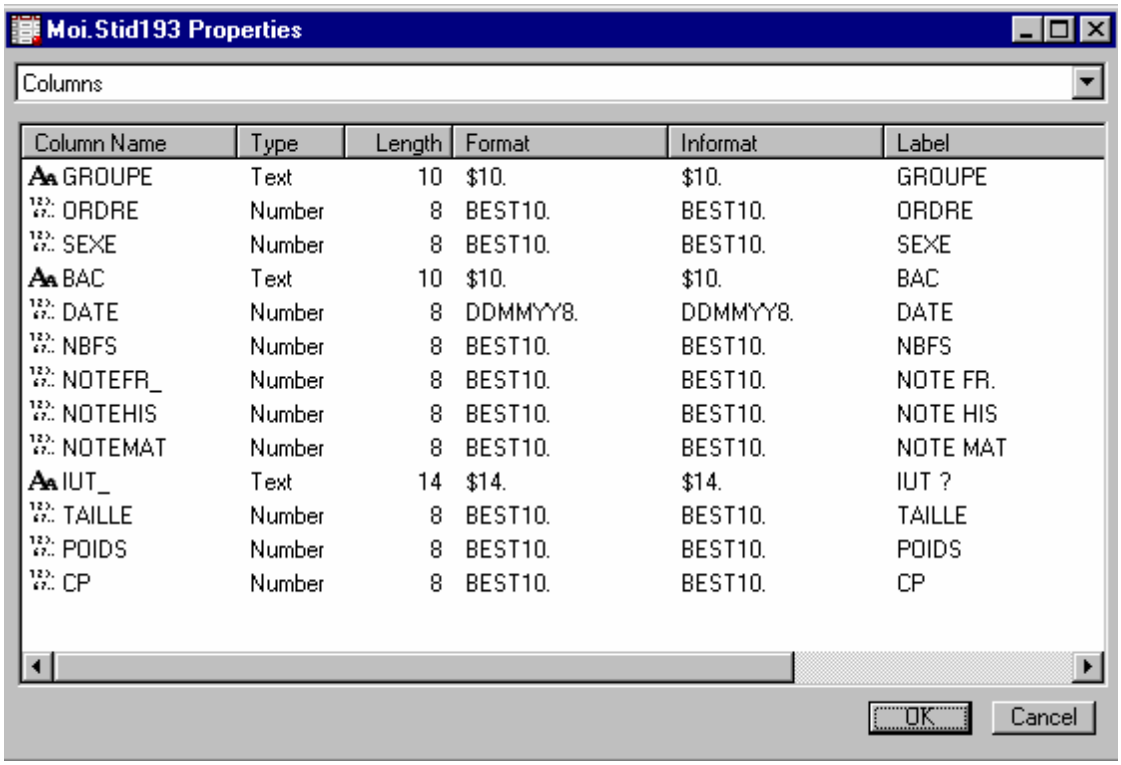
[30] Ce programme, utilisant la procédure IMPORT peut être utile lorsque vous avez un grand nombre de fichiers à importer…

4. Visualisation du fichier SAS

Il y a deux méthodes :

a) La fenêtre Explorer

Sélectionnez votre bibliothèque. Repérez le fichier STID193. Cliquez sur le bouton droit de la souris (menu contextuel) et choisissez View Columns[31] :

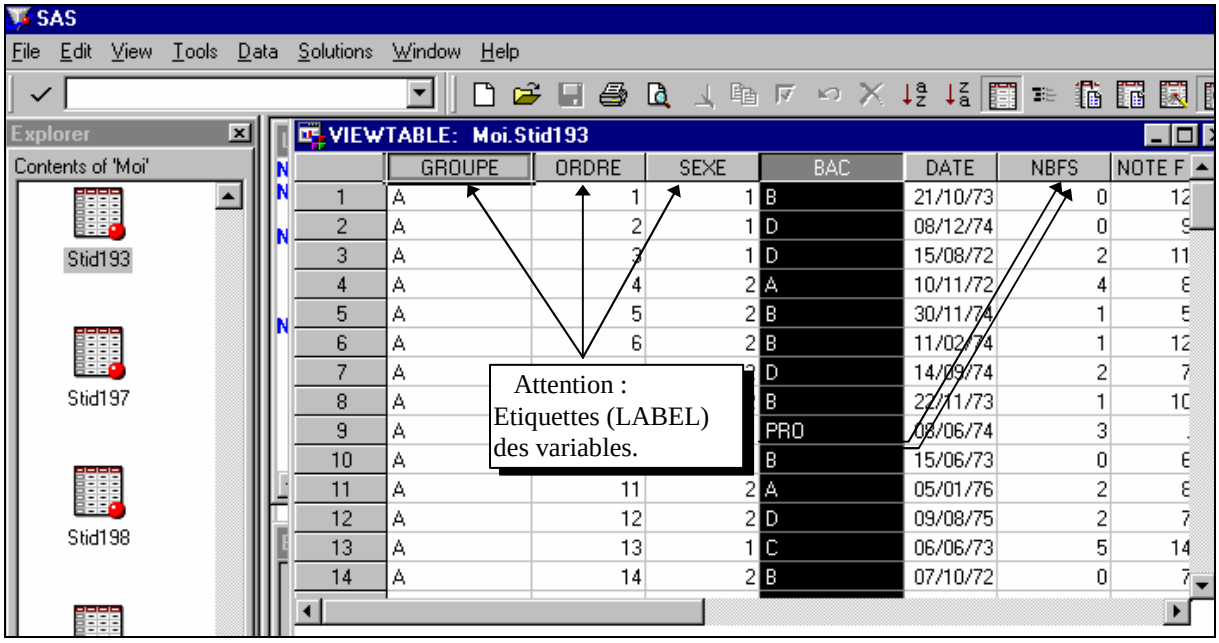


Column name	C'est le nom de la variable qui sera utilisé dans les procédures les étapes DATA etc. Son type (numérique ou caractère) est symbolisé par une icône. Longueur maxi : 32 caractères.
Length	C'est la longueur (en bytes) $ variable caractère 8 variable numérique
Format	C'est le format d'affichage de la variable. Remarquez le format de la date de naissance.
Informat	C'est le format de lecture, utile si vous importez des données. Toujours pour la date, remarquez que vous ne pouvez importer que des dates en format ddmmyy8.
Label	Etiquette de la variable **à ne pas confondre avec le nom de la variable.** Longueur maxi 256 caractères.

[31] Sivous importez unfichier EXCEL97, vous aurez peut être une différence au niveau de la date. SAS va lire une DATETIME (date heure) dont il faudra tenir compte plus tard.

Pour visualiser votre fichier, choisissez OPEN , vous visualisez alors votre fichier [32]:

Grâce au menu DATA, vous pouvez effectuer des recherche (Where), des tris (Sort) etc... Nous sommes en fait ici dans le module SAS/FSP qui permet d'effectuer des manipulations intéractives sur les fichiers de données.

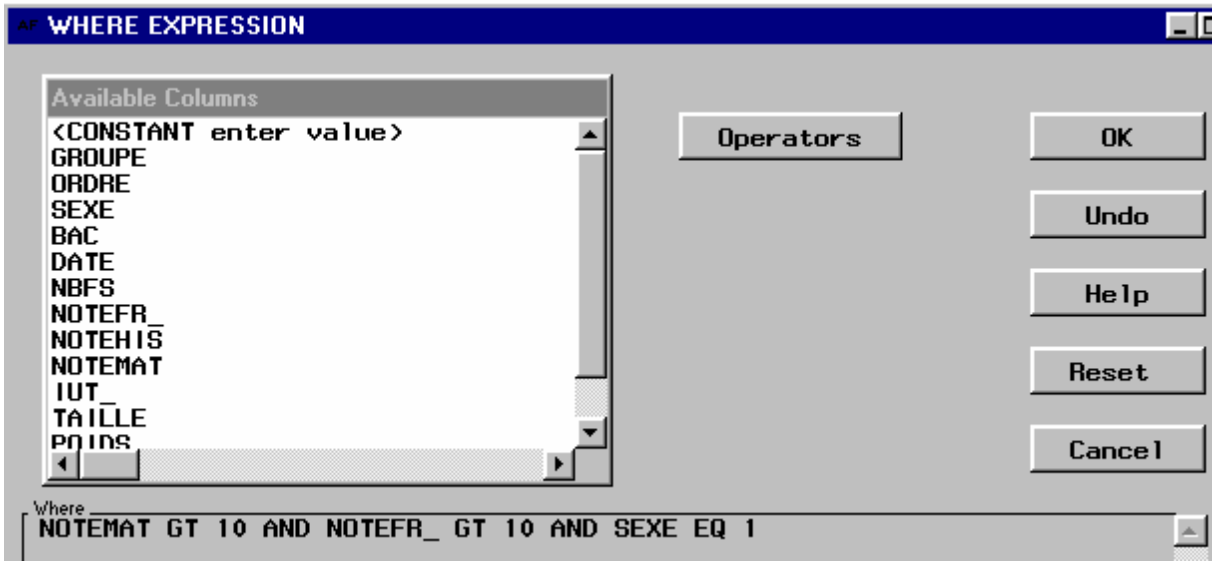


[32] Si les variables de votre fichier possèdent des étiquettes (label), ce sont les labels qui sont en tête de colonne.

Comment sélectionner une partie d'un fichier ?

Cherchons par exemple les individus de STID193 masculins ayant plus de 10 en maths et en français :

Allez dans DATA/WHERE



et tapez la close précédente en cliquant successivement sur les nom de variables et sur les operateurs (AND, OR, GT (=Greater Than plus grand que,), EQ (Equal, Egal), LT (Less Than, plus petit que), GE (great or equal = supérieur ou égal), LE (Less or equal = inferieur ou égal), NE (Non equal= différent).[33]

Validez en cliquant sur OK.

Vous obtenez les 11 individus répondant à la question :

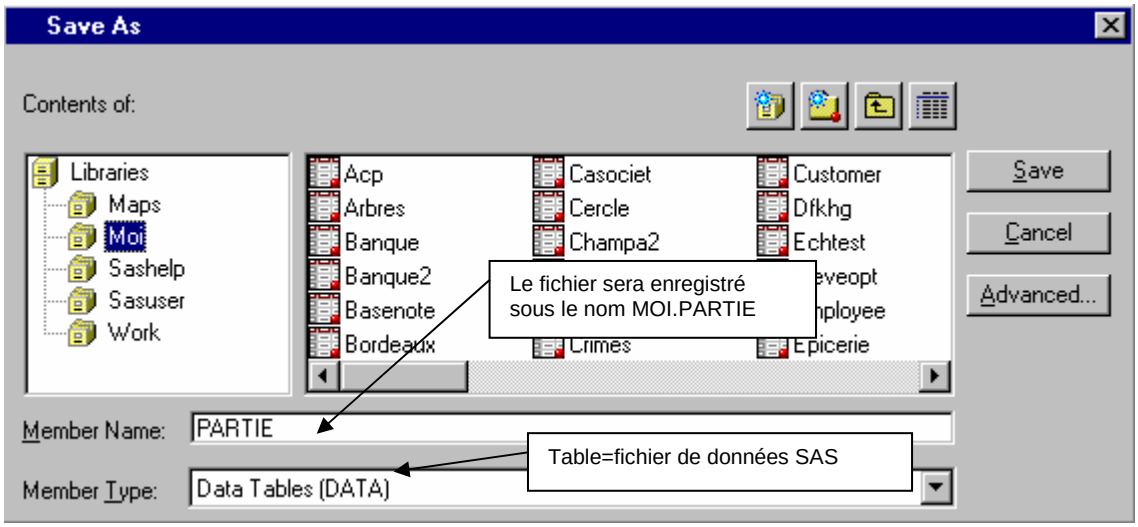
VIEWTABLE: MOI.STID19 Enter a Command

	GROUPE	ORDRE	SEXE	BAC	DATE
1	A	1	1	B	21/10/73
3	A	3	1	D	15/08/72
13	A	13	1	C	06/06/73
16	A	16	1	C	03/05/73
56	C	4	1	C	12/05/74
59	C	7	1	B	01/01/73
63	C	11	1	B	06/10/72
72	C	20	1	B	16/06/71
78	C	26	1	D	18/09/74
92	D	13	1	D	10/10/72
97	D	18	1	C	13/08/74

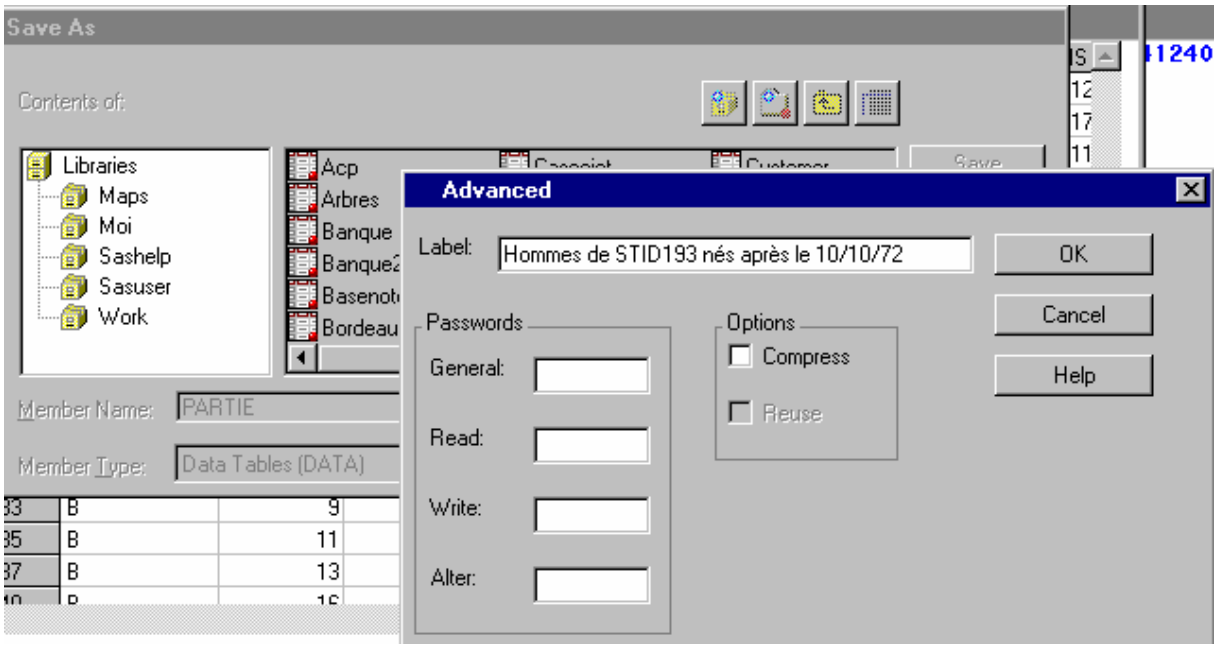
a) Cherchez les individus masculins nés après le 10 octobre 1972.

[33] Pour une constante, cliquez sur le champ CONSTANT Enter Value, puis entrez la valeur. Remarquez bien que la valeur peut être une date. '10JAN69:000:00:00'DT est une constante valide pour SAS. (Le DT sert à SAS pour identifier une date-heure : DATETIME. D pour une date seule (DATE) ; T pour une heure seule(TIME).

Vous pouvez sauvegarder le résultat de la requête dans un autre fichier de données SAS. Grâce à File/ Save As :



En cliquant sur Advanced, vous pouvez mettre un LABEL explicitant ce que contient votre fichier. Vous pouvez aussi protéger votre fichier de données en lecture, écriture etc…



b) Cherchez les individus des groupes B et C ayant une note strictement supérieure à 10 dans les trois matières et nés le 1/1/73 ou après [34]

34 Attention à la spécification de la date. Voir la note précédente.

b) Visualisation d'un fichier de données dans l'OUTPUT

C'est une méthode beaucoup plus rudimentaire à n'utiliser que pour de petits fichiers et pour cas de force majeure !

Vous pouvez donc visualiser, dans la fenêtre OUTPUT, un fichier de données en tapant dans le program Editor :

```
PROC PRINT DATA=MOI.STID193 ; RUN ;
```

Voir la PROC PRINT dans ce document pour avoir plus de détails sur sa syntaxe.

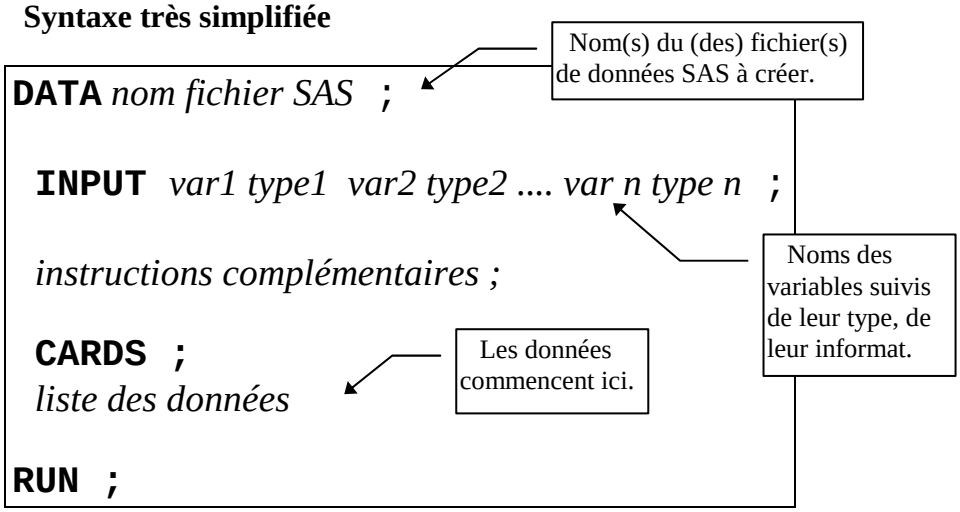
C. Fichier de données créé dans un programme SAS : étape DATA

L'étape DATA est l'étape quasi incontournable en langage SAS pour créer ou modifier un fichier de données[35]. Nous allons distinguer 3 types d'étape DATA :

1. Données incluses dans le programme. (CARDS)

Nous l'avons déjà rencontrée avec mon premier programme SAS :

a) Syntaxe très simplifiée



Le type contient les informations sur le type de la variable ($ si elle est de type texte, rien sinon), sur son format, sa longueur, sa position etc.

[35] Il y a aussi la méthode EXPLORER (File/New) plus graphique, ou des procédures qui permettent de créer un fichier de données, mais ceux-ci sont « figés » : procédure CORR rendant les coefficients de corrélations, pocédure REG les résidus etc. Seule la procédure SQL permet de créer des fichiers de données avec souplesse. Notons aussi l'ODS qui permet de réorienter toutes les sorties dans des fichiers de données SAS.

Exemples :

```
DATA work.donnee;                    Nous allons créer un nouveau fichier donnee dans work.

input groupe $ taille poids;         Le « $ » indique à SAS que groupe est une variable texte.

cards;                               Les données suivent...
A 175 85
B 167.5 75
B 190 92
C . .
C 165 57
A 166 49.5
;                                    Le ; marque la fin du jeu de données.
RUN ;                                L'étape DATA se termine par un RUN.

proc print data=work.donnee;
run;
```

Dans le programme suivant, nous introduisons les noms des individus.

Testez ce programme, quel est son inconvénient pour la variable nom ?

```
DATA work.donnee;
input nom $ groupe $ taille poids;
cards;
Jean-philippe     A 175 85
Claude             B 167.5 75
Marie-christine  B 190 92
Eric               C . .
Carmino          C 165 57
Etienne          A 166 49.5
;
run ;
Proc Print data=work.donnee ; run ;
```

Les deux points indiquent que les données sont manquantes pour la taille et le poids.

Pour y remédier, nous allons introduire un informat. Nous allons demander à SAS de lire 15 caractères et non pas 8. L'informat se nomme **$15.**

Remplacer donc l'input par : **`input nom $15. groupe $ taille poids;`**

b) Lecture et restitution de dates (Informat et Format)

De façon générale, lorsque SAS lit des données « spéciales » il faut lui préciser un format d'entrée (nommé informat) derrière le nom de la variable dans l'instruction INPUT. Vous pouvez ensuite déclarer un format de sortie (format) qui est en général lié à l'informat.

La liste des formats et informats courants figurent en annexe.[36]

Exemple

Nous allons lire la variable date d'entrée pour les personnes précédentes.

Si nous voulons lire des dates, il va falloir préciser à SAS leur « forme d'entrée » c'est à dire leur informat. En effet, il existe plusieurs façons d'écrire une date : 08/04/1997 ; 08APR97 ; Thu, April 8, 1997 etc...

Tapez le programme suivant :

```
DATA work.donnee;
input nom $15. dat_entr date8.;   Date8. est le format des dates ci-dessous cf.annexe
cards;
Jean-philippe   08JAN89           Voici les données.
Claude          05FEB88
Marie-christine 02MAR90
Eric            31DEC95
Carmino         12APR75
Etienne         10JUL85
;
run ;
proc print data=work.donnee ; run ;
```

Vous obtenez l'affichage suivant :

OBS	NOM	DAT_ENTR
1	Jean-philippe	10600
2	Claude	10262
3	Marie-christine	11018
4	Eric	13148
5	Carmino	5580
6	Etienne	9322

Contrairement à ce qu'il semble, SAS a bien lu les dates (elles sont codées en interne sous forme de nombre)[37].

Nous allons maintenant demander à SAS de les afficher convenablement en donnant un format d'affichage.[38]

---

[36] Il vous est également possible de définir vos propres formats et informats en uitlisant PROC FORMAT.

[37] Le nombre obtenu est le nombre de jours entre le 1/1/1960 et la date en question. Dans notre exemple, il y a donc 10600 jours entre le 8/1/89 et le 1/1/1960 !

[38] Notons que ce format ne change rien à la représentation interne de la date. Elle sera toujours codée sous forme de nombre. Seule son apparence changera.

```
DATA work.donnee;
input nom $15. dat_entr date8.;
format dat_entr date8.;
cards;
Jean-philippe   08JAN89
Claude          05FEB88
Marie-christine 02MAR90
Eric            31DEC95
Carmino         12APR75
Etienne         10JUL85
;
run ;
proc print data=work.donnee ; run ;
```

Nous conservons le même format pour l'affichage

Nous obtenons :

OBS	NOM	DAT_ENTR
1	Jean-philippe	08JAN89
2	Claude	05FEB88
3	Marie-christine	02MAR90
4	Eric	31DEC95
5	Carmino	12APR75
6	Etienne	10JUL85

Exercice

Modifiez le programme précédent pour obtenir l'affichage suivant dans la fenêtre OUTPUT (date à la française)

OBS	NOM	DAT_ENTR
1	Jean-philippe	08/01/89
2	Claude	05/02/88
3	Marie-christine	02/03/90
4	Eric	31/12/95
5	Carmino	12/04/75
6	Etienne	10/07/85

Modifiez ce programme pour afficher le jour (de la semaine) de la date d'entrée de chaque personne. (on pourra choisir un format adapté cf. annexe)

c) Instructions supplémentaires

Vous pouvez ajouter des instructions dans une étape data de manière à calculer de nouvelles variables à partir de variables existantes.

Exemple :

Nous voulons calculer la date de sortie (dat_sort) des individus précédents sachant qu'ils restent exactement 900 jours sur place.

Nous ajoutons les deux lignes (en gras) au programme :

```
DATA work.donnee;
input nom $15. dat_entr date8.;
format dat_entr date8.;
dat_sort=dat_entr+900;
format dat_sort date8.;

cards;
Jean-philippe   08JAN89
Claude          05FEB88
Marie-christine 02MAR90
Eric            31DEC95
Carmino         12APR75
Etienne         10JUL85
;
run ;
proc print data=work.donnee;
run;
```

Nous obtenons :

OBS	NOM	DAT_ENTR	DAT_SORT
1	Jean-philippe	08JAN89	27JUN91
2	Claude	05FEB88	24JUL90
3	Marie-christine	02MAR90	18AUG92
4	Eric	31DEC95	18JUN98
5	Carmino	12APR75	28SEP77
6	Etienne	10JUL85	27DEC87

Génial non ?

Exercice

En utilisant la fonction TODAY( ) qui donne la date courante, calculez l'ancienneté[39] en jour, puis en années des individus en ne tenant compte que de la date d'entrée.

Calculez les statistiques élémentaires sur cette variable (Proc Means)

[39] Différence entre la date d'entrée et la date courante.

2. Utilisation de fichiers de données SAS existants : Instruction SET

Dans ce paragraphe, vous apprendrez à créer de nouveaux fichiers à partir de fichiers existants, à ajouter des variables, recoder des variables etc...en utilisant l'étape DATA du langage SAS[40].

DATA *nom(s) fichier(s) SAS (options);*

SET *fichierSAS$_1$(options$_1$)…fichierSAS$_n$(options$_n$)[options point=nomvariable nobs=nomvariable end=nomvariable…];*

instructions complémentaires (IF, KEEP, DROP… ;

RUN ;

L'instruction SET ci-dessus permet de spécifier le (ou les) fichier de données SAS, éventuellement assortis d'options[41], à partir duquel on va en construire un autre. Nous allons retrouver les keep, drop... que vous venez de voir mais sous forme d'instructions et non plus d'options.

Les options de l'instruction SET (END=, POINT= , NOBS=) sont décrites un peu plus loin dans ce paragraphe.

[40] Notez aussi que la procédure DATASETS permet d'effectuer directement des modifications sur le fichier d'origine (changement de nom, de format etc... des variables d'origine)

[41] (WHERE= KEEP= etc. permettant de sélectionner certaines variables ou certains individus d'un fichier.

a) Copie d'un fichier SAS existant

(1) Copie totale en utilisant l'étape DATA

```
LIBNAME MOI 'Z:\' ;
LIBNAME PUB 'P:\LOGICIEL' ;
DATA MOI.ACP;
  SET PUB.ACP;
RUN;
```

Ces instructions permettent la création d'un fichierACP qui est la copie conforme du fichier ACP de la bibliothèque PUB. (très utile pour copier un fichier du répertoire public sur le votre)

Remarque importante: Ce n'est pas la façon la plus rapide de copier deux fichiers ! Vous pouvez, tout simplement, faire un copier coller entre les deux bibliothèques dans l'explorer !
Si vous utilisez le langage SAS la PROC COPY permet de faire ce travail.

(2) Copie partielle

C'est le réel intérêt de l'étape DATA pour « copier » des fichiers.

Nous allons retrouver ici certains termes (KEEP, DROP...) qui sont cette fois des instructions et non plus des options attachées qux fichiers de données (Voir page 64).

en ne conservant que certaines variables (KEEP)

```
DATA WORK.NOTSTID;            le fichier notstid est créé...
  SET MOI.STID193;            à partir du fichier STID193 de la bibliothèque MOI...
  KEEP NOTEMAT NOTEHIS NOTEFR_; en ne conservant que les variables notes.
RUN;
```

Remarque : Le programme :

```
DATA WORK.NOTSTID;
 SET MOI.STID193 (KEEP=NOTEFR_ NOTEHIS NOTEMAT);
RUN;
```

ferait la même chose.

en ne supprimant que certaines variables (DROP)

```
DATA work.saufsexe;
  SET MOI.STID193;
  DROP SEXE;                  On enlève la variable sexe.
RUN;
```

en conservant certaines observations (IF THEN, WHERE, OBS, FIRSTOBS)

```
DATA WORK.HOMME;
  SET MOI.STID193;
  IF SEXE=1;                  on ne conserve que les individus dont SEXE=1
RUN;
```

On peut aussi utiliser les options (cf. plus loin)

```
DATA WORK.HOMME;
  SET MOI.STID193 (where=(sexe=1));
RUN;
```

```
DATA WORK.BONMAT;
  SET MOI.STID193;
  KEEP GROUPE NOTEMAT ;
  IF NOTEMAT>=10 AND GROUPE='A';      on ne conserve que les gens du groupe A
                                      ayant la moyenne en math.
RUN;
```

Remarque : Ce IF est différent de ceux que vous avez l'habitude de voir en INFO. Il permet de faire des sélections sur des individus.

On a aussi en utilisant l'option **WHERE=** : (cf. Plus loin)

```
DATA WORK.HOMMAT;
SET MOI.STID193 (KEEP=GROUPE NOTEMAT WHERE=(GROUPE='A'
AND NOTEMAT>10));
RUN;
```

Remarque (rappel) : L'option WHERE ne peut pas être utilisée avec OBS et FIRSTOBS suivantes.

```
DATA WORK.PARTIE;
 SET MOI.STID193 (OBS=15);
RUN;
DATA WORK.EXTRAIT;
 SET MOI.STID193 (FIRSTOBS=100 OBS=106);
RUN;
```

on ne conserve que les 15 premiers individus

on ne conserve que les individus du 100ème au 106éme .

en éliminant certaines observation (delete)

```
DATA WORK.LESBONS;
  SET MOI.STID193;
  IF MEAN(NOTEFR_,NOTEHIS,NOTEMAT)<12 THEN DELETE;
RUN;
```

On élimine les gens ayant moins de 12 de moyenne.

```
DATA WORK.PRESENT;
SET MOI.STID193;
IF NMISS(NOTEFR_,NOTEHIS,NOTEMAT)>0 THEN DELETE;
RUN;
```

Cette fois on enlève tout ceux ayant au moins une note manquante...

Remarque:

Pour plus d'information sur les fonctions NMISS, MEAN consultez l'annexe (Opérateurs et fonctions).

(3) Copies multiples sur des fichiers différents OUTPUT

Il est possible de créer plusieurs fichiers à la fois en les spécifiant derrière l'instruction DATA.

L'instruction OUTPUT nomdefichier permettra ensuite d'affecter les observations dans les fichiers choisis.

```
data work.homme work.femme;
 set moi.stid193;
 if sexe=1 then output work.homme;
 if sexe=2 then output work.femme;
run;
```

Génial non ?

Exercices

- A partir du fichier STID193 importé préalablement, créez un fichier temporaire ne contenant que les gens ayant la moyenne dans les trois matières;

- Créez un fichier WORK.HOM ne contenant que les hommes de STID avec les variables taille, poids et sexe. Faites de même un fichier WORK.FEM.

- Créez trois fichiers de données ENFANT1, ENFANT2, ENFANT3 contenant les individus de STID ayant respectivement 1, 2 ou 3 frères et sœurs (variable NBFS). On effectuera ce travail en une seule étape DATA.

- Toto veut exécuter le programme suivant.

```
data essai;
 set pub.stid193;
 jour=date;
 format jour downame10.;
 if jour='Sunday';
run;

proc print data=essai(obs=10);
var groupe ordre jour;
run;
```

Il s'étonne car il ne fonctionne pas. Identifiez l'erreur de TOTO sur la notion de Format et apportez une solution. On pourra utiliser la fonction SAS WEEKDAY()[42]

[42] Attention toutefois,car la fonction WEEKDAY ne fonctionne qu'avec des variables DATE. Si vous avez une variable de type DATETIME (c'est le cas si vous avez importé votre fichier depuis EXCEL 97-2000), vous devez en extraire la date grâce à la fonction DATEPART.

b) Créations de variables, modifications, tableaux de variables

(1) Création (à partir des variables existantes)

La syntaxe est très simple, il suffit de déclarer le nom de la variable = à sa définition :

```
Data moi.stid193 ;
 Set moi.stid193 ;
 Taille_metre=taille/100 ;
Run ;
```

Ce programme crée la variable taille_metre qui est la taille en mètre (taille/100). Cette variable est ajoutée au fichier existant.

```
Data moi.stid193 ;
 Set moi.stid193 ;
NOTEMAX=MAX(NOTEFR_,NOTEHIS,NOTEMAT) ;
NOTEMIN=MIN(NOTEFR_,NOTEHIS,NOTEMAT) ;
Run ;
```

Ce programme va créer les variables notemax, notemin qui sont le max et le min pour chaque individu de leurs trois notes obtenues au bac. Ces variables seront ajoutées au même fichier MOI.STID193 ;

Remarque : Si les variables sont indicées, on peut spécifier cela autrement :

```
PMAX=MAX(POIDS1,POIDS2,POIDS3,POIDS4);
```

Peut être remplacé par : **`PMAX=MAX(OF POIDS1-POIDS4);`**

Très utile lorsqu'on a un grand nombre de variables.

Création d'une variable de type caractère ; instruction LENGTH

```
data work.essai;
*on va créer une variable identif de 4 caractères ;
length identif $4.;
set moi.stid193;

identif=put(groupe,$1.)!!put(ordre,2.0);
keep identif groupe ordre sexe taille poids;
run;
```

- Exécutez cet exemple et visualisez le résultat en faisant un VT work.essai dans la ligne de commandes.

- Comment est créée la variable Identif ?

(2) Modification

Supposons qu'un prof de maths décide de mettre 0 à ceux qui n'ont pas de notes de maths dans le fichier STID193. Nous allons modifier en conséquence les notes de maths...

Nous avons alors :

```
DATA WORK.COPY;
 SET MOI.STID193;
 IF NOTEMAT=. THEN NOTEMAT=0;
RUN;
```

Création d'un fichier temporaire copy. Nous ne voulons pas altérer l'original !

Si notemat est manquant alors notemat=0.

Plusieurs variables...

Supposons que les professeurs décident d'étendre cette manipulation à toutes les autres notes !

Il est possible de répéter le programme précédent trois fois ; toutefois comme la modification est identique, nous allons regrouper les trois variables dans un tableau.

(3) Tableau de variables

(à passer en première lecture)

Un tableau de variables est un mot clé avec un indice qui remplace un ensemble de variables, en général de même type[43]. Le tableau vous permet d'effectuer d'un seul coup un même traitement à vos variables en utilisant leur nouveau nom dans une boucle par exemple :

```
DATA WORK.COPY (DROP=I);
 SET MOI.STID193;
 ARRAY NOTES{3} NOTEFR_ NOTEHIS NOTEMAT;
 DO I=1 TO 3 ;
  IF NOTES{I}=. THEN NOTES{I}=0;
 END;
RUN;
```

on ôte la variable i du fichier copy (c'est une variable temporaire utilisée dans les calculs)

On crée le tableau notes qui a 3 variables notefr_ notehis et notemat. Notes{1} désigne notefr_ etc...

Dans cet exemple : Notes{1} est la variable NOTEFR_, Notes {2} la variable NOTEHIS etc.

Exercice : Le fichier de données ACP contient les températures annuelles de quinze villes en °C. Mettez ces températures en °F sachant que °F=1.8*°C+32.

[43] Attention, aucune nouvelle variable n'est créée. Il ne s'agit que d'un changement provisoire de nom pour alléger les algorithmes de calcul.

c) Changement d'étiquette, de nom, de format d'une variable

(à passer en première lecture)

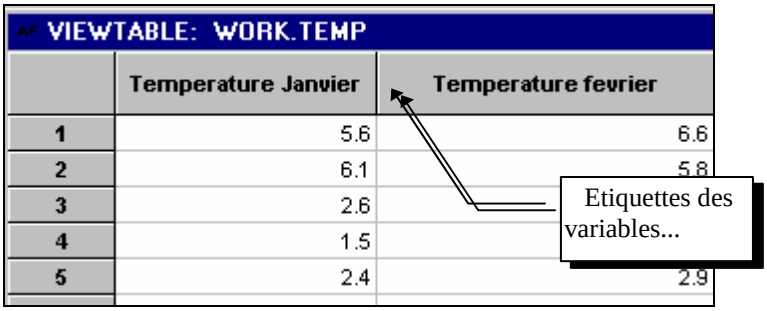
(1) Changement d'étiquette

LABEL nom de variable= 'Etiquette' ;

Cette instruction permet d'affecter des étiquettes à des variables pour avoir des sorties plus lisibles :

```
DATA WORK.TEMP;
  SET PUB.ACP;
  KEEP JAN FEV;
  LABEL JAN='TEMPERATURE JANVIER' FEV='TEMPERATURE
FEVRIER';
RUN;
```

Si vous demandez la visualisation du fichier, vous obtiendrez **:**



(2) Changement du nom d'une variable :

RENAME nom de variable à renommer(une ou plusieurs)=nouveau nom

```
LIBNAME PUB 'I :\STID9799\PUBLIC\LOGICIEL' ;
DATA WORK.TEMP;
  SET PUB.ACP;
  KEEP JAN FEV;                    On ne conserve que la t° de janvier et de février
  RENAME JAN=JANVIER FEV=FEVRIER;  La variable jan devient Janvier etc.
RUN;
```

Remarque (complément) :

Ici, nous avons recréé un fichier (TEMP) ce qui peut être couteux en temps d'exécution. Il est possible de passer par la procédure DATASETS pour effectuer ce travail :

(**Attention,** ce programme modifiera définitivement le nom de la variable NOTEFR_ ; si vous l'exécutez souvenez vous du nouveau nom ! ! !)

```
PROC DATASETS LIBRARY=MOI;
  MODIFY STID193;
  RENAME NOTEFR_=FRANCAIS;
RUN;
QUIT;
```

Ici nous renommons la variable NOTEFR_ en Français directement sur le fichier de départ.

Remarque : Pour visualiser la modification allez dans Global/Access/Display Libraries sélectionnez le fichier puis dans le menu contextuel (clic droit) choisissez la VAR Window.

Pour plus de détails sur DATASETS reportez vous au paragraphe « La procédure DATASETS » de ce document.

(3) Changement du FORMAT d'une variable

L'instruction FORMAT permet de changer le format d'affichage des variables. Ils suffit de spécifier le nom de la variable et son nouveau format. Vous pouvez l'utiliser dans une étape DATA ou dans la procédure DATASETS comme le montre l'exemple ci dessous...

Exemple :

```
PROC DATASETS LIBRARY=MOI;
  MODIFY STID193;
  FORMAT NOTEHIS NOTEMAT 4.1 TAILLE 6.2;
RUN;
QUIT;
```

Dans cet exemple, les notes de Math et Histoire-géo auront un format 4.1 et la taille un format 6.2.[44].

Pour visualiser le résultat faites un PROC PRINT et vous obtenez:

OBS	NOTEHIS	NOTEMAT	TAILLE
103	12.0	17.0	165.00
104	6.0	12.0	160.00
105	12.0	18.0	167.00
106	6.0	15.0	168.00

Ici, on voit que la variable taille est codée sur 6 caractères dont 2 décimales.

[44] Le premier chiffre indique la taille maximale du nombre et le deuxième, le nombre de décimale(s). Pour avoir plus d'informations sur les formats disponibles, allez voir en annexe.

d) Mot-clés particuliers :

Lors de l'exécution d'une étape DATA, SAS génère des variables temporaires très utile pour des traitements particuliers.

(1) _N_ compteur de l'étape DATA

_N_ est une variable prédéfinie du type compteur de boucle. Elle peut permettre de repérer le numéro de l'observation en cours de lecture dans une étape DATA.

Voici un exemple d'utilisation :

Le fichier CASOCIET (fichier de données SAS, répertoire public/logiciel) contient le chiffre d'affaire annuel d'une société de 1971 à 1996. (Variable C1) :

OBS	C1
1	9050
2	9380
3	9378
4	9680
5	10100
6	10160
7	10469
8	10738
9	10910
10	11058
11	11016
...	

Nous souhaiterions, pour rendre le fichier plus lisible , créer une variable « Année » qui renvoie l'année associée au chiffre d'affaire.
Le programme suivant répond à la question :

```
DATA WORK.ESSAI;
 SET MOI.CASOCIET;
 ANNEE=_N_+1970;
RUN;

PROC PRINT DATA=WORK.ESSAI;
RUN;
```

Tapez-le et vérifiez.
Supposons que le chiffre d'affaire de CASOCIET corresponde au chiffre d'affaires des années 1945, 1947, 1949,1951 etc. Modifiez le programme pour l'adapter à cette situation.

(2) Variables instantanées

Méditez l'exemple suivant :

```
data work.groupe;
  set moi.stid193 ;
  if sexe=1 then hom+1;
  if sexe=2 then fem+1;
  total+1;
  keep groupe sexe taille poids hom fem total;
run ;

Proc print data=work.groupe (obs=10) ;
Run ;
```

- Comment sont construites les variables hom, fem et total ?

Une utilisation de ce qui précède va être faite dans l'exemple suivant.

e) Options de l'instruction SET

Ci-dessous, nous décrivons les options de l'instruction SET. Les mots clés : END= ; POINT= ; NOBS=

(1) END=

Ce mot clef se place derrière le SET et permet de créer une variable temporaire qui prendra la valeur VRAIE lorsque l'étape DATA aura lu toutes les observations.

Syntaxe ultra simplifiée

```
SET nomdefichier END=nomdevariable ;
```

```
data work.groupe;
  set moi.stid193 end=final ;
  if sexe=1 then hom+1;
  if sexe=2 then fem+1;
  total+1;
  keep groupe sexe taille poids hom fem ;

/*nous detectons la fin du fichier*/
  if final then do;
   put hom=;
   put fem=;
  end;
 run;
```

va donner dans la LOG :

```
HOM=46
FEM=60
```

(2) POINT=

L'option POINT= de l'instruction SET permet de sélectionner l'individu dont le numéro est dans la variable suivant POINT=

Dans l'exemple suivant, nous allons électionner un individu sur 10 dans le fichier STID193.

```
data unsurdix;
 do i=1 to 110 by 10 ;
 set moi.stid193 point=i;   Nous allons lire la ieme observation de ce fichier
 if _error_ then abort;     Si elle n'existe pas (dépassement du fichier) , la
                            variable automatique _ERROR_ vaut 1, nous
                            arrêtons ABORT.
 output;                    Nous inscrivons cette observation dans le fichier
 end;                       Nous passons à la valeur de i suivante
 stop;                      INDISPENSABLE : sinon on entre en boucle infinie.[45]
 run;
```

Attention : l'option POINT= ne peut s'utiliser avec BY, WHERE, WHERE=.

[45] En effet, pour sortir de l'étape data il faut « dépasser » la fin du fichier, comme ici nous ne pointons que sur des observations existantes, nous n'y arriverons jamais ; d'où le STOP pour arréter l'étape DATA quand la boucle est finie.

(3) NOBS=

Cette option de l'instruction SET crée une variable contenant le nombre total d'observations du fichier de données.

La valeur de cette variable est affectée lors de la compilation. Vous pouvez donc vous y référer avant l'instruction SET. Cette variable n'estpas disponible en dehors de l'étape DATA qui la contient.

Nous pouvons modifier le programme précédent de la sorte :

```
data unsurdix;
 do i=1 to dernier by 10 ;
 set moi.stid193 point=i nobs=dernier;
 output;
 end;
 stop;
 run;
```

Dernier vaudra 106. Nous n'avons plus besoin de la condition d'erreur du programme précédent car nous n'allons pas dépasser la fin du fichier.

(4) Option IN=

Crée une variable prenant la valeur 1 si l'observation vient du fichier ou figurait le IN et 0 sinon.

Voir un exemple dans la concaténation de fichiers.

f) Fusion de fichiers

Si vous devez augmenter fusionner des fichiers contenant les mêmes[46] variables sur des individus différents (par ex. STID93, STID94…STID99), utilisez la fusion verticale.

Si vous devez fusionner des fichiers contenant les mêmes individus mais sur des variables différentes (par ex. ventes sur les dernier trimestre : TOTOCT, TOTNOV,TOTDEC), utilisez la fusion horizontale (MERGE).

[46] Si des variables n'existent pas dans les deux fichiers, la colonne contiendra des manquants pour les individus en question.

(1) Fusion verticale de deux fichiers (augmente le nombre d'observations (ou d'individus))

FICHIER 1

OBS	X	Y
1	23	Jules
2	54	Toto
3	123	Prof

FICHIER 2

OBS	X	Y
1	678	COUCOU
2	787	truc

FICHIER CONCATENE 1+2

SET FICHIER1 FICHIER2

OBS	X	Y
1	23	Jules
2	54	Toto
3	123	Prof
4	678	COUCOU
5	787	truc

Si WORK.HOM contient les hommes de STID et WORK.FEM les femmes, vous pouvez reconstituer un fichier TOUT, concaténation des deux précédents, de la façon suivante:

```
DATA WORK.TOUT;
  SET WORK.HOM WORK.FEM;
RUN;
```

Ce fichier contiendra l'ensemble des individus de ces deux fichiers sur les variables correspondantes.

Remarque : Si une variable se trouve dans un des fichiers sans être dans le second, les individus seront portés manquants pour cette variable dans le second fichier.

On peut également faire figurer des options derrière les noms des fichiers de données :

```
data essai;
 set moi.stid193(where=(sexe=1)) moi.stid197 (where=(sexe=2));
 run;
```

- Que fait ce programme ?

Fusion avec l'utilisation de l'option IN=

IN=Variable est une option des instructions SET et MERGE permettant de savoir d'où vient l'observation lorsque l'on fusionne plusieurs fichiers de données :

```
data work.tous;
set moi.stid193 moi.stid197(in=x);
keep annee sexe taille poids;
annee=1993;
if x=1 then annee=1997;
output;
run;
```

Ici X prend la valeur 1 lorsque l'observation vient de STID197 .

- Que fait ce programme ?

(2) Fusion horizontale simple: L'instruction MERGE

Cette instruction permet de fusionner deux fichiers (en augmentant le nombre de variables cette fois). Elle suppose que les individus (lignes) SONT LES MEMES et DANS LE MEME ORDRE[47] !

Nous ne présentons ici qu'une version simple de cette instruction:

FICHIER 1

OBS	X	Y
1	23	Jules
2	54	Toto
3	123	Prof

FICHIER 2

OBS	Z	T
1	678	COUCOU
2	787	truc
3	89	Machin

FICHIER CONCATENE

`MERGE FICHIER1 FICHIER2`

OBS	X	Y	Z	T
1	23	Jules	678	COUCOU
2	54	Toto	787	truc
3	123	Prof	89	Machin

[47] Sinon votre fichier résultat ne sera plus cohérent. L'option BY permet d'effectuer cette fusion en se basant sur une ou plusieurs variables identifiant les individus.

Exemple :

Nous allons d'abord créer deux extraits complémentaires de STID193 (M1 et M2):

```
DATA WORK.M1;
        SET MOI.STID193;
        KEEP SEXE GROUPE;  Nous ne conservons que les variables sexe et groupe
RUN;
DATA WORK.M2;
        SET MOI.STID193;
        DROP SEXE GROUPE;  Nous prenons toutes les variables sauf sexe et groupe
RUN;
```

Puis, nous allons les fusionner pour retrouver le fichier original:

```
DATA WORK.TOUT;
 MERGE WORK.M1 WORK.M2;
RUN;
```

Théoriquement TOUT=STID193!

(3) Fusion horizontale sophistiquée (MERGE avec option BY)

Prenons les deux fichiers CHOL_AVR et CHOL_OCT contenant les taux de cholesterol de quelques individus au mois d'avril puis au mois d'octobre :

Fichier CHOL_AVR :

Analyses du mois d'avril

Obs	NUM_SECU	LDL_AVR
1	1660538898013	0.96
2	1770538351009	1.65
3	2761138010001	0.89
4	2781038351025	0.67
5	2890138351006	2.20

Fichier CHOL_OCT :

Analyses du mois d'octobre

Obs	NUM_SECU	LDL_OCT
1	1660538898013	0.90
2	1770538351009	1.77
3	2650238982002	1.65
4	2781038351025	0.67
5	2890138351006	2.00

Remarquez que les patients ne sont pas toujours les mêmes !

Nous souhaitons fusionner ces deux fichiers en un seul. Chaque ligne représentant un patient.

- Utilisez « bêtement » l'option MERGE.
- Quel est le problème ici ?

Nous allons dire à SAS de fusionner les deux fichiers par rapport aux individus :

```
data ensemble;
merge moi.chol_avr moi.chol_oct;
by num_secu;
run;
```

Attention : Pour que BY fonctionne, les fichiers doivent avoir été triés par rapport à la variable contenue dans le BY (ici num_secu). Si tel n'est pas le cas, utilisez PROC SORT. (Voir page 135)

Fichiers octobre et avril avec BY

Obs	NUM_SECU	LDL_AVR	LDL_OCT
1	1660538898013	0.96	0.90
2	1770538351009	1.65	1.77
3	2650238982002	.	1.65
4	2761138010001	0.89	.
5	2781038351025	0.67	0.67
6	2890138351006	2.20	2.00

Super non ?

Syntaxe simplifiée

```
DATA nom de fichier ;
  MERGE fichier1 fichier2 fichier3… ;
  BY Variable1 Variable2 … ;
RUN ;
```

Remarques :

- Les variables dans le BY doivent figurer dans TOUS les fichiers de données à concaténer.
- Les fichiers doivent être triés par rapport aux variables figurant dans le BY ou être indexés par rapport à ces variables.
- Il est possible de spécifier un ordre décroissant… Cf. Aide en Ligne.

(4) Exercices

- Créez trois fichiers (temporaires) SAS (hommes seuls, femmes seules, hommes et femmes) contenant les variables note de maths, d'histoire-géo et de français ainsi que leur moyenne, max et min pour chaque individu.

- Créez un autre fichier SAS contenant les individus féminins ayant 3 notes et dont la moyenne générale est supérieure ou égale à 12.

- Créez **un** fichier de données contenant le groupe, ordre, taille, poids, des hommes de tous les fichiers STID : stid193, 194 etc. Vous vous arrangerez pour avoir une variable année dans le fichier qui permet de savoir de quelle année était l'individu de STID. Idem avec les femmes.

- Sélectionnez les 4eme, 7eme, 10 eme, 13eme etc. individus des fichiers STID et mettez les dans un fichier en repérant le numéro de l'année, le groupe, l'ordre et le sexe.

- Exercice récapitulatif n°1 du paragraphe PROC PRINT de la page 136. (A faire lorsque vous aurez compris PROC SORT)

g) Recodage de variables (if, then, else, select when)

(1) if...then ...

syntaxe (dans le cas d'une instruction après le test)

```
IF condition THEN instruction;
IF condition THEN instruction; ELSE instruction;
```

syntaxe s'il y en a plusieurs:

```
IF condition THEN DO;
     instruction1;
     instruction2;
END;
```

Exemples:

```
IF AGE<10 THEN DO;
          TYPE='ENFANT';
          ECOLE='PRIMAIRE';
END;
```

Que fait le programme suivant ?

```
DATA WORK.STID;
SET MOI.STID193;
IF SERIEBAC IN ('C' 'D') THEN TYPE='SCIENTIFIQUE';
                         ELSE TYPE='AMATEUR';
KEEP SERIEBAC TYPE;
RUN;
```

Remarquez la présence du type liste dans SAS. (if seriebac **in**('C' 'D') etc...) qui est très pratique.

(2) Exercices:

- Créez un programme SAS créant un fichier temporaire contenant le fichier STID auquel on ajoute une nouvelle variable SEXEA qui vaut "homme" si SEXE=1 et "femme" si SEXE=2.

- Ecrivez un programme SAS créant la variable mention dans le fichier STID. La mention est « passable » si la moyenne est entre 10 et 12, « assez bien » entre 12 et 14, « bien » entre 14 et 16 et « très bien » entre 16 et 20.

- Créez un fichier de données extrait aléatoirement de STID193, ne contenant qu'une moitié des individus environ. On pourra utiliser la fonction RANUNI(0) qui donne une réalisation d'une variable aléatoire suivant une U[0,1].

(3) Select / When

Vous avez déjà vu cette fonction en informatique.

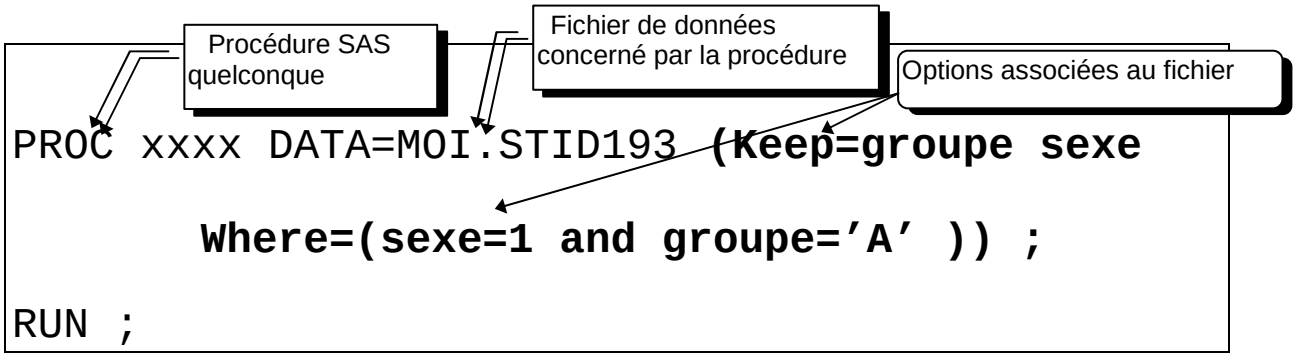
Un petit exemple vaut mieux qu'un long discours:

```
DATA WORK.STID;
 FORMAT TAILLEC $10. ;
 SET MOI.STID193;
 SELECT (SEXE);

  WHEN (1)
   IF TAILLE>190 THEN TAILLEC='GRAND';
   ELSE IF TAILLE >170 THEN TAILLEC='MOYEN';
   ELSE TAILLEC='PETIT';

  WHEN (2)
   IF TAILLE>180 THEN TAILLEC='GRANDE';
   ELSE IF TAILLE >160 THEN TAILLEC='MOYENNE';
   ELSE TAILLEC='PETITE';

 END;
KEEP TAILLE TAILLEC SEXE;
RUN;
```

Que fait ce programme ? Comment est définie taillec ?

D. Utilisation de données SAS dans les Procédures ou les étapes DATA)

Lorsque nous faisons agir une procédure SAS sur un fichier SAS, il est possible de limiter la portée de la procédure à une partie du fichier de données en utilisant des mots clés situés derrière le nom du fichier :



Ces options permettent de ne conserver qu'une partie du fichier de départ sans pour autant modifier celui-ci.

Dans l'exemple ci-dessus, la procédure xxxx ne s'appliquera qu'aux hommes du groupe A de STID193. Seules les variables Groupe et sexe sont conservées.

1. Sélection sur les variables

a) Suppression de variables : **DROP=**

DROP= *variables*

Exemple :

```
proc print data=moi.stid193(drop=notefr_ notehis notemat);
run;
```

Ici nous affichons STID193, dans la fenêtre OUTPUT, en enlevant les variables notes. L'instruction ci-dessous fait exactement le contraire.

b) Conservation de variables : **KEEP=**

Keep= *variables*

Exemple :

```
PROC PRINT DATA=MOI.STID193(KEEP=NOTEFR_ NOTEHIS
NOTEMAT);
RUN;
```

Ici nous imprimons STID193 en ne conservant que les variables NOTEMAT, NOTEHIS ET NOTEFR_.

c) Renommer des variables : **RENAME=**

rename=(*ancien_nom1=nouv_nom1 ancien-nom2=nouv_nom2...*)

Exemple :

```
PROC MEANS DATA=MOI.STID193(KEEP=NOTEMAT NOTEHIS
NOTEFR_ RENAME=(NOTEMAT=MATHS NOTEHIS=HISTOIRE
NOTEFR_=FRANCAIS));
RUN;
```

va donner

```
VARIABLE    N          MEAN       STD DEV       MINIMUM       MAXIMUM
  ------------------------------------------------------------------------
  FRANCAIS  105     8.8285714     2.3099165     4.0000000    14.0000000
  HISTOIRE   98    10.6020408     2.8420255     5.0000000    17.0000000
  MATHS     104    12.5144231     3.2226871     5.0000000    19.0000000
  ------------------------------------------------------------------------
```

2. Sélection d'individus

a) Sélection d'individus par leur n : `FIRSTOBS= OBS=`

`FIRSTOBS=` *n* `OBS=`*p*

SAS ne conserve que les individus compris entre le n^{ieme} et le p^{ieme}. On peut utiliser ces deux options séparément.

Exemple :

```
PROC MEANS DATA=MOI.STID193(KEEP=NOTEFR_ NOTEHIS
NOTEMAT FIRSTOBS=10 OBS=25);
RUN;
```

Ici nous calculons quelques statistiques sur STID193 en ne conservant que les variables notemat, notehis et notefr_ et 16 individus (entre le 10^{eme} et le 25^{eme})

b) Sélection d'individus par une condition : `WHERE=`

`WHERE=`*Condition*
Seuls les individus remplissant la condition seront sélectionnés. Cette option est extrêmement riche et donc importante à connaître.

Remarque : L'option WHERE ne peut pas être utilisée avec OBS et FIRSTOBS.

c) Opérateurs < > = ...

```
PROC PRINT DATA=MOI.STID193
(WHERE=(NOTEMAT>10 AND NOTEFR_>=12));
RUN;
```

On ne sélectionne que les individus ayant plus de 10 en maths et plus de 12 (ou 12) en français. Il n'en reste plus beaucoup ! ! !

```
PROC PRINT DATA=MOI.STID193
 (WHERE=(SERIEBAC='C' OR NOTEMAT>16));
RUN;
```

Vous pouvez bien sûr utiliser les `AND, OR , NOT`,< (ou LT),= (ou EQ) ^= (ou NE) (différent) que vous connaissez bien.

d) Utilisation de Fonctions

Il est possible d'utiliser des fonctions dans les « WHERE ». Ici, nous utilisons la fonction MEAN qui calcule la moyenne arithmétique des variables entre parenthèses.[48]

```
PROC PRINT DATA=MOI.STID193
(KEEP=GROUPE BAC SEXE NOTEFR_ NOTEHIS NOTEMAT
WHERE=(MEAN(NOTEFR_,NOTEMAT,NOTEHIS)>12));
RUN;
```

que fait le programme précédent ?

et celui-ci ?

```
PROC PRINT DATA=MOI.STID193
(KEEP=GROUPE SEXE NOTEFR_ NOTEHIS NOTEMAT
WHERE=(NMISS(NOTEFR_,NOTEMAT,NOTEHIS)>0));
RUN;
```

e) Opérateur IS MISSING

Il permet de sélectionner les individus ayant une variable manquante (ou plusieurs).

```
PROC PRINT DATA=MOI.STID193
(WHERE=(NOTEMAT IS MISSING));
RUN;
```

SAS va afficher les individus n'ayant pas de notes de note de maths.

Dans l'exemple suivant, nous utilisons l'opérateur NOT pour prendre la négation.

```
PROC PRINT DATA=MOI.STID193
(WHERE=(NOTEMAT IS NOT MISSING AND NOTEFR_ >=12));
RUN;
```

[48] Attention à ne pas confondre la fonction MEAN avec la procédure MEANS. MEAN calcule une moyenne pour chaque individu et MEANS calcule la moyenne de la classe.

f) Opérateur CONTAINS

Cet opérateur (et le suivant) sont à utiliser avec les variables alphanumériques ou textes.

« Contains » permet de ne sélectionner que les individus dont la variable (texte) contient la chaîne spécifiée.

Prenons le fichier CUSTOMER (Répertoire public) il contient les données suivantes :
(Ce fichier est détaillé dans le paragraphe sur la procédure SQL que vous verrez plus tard)

OBS	CUSTNAME	CUSTNUM	CUSTCITY
1	Beach Land	16	Ocean City
2	Coast Shop	3	Myrtle Beach
3	Coast Shop	5	Myrtle Beach
4	Coast Shop	12	Virginia Beach
5	Coast Shop	14	Charleston
6	Del Mar	3	Folly Beach
7	Del Mar	8	Charleston
8	Del Mar	11	Charleston
9	New Waves	3	Ocean City
10	New Waves	6	Virginia Beach
11	Sea Sports	8	Charleston
12	Sea Sports	20	Virginia Beach
13	Surf Mart	101	Charleston
14	Surf Mart	118	Surfside
15	Surf Mart	127	Ocean Isle
16	Surf Mart	133	Charleston

Si nous voulons sélectionner les individus dont la ville contient « Beach », nous allons taper le programme suivant :

```
LIBNAME PUB ‘Z:\LOGICIEL’ ;
PROC PRINT DATA=PUB.CUSTOMER
 (WHERE=(CUSTCITY CONTAINS 'Beach'));
RUN;
```

Nous obtenons :

OBS	CUSTNAME	CUSTNUM	CUSTCITY
2	Coast Shop	3	Myrtle Beach
3	Coast Shop	5	Myrtle Beach
4	Coast Shop	12	Virginia Beach
6	Del Mar	3	Folly Beach
10	New Waves	6	Virginia Beach
12	Sea Sports	20	Virginia Beach

Génial non !

g) Opérateur Like

Vous sélectionnez les individus dont la variable (texte) est égale (ou à peu près !) à la chaîne spécifiée.

```
PROC PRINT DATA=PUB.CUSTOMER
 (WHERE=(CUSTCITY LIKE 'Ocean City'));
RUN;
```

Vous n'allez sélectionner que les individus dont la ville est *Ocean City.* (un « = » aurait fait la même chose)

OBS	CUSTNAME	CUSTNUM	CUSTCITY
1	Beach Land	16	Ocean City
9	New Waves	3	Ocean City

```
PROC PRINT DATA=PUB.CUSTOMER
 (WHERE=(CUSTCITY LIKE 'Ocean%'));
RUN;
```

Le caractère « % » remplace **toute chaine de caractères**. Nous allons donc sélectionner toutes les villes commençant par « Ocean ».

OBS	CUSTNAME	CUSTNUM	CUSTCITY
1	Beach Land	16	Ocean City
9	New Waves	3	Ocean City
15	Surf Mart	127	Ocean Isle

De même, **le caractère « _ »** remplace **un** caractère quelconque.

Remarque : Dans le cas de fichiers volumineux, il peut être intéressant de créer des index sur le fichier ce qui peut considérablement accélerer la recherche. (cf. La procédure SQL de ce document ou le Chap. 6 §« SAS indexes »du *SAS Language Reference*)

Exercices

A)

TOTO dit qu'il n'y a aucune différence entre ces deux programmes, qu'en pensez-vous ?

```
PROC PRINT DATA=MOI.STID193 (WHERE=((SERIEBAC='C' OR
SERIEBAC='D') AND SEXE=2) );
  RUN;

PROC PRINT DATA=MOI.STID193 (WHERE=( SERIEBAC='C' OR
SERIEBAC='D' AND SEXE=2) );
  RUN;
```

B)

1) Affichez les individus ayant un bac D ou B, n'ayant aucune note manquante et dont la moyenne des trois notes est supérieure à 13. Vous n'afficherez que le Bac, le Groupe, le Sexe et les notes de ces individus.
2) Affichez les hommes de STID193 des groupes A,B et C ayant une note manquante en histoire géo ou en français ou dans les deux.
3) Affichez les femmes du groupe A de tous les fichiers STID dont la taille est supérieure à 170cm. On spécifiera les années et les numéros d'ordre des femmes sélectionnées.
4) Affichez les individus de STID193 nés après le 14/8/1973 n'ayant aucune note manquante et dont la plus grande est supérieure à 14.
5) Affichez les individus de STID193 ayant connu l'IUT grâce à un ou une amie. Examinez pour cela le contenu de la variable IUT ?

III. L'ODS : Gestion des sorties SAS

Avant de lire ce chapitre, il faut avoir les notions des procédures SORT, PRINT, UNIVARIATE et TABULATE.

Les procédures précédentes PRINT, MEANS… envoient leur résultat dans la fenêtre OUTPUT.

SAS permet d'envoyer ces résultats aussi dans un fichier HTML[49] ou directement dans une table SAS. Le but de ce praragraphe est de vous montrer comment y parvenir.

[49] Ceci présente un double intérêt. Les fichiers HTML peuvent contenir des informations sous un format très sophistiqué tout en étant lus par un simple navigateur WEB. D'autre part, on peut copier coller des tableaux HTML directement sous EXCEL et réutiliser ces données facilement !

A. Quelques notions basiques sur l'HTML

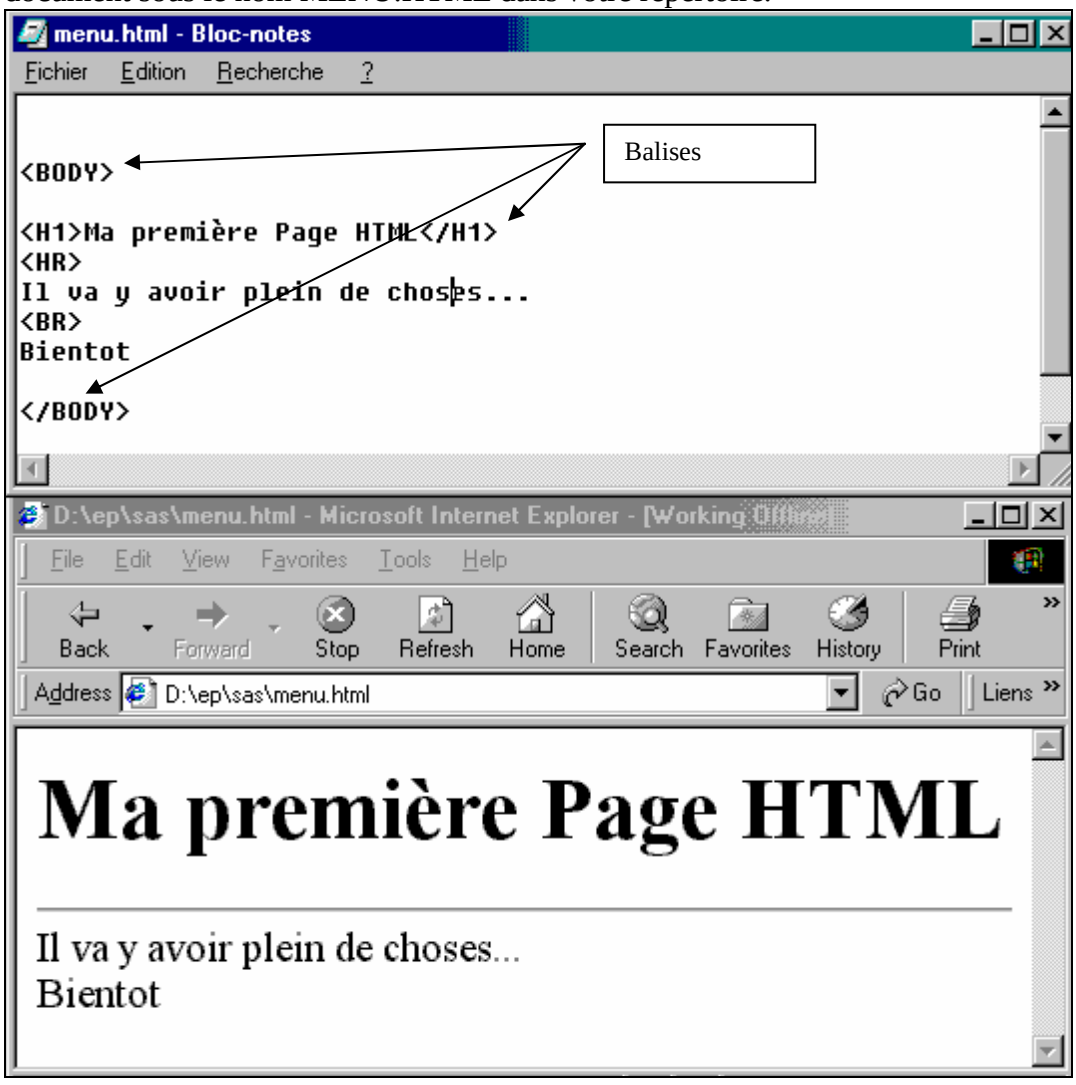
L'HTML est le langage des documents WEB. Il contient du texte, mais aussi des balises permettant de mettre en forme le texte, de pointer vers un autre document etc.

Vous pouvez construire un document HTML directement ou en utilisant un logiciel spécialisé comme FRONTPAGE.

WORD, EXCEL, SAS etc. permettent aussi de créer des documents HTML.

a) Un exemple

Tapez le texte suivant dans NOTEPAD (bloc-notes de Windows). Enregistrez le document sous le nom MENU.HTML dans votre répertoire.



Ouvrez ce document avec Internet Explorer par exemple.

Nous avons utilisé les balises suivantes :

Balise	Signification
<H1> </H1>	Caractères Grande taille. On peut remplacer le 1 par un nombre de 1 à 6 On peut aussi mettre des attributs ALIGN=CENTER pour centrer le texte
 	Retour à la ligne
<HR>	Tracé d'une ligne horizontale
<FONT FACE=…> </FONT>	Pour changer la police de caractère
<FONT COLOR=> </COLOR>	Pour changer la couleur de la police
<G> </G>	Mettre en Gras
<I> </I>	Mettre en Italique
<IMG SRC="nom fichier.gif" ALIGN=…>	Insere une image GIF et la centre si ALIGN=CENTER etc.

Exemple :

Le texte placé entre deux balises <H1> </H1> sera en grands caractères.

Essayez l'exemple suivant en remplaçant le D:\SASV801 par le répertoire SAS de votre ordinateur .

Pratique de l'écriture de code HTML : Vous laisserez NOTEPAD et INTERNET EXPLORER actifs. Vous basculerez de l'un a l'autre avec ALT+TAB. Faites Fichier/Enregistrer avec NOTEPAD et REFRESH avec internet explorer pour enregistrer et visualiser vos modifications.

```
<HEAD> <TITLE> Ma page de Menu  </TITLE> </HEAD>

<BODY BGCOLOR=YELLOW>

<H1 ALIGN=CENTER> <I> <FONT FACE='COMIC SANS MS'> <FONT
COLOR=RED>
Ma première Page HTML </FONT> </COLOR> </I></H1>
<IMG SRC="E:\sasv8\core\sasmisc\gfkids.gif" ALIGN=RIGHT>
<H2> C'est super non !!! </H2>
<HR>
Il va y avoir plein de choses. !..
<BR>
<B> Bientot </B>

</BODY>
```

b) D'autres exemples

Sur le Web vous trouverez des belles pages dont vous pourrez examiner la source (Clic droit, afficher la source)

Nous allons maintenant voir comment créer automatiquement des documents HTML avec SAS.

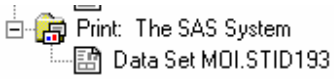
B. Utilisation de l'ODS de SAS. Objets de sortie

Certaines procédures envoient un d'autres plusieurs objets (ou section) dans la fenêtre OUTPUT.

Exemples :
La procédure PRINT ci-dessous ne va renvoyer qu'un seul objet :

```
Proc print data=moi.stid193 (obs=10) obs='Numéro' ;
Var date taille poids ;
Run ;
```

Va donner :

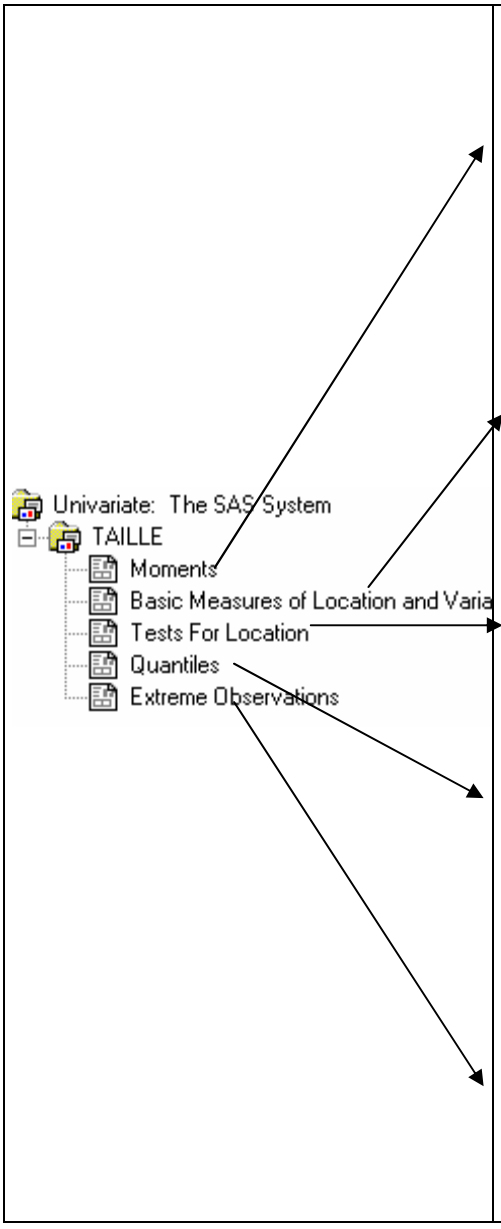


(Un seul objet : les 10 observations)

Numéro	DATE	TAILLE	POIDS
1	21/10/73	180	68
2	08/12/74	168	61
3	15/08/72	178	68
4	10/11/72	167	54
5	30/11/74	162	50
6	11/02/74	167	58
7	14/09/74	178	58
8	22/11/73	160	56
9	08/06/74	186	64
10	15/06/73	168	55

```
proc univariate data=moi.stid193;
var taille;
run;
```

Va donner 5 objets en sortie : Les moments (moyenne, écart type etc.), Les statistiques de base (de position et de dispersion), les tests de position, les quantiles et les valeurs extrêmes.



The UNIVARIATE Procedure
Variable: TAILLE (TAILLE)

Moments

N	106	Sum Weights	106
Mean	170.69	Sum Observations	18094
Std Deviation	7.839	Variance	61.4508535
Skewness	0.480	Kurtosis	-0.057081
Uncorrected SS	3095064	Corrected SS	6452.33962
Coeff Variation	4.592	Std Error Mean	0.76139676

Basic Statistical Measures

Location		Variability	
Mean	170.69	Std Deviation	7.83906
Median	170.00	Variance	61.45085
Mode	160.00	Range	41.00000
		Interquartile Range	10.00000

Tests for Location: Mu0=0

Test	-Statistic-	-----p Value------	
Student's t	t 224.1908	Pr > |t|	<.0001
Sign	M 53	Pr >= |M|	<.0001
Signed Rank	S 2835.5	Pr >= |S|	<.0001

Quantiles (Definition 5)

Quantile	Estimate	Quantile	Estimate
100% Max	196	50% Median	170
99%	187	25% Q1	165
95%	184	10%	160
90%	182	5%	160
75% Q3	175	1%	158
0% Min	155		

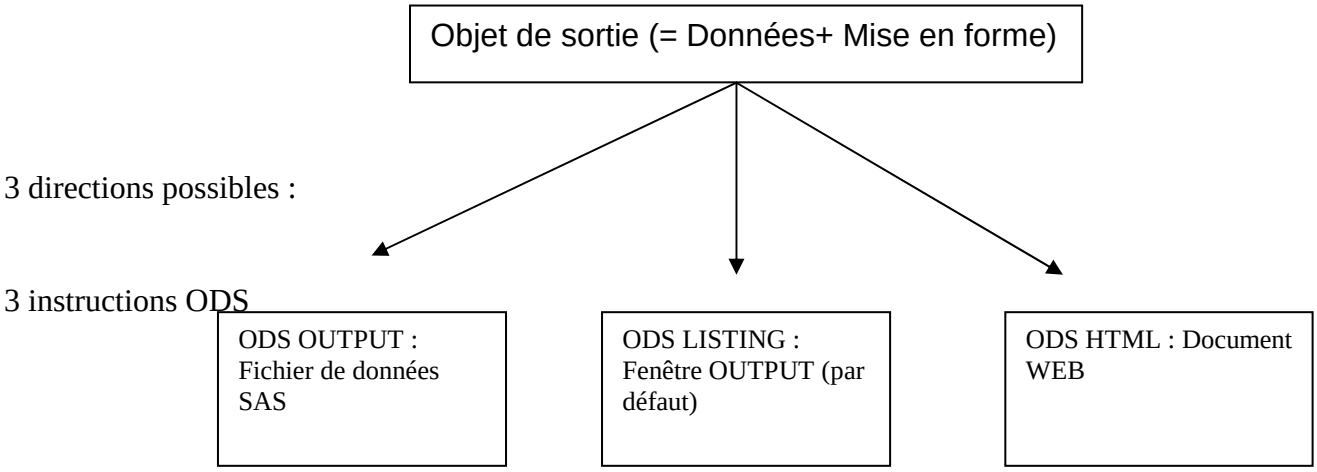
Extreme Observations

----Lowest----		----Highest---	
Value	Obs	Value	Obs
155	52	185	25
158	86	185	90
158	70	186	9
160	103	187	48
160	101	196	16

C. Trois sorties possibles

Chaque Objet de sortie se compose de données et de mise en forme appeléeTemplate. La mise en forme pourra être personnalisée.

L'instruction ODS permettra de diriger chaque Objet vers une sortie (ou plusieurs à la fois !) de notre choix :



☞ Ne confondez pas la fenêtre OUTPUT et l'ODS OUTPUT qui n'ont rien à voir…

1. Sortie HTML basique

L'intruction ODS HTML Body= « Nom de fichier HTML » va rediriger la sortie en fichier HTML.

L'instruction ODS LISTING CLOSE sert à supprimer la sortie vers l'OUTPUT.
L'instruction ODS LISTING sert à réactiver la sortie vers l'OUTPUT.

```
ods html body='c:\temp\univariate.html';
ods listing close;
proc univariate data=moi.stid193;
var taille;
run;
proc means data=moi.stid193 ;
  class groupe;
  var notemat;
  run;
ods html close;
ods listing;
```

SAS vous affiche alors le contenu du fichier HTML que vous auriez pu lire avec INTERNET EXPLORER ou NETSCAPE.

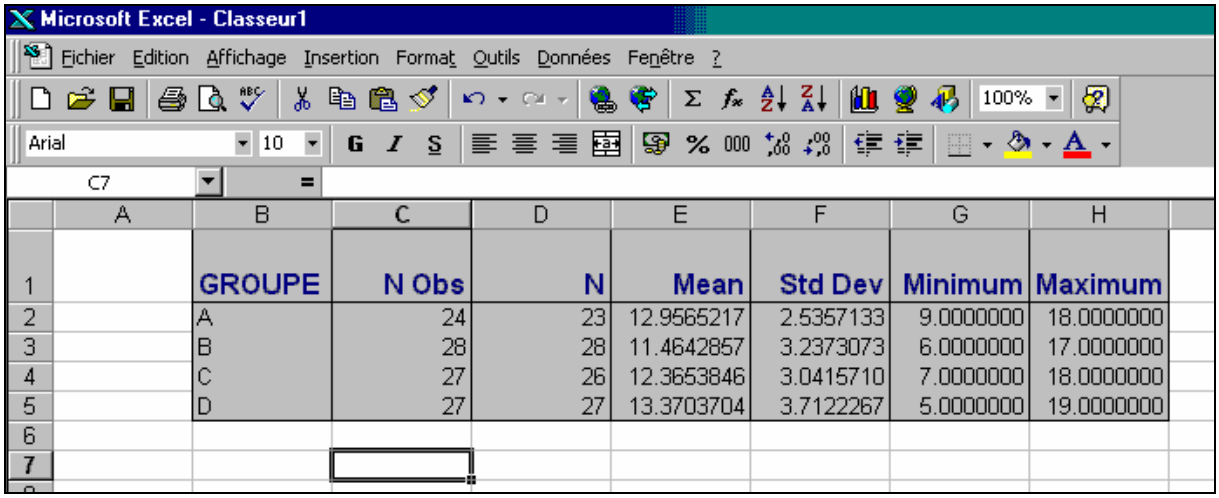
The SAS System

The UNIVARIATE Procedure
Variable: TAILLE (TAILLE)

Moments			
N	106	Sum Weights	106
Mean	170.698113	Sum Observations	18094
Std Deviation	7.83905948	Variance	61.4508535
Skewness	0.48084832	Kurtosis	-0.057081
Uncorrected SS	3095064	Corrected SS	6452.33962
Coeff Variation	4.59235274	Std Error Mean	0.76139676

Ce qui a quand même un peu plus d'allure…

Dans la suite de la sortie, vous avez la procédure MEANS dont vous pouvez copier coller les résultats sous Excel :



	A	B	C	D	E	F	G	H
1		GROUPE	N Obs	N	Mean	Std Dev	Minimum	Maximum
2		A	24	23	12.9565217	2.5357133	9.0000000	18.0000000
3		B	28	28	11.4642857	3.2373073	6.0000000	17.0000000
4		C	27	26	12.3653846	3.0415710	7.0000000	18.0000000
5		D	27	27	13.3703704	3.7122267	5.0000000	19.0000000
6								
7								

2. Sélection d'objets en sortie : ODS TRACE, ODS SELECT, ODS EXCLUDE

Comme nous l'avons vu tout à l'heure, la procédure UNIVARIATE inscrit 5 objets dans la sortie. Pour sélectionner ceux que nous voulons afficher, il faut repérer le nom des objets : c'est le but de L'instruction ODS TRACE

ODS TRACE ON <options> ;	Active le mode TRACE
ODS TRACE OFF ;	Supprime le mode TRACE (c'est l'option par défaut)

Les options étant

LABEL pour indiquer les chemin de l'objet
LISTING pour mettre les noms des objets avant les objets dans les sorties.

Exemple :

```
ods listing;                    Pour diriger les objets vers la fenêtre OUTPUT
ods trace on;                   Pour activer le mode TRACE
proc univariate data=moi.stid193;
var taille;
run;
ods trace off;
```

Va donner dans la LOG les noms de nos 5 objets :

```
Output Added:
-------------
Name:       Moments
Label:      Moments
Template:   base.univariate.Moments
Path:       Univariate.TAILLE.Moments
-------------

Output Added:
-------------
Name:       BasicMeasures
Label:      Basic Measures of Location and Variability
Template:   base.univariate.Measures
Path:       Univariate.TAILLE.BasicMeasures
-------------

Output Added:
-------------
Name:       TestsForLocation
Label:      Tests For Location
Template:   base.univariate.Location
Path:       Univariate.TAILLE.TestsForLocation
-------------

Output Added:
-------------
Name:       Quantiles
Label:      Quantiles
Template:   base.univariate.Quantiles
Path:       Univariate.TAILLE.Quantiles
-------------

Output Added:
-------------
Name:       ExtremeObs
Label:      Extreme Observations
Template:   base.univariate.ExtObs
Path:       Univariate.TAILLE.ExtremeObs
-------------
```

Ces noms (en gras) sont TRES IMPORTANTS car ce sont eux dont on va se servir dans la suite pour sélectionner les objets à afficher ! ! !

Exercice

- Compliquez la procédure UNIVARIATE en ajoutant une option PLOTS par exemple qui permet d'obtenir un box plot et un graphique de normalité :

```
ods trace on;
proc univariate data=moi.stid193 plots;
var taille;
run;
ods trace off;
```

Combien d'objets figurent dans la sortie cette fois ci ? Quel est le nom du dernier ?

- Redirigez la sortie en un document HTML, cela change-t-il quelque chose au nom des objets ?

Passons à ce qui nous intéresse :

Sélection des objets :

ODS <Destination> SELECT *noms des objets* | ALL |NONE ;

Exclusion d'objets :

ODS <Destination> EXCLUDE *noms des objets* | ALL |NONE ;

Pour savoir ou on en est :
ODS <destination> SHOW

Exemple :

```
ods listing select quantiles basicmeasures;     Sélection d'objets
ods listing show;                               Pour savoir où on en est
proc univariate data=moi.stid193 plots;
var taille;
run;
```

ODS LISTING SHOW va nous donner dans la LOG :

```
  ods listing show;
Current LISTING select list is:
1. quantiles
2. basicmeasures
```

cela nous confirme ce que nous souhaitions avoir.

Dans la fenêtre OUTPUT vous n'avez que ces deux objets.

Exercice

Le programme

```
proc reg data=moi.stid193;
model poids=taille;
run;
quit ;
```

permet d'effectuer une regression linéaire POIDS=a+b*TAILLE.

- Combien d'objets va créer ce programme ? Quels sont leurs noms ?

- Ecrivez un programme permettant de ne mettre dans la fenêtre OUTPUT que l'estimation des paramètres. a= ? b= ?

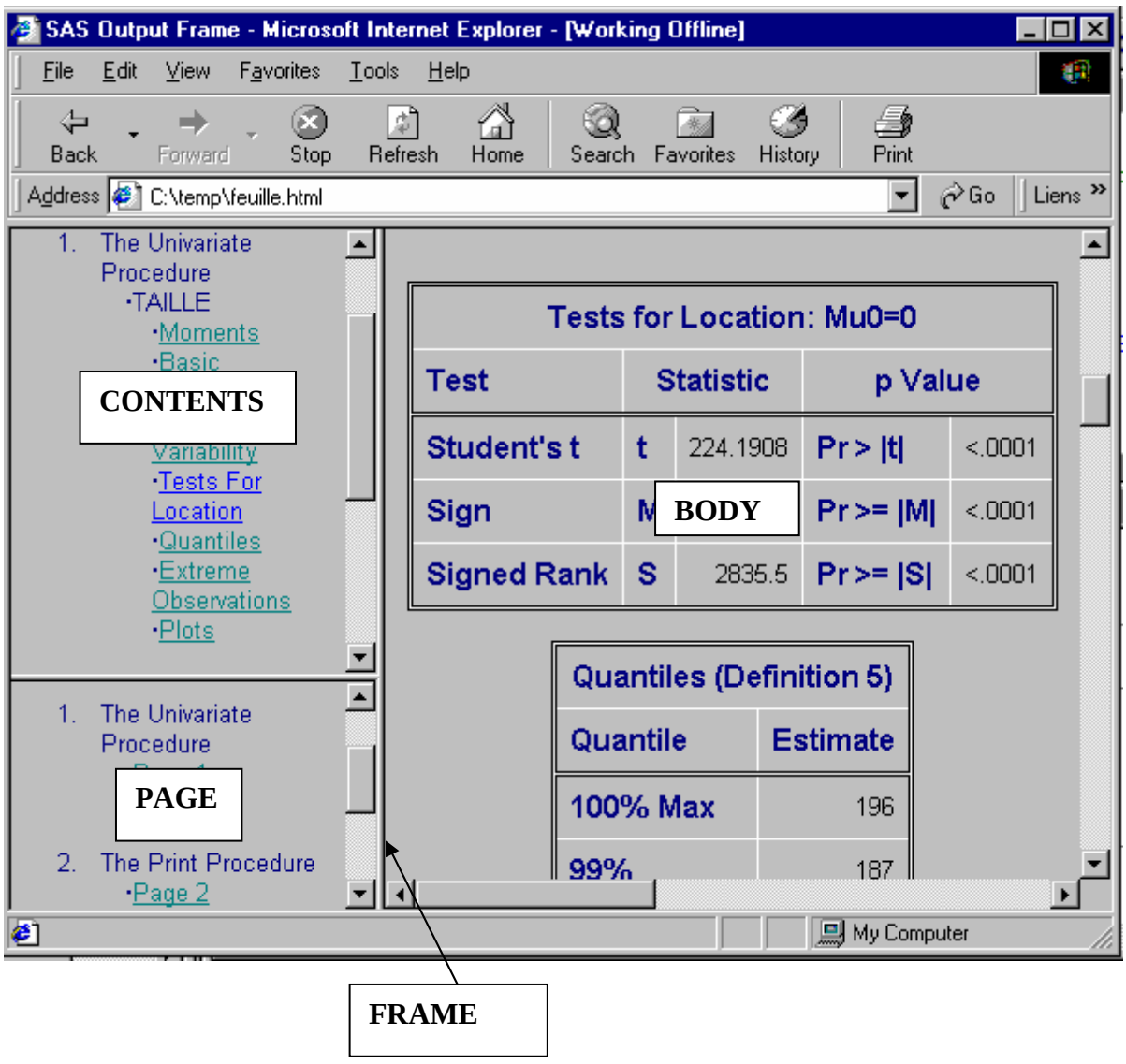
- Même chose mais dans un document HTML. (vous indiquerez deux instructions ODS HTML une pour indiquer le fichier de sortie (BODY), une autre pour sélectionner les objets (SELECT)).

3. Sorties HTML sophistiquées

a) Structure d'une Feuille HTML

Pour SAS, votre feuille HTML contient 3 éléments :

- Le corps (BODY) ce sont toutes les sorties de SAS en HTML.
- La table des matières (CONTENTS) contient le nom des objets de chaque Page du BODY.
- La table des Pages (PAGE) qui contient le titre de chaque Page et son numéro.



Ici, notre sortie comporte deux pages. LA première page contient la sortie d'UNIVARIATE. Cette sortie d'UNIVARIATE comporte plusieurs objets. Nous sommes en train de visualiser les TESTS FOR LOCATIONS.

Pour chacun des éléments précédents, SAS vous demande un nom de fichier HTML.

Pour éviter de taper le chemin de chaque fichier , vous pouvez utiliser la commande PATH=.

Exemple :

```
ods listing close;                  On ferme l'OUTPUT
ods html path='c:\temp'             On ouvre la sortie HTML en c:\temp
         body='corps.html'          On stocke différentes parties dans 4 fichiers
         contents='contenu.html'
         page='page.html'
         frame='feuille.html' ;
proc univariate data=moi.stid193 plots;
var taille;
run;
proc print data=moi.stid193 (obs=10) ;
run;
ods html close;          On ferme les fichiers HTML
ods listing;             On ouvre la feuille OUTPUT pour la suite.
```

Remarques : En fait Corps.html va être en c:\temp\corps.html etc.

Un seul des fichiers précédents est essentiel : c'est BODY qui contient tous les résultats.

Application

- Tapez le programme précédent.
- Exécutez le.
- Depuis Internet Explorer chargez la page Feuille.html. Amusez vous à vous promener dans cette sortie. Chargez les autres fichiers HTML crées par SAS.

b) Changement de Style d'une feuille HTML

Pour personnaliser les sorties HTML précédente, il suffit d'ajouter l'option STYLE= dans les instructions précédentes :

```
ods listing close;
ods html path='c:\temp'
         body='corps.html'
         contents='contenu.html'
         page='page.html'
         frame='feuille.html'
         style=brown;
proc univariate data=moi.stid193 plots;
var taille;
run;
proc print data=moi.stid193 (obs=10) ;
run;
ods html close;
ods listing;
```

- Faites l'essai !

Divers modèles sont fournis par SAS :

DEFAULT
BEIGE
BRICK
BROWN
D3D
MINIMAL
STATDOC

- Essayez les avec l'exemple précédent

Pour créer vos propres styles, il faut avoir recours à une nouvelle procédure PROC TEMPLATE.

c) Personnalisation des titres et notes de bas de page

Dans l'instruction Title, ou Footnote, nous pouvez indiquer des éléments HTML permettant de modifier les polices des titres, notes de bas de page :

Syntaxe :

Title '<FONT *nom de l'attribut= « valeur » > texte du titre* </FONT> ';

Quelques attributs et leur valeur :

FONT FACE= *nom de la police*
Pour changer la police de carcatères (Arial, Times etc.)

FONT SIZE= taille
Pour changer la taille de la police (1 à 7...)

FONT STYLE= *style de la police*
Pour changer le style de la police de carcatères (Italic, Roman…)

FONT WEIGHT= *gras ou non*
Pour changer le style (Medium, Bold…)

FONT WIDTH=*espacement de la police :* Normal, Narrow ou Wide

De même avec FOOTNOTE.

Exemple :

```
ods html path='c:\temp'
         body='body1.html';
title '<font face="Arial" color="green" weight="bold" size=6
> Dix individus de STID année 93 </font>';
proc print data=moi.stid193 (obs=10);
var groupe sexe taille poids;
run;
title ; /*pour effacer le titre pour la suite */
```

- Exécutez cet exemple.
- Créez une note de bas de page Courier Rouge de taille 4 indiquant d'où viennent les données.

d) Utilisation de STYLES dans la procédure TABULATE

Nous allons utiliser une procédure TABULATE pour illustrer ces notions :

```
title 'Synthèse des résultats par Groupe';
proc tabulate data=moi.stid193 format=6.1;
class groupe ;
var notemat notehis notefr_;
table groupe*(notemat notehis notefr_), min  median max /
 box={label='Notes du BAC'};
run;
title ;
```

donne dans la fenêtre OUTPUT :

Synthèse des résultats par Groupe

Notes du BAC		Min	Median	Max
GROUPE				
A	NOTE MAT	9.0	13.0	18.0
	NOTE HIS	5.0	11.5	17.0
	NOTE FR.	5.0	8.0	14.0
B	NOTE MAT	6.0	11.0	17.0
	NOTE HIS	6.0	11.0	16.0
	NOTE FR.	4.0	9.0	14.0
C	NOTE MAT	7.0	11.5	18.0
	NOTE HIS	6.0	10.0	17.0
	NOTE FR.	5.0	8.0	14.0
D	NOTE MAT	5.0	13.0	19.0
	NOTE HIS	5.0	11.0	16.0
	NOTE FR.	5.0	8.0	12.0

- Redirigez cette sortie dans un fichier HTML.

Options à placer derrières les commandes de la procédure TABULATE…
Pour changer les couleurs et la police des éléments d'une page HTML

STYLE= {Background=*couleur*} ;
Pour changer la couleur de l'arrière plan.

STYLE={Foreground=*couleur*} ;
Pour changer la couleur du texte.

STYLE={FONT_FACE= *nom de la police}*
Pour changer la police de carcatères (Arial, Times etc.)

STYLE={FONT_SIZE= taille*}*
Pour changer la taille de la police (1 à 7...)

STYLE={FONT_STYLE= *style de la police}*
Pour changer le style de la police de carcatères (Italic, Roman…)

STYLE={FONT_WEIGHT= *gras ou non}*
Pour changer le style (Medium, Bold…)

STYLE={FONT_WIDTH=*espacement de la police :* normal, Narrow ou Wide

Quelques couleurs possibles étant :

Red, Pink, Orange, Yellow, Yellow-Green, Green, Blue, purple, Black, White, Cyan :

Exemple :

Reprenons l'exemple précédent en changeant les couleurs de fond et de caractères pour différents éléments de la sortie précédente :

```
ods listing close;
ods html body='c:\temp\tabulate.html';
proc tabulate data=moi.stid193 format=6.1
style={background=yellow foreground=red};
title 'Synthèse des résultats par Groupe';
class groupe / style={background=brown};
classlev groupe / style={Background=purple foreground=red
font_size=30};
keyword min median max / style={background=red
font_weight=bold};
var notemat / style={background=pink foreground=red
font_face=times font_style=italic};
var notefr_ / style={background=pink foreground=green};
var notehis / style={background=pink foreground=cyan};
table groupe*(notemat notehis notefr_), min  median max /
 box={label='Notes du BAC'};
run;
ods html close;
ods listing ;
```

- Exécutez cette sortie , à quoi servent les lignes CLASSLEV, KEYWORD ?

STYLES des cellules « Parents » :

L'inconvénient de la sortie précédente, c'est que le style des cellules calculées (nombres) ne correspond pas au style des cellules contenant les noms des notes. Nous pouvons automatiquement les affecter en utilisant l'option :

*STYLE=<PARENT>

```
table groupe*(notemat notehis notefr_)*{style=<parent>}, min
median max / box={label='Notes du BAC'};
```

- Voyez la différence et reconnaissez le goût certain de votre prof pour l'harmonie des couleurs !

e) Coloration conditionnelle des cellules : utilisation de FORMAT

Cet exemple assez spectaculaire peut vous montrer l'utilisation des STYLES différentes selon la valeur de la cellule. Pour cela, nous allons utiliser un Format créé spécialement. Si vous n'êtes pas à l'aise avec les formats, allez voir le paragraphe correspondant dans les annexes et la PROC FORMAT.

Nous allons voir ici comment colorier le fond de la cellule

en ROUGE si le poids est supérieur à 65
en JAUNE s'il est entre 60 et 65
en VERT, s'il est inférieur à 60.

Création d'un nouveau Format

Pour cela, nous allons définir un nouveau Format appelé **FOND.**

```
proc format ;
value fond low-60 = 'Green' 60<-65='Yellow' 65<-High ='Red';
run;
```

Fond. prend donc les valeurs Green , Yellow et Red selon les valeurs de la variable à laquelle nous allons attribuer ce format :

Amusons nous à afficher les Poids des 10 premières personnes de STID avec ce format :

```
proc print data=moi.stid193 (obs=10);
format poids fond.;
var poids;
run;
```

On a :

Obs	POIDS
1	Red
2	Yellow
3	Red
4	Green
5	Green
6	Green
7	Green
8	Green
9	Yellow
10	Green

Rigolo non ?

Remarque : La valeur de la variable POIDS n'a pas changé ! On lui a juste appliqué un masque ! En interne, les valeurs sont inchangées…

Utilisons ce format dans les Styles :

```
ods listing close;
ods html body='c:\temp\corps.html' style=brown;

proc tabulate data=moi.stid193;
title 'Répartition des Poids selon le sexe';
class groupe sexe;
var poids;
table groupe*(sexe*Poids)*{style={background=fond.
foreground=black font_weight=bold}},mean median;
run;

ods html close;
ods listing;
```

Visualisez le résultat.

Rigolo non ?

f) Exercice récapitulatif :

Copier les fichiers FRAME1.HTML ; BODY1.HTML, CONTENT1.HTML, PAGE1.HTML du répertoire public en C:\TEMP

Ouvrez le fichier FRAME1.HTML avec INTERNET EXPLORER.

Retrouvez le programme SAS capable de produire ces fichiers HTML ! ! !

4. Sorties HTML pour les graphiques

Les graphiques ne sont pas simples à gérer sous SAS. Le module GRAPH permet d'obtenir des graphiques présentables mais au prix d'une programmation complexe. Certains outils (Graph n Go) ou modules (ASSIST, *Enterprise Guide*) permettent de créer des graphiques simples.

a) Assistant graphique : Graph n Go

Sinon, vous pouvez utiliser l'utilitaire GRAPH-N-GO (Menu SOLUTIONS/REPORTING) qui est un assistant graphique.[50]

Il permet d'effectuer des graphiques simples en cliquant sur des boutons puis de les exporter en HTML (fixe, JAVA, ou ACTIVE X). Vous pouvez aussi récupérer le code SAS ayant permis de faire les graphiques en question.

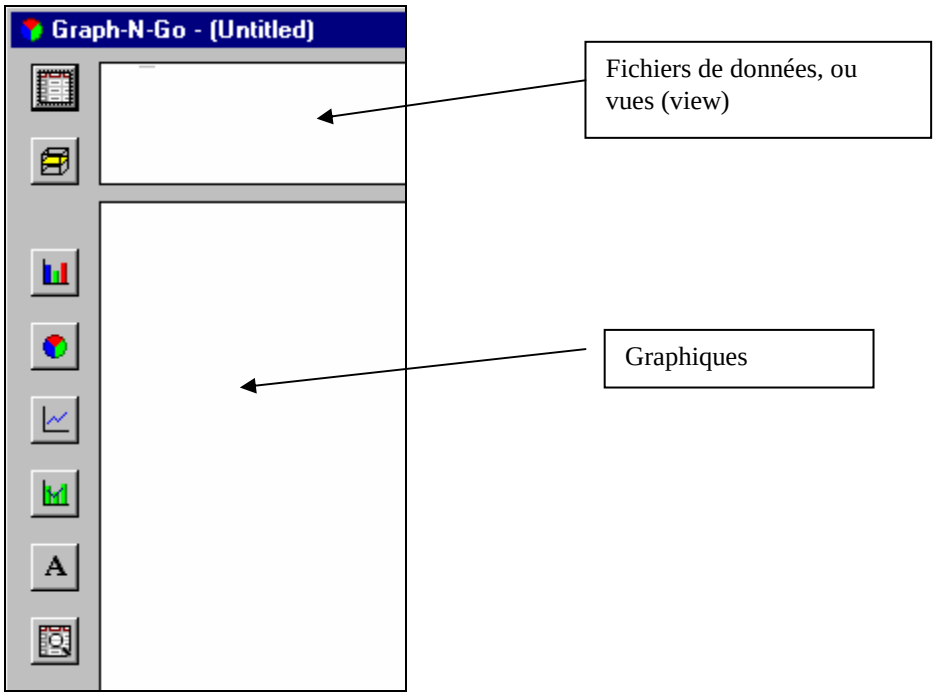
Les données doivent figurer dans des tables SAS dont vous pouvez extraire des parties.

[50] Signalons aussi SAS ENTERPRISE GUIDE mais il necessite un module supplémentaire. Graph n Go ne nécessite que les modules BASE et GRAPH.

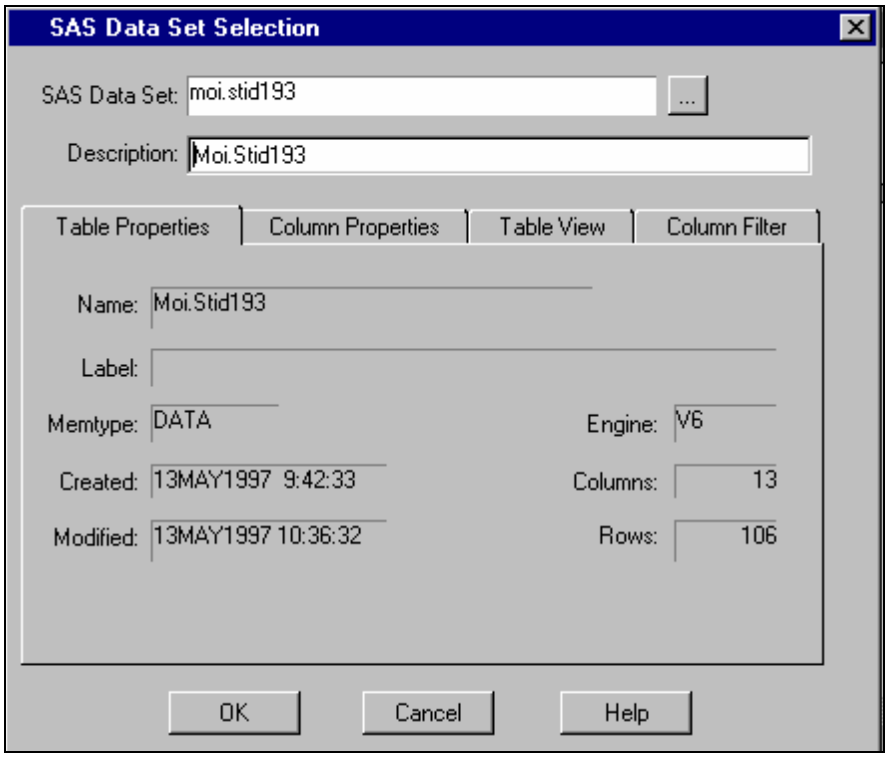
b) Diagramme à Bande HTML Interactif avec Graph n Go

Nous allons créer un petit graphique illustrant la répartition des Bacs en STID193. Nous allons récupérer le code permettant de faire ce graphique. Puis nous allons l'exporter en Fichier HTML interactif Active X.

Activez cet outil (SOLUTIONS/REPORTING)



Nous allons d'abord sélectionner un fichier de données (STID193). Pour cela cliquez sur le premier bouton en haut à gauche. (New SAS Data Set Model)

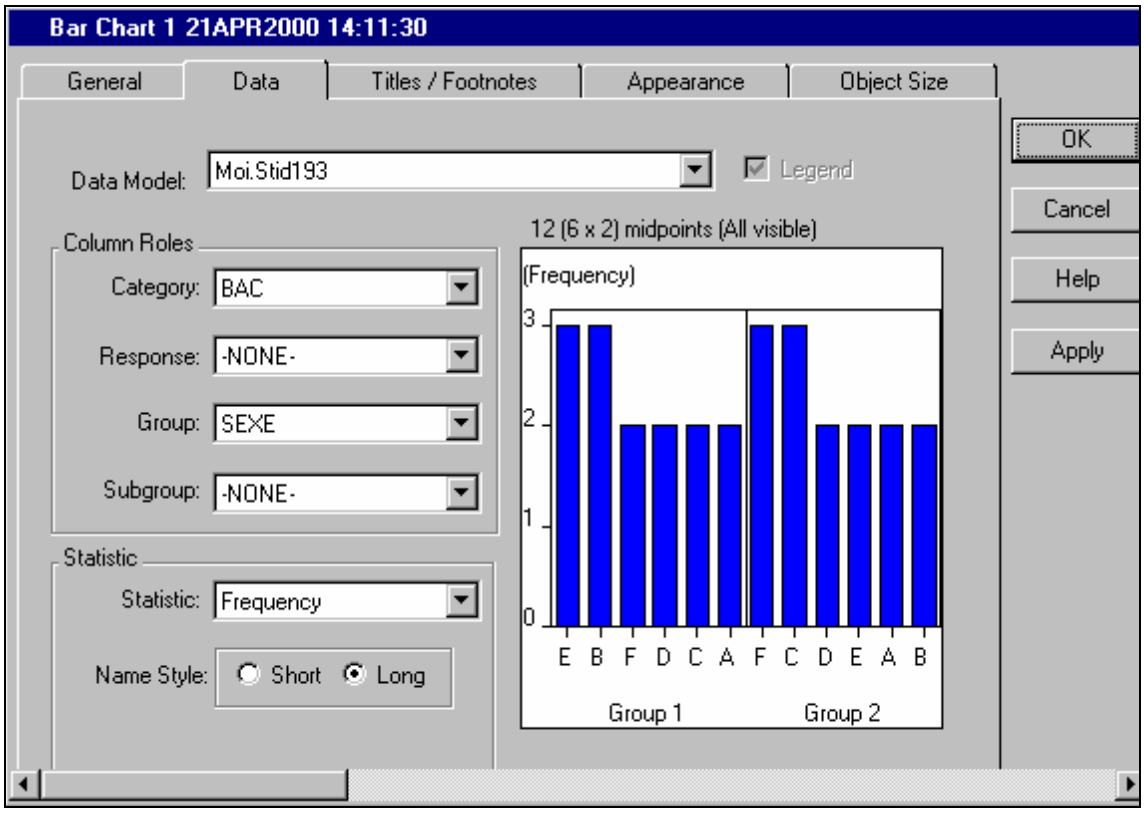


Sélectionnez le fichier STID193. Vous voyez que vous pouvez sélectionner certaines colonnes (Column Filter) si vous le souhaitez.

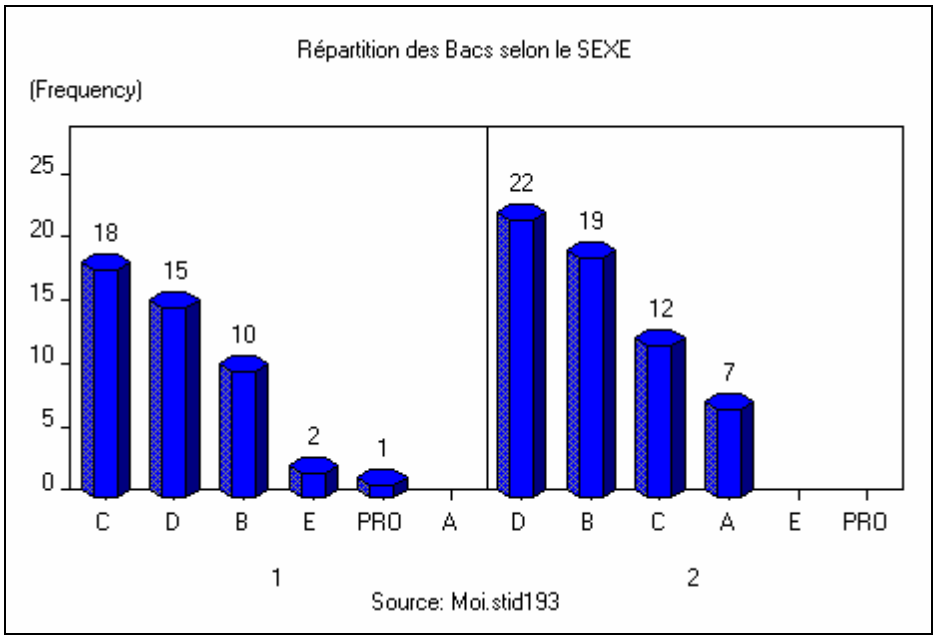
Validez, vous avez maintenant dans la partie fichier de données :

Nous allons maintenant effectuer un diagramme à bande.

Cliquez sur l'outil « BAR CHART ». Glissez le cadre où vous souhaitez. Double cliquez dessus.



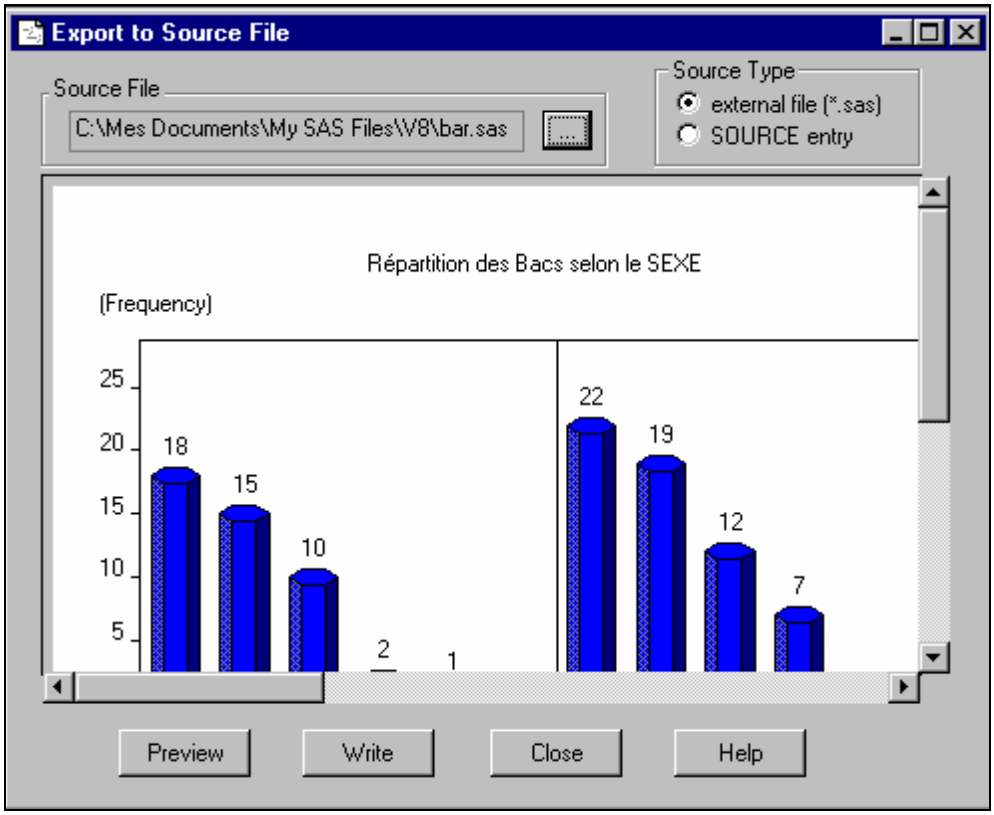
Choisissez les options de manière à obtenir (après avoir agrandi le graphique avec GROW – du menu contextuel) :



Récupération du Code SAS

Dans le menu contextuel, choisissez EXPORT puis

Vous obtenez alors :



Cliquez sur WRITE pour sauver le fichier et PREVIEW pour le visualiser. Vous aurez les commandes SAS qu'il fallait donner pour obtenir ce graphique.

En gras figurent les commandes graphiques fondamentales.

```
/* Graph-N-Go SAS/Graph Code Generation for
**
** SAS products required: Base, SAS/Graph (Version 8 or later)
** Code generated on: 21APR2000  15:29:55
**
** Notes: There may be differences in appearance of the graph
**        generated by the code below and the Graph-N-Go viewer.
**
**        To make code modifications consult the documentation
**        for these statements: GCHART, GPLOT, ODS,
**        GOPTIONS, AXIS, LEGEND, SYMBOL, TITLE, FOOTNOTE.
**
**        To route output to a graphics device other than your monitor,
**        modify the source code below to change the device driver by
**          1) removing the asterisk preceding GOPTIONS DEVICE=JAVA;
**          2) changing JAVA to some other valid device.
**
**        To create an interactive HTML file, modify the source code below
**        to enable ODS output by
**          1) removing the asterisks from the two ODS statments and
**             the asterisk preceding GOPTIONS DEVICE=JAVA;
**          2) changing DEVICE=JAVA to DEVICE=ACTIVEX if you want to create
**             an ActiveX control rather than a Java applet
**          3) verifying or changing the ODS FILE= option so it names an
**             output HTML file.
*/


/* Begin ODS output */
* ods html file="C:\Mes Documents\My SAS Files\V8\bar.html"
           parameters=("DisableDrillDown"="True"
                       "ShowBackDrop"="False"
                       "BackColor"="#FFFFFF"
                       "BackDropColor"="#FFFFFF"
                       "FreqName"="BAC"
                       "FreqDesc"="(Frequency)"
                       "FreqFmt"="BEST."

"MenuRemove"="File,Variables,Options:Drilldown,Graph:Image,Graph:Navigate,Graph:Type,Legend"
                       );


/* Set the SAS/Graph options */
goptions reset=all hpos=40
         ctext=CX000000 ftext="MS Sans Serif"
         colors=(CX0000FF CXFF0000 CX008080 CX00FF00 CXFF00FF CXFFFF00 CX00FFFF CX800000
                 CX008000 CX800080 CX000080 CX808000 CXFFFFFF CX808080 CXC0C0C0 CX000000);

/* Set the Titles/Footnotes */
title1 justify=center color=CX000000 font="MS Sans Serif" height=8 pt "Répartition des Bacs
selon le sexe";
footnote1 justify=center color=CX000000 font="MS Sans Serif" height=8 pt "Source: Stid193";

/* Set the SAS/Graph device driver */
* goptions device=JAVA xpixels=531 ypixels=346;

/* AXIS1 describes axis for Category variable BAC */
/* AXIS2 describes axis for Response statistic FREQ */
axis1 minor=none  label=("BAC") ;
axis2 minor=none  label=("(Frequency)")
      order=(0 to 25 by 5) ;
proc gchart data=MOI.STID193;
   vbar BAC /
        type=FREQ  maxis=axis1 descending
        discrete frame cframe=CXFFFFFF
        woutline=1 coutline=CX000000 caxis=CX000000
        raxis=axis2   group=SEXE G100  ;
run;
quit;

/* Reset all graphics options */
goptions reset=all;

/* End ODS output */
* ods html close;
```

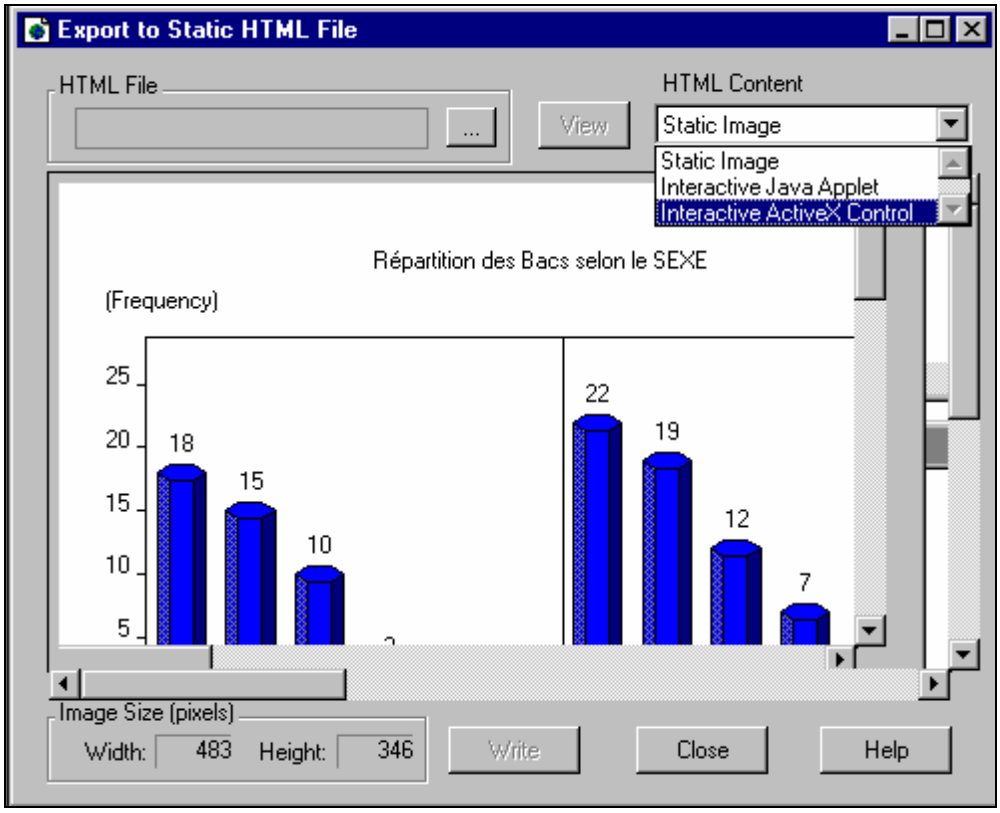
Exportation du graphique vers un fichier HTML

Choisissez EXPORT puis HTML File

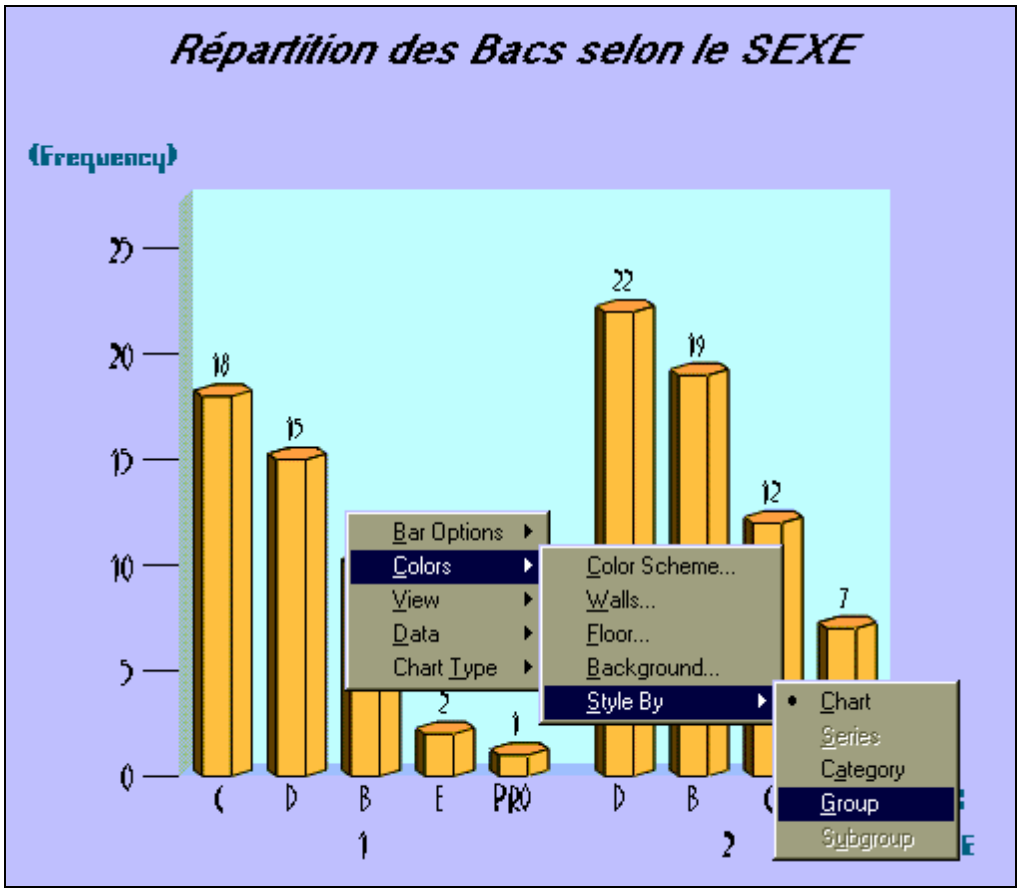
3 options s'offrent à vous :

- Static Image : Fichier HTML fixe. On ne peut modifier le graphique.
- Interactive JAVA Applet : Fichier HTML interactif puisqu'en fait SAS va générer un programme JAVA permettant à votre Browser Internet de rendre votre grqphique modifiable.
- Interactive Active X Control : Idem.

Remarque : Les deux dernières options requièrent les bibliothèques (JAVA, Active X) adéquates pour Windows. Elles sont automatiquement installées si SAS a été installé dans les règles sur votre Micro.



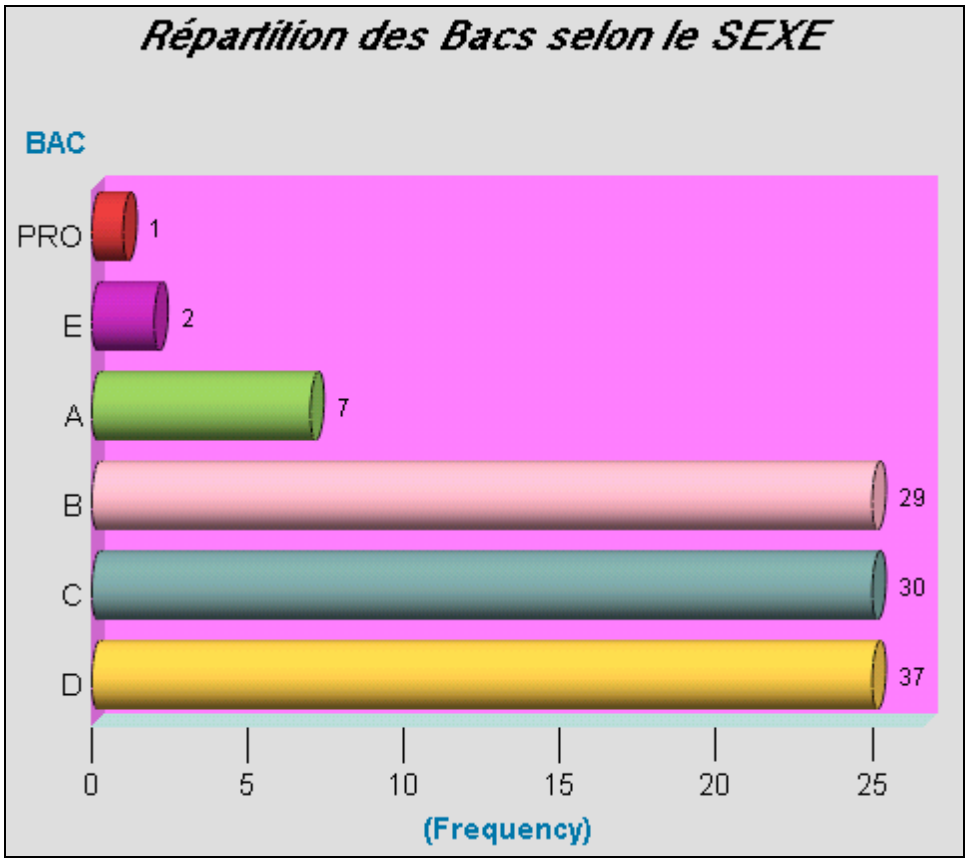
- Choisissons la troisième option.
- Editez alors le fichier HTML avec SAS ou votre Browser Internet habituel.
- Cliquez sur le bouton droit de la souris



Comme vous le voyez, nous pouvons modifier les couleurs, les polices, le type de graphique…

Amusant non ?

Modifiez le graphique précédent pour faire apparaître le graphique suivant (sans retoucher à SAS !)


Répartition des Bacs selon le SEXE
BAC
PRO
E
A
B
C
D
1
2
7
29
30
37
0
5
10
15
20
25
(Frequency)

5. Sorties vers des fichiers de données

a) Fichiers simples

La version 6 de SAS ne permettait de sauvegarder dans les fichiers de données qu'une partie des sorties des procédures. Avec la V8, TOUTES les sorties des procédures peuvent être exportée dans des fichiers de données.

Il est possible de ne mettre qu'une partie des sortie en spécifiant le nom des objets.

Syntaxe simplifiée :

ODS OUTPUT *nom de l'objet* (options) = nom du fichier de données SAS ;

Exemple :
La procédure Univariate renvoie (en général) 5 objets :Moments, BasicMeasures, TestsForLocation, Quantiles, Extremeobs.

Pour sauvegarder ces données dans des fichiers, il suffit de l'indiquer derrière ODS OUTPUT comme le montre le programme ci-dessous :

```
ods listing close;        On ferme la sortie dans la fenêtre OUTPUT classique

/*On redirige la sortie vers des fichiers de données*/

ods output Moments=work.moments
           BasicMeasures=Work.statdebase
           TestsForLocation=Work.test
           Quantiles=work.quantiles
           Extremeobs=work.extremes;

Proc univariate data=moi.stid193;
var notemat notehis notefr_;
run;

ods listing;        On réactive la sortie OUTPUT classique pour visualiser les fichiers créés.

/*Visualisation de deux fichiers */

proc print data=work.moments;
title 'Fichier Work.moments';
run;

proc print data=work.quantiles;
title 'Fichier Work.quantiles';
run;
```

va donner pour le fichier WORK. QUANTILES

Fichier Work.quantiles

Obs	VarName	Quantile	Estimate
1	NOTEMAT	100% Max	19.00
2	NOTEMAT	99%	18.00
3	NOTEMAT	95%	18.00
4	NOTEMAT	90%	17.00
5	NOTEMAT	75% Q3	15.00
6	NOTEMAT	50% Median	12.25
7	NOTEMAT	25% Q1	10.00
8	NOTEMAT	10%	9.00
9	NOTEMAT	5%	7.00
10	NOTEMAT	1%	6.00
11	NOTEMAT	0% Min	5.00
12	NOTEHIS	100% Max	17.00
13	NOTEHIS	99%	17.00
14	NOTEHIS	95%	16.00
15	NOTEHIS	90%	15.00
16	NOTEHIS	75% Q3	12.00
17	NOTEHIS	50% Median	11.00
18	NOTEHIS	25% Q1	9.00
19	NOTEHIS	10%	6.00
20	NOTEHIS	5%	6.00
21	NOTEHIS	1%	5.00
22	NOTEHIS	0% Min	5.00
23	NOTEFR_	100% Max	14.00
24	NOTEFR_	99%	14.00
25	NOTEFR_	95%	13.00
26	NOTEFR_	90%	12.00
27	NOTEFR_	75% Q3	10.00
28	NOTEFR_	50% Median	9.00
29	NOTEFR_	25% Q1	7.00
30	NOTEFR_	10%	6.00
31	NOTEFR_	5%	6.00
32	NOTEFR_	1%	5.00
33	NOTEFR_	0% Min	4.00

Vous pouvez aussi le visualiser avec l'EXPLORER de SAS en allant dans la bibliothèque WORK.

- Visualisez les autres fichiers de données.

b) Fichiers multiples (Option MATCH_ALL)

Lorsque vous utilisez une OPTION BY dans une procédure, vous allez créer un fichier de données contenant toutes les occurrences du nom de l'objet. En clair :

```
/* On trie le fichier par rapport au groupe*/
proc sort data=moi.stid193 out=work.stidtri;
 by groupe;
 run;

/* On se prepare a sauvegarder les Quantiles*/

ods listing close;
ods output Quantiles=work.quantiles;

proc univariate data=work.stidtri;
var notemat notehis;
by groupe;
run;

ods listing;

proc print data=work.quantiles;
title 'Quantiles des notes de Maths et d''Histoire
par Groupe';
run;
```

Va donner :

Quantiles des notes de Maths et d'Histoire par Groupe

Obs	GROUPE	VarName	Quantile	Estimate
1	A	NOTEMAT	100% Max	18.0
2	A	NOTEMAT	99%	18.0
3	A	NOTEMAT	95%	17.0
4	A	NOTEMAT	90%	16.5
5	A	NOTEMAT	75% Q3	15.0
6	A	NOTEMAT	50% Median	13.0
7	A	NOTEMAT	25% Q1	10.0
8	A	NOTEMAT	10%	10.0
9	A	NOTEMAT	5%	9.0
10	A	NOTEMAT	1%	9.0
11	A	NOTEMAT	0% Min	9.0
12	A	NOTEHIS	100% Max	17.0
13	A	NOTEHIS	99%	17.0
14	A	NOTEHIS	95%	15.0
15	A	NOTEHIS	90%	14.0
16	A	NOTEHIS	75% Q3	12.0
17	A	NOTEHIS	50% Median	11.5
18	A	NOTEHIS	25% Q1	9.0
19	A	NOTEHIS	10%	8.0
20	A	NOTEHIS	5%	6.0
21	A	NOTEHIS	1%	5.0
22	A	NOTEHIS	0% Min	5.0
23	B	NOTEMAT	100% Max	17.0
24	B	NOTEMAT	99%	17.0
25	B	NOTEMAT	95%	17.0
26	B	NOTEMAT	90%	17.0
27	B	NOTEMAT	75% Q3	13.5
28	B	NOTEMAT	50% Median	11.0
29	B	NOTEMAT	25% Q1	9.5
30	B	NOTEMAT	10%	7.0
31	B	NOTEMAT	5%	6.0
32	B	NOTEMAT	1%	6.0
33	B	NOTEMAT	0% Min	6.0
34	B	NOTEHIS	100% Max	16.0
35	B	NOTEHIS	99%	16.0
36	B	NOTEHIS	95%	16.0
37	B	NOTEHIS	90%	15.0
38	B	NOTEHIS	75% Q3	13.0
39	B	NOTEHIS	50% Median	11.0
40	B	NOTEHIS	25% Q1	9.0
41	B	NOTEHIS	10%	8.0
42	B	NOTEHIS	5%	6.0
43	B	NOTEHIS	1%	6.0
44	B	NOTEHIS	0% Min	6.0
45	C	NOTEMAT	100% Max	18.0
46	C	NOTEMAT	99%	18.0
47	C	NOTEMAT	95%	17.0

….à suivre

…(Tous les groupes sont dans un seul fichier)

En ajoutant (Match_all) derrière le nom de l'objet, SAS va créer un fichier de données différent à chaque changement de groupe et de variable :

```
ods listing close;
ods output Quantiles(Match_all)=work.quantiles;

proc univariate data=work.stidtri;
var notemat notehis;
by groupe;
run;

ods listing;
proc print data=work.quantiles;
title 'Quantiles des notes de Maths du Groupe A';
run;
proc print data=work.quantiles1;
title 'Quantiles des notes d''Histoire du Groupe A';
run;
proc print data=work.quantiles2;
title 'Quantiles des notes de Maths du Groupe B';
run;
```

Le dernier PRINT va donner :

Quantiles des notes de Maths du Groupe B

Obs	GROUPE	VarName	Quantile	Estimate
1	B	NOTEMAT	100% Max	17.0
2	B	NOTEMAT	99%	17.0
3	B	NOTEMAT	95%	17.0
4	B	NOTEMAT	90%	17.0
5	B	NOTEMAT	75% Q3	13.5
6	B	NOTEMAT	50% Median	11.0
7	B	NOTEMAT	25% Q1	9.5
8	B	NOTEMAT	10%	7.0
9	B	NOTEMAT	5%	6.0
10	B	NOTEMAT	1%	6.0
11	B	NOTEMAT	0% Min	6.0

- Combien de fichiers de données seront créés ici ?
- Comment sont ils nommés ?

IV. Analyse interactive de données : SAS/INSIGHT

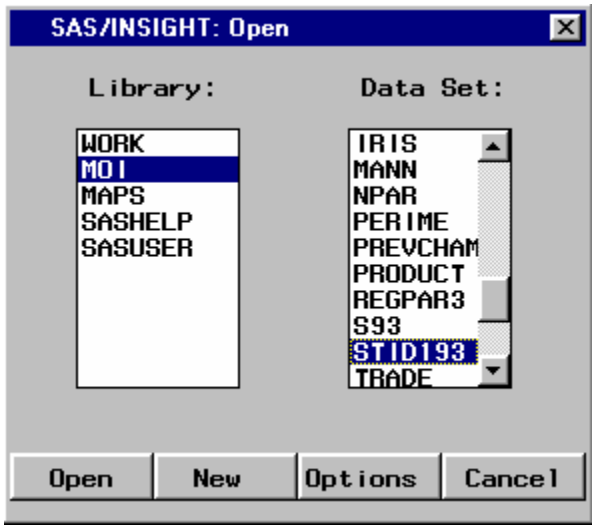
Ce module inclus dans l'outil de DATA MINING de SAS (Enterprise MINER) permet d'effectuer de puissantes analyses interactives de données, essentiellement sur des données quantitatives.[51]

Le grand intérêt de ce module réside dans le fait de pouvoir faire rapidement de nombreux graphiques interactifs, de repérer, d'exclure ou d'inclure de nouveaux individus. Tous les calculs sont alors modifiés en conséquence.

On peut effectuer des régressions, des analyses en composantes principales... bref un très bon outil.

A. Ouverture d'une table

Pour le mettre en œuvre allez dans `SOLUTIONS/ANALYSISINTERACTIVE DATA ANALYSIS` ; une fenêtre apparaît dans laquelle vous allez spécifier le nom du fichier que vous souhaitez analyser[52] :



Cliquez sur OPEN pour continuer.

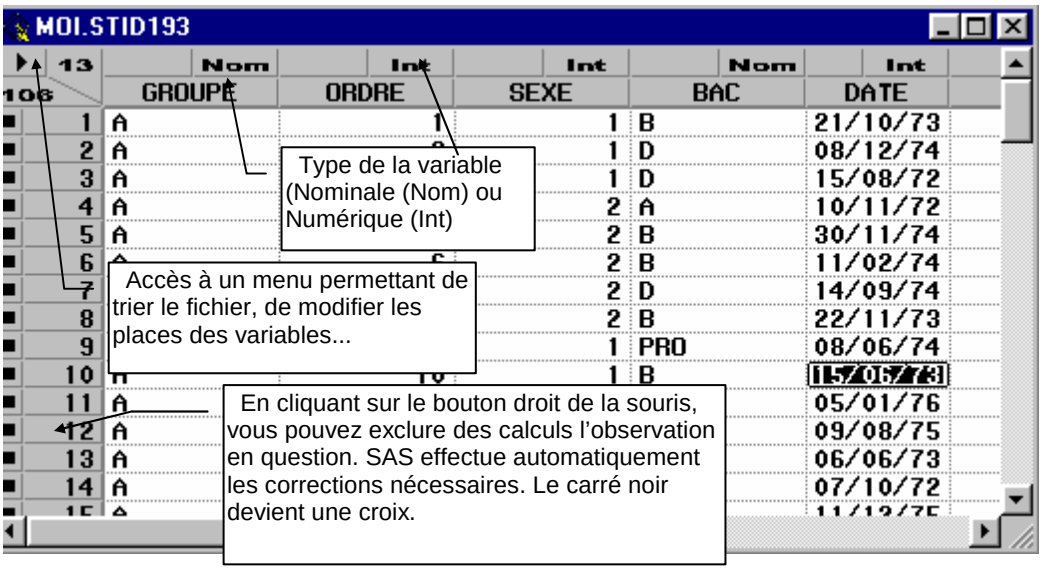
Remarque : Il est aussi possible d'ouvrir INSIGHT avec PROC INSIGHT.

```
proc insight data=moi.stid193;
run;
```

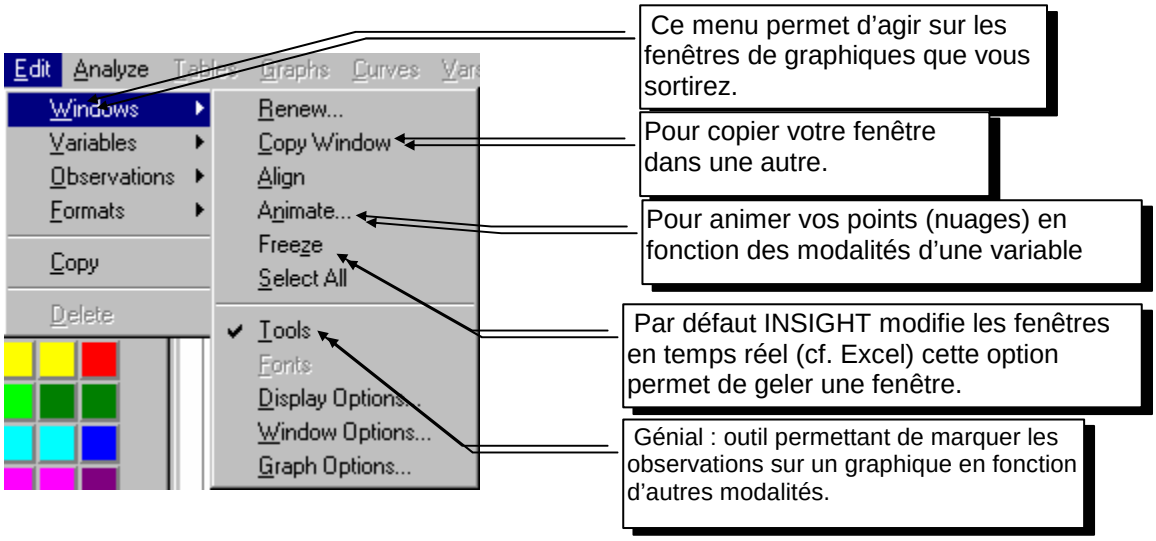
Vous obtenez :

[51] Histogrammes, Boxplots, Diagrammes à bandes, fonctions de répartition, ajustement de lois (test de normalité), Régression simple et multiple, ACP etc.

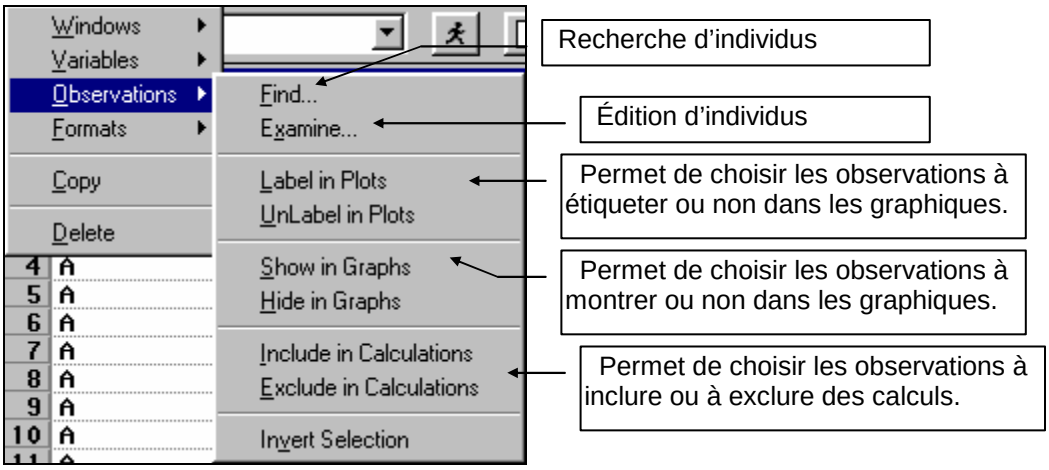
[52] Vous devez avoir déclaré la bibliothèque contenant le fichier que vous souhaitez analyser.


MOI.STID193
Nom
Int
Int
Nom
Int
GROUPE
ORDRE
SEXE
BAC
DATE
Type de la variable (Nominale (Nom) ou Numérique (Int)
Accès à un menu permettant de trier le fichier, de modifier les places des variables...
En cliquant sur le bouton droit de la souris, vous pouvez exclure des calculs l'observation en question. SAS effectue automatiquement les corrections nécessaires. Le carré noir devient une croix.

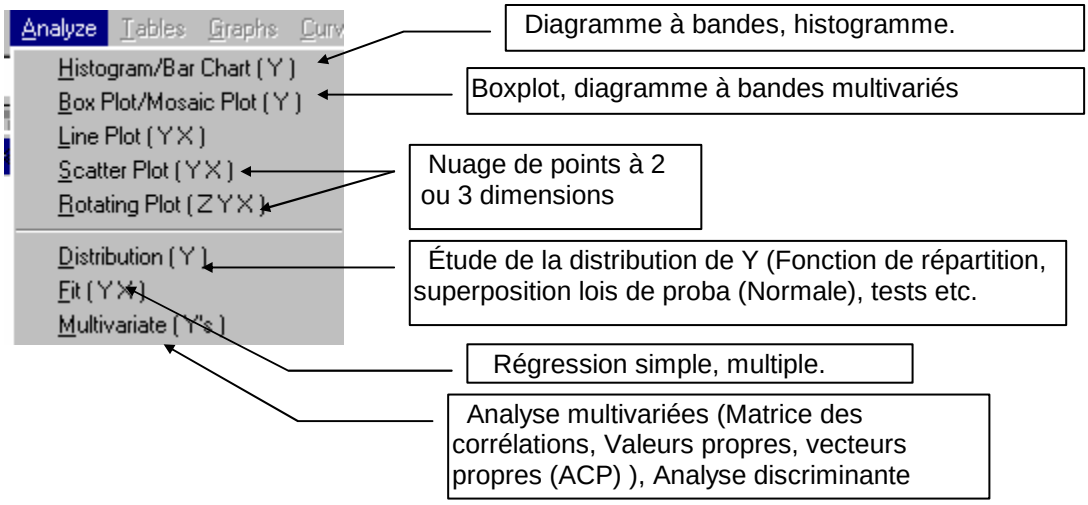
1. Aperçu rapide de quelques menus



Le menu Edit permet aussi d'agir sur les variables (Transformation (Log, Exponentielle,...) sur les observations :



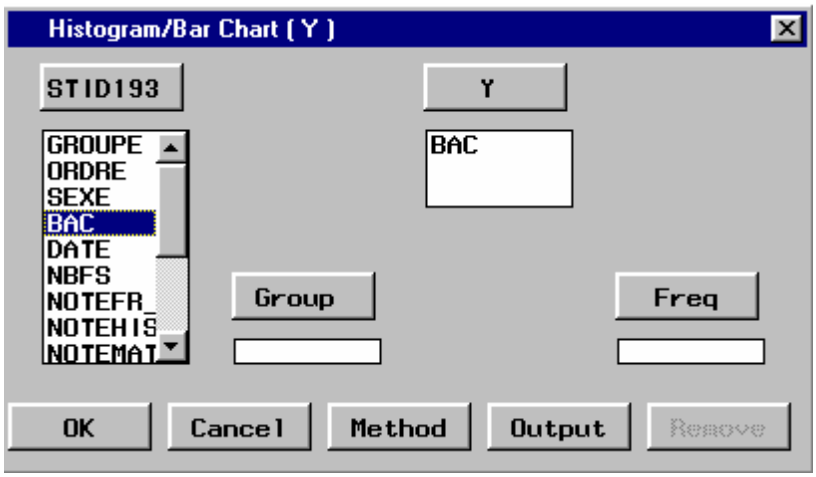
Le menu Analyse permet d'effectuer des calculs et des graphiques de différents types :



B. Analyse d'une Variable qualitative

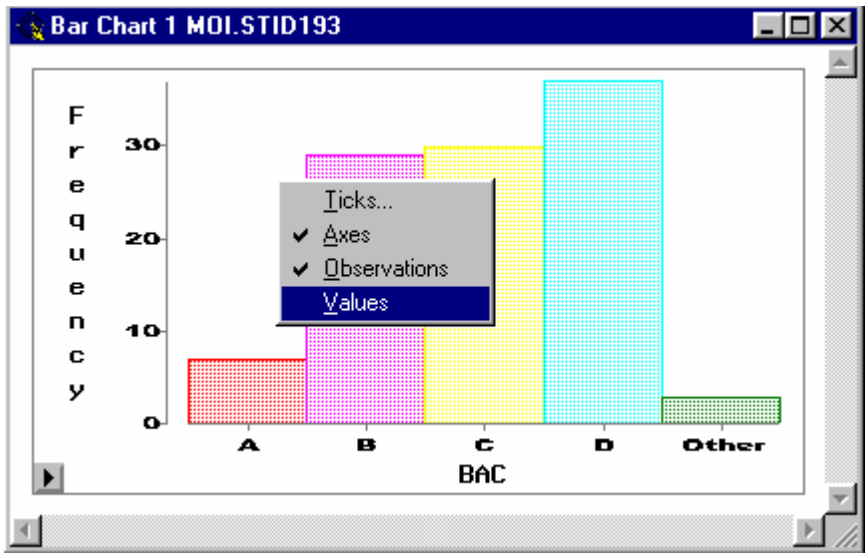
Nous allons voir comment tracer un diagramme à bandes et obtenir un tri à plat .

Choisissez la première option du menu précédent, nous allons analyser la variable BAC. Pour cela cliquez sur le nom de cette variable, puis sur Y et validez.



Remarque : Vous pouvez aussi sélectionner, depuis la feuille de données, une ou plusieurs colonnes du fichier de données en cliquant sur leur nom et choisir ensuite le menu Histogram/Bar Chart. Vous aurez alors directement tous les graphiques concernés dans une fenêtre.

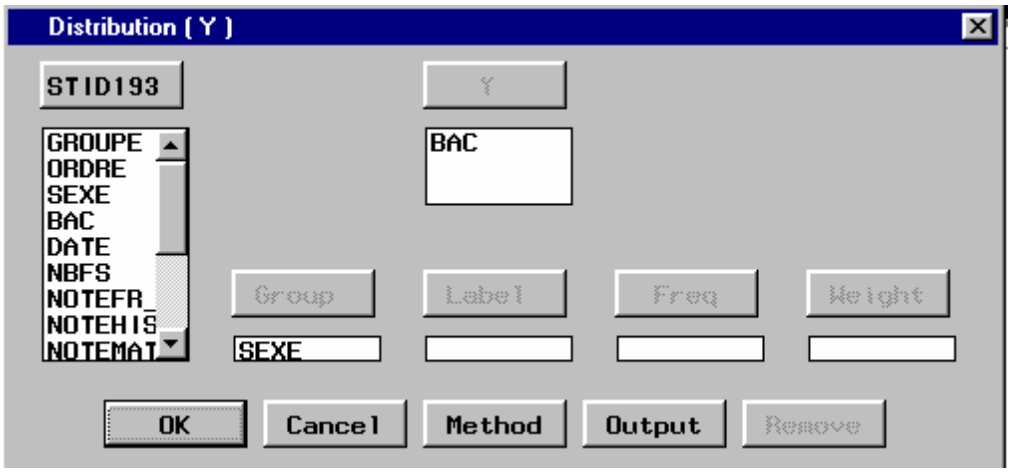
Vous obtenez [53]



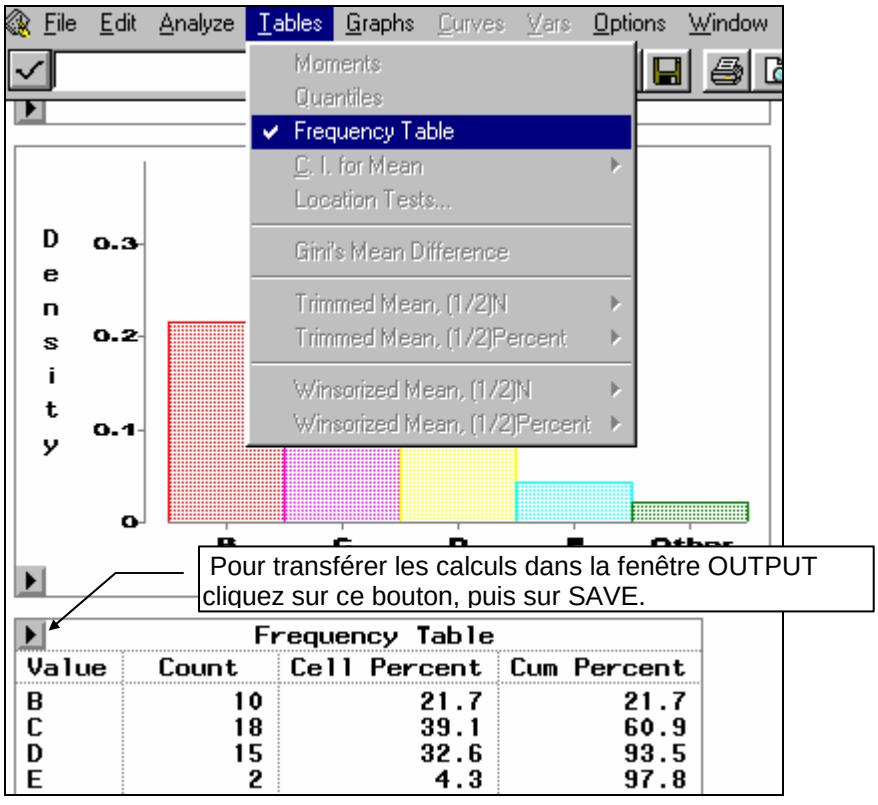
[53] Remarquez que SAS a mis dans une colonne OTHER toutes les modalités « rares ». Il suffit de cliquer sur le bouton METHOD de la boîte précédente pour paramétrer la « rareté ». De plus, vous pouvez en cliquant sur les élément du graphique afficher les effectifs correspondants...

En cliquant sur le bouton droit de la souris, vous faites apparaître un menu contextuel avec lequel vous pouvez afficher les valeurs de la répartition (Values...)

Pour obtenir le tri à plat de la répartition allez dans Analyse/Distribution Y, pour compliquer, nous mettons SEXE comme variable de Groupement.



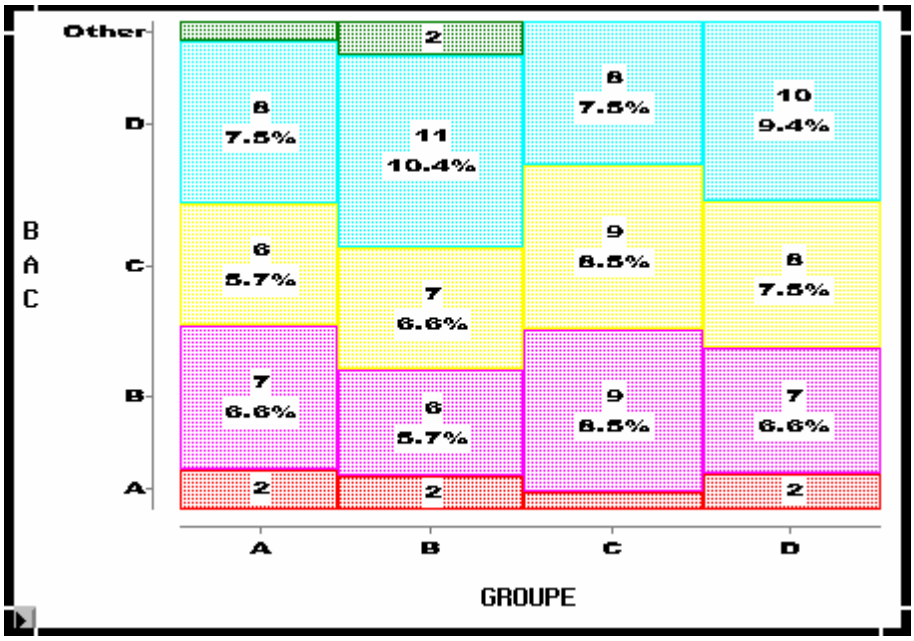
Nous obtenons un graphique du même type que le précédent.



En allant dans Tables/ Frequency counts, SAS ajoute de tableaux de fréquences un pour les hommes et un autre pour les femmes.

Exercice

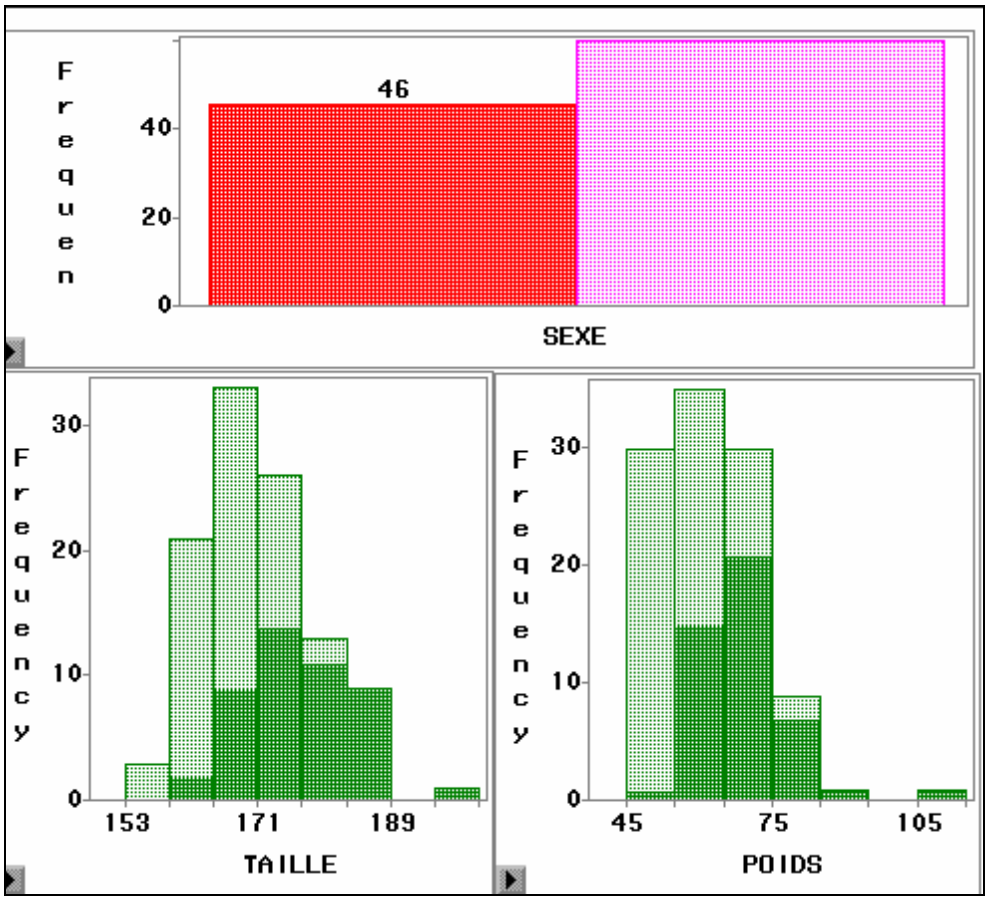
1. Reproduisez le graphique suivant en utilisant la commande MOSAIC PLOT qui permet de faire des diagrammes à bandes à deux dimensions :



Etude graphique de la liaison entre deux variables :

2. Sélectionnez les colonnes SEXE, TAILLE et POIDS de la feuille de données en cliquant sur leur nom avec la touche CTRL maintenue enfoncée. (Assurez vous au préalable que SEXE est bien considérée comme nominale).

Activez ensuite le menu Histogram/Bar Chart. Cliquez ensuite sur la barre du graphique représentant les hommes. Interprétez le résultat obtenu.



Ceci fonctionne sur toutes les fenêtres crées par INSIGHT sauf celles qui sont gelées (Menu Freeze de Windows)

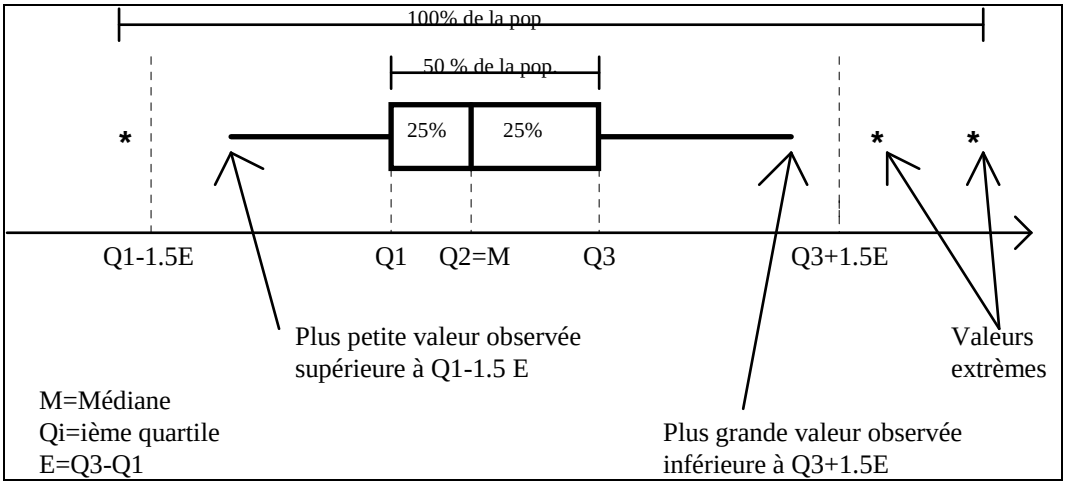
3. Donnez un graphique démontrant l'absence (ou la présence) de liaison entre les variables SEXE et GROUPE.

C. Variable quantitative ; Analyse univariée

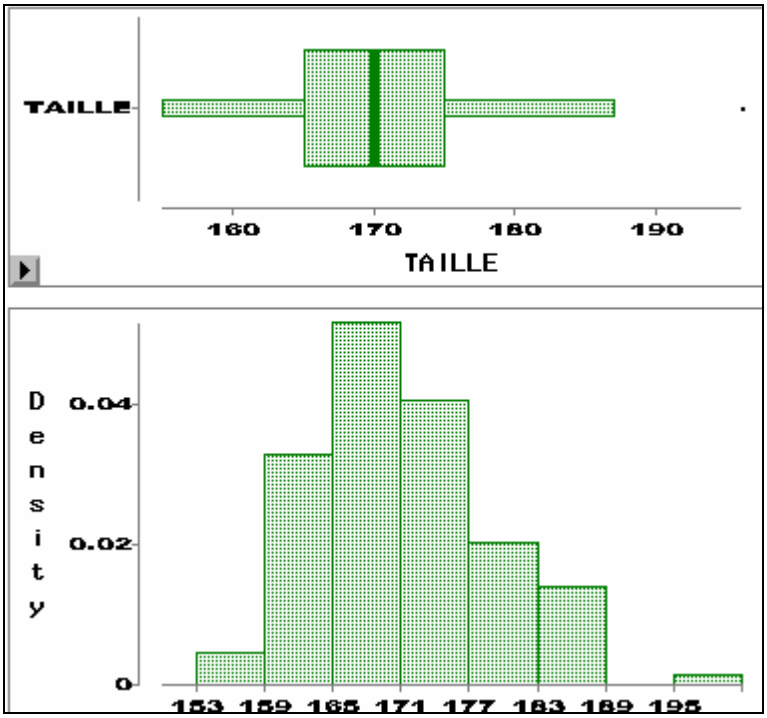
1. Boxplots, histogrammes, moments

Analysons la variable TAILLE. Pour cela allons dans Analyse/Distribution Y et choisissons la variable Taille :

Rappel de la définition du Box Plot :



Sortie SAS :



Moments

N	106.0000	Sum Wgts	106.0000
Mean	170.6981	Sum	18094.0000
Std Dev	7.8391	Variance	61.4509
Skewness	0.4808	Kurtosis	-0.0571
USS	3095064.00	CSS	6452.3396
CV	4.5924	Std Mean	0.7614

Quantiles

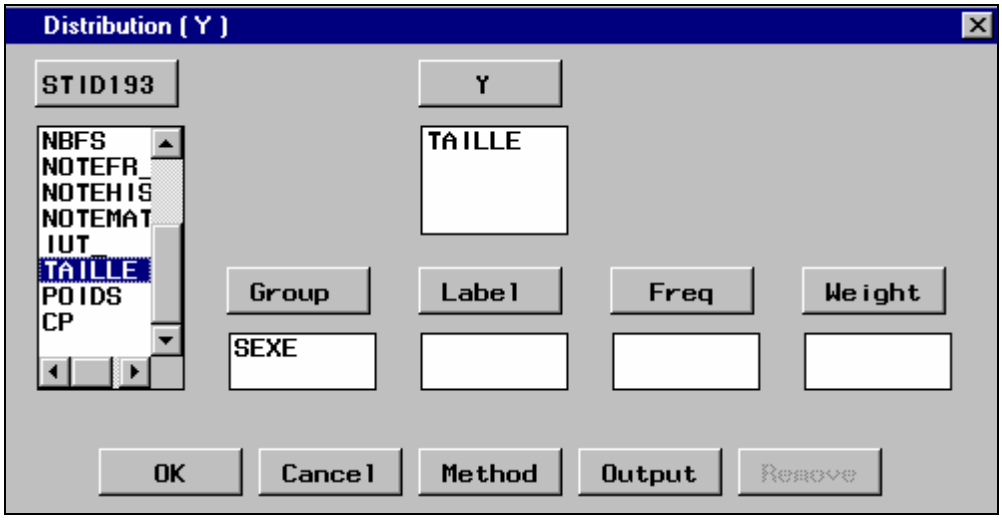
100% Max	196.0000	99.0%	187.0000
75% Q3	175.0000	97.5%	186.0000
50% Med	170.0000	95.0%	184.0000
25% Q1	165.0000	90.0%	182.0000
0% Min	155.0000	10.0%	160.0000
Range	41.0000	5.0%	160.0000
Q3-Q1	10.0000	2.5%	158.0000
Mode	160.0000	1.0%	158.0000

- Interprétez les éléments ci dessus. (Range, Qi etc.) Vous rappellerez les définitions de ces moments.

- Identifiez l'individu hors norme (Box Plot) en cliquant dessus. Refaites les calculs sans lui. (Menu Edit/Observations/Exclude in Calculations, vous pouvez aussi le sortir du graphique Hide in Graph)

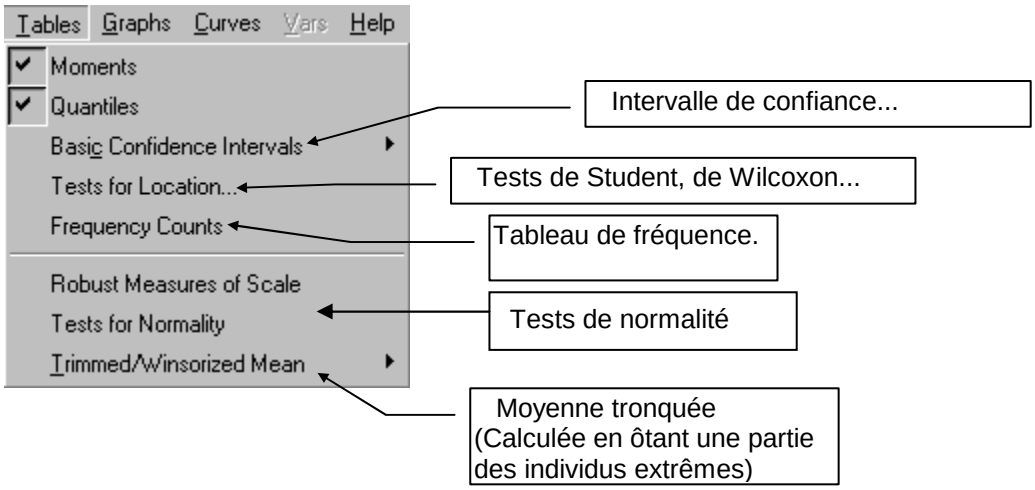
Analyse par groupe

Pour distinguer selon les sexes l'étude précédente, il suffit d'indiquer la variable SEXE dans la case GROUP :



Compléments :

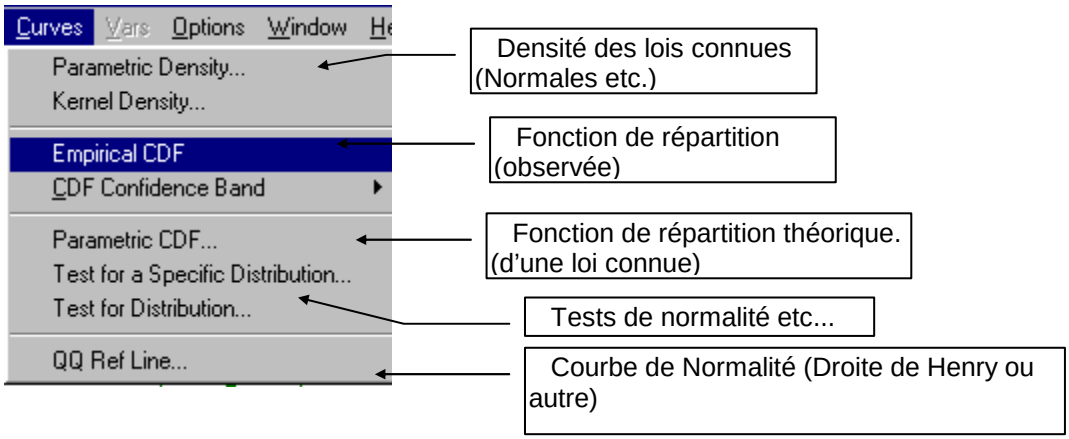
Dans la fenêtre précédente, il vous est possible d 'intégrer différents éléments : (ceux déjà affichés sont indiqués par ☑)



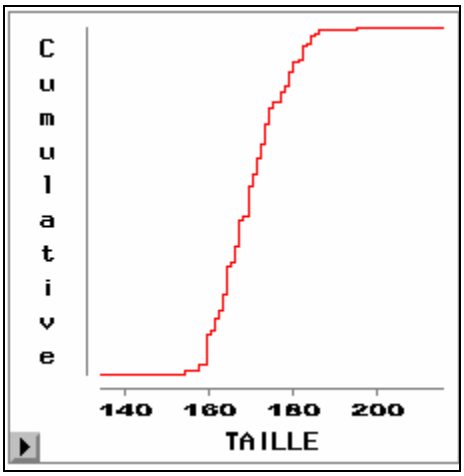
Application : Donnez les statistiques élémentaires sur la variable POIDS en distinguant hommes et femmes. Amusez vous à retirer l'individu le plus pesant et regardez les changements dans les calculs.

2. Fonction de répartition

En allant dans le menu CURVES, vous pouvez intégrer des courbes supplémentaires à la sortie précédentes :

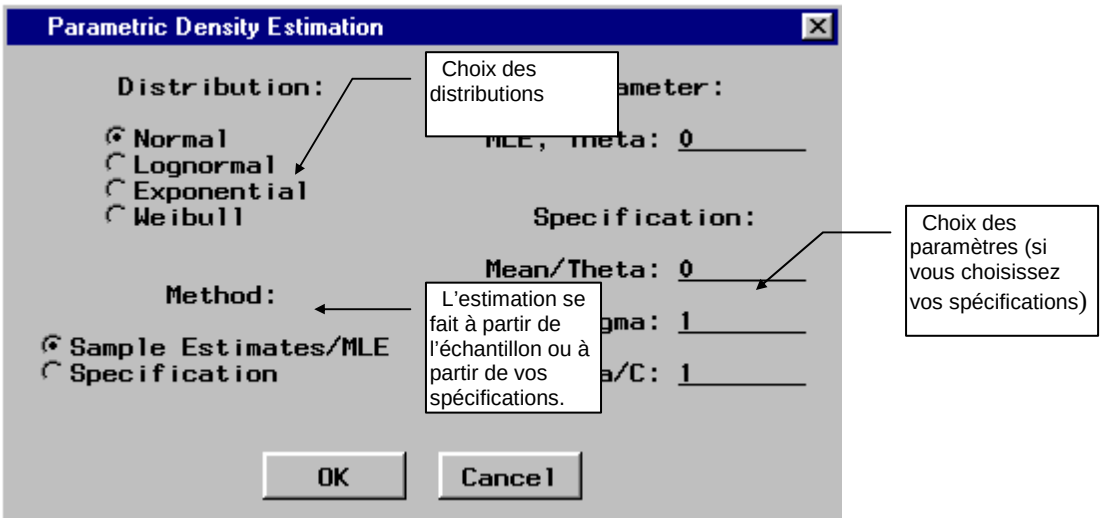


En choisissant Empirical CDF, vous avez la fonction de répartition suivante :

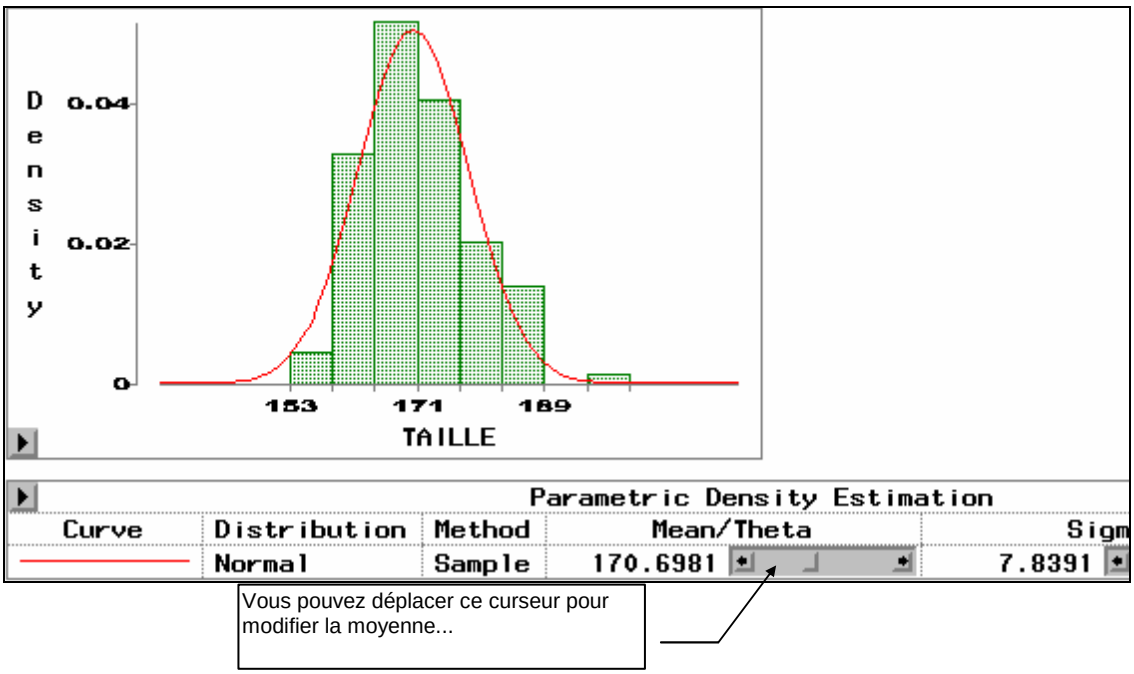


3. Densité de probabilité

Vous pouvez obtenir la superposition de la densité de la loi Gaussienne sur l'histogramme précédent :



Cliquons simplement sur OK



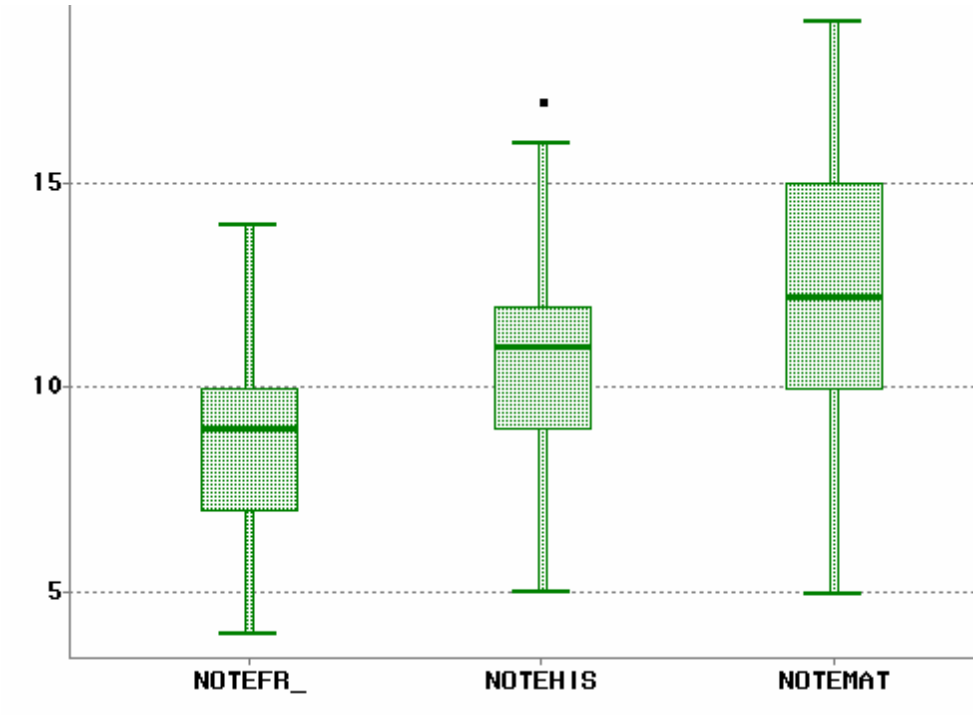
Vous pouvez agir sur les curseurs pour modifier la moyenne et l'écart type. La courbe rouge se modifiera en temps réel.

Complément

Vous pourrez tester l'ajustement de la courbe précédente en allant dans Curves/Test for distribution.
Dans l'exercice suivant, vous allez voir les possibilités de la commande Box plot/ Mosaic Plot

Exercice

En allant dans Analyse/ Box Plot, reproduisez le graphique suivant :



- Identifiez les deux individus hors normes d'histoire géographie.

- Faites des Boxplots illustrant la répartition des notes de maths selon les groupes. Le niveau en maths vous semble-t-il homogène selon les groupes ?

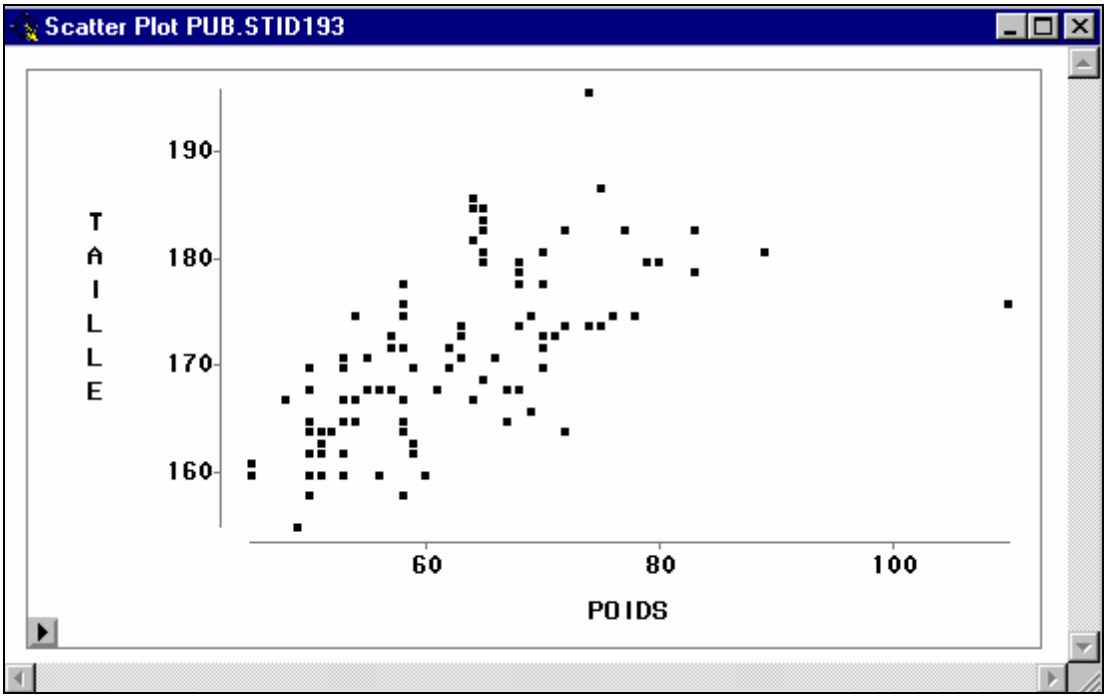
- Y a-t-il une liaison entre les notes et la variables SEXE ? Quel graphique peut-on faire pour s'en rendre compte ?

D. Etude de plusieurs variables quantitatives

1. Nuage de points (scatter plot)

Cette commande permet de tracer des nuages de points. Il suffit d'entrer la variable X et la variable Y.

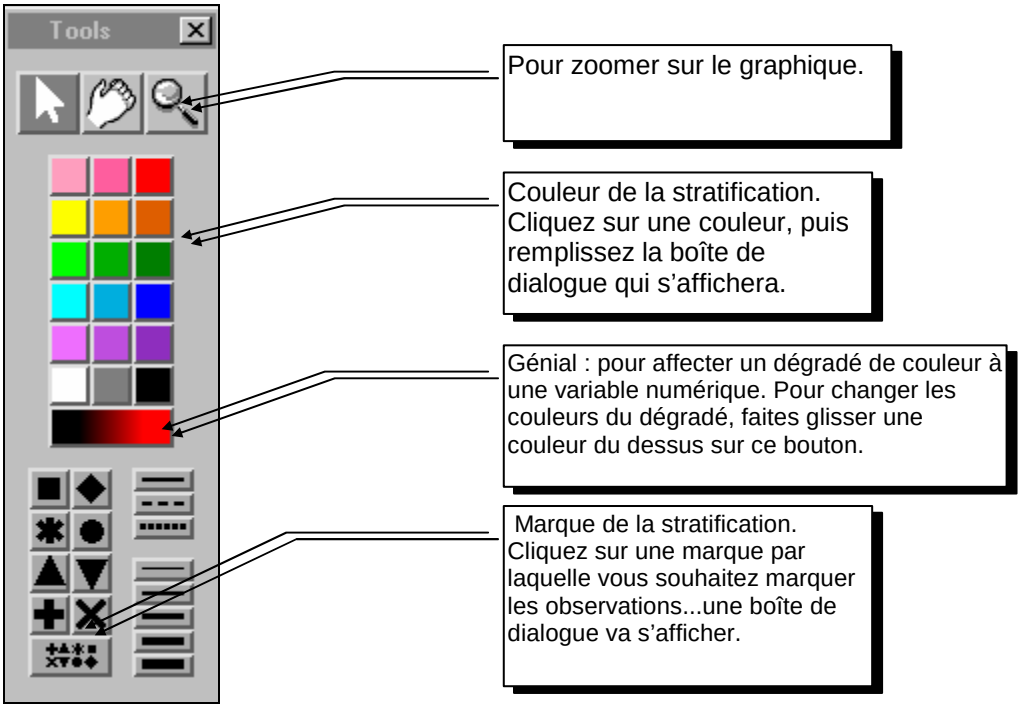
Effectuez un nuage taille poids pour les STID193.



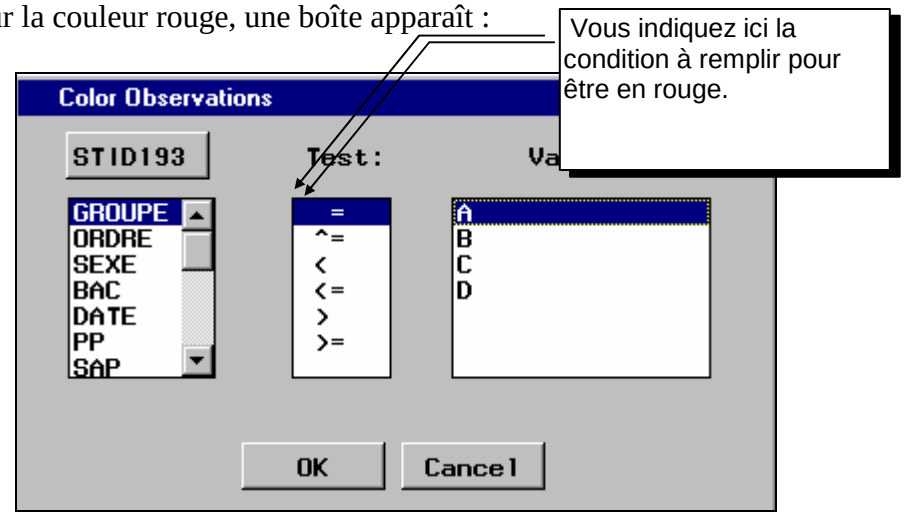
2. Stratification par une variable qualitative, ou quantitative agrégée (TOOL)

Nous souhaitons savoir où se placent les individus du groupe A sur ce graphique.

Nous allons activer la boîte EDIT/WINDOWS/TOOL

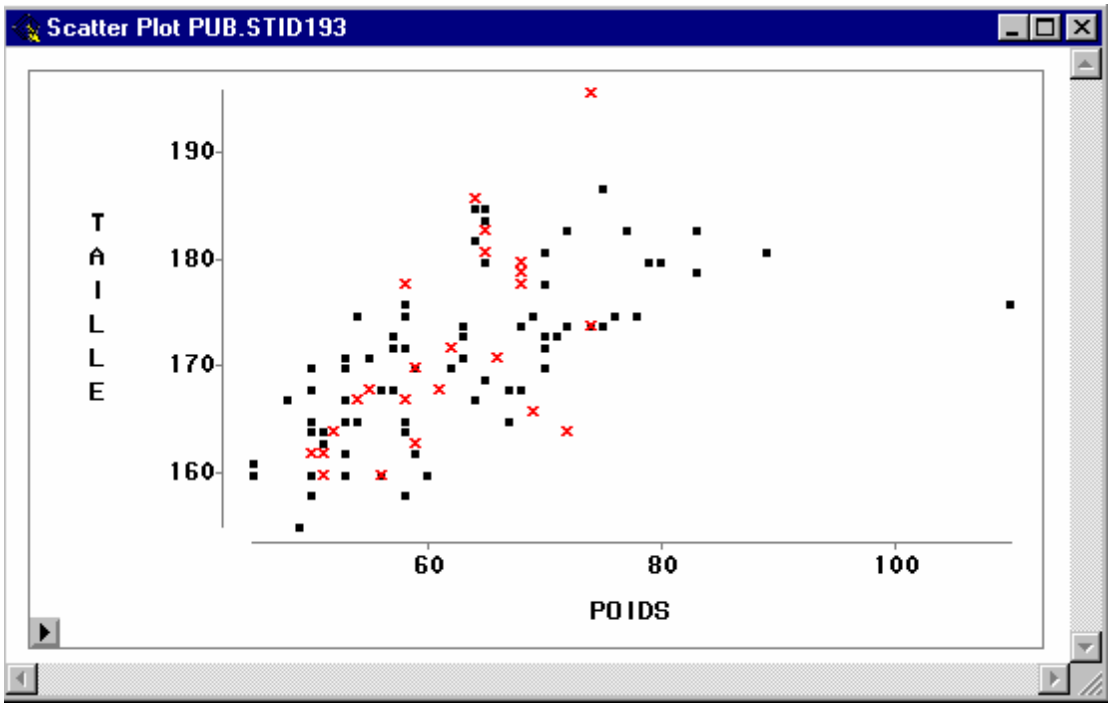


Cliquez sur la couleur rouge, une boîte apparaît :



Mettez en rouge les individus du groupe A. Mettez leur également un symbole en croix.

Le graphique apparaît.



Exercice :

1. Sur le graphique Taille Poids, faites figurer les hommes avec un point bleu et les femmes avec une croix rouge. Activez le Zoom et faites le fonctionner.[54]

2. Représenter graphiquement les trois notes des individus sur un même graphique. On représentera deux notes sur un nuage de points et la troisième en forme d'un dégradé de couleurs. Du rouge pour les individus ayant une mauvaise note (dans la troisième matière) au vert pour les bons . (Cf. Windows/Tool)

Existe il une liaison entre les trois notes ? (les bons en français sont ils automatiquement bons ou mauvais en maths ?...)

3. Le fichier PUB.BANQUE contient des informations sur 50 clients dune banque. La variable SOLD contient le solde moyen sur le compte courant, la variable DEPO contient les DEPO effectués l'an passé sur les comptes d'épargne. La variable NBPR contient le nombre de produits bancaires possédés par le client.

Faites un nuage de points SOLD DEPO en stratifiant par la variable NBPR. (Vous représenterez en dégradé de couleurs les individus suivant leur valeur de NBPR) Méditez...Existe-il une liaison entre ces trois variables ?

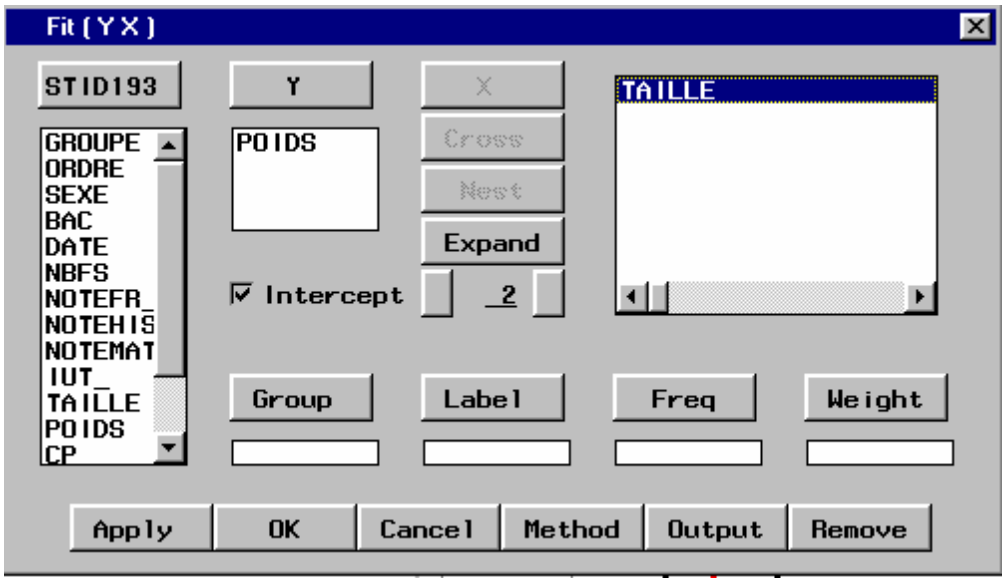
[54] Le curseur de la souris se transforme en loupe. Grosse loupe : en cliquant vous aggrandissez, Petite loupe : en cliquant, vous rétrécissez.

3. Régression (Fit XY)

Cette commande permet de faire des régressions linéaires de tous ordres[55], des ajustement polynomiaux, exponentiels...

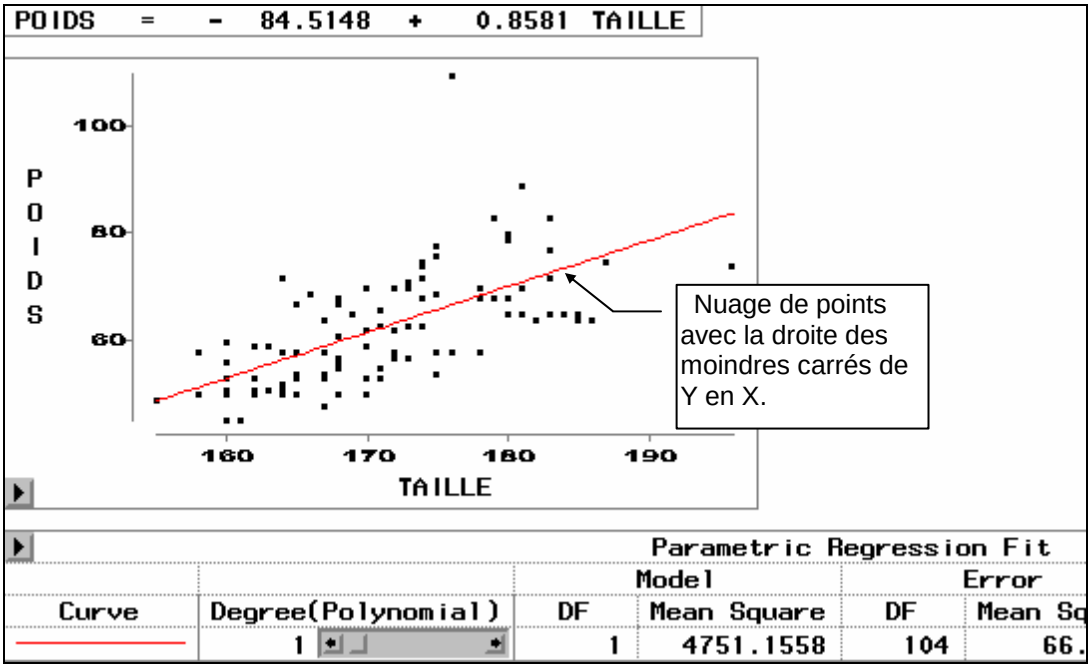
Nous allons étudier la liaison entre deux variables quantitatives : la taille et le poids des Stid193.

Allez dans Analyse/Fit XY et complétez la boîte comme suit :



Vous obtenez les résultats suivants :

55



Vous pouvez vous amuser à déplacer le curseur ci-dessus vers la droite pour augmenter le degré du polynôme ajustant Y en X.[56]

Tout est recalculé automatiquement, y compris le nuage des résidus.

Coefficients de corrélations

Il vous est possible de calculer directement le coefficient de corrélation entre les variables quantitatives en allant dans Analyse/Multivariate.

[56] Si vous passez au degré 2, vous aurez une parabole au lieu d'une droite d'ajustement etc.

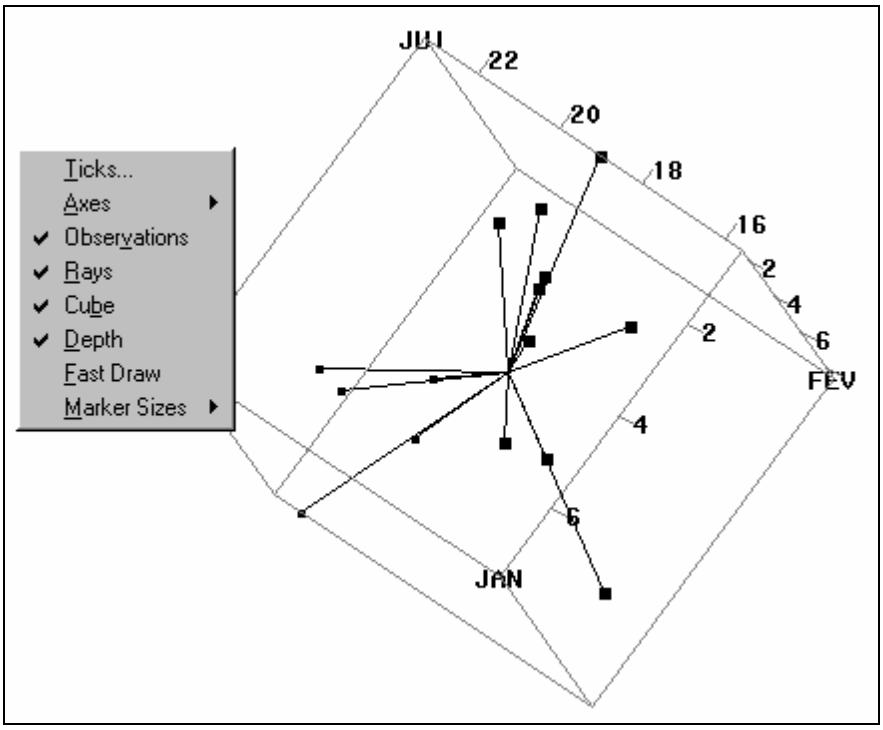
4. Représentation 3D interactive

Ce que nous allons voir ici est plus spectaculaire qu'utile...

Nous allons représenter graphiquement les données du fichier ACP (températures moyennes annuelles de 15 villes de France).

Chargez le fichier ACP. Allez dans Analyse/Rotating Plot ZXY. Choisissez les pour axes : JAN, FEV et JUI qui sont les variables Janvier, Février et Juillet.

Reproduisez le graphique suivant en donnant la signification des options du menu contextuel : Rays, Depth...



- Faites « tourner » le graphique en positionnant le curseur de la souris légèrement en dehors du graphique, il se transforme alors en main. Laissez alors le bouton gauche enfoncé et déplacez la souris.... Génial non ?

- Quelle est la particularité géométrique du nuage de points précédent ? Est-il vraiment tri - dimensionné ?

5. Lancement d'INSIGHT avec le langage SAS

Il est possible de lancer SAS INSIGHT directement à partir du langage SAS avec la procédure PROC INSIGHT.

```
proc insight data=moi.stid193;
  rotate notemat*notehis*notefr_;
run;
```

Vous permet de faire un graphique 3D des notes…

Exercice 1

- Chargez le fichier STID198

- Représentez le nuage Taille Poids (Scatter Plot) et faites un dégradé proportionnel à la pointure (Edit/Window/Tools). Existe-il une liaison entre ces 3 variables.

- Nous allons essayer de prédire la Pointure en fonction de la taille et du poids. Nous allons ajuster un modèle linéaire. Dans le menu FIT XY Mettez Pointure comme variable à expliquer (Y) et TAILLE et POIDS comme predicteurs (X)

- Que représente le graphique avec une grille ? Vous pourrez le faire tourner pour mieux vous rendre compte de la nature de la surface.

Exercice 2

- Chargez le fichier ECHXYZ1. Représentez le nuage 2D XY et le nuage 3D XYZ. Identifiez ensuite visuellement sur les deux graphiques les individus dont le Z est supérieur à 0.9.(windows tool)

- Quelle est la nature des lignes de niveau ?

- La liaison entre Z et X,Y est elle linéaire ?

V. Quelques procédures statistiques

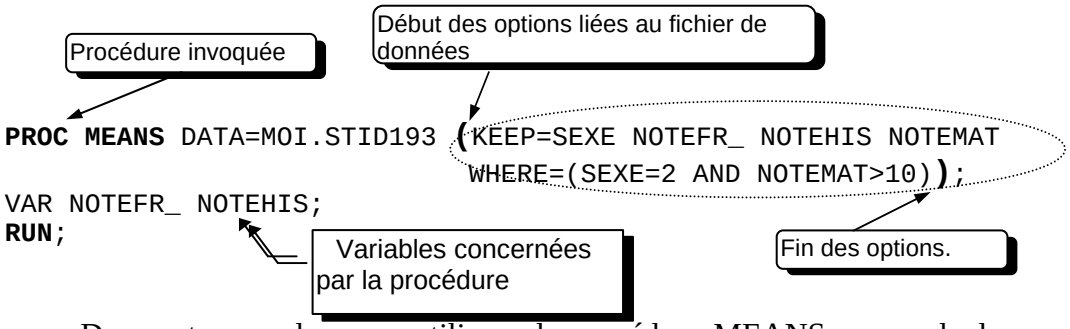
Nous avons vu dans les pages précédentes quelques éléments du langage SAS pour gérer les fichiers de données. Un fichier de données existe souvent dans le but d'être traité, analysé. C'est l'objet des procédures SAS.

Les procédures SAS sont très nombreuses , très puissantes et couvre un très large champ d'applications. Cf. fin de ce document.

L'appel d'une procédure SAS se fait en général par les instructions suivantes :

PROC *nom de la procédure* **DATA=***nom du fichier concerné par le traitement options* **;**
instructions liées à la procédure (modèle pour une régression etc...) **;**
RUN ;

Exemple :



Dans cet exemple, nous utilisons la procédure MEANS pour calculer quelques statistiques simples sur une partie du fichier MOI.STID193.

On ne conserve que les variables sexe, NOTEFR_, NOTEHIS et NOTEMAT et les individus de sexe féminin (2) dont la note de maths est strictement supérieure à 10. Nous calculons, pour ces individus, les moyennes, mini, maxi, écart types des notes de français et d'histoire-géo.

A. SORT (Trier des fichiers)

Elle permet de trier les observations d'un fichier dans l'ordre croissant (par défaut) selon un ou plusieurs critères.
Elle est indispensable pour utiliser l'option BY[57] dans les certaines procédures comme PRINT, UNIVARIATE.

Syntaxe:

Pour trier dans l'ordre croissant

```
PROC SORT (options);
BY variables;
RUN;
```

Pour trier dans l'ordre décroissant

```
PROC SORT(options);
BY DESCENDING variables;
RUN;
```

Les options principales étant:

DATA=nom du fichier (si ce n'est pas le fichier en cours);
OUT=nom du fichier où sera mis le résultat (si ce n'est pas le même)

Exemple:

```
PROC SORT DATA=MONLIB.STID193 OUT=WORK.STIDTRI;
BY SEXE TAILLE;
RUN;

PROC PRINT DATA=WORK.STIDTRI;
RUN;
```

Ce programme va trier et afficher le fichier trié selon les critères Sexe et Taille.

Exercice:

Ecrivez un programme triant STID193 par le BAC et le nombre de frères et soeurs, puis par le nombre de frères et soeurs et le bac. (Vous mettrez le résultat de ce tri dans un fichier temporaire STIDTRI). Visualisez le fichier obtenu dans le deux cas. Y a-t-il une différence ?

---

[57] Cette option permet de répéter des calculs dans différentes sous populations. Par exemple PROC MEANS ; VAR MATH ; BY SEXE ; RUN ; effectuera des statistiques élémentaires sur les notes de maths en distinguant homme et femme. Nous aurons donc deux tableaux de sortie.

B. PRINT (Afficher un fichier dans l'OUTPUT)

Proc PRINT permet d'éditer tout ou partie de votre fichier à partir de certaines variables.

Syntaxe:

PROC PRINT (options)
VAR variables; (Les variables qui seront affichées)
BY variables; (Permet de distinguer selon les sous-populations, **le tableau devra avoir été trié avec SORT préalablement**)
PAGEBY variables; (Provoquera un changement de page à chaque nouvelle modalité de la variable considérée. Cette variable doit aussi figurer dans BY)
SUM variables; (Variables dont on veut effectuer la sommation (salaires, dépenses...))

Les options étant:

DATA= nom du fichier à éditer (sinon c'est le fichier courant)
N=*'texte à afficher'* nombre d'observations à la fin de l'édition ou de chaque "BY" précédée par le texte à afficher.
OBS=*'nom de la colonne'* change l'entête de la colonne OBS selon le nom spécifié.
ROUND arrondit les valeurs
LABEL variable='intitulé' permet d'éditer des intitulés de variables
SPLIT=sert à définir un caractère de saut à la ligne pour Label (cf exemple3)
NOOBS pas d'édition du numéro d'identification de chaque individu.

Exemples:

ex1:

```
PROC PRINT DATA=PUB.STID193;
VAR GROUPE SEXE TAILLE POIDS;
RUN;
```

Ajoutons quelques options :

```
PROC PRINT DATA=MOI.STID193 N='Nombre d individus:' OBS='Numéro '
;
VAR GROUPE SEXE TAILLE POIDS;
RUN;
```

Va donner : (extrait de la sortie)

Numéro	GROUPE	SEXE	TAILLE	POIDS
101	D	2	160	60
102	D	2	165	58
103	D	2	160	53
104	D	2	167	48
105	D	1	172	57
106	D	2	168	50

Nombre d individus:106

ex2:(Remarquez la présence de PROC SORT...à cause de l'utilisation de BY dans PRINT)

```
PROC SORT DATA=MOI.STID193 OUT=WORK.STIDTRI;
BY GROUPE;
RUN;

PROC PRINT DATA=WORK.STIDTRI obs='Numéro' N='Nombre d''individus
dans ce groupe';
VAR SEXE TAILLE POIDS ;
BY GROUPE;
SUM POIDS ;
RUN;
```

Que fait le programme précédent ?

ex3:

```
PROC SORT DATA=PUB.STID193 OUT=WORK.STIDTRI;
BY BAC SEXE;
RUN;

PROC PRINT DATA=WORK.STIDTRI NOOBS;
VAR TAILLE POIDS;
BY BAC SEXE;
LABEL TAILLE='Taille en cm'
      POIDS='Poids en kg';
RUN;
```

Exercices récapitulatifs :

I) Les fichiers NOTE1T, 2T et 3T contiennent les notes de deux groupes de STID en maths au premier,deuxième et troisième trimestres. Il y a des notes manquantes, c'est pour cela que les fichiers n'ont pas le même nombre d'individus.

Chaque individu est identifié par son groupe et son ordre dans le groupe. (GROUPE et ORDRE).

Fusionnez ces trois fichiers dans un seul nommé ENSEMBLE. Calculez les moyennes, mini, maxi de chaque individu. Sortez des statistiques élémentaires sur la promo (moyenne, mediane, quantile au premier, deuxieme et troisieme trimestre puis sur la moyenne annuelle).

II)

Créez un fichier de données SAS ne contenant que les individus de STID193 qui fêteront leur anniversaire d'ici 30 jours (à compter d'aujourd'hui)..

Vous ne conserverez que les variables Groupe, Sexe, Ordre (dans le groupe).

Vous ordonnerez les individus par rapport à la date d'anniversaire (du plus proche au plus éloigné).

Indications pour le II) :

- Vous pouvez créer, à partir de la date de naissance (DATE) et de la date d'aujourd'hui (TODAY()), la date de l'anniversaire de chaque personne. (Voir en annexe les fonctions de date et heure : MDY, YEAR, DAY, MONTH entre autres...)

- Pour le reste, SAS peut calculer la différence entre deux dates. Vous pourrez jouer sur cette différence pour récupérer les individus concernés.

C. TABULATE

Elle permet d'effectuer des statistiques élémentaires et de les afficher en tableaux. Nous pouvons obtenir des tri-croisés très sophistiqués ! ! ![58]

Il est possible d'effectuer des statistiques élémentaires et de les afficher sous forme tabulée..

Syntaxe simplifiée

```
Proc TABULATE <options> ;
  Class variables ;      variables de classe qui seront utilisées dans Table (qualitatives ou quant. Discrètes)
  Var variables ;        variables à analyser (quantitatives) Ne rien mettre pour un tri croisé banal
  Table description de la table à effectuer ;
  By variables ;
  Format var1 format1 var2 format 2... ;      cf. format (annexe)
  Label var1='étiquette1' etc. ;
  Weight variable ;      variable de poids à affecter à chaque individu.
Run ;
```

Les principales options étant :

Data=*nom du fichier de données SAS*
Depth=*niveau maxi de profondeur de la table ;*
Format=*format de chaque cellule du tableau*
Missing Les manquants constituent modalité à part entière. Si cette option n'est pas utilisée les manquants ne sont pas inclus dans les modalités des variables de classement.
Noseps élimine les séparateurs horizontaux
Order=
- **Data** Les modalités des variables de classement sont classées par ordre d'apparition dans le fichier original
- **Freq** Les modalités sont classées par ordre décroissant d'effectifs
- **Internal** (c'est l'option par défaut)Les modalités sont classées par ordre croissant (alphanumérique)

VARDEF= Vous indiquez ici le diviseur utilisé pour le calcul de la variance:

DF	**(n-1) (Choisi par défaut)**
N	(n)
WDF	(somme des poids moins 1)
WEIGHT	(somme des poids)

Vous allez comprendre la syntaxe de commande « table » à l'aide des exemples ci-dessous.

[58] La syntaxe n'est pas toujours simple ! L'outil Enterprise Guide peut alors se révéler utile si l'on est complêtement réfractaire au langage SAS

Exemples de tri croisés simples :

Le programme suivant permet de dresser un tri croisé Sexe*Groupe.

```
proc tabulate data=sasuser.stid193;
class groupe sexe;
```
car nous allons effectuer un tri croisé avec les variables groupe et sexe

```
table groupe, sexe;
run;
```

	SEXE	
	1	2
	N	N
GROUPE		
A	10.00	14.00
B	13.00	15.00
C	12.00	15.00
D	11.00	16.00

Le N à l'affichage indique que SAS a calculé, pour chaque cellule, l'effectif de non manquants.

Comparez avec le programme suivant :

```
proc tabulate data=moi.stid193;
class groupe sexe;
```
car nous allons effectuer un tri croisé avec les variables groupe et sexe

```
table groupe*sexe / condense;
```
le « /condense » demande à SAS de condenser l'affichage au maximum pour limiter le nombre de pages à afficher.

```
run;
```

On a à l'affichage :

GROUPE				
A		B		C
SEXE		SEXE		SEXE
1	2	1	2	1
N	N	N	N	N
10.00	14.00	13.00	15.00	12.00

(CONTINUED)

GROUPE		
C	D	
SEXE	SEXE	
2	1	2
N	N	N
15.00	11.00	16.00

Le N à l'affichage indique que SAS a calculé, pour chaque cellule, l'effectif de non manquants.

Syntaxe simplifiée de la commande table

Elle est obligatoire dans la procédure Tabulate. Elle contient des expressions de une à trois dimensions (var1*var2*var3) séparées par des virgules et éventuellement terminée par un « / » suivi des options.

Les variables indiquées dans « table » sont soient déclarées dans var soit dans class mais pas dans les deux.

Les expressions peuvent être du type :

element*element (croisé)
element element (concaténation=cellules adjacentes)
(element element) (agrégation)

Les éléments sont des variables statistiques ou des statistiques MEAN, SUM, N, NMISS, VAR etc. ***(voir plus loin)***t

Exemple

```
proc tabulate data=sasuser.stid193;
class groupe;
var mat;
table groupe,mean*mat / condense;
run;
```

va donner :

	MEAN
	MAT
GROUPE	
A	12.96
B	11.46
C	12.37
D	13.37

Nous avons la moyenne des notes de maths selon les groupes.

Statistiques disponibles

Symbole SAS	signification
N:	effectif
SUMWGT:	la somme des poids (var WEIGHT)
SUM:	la somme.
MEAN:	la moyenne
VAR:	la variance
STD:	la déviation standard
STDERR:	erreur standard sur la moyenne
Range:	L'étendue (max-min)
PCTN	Pourcentage
PCTSUM	Pourcentage (somme)
USS:	somme des carrés des xi
CSS:	somme des carrés des écarts à la moyenne (cf annexe I)
CV:	coefficient de variation (cf annexe I)
T:	statistique de Student pour tester $\mu=0$
PRT:	c'est le P correspondant au test précédent (bilatéral)

Exercices

Ecrire des programmes SAS donnant les affichages suivants :

a)

	MEAN	STD
	MAT	MAT
GROUPE		
A	12.96	2.54
B	11.46	3.24
C	12.37	3.04
D	13.37	3.71

b)

	MAX		MIN	
	MAT	FRA	MAT	FRA
GROUPE				
A	18.00	14.00	9.00	5.00
B	17.00	14.00	6.00	4.00
C	18.00	14.00	7.00	5.00
D	19.00	12.00	5.00	5.00

c)

	MAX		MIN	
	MAT	FRA	MAT	FRA
GROUPE				
A	18.00	14.00	9.00	5.00
B	17.00	14.00	6.00	4.00
C	18.00	14.00	7.00	5.00
D	19.00	12.00	5.00	5.00
SEXE				
1	18.00	14.00	6.00	5.00
2	19.00	14.00	5.00	4.00

d)

		MEAN	
		MAT	FRA
GROUPE	SEXE		
A	1	13.28	9.89
	2	12.75	8.29
B	1	11.62	8.85
	2	11.33	9.27
C	1	12.25	9.67
	2	12.46	8.20
D	1	14.09	8.73
	2	12.88	8.31

Complément : Quelques options de la commande table :

`Condense` Pour limiter le nombre de pages à afficher

`Misstext='...'` Pour indiquer le texte à afficher dans les cellules contenant les valeurs manquantes.

`Printmiss` Par défaut SAS n'affiche pas les cellules vides, l'option en question permet de les afficher quand même.

D. RANK (Calculs de rangs)

Comme son nom l'indique, elle permet de calculer les rangs de variables quantitative.

Elle peut aussi découper en classes de mêmes effectifs une série de données.

Syntaxe:

```
PROC RANK (options);
  VAR variables; (Les variables dont nous voulons le calcul des rangs. (sinon toutes !)
  RANKS nouvelle liste de variables  (variables ou seront stockés les rangs)
  BY variables;  (Permet la création de sous-groupes... le fichier devra avoir été trié avant)
RUN ;
```

Les options principales étant:

`DATA=` Nom du fichier de données (si ce n'est pas le fichier courant)

`OUT=` Fichier de données contenant les rangs.

`DESCENDING` Pour calculer les rangs dans l'ordre décroissant (croissant par défaut)

`GROUPS = n` Permet d'obtenir un découpage en n classes de même effectif autant que faire se peut. Pour n=4 on obtient les quartiles, n=100 les centiles etc.

`FRACTION` Les rangs sont divisés par l'effectif. On obtient donc la fonction de répartition des données en question.

`TIES=MEAN/HIGH/LOW`

Pour spécifier la procédure à suivre en cas d'ex aequo. MEAN (par défaut sauf si FRACTION est utilisé, c'est alors HIGH) attribue la moyenne des rangs, HIGH le plus grand et LOW le plus petit.

`NORMAL=BLOM` Pour btenir les « scores normaux » qui sont une fonction linéaire des valeurs initiales si la distribution initiale est normale. Ce sont les valeurs que l'on calcule construire une de Henry. On les obtient en calculant $y_i = \Phi^{-1}\left(\frac{r_i - 3/8}{n + 1/4}\right)$ où n est l'effectif, r_i le rang et Φ la fonction de répartition d'une N(0,1).

Exemple:

```
DATA COPY;
 KEEP GROUPE SEXE TAILLE;
 SET PUB.STID193;
RUN;
PROC RANK DATA=COPY;
 VAR TAILLE;
 RANKS RANGTAIL;
RUN;
PROC PRINT;
RUN;
```

Nous faisons une copie du fichier original...

Nous calculons les rangs de la variable taille en stockant le résultat dans rangtail.

ce qui donne :

OBS	GROUPE	SEXE	TAILLE	RANGTAIL
1	A	1	180.00	90.5
2	A	1	168.00	43.5
3	A	1	178.00	85.0
4	A	1	186.00	104.0

E. UNIVARIATE (Analyse univariée)

Pour **l'étude (univariée) de variables quantitatives.** Elle est beaucoup plus complète que la procédure MEANS citée précédemment.[59]

1. Syntaxe:

```
PROC UNIVARIATE (options);
  VAR variables; (Les variables dont nous voulons l'étude)
  BY variables;  (Permet la création de sous-groupes... le fichier devra avoir été trié avant)
  WEIGHT variable;(variable contenant les "poids" de chaque individu. Il vaut 1 par défaut)
  OUTPUT OUT=FICH.SAS (fichier contenant autant d'individus que de modalités de BY et
                       comme variables, les statistiques données par UNIVARIATE) cf. ex
RUN;
```

Les options principales étant:

DATA=	Nom du fichier de données (si ce n'est pas le fichier courant)
NOPRINT	
NORMAL	Effectue un test de normalité (Shapiro-Wilk si n<2000, Kolmogorov si n>2000)
ROUND=	Spécifie la façon d'arrondir les variables.
FREQ	(Edition d'une table avec les valeurs des variables, les %, les % cumulés...)
PLOT	Produit des "graphiques" (Box plot, Stem and leaf...)
VARDEF=	Vous indiquez ici le diviseur utilisé pour le calcul de la variance:
	DF (n-1) (Choisi par défaut)
	N (n)
	WDF (somme des poids moins 1)
	WEIGHT (somme des poids)

Exemple:

```
PROC UNIVARIATE DATA=MONLIB.STID193 NORMAL FREQ;
  VAR TAILLE;
  BY BAC;
  OUTPUT OUT=WORK.SORTIE MEAN=MOY ;
RUN;
```

Ce programme va sortir des statistiques (incluant la normalité et un tableau de fréquence) sur la variable taille en distinguant selon le bac.
Un fichier SAS (WORK.SORTIE) sera constitué, il contiendra la variable bac et les moyennes des tailles dans les différents bacs.
Il suppose le fichier trié compte tenu de la remarque suivante:

[59] Pour des analyses simples de débrousaillage des données, le module SAS/INSIGHT peut être très intéressant à utiliser. Cf. paragraphe SAS/INSIGHT.

Application: Tapez le programme SAS suivant.

```
PROC UNIVARIATE DATA=MONLIB.STID193 NORMAL PLOT FREQ;
VAR TAILLE;
RUN;
```

Examinez tout l'output et comprenez sa signification en vous aidant du tableau suivant:

SORTIE SAS (exemple)		**Symbole SAS utilisé dans l'output**	**signification**
N	106	N:	effectif
Sum Wgts	106	SUMWGT:	la somme des poids (var WEIGHT)
Sum	18094	SUM:	la somme.
Mean	170.6981	MEAN:	la moyenne
Variance	61.45085	VAR:	la variance
Std Dev	7.839059	STD:	la déviation standard (en 1/n-1)
Skewness	0.480848	Skewness:	coefficient mesurant l'asymétrie (cf annexe)
Kurtosis	-0.05708	Kurtosis:	coefficient mesurant l'aplatissement (cf annexe)
USS	3095064	USS:	somme des carrés des xi
CSS	6452.34	CSS:	somme des carrés des écarts à la moyenne (cf annexe I)
CV	4.592353	CV:	coefficient de variation (cf annexe I)
Std Mean	0.761397	STDMEAN:	erreur standard sur la moyenne
T:Mean=0	224.1908	T:	statistique de Student pour tester $\mu=0$
Pr>|T|	0.0001	PROBT:	c'est le P correspondant au test précédent (bilatéral)
Num ^= 0	106		Nombre d'observations non nulles
Num > 0	106		Nombre d'observations strictement positives
M(Sign)	53	MSIGN:	Statistique. utilisée pour tester la nullité de la médiane
Pr>=|M|	0.0001	PROBM:	c'est le P correspondant
Sgn Rank	2835.5	Signrank:	statistique du test des rangs (Wilcoxon) cf ci-dessous
Pr>=|S|	0.0001	PROBS:	c'est le P correspondant.
W:Normal	0.967964	NORMAL:	c'est la stat. utilisée pour tester la normalité (cf ci dessous)
Pr<W	0.0865	PROBN:	et le P correspondant

2. Détails

a) Test de Normalité

NORMAL : statistique testant la normalité de la distribution.

Nous voulons vérifier, à partir des données de l'échantillon, si la distribution observée est compatible avec l'hypothèse d'une distribution Gaussienne (H0).

La statistique employée est celle de Shapiro-Wilk si n<2000 et celle de Kolmogorov-Smirnov pour les valeurs de n supérieures. Vous avez également le P correspondant (PROBN). *Ici P>0.05, nous acceptons l'hypothèse de normalité.*

b) Test de comparaison d'une moyenne à une valeur fixée (0)

T: C'est la statistique du test de Student

$H_0: \mu = \mu_0$ contre $H_1: \mu \neq \mu_0$ (bilatéral) avec pour SAS ici $\mu_0 = 0$

(1 Sample t de Minitab) H0:moyenne=0 contre H1: moyenne≠0. Elle suppose la population normale et l'échantillon aléatoire.

Méthode de calcul: Pour le test, SAS calcule la statistique $T = \frac{\bar{x} - \mu_0}{s / \sqrt{n}}$ avec

$s = \sqrt{\frac{1}{n-1} \sum_i (x_i - \bar{x})^2}$ et $\mu_0 = 0$ la valeur P correspondante. (Si P>0.05, on ne peut rejeter H0 au niveau de signification de 5%, on rejette sinon...)

Vous devez toujours vous ramener à la nullité de la moyenne à tester ou utiliser la PROC TTEST avec l'option H0. Voir page 152

c) Test de Wilcoxon (comparaison d'une médiane à une valeur fixée)

SIGNRANK (Sign rank test, Wilcoxon) c'est la statistique de test calculée par $\sum r_i^+ - n(n+1)/4$ où r_i^+ est le rang de $|x_i|$ obtenu après avoir enlevé les xi=0.

Il permet de tester la nullité de la médiane ou de la moyenne (H0: mediane=0) Nous avons le P correspondant (PROBS).(*Ici, nous rejetons H0 car p<0.05*). Ce test **suppose la distribution symétrique.** Si tel n'est pas le cas, il faut utiliser le test suivant:

MSIGN (sign-test) c'est la statistique de test calculée par M=p-n/2 où n est le nombre de valeurs non nulles et p le nombre de valeurs strictement positives. Cette statistique est utilisée pour tester la nullité de la médiane (H0: mediane=0). Nous avons le P correspondant (PROBM). *Dans le cas présent P<0.05, nous rejetons l'hypothèse nulle*.

3. Exercices

- Testez la normalité des tailles des STID193 (pour l'ensemble puis en distinguant selon les sexes, il sera nécessaire d'utiliser BY et donc de trier le fichier avant !).

- Editez un Stem and Leaf des poids pour les femmes(option Plot).

- En supposant l'échantillon des Stid grenoblois comme un échantillon aléatoire (extrait de la population des Stid de France). Testez l'hypothèse H0: Moyenne taille femme Stid france=165 contre H1 (différent)(Il sera nécessaire de créer une variable Taille - 165 via les instructions DATA et SET pour utiliser UNIVARIATE)

- Comment trouver la P value du test unilatéral obtenu avec H1 :Moy>165 ?

- Que fait le programme suivant ?

```
PROC SORT DATA=MONLIB.STID193 OUT=WORK.TATES;
BY BAC SEXE;
RUN;
PROC UNIVARIATE NORMAL FREQ DATA=WORK.TATES;
VAR TAILLE;
OUTPUT OUT=WORK.ESSAI N=EFF MEAN=MOY NORMAL=TESNORM
PROBN=PNORM;
BY BAC SEXE;
RUN;
PROC PRINT DATA=WORK.ESSAI NOOBS;
SUM EFF;
RUN;
```

F. TTEST (Tests de Student à un ou deux échantillons, appariés ou non)

Cette procédure effectue des tests de Student de comparaison d'une moyenne à une valeur fixée ou comparaison de deux moyennes - dans le cas apparié ou non - ainsi . Dans le cas de comparaison de deux moyennes, SAS effectue aussi un test de Fisherd'égalité de variances. Les deux échantillons figurent dans une colonne (variable) et se distingue par une autre variable spécifiée dans CLASS. [60]

1. Syntaxe simplifiée

PROC TTEST (options);
CLASS variable; (C'est la variable qui permet de distinguer les deux échantillons pour un TTEST à deux échantillons)
PAIRED variable1*variable2 (pour identifier les variables appariées Var1-Var2 va être calculé)
VAR variable(s); (variable à tester, il peut y en avoir plusieurs)
BY variable(s); (donne des tests séparés selon les populations définies par ces variables)
RUN;

(options)
DATA= nom du fichier de données SAS
H0= nombre c'est la valeur de la moyenne à tester.

Exemples

Comparaison d'une moyenne à une valeur fixée

```
PROC TTEST DATA=MOI.STID193 H0=170 ;
VAR TAILLE ;
RUN ;
```

Compare la moyenne de la taille des STID à 170.

Comparaison de deux moyennes (échantillons appariés)

```
PROC TTEST DATA=MOI.CHOLES;
PAIRED AVANT APRES;
RUN;
```

Compare les moyennes de AVANT et de APRES en testant la nullité de leur différence.

Comparaison de deux moyennes (échantillons indépendants)

```
PROC TTEST DATA=MONLIB.STID193;
CLASS SEXE;
VAR TAILLE;
RUN;
```

[60] (Si vos deux échantillons figurent dans deux colonnes distinctes, il faut vous débrouiller pour les mettre dans une seule et créer une variable qui les distinguera)

(Effectue un test de comparaisons des moyennes des tailles des hommes et des femmes).

2. Rappels théoriques

a) Test de comparaison d'une moyenne à une valeur fixée (μ0)

T: C'est la statistique du test de Student

$H_0: \mu = \mu_0$ contre $H_1: \mu \neq \mu_0$ (bilatéral)

(1 Sample t de Minitab) H0:moyenne=μ0 contre H1: moyenne≠0. Elle suppose la population normale et l'échantillon aléatoire.

Méthode de calcul: Pour le test, SAS calcule la statistique $T = \frac{\bar{x} - \mu_0}{s / \sqrt{n}}$ avec $s = \sqrt{\frac{1}{n-1}\sum_i (x_i - \bar{x})^2}$ et $\mu_0 = 0$ la valeur P correspondante. (Si P>0.05, on ne peut rejeter H0 au niveau de signification de 5%, on rejette sinon...)

b) Test de comparaison de deux moyennes (échantillons appariés)

On se ramene au cas précédent en calculant la différence entre les deux variables et en testant la nullité de cette différence.

c) Test de comparaison de deux moyennes (échantillons indépendants)

Hypothèses: Les populations sont supposées normales et les échantillons qui en sont issus sont supposés aléatoires et **indépendants**. De plus, il faut s'assurer de l'égalité ou de l'inégalité des variances pour bien interpréter ce test.

Pour des variances inégales, l'écart type de $\bar{x}_1 - \bar{x}_2$ est estimé par $s = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$

Le test est basé sur la statistique $T = \frac{\bar{x}_1 - \bar{x}_2}{s}$ avec un nombre de degrés de libertés donné par $\frac{(V1+V2)^2}{(V1^2/(n_1-1)) + (V2^2/(n_2-1))}$ (arrondi à l'entier le plus proche) avec $V1 = \frac{s_1^2}{n_1}$ et $V2 = \frac{s_2^2}{n_2}$.

Pour des variances égales, la variance commune est estimée par $s^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1 + n_2 - 2}$ et l'écart type de $\bar{x}_1 - \bar{x}_2$ est estimé par $s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$ Le test est basé sur la statistique $T = \frac{\bar{x}_1 - \bar{x}_2}{s}$ avec $n_1 + n_2 - 2$ degrés de libertés.

Test de Fisher F': SAS effectue un test d'égalité de variances pour vous aider à ne pas vous tromper dans l'interprétation du test de Student. H0: $\sigma_1^2 = \sigma_2^2$ contre H1: $\sigma_1^2 \neq \sigma_2^2$**.** Parmi les deux rapports de variances possibles, on calcule celui qui est supérieur à l'unité. Par exemple si $Stdev1^2 > Stdev2^2$ on calcule $F_{obs} = \frac{Stdev1^2}{Stdev2^2}$ et on rejette l'hypothèse nulle lorsque $F_{obs} > F_{1-\alpha/2}$ où $F_{1-\alpha/2}$ est de fractile d'ordre 1-α/2 d'une loi de Fischer à k1 et k2 d.d.l. où k1 est le nombre de degrés de libertés associés au numérateur et k2 le nombre de ddl associés au dénominateur. SAS indique le P correspondant.

3. Exercices:

a) TTEST à un échantillon

Le fichier COLA (répertoire Public) contient les réponses à un sondage effectué sur 40 personnes prises au hasard à la sortie d'un grand magasin de la banlieue parisienne. 4 questions furent posées:

Q1 Marque préférée de Cola ? 0: Coca 1: Pepsi (dans Colonne n°1)
Q2 Avez-vous déjà acheté Coca-cola ? 0: Non 1: Oui (dans Colonne n°2)
Q3 Aimez-vous les boissons sucrées ? 1: Oui 2: Indifférent 3: Non (dans Colonne n°3)
Q4 Combien de litres de boissons au Cola votre famille a-t-elle consommés le mois dernier ? (dans la colonne n°4)

Les résultats de cette enquête sont-ils compatibles avec une hypothèse de consommation moyenne de 5 litres par mois au niveau de la population étudiée (on pensera d'abord à vérifier la normalité de la VA en question)

b) Deux Echantillons indépendants

Voici un relevé de la teneur en K2O sur 20 échantillons de type 1 et 10 échantillons de type 2 . Peut-on dire (au seuil de 5%) que la moyenne des échantillons de type 1 est égale à celle des échantillons de type 2 ? On détaillera le raisonnement ainsi que les valeurs données par SAS.

	Echantillons	**de type 1**	**(20 valeurs)**	**Echantillons**	**de type 2 (10 val)**
0.80	0.92	1.28	1.52	0.96	1.08
0.84	1.00	1.40	1.56	1.00	1.08
0.88	1.04	1.48	1.88	1.04	1.16
0.88	1.20	1.48	1.92	1.04	1.20
0.92	1.24	1.48	2.20	1.08	1.28

c) Cas d'échantillons appariés

Dans une forêt on choisit 12 arbres au hasard que l'on mesure debout. Ensuite, on les abats, puis on les mesure à nouveau. Chaque arbre a donc été mesuré 2 fois. On veut tester l'égalité des moyennes de ces deux séries pour comparer les deux méthodes de mesure. Y a-t-il une différence significative au seuil de 5% ?

debout	20.4	25.4	25.6	25.6	26.6	28.6	28.7	29	29.8	30.5	30.9	31.1
abattus	21.7	26.3	26.8	28.1	26.2	27.3	29.5	32	30.9	32.3	32.3	31.7

G. FREQ (tris à plat, tris croisés, test d'indépendance du chi2)

Elle produit des tableaux croisés à une ou plusieurs dimensions. Pour les tableaux à deux dimensions, PROC FREQ calcule des coefficients de liaison et effectue des Tests (du Chi 2 entre autres).

Nous n'allons étudier ici qu'une syntaxe simplifiée. Pour plus de détails, reportez vous au SAS *Procédures Guide* (Bibliographie)

1. Syntaxe simplifiée

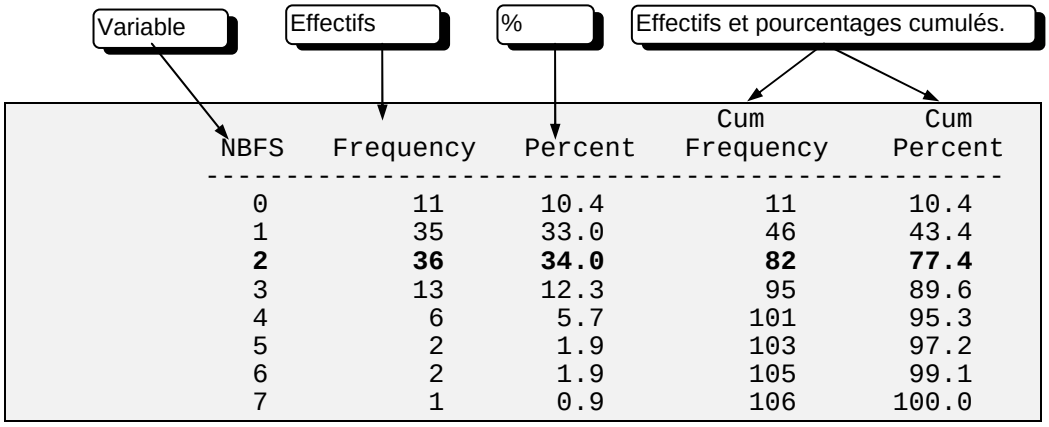
PROC FREQ (data=...);
TABLES 'voir syntaxe plus bas' / option (2);
BY *variables*; (pour distinguer selon les sous-populations...)
WEIGHT *variable*; (pour ajouter un poids à chaque individu, =1 par défaut)
RUN;

2. Exemples

```
PROC FREQ  DATA=MONLIB.STID193;
TABLES NBFS SEXE*GROUPE ;
RUN;
```

Cette commande effectue un tri à plat de nbfs et un tri croisé des variables sexe et groupe :

On obtient la sortie suivante:



NBFS	Frequency	Percent	Cum Frequency	Cum Percent
0	11	10.4	11	10.4
1	35	33.0	46	43.4
2	**36**	**34.0**	**82**	**77.4**
3	13	12.3	95	89.6
4	6	5.7	101	95.3
5	2	1.9	103	97.2
6	2	1.9	105	99.1
7	1	0.9	106	100.0

Ainsi 36 étudiants ont deux frères et soeurs (34% de la population) .
Et 82 étudiants ont au plus deux frères et soeurs (77.4% de la population).

TABLE OF SEXE BY GROUPE

SEXE GROUPE

Légende des cellules du tableau.

SEXE	A	B	C	D	Total
1	10	13	12	11	46
	9.43	12.26	11.32	10.38	43.40
	21.74	28.26	26.09	23.91	
	41.67	46.43	44.44	40.74	
Total	24	28	27	27	106
	22.64	26.42	25.47	25.47	100.00

SEXE	A	B	C	D	Total
2	14	15	15	16	60
	13.21	14.15	14.15	15.09	56.60
	23.33	25.00	25.00	26.67	
	58.33	53.57	55.56	59.26	
Total	24	28	27	27	106
	22.64	26.42	25.47	25.47	100.00

Dans chaque cellule figure l'effectif, le pourcentage et les fréquences conditionnelles (23.33% des femmes sont dans le groupe A, et 53.57% des élèves du groupe B sont des femmes)

Autre exemple

```
PROC FREQ;
tables sexe*(bac groupe);
run;
```

Cette commande effectue deux tri croisés: un sexe*bac et un sexe*groupe, elle est équivalente à sexe*bac sexe*groupe.

3. Quelques options de la commande TABLES

pour enlever certaines informations des cellules de la table:

NOCOL: pour enlever l'affichage des % en colonne (Col Percent) des cellules
NOROW: idem avec les lignes
NOPERCENT: pour enlever le pourcentage (percent) de chaque cellule.
NOFREQ: idem avec les effectifs.
NOPRINT: n'affiche pas de table du tout (utile si vous n'avez besoin que du Chi2)

pour ajouter des informations dans les cases de la table:

CUMCOL: affiche les % cumulés en colonne.
EXPECTED: affiche la valeur espérée de la cellule dans le cas d'indépendance entre les variables lignes et colonne.(cf. calcul du chi2).
DEVIATION: affiche l'écart entre EXPECTED et l'effectif observé.
CELLCHI2: affiche la contribution au chi2 de la cellule.

pour effectuer des analyses statistiques

CHISQ: effectue divers tests du chi2 dont celui de Pearson (celui que vous connaissez). l'hypothèse H0 étant l'indépendance des deux critères de classification. Le chi2 (de Pearson) est celui qui affiché en premier.

Pour sauvegarder les résultats dans un fichier

OUT=*nom de fichier SAS* ; SAS va alors créer un fichier de données contenant les variables de la dernière instruction de TABLE ainsi que les variables COUNT et PERCENT.

Exemple 1

```
PROC FREQ DATA=MONLIB.STID193;
TABLES SEXE*GROUPE / NOROW NOCOL OUT=SORTIE;
RUN;
```

Ce programme effectue un tri croisé sexe groupe. Chaque cellule ne contient que l'effectif et le pourcentage des modalités associées. Les résultats sont stockés dans le fichier WORK.SORTIE. [61] Il est ensuite possible d'exporter ce fichier sous Excel pour effectuer des graphiques illustrant le tableau croisé.[62]

Exemple 2

```
PROC FREQ DATA=MONLIB.DONNEE;
TABLES MARPREF*AIMCOLA /EXPECTED DEVIATION CHISQ NOROW NOCOL
NOPERCENT;
RUN;
```

Ce programme effectue un tri croisé entre marpref et aimcola avec dans chaque cellule l'effectif observé, l'effectif attendu (expected), l'écart entre les deux

[61] Quel que soit le contenu des cellules, ce fichier ne contiendra que l'effectif et la frequence pour tous les couples de modalités considérées.

[62] Il faut alors lancer la commande tableau croisé dynamique d'Excel pour retrouver un tri croisé à partir du fichier exporté.

(deviation). Un test du chi2 est effectué (chisq) comme le montre l'extrait de la sortie ci-dessous. Notons que le chi2 vaut 0.714. La signification P du test vaut 0.700>0.05, on accepte H0: indépendance entre les deux critères.

Statistic	DF	Value	Prob
Chi-Square	**2**	**0.714**	**0.700**
Likelihood Ratio Chi-Square	2	0.726	0.695
Mantel-Haenszel Chi-Square	1	0.024	0.877

Remarque : Si les effectifs théoriques sont insuffisants pour le test du chi deux, on procède souvent à des regroupements. Il suffit, pour SAS, de changer le format d'affichage des variables pour effectuer ces regroupements. Ceci évite de modifier les données originales.

4. Exercice

Chargez le fichier cola (répertoire Public) qui contient les réponses à un sondage effectué sur 40 personnes prises au hasard à la sortie d'un grand magasin de la banlieue parisienne. 4 questions furent posées:

Q1 Marque préférée de Cola ? 0: Coca 1: Pepsi (dans C1)
Q2 Avez-vous déjà acheté Coca-cola ? 0: Non 1: Oui (dans C2)
Q3 Aimez-vous les boissons sucrées ? 1: Oui 2: Indifférent 3: Non (dans C3)
Q4 Combien de litres de boissons au Cola votre famille a-t-elle consommés le mois dernier ? (dans C4)

Existe-t-il une liaison entre Q1 et Q3 ? (Vous donnerez un tableau croisé avec un contenu de cellule approprié et la valeur du chi 2)
Idem entre Q1 et Q2 ? Acheter Coca est-ce l'adopter ?

5. Cas Particulier important, TEST du chi2 sur un tri croisé existant

Ceci est très intéressant lorsque vous n'avez pas les données brutes mais seulement un tri croisé se rapportant à ces données:

Des bouquets floraux de *Golden Delicious* ont été soumis en nombre sensiblement égaux à quatre traitements donnés, et on a compté le nombre de fruits produits dans chaque cas afin de vérifier s'il existe ou non une relation entre les différents traitements et le nombre de fruits produits (la fructification).

	Nombre de fruits produits		
Traitements	**0**	**1**	**plus**
A	203	150	6
B	266	112	1
C	258	126	2
D	196	168	17

(exemple de lecture: 203 bouquets ayant subi le trait. A n'ont produit aucun fruit)

En utilisant un test du χ^2 d'indépendance, nous voulons vérifier si une relation existe entre les deux critères de classification.

Nous n'avons pas le fichier des données originales ici. Nous allons en construire 1 dont le tableau ci dessus serait le tri-croisé associé.

L'idée est simple, nous mettons un individu dans chaque couple possible de modalité et nous affectons comme poids (Weight) à chacun de ces individus l'effectif correspondant. Ainsi, nous mettons un individu ayant subi le traitement A et ayant porté 0 fruit et nous lui donnons un poids de 203 (il sera compté 203 fois !)

```
data golden;                    C'est le nom du fichier
  do trait=1 to 4;              Nous notons les traitements 1 2 3 et 4. (au lieu de A,B,C et D)
    do nbfruit=0 to 2;
      input wt @@;              La double @ (ALTGR 0) permet de mettre chaque valeur dans le
                                fichier avant de continuer la boucle
      output;                   Idem
    end;
  end;
cards;
203  150  6
266  112  1          (Vous écrivez votre tableau ligne par ligne)
258  126  2
196  168  17
;
run;
proc freq;
weight wt;           (Etape importante, nous disons à SAS que chaque individu compte wt fois)
tables trait*nbfruit /chisq noprint;(Nous ne voulons pas l'affichage de la table)
run;
```

Ce programme vous permet d'obtenir un chi2 de 53.72. Tapez ce programme pour vérifier et concluez.

6. Rappels théoriques sur le test d'indépendance du χ^2

Considérons deux variables qualitatives X et Y prenant respectivement p et q modalités (notées 1,2,3...,p et 1,2,3...,q) et observées sur une même population. On dit que ces deux variables sont indépendantes en probabilité si: P(X=i et Y=j)=P(X=i)P(X=j) pour tout couple de modalité (i,j).

Si l'on note nij le nombre d'individus dont le caractère X vaut i et Y vaut j, ni. le nombre d'individus dont le caractère X vaut i et n.j le nombre d'individu dont le caractère Y vaut j on doit avoir en cas d'indépendance: nij/n=ni./n * n.j/n où encore nij=(ni.*n.j)/n

Nous allons tester cette hypothèse d'indépendance (H0) entre X et Y à partir d'un tableau croisé construit sur un échantillon de taille n.

Tableau des effectifs observés

X\Y	1	...	j	...	q	total
1	n11		n1j		n1q	n1.
...		...		...		
i	ni1	...	nij	...	niq	ni.
...		...		...		
p	np1	...	npj	...	npq	np.
Total	n.1	...	n.j	...	n.q	n

Tableau des effectifs théoriques (si H0 vraie)

X\Y	1	...	j	...	q	total
1	$\frac{n_{1.} \times n_{.1}}{n}$		$\frac{n_{1.} \times n_{.j}}{n}$		$\frac{n_{1.} \times n_{.q}}{n}$	n1.
...		...		...		
i		...	$\frac{n_{i.} \times n_{.j}}{n}$	...		ni.
...		...		...		
n		...		...		np.
Total	n.1	...	n.j	...	n.q	n

$$T=\sum_{i,j}\frac{\left(n_{ij}-\frac{n_{i.}n_{.j}}{n}\right)^2}{\frac{n_{i.}n_{.j}}{n}}=\sum_{i,j}\frac{n_{ij}^2}{\frac{n_{i.}n_{.j}}{n}}-n$$

La statistique utilisée mesure "l'écart" entre la distribution observée et la distribution théorique attendue (en cas d'indépendance des deux critères: H0 vraie)

Si les effectifs théoriques de chacune des cases sont supérieurs ou égaux à 5 alors T suit approximativement une loi du χ^2 à (p-1)(q-1) degrés de liberté.
On rejette H0 lorsque la statistique T est trop grande: T> $\chi^2_{(1-\alpha)}[(p-1)(q-1)]$ où $\chi^2_{(1-\alpha)}[(p-1)(q-1)]$ est le fractile d'ordre 1-α d'une loi du χ^2 à (p-1)(q-1) degrés de liberté.

ATTENTION: Les effectifs théoriques des cases doivent être ≥ 5 pour appliquer cette méthode. (Dans le cas contraire, il faut opérer des regroupements)

H. ANOVA et GLM, Analyse de la variance

1. Un exemple

Le fichier ARBRES donne la distribution des hauteurs d'arbres observées pour une même catégorie d'arbres dans 3 types de forêts (notés 1,2 et 3).

Nous voulons savoir si la hauteur des arbres dépend du type de forêt dans laquelle on les a trouvés. Autrement dit, on veut savoir si la hauteur moyenne dépend du type de forêt.[63] L'analyse de la variance à un critère est typiquement adaptée à ce genre de problème.

2. ANOVA à un critère

Elle a pour but de mesurer la liaison entre une variable qualitative (le type d'arbre) et une variable quantitative (la hauteur). Cela revient à comparer les moyennes de k populations qui son supposées **normales**[64] **et de même variance**[65], à partir **d'échantillons aléatoires et indépendants** les uns des autres.

a) Principe de l'ANOVA

Soit Y une variable quantitative et une variable qualitative X prenant k modalités (1,2,...k). Nous supposons que Y suit une loi normale $N(\mu_i, \sigma^2)$ sur chaque sous-population Pi définie par X=i.

La variance totale se décompose de la façon suivante:

$$\frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\overline{y})^2 = \frac{1}{n}\sum_{i=1}^{k} n_i(\overline{y}_i-\overline{y})^2 + \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\overline{y}_i)^2$$

Variance totale = Variance inter + Variance intra

où n_i désigne le nombre d'individus de la sous population Pi, n le nombre total d'individus, $\overline{y}_i$ la moyenne de y dans Pi et $\overline{y}$ la moyenne générale de Y.

Plus la liaison entre X et Y est forte, plus la part de la variance inter est importante et plus la variance intra est faible. La variance intra comptabilise la partie de la variation de Y non expliquée par X.

On définit **le rapport de corrélation R^2 par Variance inter / Variance totale**. Ce nombre, compris entre 0 et 1 mesure l'intensité de la liaison entre X et Y. S'il vaut zéro il n'y a aucune liaison entre X et Y. Inversement, s'il vaut 1, la liaison est parfaite.

---

63 On dit alors que la variable hauteur est discriminante par rapport au type d'arbre.

64 Bien que la normalité des sous-populations fasse partie des hypothèses d'application du test précédent, il faut reconnaître que l'analyse de la variance est peu sensible, dans l'ensemble, à la non-normalité des des populations considérées. Il suffit en pratique d'éviter d'employer l'analyse lorsque les populations-parents sont très différentes des distributions normales, et lorsque ces distributions sont de forme très différentes d'une sous-population à une autre (dissymétries de sens opposés par exemple), surtout pour de petits échantillons. (Dagnélie TMS 2)

65 De même, l'hypothèse d'égalité des variances est d'importance relativement secondaire lorsque les effectifs des échantillons sont d'effectifs tous égaux (ni=nj) (cf le 2 Sample T TEST) Par contre dans le cas d'échantillons d'effectifs inégaux on doit s'assurer de la validité de cette hypothèse (cf T.Test) surtout lorsque les échantillons d'effectifs les plus réduits correspondent aux populations de variance maximum. (Dagnélie TMS 2)

Test d'égalité des moyennes

H0: $\mu_1 = \mu_2 = \ldots = \mu_k$ contre H1: au moins un μ_i est différent des autres.

Statistique utilisée: $F = \dfrac{\frac{1}{k-1}\sum_{i=1}^{k} n_i(\bar{y}_i - \bar{y})^2}{\frac{1}{n-k}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2}$

Lorsque H0 est vraie F suit une loi de Fisher à (k-1, n-k) degrés de liberté. On rejette H0 lorsque Fobs est supérieur au fractile d'ordre 1-α de la loi de Fisher correspondante.

Tableau d'analyse de la variance

La plupart des logiciels présente leurs sorties de la façon suivante:

Source de variation	**Degrés de liberté** (DF)	**Somme des carrés** (sum of squares)	**Carrés moyens** (mean square)	**F**
Inter-classes	k-1	$\sum_{i=1}^{k} n_i(\bar{y}_i - \bar{y})^2$	$\frac{1}{k-1}\sum_{i=1}^{k} n_i(\bar{y}_i - \bar{y})^2$	$F = \dfrac{\frac{1}{k-1}\sum_{i=1}^{k} n_i(\bar{y}_i - \bar{y})^2}{\frac{1}{n-k}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2}$
Intra-classes	n-k	$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2$	$\frac{1}{n-k}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2$	
Total	n-1	$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y})^2$		

Les logiciels reproduisent la table précédente et donne également la p-valeur correspondante.

Règle de décision :

Si F est supérieur à la valeur critique (ou si p est inférieur à la valeur de référence (0.05 en général)), on rejette H0. Les moyennes entre différentes classes sont significativement différentes, la variable X a donc une influence sur Y ou encore, le pouvoir discriminant de X est significatif.

b) Test d'égalité de variances (BARTLETT)

Ajoutons que SAS peut effectuer un TEST de comparaison de variances pour s'assurer que l'ANOVA est applicable raisonnablement.

Un test proposé par SAS est le test de Bartlett.

Comme le F-test, le test de Bartlett est très sensible à la non-normalité des populations parents quels que soient les effectifs des échantillons. De plus il s'agit d'une méthode approximative qui n'est satisfaisante que lorsque les effectifs des p échantillons (aléatoires, indépendants) n1,...,np sont suffisamment grands ($n_i \geq 4$) et que p n'est pas trop élevé par rapport aux ni.

Notations : Nous disposons de p échantillons aléatoires, simples et indépendants, d'effectifs n1, n2,... np:
x11,x12,...,x1p pour le premier échantillon puis x21,x22,..., x2p pour le deuxième puis...
xp1,xp2,...,xpp pour le dernier échantillon.

$SCE_i = \sum_{j=1}^{n_i} x_{ij}^2 - \frac{1}{n_i}(\sum_{j=1}^{n_i} x_{ij})^2$ et $\hat{\sigma}_i^2 = \frac{SCE_i}{n_i - 1}$; $SCE = \sum_{i=1}^{p} SCE_i$ et $\hat{\sigma}^2 = \frac{SCE}{n-p}$ (n: effectif total)

Nous testons $H0$: $\sigma_1^2 = \sigma_2^2 = \ldots = \sigma_p^2$ contre H1: *une des variances est différentes des autres*.

Si l'hypothèse nulle $H0$: $\sigma_1^2 = \sigma_2^2 = \ldots = \sigma_p^2$ est vraie $\hat{\sigma}^2$ est une estimation non biaisée de la variance commune.

La statistique de test est: $T = \frac{(n-p)Ln(\hat{\sigma}^2) - \sum_{i=1}^{p}(n_i - 1)Ln(\hat{\sigma}_i^2)}{1 + \frac{1}{3(p-1)}\left[\sum_{i=1}^{p}\frac{1}{n_i - 1} - \frac{1}{n-p}\right]}$ qui suit approximativement une loi du χ^2 à p-1 degrés de libertés.

On rejette H0 lorsque $t_{obs} \geq \chi^2_{1-\alpha}$ avec p-1 degrés de liberté.

3. Mise en pratique sous SAS

SAS peut effectuer avec une même procédure une Anova et un test d'égalité de variances.[66]

PROC ANOVA ou PROC GLM ?

Les procédures ANOVA et GLM traitent en particulier les analyses de variance à un critère. ANOVA suppose que vos données sont équilibrées, c'est à dire que les sous-populations sont de même effectif (ce qui fait gagner du temps et de la mémoire), GLM s'utilise dans tous les cas mais elle sera beaucoup moins rapide qu' ANOVA dans le cas de données équilibrées.[67]

a) Syntaxe simplifiée.

Appelons Y la variable réponse (quantitative) et X la variable qualitative (Facteur) servant à définir les sous populations.

PROC ANOVA ou GLM; (cela dépend si vos données sont équilibrées ou non)
CLASS *X* variables (qualitatives en gén.) qui définissent vos sous-populations
MODEL *Y= X ;* (modèle d'étude)

MEANS X */ options ;* Pour calculer des moyennes de la variable réponse selon les catégories de la variable qualitative spécifiée. Cette dernière doit figurer dans le modèle. Means permet également de faire un test d'égalité de variances.[68]

Options de Means : **HOVTEST= BARTLETT ou BF ou LEVENE ou OBRIEN**

Pour effectuer un test d'égalité de variance de Bartlett ou de Brown-Forsythe (plus puissant) ou Levene ou O'brien (Levene modifié)

RUN;
QUIT[69];

---

66 Pour les heureux possesseurs de la V 6.12. Pour les autres, il faut utiliser un programme à part pour tester l'égalité des variances. A la fin de ce paragraphe, vous trouverez le code du programme effectuant un test de Bartlett.

67 ANOVA fonctionne pour des données non équilibrées seulement s'il n'y a qu'un facteur (ce qui est le cas ici), mais autant prendre de bonnes habitudes dès maintenant...

68 Nous savons en effet que pour l'hypothèse d'égalité des variances est nécessaire à l'ANOVA.

69 Le Quit est nécessaire car la procédure ANOVA (ou GLM) est intéractive. C'est à dire que vous pouvez, même après l'exécution de l'instruction (RUN compris) ajouter des commandes supplémentaires (tests etc...) Cela à l'avantage d'éviter de refaire tous les calculs, qui peuvent être très long, juste pour une sous-commande que l'on aimerait ajouter.

b) Pour notre exemple...

Le programme suivant effectue une ANOVA sur le fichier Moi.ARBRES.

Nous allons tester l'égalité des moyennes des hauteurs (Y=HAUT) en fonction du type de forêt (X=TYPE) . Nous effectuerons également un test de Bartlett d'égalité de variances.

```
proc anova data=moi.arbres;
 class type;
 model haut=type;                    Choix du modèle.
 means type/ hovtest=bartlett;       Effectue un test d'égalité de variances en prime.
run;
quit;
```

Ce qui donne :

```
Analysis of Variance Procedure

Dependent Variable: HAUT

Source                   DF    Sum of Squares     F Value      Pr > F

Model                     2       48.88143511        7.12      0.0026

Error                    34      116.64883516

Corrected Total          36      165.53027027

                  R-Square             C.V.              HAUT Mean

                  0.295302          7.413828             24.9837838
```

P valeur du test ANOVA. Si P<5%, on rejette l'égalité des moyennes au risque de 5%.

Interprétation

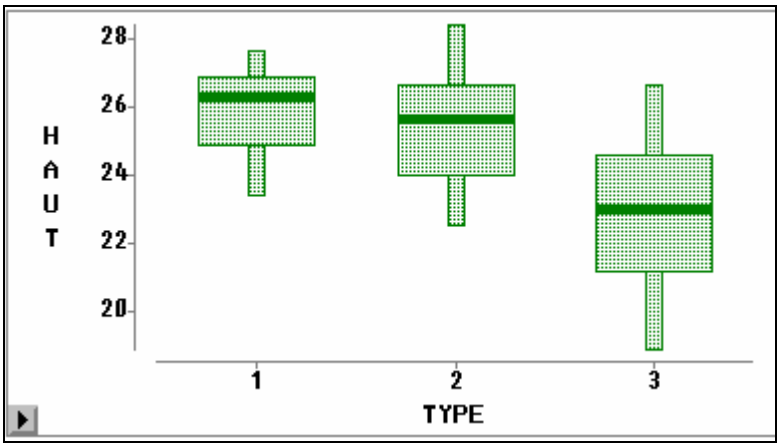
Nous retrouvons la table d'analyse de la variance avec la variable HAUT qui correspond à la variable quantitative Y précédente. (Dependent variable) et quelques informations supplémentaires:

Pr > F:	C'est la signification du test précédent.(<0.05 on rejette H0)
R-Square:	C'est le rapport de corrélation. $\eta^2(X,Y)$ =Somme des carrés inter classe/ Somme des carrés totale)
C.V.:	C'est le coefficient de variation (100*Root MSE/Mean)
Root MSE:	Ecart-type estimé de Y
Mean:	Moyenne de la variable HAUT.

Dans l'exemple ci-dessus, nous rejetons donc l'hypothèse d'égalité des moyennes à condition que les hypothèses d'application de ce test soient remplies. Ci-dessous, nous pourrons vérifier une de ces hypothèses : l'égalité des variances

Remarque :

Le module SAS/INSIGHT peut vous permettre d'obtenir un graphique illustrant cette différence. Allez Dans Global/Analyse/Interactive Data Analysis puis Boxplot [70]



Test d'égalité de variances :

Bartlett's Test for Equality of HAUT Variance			
Source	DF	Chisq Value	Prob>Chisq
TYPE	2	3.4657	0.1768

Cette sortie nous permet de voir que l'égalité des variances semble vérifiée au seuil de 5%. (P=0.1768 > 0.05).[71]

SAS termine en calculant les moyennes et les écart types des hauteurs des arbres dans les trois types de forêts :

Analysis of Variance Procedure

Level of TYPE	N	HAUT Mean	HAUT SD
1	13	25.9692308	1.35854372
2	14	25.3857143	1.77324437
3	10	23.1400000	2.44094699

Nous voyons clairement que les arbres des forêts de type 3 semblent moins hauts que les autres.

[70] Vous pouvez également retrouver l'analyse de la variance, mais la variable type doit être déclarée Nominale avant de faire Analyse/Fit XY

[71] Il faut rester prudent quand même. Le test de Bartlett n'est pas très puissant. (cf. TESTS sous Minitab) Le test de Brown Forsythe est meilleur. (Cf. doc SAS)

4. Exercices

I) La société *FRED&NUCCI frères* spécialiste vinicole réputé de la région bordelaise effectue une étude pour relier la qualité de leur vins en fonction des caractéristiques météorologiques.

Les données sont dans les fichiers Minitab et SAS BORDEAUX.MTW, BORDEAUX (répertoire Public)
C1 : Somme des t° moyennes journalières (en °C)
C2 : Durée d'insolation (en h)
C3 : Nombre de jour de grande chaleur
C4 : Hauteur des pluies.
C5 : Qualité du vin : 1 Bon, 2 Moyen 3 : Médiocre.

La qualité du vin est-elle liée aux variables C1, C2 et C3 ? Vous illustrerez votre raisonnement par des graphiques et des Analyses de variance. On supposera la normalité vérifiée.

II) *Une usine fabrique des billes d'acier selon 3 procédés différents. Nous avons prélevé aléatoirement 3 échantillons de pièces (1 par méthode) et nous avons mesuré leur longueur. Les résultats sont dans le fichier: EXANOVA2.TXT.*

Peut-on considérer que les moyennes des pièces fabriquées selon ces trois méthodes sont les mêmes ? Une ANOVA est-elle indiquée pour résoudre ce problème ?

On détaillera le raisonnement.

APPENDICE : Test de Bartlett sous SAS (version antérieure à 6.12)

Mise en œuvre sous SAS

Voici ci-dessous un programme extrait de la documentation SAS qui vous permet d'effectuer un test de Bartlett fort utile pour s'assurer de l'égalité des variances.

Ce programme est en général fourni avec SAS en SAS/STAT sous le nom « Bartlett.sas ».

Le voici ci-après pour le fichier moi.arbres :

```
PROC SUMMARY NWAY
        data=moi.arbres ; /* calcule et stocke la variance et le */
  CLASS type;             /* nombre d'observations pour chaque niveau*/
  VAR haut;

/* Les lignes ci-dessous ne doivent pas être modifiées */
  OUTPUT OUT=WEARVAR VAR=VARIANCE N=NUM;
  RUN;
DATA _NULL_;
  SET WEARVAR END=EOF;
  LOGVARI=LOG(VARIANCE);
  N=NUM-1;           /* degrés de libertés pour chaque niveau */
  SLOGVAR+LOGVARI*N;
  TOTN+N;
  NVAR=N*VARIANCE;
  SNVAR+NVAR;
  A+1;               /* nombre de niveaux                     */
  SFRACT+1/N;
  IF EOF THEN DO;
    M=TOTN*LOG(SNVAR/TOTN)-SLOGVAR;
    C=1+(1/(3*(A-1)))*(SFRACT-1/TOTN);
    CHISQ=M/C;
    PROBCHI=PROBCHI(CHISQ,(A-1));
    ALPHA=1-PROBCHI;
    FILE PRINT;
    PUT 'TEST DE BARTLETT: CHI-SQUARE=' CHISQ '  ALPHA=' ALPHA '.';
    END;
    RUN;
```

Il donne

TEST DE BARTLETT: CHI-SQUARE=3.4657438688 ALPHA=0.1767759912 .

Nous acceptons donc l'égalité des variances au niveau 0.05.

Vous pouvez aisément le modifier en changeant les lignes 2, 3 et 4 de ce programme. (Ligne 2 : nom du fichier, Ligne 3 : Variable qualitative définissant les sous populations, Ligne 4 : Varaible quantitative)

5. ANOVA à deux critères de classification (modèle fixe)

Nous avons vu que l'analyse de la variance à un critère permet de diviser la variation totale en deux composantes: l'une factorielle, l'autre résiduelle. Nous généralisons ceci pour deux critères de classification. Notons A et B les deux facteurs en question.

a) Tableau d'analyse de la variance

La plupart des logiciels présentent leur sortie sous cette forme:

Source de variation	Degrés de liberté	Sommes des carrés (sum of squares)	Carrés moyens (mean square)	F	P
Facteur A	p-1	SSA	MSA=SSA/(p-1)	Fa=MSA/MSE	
Facteur B	q-1	SSB	MSB	Fb	
Interaction A*B	(p-1)(q-1)	SSAB	MSAB	Fab	
Variation résiduelle	pq(n-1)	SSE	MSE		
Total	pqn-1	SST	MST		

p:nombre de modalités de A
q: nombre de modalités de B
n: nombre d'observations par case (supposé le même)

Nous vous renvoyons à la fiche de cours pour plus de renseignements.

b) Hypothèses générales

Les conditions d'application de l'analyse de la variance à deux critères sont semblables à celles exposées pour la variance à un critère. Il faut être tout aussi vigilant !

c) Mise en oeuvre sous Minitab et sous SAS

Chargez le fichier OPER.MTW (sous Minitab) qui contient les données de l'exercice II de la fiche 4C de statistique.
Nous allons analyser l'influence des facteurs machines et opérateurs sur le temps de réalisation d'un certain travail.

Chaque cellule contenant le même nombre de données (on dit alors que les données sont équilibrées, balanced en anglais) nous pouvons utiliser la commande STAT/ ANOVA /BALANCED ANOVA de Minitab, ou la commande PROC ANOVA de SAS.

Si les données ne sont pas équilibrées, il faut utiliser STAT/ANOVA/GLM pour Minitab et PROC GLM pour SAS.

Il n'y a aucune différence entre GLM et ANOVA dans le cas de données équilibrées si ce n'est le temps de calcul !

Sous Minitab, allez dans STAT/ANOVA/BALANCED ANOVA. Choisissez la variable de réponse.

Ensuite, Minitab vous demande le modèle de l'étude.

Comment spécifier un modèle ? (sous SAS ou sous Minitab)

Le modèle est introduit de la même façon sous Minitab et sous SAS.

Exemples:

à 2 facteurs notés A et B

A B A*B :	modéle complet.
A | B	modèle complet (revient au même que le précedent)
A A*B	on néglige à priori l'effet additif de B.
A B	on néglige l'interaction.

à 3 facteur A,B et C cela donne

A|B|C	modèle complet
A|B|C - A*B*C	modèle A B C A*B A*C B*C (minitab)

d) Syntaxe sous SAS.

PROC ANOVA ou GLM; (cela dépend si vos données sont équilibrées ou non)
CLASSES *variables*; (variables définissant vos sous-populations)
MODEL *variable réponse= modèle*; (modèle d'étude, ici ce serait Y=X)

BY *variables* (pour distinguer des sous populations selon les variables indiquées, le fichier est supposé trié par rapport aux variables indiquées...)

RUN; (pour lancer la procédure)
QUIT; (pour quitter définitivement la procédure)

PUBLIC contient le fichier OPER.TXT. Importez le sous SAS et écrivez la procédure ANOVA correspondante.

Quels sont les renseignements supplémentaires apportés par SAS ?

Vous devez obtenir la sortie suivante:

Analysis of Variance Procedure

Dependent Variable: TEMPS

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	11	888.000000	80.727273	2.16	0.0554
Error	24	896.000000	37.333333		
Corrected Total	35	1784.000000			

R-Square	C.V.	Root MSE	TEMPS Mean
0.497758	11.98059	6.11010	51.0000

Dependent Variable: TEMPS

Source	DF	Anova SS	Mean Square	F Value	Pr > F
MACH	3	756.000000	252.000000	6.75	0.0018
OPER	2	96.000000	48.000000	1.29	0.2948
MACH*OPER	6	36.000000	6.000000	0.16	0.9848

Dans un premier temps SAS fait un test permettant de vérifier globalement la validité du modèle (ce que ne fait pas Minitab)
Dans la deuxième partie, SAS effectue trois tests pour vérifier l'influence des facteurs et de l'interaction sur le modèle.

Le modèle vous paraît-il adapté ici ? Quels sont les facteurs influents ?

I. NPAR1WAY :Quelques méthodes non paramétriques

1. Préliminaires

« *On qualifie de <u>non paramétriques</u> les méthodes statistiques qui sont applicables dans des conditions générales, quant aux distributions des populations parents* » Dagnélie TMS p377. En anglais on traduit « *non parametric ou distribution free* » ce qui est plus parlant.

Elles sont utilisées lorsque les populations parents ne suivent pas une loi normale par exemple. Par contre, d'autres conditions peuvent être exigées (indépendance des échantillons, distributions continues...) Ces tests sont aussi en général moins puissants que leurs confrères paramétriques.

2. Test de Kolmogorov-Smirnov

(comparaison de distributions)

a) But

Ce test a pour but de comparer 2 distributions entre elles. Il est plus complet que le T test qui ne compare que la position de deux distributions (nous testons alors l'égalité des moyennes).

b) Principe

Hypothèses d'application: Nous considérons une variable quantitative observée sur deux populations. Nous avons deux échantillons (de taille n_1 et n_2) constitués d'observations indépendantes respectivement d'une loi G1 et d'une loi G2 . Les distributions Gi sont supposées continues.
H0 : G1=G2 **H1** : $G1 \neq G2$

Nous comparons les fonctions de répartitions(F_1 et F_2) des deux échantillons. D'une manière plus précise, on calcule l'écart maximum entre F_1 et F_2 :

$$D = \sup \left| F_1(x_j) - F_2(x_j) \right|$$

et l'on compare cette valeur à des valeurs critiques particulières.
Si D>c on rejette H0 au risque α

Remarque : Le test présenté ici est bilatéral. Il existe aussi une version unilatérale (cf. vos fiches de Stat)

c) Mise en pratique sur SAS (Proc NPAR1WAY).

Note importante :

SAS calcule en fait plusieurs statistiques dont celle de Kolmogorov présentée ci-dessus. Nous n'avons présenté que la comparaison de deux échantillons. SAS peut comparer n échantillons. Il calcule alors la fonction de répartition moyenne F de tout l'échantillon et calcule des « différences » entre les F_i et F.

Calculs pour Kolmogorov-Smirnov

On a : $F = \frac{1}{n}\sum_i n_i F_i$ $KS = \max_i \sqrt{\sum_i \frac{n_i}{n}(F_i(x_j) - F(x_j))^2}$ et $KSa = KS\sqrt{n}$.

En fait, dans le cas de deux échantillons, on a : KS=$\sqrt{\frac{n_1 n_2}{n^2}}D$. Et $KSa = KS\sqrt{n}$. SAS calcule ensuite le P correspondant à Ksa. (pour 2 échantillons)

Calculs pour Kuiper

Les hypothèses sont analogues à celles du test précédent.

Dans le cas de deux échantillons, SAS donne la statistique de Kuiper calculée par :
$K = \max_j (F_1(x_j) - F_2(x_j)) - \min_j (F_1(x_j) - F_2(x_j))$ et la valeur asymptotique Ka= $K\sqrt{\frac{n_1 n_2}{n}}$. Si Ka>c, on rejette H0 (l'identité des distributions). SAS calcul le P relatif à Ka.

Comparaison des tests de Kuiper et de Kolmogorov :

« *Des études par simulation ont montré que le test de Kuiper était plus puissant que celui de Kolmogorov-Smirnov* » (Statistique non paramétrique et robustesse, Lecoutre-Tassi P328)

Syntaxe de la procédure PROC NPAR1WAY

PROC NPAR1WAY <*options*> ;	
CLASS *variables ;*	On indique ici la variable de classification (qualitative en général)
BY *variables ;*	Pour effectuer si nécessaire plusieurs analyse d'un coup selon les modalités des variables indiquées ici.
VAR *variables ;*	On indique ici la variable de réponse à utiliser dans les tests. Par défaut SAS utilisera toutes les variables quantitatives.
RUN ;	

Les principales options étant :

DATA= *Le nom du fichier de données SAS à utiliser*

EDF *Pour indiquer à SAS d'effectuer des tests utilisant les fonctions de répartitions : Kolmogorov, Cramer Von Mises, Kuiper.*

WILCOXON *Pour obtenir des méthodes sur les rangs (Wilcoxon Mann et Whitney, Kruskal-Wallis)*

d) Exemple

Nous allons tester l'égalité des distributions des tailles des hommes et des femmes de STID de France dont un échantillon (supposé aléatoire) est contenu dans le fichier STID193.

```
Libname moi 'Z :\TOTO' ;
Libname moi 'Z :\TOTO' ;

proc npar1way data=moi.stid193 edf;
   class sexe;
   var taille;
run;
```

On obtient alors :

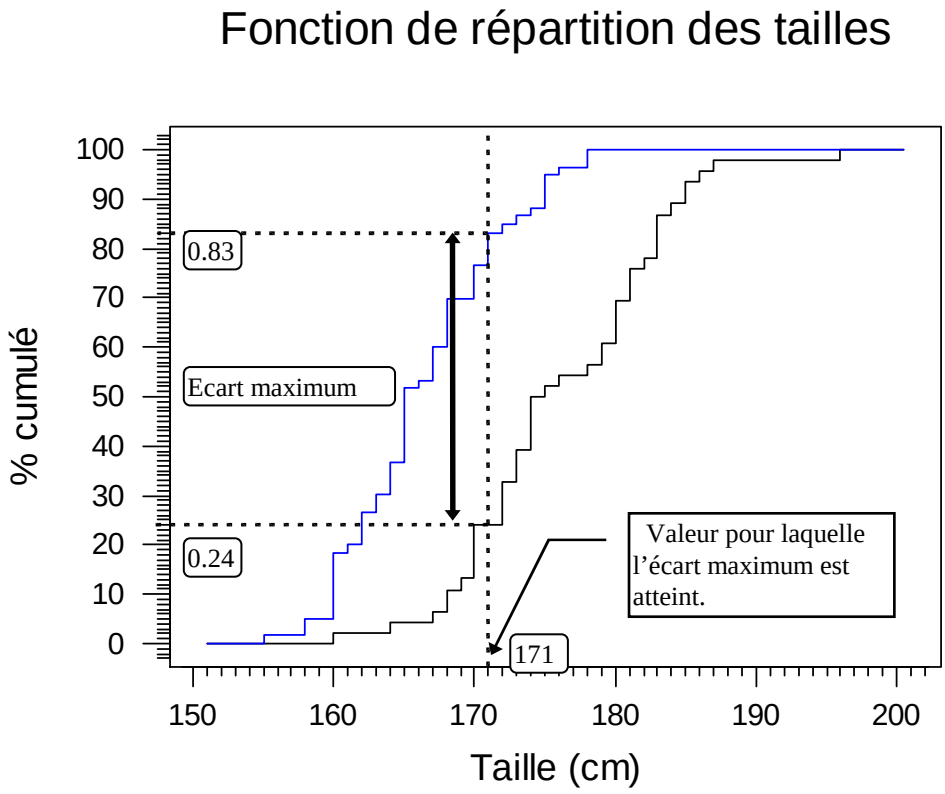
```
        Kolmogorov-Smirnov Test for Variable TAILLE
               Classified by Variable SEXE

                                        EDF              Deviation
                                                          from Mean
 SEXE                 N          at maximum          at maximum
 1                   46          0.239130435          -2.28117743
 2                   60          0.833333333           1.99738764
 --------          ----          -----------
                    106          0.575471698
Maximum Deviation occurred at Observation   59
Value of TAILLE   at maximum  171.000000

        Kolmogorov-Smirnov 2-Sample Test (Asymptotic)
              KS = 0.294499    D = 0.594203
          KSa =  3.03205    Prob > KSa =   0.0001
```

Fonction de répartition.

P valeur

Stat de test.

Comme nous le voyons sur le graphique ci-dessous, l'écart maximum entre les deux fonctions de répartition est obtenu pour T=171cm. On a alors Pour les homme F(171)=0.239 et pour les femmes F(171)=0.833.



On obtient D=0.5942. et KS=0.294499. Le calcul de la signification du test est effectué sur Ksa.

La signification du test montre que l'on rejette H0 (identité des distributions) au risque 0.0001.

Le test de Kuiper confirme-t-il ce résultat ?

```
                Kuiper Test for Variable TAILLE
                 Classified by Variable SEXE
                                        Deviation
 SEXE                    N              from Mean
 1                      46             0.000000000
 2                      60             0.594202899

                Kuiper 2-Sample Test (Asymptotic)
 K = 0.594203    Ka =  3.03205    Prob > Ka =   0.0001
```

e) Exercice

On fabrique des pièces métalliques selon deux procédés différents. On tire deux échantillons aléatoirement, on mesure les pièces obtenues. Les résultats figurent dans le fichier NPAR (format SAS). Il contient les deux échantillons aléatoires indépendants de longueurs (LONG) de pièces métalliques fabriquées selon deux méthodes différentes (numérotées 1 et 3).
Peut-on considérer que les deux méthodes donnent des résultats identiques quant à la distribution des pièces ? (Vous effectuerez les tests de Kolmogorov et de Kuiper)

Une étude complémentaire a montré que les deux distributions peuvent être considérées comme normales. Quel test pouvez-vous effectuer ? Que concluez-vous ?

Complément : Aurait-on pu effectuer un Test de Mann et Whitney ici ?

3. Test de Mann et Whitney (ou Wilcoxon ou White)

(Comparaison de 2 distributions quant à leur position)

a) But

Ce test a essentiellement pour objet de comparer deux populations en ce qui concerne plus particulièrement leur position et est donc comparable au T test de comparaison de moyennes vu dans un chapitre précédent.

Contrairement au T Test, il ne suppose pas la normalité des populations parents mais il suppose des distributions de même « forme ». Il est un tout petit peu moins puissant que le Ttest dans les conditions d'application de ce dernier.

b) Mise en oeuvre

Hypothèses d'application (cf. fiche de cours): Nous considérons une variable numérique ordinale observée sur deux populations indépendantes. Nous avons deux échantillons (de taille n_1 et n_2 , on suppose $n_1 \leq n_2$) constitués d'observations indépendantes respectivement d'une loi G1 et d'une loi G2 . Les distributions Gi sont supposées continues et leurs fonctions de répartitions ne doivent pas se « chevaucher » (Minitab suppose qu'elles ont la même forme donc des variances égales)

H0 : G1=G2 **H1** : $G1 \neq G2$

Réalisation:

La réalisation de ce test est basée sur le classement de l'ensemble des observations par ordre croissant, la détermination du rang de chacune d'elles, et le calcul de la somme des rangs (Y1 et Y2) pour chaque échantillon. (en fait seul le calcul de Y1 est nécessaire)

Le principe du test consiste à rejeter l'hypothèse d'identité des deux distributions lorsque la valeur observée par Y1 s'écarte trop de la valeur attendue correspondante.

Pour des effectifs suffisamment élevés ($n_1+n_2>30$) la distribution de Y1 est approximativement normale de moyenne $n_1(n_1+n_2+1)/2$ et de variance $n_1n_2(n_1+n_2+1)/12$.

La quantité $u_{obs} = \dfrac{\left|Y_1 - n_1(n_1+n_2+1)/2\right|}{\sqrt{n_1 n_2 (n_1+n_2+1)/12}}$ suit approximativement, en valeur absolue une N(0,1).

Correction en cas exaequo : « *La présence de valeurs identiques dans les deux échantillons diminue en fait la variabilité de la somme des rangs : la variance de ces sommes doit donc être corrigée en conséquence. L'importance de cette correction est en général secondaire.* » Dagnélie TMS2 p 384.

c) Exemple sous SAS

Voici un exemple de sortie sous SAS. (Il ne faut pas oublier de mettre l'option WILCOXON dans la procédure NPAR1WAY)

```
              N P A R 1 W A Y  P R O C E D U R E

          Wilcoxon Scores (Rank Sums) for Variable HAUT
                   Classified by Variable TYPE

                        Sum of     Expected      Std Dev         Mean
TYPE          N         Scores     Under H0     Under H0        Score

1            13    204.500000        182.0   20.5948573   15.7307692
2            14    173.500000        196.0   20.5948573   12.3928571
                   Average Scores were used for Ties
       Wilcoxon 2-Sample Test (Normal Approximation)
       (with Continuity Correction of .5)
       S=  204.500    Z=  1.06823    Prob > |Z| =   0.2854
```

SAS calcule bien sûr la somme des rangs (sum of scores), les sommes attendues si H0 est vraie (expected under H0).

« `Average Scores were used for Ties` » signifie que des rangs affectés aux exaequo correspondent à la moyenne des rangs correspondants.

S Correspond au Y1 précédent (somme des rangs du plus petit échantillon)
Z Statistique de Test.
P Le P correspondant.

Ici, on accepte l'hypothèse H0 d'identité des distributions.

Remarque importante : SAS effectue aussi un test de Kruskal et Wallis. Ce test, que nous verrons plus loin, généralise Mann et Whitney à k échantillons. Cela dit, il s'applique lorsque k=2 et est alors équivalent au test de Mann et Whitney.

Minitab donne quant à lui:

```
          Mann-Whitney Confidence Interval and Test
type1      N =  13     Median =      26.300
type2      N =  14     Median =      25.650
Point estimate for ETA1-ETA2 is       0.600
95.1 Percent C.I. for ETA1-ETA2 is (-0.700,2.000)
W = 204.5
Test of ETA1 = ETA2  vs.  ETA1 ~= ETA2 is significant at
0.2857
The test is significant at 0.2854 (adjusted for ties)
Cannot reject at alpha = 0.05
```

Note : Minitab interprète ce test comme un test d'égalité de médianes.

d) Exercice

1°) Nous avons calculé les durées de vie de composants électroniques fabriqués selon deux méthodes différentes.
Nous avons extrait deux échantillons indépendants. Ils sont consignés dans le fichier NONPARA (au format SAS). (NONPARA.MTW Minitab)
On veut savoir si les deux méthodes donnent des résultats analogues.

Peut-on effectuer un test paramétrique ici ?
Que donne Mann et Whitney ici ? (Vous vérifierez les hypothèses d'application sous Minitab)
(Pourrait-on effectuer un test T ? Que donne-t-il ici ?)

2°) Pourrait-on appliquer Mann et Whitney à NPAR.MTW vu précédemment ? Que donne-t-il ?

4. Le test de Kruskal et Wallis

(comparaison de k distributions quant à leurs positions)

a) But

Il généralise Mann et Whitney pour k échantillons. Il est, comme le test précédent, basé sur le classement des observations, la détermination de leur rang et les calcul des sommes Yi associées.

b) Mise en œuvre

Hypothèses : Nous considérons une variable numérique ordinale observée sur k populations indépendantes. Nous avons k échantillons (de tailles n_i) constitués d'observations indépendantes de loi Gi. Les distributions Gi sont supposées continues et de même forme (cf. Mann et Whitney).
H0 : G1=...=Gk **H1** : H0 ^{c}

On calcule $\chi^2_{obs} = \frac{12}{n(n+1)} \sum_{i=1}^{k} \frac{Y_i^2}{n_i} - 3(n+1)$ avec $n = \sum n_i$

Cette quantité suit un chi2 (k-1) pour n suffisamment grand (15).

On rejette H0 lorsque cette valeur est « trop grande ».

```
              N P A R 1 W A Y  P R O C E D U R E
          Wilcoxon Scores (Rank Sums) for Variable DUREE
                   Classified by Variable TYPE

                        Sum of     Expected      Std Dev         Mean
TYPE          N         Scores     Under H0     Under H0        Score
1            30   1050.50000        915.0   67.6359272   35.0166667
2            30    779.50000        915.0   67.6359272   25.9833333
                  Average Scores were used for Ties
      Wilcoxon 2-Sample Test (Normal Approximation)
      (with Continuity Correction of .5)
      S=  1050.50    Z=  1.99598    Prob > |Z| =   0.0459
      T-Test approx. Significance =     0.0506

      Kruskal-Wallis Test (Chi-Square Approximation)
      CHISQ=  4.0135    DF=  1    Prob > CHISQ=     0.0451
```

Voici un exemple de sortie avec le fichier précédent. Nous pouvons remarquer que les conclusions sont analogues à celles du test de Mann et Whitney. Ceci qui est logique (ces tests sont équivalents pour k=2)

J. **CORR ,** calcul des coefficients de corrélations

Objet

Proc CORR est utilisée pour calculer les coefficients de corrélations entre des variables. Par défaut le coefficient de corrélation linéaire de Pearson[72] donné par $r_{xy} = \frac{\text{cov}(X,Y)}{\sqrt{\text{var}(X)\,\text{var}(Y)}}$ est calculé pour chaque couple de variables[73].

1. Syntaxe simplifiée

```
PROC CORR options1;

  VAR variables;      Liste des variables sur lesquelles se feront les calculs des coefficients
  BY variables;       Pour effectuer un calcul pour chaque sous-population définie par BY
  FREQ variables;     Chaque observation est comptée n fois où n est la valeur de la variable pour
                      l'observation correspondante
WITH variables;       Pour calculer les corrélations entre ces variables et celles de VAR. cf exemple
RUN;
```

Exemples

```
PROC CORR DATA=moi.stid193 NOPROB;   (NOPROB: pour ne pas afficher le test)
  VAR taille poids;
RUN;
```

Ce petit programme effectuera le calcul du coefficient de corrélation linéaire (de Pearson donné par $r_{xy} = \frac{\text{cov}(X,Y)}{\sqrt{\text{var}(X)\,\text{var}(Y)}}$) entre les variables Taille et Poids du fichier Stid193. Le résultat est donné ci-dessous sous forme matricielle. Il est précédé par des statistiques élémentaires:

```
                          Simple Statistics

Variable          N       Mean    Std Dev        Sum    Minimum    Maximum

TAILLE           46   176.3043     6.9694       8110   160.0000   196.0000
POIDS            46    68.5217    10.0504       3152    53.0000   110.0000

               Pearson Correlation Coefficients  / N = 106

                                TAILLE            POIDS
                TAILLE          1.00000           0.63779
                POIDS           0.63779           1.00000
```

2. Test de nullité

```
PROC CORR DATA=moi.stid193 ;
  VAR taille poids;
```

[72] Ce nombre, calculé pour des variables quantitatives en général, est compris entre -1 et 1, caractérise l'intensité de la liaison linéaire entre deux variables. Plus il est proche de 1 (ou proche de -1) plus la liaison linéaire est forte.

[73] Le module SAS/INSIGHT vous permet de faire ce calcul. Cf. ce document paragraphe SAS/INSIGHT, option Multivariate (Y's).

```
RUN;
```

SAS effectue en plus un test sur le coefficient de corrélation.

Soit ρ les coefficient de corrélation entre deux variables X et Y sur une population P.

Test: H0: $\rho = 0$ (il n'y a pas de correlation) contre H1: $\rho \neq 0$.

Hypothèses:

> *' D'une part, on peut considérer la première variable dite dépendante, exprimée en fonction d'une autre variable dite indépendante. Dans ce cas, on suppose que la variable dépendante est* **normale et de variance constante et que la régression est linéaire**.
> *D'autre part, on peut considérer deux variables interdépendantes, dont on suppose que la distribution commune est* **une distribution normale à deux dimensions**. *C'est en général le deuxième cas qui est retenu.'* Dagnélie TMS2

La statistique de test est T= $\frac{R\sqrt{n-2}}{\sqrt{1-R^2}}$ qui sous H0 suit une loi de Student à n-2 degrés de libertés.
(n désigne l'effectif de l'échantillon et R la variable aléatoire associée au coefficient de corrélation)

En pratique

On calcule $t_{obs} = \frac{|r_{xy}|\sqrt{n-2}}{\sqrt{1-r_{xy}^2}}$ et on rejette H0 lorsque cette valeur est trop élevée, c'est à dire supérieure à $t_{1-\alpha/2}$ pour un niveau de signification α avec n-2 degrés de libertés.

Dans la pratique SAS donne le P correspondant. On rejette H0 lorsque P est supérieur à α=0.05 (en général)

Correlation Analysis

Pearson Correlation Coefficients / Prob > |R| under Ho: Rho=0 / N = 106

	TAILLE	POIDS
TAILLE	1.00000	0.63779
	0.0	**0.0001**
POIDS	0.63779	1.00000
	0.0001	**0.0**

Ici P=0.0001, on accepte donc H1: $\rho \neq 0$. On rejette l'hypothèse de nullité du coefficient de corrélation entre la taille et le poids de la population des STID de France. [74]

Exercice et exemple

[74] En considérant les STID grenoblois comme un échantillon aléatoire représentatif de la population des STID de france et en supposant les hypothèses citées plus haut remplies.

```
PROC SORT DATA=moi.stid193 out=stidtri;
BY sexe;
RUN;

PROC CORR data=stidtri;
  VAR taille poids;
  BY sexe;
RUN;
```

(Pour utiliser le BY dans PROC CORR)

Que fait le programme précédent. Quels sont les tests effectués ? Que concluent-ils ?

Utilisation de WITH

```
PROC CORR;
  VAR a b;
  WITH x y z;
RUN;
```

Calculera les corrélations X A, Y A, Z A, X B, Y B, Z B.

Quelques options de PROC CORR

Fichier de données

DATA=*nom du fichier de données SAS*
Fichier de données SAS sur lequel s'effectuera le calcul.

Sauvegarde des résultats

OUTP=*nom de fichier SAS*
Créée un fichier SAS contenant les coefficients de correlation de Pearson.
OUTS=*nom de fichier SAS*
Créée un fichier SAS contenant les coefficients de correlation de Spearman.
OUTK=*nom de fichier SAS*
Créée un fichier SAS contenant les coefficients de correlation de Kendall.

Affichage des résultats

BEST=*nombre n*
Affiche les n correlations les plus élevées (en valeur absolue) pour chaque variable.
RANK
Affiche les coefficients de corrélation du plus grand au plus petit (en valeur absolue). Si cette option n'est pas choisie, ils s'affichent sous forme d'une matrice.

NOPRINT
Aucun affichage des résultats.
NOSIMPLE
Pas d'affichage des statistiques descriptives effectuées sur les variables.
NOPROB
Pas d'affichage de la signification P du test H0: $\rho = 0$ contre H1: $\rho \neq 0$.

Statistiques particulières

PEARSON
Affiche le coefficient de corrélation linéaire. (option valide par défaut)

SPEARMAN
Affiche les coefficients de Spearman.

$$\theta = \frac{\sum (R_i - \overline{R})(S_i - \overline{S})}{\sqrt{\sum (R_i - \overline{R})^2 \sum (S_i - \overline{S})^2}}$$

où Ri est le rang de la ième valeur observée de X
et Si le rang de la ième valeur observée de Y.
(cela revient à substituer chaque valeur par son rang).

Notes: en cas d'ex aequo c'est la moyenne qui est prise

Dans la pratique, on utilise aussi la formule $1-\frac{6\sum d_i^2}{n(n^2-1)}$ où d_i est la différence de rangs relative aux différents individus.

COV

Affiche les covariances.

Diviseur utilisé pour le calcul de la variance

VARDEF=

DF Degrés de libertés (n-1) C'est l'option par défaut.
N Nombre d'observations (n)
WEIGHT Somme des poids
WDF Somme des poids -1.

K. PRINCOMP, Analyse en Composantes Principales

1. Syntaxe simplifiée

```
PROC PRINCOMP options;
VAR variables;
BY variables;
FREQ variables;
RUN;
```

variables (actives) sur lesquelles s'effectue l'ACP

décompose la population en sous populations identifiées par les modalités de BY

Chaque observation est comptée n fois où n est la valeur de la variable pour l'observation correspondante

Certaines options de PRINCOMP

Fichier de données

DATA=*nom du fichier de données SAS*

Fichier de données SAS sur lequel s'effectuera le calcul.

Sauvegarde des résultats

OUT=*nom de fichier SAS*

Créée un fichier SAS contenant toutes les données originales ainsi que les coordonnées des projections des individus sur les différentes composantes principales.

OUTSTAT=*nom de fichier SAS*

Créée un fichier SAS contenant les moyennes, écart-types, valeurs propres, vecteurs propres.

Paramétrage de l'analyse

N=*nombre entier*

Spécifie le nombre de composantes principales à calculer.

COVARIANCE ou COV

Effectue le calcul des composantes principales à partir de la matrice de variance covariance et non plus à partir de la matrice des corrélations. "*Cette option ne doit pas être utilisée à moins que les unités dans lesquelles sont exprimées les variables soient comparables ou à moins que les variables soient centrées réduites* ". SAS/STAT User Guide p.1244.

NOPRINT

Pas d'affichage des résultats

NOINT

La matrice des corrélations (ou de covariance) ne sera pas corrigée par rapport à la moyenne

Diviseur utilisé pour le calcul de la variance

VARDEF=

DF Degrés de libertés (n-1) C'est l'option par défaut. (N si NOINT est spécifiée)
N Nombre d'observations (n)
WEIGHT Somme des poids
WDF Somme des poids -1.(Somme des poids si NOINT est spécifiée)

Divers

PREFIX=*nom*

spécifie un prefixe pour nommer les composantes principales. Par défaut elles sont notées PRIN1, PRIN2 etc... Si PREFIX = Y est spécifié, les composantes principales seront notées Y1, Y2,... etc

2. Exercice

Prenez le fichier SAS ACP qui se trouve dans les répertoires habituels. Il donne les moyennes mensuelles des températures de 15 villes de France calculées sur 30 ans de 1930 à 1961. Sont indiquées également dans le fichier la latitude, la longitude, la moyenne annuelle et l'amplitude thermique de ces 15 villes.

Tapez le programme suivant:

```
PROC PRINCOMP DATA=PUB.ACP ;          (déclarez Pub correctement)
VAR JAN FEV MAR AVR MAI JUN JUI AOU SEP OCT NOV DEC;
RUN;
```

Et vérifiez que vous obtenez bien entre autres:

Principal Component Analysis

Eigenvalues of the Correlation Matrix

	Eigenvalue	Difference	Proportion	Cumulative
PRIN1	9.58178	7.30536	0.798482	0.79848
PRIN2	2.27642	2.20640	0.189702	0.98818
PRIN3	0.07001	0.03034	0.005835	0.99402
PRIN4	0.03967	0.02563	0.003306	0.99732
PRIN5	0.01405	0.00606	0.001170	0.99849
PRIN6	0.00798	0.00193	0.000665	0.99916
PRIN7	0.00605	0.00430	0.000504	0.99966
PRIN8	0.00175	0.00025	0.000146	0.99981
PRIN9	0.00149	0.00100	0.000124	0.99993
PRIN10	0.00049	0.00021	0.000041	0.99997
PRIN11	0.00029	0.00027	0.000024	1.00000
PRIN12	0.00002	.	0.000002	1.00000

Que contient ce tableau ? Quels sont les éléments donnés par SAS dans l'OUTPUT ?

Quelle est l'inertie expliquée par le premier axe ? le deuxième ?
Que pensez-vous de la qualité globale du premier plan principal ?
Combien de composantes allez-vous retenir ?

a) Calcul des coordonnées des individus sur les axes principaux

Modifiez le programme précédent comme suit:

```
PROC PRINCOMP DATA=MOI.ACP OUT=WORK.ESSAI;
 VAR JAN FEV MAR AVR MAI JUN JUI AOU SEP OCT NOV DEC;
RUN;
```

Que fait de plus ce programme par rapport au précédent ? Quels sont les données supplémentaires figurant dans le fichier SAS work.essai?

Exécutez ensuite:

```
PROC PRINT DATA=WORK.ESSAI;
 VAR NOM PRIN1 PRIN2 ;
RUN;
```

NOTE: la variable nom contient le nom de chaque ville

Vous devez obtenir:

OBS	NOM	PRIN1	PRIN2
1	bordeaux	3.01489	0.10559
2	brest	-2.19110	3.95451
3	clermont	-1.66741	-0.57244
4	grenoble	-1.47740	-1.63071
5	lille	-4.07384	0.57502
6	lyon	-0.80663	-1.72759
7	marseille	4.66885	-0.80070
8	montpellier	4.00667	-0.42059
9	nantes	-0.27175	1.07677
10	nice	5.80335	0.76254
11	paris	-1.19983	-0.15104
12	rennes	-1.38987	1.61446
13	strasbourg	-3.96639	-2.09860
14	toulouse	1.67730	-0.13151
15	vichy	-2.12684	-0.55571

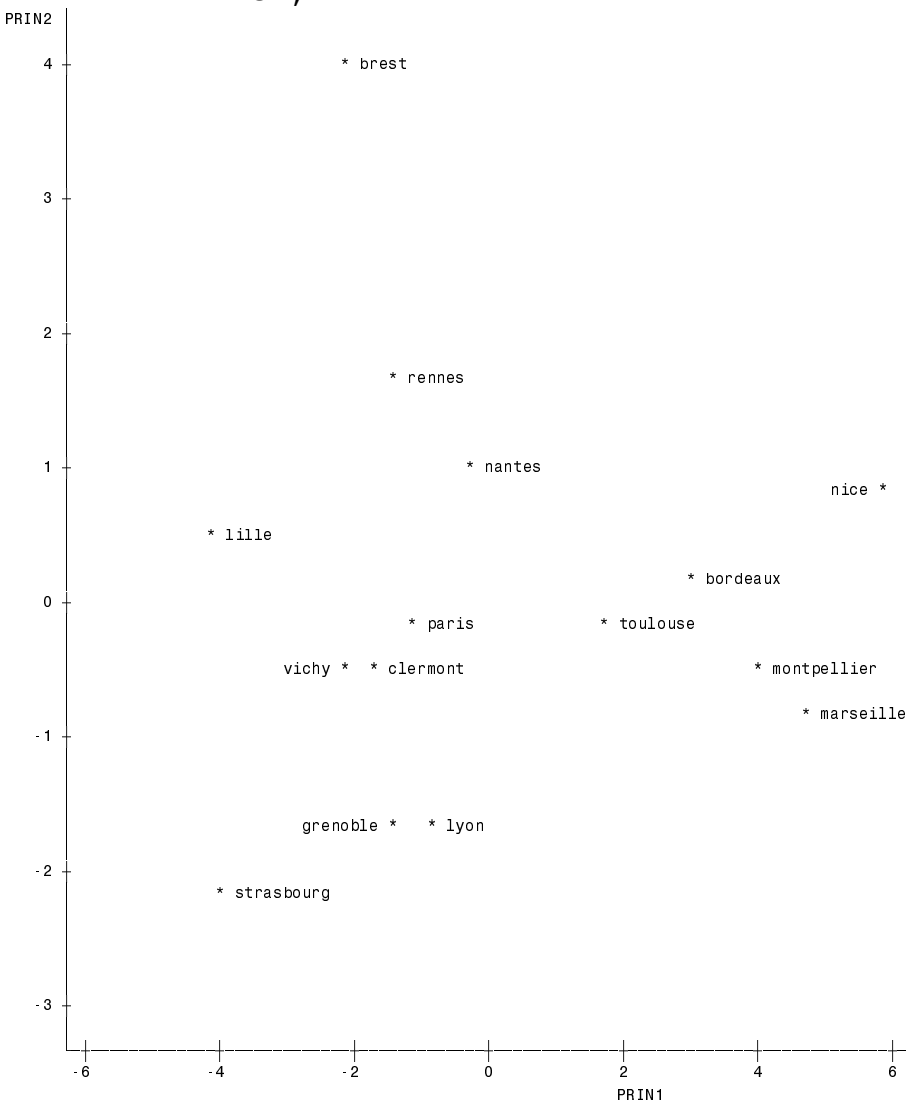
Donnez la signification des nombres en face de chaque ville .[75]

[75] SAS a la fâcheuse manie de donner le nom de PRIN1 au premier axe principal et à la première composante principale. Attention aux confusions.

b) Représentation du premier plan principal

Vous pouvez obtenir une représentation graphique basse résolution avec les instructions qui suivent :

```
PROC PLOT DATA=WORK.ESSAI;
 PLOT PRIN2*PRIN1='*' $ NOM;
RUN;
```



Il est possible d'obtenir une représentation en haute résolution[76] en utilisant GPLOT.

Il va falloir créer un fichier `Work.annoter` qui nous permettra d'afficher les noms des villes sur le graphique :

[76] On peut se demander pourquoi présenter à la fois un graphique en haute et en basse résolution. Les graphiques en basse résolution, s'ils sont moins beaux, ont le mérite d'être disponible sur toutes les plateformes de SAS ce qui n'est pas le cas pour les graphiques en haute résolution. De plus, ces derniers occupent beaucoup d'espace mémoire. Signalons enfin que le copier-coller de SAS vers Word des graphiques en haute résolution ne fonctionne pas (pour l'impression). Il faut passer par un logiciel de dessin (Paint ou Paintbrush) et sous Word choisir Insert/Image.

```
DATA WORK.ANNOTER;
 SET WORK.ESSAI;
 X=PRIN1;
 Y=PRIN2;
 TEXT=NOM;
 SIZE=1;
 XSYS='2';
 YSYS='2';
 LABEL Y='AXE 2'
       X='AXE 1';
 KEEP X Y XSYS YSYS TEXT SIZE;
RUN;

TITLE 'PREMIER PLAN PRINCIPAL';
PROC GPLOT DATA=WORK.ANNOTER;
 PLOT Y*X=1 / ANNOTATE=WORK.ANNOTER HREF=0 VREF=0;
RUN;
QUIT;
```

On part du fichier ESSAI

On appelle x et y Prin1 et Prin2.

Le texte à afficher est le nom de la ville

Taille du texte.

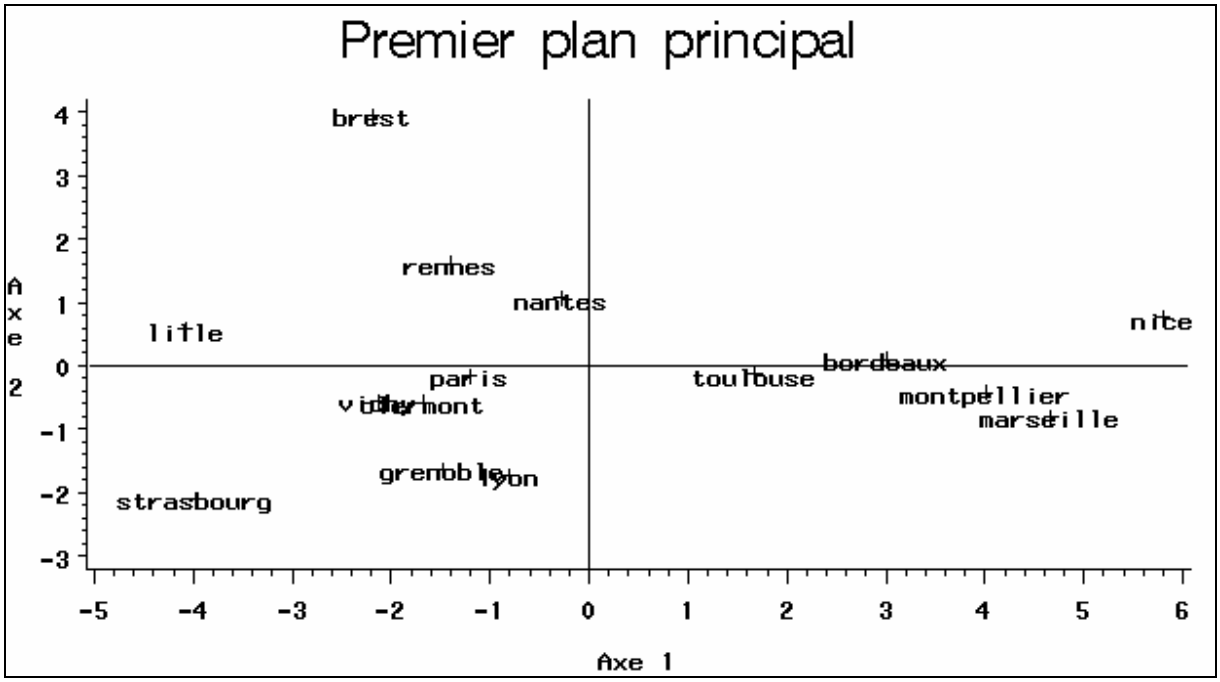
Pour mettre les noms des villes (Text) en (x,y)

Etiquette des axes

On ne conserve que les variables utiles

TITRE DU GRAPHIQUE.

Pour tracer les axes passant par O

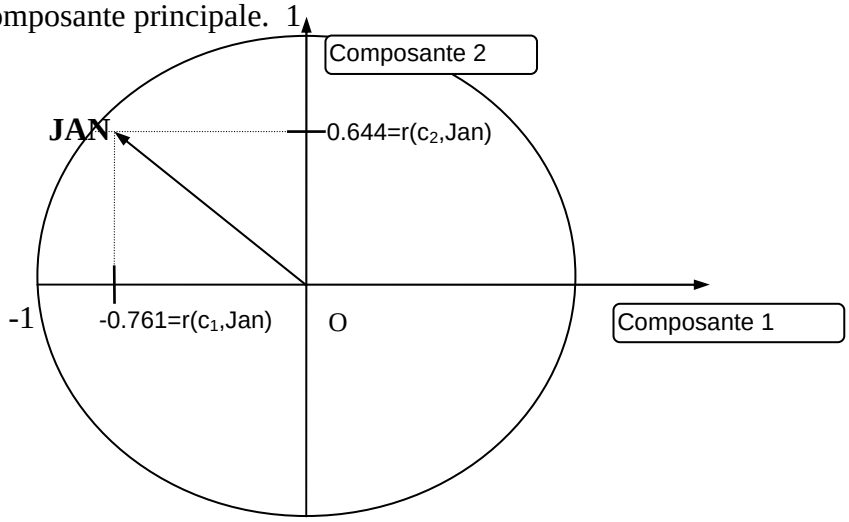


Si le graphique précédent ne vous plait pas, vous pouvez également transférer les données sous Excel (File/Export ou liaison DDE) et effectuer le graphique sous Excel.

c) Interprétation des composantes principales - Cercle des corrélations

Pour interpréter les composantes principales, il peut être intéressant de calculer le coefficient de corrélation entre chaque composante et les variables du fichier. Nous distinguerons deux types de variables, les variables utilisées dans les calculs et les autres appelées *variables supplémentaires*.

Nous allons ensuite représenter graphiquement chaque variable par un point dont les coordonnées sont les coefficients de corrélation avec la première puis la deuxième composante principale.



Ainsi Cor(Composante1,Jan)= -0.761 et Cor(Composante2,Jan)=0.644 donc la variable « Janvier » aura pour coordonnées (-0.761,0.644).

Remarque fondamentale :

On peut montrer aisément que, dans le cas de l'ACP normée, le cercle des corrélations précédent n'est pas qu'une simple représentation graphique mais également la **projection de l'ensemble des variables centrées réduites dans le plan engendré par les deux composantes principales.**
Une variable sera bien représentée par cette projection lorsque son point variable sera proche de la circonférence.[77]

Nous allons calculer les corrélations (PROC CORR) entre les composantes principales (PRIN1 et PRIN2) et les variables statistiques dont nous disposons.

[77] Soit L_i la longueur de Vi le ieme vecteur variable projeté. Alors $L_i=r^2(C_1,Vi)+ r^2(C_2,Vi)$ or C_1 et C_2 ne sont pas corrélées donc $L_i=r^2(V_i ; C_1, C_2)$ le coefficient de détermination entre Vi et (C_1, C_2). Lorsque V_i est combinaison linéaire de (C_1, C_2) alors $L_i=r^2(V_i ; C_1, C_2)=1$ donc le point variable est sur le cercle et réciproquement. Dans la pratique il faut de méfier de l'interprétation de la proximité entre des points variables si ceux-ci ne sont pas proches de la circonférence.

Vous pourrez utiliser le programme suivant, en le complétant, pour calculer de nouvelles variables utiles pour l'interprétation.

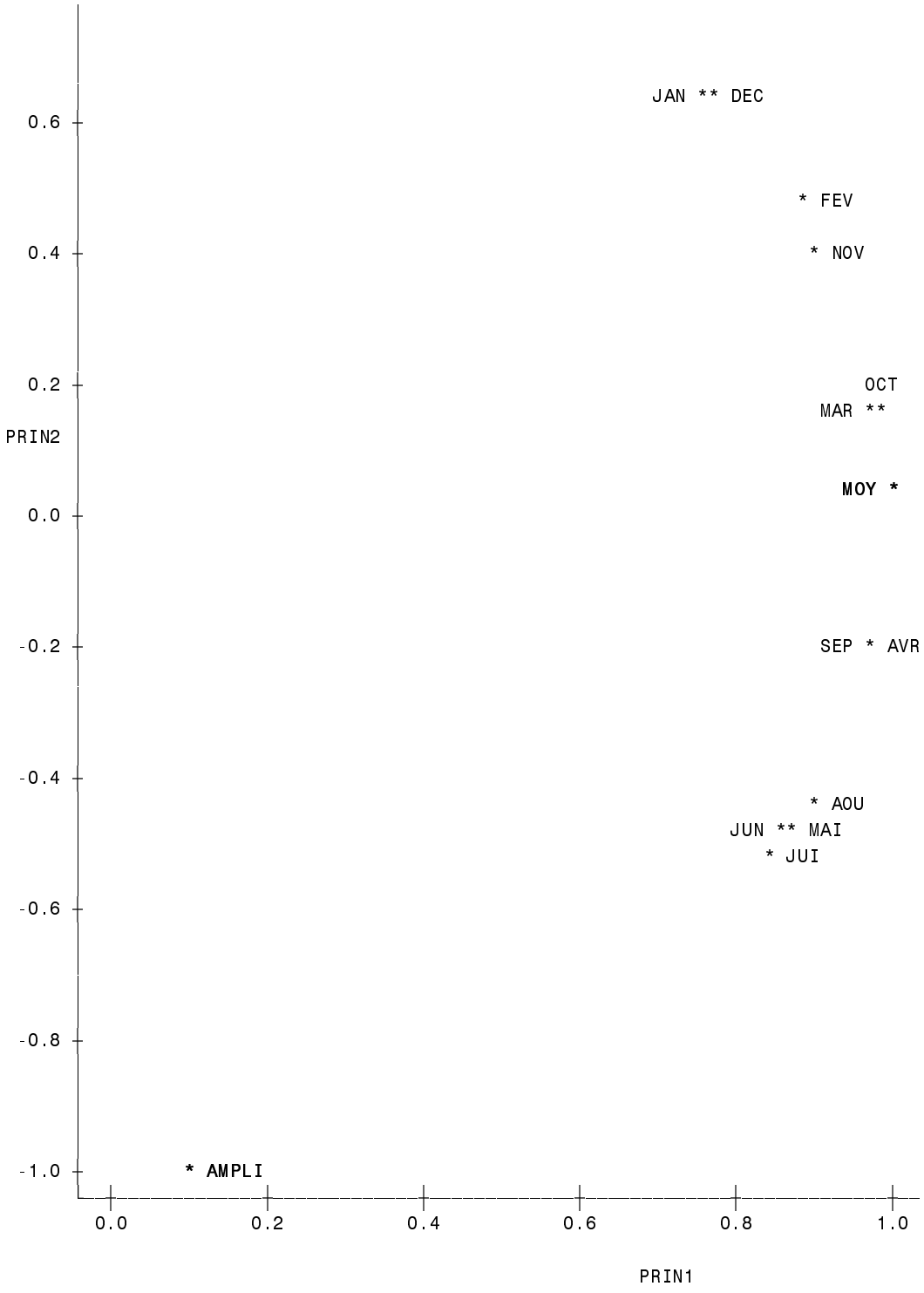
```
PROC PRINCOMP DATA=PUB.ACP OUT=ESSAI NOPRINT;
 VAR JAN FEV MAR AVR MAI JUN JUI AOU SEP OCT NOV DEC;
RUN;

DATA FINAL;
 SET ESSAI;
 MOY=MEAN(JAN,FEV,MAR,...,DEC);
 AMPLI=MAX(JAN,FEV,MAR,...,DEC)-MIN(JAN,FEV,MAR,...,DEC);
RUN;
PROC CORR DATA=FINAL OUTP=ESSAI2 ;
 VAR PRIN1 PRIN2;
 WITH JAN FEV MAR AVR MAI JUN JUI AOU SEP OCT NOV DEC MOY AMPLI;
RUN;
PROC PRINT DATA=ESSAI2 ; RUN ;
/* la procédure précédente affiche les coefficients de corr.*/
/* entre les composantes principales et les variables de
températures ainsi que la p valeur du test de nullité */

/* Création d'un fichier de données en vue de dessiner un cercle*/
/* des corrélations*/

DATA MOI.CERCLE;
 SET ESSAI2 (WHERE=(_TYPE_='CORR'));
 DROP _TYPE_;
RUN;
PROC PLOT DATA=MOI.CERCLE;
 PLOT PRIN2*PRIN1= '*' $ _NAME_;
RUN;
```

Le fichier ESSAI2 contient entre autres les coefficients de corrélation entre les composantes principales et les variables du fichier :

	PRIN1	PRIN2
JAN	0.76124	0.64434
FEV	0.88046	0.46908
MAR	0.96877	0.15601
AVR	0.96934	-0.20367
MAI	0.87276	-0.47471
JUN	0.86357	-0.49935
JUI	0.84153	-0.53142
AOU	0.89861	-0.42994
SEP	0.97403	-0.20810
OCT	0.98016	0.17046
NOV	0.90375	0.41393
DEC	0.77433	0.62430
MOY	**0.99972**	**0.02135**
AMPLI	**0.10106**	**-0.98568**

Variables supplémentaires

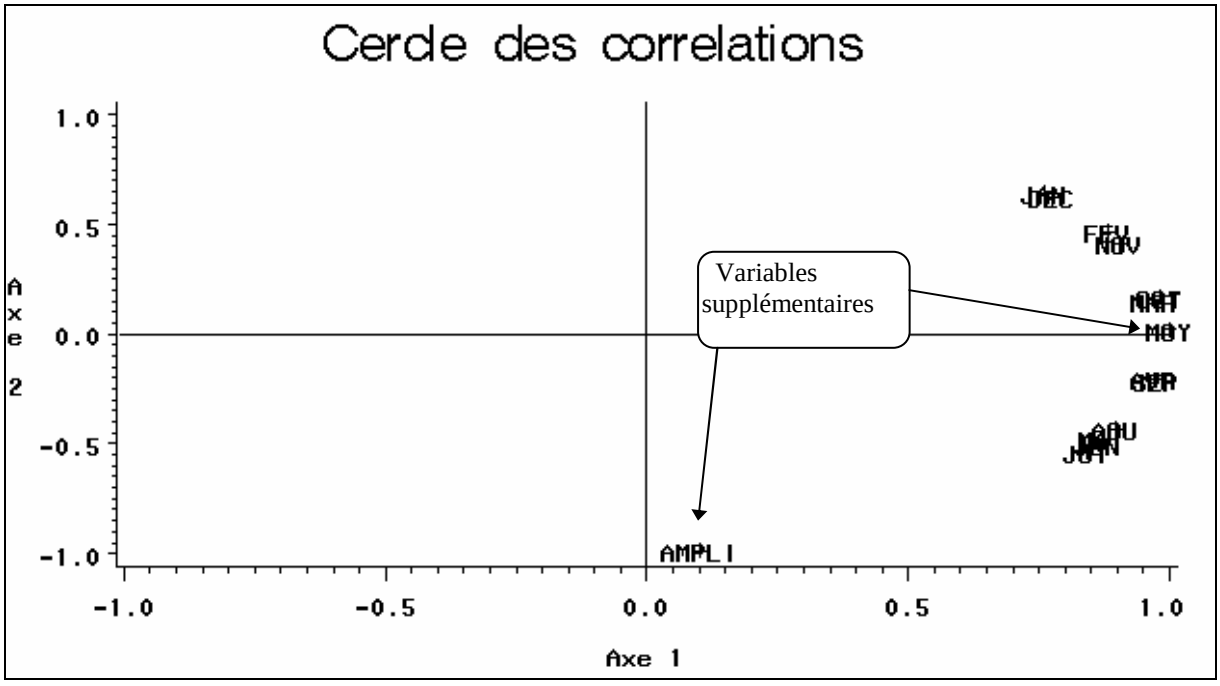
Nous pouvons représenter cela graphiquement ce qui donne :


JAN ** DEC
* FEV
* NOV
OCT
MAR **
MOY *
SEP * AVR
* AOU
JUN ** MAI
* JUI
* AMPLI
0.6
0.4
0.2
0.0
-0.2
-0.4
-0.6
-0.8
-1.0
PRIN2
0.0
0.2
0.4
0.6
0.8
1.0
PRIN1

et pour la haute résolution[78], nous devons modifier le fichier Moi.cercle.

```
…
PROC CORR DATA=FINAL OUTP=ESSAI2 NOPRINT;
 VAR PRIN1 PRIN2;
 WITH JAN FEV MAR AVR MAI JUN JUI AOU SEP OCT NOV DEC MOY AMPLI;
RUN;

DATA MOI.CERCLE;
 SET ESSAI2 (WHERE=(_TYPE_='CORR'));
 X=PRIN1;
 Y=PRIN2;
 TEXT=_NAME_;
 XSYS='2';
 YSYS='2';
 LABEL X='AXE 1'
       Y='AXE 2';
 KEEP X Y XSYS YSYS TEXT;
 RUN;
TITLE "CERCLE DES CORRELATIONS";
PROC GPLOT DATA=MOI.CERCLE;
 PLOT Y*X=1 / ANNOTATE=MOI.CERCLE HREF=0 VREF=0
              HAXIS=-1 TO 1 BY .5 VAXIS=-1 TO 1 BY .5;
RUN;
QUIT;
```

Nous spécifions des axes horizontaux et verticaux sur [-1 ;1]



Les variables supplémentaires choisies nous permettent d'interpréter facilement les axes principaux.

[78] Pour transférer ce graphique en haute résolution de SAS vers Word, il faut passer par Paint (Edition/Copier vers puis sous Word Insere/Image) sinon, l'impression nous réserve quelques surprises...

d) Calcul des contributions des individus

Il est nécessaire de s'assurer que la direction des axes principaux n'a pas été fixée par un petit nombre d'individus atypiques. Pour cela, on calcule les contributions des individus pour chaque axe principal.

Calcul

On a vu que l'inertie du nuage projeté sur un axe est égal à la valeur propre associée à cet axe. Cette inertie est aussi égale à la moyenne des carrés des coordonnées des individus sur cet axe. **(voir la remarque ci-dessous pour certains logiciels)**

Ainsi chaque individu contribue à l'inertie de l'axe en question. Plus il est éloigné de l'origine plus il contribue, et plus il est proche de O moins il contribue.

On a donc la contribution de l'individu j à l'axe i :

Contribution de l'individu j à l'*axe i* $= \frac{x_{j,i}^2}{(n-1)\alpha_i}$ où α_i est la valeur propre associée à l'axe *i*, $x_{j,i}$ la coordonnée de l'individu j sur l'axe i et *n* le nombre d'individus.
La somme de ces contributions étant égale à 1 pour chaque axe.

(**Rq :**On peut multiplier ce nombre par 100 pour obtenir des %)

Remarque : Pour certains logiciels (comme SAS et Minitab) on a $\frac{\sum_j x_{j,i}^2}{n-1} = \alpha_i$, il se peut donc que les coordonnées sur les axes fournies par d'autres logiciels (Statlab) ne correspondent pas avec celles fournies par SAS et Minitab. Par contre les qualités, les contributions sont rigoureusement identiques !

Ce calcul permet de repérer les individus contribuant fortement à des axes... remettant ainsi fortement an cause leur interprétation.

Mise en pratique

Effectuez le calcul sous SAS en ajoutant 2 lignes au programme précédent (contri1=... contri2=...) et vérifiez que l'on obtient bien pour les deux premières villes :

	contribution à l'axe 1 (en %)	**contribution à l'axe 2 (en %)**
Bordeaux	**6.7759**	**0.0350**
Brest	**3.5789**	**49.0692**

La ville de Bordeaux contribue à 6.8% de l'inertie du premier axe.
Ceci montre aussi que la ville de Brest contribue à 49% de l'inertie du deuxième axe principal ce qui est beaucoup.

Il serait intéressant de placer cette ville en « individu supplémentaire » et de regarder si nous conservons l'interprétation précédente.

e) Individu supplémentaire

Comme nous l'avons dit, nous allons mettre Brest en individu supplémentaire. Brest ne sera donc plus utilisé dans les calculs, mais nous l'afficherons dans le premier plan principal.

Pour cela, il suffit de lui affecter un poids négatif.

```
DATA WORK.ACPBREST;
 SET MOI.ACP;
 W=1;
 IF NOM='BREST' THEN W=-1;
RUN;

PROC PRINCOMP DATA=WORK.ACPBREST OUT=WORK.ESSAI;
 WEIGHT W;
 VAR JAN FEV MAR AVR MAI JUN JUI AOU SEP OCT NOV DEC;
RUN;


 DATA WORK.ANNOTER;
 SET WORK.ESSAI;
 X=PRIN1;
 Y=PRIN2;
 TEXT=NOM;
 SIZE=1;
 XSYS='2';
 YSYS='2';
 COLOR='BLUE';
 IF W=-1 THEN COLOR='RED';
 LABEL Y='AXE 2'
       X='AXE 1';
 KEEP X Y XSYS YSYS TEXT SIZE COLOR;
RUN;

TITLE 'PREMIER PLAN PRINCIPAL';
PROC GPLOT DATA=WORK.ANNOTER;
 PLOT Y*X=1 / ANNOTATE=WORK.ANNOTER HREF=0 VREF=0;
RUN;
QUIT;
```

L'ACP s'effectuera en tenant compte du poids de chaque individu.

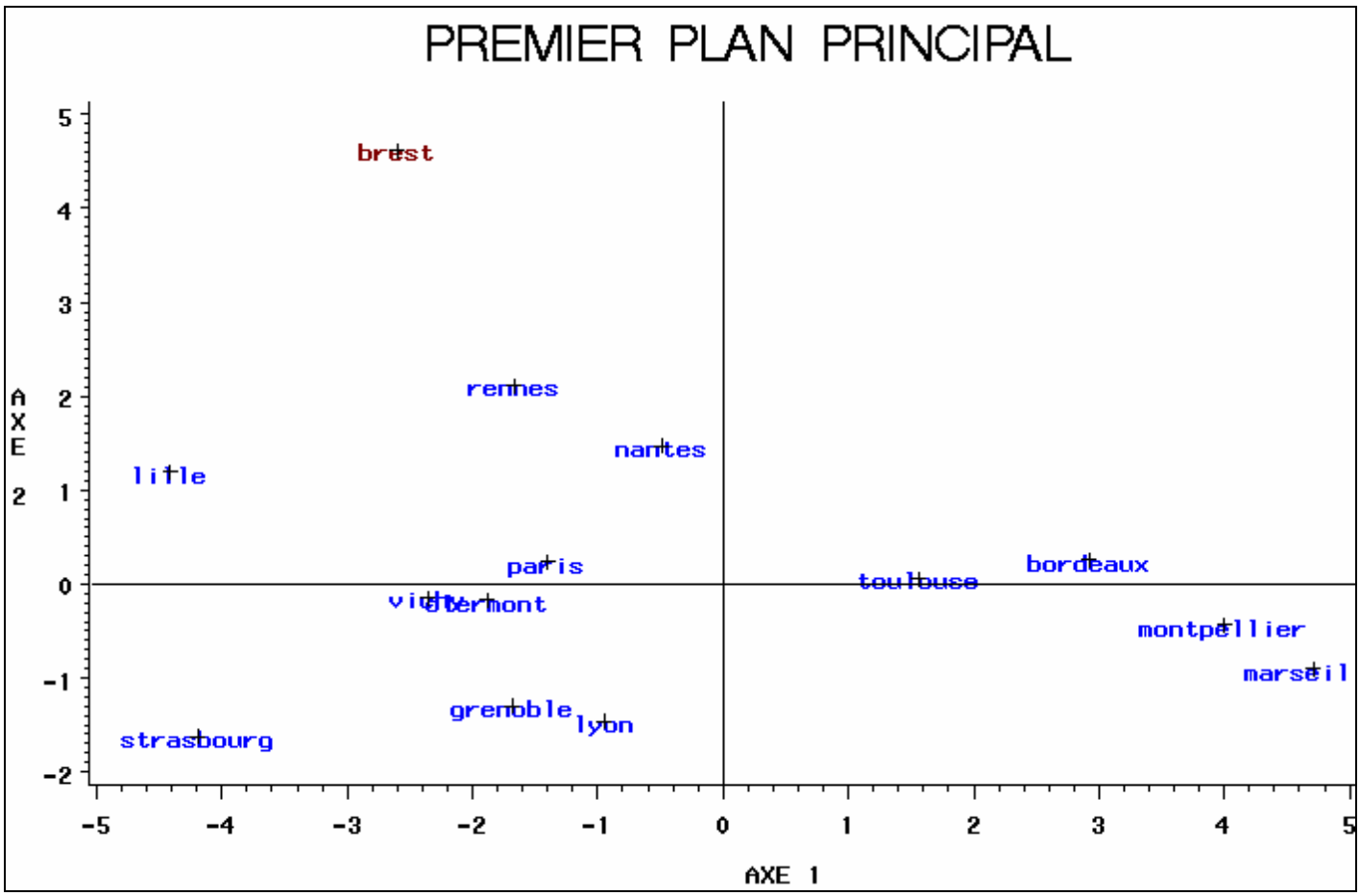
Nous affectons Brest d'un poids négatif

L'individu supplémentaire sera affiché en rouge (les autres en bleu)

Nous obtenons :

	Eigenvalue	Difference	Proportion	Cumulative
PRIN1	10.6035	9.35905	0.883625	0.88362
PRIN2	1.2444	1.16837	0.103704	0.98733
PRIN3	0.0761	0.03066	0.006340	0.99367
PRIN4	0.0454	0.02987	0.003785	0.99745

La représentation graphique est la suivante :

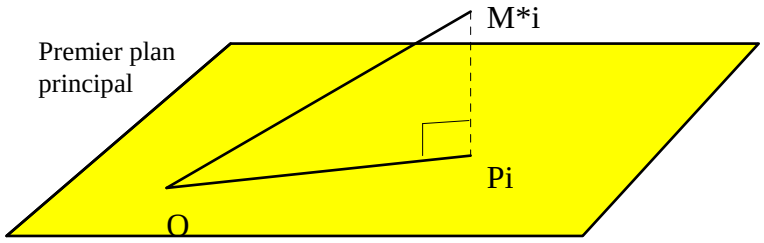


Les interprétations ne sont pas modifiées. Nous allons maintenant calculer les qualités de représentation des individus.

f) Qualité de représentation des individus sur le premier plan principal

Nous avons déjà vu que la qualité globale était excellente, nous allons maintenant regarder pour chaque individu ce qu'il en est. Un point sera bien représenté lorsqu'il sera "proche" du plan en question. Cette qualité se mesure à l'aide de $\frac{\left\|\vec{OP_i}\right\|^2}{\left\|\vec{OM_i^*}\right\|^2} = \frac{Y_1(i)^2 + Y_2(i)^2}{\left\|\vec{OM_i^*}\right\|^2}$. où M*i est le ième individu (centré réduit) et Pi sa projection sur le premier plan principal. (Pi a pour coordonnées (Y1(i), Y2(i)) dans le premier plan principal)

Que représente ce nombre ? (Vous pourrez vous aider du graphique qui suit)



Plus ce nombre est proche de 1, et plus M*i est proche de Pi, et meilleure est la qualité de représentation de l'individu en question.

Pour calculer ce nombre, il faut centrer et réduire les variables[79], c'est ce que fait la procédure suivante: (PROC STANDARD)

[79] En effet, comme nous avons pris la matrice des corrélations, cela revient à dire que nous avons centrés réduits nos variables de départ avant d'effectuer les calculs. Le plan principal retenu est donc le plan donnant une image déformant le moins le nuage des individus (centrés réduits) initial. Pour obtenir plus d'informations sur le choix entre la matrice de corrélation et la matrice de covariance, nous vous renvoyons à la bibliographie: (Escoffier et Pages notamment.)

L. **STANDARD** , normalisation de variables

(pour obtenir entre autres, des variables centrées réduites...)

Syntaxe simplifiée

`PROC STANDARD` *options*`;`	DATA, OUT, STD et MEAN sont les options les plus fondamentales
`VAR` *variables*`;`	Variables à transformer
`BY` *variables*`;`	Sens habituel
`FREQ` *variable*	*Idem*
`RUN;`	

Quelques options de la procédure Standard

Fichier de données

DATA=*nom du fichier de données SAS*

Fichier de données SAS sur lequel s'effectuera le calcul.

Sauvegarde des résultats

OUT=*nom de fichier SAS*

Créée un fichier SAS contenant toutes les données originales ainsi que les variables centrées réduites. **Si cette option n'est pas spécifiée, SAS va créer de lui-même un fichier de données résultat**.

Paramétrisation

MEAN= moyenne des nouvelles variables. (par défaut la moyenne de la variable initiale)
STD= ecart-type des nouvelles variables; (par défaut l'écart-type de la variable initiale)

NOTE: Si MEAN et STD ne sont pas précisées, les nouvelles variables seront rigoureusement identiques aux précédentes.

REPLACE les valeur manquantes seront remplacée par la valeur MEAN

Affichage des résultats

PRINT

Affiche les moyennes, écart-types pour les variables transformées.

Diviseur utilisé pour le calcul de la variance

VARDEF=

DF Degrés de libertés (n-1) C'est l'option par défaut.
N Nombre d'observations (n)
WEIGHT Somme des poids
WDF Somme des poids -1.

Exemple

```
PROC STANDARD data=moi.essai out=moi.essai mean=0 std=1;
VAR JAN FEV;
RUN;
```

Ce programme va centrer réduire les variables jan et fev du fichier moi.essai.

Note: En choisissant OUT=MOI.ESSAI, j'impose à SAS de remplacer les variables initiales de ce fichier par les variables centrées réduites. [80]

Application, qualité de représentation des individus sur le premier plan principal

Nous allons centrer réduire les variables Janvier à décembre puis nous allons créer un nouveau fichier contenant la somme des carrés de chaque ligne. Nous aurons ainsi le dénominateur de la quantité à calculer.

Comprenez ce que fait le programme suivant:

```
PROC PRINCOMP DATA=PUB.ACP OUT=WORK.ESSAI;
VAR JAN FEV MAR AVR MAI JUN JUI AOU SEP OCT NOV DEC;
RUN;
PROC STANDARD DATA=WORK.ESSAI OUT=WORK.ESSAI MEAN=0 STD=1;
VAR JAN FEV MAR AVR MAI JUN JUI AOU SEP OCT NOV DEC;
RUN;
DATA WORK.ESSAI2;
 SET WORK.ESSAI;
 DENOM= USS(JAN,FEV,MAR,AVR,MAI,JUN,JUI,AOU,SEP,OCT,NOV,DEC);
 NUMER=USS(PRIN1,PRIN2);
 QUAL=NUMER/DENOM;
 KEEP NUMER DENOM QUAL NOM;
RUN;
```

Complétez votre programme pour obtenir la qualité de représentation des individus dans le premier plan principal.

[80] Si je ne souhaite pas perdre les données initiales, je change de nom du fichier, si je ne mets rien, SAS va créer un nouveau fichier pour y stocker les résultats.

Vous vérifierez avec le tableau ci-dessous.

OBS	NOM	DENOM	NUMER	QUAL
1	bordeaux	9.6014	9.1007	0.94785
2	brest	20.4850	20.4391	0.99776
3	clermont	3.1598	3.1080	0.98359
4	grenoble	5.0886	4.8419	0.95152
5	lille	17.0827	16.9268	0.99088
6	lyon	3.6525	3.6352	0.99526
7	marseille	22.6076	22.4393	0.99255
8	montpellier	16.2854	16.2303	0.99662
9	nantes	1.3081	1.2333	0.94278
10	nice	34.3644	34.2603	0.99697
11	paris	1.6187	1.4624	0.90345
12	rennes	4.6010	4.5382	0.98636
13	strasbourg	20.2825	20.1363	0.99279
14	toulouse	2.9534	2.8306	0.95841
15	vichy	4.9088	4.8323	0.98442

M. CLUSTER : Classification d'individus

C'est une méthode complémentaire à l'ACP c'est pourquoi nous allons la présenter ici.

Dans le paragraphe précédent (Proc Princomp), vous avez établi une typologie des villes à partir des données de températures. Nous allons essayer de retrouver cette typologie en utilisant l'algorithme de classification ascendante hiérarchique.

1. But

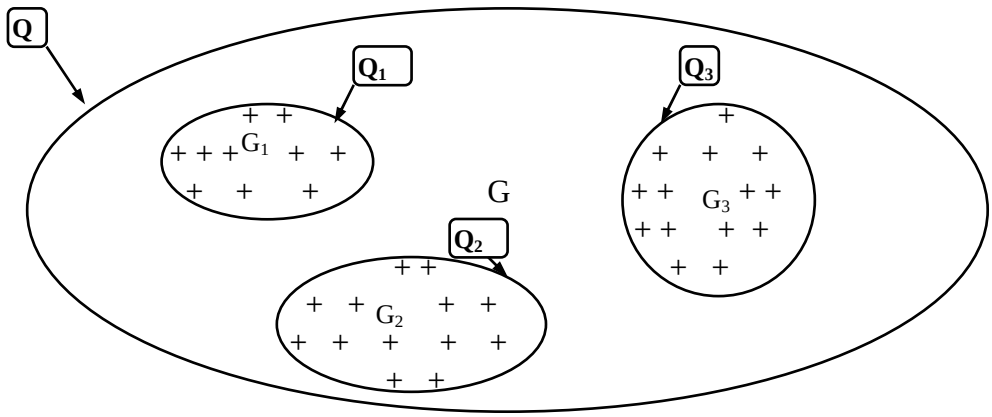
Le but de la méthode est de construire une partition de l'ensemble des individus de telle sorte que les individus d'une même classe soient « proches » et ceux issus classes distinctes soient « éloignés ».

2. Choix de la distance

Pour savoir si des individus sont proches ou éloignés, il faut mesurer la distance qui les sépare. Nous devons donc choisir une distance. Pour nos données quantitatives, nous choisirons la distance euclidienne qui est celle que vous connaissez bien.[81]

3. Qualité de la typologie

Considérons notre population initiale Q décomposée en p classes $Q_1,Q_2,...,Q_p$. Nous supposons que la sous-population Q_i d'effectif n_i a pour barycentre le point G_i. Enfin, notons G le barycentre du nuage complet.



Pour mesurer la qualité d'un découpage en classe, nous allons utiliser une décomposition de l'inertie totale d'un nuage de point qui ressemble beaucoup à la décomposition de la variance :

[81] Pour des données qualitatives ayant plus de deux modalités, la distance du chi deux semble plus indiquée. Cf. Saporta P257

$$I_{tot} = \underbrace{\frac{1}{n-1}\sum_{i=1}^{p} n_i d(G_i, G)^2}_{\text{Inertie inter-classes}} + \underbrace{\frac{1}{n-1}\sum_{i=1}^{p} n_i Inertie(Q_i)}_{\text{Inertie intra-classes}}$$

Lorsqu'il n'y a qu'une seule classe, l'inertie inter-classes est nulle, l'inertie totale est égale à l'inertie intra-classes. Inversement lorsque chaque classe ne contient qu'un individu, l'inertie intra-classes est nulle et l'inertie inter-classes est égale à l'inertie totale. L'inertie intra-classe mesure l'homogénéité des classes. Plus elle est faible, plus les individus sont proches les uns des autres dans les classes donc plus les classes sont homogènes.

On mesure la qualité d'une partition par le rapport $\frac{\text{Inertie inter - classes}}{\text{Inertie totale}}$ qui doit être le « plus élevé possible ».

4. Algorithme

Nous allons partir du découpage maximum (un individu par classe), l'inertie intra-classes est donc nulle, les classes ont une homogénéité parfaite !

Nous allons à chaque étape regrouper les classes les plus « proches » de façon à **augmenter le moins possible l'inertie intra-classes**.[82]

A la dernière étape, nous n'aurons qu'une seule classe. Il est évident que le « meilleur » découpage en classe se trouve entre ces deux étapes extrêmes. Dans la pratique, nous nous arrêtons lorsque nous voyons une augmentation « brutale » du critère.

Lorsque les classes Q_i et Q_j sont regroupées la diminution d'inertie inter-classes est égale à $d(Q_i, Q_j) = \frac{n_i n_j}{(n-1)(n_i + n_j)} d(G_i, G_j)^2$. A chaque étape de l'algorithme, nous allons chercher a regrouper les classes qui minimisent cette quantité c'est ce qu'on appelle **le critère de Ward**.

[82] Ou, ce qui revient au même, à diminuer le moins possible l'inertie inter-classes.

5. Mise en œuvre (Proc CLUSTER)

a) Syntaxe simplifiée

```
PROC CLUSTER METHOD=nom de méthode  options ;
  VAR variables ;
  BY variable ;
RUN ;
```

Choix de la méthode

SAS connaît 11 méthodes différentes pour effectuer la classification.

AVERAGE, CENTROID, COMPLETE, DENSITY, EML, FLEXIBLE, MCQUITTY, MEDIAN, SINGLE, TWOSTAGE, WARD.

Dans l'exemple qui nous intéresse, nous allons utiliser le critère de Ward.

Les principales options étant

DATA= *Nom de fichier de données SAS*
Pour spécifier un nom de fichier de données à traiter.

OUTTREE=*Nom de fichier de données SAS*
Pour indiquer un fichier de données où SAS mettra les résultats des calculs de la procédure CLUSTER. Ces résultats pourront être récupérés par la procédure PROC TREE qui permet d'effectuer un découpage en classes des données, de tracer un pseudo dendrogramme[83] etc.

STANDARD
Pour demander à SAS de travailler sur les données centrées réduites.

NOTIE
Pour demander à SAS de ne pas vérifier l'existence d' exaequos.[84]

RSQUARE
Pour afficher le R^2 qui est l'indice mesurant la qualité de la classification dont nous parlions plus haut (Inertie inter classes/Inertie totale). Cette option est automatiquement activée pour METHOD=WARD. SAS affiche également un « R^2 partiel » qui est en fait la perte de R^2 à chaque étape.[85]

NOPRINT
Supprime l'affichage.

---

[83] « Pseudo » car il faut vraiment beaucoup d'imagination pour le reconnaître !

[84] Par défaut, SAS vérifie s'il y a des exaequos dans les distances entre classes ce qui ralentit l'exécution de l'algorithme. Cette option supprime cette vérification.

[85] Ce R^2 partiel est très pratique pour détecter les sauts importants de l'indicateur et donc le nombre de classes à conserver en final.

SIMPLE

Affiche des statistiques de base sur les variables initiales.

b) Le programme

Nous allons donc taper le programme suivant :

```
proc cluster data=pub.acp method=ward standard;
 var jan fev mar avr mai jun jui aou sep oct nov dec;
run;
```

Les deux mots WARD et STANDARD sont fondamentaux ici bien entendu.

c) Résultats

SAS donne alors la diagonalisation de la matrice des corrélations.[86]

Ward's Minimum Variance Cluster Analysis

Eigenvalues of the Correlation Matrix

	Eigenvalue	Difference	Proportion	Cumulative
1	9.58178	7.30536	0.798482	0.79848
2	2.27642	2.20640	0.189702	0.98818
3	0.07001	0.03034	0.005835	0.99402
4	0.03967	0.02563	0.003306	0.99732
5	0.01405	0.00606	0.001170	0.99849
6	0.00798	0.00193	0.000665	0.99916
7	0.00605	0.00430	0.000504	0.99966
8	0.00175	0.00025	0.000146	0.99981
9	0.00149	0.00100	0.000124	0.99993
10	0.00049	0.00021	0.000041	0.99997
11	0.00029	0.00027	0.000024	1.00000
12	0.00002	.	0.000002	1.00000

The data have been standardized to mean 0 and variance 1

Etapes de l'algorithme

Nombre de classes

Classes réunies

Effectif de la nouvelle classe

Diminution de R^2 à chaque étape

Inertie inter-cl/ Inertie totate.

Number of Clusters	--Clusters Joined--		Frequency of New Cluster	Semipartial R-Squared	R-Squared
14	OB3	OB15	2	0.000812	0.999188
13	OB7	OB8	2	0.002052	0.997137
12	OB4	OB6	2	0.002141	0.994996
11	CL14	OB11	3	0.003951	0.991044
10	OB9	OB12	2	0.004746	0.986298
9	OB1	OB14	2	0.007447	0.978851
8	CL11	CL12	5	0.013823	0.965028
7	CL13	OB10	3	0.016475	0.948552
6	OB5	OB13	2	0.022101	0.926452
5	OB2	CL10	3	0.035062	0.891390
4	CL9	CL7	5	0.046949	0.844440
3	CL8	CL6	7	0.057419	0.787021
2	CL5	CL3	10	0.130226	0.656796
1	CL4	CL2	15	0.656796	0.000000

Nous partons avec un R^2 à 100%. L'inertie inter classe est en effet égale à l'inertie totale.

Lisons l'étape 1 de l'algorithme [87]: Les individus les plus proches sont le 3 et le 15. Les classes 3 et 15 sont donc réunies pour former une nouvelle classe : CL14 dans laquelle il y a deux individus. Il n'y a donc plus que 14 classes. Cette réunion diminue le R^2 de 0.000812 pour le placer à 0.999188.

A l'étape 2, ce sont les individus 7 et 8 qui sont réunis pour former CL13 etc.

Nous voyons clairement un saut brutal de R^2 à l'étape 13. Le « *semi partial* R^2 » passe de 0.05 à 0.13. **Ceci nous confirme dans une partition à 3 classes** (étape 12) que nous avions cru discerner lors de l'ACP.

[86] Nous avons déjà interprété ce type de sortie en ACP.

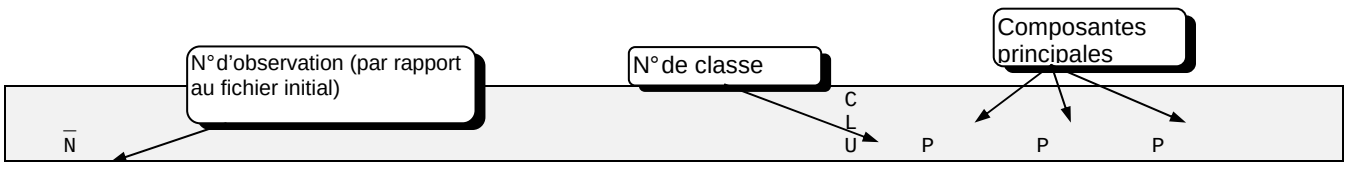
[87] Rappel : A l'étape 0, il y a 15 classes, un seul individu par classe.

d) Partition en classes utilisation de PROC TREE

Pour obtenir une variable contenant les affectations des individus dans les classes, il va falloir demander à SAS de créer un fichier de résultats dans PROC CLUSTER et le récupérer pour être traiter par la procédure TREE qui elle même donnera un fichier avec la classe de chaque individu ![88]

```
/*Nous effectuons le decoupage en classes par la methode de Ward*/
PROC CLUSTER DATA=PUB.ACP METHOD=WARD STANDARD OUTTREE=RESULT2;
 VAR JAN FEV MAR AVR MAI JUN JUI AOU SEP OCT NOV DEC ;
RUN;

/*On récupère les resultats précédents pour construire*/
/* le découpage en classes (nclusters=3 précise que */
/* nous voulons trois classes ici)*/
/* Le COPY effectue la copie des variables jan fev etc.*/
/* dans le fichier de résultats*/

PROC TREE DATA=RESULT2 OUT=RESULT3 NCLUSTERS=3 NOPRINT;
COPY JAN FEV MAR AVR MAI JUN JUI AOU SEP OCT NOV DEC;
RUN;

/*Nous recalculons les composantes principales pour pouvoir */
/*faire dessiner le premier plan principal en visualisant les 3*/
/*classes calculees ci-dessus*/

PROC PRINCOMP DATA=RESULT3 OUT=RESULT4;
 VAR JAN FEV MAR AVR MAI JUN JUI AOU SEP OCT NOV DEC;
RUN;

/*Nous creons le fichier temporaire annoter qui va nous permettre*/
/* de personnaliser notre graphique */
DATA WORK.ANNOTER;
 SET RESULT4;
 X=PRIN1;
 Y=PRIN2;
 TEXT=_NAME_;  /*CONTIENT LE N° D'OBSERVATION*/
 SIZE=1;
 XSYS='2';
 YSYS='2';
 IF CLUSTER=1 THEN COLOR='BLUE'; /*couleurs differentes selon  le groupe*/
 IF CLUSTER=2 THEN COLOR='RED';
 IF CLUSTER=3 THEN COLOR='CYAN';
 LABEL Y='AXE 2'
       X='AXE 1';
 KEEP X Y XSYS YSYS TEXT SIZE CLUSTER COLOR ;
RUN;

TITLE 'PREMIER PLAN PRINCIPAL';
PROC GPLOT DATA=WORK.ANNOTER;
 PLOT Y*X=CLUSTER / ANNOTATE=WORK.ANNOTER HREF=0 VREF=0;
RUN;
QUIT;
```

Le fichier RESULT4 contient entre autres :

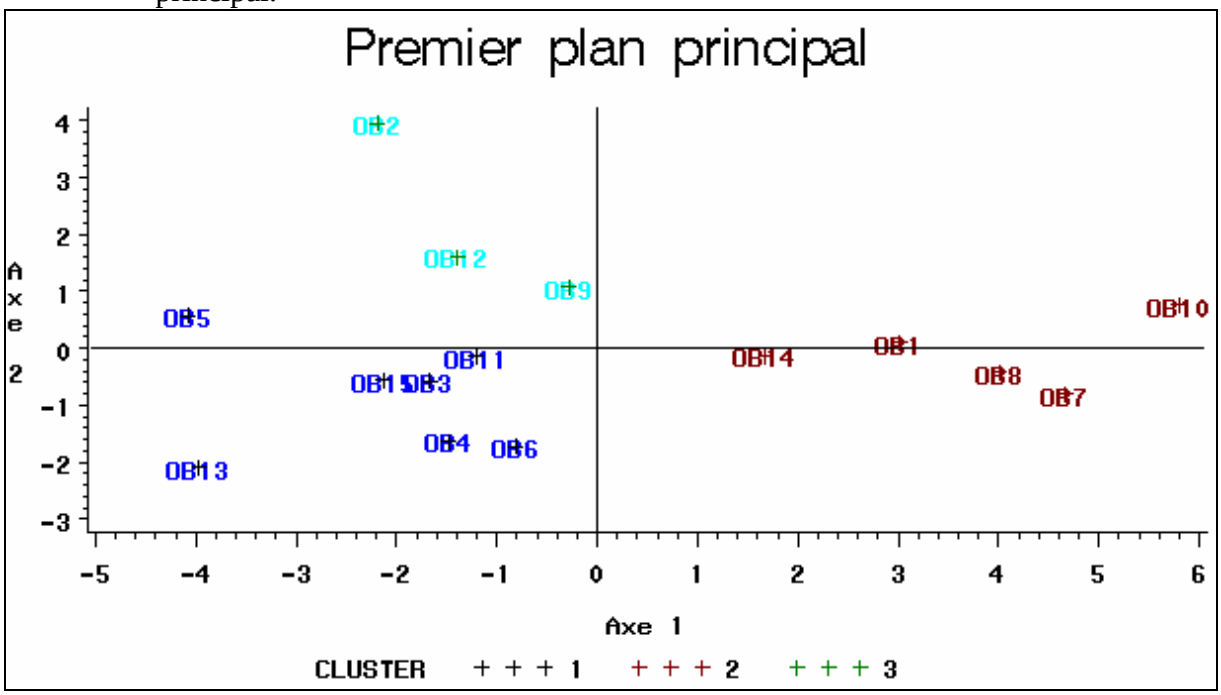


[88] La simplicité légendaire de SAS...

	OBS	AME_JAN	FEV	MAR	AVR	MAI	JUN	JUI	AOU	SEP	OCT	NOV	DEC	STER	RIN1	RIN2	RIN3
1	OB3	2.6	3.7	7.5	10.3	13.8	17.3	19.4	19.1	16.2	11.2	6.6	3.6	1	-1.66741	-0.57244	-0.01846
2	OB15	2.4	3.4	7.1	9.9	13.6	17.1	19.3	18.8	16.0	11.0	6.6	3.4	1	-2.12684	-0.55571	0.19877
3	OB7	5.5	6.6	10.0	13.0	16.8	20.8	23.3	22.8	19.9	15.0	10.2	6.9	2	4.66885	-0.80070	0.34449
4	OB8	5.6	6.7	9.9	12.8	16.2	20.1	22.7	22.3	19.3	14.6	10.0	6.5	2	4.00667	-0.42059	0.17704
5	OB4	1.5	3.2	7.7	10.6	14.5	17.8	20.1	19.5	16.7	11.4	6.5	2.3	1	-1.47740	-1.63071	-0.13283
6	OB6	2.1	3.3	7.7	10.9	14.9	18.5	20.7	20.1	16.9	11.4	6.7	3.1	1	-0.80663	-1.72759	-0.02220
7	OB11	3.4	4.1	7.6	10.7	14.3	17.5	19.1	18.7	16.0	11.4	7.1	4.3	1	-1.19983	-0.15104	-0.21079
8	OB9	5.0	5.3	8.4	10.8	13.9	17.2	18.8	18.6	16.4	12.2	8.2	5.5	3	-0.27175	1.07677	-0.22045
9	OB12	4.8	5.3	7.9	10.1	13.1	16.2	17.9	17.8	15.7	11.6	7.8	5.4	3	-1.38987	1.61446	-0.14715
10	OB1	5.6	6.6	10.3	12.8	15.8	19.3	20.9	21.0	18.6	13.8	9.1	6.2	2	3.01489	0.10559	-0.69639
11	OB14	4.7	5.6	9.2	11.6	14.9	18.7	20.9	20.9	18.3	13.3	8.6	5.5	2	1.67730	-0.13151	0.04190
12	OB10	7.5	8.5	10.8	13.3	16.7	20.1	22.7	22.5	20.3	16.0	11.5	8.2	2	5.80335	0.76254	0.19787
13	OB5	2.4	2.9	6.0	8.9	12.4	15.3	17.1	17.1	14.7	10.4	6.1	3.5	1	-4.07384	0.57502	0.34381
14	OB13	0.4	1.5	5.6	9.8	14.0	17.2	19.0	18.3	15.1	9.5	4.9	1.3	1	-3.96639	-2.09860	0.03340
15	OB2	6.1	5.8	7.8	9.2	11.6	14.4	15.6	16.0	14.7	12.0	9.0	7.0	3	-2.19110	3.95451	0.11099

ainsi que les autres composantes principales. Nous voyons ainsi que la première classe contient les individus 3,15,11,4,6,5 et 13 etc.

Nous pouvons maintenant visualiser le découpage en classes sur le premier plan principal.



Ce qui correspond bien à ce que nous avions trouvé en ACP.

6. Exercice

Le fichier « Banque » (répertoire PUBLIC) contient les informations suivantes concernant 50 clients de la banque SGCM.

SOLD: Solde moyen du compte courant (en F)
CHEQ: Montant moyen des chèques tirés lors du dernier semestre (en F.)
NDEC: Nombre de mois avec découvert sur le compte courant lors de l'année précédente.
MDEC: Montant cumulé des découverts sur le compte courant lors de l'année précédente(en KF)
NBPR: Nombre de produits de la banque utilisés en plus du compte courant.
NEMP: Nombre d'emprunts divers effectués lors des cinq dernières années.
MEMP: Montant total des emprunts effectués lors des cinq dernières années (en KF)
VADD Pourcentage de variation des dépôts d'épargne (pour les douze derniers mois)
DEPO: Montant total des dépôts effectués l'année précédente sur les comptes d'épargne (en KF)
RETR: Montant total des retraits effectués sur les comptes d'épargne l'année précédente (en KF)
VARR: Pourcentage de variation des retraits sur les comptes d'épargne (pour les 12 derniers mois)
TAIL: Taille du ménage du titulaire du compte courant
AGEC: Age du client titulaire du compte courant.

Problème :

La banque souhaite définir des types homogènes de clients afin de pouvoir élaborer des politiques différenciées pour chacun d'eux.

A vous...

1°) Effectuez une ACP normée sur les variables bancaires (les 11 premières). Combien de composantes allez-vous retenir a priori ?
2°) Calculer les corrélations entre les composantes principales retenues et les variables du fichier, en déduire une interprétation « bancaire » de ces composantes principales.
3°) Calculez les contributions des individus aux axes. Pourquoi ne peut-on pas utiliser le 4^{eme} axe pour une interprétation globale ?
4°) Avec les éléments que vous avez, identifiez les grands types de clients de cette banque.
5°) Confirmez votre typologie par une classification ascendante hiérarchique.

N. CORRESP Analyse des correspondances simples

C'est une méthode permettant d'analyser la liaison entre **deux variables qualitatives** A et B avec respectivement p et q modalités. Elle permet d'affiner le test du chi2 d'indépendance que vous connaissez bien. On note *n* le nb total d'individus.

Exemple : Tableau de la répartition de 10000 (=n) étudiants en fonction de la CSP de leur père en 1975-1976.

	Droit	Eco	Lettres	Sciences	Médecine	Pharma	Pluri	IUT	**Total**
Exploitant	80	36	134	99	65	28	11	58	**511**
Salarié ag.	6	2	15	6	4	1	1	4	**39**
Patron	168	74	312	137	208	53	21	62	**1035**
Cadre sup.	470	191	806	400	876	164	45	79	**3031**
Ca. Moy.	236	99	493	264	281	56	36	87	**1552**
Employ	145	52	281	133	135	30	20	54	**850**
Ouvrier	166	64	401	193	127	23	28	129	**1131**
Pers. serv.	16	6	27	11	8	2	2	8	**80**
Autres	305	115	624	247	301	47	42	90	**1771**
Total	**1592**	**639**	**3093**	**1490**	**2005**	**404**	**206**	**571**	**10000**

Nous étudions donc ici une population de 10000 personnes sur lesquelles agissent deux variables qualitatives : A : CSP Père et B : Type d'étude.
La variable A comporte 9 modalités (p=9) et B 8 modalités (q=8).

Plan d'étude

Du tableau précédent, nous pouvons tirer trois tableaux de fréquences : Celui des fréquences totales, des fréquences lignes et des fréquences colonnes. Nous pourrons ensuite, à l'aide d'une ACP sur les tableaux de fréquences lignes et d'une autre sur les fréquences colonnes, synthétiser la liaison entre nos variables.
Du tableau de contingence précédent, nous pouvons déduire le tableau de fréquences en divisant les effectifs par 10000.

	Droit	Eco	Lettres	Sciences	Médecine	Pharma	Pluri	IUT	**Total**
Exploitant	0,0080	0,0036	0,0134	0,0099	0,0065	0,0028	0,0011	0,0058	**0,0511**
Salarié ag.	0,0006	0,0002	0,0015	0,0006	0,0004	0,0001	0,0001	0,0004	**0,0039**
Patron	0,0168	0,0074	0,0312	0,0137	0,0208	0,0053	0,0021	0,0062	**0,1035**
Cadre sup.	0,0470	0,0191	0,0806	0,0400	0,0876	0,0164	0,0045	0,0079	**0,3031**
Ca. Moy.	0,0236	0,0099	0,0493	0,0264	0,0281	0,0056	0,0036	0,0087	**0,1552**
Employ	0,0145	0,0052	0,0281	0,0133	0,0135	0,0030	0,0020	0,0054	**0,0850**
Ouvrier	0,0166	0,0064	0,0401	0,0193	0,0127	0,0023	0,0028	0,0129	**0,1131**
Pers. serv.	0,0016	0,0006	0,0027	0,0011	0,0008	0,0002	0,0002	0,0008	**0,0080**
Autres	0,0305	0,0115	0,0624	0,0247	0,0301	0,0047	0,0042	0,0090	**0,1771**
Total	**0,1592**	**0,0639**	**0,3093**	**0,1490**	**0,2005**	**0,0404**	**0,0206**	**0,0571**	**1**

On peut voir que 5,11% des étudiants de l'échantillon ont un père exploitant et que 20% font des études de médecine. On peut également lire que 1,64% des individus ont un père cadre sup. et font des études de médecine.

1. Étude des profils lignes

a) Tableau des profils lignes

Nous pouvons diviser chaque ligne par le total correspondant pour obtenir des fréquences lignes [89]:

	Droit	Eco	Lettres	Sciences	Médecine	Pharma	Pluri	IUT	**Total**
Exploitant	0,1566	0,0705	0,2622	0,1937	0,1272	0,0548	0,0215	0,1135	**1**
Salarié ag.	0,1538	0,0513	0,3846	0,1538	0,1026	0,0256	0,0256	0,1026	**1**
Patron	0,1623	0,0715	0,3014	0,1324	0,2010	0,0512	0,0203	0,0599	**1**
Cadre sup.	0,1551	0,0630	0,2659	0,1320	0,2890	0,0541	0,0148	0,0261	**1**
Ca. Moy.	0,1521	0,0638	0,3177	0,1701	0,1811	0,0361	0,0232	0,0561	**1**
Employ	0,1706	0,0612	0,3306	0,1565	0,1588	0,0353	0,0235	0,0635	**1**
Ouvrier	0,1468	0,0566	0,3546	0,1706	0,1123	0,0203	0,0248	0,1141	**1**
Pers. serv.	0,2000	0,0750	0,3375	0,1375	0,1000	0,0250	0,0250	0,1000	**1**
Autres	0,1722	0,0649	0,3523	0,1395	0,1700	0,0265	0,0237	0,0508	**1**

Ces nombres correspondent en fait les probabilités conditionnelles sachant la CSP. Par exemple, pour notre échantillon, la probabilité pour un étudiant de faire des études de médecine sachant que son père est cadre sup. est de 28.9% , elle n'est que de 10% avec un père ouvrier. Nous voyons donc apparaître une liaison entre les deux variables.[90]

Nous pouvons donc analyser la liaison entre les deux variables qualitatives CSP et ETUDE en étudiants les différences entre les profils lignes. Il y a indépendance parfaite entre les deux si tous les profils lignes sont identiques.[91]

Sous SAS, nous pouvons obtenir les profils lignes en utilisant la procédure FREQ ou en utilisant la procédure CORRESP avec l'option RP (=Row Profile).

```
PROC CORRESP DATA=MOI.ETUD RP SHORT;
 VAR DRO ECO LET SCI MED PHA PLU IUT;
 ID CSP;
RUN;
```

qui nous donne entre autres :

The Correspondence Analysis Procedure

[89] Tous les calculs suivants (Profils lignes, colonnes), effectifs observés, théoriques (si indépendance parfaite), chi2 peuvent être obtenus en dans la commande Stat/Table/Simple correspondance Analysis/ Indiquer les colonnes du tableau de contingence dans Column of contingency table puis cliquez sur RESULT vous pouvez alors choisir Row Profile (profil ligne), Column profile, expected (effectifs attendus...)

[90] Nous pouvons quantifier cette liaison en calculant les contributions au chi2 (cf. Test du chi2 d'indépendance)

[91] C'est à dire que toutes les lignes du tableau précédent sont égales. Ainsi, pour notre échantillon, la probabilité de faire médecine est la même pour toutes les CSP. Ici, ce n'est visiblement pas le cas.

Row Profiles

	DRO	ECO	LET	SCI	MED	PHA	PLU	IUT
Exploitant	0.156556	0.070450	0.262231	0.193738	0.127202	0.054795	0.021526	0.113503
Salarié ag.	0.153846	0.051282	0.384615	0.153846	0.102564	0.025641	0.025641	0.102564
Patron	0.162319	0.071498	0.301449	0.132367	0.200966	0.051208	0.020290	0.059903
Cadre sup.	0.155064	0.063016	0.265919	0.131970	0.289014	0.054108	0.014847	0.026064
Cadre Moyen	0.152062	0.063789	0.317655	0.170103	0.181057	0.036082	0.023196	0.056057
Employ	0.170588	0.061176	0.330588	0.156471	0.158824	0.035294	0.023529	0.063529
Ouvrier	0.146773	0.056587	0.354553	0.170645	0.112290	0.020336	0.024757	0.114058
Pers. service	0.200000	0.075000	0.337500	0.137500	0.100000	0.025000	0.025000	0.100000
Autres	0.172219	0.064935	0.352343	0.139469	0.169960	0.026539	0.023715	0.050819

b) ACP sur le tableau des profils lignes

Chaque ligne du tableau précédent représente une CSP. Chaque CSP peut donc être considérée comme un point de R^q. On peut montrer que l'inertie[92] du nuage de points ainsi formé est égale à χ^2/n où χ^2 est égale à la statistique du test du chi2 d'indépendance que vous connaissez bien.

Dans le cas d'une indépendance parfaite entre les deux variables, le χ^2 est nul, tous les points représentant les profils lignes sont confondus. L'inertie du nuage est nulle, le nuage étant réduit à un point.

Comme en ACP classique, nous allons chercher des axes principaux sur lesquels nous allons projeter notre nuage en conservant le mieux possible l'inertie initiale.

Sous SAS, il suffit de taper le programme suivant :

```
PROC CORRESP DATA=MOI.ETUD RP;
 VAR DRO ECO LET SCI MED PHA PLU IUT;
 ID CSP;
RUN;
```

[92] C'est la somme des carrés des distances des points du nuage au barycentre du nuage. La distance employée ici est la distance du chi2. $D^2(X,Y)=\Sigma(x_j-y_j)^2/f_{.j}$

(1) Nombre d'axes à retenir

SAS donne :

The Correspondence Analysis Procedure

Inertia and Chi-Square Decomposition

Singular Values	Principal Inertias	Chi-Squares	Percents	17 34 51 68 85
0.19934	0.03974	397.372	83.72%	**********
0.07384	0.00545	54.517	11.49%	***
0.03361	0.00113	11.297	2.38%	*
0.03099	0.00096	9.604	2.02%	*
0.01154	0.00013	1.332	0.28%	
0.00732	0.00005	0.536	0.11%	
0.00103	0.00000	0.011	0.00%	
	0.04747	474.668	(Degrees of Freedom = 56)	

L'inertie totale est égale à 0.0475=474668/10000 (χ^2=474668)
Nous voyons que 95% de l'inertie du nuage initial des Profil lignes est expliqué par les deux premiers axes dont 83.7% pour le premier axe. Il est clair que le premier axe est essentiel. Nous conserverons aussi le 2eme axe mais ne nous faisons pas trop d'illusions sur lui.

SAS donne ensuite les contributions des CSP à la construction des axes et d'autres statistiques fort intéressantes :

Row Coordinates[93]

	Dim1	Dim2
Exploitant	0.232788	0.226391
Salarié ag.	0.322778	-.032999
Patron	-.019711	0.029833
Cadre sup.	-.263519	0.023860
Cadre Moyen	0.049553	-.011446
Employ	0.103477	-.033166
Ouvrier	0.334848	0.029924
Pers. service	0.291183	-.021171
Autres	0.068162	-.115071

Summary Statistics for the Row Points

	Quality	Mass	Inertia
Exploitant	0.967440	0.051100	0.117335
Salarié ag.	0.941372	0.003900	0.009188
Patron	0.211148	0.103500	0.013203
Cadre sup.	0.995147	0.303100	0.449242
Cadre Moyen	0.415316	0.155200	0.020363
Employ	0.893202	0.085000	0.023672
Ouvrier	0.957777	0.113100	0.281163
Pers. service	0.809763	0.008000	0.017740
Autres	0.980096	0.177100	0.068094

Partial Contributions to Inertia for the Row Points

	Dim1	Dim2
Exploitant	0.069686	0.480406
Salarié ag.	0.010225	0.000779
Patron	0.001012	0.016897
Cadre sup.	0.529681	0.031651
Cadre Moyen	0.009590	0.003730
Employ	0.022904	0.017150
Ouvrier	0.319125	0.018577
Pers. service	0.017070	0.000658
Autres	0.020707	0.430153

Squared Cosines for the Row Points

	Dim1	Dim2
Exploitant	0.497195	0.470245
Salarié ag.	0.931634	0.009738
Patron	0.064163	0.146984
Cadre sup.	0.987055	0.008092
Cadre Moyen	0.394280	0.021036
Employ	0.809994	0.083208
Ouvrier	0.950188	0.007589
Pers. service	0.805505	0.004258
Autres	0.254570	0.725526

[93] Remarquons qu'elles sont de signe opposé aux coordonnées de Minitab. Ceci vient du choix des vecteurs propres dans la diagonalisation et cela ne change rien au résultat final.

Légende :

Minitab	**SAS**	**Interprétation**
Qual	Quality	Qualité globale de représentation (somme des cos carrés)
Mass	Mass	% d'individus ayant choisi la modalité ($f_{i,.}$)
Inert	Inertia	Part d'inertie totale expliquée par la modalité. ($f_{i,.} \frac{\sum_j coord(\text{mod i sur axe j})^2}{\chi^2 / n}$)
Coord	Coordinates	Coordonnée de la modalité sur l'axe.
Corr	Squared cosine	Qualité de représentation de l'individu sur l'axe. (Cosinus carré de l'angle entre l'individu et l'axe). Plus ce nombre est proche de 1 meilleure est cette qualité.
Contr	Partial contribution to inertia	Part d'inertie de l'axe expliquée par la modalité. C'est une quantité essentielle pour l'interprétation des axes. (= $\frac{f_{i,.}}{vp_j}$ coord(modalité i sur axe j)²) vp_j=valeur propre axe j.

(2) Etude de l'axe 1 (Component 1)

Comme nous l'avons dit, il conserve 83,7% de l'inertie initiale. Il est donc essentiel.

Pour interpréter l'axe 1, on va chercher les individus **ayant une forte contribution** à l'axe 1 et regarder le signe de leur coordonnée sur l'axe.

Individus à forte contribution sur l'axe 1 avec une coordonnée négative.	**Individus à forte contribution sur l'axe 1 avec une coordonnée positive.**
Ouvrier (31%)	Cadre sup.(53%)

L'axe 1 oppose donc la CSP Ouvrier à la CSP Cadre Sup. Vous pouvez l'interpréter facilement.[94]

[94] (Au besoin aidez vous du graphique de la page suivante en utilisant uniquement les points ayant une **bonne qualité de représentation** (Corr, ou squared cosines)).

(3) Etude de l'axe 2 (Component 2)

Quel pourcentage d'inertie conserve-t-il ? Faites l'étude précédente pour l'axe 2.

Individus à forte contribution sur l'axe 2 avec une coordonnée négative.	**Individus à forte contribution sur l'axe 2 avec une coordonnée positive.**

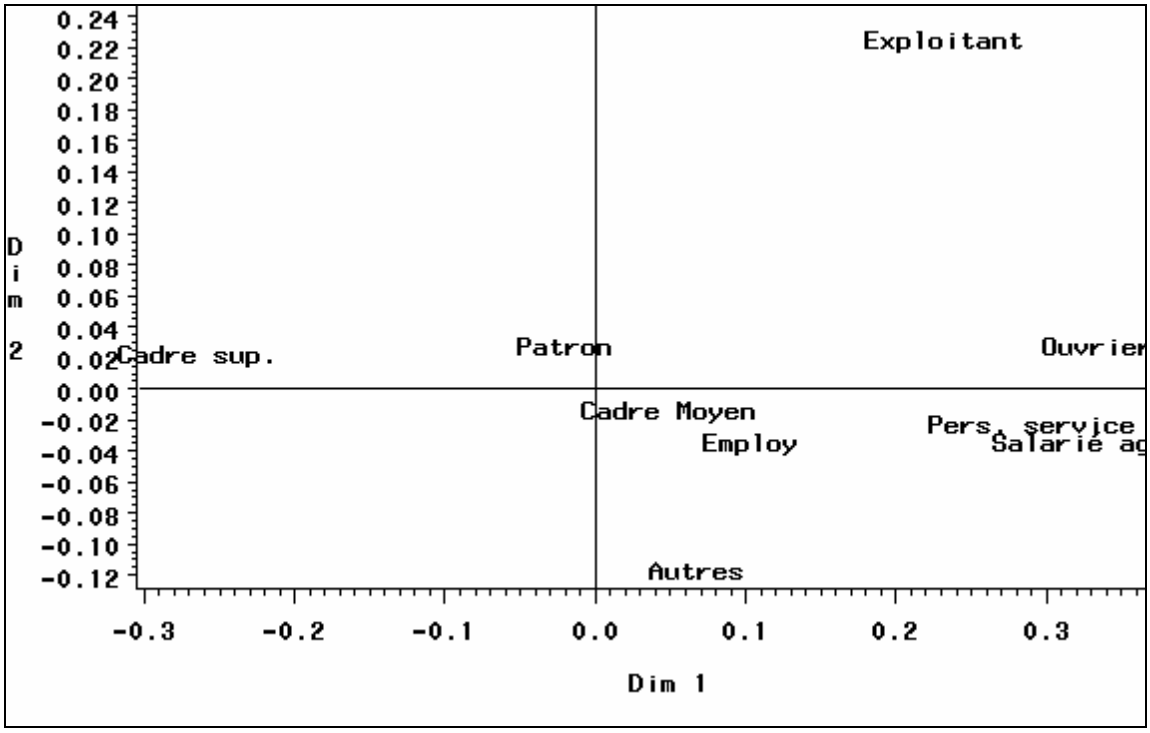
Peut-on trouver une interprétation ici ?[95] Vous pourrez vous aider du graphique suivant.

(4) Projection du nuage dans le plan des deux premiers axes.

Pour SAS, il faut taper :

```
/*On relance l'AFC en stockant les résultats dans Work.corr*/
PROC CORRESP DATA=MOI.ETUD OUTC=CORR;
 VAR DRO ECO LET SCI MED PHA PLU IUT;
 ID CSP;
RUN;

/*Modification de work.corr pour annoter notre graphique*/
DATA CORR;
 SET CORR;
 IF _TYPE_='OBS';  /*Nous ne prenons que les CSP*/
 Y=DIM2;           /*Axe 2 vertical, Axe 1 horizontal*/
 X=DIM1;
 XSYS='2';
 YSYS='2';
 TEXT=CSP;        /*CSP contient les noms des categories*/
 SIZE=1;
 LABEL Y='DIM 2'
       X='DIM 1';
 KEEP X Y TEXT XSYS YSYS SIZE;
RUN;
/*Nous lancons ensuite GPLOT pour avoir un graphique haute res*/

PROC GPLOT DATA=CORR;
 SYMBOL1 V=NONE;
 PLOT Y*X=1/ ANNOTATE=CORR FRAME HREF=0 VREF=0;
RUN;
QUIT;
```

[95] La difficulté vient du fait que l'inertie conservée par cet axe est faible.



Nous pouvons utiliser le graphique précédent **uniquement avec les points bien représentés.** (QUAL(ou Squared cosine) important, QUAL=CORR axe1 + CORR axe 2).[96]

Deux points (bien représentés) et proches sur ce graphique signifie que les modalités correspondantes ont des profils qui se ressemblent ou encore que les barycentre des individus ayant choisis ces modalités sont proches.

Il est maintenant aisé d'interpréter les deux axes relativement aux CSP.

Nous allons maintenant analyser la liaison en privilégiant la variable Type d'étude.

[96] Ici seuls Patron et Cadre Moyen sont mal représentés. Il ne faut donc pas les inclure dans une interprétation.

2. Etude des profils colonnes

a) Tableau des profils colonnes

Nous pouvons diviser chaque ligne par le total correspondant pour obtenir des fréquences lignes :

Dans le tableau précédent nous avons privilégié les CSP. Nous allons maintenant nous intéresser à la variable ETUDE :

De la même façon, nous pouvons construire le tableau des fréquences colonnes en divisant chaque élément par la somme de sa colonne.

	Droit	Eco.	Lettres	Sciences	Médecine	Pharma	Pluri	IUT
Exploitant	0,0503	0,0563	0,0433	0,0664	0,0324	0,0693	0,0534	0,1016
Salarié ag.	0,0038	0,0031	0,0048	0,0040	0,0020	0,0025	0,0049	0,0070
Patron	0,1055	0,1158	0,1009	0,0919	0,1037	0,1312	0,1019	0,1086
Cadre sup.	0,2952	0,2989	0,2606	0,2685	0,4369	0,4059	0,2184	0,1384
Ca. Moy.	0,1482	0,1549	0,1594	0,1772	0,1401	0,1386	0,1748	0,1524
Employ	0,0911	0,0814	0,0909	0,0893	0,0673	0,0743	0,0971	0,0946
Ouvrier	0,1043	0,1002	0,1296	0,1295	0,0633	0,0569	0,1359	0,2259
Pers. serv.	0,0101	0,0094	0,0087	0,0074	0,0040	0,0050	0,0097	0,0140
Autres	0,1916	0,1800	0,2017	0,1658	0,1501	0,1163	0,2039	0,1576
Total	**1,0000**	**1,0000**	**1,0000**	**1,0000**	**1,0000**	**1,0000**	**1,0000**	**1,0000**

Sous SAS, il suffit de taper le programme suivant pour obtenir ce tableau. (CP=column profile)

```
PROC CORRESP DATA=MOI.ETUD CP;
 VAR DRO ECO LET SCI MED PHA PLU IUT;
 ID CSP;
RUN;
```

Ainsi, les chiffres précédents correspondent aux probabilités conditionnelles sachant le type d'étude. Par exemple, 9% des étudiants en Droit ont un père Employé.

Comme précédemment, s'il y a indépendance parfaite entre les deux variables, les profils colonnes doivent tous être égaux.[97]

Dans ce tableau ce sont les types d'études qui sont mis en avant.

[97] Dans ce cas, on montre que les profils lignes sont eux aussi égaux.

b) ACP sur les tableaux de profils colonnes

Comme précédemment, nous allons effectuer une ACP sur le tableau des profils colonnes. **Nous pouvons montrer le résultat suivant : Les valeurs propres obtenues sont les mêmes, de plus, les axes principaux des deux ACP sont en relation directe.** Nous pourrons donc ensuite « superposer » les deux analyses.

(1) Etude de l'axe 1 (Component 1)

Column Coordinates

	Dim1	Dim2
DRO	0.003548	-.032659
ECO	-.011480	0.011575
LET	0.090282	-.074258
SCI	0.095562	0.050293
MED	-.313387	0.020203
PHA	-.249015	0.181087
PLU	0.177417	-.062809
IUT	0.477149	0.172698

Summary Statistics for the Column Points

	Quality	Mass	Inertia
DRO	0.316593	0.159200	0.011433
ECO	0.064286	0.063900	0.005566
LET	0.984721	0.309300	0.090425
SCI	0.713400	0.149000	0.051312
MED	0.986455	0.200500	0.422290
PHA	0.926296	0.040400	0.087107
PLU	0.952004	0.020600	0.016148
IUT	0.981101	0.057100	0.315720

Partial Contributions to Inertia for the Column Points

	Dim1	Dim2
DRO	0.000050	0.031147
ECO	0.000212	0.001570
LET	0.063444	0.312846
SCI	0.034242	0.069130
MED	0.495541	0.015011
PHA	0.063043	0.243011
PLU	0.016318	0.014907
IUT	0.327150	0.312378

Squared Cosines for the Column Points

	Dim1	Dim2
DRO	0.003694	0.312899
ECO	0.031876	0.032409
LET	0.587362	0.397359
SCI	0.558664	0.154736
MED	0.982373	0.004083
PHA	0.605882	0.320414
PLU	0.845978	0.106026
IUT	0.867464	0.113637

Pour interpréter l'axe 1, on va chercher les individus ayant une forte contribution[98] à l'axe 1 et regarder le signe de leur coordonnée sur l'axe.

Complétez le tableau suivant :

Individus à forte contribution sur l'axe 1 avec une coordonnée négative.	**Individus à forte contribution sur l'axe 1 avec une coordonnée positive.**

Interprétez l'axe 1 [99]:

(2) Etude de l'axe 2 (Component 2)

De même essayez d'interpréter l'axe 2 sans perdre de vue son faible pourcentage d'inertie expliquée.

Complétez le tableau suivant :

Individus à forte contribution sur l'axe 2 avec une coordonnée négative.	**Individus à forte contribution sur l'axe 2 avec une coordonnée positive.**

Interprétez l'axe 2:

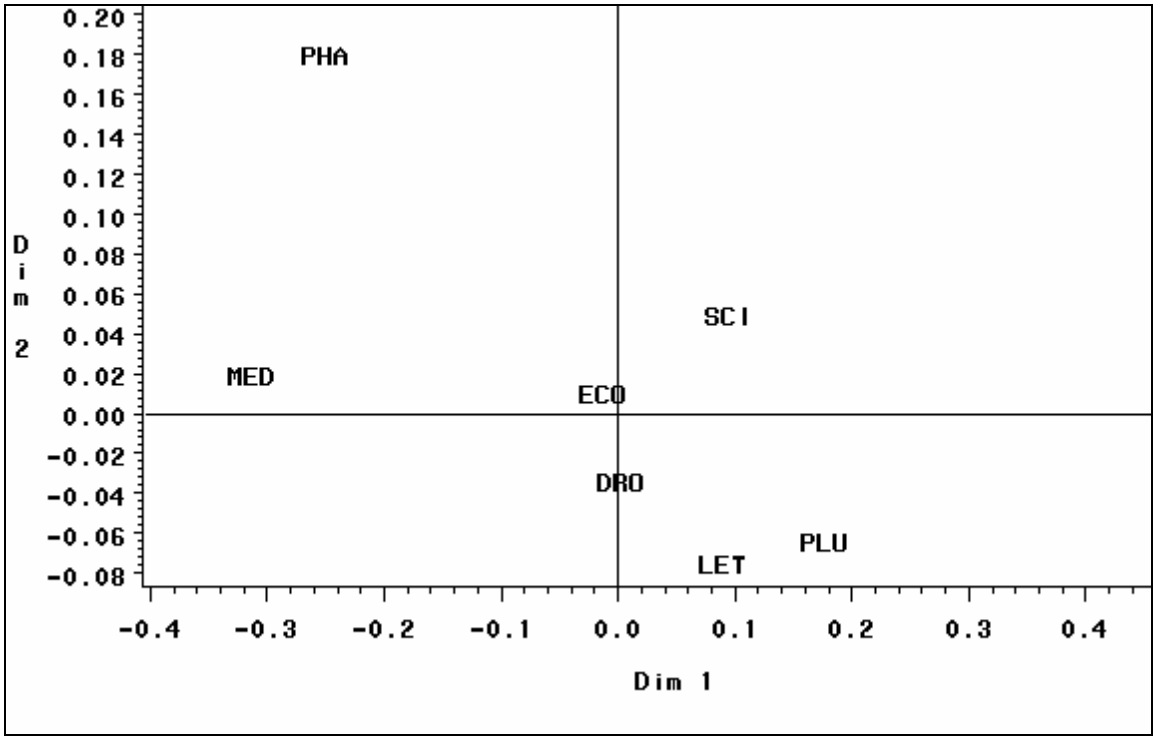
[98] Lorsque certains individus ont une trop forte contribution, on recommence l'étude en plaçant ces individus en supplémentaires. Ils ne sont plus utilisés dans les calculs mais peuvent être visualisés sur le graphique et peuvent être utilisé pour l'interprétation.

[99] En vous servant aussi de la représentation graphique des pages suivantes pour les individus bien représentés par cet axe (CORR important)

(3) Projection du nuage dans le plan des deux premiers axes.

Sous SAS, il suffit de modifier la ligne IF _TYPE_='OBS' par IF _TYPE_='VAR'.

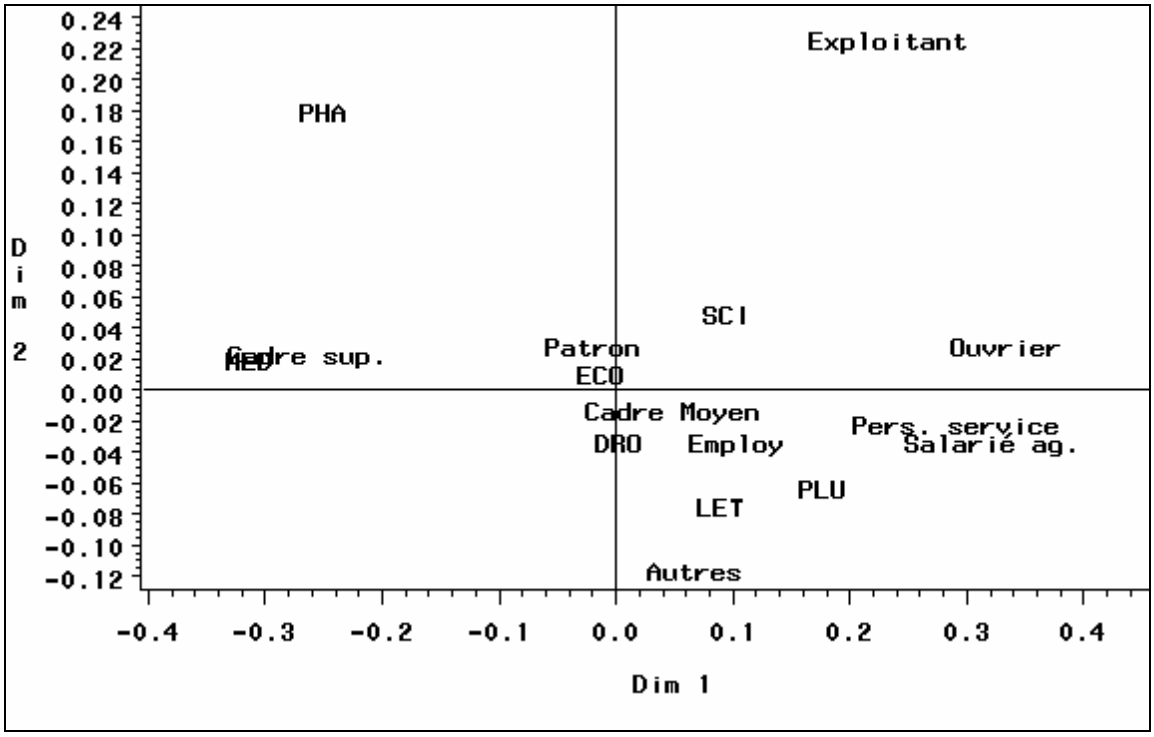
```
PROC CORRESP DATA=MOI.ETUD OUTC=CORR;
 VAR DRO ECO LET SCI MED PHA PLU IUT;
 ID CSP;
RUN;

DATA CORR;
 SET CORR;
 IF _TYPE_='VAR';
 Y=DIM2;
 X=DIM1;
 XSYS='2';
 YSYS='2';
 TEXT=CSP;
 SIZE=1;
 LABEL Y='DIM 2'
       X='DIM 1';
 KEEP X Y TEXT XSYS YSYS SIZE;
RUN;

PROC GPLOT DATA=CORR;
 SYMBOL1 V=NONE;
 PLOT Y*X=1/ ANNOTATE=CORR FRAME HREF=0 VREF=0;
RUN;
QUIT;
```



3. Lien entre les deux analyses

Les valeurs propres des deux ACP précédentes sont les mêmes comme nous l'avons déjà dit. On peut montrer que les axes principaux de l'une sont les composantes principales de l'autre à un facteur multiplicatif près.
Ceci nous autorise à superposer les deux représentations graphiques précédentes mais en restant très prudent dans l'interprétation d'un tel graphique.

Pour SAS, il suffit de supprimer la ligne IF _TYPE_= du programme précédent.



ATTENTION

Dans le graphique précédent, nous avons superposé des individus qui n'appartiennent pas au même espace. En conséquence, il est dangereux d'interpréter la proximité entre deux points (même bien représentés[100]) associés à deux variables distinctes[101]. Par contre, il est possible d'interpréter la proximité entre deux points (bien représentés[102]) associés à une même variable

Exemple : Les profils « Salarié agricole » et « personnel de service » sont proches ce qui veut dire que les barycentres des étudiants dont le père est salarié agricole est proche du barycentre des étudiants dont le père est « Personnel de service ». Par contre la proximité de « Cadre » et de « Médecine » ne peut être utilisée ; tout au moins directement[103].

[100] Avec un QUAL élevé.

[101] Tenenhauss utilise les représentations barycentriques pour interpréter ces proximités. Ceci dépasse le cadre de ce cours.

[102] Idem

[103] En fait, on peut relier ces deux choses en passant par l'intermédiaire de l'axe principal (qui est directement lié entre les deux ACP).

Par contre nous pouvons maintenant donner une interprétation complète de l'axe 1 et de l'axe 2.

Brève synthèse des analyses des profils lignes et colonnes

« Le premier axe oppose les études de médecine, caractéristiques des fils de professions libérales et cadres sup, aux études en IUT, caractéristiques des fils d'ouvrier.

Le deuxième axe oppose les fils d'exploitant agricole à ceux de la CSP « Autres » et les études de pharmacie et d'IUT aux études de lettres. »

(*Saporta* p208)

4. Syntaxe de PROC CORRESP sous SAS

```
PROC CORRESP options ;
  TABLES  variables lignes  variables colonnes (Données brutes)
 ou
  VAR  Variables ;                    (Données en tableau de contingence)

  ID variable ;                (variable identificatrice[104] avec VAR seulement)
  SUPPLEMENTARY  variables ;   (Variables supplémentaires.)
  WEIGHT  variable ;           Sert à spécifier les poids des individus. Les poids négatifs
                               désignent des individus supplémentaires.
  BY  Variables  ;             idem...
RUN ;
```

Les principales options étant :

Fichiers d'entrée sortie

DATA= *nom de fichier*
Nom du fichier de données contenant les données sous forme brute ou tabulaire.

OUTC= *nom de fichier*
Nom du fichier où seront stockés les résultats de l'analyse en plus du fichier original. (Coordonnées des modalités etc.)

OUTF= *nom de fichier*
Nom du fichier où seront stockés les fréquences etc.

Options de calcul

DIMENS=*nombre*
Nombre de dimensions ou axes à calculer. **2 par défaut.**
MCA
Pour effectuer un analyse des correspondances multiples.

Options d'affichage

ALL
Équivalent à OBSERVED, RP,CP, CELLCHI2, EXPECTED, DEVIATION

CELLCHI2
Contribution au chi2 de chaque cellule.

[104] Permet de mettre des étiquettes aux lignes du tableau de données.

CP
: Profils colonnes

DEVIATION
: Différence entre les valeurs prédites et observées.

EXPECTED
: Valeurs prédites (en cas d'indépendance totale entre les deux variables)

NOCOLUMN
: Pas d'affichage des coordonnées des colonnes.

NOPRINT
: Pas d'affichage

NOROW
: Pas d'affichage des lignes.

OBSERVED
: Valeurs observées.

RP
: Profils colonnes

SHORT
: Supprime l'affichage des statistiques sur les points et coordonnées.

O. CORRESP Analyse des Correspondance Multiples

Dans le chapitre précédent, nous avons vu comment analyser la liaison entre deux variables qualitatives en utilisant l'analyse des correspondances simples. L'analyse des correspondances multiples étend l'étude précédente à l'étude de p variables qualitatives. [105] Une ACM est une analyse factorielle des correspondance du tableau disjonctif complet.

1. Tableau disjonctif complet

Ce tableau est un codage particulier qui permet de n'avoir dans chaque colonne qu'une modalité et une seule des variables à étudier. Il y a donc autant de colonnes que de modalité. Chaque colonne n'est composée que de 0 ou de 1.

Prenons les données suivantes :

	Sexe	Groupe	Profil
individu 1	1	A	Bon
individu 2	2	B	Bon
individu 3	1	A	Mauvais
individu 4	1	C	Moyen
individu 5	2	C	Moyen

Nous avons 5 individus sur lesquels agissent 3 variables Sexe (2 modalités) ; Groupe(3 modalités) et Profil (3 modalités).

Nous allons recoder le tableau précédent de la façon suivante :

Individu	Sexe=1	Sexe=2	Groupe A	Groupe B	Groupe C	Bon profil	Profil moyen	Mauvais profil
1	1	0	1	0	0	1	0	0
2	0	1	0	1	0	1	0	0
3	1	0	1	0	0	0	0	1
4	1	0	0	0	1	0	1	0
5	0	1	0	0	1	0	1	0

Chaque variable a été décomposée en utilisant la fonction indicatrice de ses modalités. **C'est le tableau disjonctif complet** (TDC en abrégé)

Chaque colonne de notre tableau représente donc une modalité (ou catégorie) et chaque ligne un individu.

Effectuer une ACM consiste à effectuer une ACS sur le tableau disjonctif complet.

[105] Mathématiquement parlant, l'ACM résume p variables qualitatives par des variables numériques de variance maximale et les plus corrélées possibles (au sens du rapport de corrélation) avec les variables initiales.

2. Exemple

Nous allons étudier une population de 27 chiens sur lesquels agissent 7 variables :

Taille (-,+ ou ++)
Poids (-,+ ou ++) (1, 2 ou 3)
Vélocité (-,+ ou ++) (1, 2 ou 3)
Intelligence (-,+ ou ++) (1, 2 ou 3)
Affection (Oui (1)ou Non(0))
Agressivité (Oui(1) ou Non(0))
Fonction (Compagnie, Chasse, Utilité)

Nom	Taille	Poids	Vélocité	Intelligence	Affection	Agressivité	Utilité
BEAUCERON	Taille3	Poids2	vel3	int2	oui	oui	Utile
BASSET	Taille1	Poids1	vel1	int1	non	oui	Chasse
BERGER_ALLEMAND	Taille3	Poids2	vel3	int3	oui	oui	Utile
BOXER	Taille2	Poids2	vel2	int2	oui	oui	Compagnie
BULL-DOG	Taille1	Poids1	vel1	int2	oui	non	Compagnie
BULL-MASTIFF	Taille3	Poids3	vel1	int3	non	oui	Utile
CANICHE	Taille1	Poids1	vel2	int3	oui	non	Compagnie
CHIHUAHUA	Taille1	Poids1	vel1	int1	oui	non	Compagnie
COKER	Taille2	Poids1	vel1	int2	oui	oui	Compagnie
COLLEY	Taille3	Poids2	vel3	int2	oui	non	Compagnie
DALMATIEN	Taille2	Poids2	vel2	int2	oui	non	Compagnie
DOBERMANN	Taille3	Poids2	vel3	int3	non	oui	Utile
DOGUE_ALLEMAND	Taille3	Poids3	vel3	int1	non	oui	Utile
EPAGNEUL_BRETON	Taille2	Poids2	vel2	int3	oui	non	Chasse
EPAGNEUL_FRANCAIS	Taille3	Poids2	vel2	int2	non	non	Chasse
FOX-HOUND	Taille3	Poids2	vel3	int1	non	oui	Chasse
FOX-TERRIER	Taille1	Poids1	vel2	int2	oui	oui	Compagnie
GRAND_BLEU_DE_GASCOG	Taille3	Poids2	vel2	int1	non	oui	Chasse
LABRADOR	Taille2	Poids2	vel2	int2	oui	non	Chasse
LEVRIER	Taille3	Poids2	vel3	int1	non	non	Chasse
MASTIFF	Taille3	Poids3	vel1	int1	non	oui	Utile
PEKINOIS	Taille1	Poids1	vel1	int1	oui	non	Compagnie
POINTER	Taille3	Poids2	vel3	int3	non	non	Chasse
SAINT-BERNARD	Taille3	Poids3	vel1	int2	non	oui	Utile
SETTER	Taille3	Poids2	vel3	int2	non	non	Chasse
TECKEL	Taille1	Poids1	vel1	int2	oui	non	Compagnie
TERRE-NEUVE	Taille3	Poids3	vel1	int2	non	non	Utile

Nous allons effectuer une ACM sur 6 variables actives (les 6 premières). Ceci revient à effectuer une ACP sur le tableau disjonctif complet.[106]

[106] La septième variable (illustrative) peut être mise en « variable supplémentaire ». Cela permet de l'utiliser dans l'interprétation sans qu'elle ne joue de rôle dans la détermination des axes.

Pour SAS, on peut lancer la programme suivant :

```
proc corresp data=pub.chiens mca obs all;
tables taille -- agressiv;
run;
```

Nous obtenons les résultats suivants :

The Correspondence Analysis Procedure

Inertia and Chi-Square Decomposition

Singular Values	Principal Inertias	Chi-Squares	Percents	6 12 18 24 30
0.69398	0.48161	139.417	28.90%	**********
0.62027	0.38474	111.375	23.08%	**********
0.45930	0.21095	61.068	12.66%	**********
0.39693	0.15755	45.609	9.45%	********
0.38747	0.15013	43.461	9.01%	********
0.35113	0.12330	35.692	7.40%	******
0.28542	0.08146	23.582	4.89%	****
0.21370	0.04567	13.221	2.74%	**
0.15343	0.02354	6.815	1.41%	*
0.08782	0.00771	2.233	0.46%	
	1.66667	482.471 (Degrees of Freedom = 225)		

a) Nombre de composantes à retenir

Différence essentielle entre l'AFC simple et l'ACM :

> Contrairement au cas précédent (AFC), l'inertie totale du nuage n'est plus liée à la structure de la liaison entre les variables. On peut montrer qu'elle est égale à a/b-1 où a est le nombre de modalité des variables actives et p le nombre de variables actives. Ici, nous avons 6 variables actives définissant 3+3+3+3+2+2=16 modalités. D'où I=16/6-1≈1.667

Il en résulte que **les valeurs propres et les pourcentages d'inertie expliqués par les axes n'ont qu'un intérêt relatif en ACM.**

Pour déterminer le nombre de composantes à retenir, nous pouvons utiliser la règle suivante :

> La moyenne des valeurs propres vaut 1/p où p est le nombre de variables actives. Comme en ACP, une méthode possible pour choisir le nombre de composantes est de ne conserver que les valeurs propres supérieures à 1/p. Ici, nous n'en retiendrions que trois au maximum (1/6≅0.167).

Toutefois, nous remarquons une chute brutale de l'inertie après la deuxième, nous nous contenterons donc de deux composantes.

b) Interprétation des axes

Column Coordinates[107]

	Dim1	Dim2
ta+	0.85109	-1.23172
ta++	-0.83668	-0.02058
ta-	1.18496	0.92390
Poids+	-0.30541	-0.81888
Poids++	-1.01513	0.97390
Poids-	1.16892	0.82434
vel+	0.60369	-0.88781
vel++	-0.89210	-0.37183
vel-	0.31994	1.04490
int+	0.36944	-0.28550
int++	-0.33507	-0.45948
int-	-0.34905	0.80855
af+	0.77550	-0.26694
af-	-0.83515	0.28747
ag+	-0.43154	0.20920
ag-	0.40071	-0.19425

Summary Statistics for the Column Points

	Quality	Mass	Inertia
ta+	0.509428	0.030864	0.081481
ta++	0.875561	0.092593	0.044444
ta-	0.790197	0.043210	0.074074
Poids+	0.822586	0.086420	0.048148
Poids++	0.449768	0.030864	0.081481
Poids-	0.861437	0.049383	0.070370
vel+	0.485327	0.049383	0.070370
vel++	0.467051	0.055556	0.066667
vel-	0.702458	0.061728	0.062963
int+	0.202428	0.080247	0.051852
int++	0.092398	0.037037	0.077778
int-	0.326566	0.049383	0.070370
af+	0.724392	0.086420	0.048148
af-	0.724392	0.080247	0.051852
ag+	0.213561	0.080247	0.051852
ag-	0.213561	0.086420	0.048148

Partial Contributions to Inertia for the Column Points

	Dim1	Dim2
ta+	0.046421	0.121707
ta++	0.134585	0.000102
ta-	0.125978	0.095866
Poids+	0.016737	0.150621
Poids++	0.066040	0.076089
Poids-	0.140104	0.087222
vel+	0.037369	0.101171
vel++	0.091804	0.019964
vel-	0.013120	0.175174
int+	0.022742	0.017001
int++	0.008634	0.020324
int-	0.012492	0.083913
af+	0.107915	0.016005

[107] On remarque qu'elles sont de signes opposés à celles de Minitab ce qui n'a aucune importance sur l'interprétation finale. Ceci est du au choix des vecteurs propres.

af-	0.116216	0.017236
ag+	0.031030	0.009128
ag-	0.028813	0.008476

Squared Cosines for the Column Points

	Dim1	Dim2
ta+	0.164625	0.344803
ta++	0.875032	0.000529
ta-	0.491442	0.298755
Poids+	0.100447	0.722139
Poids++	0.234204	0.215564
Poids-	0.575313	0.286124
vel+	0.153447	0.331879
vel++	0.397921	0.069130
vel-	0.060213	0.642245
int+	0.126739	0.075690
int++	0.032077	0.060321
int-	0.051298	0.275268
af+	0.647656	0.076736
af-	0.647656	0.076736
ag+	0.172924	0.040637
ag-	0.172924	0.040637

Légende :

Minitab	**SAS**	**Interprétation**
Qual	Quality	Qualité globale de représentation (somme des cos carrés)
Mass	Mass	% d'individus ayant choisi la modalité ($f_{i,.}$)[108]
Inert	Inertia	Part d'inertie totale expliquée par la modalité. ($f_{i,.} \frac{\sum_j coord(\text{mod i sur axe j})^2}{\text{inertie totale}}$) Voir note[109]
Coord	Coordinates	Coordonnée de la modalité sur l'axe.
Corr	Squared cosine	Qualité de représentation de l'individu sur l'axe. (Cosinus carré de l'angle entre l'individu et l'axe). Plus ce nombre est proche de 1 meilleure est cette qualité.
Contr	Partial contribution to inertia	Part d'inertie de l'axe expliquée par la modalité. C'est une quantité essentielle pour l'interprétation des axes. (= $\frac{f_{i,.}}{vp_j}$ coord(modalité i sur axe j)²) vp_j=valeur propre axe j.

[108] Permet de détecter, entre autres, les modalités rares. Ces dernières peuvent jouer un rôle excessif dans la construction des axes. Voir note suivante.

[109] L'inertie engendrée par une modalité est d'autant plus grande que cette modalité est rare (Mass petit). Il faut donc être très méfiant envers les modalités rares qui risquent de jouer un rôle excessif dans la construction des axes. Une solution consiste a donner le statut de variable supplémentaires aux variables possédant des modalités trop rares ; il est aussi possible d'effectuer des recodages.

Nous allons repérer les modalités ayant une forte contribution à l'axe 1 en distinguant le signe des coordonnées sur l'axe. Il faut aussi être attentif aux modalités trop rares (cf. note).

Complétez le tableau suivant :

Modalités à forte contribution sur l'axe 1 avec une coordonnée négative.	**Modalités à forte contribution sur l'axe 1 avec une coordonnée positive.**
Ta - (12.6%) P -(14%) Af + (10.8%) ...	

Interprétez l'axe1.

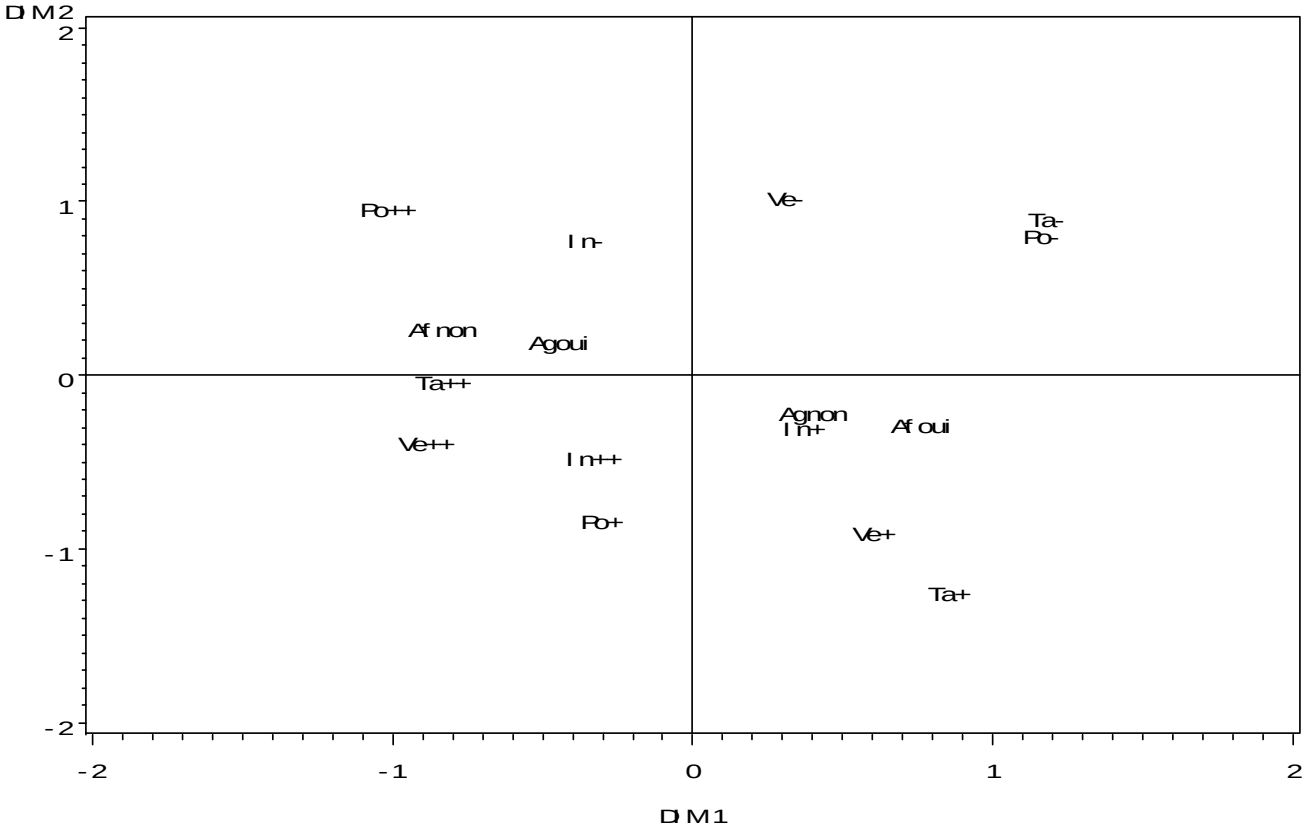
Faites de même pour l'axe 2.

Modalités à forte contribution sur l'axe 2 avec une coordonnée négative.	**Modalités à forte contribution sur l'axe 2 avec une coordonnée positive.**
...	

c) Représentation graphique.

Sous SAS :

```
proc corresp data=pub.chiens mca obs all outc=corr;
tables taille -- agressiv;
run;

data corr;
set corr;
if _type_='VAR';
y=dim2;
x=dim1;
XSYS='2';
Ysys='2';
text=_NAME_;
size=1;
LABEL Y='DIM 2' X='DIM 1';
keep x y text xsys ysys size;
run;

proc gplot data=corr;
symbol1 V=NONE;
 PLOT Y*X=1 / ANNOTATE=corr frame href=0 vref=0;
 run; quit;
```

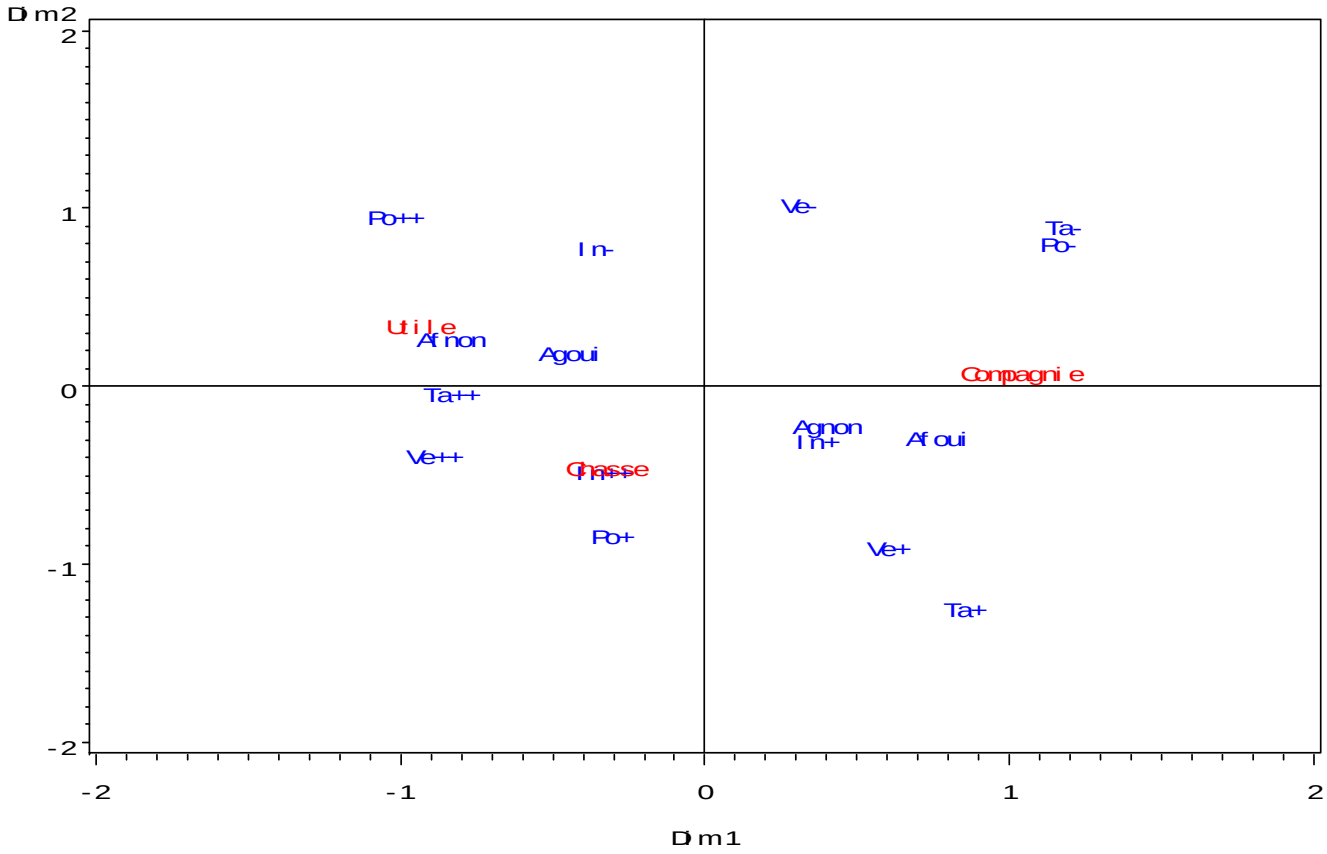


Comme en AFC simple, les points représentants les modalités sont les barycentres des individus qui possèdent cette modalité. Leur proximité doit donc être interprétée avec prudence.

Intervention de variables supplémentaires pour l'interprétation

Sous SAS, vous pouvez demander à faire figurer la variable « Utilité » sur le graphique pour facilité l'interprétation des axes (sans que la variable n'intervienne dans les calculs).

```
proc corresp data=Pub.chiens mca obs all outc=corr;
 tables taille--agressiv utilite;
 supplementary utilite;
 run;

 data corr;
 set corr;
 if _type_='VAR' or _type_='SUPVAR';
 if _type_='VAR' then color='BLUE' ;
 if _type_='SUPVAR' then color='RED';

 y=dim2;
 x=dim1;
 xsys='2';
ysys='2';
 text=_name_;
 size=1;
 label y='Dim 2'
       x='Dim 1';
 keep x y text xsys ysys size color;
run;

proc gplot data=corr;
 symbol1 v=none;
 plot y*x=1/ annotate=corr frame href=0 vref=0;
run;
quit;
```



Vous pouvez maintenant donner une interprétation complète des deux axes.

Visualisation des individus.

Il faudrait maintenant compléter cette étude par l'étude des individus mais SAS n'a pas prévu cela.

Nous allons donc ruser en effectuant une analyse des correspondances simples sur le tableau disjonctif complet.

Le fichier CHIENSI contient les variables indicatrices des variables précédentes :

Nom	cat	ta1	ta2	ta3	po1	po2	po3	ve1	ve2	ve3	In1	In2	In3	Af0	Af1	Ag0	Ag1
BEAUCERON	**Ta-**	**0**	**0**	**1**	**0**	**1**	**0**	**0**	**0**	**1**	**0**	**1**	**0**	**0**	**1**	**0**	**1**
BASSET	**ta+**	**1**	**0**	**0**	**1**	**0**	**0**	**1**	**0**	**0**	**1**	**0**	**0**	**1**	**0**	**0**	**1**
BERGER_ALLEMAND	**ta++**	**0**	**0**	**1**	**0**	**1**	**0**	**0**	**0**	**1**	**0**	**0**	**1**	**0**	**1**	**0**	**1**
BOXER	**p-**	**0**	**1**	**0**	**0**	**1**	**0**	**0**	**1**	**0**	**0**	**1**	**0**	**0**	**1**	**0**	**1**
BULL-DOG	**p+**	**1**	**0**	**0**	**1**	**0**	**0**	**1**	**0**	**0**	**0**	**1**	**0**	**0**	**1**	**1**	**0**
BULL-MASTIFF	**p++**	**0**	**0**	**1**	**0**	**0**	**1**	**1**	**0**	**0**	**0**	**0**	**1**	**1**	**0**	**0**	**1**
CANICHE	**v-**	**1**	**0**	**0**	**1**	**0**	**0**	**0**	**1**	**0**	**0**	**0**	**1**	**0**	**1**	**1**	**0**

Nous allons donc effectuer une analyse des correspondances simples sur ces données ce qui revient à faire une analyse des correspondances multiples.

```
proc corresp data=moi.chiensi outc=corr;
 var ta1 ta2 ta3 po1 po2 po3 ve1 ve2 ve3
       in1 in2 in3 af0 af1 ag0 ag1 ut1 ut2 ut3;
 id nom;
 supplementary ut1 ut2 ut3 ;
 run;
```

Nous obtenons :

Row Coordinates

	Dim1	Dim2
BEAUCERON	-0.31720	-0.41770
BASSET	0.25411	1.10123
BERGER_ALLEMAND	-0.48640	-0.46445
BOXER	0.44736	-0.88178
BULL-DOG	1.01335	0.54988
BULL-MASTIFF	-0.75257	0.54691
CANICHE	0.91230	-0.01619
CHIHUAHUA	0.84080	0.84385
COKER	0.73330	0.07907
COLLEY	-0.11733	-0.52611
DALMATIEN	0.64724	-0.99018
DOBERMANN	-0.87321	-0.31548
DOGUE_ALLEMAND	-1.04702	0.50696
EPAGNEUL_BRETON	0.47804	-1.03693
EPAGNEUL_FRANCAI	-0.14491	-0.51578
FOX-HOUND	-0.87657	0.02524
FOX-TERRIER	0.88162	0.13897
GRAND_BLEU_DE_GA	-0.51734	-0.11340
LABRADOR	0.64724	-0.99018
LEVRIER	-0.67669	-0.08317

MASTIFF	-0.75593	0.88763
PEKINOIS	0.84080	0.84385
POINTER	-0.67334	-0.42389
SAINT-BERNARD	-0.58338	0.59366
SETTER	-0.50414	-0.37714
TECKEL	1.01335	0.54988
TERRE-NEUVE	-0.38350	0.48525

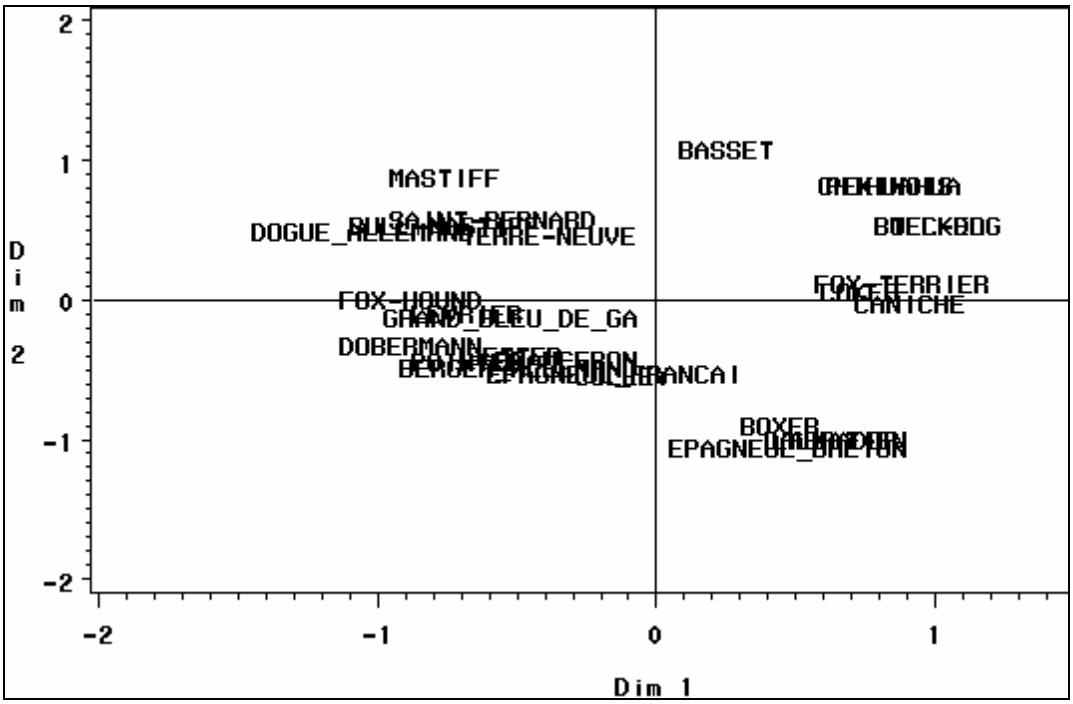
Partial Contributions to Inertia for the Row Points

	Dim1	Dim2
BEAUCERON	0.007738	0.016796
BASSET	0.004966	0.116742
BERGER_ALLEMAND	0.018194	0.020766
BOXER	0.015391	0.074850
BULL-DOG	0.078971	0.029108
BULL-MASTIFF	0.043556	0.028794
CANICHE	0.064006	0.000025
CHIHUAHUA	0.054366	0.068550
COKER	0.041353	0.000602
COLLEY	0.001059	0.026645
DALMATIEN	0.032216	0.094385
DOBERMANN	0.058638	0.009581
DOGUE_ALLEMAND	0.084305	0.024741
EPAGNEUL_BRETON	0.017574	0.103508
EPAGNEUL_FRANCAI	0.001615	0.025610
FOX-HOUND	0.059090	0.000061
FOX-TERRIER	0.059774	0.001859
GRAND_BLEU_DE_GA	0.020582	0.001238
LABRADOR	0.032216	0.094385
LEVRIER	0.035215	0.000666
MASTIFF	0.043945	0.075847
PEKINOIS	0.054366	0.068550
POINTER	0.034866	0.017297
SAINT-BERNARD	0.026173	0.033927
SETTER	0.019545	0.013692
TECKEL	0.078971	0.029108
TERRE-NEUVE	0.011311	0.022668

Nous pouvons faire une représentation graphique :

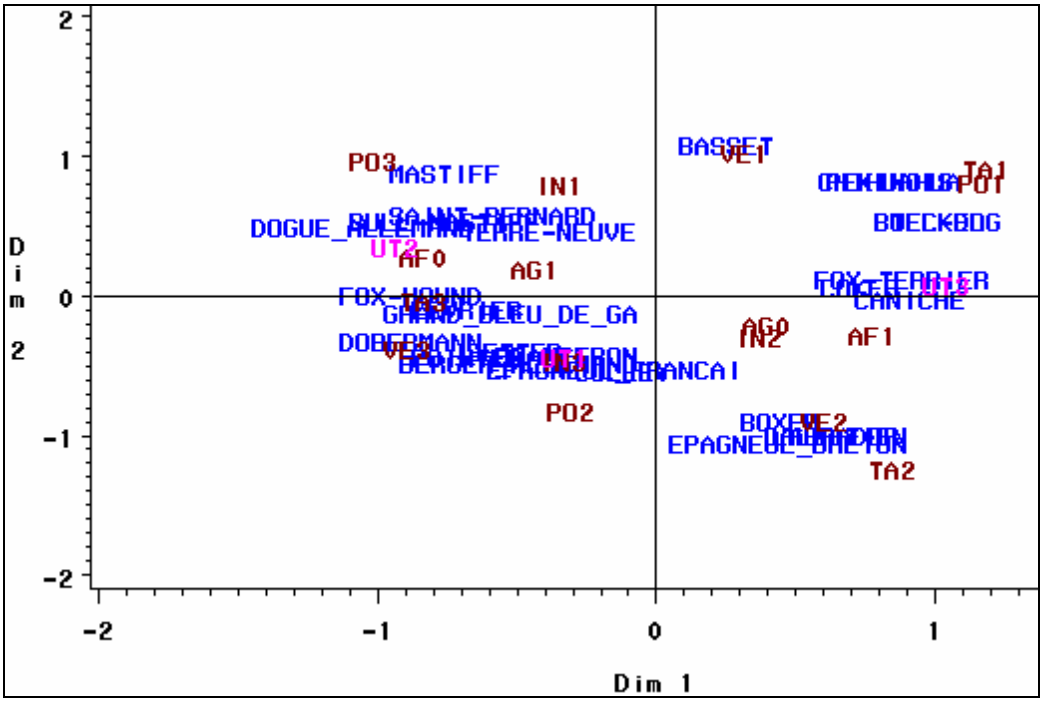
```
PROC CORRESP DATA=MOI.CHIENSI OUTC=CORR;
 VAR TA1 TA2 TA3 PO1 PO2 PO3 VE1 VE2 VE3
       IN1 IN2 IN3 AF0 AF1 AG0 AG1 UT1 UT2
UT3;
 ID NOM;
 SUPPLEMENTARY UT1 UT2 UT3 ;
 RUN;
DATA CORR;
 SET CORR;
 IF _TYPE_='OBS';
 Y=DIM2;
 X=DIM1;
 XSYS='2';
 YSYS='2';
 TEXT=NOM;
 COLOR='BLUE';
 SIZE=1;
 LABEL Y='DIM 2'
       X='DIM 1';
 KEEP X Y TEXT XSYS YSYS SIZE COLOR;
RUN;
```

```
PROC GPLOT DATA=CORR;
 SYMBOL1 V=NONE;
 PLOT Y*X=1/ ANNOTATE=CORR FRAME HREF=0
VREF=0; RUN; QUIT;
```



Pour obtenir individus et variables sur le graphique, il suffit de compléter le programme comme suit :

```
PROC CORRESP DATA=MOI.CHIENSI OUTC=CORR;
 VAR TA1 TA2 TA3 PO1 PO2 PO3 VE1 VE2 VE3
       IN1 IN2 IN3 AF0 AF1 AG0 AG1 UT1 UT2 UT3;
 ID NOM;
 SUPPLEMENTARY UT1 UT2 UT3 ;
 RUN;

DATA CORR;
 SET CORR;
 Y=DIM2;
 X=DIM1;
 XSYS='2';
 YSYS='2';
 TEXT=NOM;
 COLOR='BLUE';
 IF _TYPE_='VAR' THEN COLOR='RED';
 IF _TYPE_='SUPVAR' THEN COLOR='PINK';
 SIZE=1;
 LABEL Y='DIM 2'
       X='DIM 1';
 KEEP X Y TEXT XSYS YSYS SIZE COLOR;
RUN;

PROC GPLOT DATA=CORR;
 SYMBOL1 V=NONE;
 PLOT Y*X=1/ ANNOTATE=CORR FRAME HREF=0 VREF=0;
RUN;
QUIT;
```



Ici comme en AFC, il faut rester prudent quant aux distances entre « chiens » et « catégories ».

Néanmoins, ce graphique permet de compléter l’interprétation précédente.

Complément : Individus supplémentaires

Lorsque nous voulons sortir des individus atypiques des calculs mais les faires figurer sur les graphiques (on parle alors d'individus supplémentaires), il suffit de leur appliquer un poids négatif.

Exemple : Nous allons refaire les calculs en plaçant le BASSET en individu supplémentaire.

Cela donne :

```
DATA CHIENSI2;
 SET MOI.CHIENSI;
 W=1;
 IF NOM='BASSET' THEN W=-1;
RUN;

PROC CORRESP DATA=CHIENSI2 OUTC=CORR;
 WEIGHT W;
 VAR TA1 TA2 TA3 PO1 PO2 PO3 VE1 VE2 VE3
       IN1 IN2 IN3 AF0 AF1 AG0 AG1 UT1 UT2 UT3;
 ID NOM;
 SUPPLEMENTARY UT1 UT2 UT3 ;
 RUN;

DATA CORR;
 SET CORR;
 Y=DIM2;
 X=DIM1;
 XSYS='2';
 YSYS='2';
 TEXT=NOM;
 COLOR='BLUE';
 IF _TYPE_='VAR' THEN COLOR='RED';
 IF _TYPE_='SUPVAR' THEN COLOR='PINK';
 IF _TYPE_='SUPOBS' THEN COLOR='GREEN';
 SIZE=1;
 LABEL Y='DIM 2'
       X='DIM 1';
 KEEP X Y TEXT XSYS YSYS SIZE COLOR;
RUN;

PROC GPLOT DATA=CORR;
 SYMBOL1 V=NONE;
 PLOT Y*X=1/ ANNOTATE=CORR FRAME HREF=0 VREF=0;
RUN;
QUIT;
```

Nous créons un nouveau fichier contenant les données précédentes et une variable W valant 1 pour tous sauf pour le Basset

Cette instruction demande à SAS d'affecter le poids W à chaque observation.

Choix des couleurs : SUPOBS pour l'observation supplémentaire.

L’OUTPUT nous montre que SAS a bien enregistré cela :

Supplementary Row Coordinates

	Dim1	Dim2
BASSET	0.27397	1.06732

Summary Statistics for the Row Points

	Quality	Mass	Inertia
BEAUCERON	0.213291	0.038462	0.025012
BERGER_ALLEMAND	0.269393	0.038462	0.033987
BOXER	0.543157	0.038462	0.039734

.....

P. DISCRIM : L'Analyse discriminante

Nous allons présenter ici une technique statistique très accessible dont les applications sont innombrables. On peut la voir comme une analyse en composante principale particulière. (voir plus loin)

On considère une population P de n individus divisée en k classes P_j à l'aide d'une variable qualitative Y. Sur chaque individu agissent p variables numériques X1, X2,... Xp. Notez bien que l'on connaît le découpage en classe de la population.[110]

On distingue 2 aspects en analyse discriminante :

Un aspect descriptif (géométrique):
On va rechercher les combinaisons linéaires des variables Xi qui permettent de séparer le « mieux possible » les k classes. On peut montrer que cela revient à effectuer une ACP des k centres de gravité avec une métrique particulière. Les variables discriminantes ainsi construites sont donc non corrélées entre elles.
On l'appelle analyse factorielle discriminante en français, et analyse discriminante canonique en anglais. (PROC CANDISC de SAS)[111]

Un aspect aide à la décision (probabiliste): un nouvel individu se présente pour lequel on connaît les valeurs des Xi. Dans quelle classe a-t-il le plus de chance d'appartenir ? C'est l'analyse discriminante bayesienne qui permet de répondre à cette question. (Analyse discriminante pour les logiciels Américains, PROC DISCRIM (SAS) et Stat/Multivariate/Discriminant Analysis pour Minitab)

[110] A ne pas confondre avec les méthodes de classification hiérarchique qui servent à établir un découpage en classe d'une population donnée.
[111] Minitab n'effectue pas cette analyse. Il se contente de l'analyse discriminante bayesienne.

1. L'analyse factorielle discriminante

a) Présentation sommaire

Elle consiste à rechercher une première variable V1, combinaison linéaire des X_i **centrés réduits,** ayant un « pouvoir discriminant » maximum puis, une deuxième variable V_2 non corrélée avec V_1 au « pouvoir discriminant » maximum et ainsi de suite. Géométriquement, effectuer une analyse discriminante canonique revient à trouver les axes sur lesquels la projection du nuage sépare « au mieux » les k groupes. Comme nous le disions précédemment, elle revient à effectuer une ACP des k centres de gravité des classes avec une métrique particulière.

(1) Mesure du pouvoir discriminant

Dans la paragraphe sur l'Analyse de la variance à un facteur (ANOVA), nous avons vu comment mesurer l'influence d'un facteur sur une variable. Il est naturel de l'utiliser pour mesurer le pouvoir discriminant.

La variance totale se décompose de la façon suivante:

$$\frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij}-\bar{x})^2=\frac{1}{n}\sum_{i=1}^{k}n_i(\bar{x}_i-\bar{x})^2+\frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij}-\bar{x}_i)^2$$

Variance totale = Variance inter + Variance intra

(où n_i désigne le nombre d'individus de la sous population Pi, n le nombre total d'individus, $\bar{x}_i$ la moyenne de X dans Pi et $\bar{x}$ la moyenne générale de X)

Plus la liaison entre Y et X est forte, plus la part de la variance inter est importante et plus la variance intra est faible. La variance intra comptabilise la partie de la variation de X non expliquée par Y.

On définit également $$F=\frac{\frac{1}{k-1}\sum_{i=1}^{k}n_i(\bar{x}_i-\bar{x})^2}{\frac{1}{n-k}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij}-\bar{x}_i)^2}$$

Lorsque H0 est vraie (moyennes égales ou pouvoir discriminant néant de X sur Y) F suit une loi de Fisher à (k-1, n-k) degrés de liberté. On rejette H0 lorsque Fobs est supérieur au fractile d'ordre 1-α de la loi de Fisher correspondante. La variable X a alors un pouvoir discriminant significatif.

On mesure le pouvoir discriminant d'une variable V1 sur Y en utilisant le **rapport de corrélation** :

$\eta^2(V_1,Y)$ =Somme des carrés inter classe/ Somme des carrés totale

Plus ce nombre est grand (proche de 1) plus la variable X est discriminante, plus il est faible (proche de 0) moins la variable X n’est liée à Y. On utilisera ce critère pour trouver V1 et V2.

(2) Mise en pratique (sous SAS)[112]

La société *FRED&NUCCI frères* spécialiste vinicole réputé de la région bordelaise effectue une étude pour relier la qualité de leur vins en fonction des caractéristiques météorologiques.

Les données sont dans les fichiers Minitab et SAS BORDEAUX.MTW, BORDEAUX (répertoire Public)
X1 : Somme des t° moyennes journalières (en °C)
X2 : Durée d'insolation (en h)
X3 : Nombre de jour de grande chaleur
X4 : Hauteur des pluies.

Y : Qualité du vin : 1 Bon, 2 Moyen 3 : Médiocre.

Annee	X1	X2	X3	X4	Qual
1924	3064	1201	10	361	2
1925	3000	1053	11	338	3
1926	3155	1133	19	393	2
1927	3085	970	4	467	3
1928	3245	1258	36	294	1
1929	3267	1386	35	225	1
1930	3080	966	13	417	3
1931	2974	1189	12	488	3
1932	3038	1103	14	677	3
1933	3318	1310	29	427	2
1934	3317	1362	25	326	1
1935	3182	1171	28	326	3
1936	2998	1102	9	349	3
1937	3221	1424	21	382	1
1938	3019	1230	16	275	2
1939	3022	1285	9	303	2
1940	3094	1329	11	339	2
1941	3009	1210	15	536	3
1942	3227	1331	21	414	2
1943	3308	1366	24	282	1
1944	3212	1289	17	302	2
1945	3361	1444	25	253	1
1946	3061	1175	12	261	2
1947	3478	1317	42	259	1
1948	3126	1248	11	315	2
1949	3458	1508	43	286	1
1950	3252	1361	26	346	2
1951	3052	1186	14	443	3
1952	3270	1399	24	306	1
1953	3198	1259	20	367	1
1954	2904	1164	6	311	3
1955	3247	1277	19	375	1
1956	3083	1195	5	441	3
1957	3043	1208	14	371	3

Nous supposerons la normalité vérifiée.
Calculons le rapport de corrélation des variables du fichier BORDEAUX.MTW en tapant le programme suivant :

```
proc candisc data=moi.bordeaux anova;
 var x1 x2 x3 x4;
 class qual;
run;
```

[112] Minitab ne sait pas effectuer directement l'analyse discriminante canonique.

SAS calcule ces rapports de corrélation et teste leur signification (sous hypothèse de normalité)

Canonical Discriminant Analysis

Univariate Test Statistics

F Statistics, Num DF= 2 Den DF= 31

Variable	Total STD	Pooled STD	Between STD	R-Squared	RSQ/ (1-RSQ)	F	Pr > F
X1	141.1843	87.5697	136.1337	**0.638605**	1.7671	27.3893	0.0001
X2	126.6230	80.7610	120.0935	**0.617857**	1.6168	25.0607	0.0001
X3	10.0166	7.3273	8.5231	**0.497312**	0.9893	15.3342	0.0001
X4	91.4016	75.8817	65.4816	**0.352537**	0.5445	8.4396	0.0012

Average R-Squared: Unweighted = 0.5265777 Weighted by Variance = 0.5769962

SAS donne les rapports de corrélation (`R-Squared`), le F correspondant et le P du test précédent. Si H0 est vraie, le pouvoir discriminant n'est pas significatif.

Nous voyons que toutes les variables ont un pouvoir discriminant significatif sur la qualité du vin Y. X1 a le plus fort.

(3) Nombre de variables discriminantes à prendre en compte

Nous allons maintenant chercher de nouvelles variables, combinaison linéaires des précédentes, non corrélées entre elles ayant un pouvoir discriminant maximum. Comme nous avons 3 classes, il n'y aura que deux variables :

Total-Sample Standardized Canonical Coefficients

	CAN1	CAN2	
X1	1.209391427	-0.006529859	X1
X2	0.857727422	-0.674810955	X2
X3	-0.270993045	1.278475787	X3
X4	-0.536131215	0.564364371	X4

On a donc CAN1=1.209*X1*+0.858*X2*-0.271*X3*-0.536*X4*
(Xi* désigne Xi centré réduit.)

Sous hypothèse de normalité et d'égalité des matrices de variances covariance entre les différents groupes, on peut effectuer un test pour nous aider.

Notons η_i=Racine($\eta^2(V_i,Y)$)

On teste **H0 :la nullité des q derniers rapport de corrélation** en utilisant la statistique $\Lambda_q = \prod_{i=p-q}^{p-1}(1-\eta_i^2)$. Lorsque ce nombre est trop petit, on rejette H0.

Λ_{p-1} s'appelle le lambda de Wilks. Il mesure le pouvoir discriminant global des p variables X_i

Multivariate Statistics and F Approximations

S=2 M=0.5 N=13

Statistic	Value	F	Num DF	Den DF	Pr > F
Wilks' Lambda	**0.20526297**	**8.4505**	**8**	**56**	**0.0001**
Pillai's Trace	0.88800132	5.7896	8	58	0.0001
Hotelling-Lawley Trace	3.41743451	11.5338	8	54	0.0001
Roy's Greatest Root	3.27886049	23.7717	4	29	0.0001

NOTE: F Statistic for Roy's Greatest Root is an upper bound.
NOTE: F Statistic for Wilks' Lambda is exact.

Nous voyons que globalement les variables ont un pouvoir discriminant significatif. (sous réserve que la multinormalité et l'égalité des matrices de variances covariances soient vérifiées cf. POOL=TEST de PROC DISCRIM)

Canonical Discriminant Analysis

	Canonical Correlation	Adjusted Canonical Correlation	Approx Standard Error	**Squared Canonical Correlation**
1	0.875382	0.861944	0.040683	**0.766293**
2	0.348867	0.280587	0.152891	**0.121708**

Test of H0: The canonical correlations in the current row and all that follow are zero

	Likelihood Ratio	Approx F	Num DF	Den DF	Pr > F
1	0.20526297	8.4505	8	56	**0.0001**
2	0.87829160	1.3395	3	29	**0.2808**

SAS donne $\boldsymbol{\eta^2_1=0.766}$. On voit que le pouvoir discriminant de V_1 est meilleur que celui de X_1 (et que les autres X_i)

Les deux tests suivants montrent que nous ne devons pas aller au delà de V_1 dans les composantes canoniques. Nous ne retiendrons donc qu'une variable discriminante.

(4) Interprétation de la variable discriminante obtenue

Nous pouvons calculer cette nouvelle variable :

```
proc candisc data=moi.bordeaux all out=essai;
 var x1 x2 x3 x4;
 class qual;
run;
```

Le fichier ESSAI contient les nouvelles variables CAN1 et CAN2.

SAS calcule aussi les corrélations intra et inter avec cette nouvelle variable.

Total Canonical Structure **(Corrélations CAN1,Xj)**

	CAN1	CAN2	
X1	0.900589	0.374779	X1
X2	0.896744	-0.116190	X2
X3	0.770513	0.590030	X3
X4	-0.662815	0.361294	X4

Between Canonical Structure **Corrélations interclasses**

	CAN1	CAN2	
X1	0.986524	0.163614	X1
X2	0.998669	-0.051569	X2
X3	0.956452	0.291891	X3
X4	-0.977208	0.212284	X4

Pooled Within Canonical Structure **Corrélations intra classes.**

	CAN1	CAN2	
X1	0.724221	0.584256	X1
X2	0.701280	-0.176148	X2
X3	0.525372	0.779910	X3
X4	-0.398218	0.420797	X4

Nous allons trier le fichier essai par rapport à CAN1 puis, nous l'affichons :

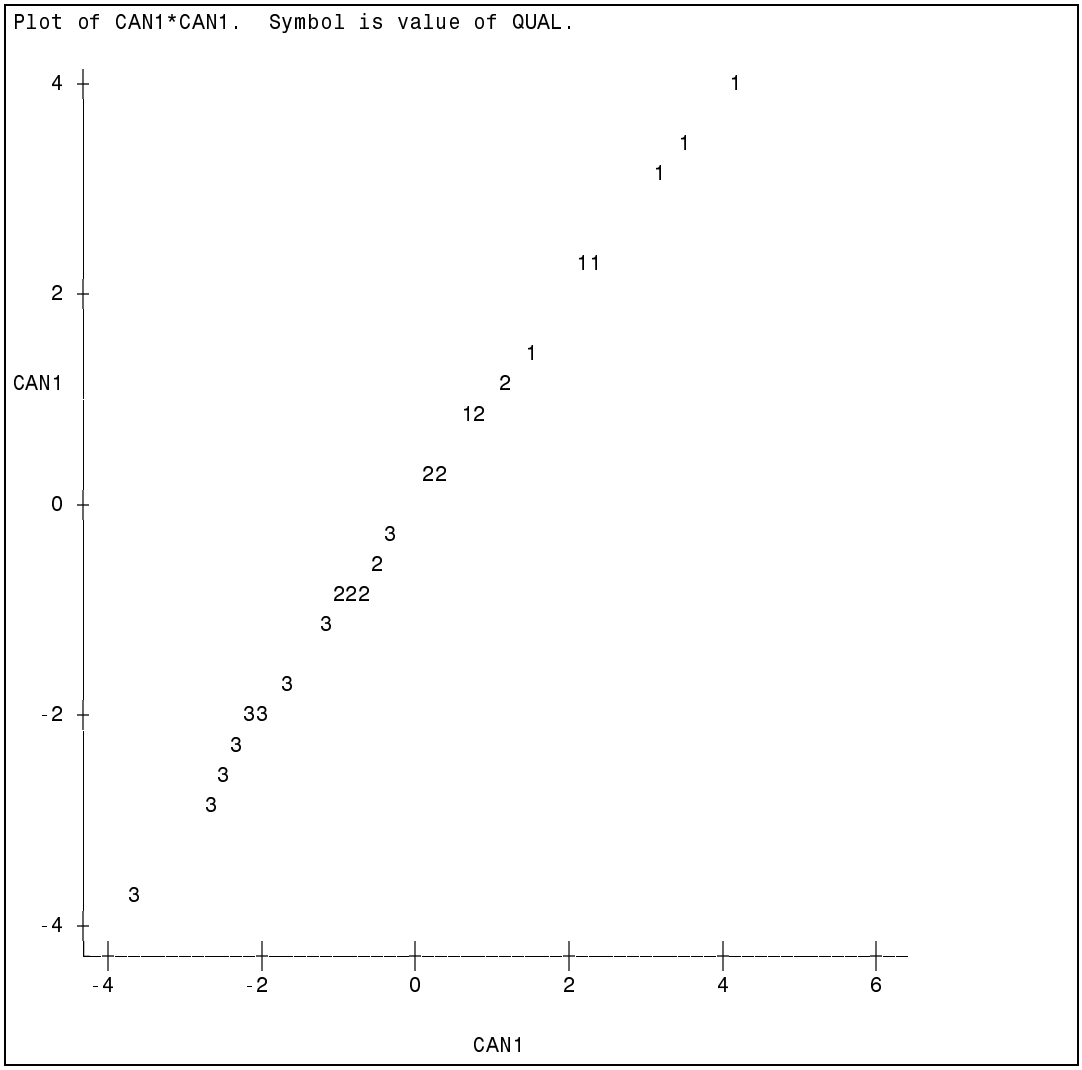
OBS	ANNEE	QUAL	CAN1
1	1932	3	-3.73088
2	1930	3	-2.74699
3	1927	3	-2.72686
4	1931	3	-2.53383
5	1941	3	-2.45448
6	1925	3	-2.32546
7	1954	3	-2.10225
8	1936	3	-2.02108
9	1951	3	-1.67615
10	1957	3	-1.18190
11	1956	3	-1.09442
12	1926	2	-0.99486
13	1924	2	-0.88255
14	1938	2	-0.72946
15	1946	2	-0.55191
16	1935	3	-0.35666
17	1939	2	-0.30606
18	1948	2	0.20968
19	1940	2	0.34347
20	1953	1	0.35244
21	1928	1	0.74360
22	1942	2	0.78584
23	1955	1	0.87424
24	1933	2	1.13041
25	1944	2	1.13802
26	1950	2	1.46680
27	1937	1	1.55211
28	1952	1	2.16713
29	1934	1	2.17473
30	1929	1	2.23089
31	1943	1	2.40988
32	1947	1	3.18211
33	1945	1	3.53529
34	1949	1	4.11917

Nous voyons que CAN1 peut être vue comme une note mesurant la mauvaise qualité du vin. (Plus CAN1 est petite, meilleur est le vin !)

Les vins dont la « note » est inférieure à -1 sont mauvais, ceux dont la note est supérieure à 1.5 sont bons. Cette règle ne donne pas de trop mauvais résultats sur cet exemple.

Le programme suivant, nous permet de visualiser la variable CAN1

```
PROC PLOT DATA=ESSAI;
PLOT CAN1*CAN1=QUAL;
RUN;
```



Utilisation : Exercice

Que prévoyez-vous pour les vins suivants ?

	Température	Insolation	Chaleur	Pluie	Can1
1958	3000	1050	10	300	
1959	3200	1300	35	225	
1960	3224	1211	18	301	

Vous calculerez la valeur de CAN1 dans les 3 cas.

Attention la formule donnant CAN1 utilise les variables centrées réduites. PROC MEANS peut vous donner les moyennes et écart types des Xi) :

Variable	Label	N	Mean	Std Dev
X1	X1	34	3157.88	141.1843336
X2	X2	34	1247.32	126.6229719
X3	X3	34	18.8235294	10.0165638
X4	X4	34	360.4411765	91.4016084

Remarque : Pour affecter de nouveaux individus dans des classes, il est préférable d'utiliser l'analyse bayesienne (cf. paragraphe suivant) beaucoup plus précise.

2. L'analyse discriminante Bayesienne

a) Principe

Nous allons calculer les probabilités d'appartenance aux différentes classes de chaque individu :
$p_j(x)=Prob(Y=j/X_1=x_1,...,X_p=x_p)$. Nous pourrons ensuite affecter l'observation $x=(x_1,x_2,...,x_p)$ à la classe la plus probable.

Hypothèses [113]: On suppose que le vecteur $X=(X_1,X_2,...,X_p)$ suit une loi multinormale $N_p(\mu_j,\Sigma_j)$ pour chaque classe P_j.
Nous nous limiterons au cas linéaire, c'est à dire que nous supposerons de plus que les matrices de variances Σ_j sont égales[114] à Σ.

Sous ces hypothèses, on a $p_j(x)=\frac{e^{f_j(x)}}{\sum_{i=1}^{p} e^{f_i(x)}}$ avec $f_i(x)=-\frac{1}{2}\mu_i\Sigma^{-1}\,{}^t\mu_i+\mu_i\Sigma^{-1}\,{}^t x$

Les fonctions $f_i(x)$ s'appellent les fonctions discriminantes. En pratique, pour estimer les $\hat{f}_i(x)$, on estime Σ par la matrice S de variances covariances intra-classe et μ_i par le vecteur des moyennes des observations dans la classe i.

On trouvera les formules de calcul pour les estimations dans *Saporta*.

On peut alors calculer la probabilité estimée : $\hat{p}_j(x)=\frac{e^{\hat{f}_j(x)}}{\sum_{i=1}^{p} e^{\hat{f}_i(x)}}$

b) Mise en pratique

Reprenons l'exemple de la société *FRED&NUCCI frères* spécialiste vinicole réputé.

---

[113] Nous allons nous limiter ici au cas gaussien. Il est possible d'effectuer une analyse discriminante non paramétrique. Cf. Saporta p419 et PROC DISCRIM de SAS.

[114] Si ce n'est pas le cas, il faut envisager la règle quadratique en principe. Nous allons présenter plus loin un test qui peut aider à effectuer le choix.

(1) Sous SAS (PROC DISCRIM)

La procédure Discrim est très complète et peut mener à bien l'analyse discriminante Bayesienne.

(a) Analyse élémentaire

Tapons le programme suivant :

```
proc discrim data=moi.bordeaux ;
var x1 x2 x3 x4;                 Variables concernées par l'étude
 class qual;                     Variable distinguant les groupes.
run;
```

Nous supposerons l'égalité des matrices de variance-covariance vérifiée. Nous donnerons des éléments de réponse pour cette vérification dans le paragraphe suivant.

SAS donne dans la fenêtre OUTPUT quelques statistiques de base puis les fonctions discriminantes estimées
:

```
Discriminant Analysis     Linear Discriminant Function

                                -1 _                                  -1 _
     Constant = -.5 X' COV   X      Coefficient Vector = COV   X
                    j        j                                      j

                                QUAL
                          1              2              3     Label

   CONSTANT          -1350          -1284          -1212
   X1              0.81768        0.80079        0.78169     X1
   X2              0.15409        0.14489        0.12590     X2
   X3             -7.00975       -7.05649       -6.90255     X3
   X4             -0.04629       -0.03955       -0.02196     X4
```

On a donc entre autres :

$\hat{f}_1(x) = -1350+0.81768*X1+0.15409*X2-7.00975*X3-0.04629*X4.$

Exprimez les autres fonctions discriminantes.

Nous pouvons nous servir de cela pour calculer la probabilité que le bordeaux 1958 (x1=3000 x2=1050 x3=10 x4=300) soit bon.

$$\hat{p}_1(x) = \frac{\exp(\hat{f}_1(x))}{\exp(\hat{f}_1(x)) + \exp(\hat{f}_2(x)) + \exp(\hat{f}_3(x))} = 0.0001 \quad (x=(3000,1050,10,300))$$

Il y a donc 1 chances sur 10000 pour que le bordeaux 58 soit bon !

Vérifiez ce calcul et calculez la probabilité que le Bordeaux 58 soit moyen, puis médiocre. Dans quelle catégorie va-t-on le mettre ?

Nous pouvons demander à SAS d'effectuer les calculs précédents (quand même !).

Il faut créer un fichier de données SAS contenant l'observation à classer et spécifier grâce à l'option TESTDATA= le nom du fichier. L'option TESTLIST demande à SAS d'afficher dans l'OUTPUT les résultats de classement pour ces nouvelles observations.

```
DATA WORK.ACLASSER;
 INPUT ANNEE X1 X2 X3 X4;
 CARDS;
 1958 3000 1100 20 300
 ;
RUN;

PROC DISCRIM DATA=MOI.BORDEAUX
TESTDATA=WORK.ACLASSER TESTLIST;
 CLASS QUAL;
 VAR X1 X2 X3 X4;
RUN;
```

Vérifiez vos calculs avec ceux de SAS. Que dire des vins suivants ?

	Température	Insolation	Chaleur	Pluie
1959	3200	1300	35	225
1960	3224	1211	18	301

Vous venez de voir une utilisation essentielle de l'analyse discriminante : la prévision pour de nouvelles observations.

(b) Mesures d'efficacité du classement : méthodes de resubstitution et de validation croisée

Nous venons de voir que nous pouvions classer une observation externe grâce au calcul des fonctions discriminantes.

Nous allons nous servir des règles établies pour reclasser les observations du fichier original et vérifier qu'elles sont bien classées la où elles le devraient. Ceci est un critère nous permettant de mesurer la qualité de l'analyse. Plus le nombre de mal classés sera faible meilleure sera cette qualité.

Méthode de resubstitution

SAS fournit dans la fenêtre OUTPUT le résultat de ce reclassement :

Resubstitution Summary using Linear Discriminant Function

Number of Observations and Percent Classified into QUAL:

From QUAL	1	2	3	Total
1	9	2	0	11
	81.82	18.18	0.00	100.00
2	**2**	8	1	11
	18.18	72.73	9.09	100.00
3	0	2	10	12
	0.00	16.67	83.33	100.00
Total	11	12	11	34
Percent	32.35	35.29	32.35	100.00
Priors	0.3333	0.3333	0.3333	

Error Count Estimates for QUAL:

	1	2	3	Total
Rate	0.1818	0.2727	0.1667	**0.2071**
Priors	0.3333	0.3333	0.3333	

SAS nous apprend ici que 2 vins de qualité 2 ont été classés en qualité 1, ce qui est une erreur. Par contre 10 vins de qualité 3 ont bien été classés en qualité 3. Globalement, le taux d'erreur apparent est de 20.7%.

Pour obtenir le classement des observations, il suffit d'ajouter l'option LIST et LISTERR pour obtenir uniquement les mal classées (*misclassified observation*) par la méthode de resubstitution :

```
PROC DISCRIM DATA=MOI.BORDEAUX LIST ;
VAR X1 X2 X3 X4;
 CLASS QUAL;
RUN;
```

On obtient alors :

Discriminant Analysis Classification Results for Calibration Data: MOI.BORDEAUX

Resubstitution Results using Linear Discriminant Function

Obs	From	Classified	Posterior Probability of Membership in QUAL:

	QUAL	into QUAL		1	2	3
1	2	2		0.0069	0.6679	0.3252
2	3	3		0.0000	0.0447	0.9553
3	2	3	*	0.0098	0.3108	0.6794
4	3	3		0.0000	0.0147	0.9853
5	1	1		0.6434	0.3279	0.0287
6	1	1		0.9334	0.0665	0.0001
7	3	3		0.0000	0.0075	0.9925
8	3	3		0.0000	0.0226	0.9774
9	3	3		0.0000	0.0004	0.9996
10	2	1	*	0.7564	0.2368	0.0068
11	1	1		0.8924	0.1074	0.0001
12	3	2	*	0.0866	0.5131	0.4003
13	3	3		0.0002	0.1134	0.8864
14	1	1		0.6222	0.3768	0.0010
15	2	2		0.0108	0.7231	0.2661
16	2	2		0.0143	0.9228	0.0629
17	2	2		0.0622	0.9202	0.0176
18	3	3		0.0000	0.0179	0.9821
19	2	2		0.3791	0.6056	0.0153
20	1	1		0.9048	0.0952	0.0000
21	2	2		0.3669	0.6295	0.0036
22	1	1		0.9838	0.0162	0.0000
23	2	2		0.0137	0.8233	0.1630
24	1	1		0.9966	0.0034	0.0000
25	2	2		0.0590	0.9120	0.0289
26	1	1		0.9990	0.0010	0.0000
27	2	1	*	0.7143	0.2841	0.0016
28	3	3		0.0007	0.1355	0.8638
29	1	1		0.8521	0.1478	0.0001
30	1	2	*	0.1957	0.7553	0.0490
31	3	3		0.0001	0.1760	0.8239
32	1	2	*	0.3833	0.6059	0.0108
33	3	2	*	0.0036	0.5799	0.4165
34	3	3		0.0036	0.4143	0.5821

* Misclassified observation

Nous voyons que les vins 10 et 27 ont été reclassés en qualité 1 alors qu'ils étaient de qualité 2. SAS donne également le calcul des probabilités pour chaque vin.

Au total, 20.7% des observations ont été mal classées par la méthode de resubstitution.

Cela dit, cette méthode de calcul de l'erreur possède le grave défaut suivant : elle se sert de mêmes valeurs pour construire le modèle et vérifier sa validité. L'erreur est donc sous évaluée.

Validation croisée

Pour éviter ce piège dans la mesure de l'erreur, nous utilisons la méthode dite de **validation croisée.**

En activant l'option (CROSSVALIDATE), SAS va classer chaque observation en effectuant au préalable une analyse discriminante sur le fichier des n-1 observations restantes. Ainsi, l'observation que l'on classe n'est jamais dans les données servant à construire les fonctions discriminantes. Le temps de calcul est évidemment plus long mais on évite le problème cité plus haut.

Nous pouvons ajouter l'option CROSSLIST ou CROSSLISTERR pour lister les observations mal classées.

```
PROC DISCRIM DATA=MOI.BORDEAUX CROSSVALIDATE CROSSLISTERR;
VAR X1 X2 X3 X4;
 CLASS QUAL;
RUN;
```

Cross-validation Summary using Linear Discriminant Function

Number of Observations and Percent Classified into QUAL:

From QUAL	1	2	3	Total
1	7	4	0	11
	63.64	36.36	0.00	100.00
2	2	8	1	11
	18.18	72.73	9.09	100.00
3	0	2	10	12
	0.00	16.67	83.33	100.00
Total	9	14	11	34
Percent	26.47	41.18	32.35	100.00
Priors	0.3333	0.3333	0.3333	

Error Count Estimates for QUAL:

	1	2	3	Total
Rate	0.3636	0.2727	0.1667	**0.2677**
Priors	0.3333	0.3333	0.3333	

Nous voyons que le taux d'erreur apparent a augmenté ce qui est logique. Ce taux est certainement plus proche de la réalité.

Cross-validation Results using Linear Discriminant Function

Obs	From QUAL	Classified into QUAL		Posterior Probability of Membership in QUAL: 1	2	3
3	2	3	*	0.0081	0.1822	0.8096
5	1	2	*	0.2537	0.6212	0.1251
10	2	1	*	0.9086	0.0869	0.0045
12	3	2	*	0.1589	0.7044	0.1367
14	1	2	*	0.4726	0.5256	0.0018
27	2	1	*	0.7873	0.2114	0.0014
30	1	2	*	0.1255	0.8323	0.0422
32	1	2	*	0.2728	0.7140	0.0132
33	3	2	*	0.0061	0.7874	0.2065

* Misclassified observation

Voici la liste des observations mal classées selon la méthode de validation croisée.

Les résultats précédents supposent l'hypothèse d'égalité des matrices de variances covariances (fonctions discriminantes linéaires).

Nous pouvons tester cette hypothèse grâce à une option de SAS.

(c) Test d'égalité des matrices Σ_j

SAS propose un test[115] pour valider l'hypothèse d'égalité des matrices Σ_j. Il est activé grâce à l'option POOL=TEST.

Entrons simplement les instructions suivantes :

```
PROC DISCRIM DATA=MOI.BORDEAUX
 POOL=TEST;

 VAR X1 X2 X3 X4;
 CLASS QUAL;
RUN;
```

Cette option demande à SAS de tester l'égalité des matrice de variances covariances entre les différents groupes.
Variables concernées par l'étude
Variable distinguant les groupes.

SAS précise dans la sortie, la statistique utilisée et le résultat du test.

```
          Test of Homogeneity of Within Covariance Matrices
   Notation: K = Number of Groups
             P = Number of Variables
             N = Total Number of Observations - Number of Groups
          N(i) = Number of Observations in the i'th Group - 1

                        __                              N(i)/2
                       ||  |Within SS Matrix(i)|
             V    = ----------------------------------
                                                 N/2
                           |Pooled SS Matrix|

                          _                          _      2
                         |       1        1   |  2P + 3P - 1
             RHO  = 1.0 - | SUM -----  -  ---  | -------------
                         |_     N(i)      N  _|  6(P+1)(K-1)
             DF   = .5(K-1)P(P+1)

                                    _                      _
                                   |    PN/2                |
                                   |   N        V           |
Under null hypothesis:  -2 RHO ln  | ------------------     |
                                   |    __       PN(i)/2    |
                                   |_  ||  N(i)             _|
is distributed approximately as chi-square(DF)

Test Chi-Square Value =    24.909800
with     20 DF      Prob > Chi-Sq = 0.2049
Since the chi-square value is not significant at the  0.1 level,
a  pooled  covariance  matrix  will  be  used  in  the  discriminant
function.
Reference: Morrison, D.F. (1976) Multivariate Statistical Methods p252.
```

[115] « Si on rejette l'hypothèse d'égalité doit-on pour autant utiliser les règles quadratiques ?. Cela n'est pas sûr dans tous les cas. Le test précédent n'est pas parfaitement fiable, ensuite l'usage de règles quadratiques implique l'estimation de bien plus de paramètres que la règle linéaire... Lorsque les échantillons sont de petite taille, les fonctions obtenues sont très peu robustes et il vaut mieux utiliser une règle linéaire malgré tout. » Saporta P423

On remarque ici que l'on accepte l'égalité. SAS nous prévient qu'il utilisera la matrice adéquate.[116]

(d) Syntaxe simplifiée de PROC DISCRIM

```
PROC DISCRIM options ;
  CLASS variable ;      Variable servant à définir les classes (Qualité du vin dans l'exemple précédent)
  VAR variable ;        Variables numériques sur lesquelles on effectuera l'analyse discriminante.
  PRIORS probabilités ; Pour spécifier les probabilités d'appartenance aux différentes classes. EQUAL, PROPORTIONAL, 'modalité1'=proba 'modalité2=proba etc.)
  WEIGHT variable ;     Variable contenant le poids de chaque individu.

RUN ;
```

Les principales options étant :

Corrélations

BCORR
: Affiche les corrélations inter classes pour les variables. Un test de signification est également effectué.

PCORR
: Affiche les corrélations intra-classe pour chaque variable. Un test de signification est également effectué.

WCORR
: Affiche les corrélations intra-classe pour chaque variable et pour chaque classe. Un test de signification est également effectué.

Analyse factorielle discriminante

CANONICAL
: Effectue une analyse factorielle discriminante (canonical discriminant analysis en anglais). Nous en reparlerons plus loin.

NCAN=*nombre*
: Nombre de variables canoniques (cf. analyse factorielle discriminanteà à calculer.

Méthode de resubstitution ou de validation croisée.

[116] Si l'on rejette H0, la règle quadratique est utilisée, sinon c'est la règle linéaire.

CROSSLIST
Affiche la classification des observations du fichier en utilisant la méthode de validation croisée.

CROSSLISTERR
Idem mais en affichant uniquement les individus mal classés.

CROSSVALIDATE
Active la validation croisée.

OUTCROSS=fichier de données
Permet de créer un fichier de données SAS contenant le fichier original, les probabilités calculées par l'analyse discriminante et le reclassement de chaque observation par la méthode de validation croisée. Si l'option CANONICAL est activée, SAS inclut également les variables canoniques dans ce fichier.

LIST
Affiche la classification des observations du fichier en utilisant la méthode de resubstitution.

LISTERR
Idem mais en affichant uniquement les individus mal classés.

NOCLASSIFY
Pas de classification.

Homogénéité des matrices de variance-covariance

POOL=NO ou TEST ou YES
Permet d'indiquer à SAS l'égalité des matrices de covariances (YES par défaut), l'inégalité (NO) ou encore d'effectuer un test pour pouvoir statuer.
Rappelons que si ces matrices sont égales, les fonctions discriminantes linéaires sont évaluées, sinon, ce sont les fonctions discriminantes quadratiques.

SLPOOL=p
Spécifie le niveau de signification pour le test précédent (POOL=TEST). Par défaut p=0.1

Classement de nouvelles observations

TESTDATA=*fichier de données SAS*
Lorsque l'on veut prévoir le classement de nouvelles observations, il est possible de les mettre dans un fichier de données SAS et d'indiquer son nom après un TESTDATA=

TESTLIST

Affiche le classement des nouvelles observations avec le probabilité associées.

TESTOUT=fichier de données SAS

Permet de créer un fichier de données SAS contenant le fichier TESTDATA= ainsi que le classement et les probabilités d'appartenance aux différentes classes. Si l'option CANONICAL est activée, SAS affiche les valeurs pour chaque variable canonique calculée.

Divers

DATA= fichier de données SAS servant à effectuer l'analyse discriminante.

OUT= fichier de données SAS

Permet de créer un fichier de données SAS contenant le fichier original, les probabilités calculées par l'analyse discriminante et le reclassement de chaque observation par la méthode de resubstitution. Si l'option CANONICAL est activée, SAS inclut également les variables canoniques dans ce fichier.

ALL

Active toutes les options d'affichage.

NOPRINT

Pas d'affichage.

ANOVA

Effectue une analyse de variance testant la signification des rapports de corrélations (entre les Xi et Y). Ceci permet de mesurer le pouvoir discriminant des variables du fichier. De façon équivalente, ceci permet aussi de voir si les moyennes des Xi différent significativement selon les classes.

MANOVA

Affiche des statistiques testant l'égalité des moyennes μ_i. Ceci permet de mesurer le pouvoir discriminant global des variables du fichier (lambda de Wilk).

METHOD =<u>NORMAL</u> ou NPAR

Spécifie la méthode utilisée : paramétrique (NORMAL) si l'hypothèse de normalité est vérifiée (on aura alors la règle quadratique ou linéaire selon l'inégalité ou l'égalité des matrices de variances covariances). Non paramétrique (NPAR) si cette normalité n'est pas vérifiée.

(2) Exercice

Voici un excellent exemple tiré de *Saporta*.

101 victimes d'infarctus du myocarde (51 décéderont (Prono=0), 50 survivront (prono=1)) sur lesquels ont été mesurées à leur admission 7 variables (fréquence cardiaque, index cardiaque, index systolique, pression diastolique, pression artérielle pulmonaire, pression ventriculaire, résistance pulmonaire)

Les données sont dans le fichier PUB.INFARCTU

On supposera la multinormalité et l'égalité des matrices de variances covariances vérifiées.

Mesurez le pouvoir discriminant de chacune des variables. Effectuez une analyse discriminante canonique sur les données précédentes.

Effectuez une analyse discriminante bayesienne. Que donne la méthode de validation croisée ? (% de bien classés, observations mal classées)
Que dire des malades suivants :

X1	*X2*	*X3*	*X4*	*X5*	*X6*	*X7*
110	*1.5*	*15*	*24*	*30*	*5.5*	*1490*
125	*3.37*	*26.7*	*17*	*29*	*6*	*800*

(3) Compléments

Nous n'avons fait qu'effleurer l'analyse discriminante dans ce qui précède. Nous n'avons pas parlé de l'analyse discriminante pas à pas, des méthodes non paramétriques.

Pour plus d'informations, nous vous renvoyons à la bibliographie, spécialement les ouvrages de M. Tenenhauss et de G. Saporta.

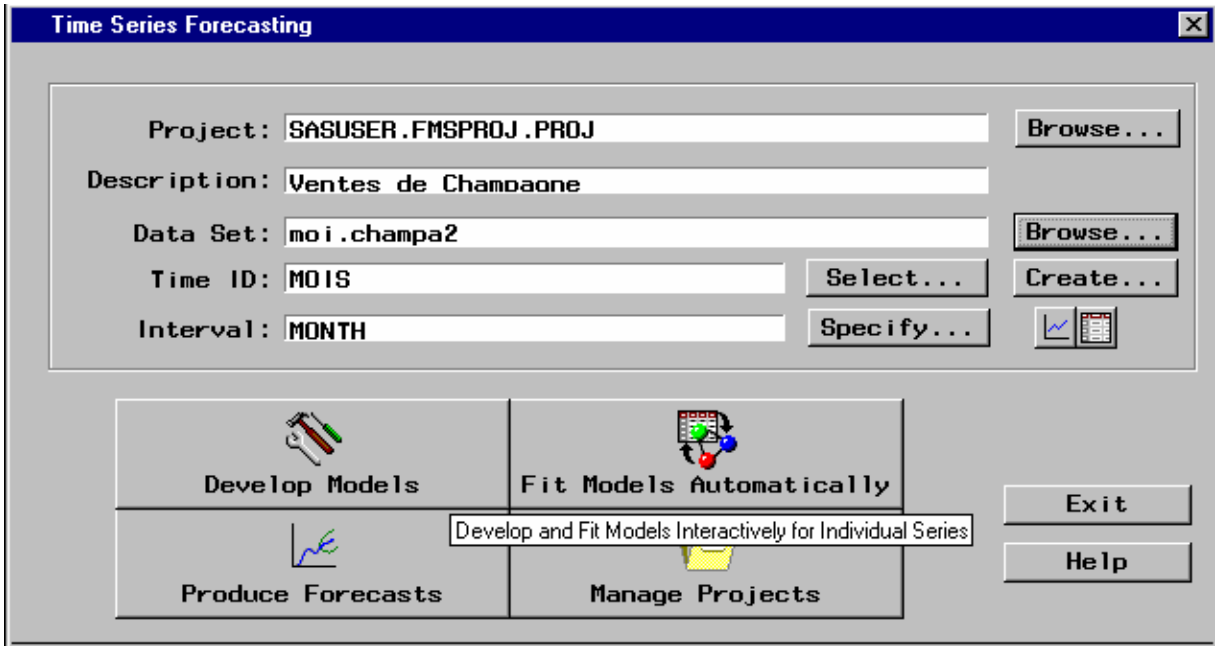
Q. **La commande FORECAST** (Etude de séries chronologiques)

Les heureux possesseurs du module SAS/ETS peuvent étudier les séries chronologiques de façon très approfondies sous SAS. Nous allons, dans ce paragraphe, regarder quelques unes des possibilités de la commande FORECAST qui permet de manipuler les séries chronologiques sans connaître le langage SAS.

Prenons le fichier CHAMPA2 qui contient les ventes mensuelles de champagne (nombre de bouteilles) entre Jan 1962 et Sept 1969.

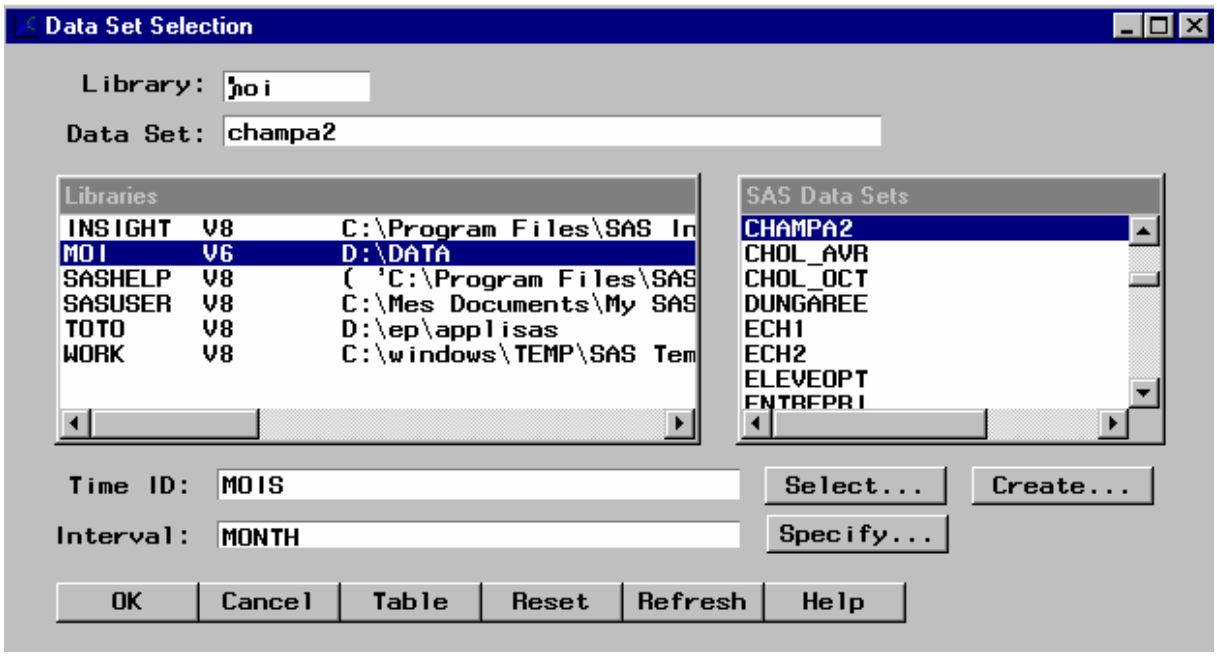
Lancez SAS, Allez dans SOLUTIONS/ANALYSIS/TIMES SERIES FORECASTING SYSTEM (ou tapez Forecast dans la ligne de commande) :

Vous voyez apparaître :



Nous allons compléter les champs les plus importants : Le nom de la table SAS (ici Champa2), de la variable chronologique (MOIS).

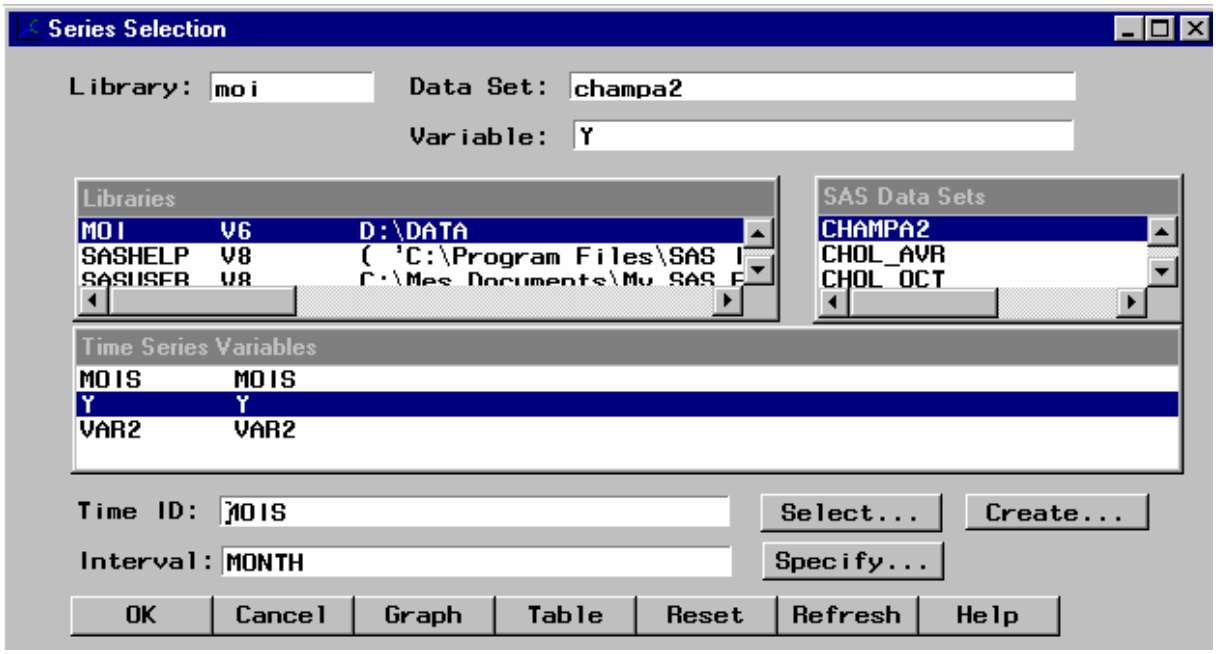
Pour cela, nous cliquons sur BROWSE en face du nom du fichier de données :



Nous entrons le nom du fichier ainsi que le nom de la variable chronologique et nous validons.

Ensuite, nous cliquons sur le bouton 

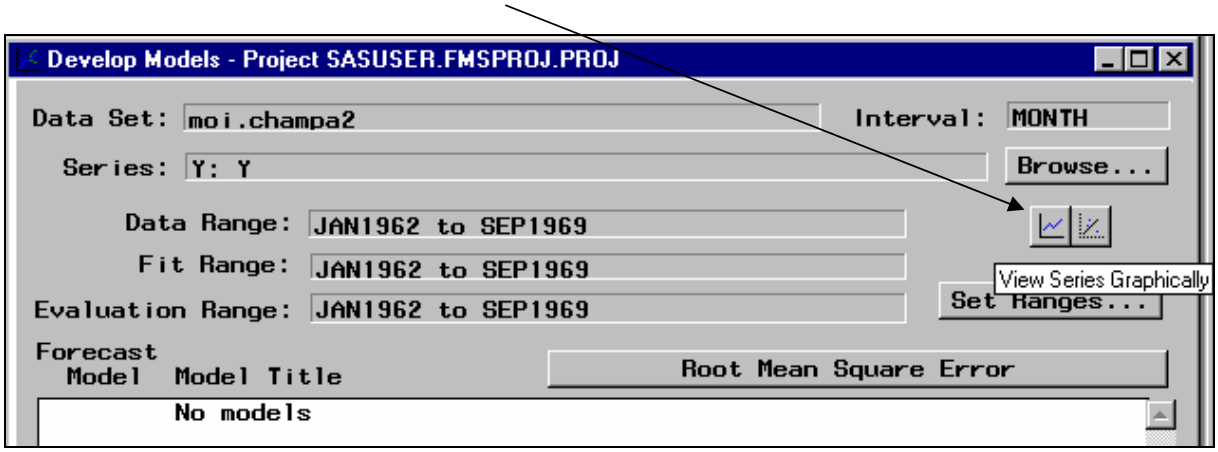
SAS nous demande la nom de la variable à modéliser : (ici Y)



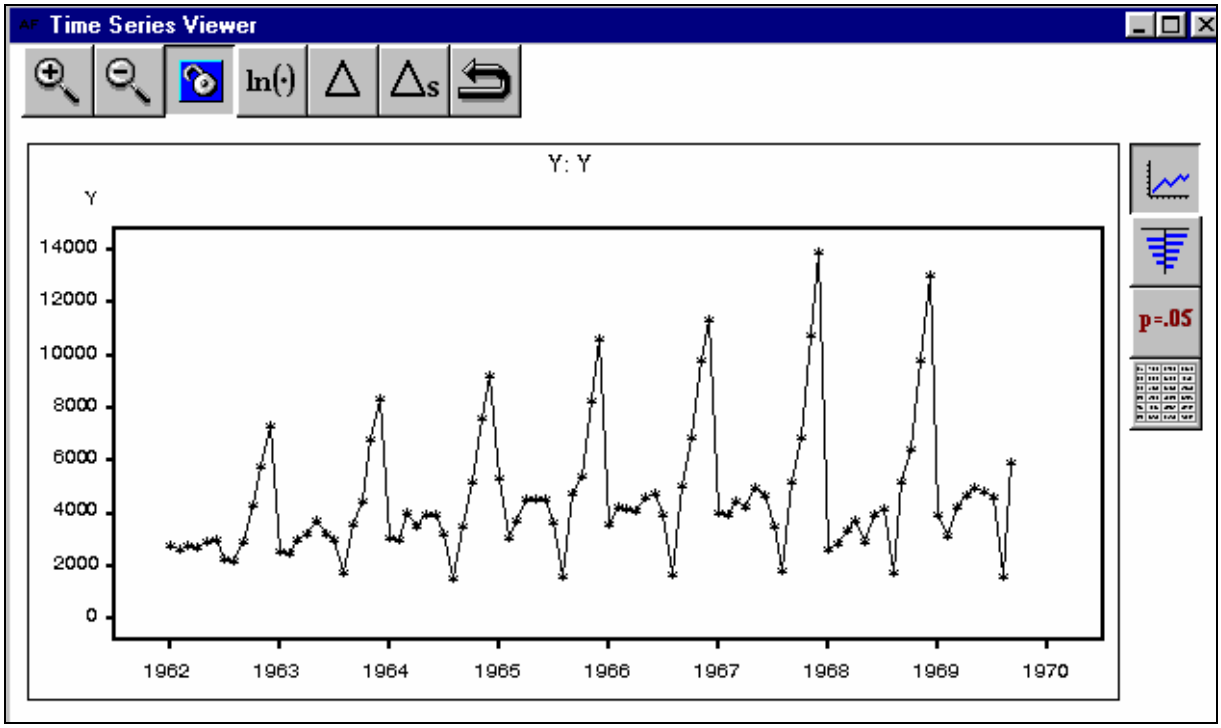
Validez en cliquant sur OK.

1. Visualisation de la série

Cliquez ensuite sur le **bouton** pour que nous puissions visualiser la série.



Vous obtenez



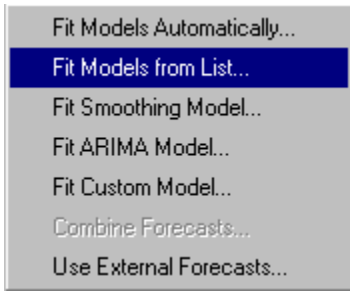
Nous voyons clairement un modèle saisonnier de type multiplicatif. Vous pouvez zoomer sur les zones qui vous interessent en cliquant sur les loupes…

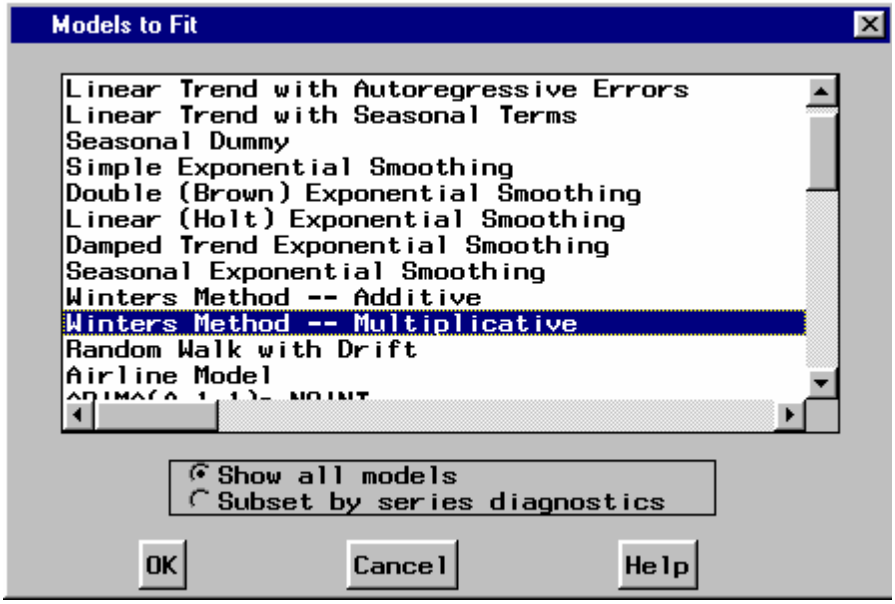
- Cliquez sur le bouton goback (flèche noire) pour revenir à l'écran précédent.

Dans l'optique d'effectuer des prévisions à court terme, nous allons effectuer un lissage de la série précédente.

2. Choix d'un modèle de lissage

Maintenant, nous allons choisir le modèle du lissage que nous souhaitons prendre en cliquant sur le bouton droit de la souris sans la zone des MODELS (en dessous du « No Models ») choisissez « Fit Models from List »



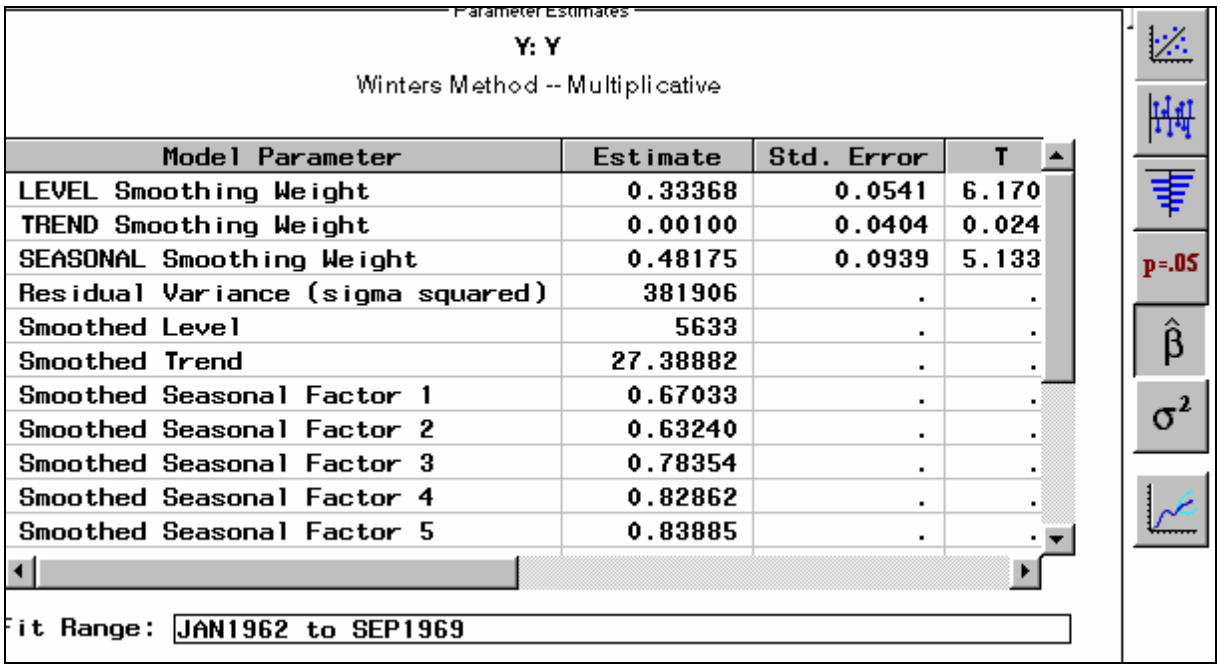


Choisissons le modèle Winters (multiplicatif) et validons.[117]

[117] Notez que SAS possède les lissages exponentiel simple et double (Holt).

3. Estimation des paramètres

SAS effectue alors les calculs. Lorsqu'il a terminé, cliquez sur le bouton « View selected Model graphically » (à coté du bouton View Serie de tout à l'heure), puis sur le bouton β pour obtenir les estimations des paramètres :



Model Parameter	Estimate	Std. Error	T
LEVEL Smoothing Weight	0.33368	0.0541	6.170
TREND Smoothing Weight	0.00100	0.0404	0.024
SEASONAL Smoothing Weight	0.48175	0.0939	5.133
Residual Variance (sigma squared)	381906	.	.
Smoothed Level	5633	.	.
Smoothed Trend	27.38882	.	.
Smoothed Seasonal Factor 1	0.67033	.	.
Smoothed Seasonal Factor 2	0.63240	.	.
Smoothed Seasonal Factor 3	0.78354	.	.
Smoothed Seasonal Factor 4	0.82862	.	.
Smoothed Seasonal Factor 5	0.83885	.	.

SAS indique ensuite les valeurs des paramètres Niveau, Pente et les coefficients saisonniers.

Nous avons donc ici : α=0.33368 β=0.00100 γ=0.48175.

La variance résiduelle est $\hat{\sigma}^2 = \frac{1}{n-1}\sum (Y_t - \hat{Y}_{t-1}(1))^2 \approx 381906$

Quels sont les mois où les coefficients saisonniers sont les plus élevés ? Ceci était il prévisible ?

4. Précision de l'ajustement

En cliquant sur σ^2 nous obtenons les statistiques d'ajustement ci-dessous :

Statistics of Fit

Y: Y

Winters Method -- Multiplicative

Statistic of Fit	Value
Mean Square Error	369586.1
Root Mean Square Error	607.93594
Mean Absolute Percent Error	10.94703
Mean Absolute Error	434.10945
R-Square	0.940

avec MSE=$\frac{1}{n-k}\sum_{t=1}^{n}(y_t-\hat{y}_t)^2$ où k est le nombre de paramètres du modèle.

RMSE=$\sqrt{MSE}$ MAPE=$\frac{100}{n}\sum_{t=1}^{n}\left|\frac{y_t-\hat{y}_t}{y_t}\right|$ MAE=$\frac{1}{n}\sum_{t=1}^{n}|y_t-\hat{y}_t|$

et R-Square=$1-\frac{\sum_{t=1}^{n}(y_t-\bar{y})^2}{\sum_{t=1}^{n}(y_t-\hat{y}_t)^2}$. (Ce nombre peut être négatif si le modèle est particulièrement mauvais)

5. Calcul des prévisions

Pour obtenir des prévisions, cliquez sur le dernier bouton de la liste (symbolisant un tableau de valeurs).

Vous pouvez lire entre autres :

MOIS	ACTUAL	PREDICT	UPPER	LOWER	ERROR
05/01/69	5010	4538	5750	3327	471.6425
06/01/69	4874	5251	6462	4040	-376.8926
07/01/69	4633	4432	5644	3221	200.6538
08/01/69	1649	2079	3291	868.1798	-430.4084
09/01/69	5951	5362	6573	4151	589.1950
10/01/69	.	6981	8192	5770	.
11/01/69	.	10390	11742	9039	.

Vous voyez par exemple que pour octobre 69, SAS prévoit une vente de 6981 bouteilles (PREDICT).

Nous donnons ci-dessous les **chiffres exacts** de vente de bouteilles d'octobre à août 1970. Comparez avec les prévisions de SAS. Qu'en pensez-vous ?

94	10/01/69	6981
95	11/01/69	9851
96	12/01/69	12670
97	01/01/70	4348
98	02/01/70	3564
99	03/01/70	4577
100	04/01/70	4788
101	05/01/70	4618
102	06/01/70	5312
103	07/01/70	4298

Exercice

Le fichier VENTE2 contient les chiffres des ventes mensuelles d'un produit (en milliers d'unités) de 1985 à 1988. Nous souhaitons effectuer des prévisions pour l'année 1989.
La série comporte elle une tendance, une saisonnalité ?
Quelle méthode de lissage est adéquate ici ?

Si on laisse SAS choisir entre différentes méthodes de lissage (Choisissez Fit Automatic Model au lieu de Select from list) laquelle retient il ?

Vous ferez les calculs sous SAS. Donnez les prévisions pour l'année 1989.

VI. Quelques procédures de gestion de fichiers

A. FORMAT (Créer de nouveaux formats)

1. Objet

Dans l'annexe sur les formats, vous pourrez voir un grand nombre de formats existants pour les dates, l'heure, les nombres. Il vous est aussi possible, sous SAS, de définir vos propres formats et de les utiliser en faisant simplement référence à leur nom comme n'importe quel autre format prédéfini. La procédure Format effectue ce travail.

Les formats personnalisés ainsi créés sont stockés dans un catalogue SAS temporaire: WORK.FORMATS. [118]

> Une confusion à ne pas commettre : Lorsqu'une variable est reformatée, sa valeur interne ne change pas, seule son apparence change. Pour un recodage, la valeur de la variable change (En général).[119]

2. Syntaxe simplifiée

```
PROC FORMAT <options> ;

  VALUE nom champ de valeurs='valeur formatée' ;      Définit les formats d'affichage
  INVALUE nom champ de valeurs='valeur formatée';     Définit ceux d'entrées (input)
  PICTURE nom champ de valeurs='masque d'affichage';  Pour présenter les données
                                                      numériques en ajoutant des symboles ($25,152 ; 12-25-52...)
  SELECT entrées ;                                    Ne sélectionne dans le catalogue FORMATS que les noms de formats ou
                                                      d'informats indiqués ici. (entrées = nom de formats ici)
  EXCLUDE entrées ;                                   N'exclut du catalogue FORMATS que les noms de format ou d'informats
                                                      indiqués ici
RUN ;
```

---

[118] Si vous voulez les conserver, utilisez l'option LIBRARY= ci dessous (pour les récupérer, utilisez « LIBNAME LIBRARY 'votre bibliothèque' ; »).

[119] **Remarque :** La procédure FREQ utilise la valeur de la variable formatée pour définir les classes du tri croisé. Il suffit de changer de format pour effectuer des regroupements en classes ou le contraire. L'avantage de ceci est de ne jamais toucher au fichier original.

Les champs de valeurs

Ils peuvent être de quatre types

Syntaxe	Exemple	Signification
valeur	12	Seule 12 sera concernée
valeur,valeur,...,valeur	14,18,25,36	Seules 12,18,25 et 36 seront concernées
valeur-valeur	12-48 12<-48 12-<48	Tous les nombres entre 12 et 48 (inclus) seront concernés. [12 ;48] correspond à]12 ;48] correspond à [12 ; 48[
Champ,champ,...	14, 12-48, 57 ,57-<62	Tous ces champs seront concernés.

Pour inclure ou exclure certaines valeurs, vous pouvez utiliser « < ».

Ainsi, `invalue anota 0-<10='Faible' 10-12='Moyen' 12<-20='Bon' ;`

associera à 10 et à 12 'moyen'.

Les mots clés **Low**, **High** et **Other** peuvent aussi être utilisé dans les champs de valeurs :

```
invalue anota low-<10='Faible' 10-12='Moyen' 12<-high='Bon' ;
```

Les options principales étant

LIBRARY= *nom de bibliothèque*

On donne ici le nom de la bibliothèque contenant un catalogue (SAS catalog) nommé FORMATS qui, comme son nom l'indique, contient tous les formats créés avec la procédure FORMAT.

Par défauts, les formats sont stockés dans un catalogue nommé FORMATS dans la bibliothèque WORK (temporaire). Ils seront donc perdus lorsque SAS sera désactivé.

Pour conserver vos formats, indiquez votre libname derrière library=.

Pour les récupérer déclarer votre bibliothèque sous le nom LIBRARY :

```
LIBNAME LIBRARY 'G :\STID9799\LOGICIEL\etc...' ;
```

CNTLOUT=*nom de fichier SAS*

Permet de stocker dans un fichier de données SAS des informations sur les formats contenus dans FORMATS. (valeurs, domaine de validité etc...)

Vous pouvez récupérer tout ou partie de ces informations pour redéfinir de nouveaux formats grâce à une autre option `CNTLIN=`*nom de fichier SAS*.

3. Exemples

Nous allons créer un nouveau format nommé « fsexe. » qui transforme « 1 » en « Masculin » « 2 » en « Féminin » et tous les autres nombres en « Non spécifié ».

Nous afficherons ensuite le fichier STID193 en utilisant le format « Fsexe. ».

```
Proc format ;
 value  fsexe
 1='masculin' 2='Féminin' other='Non spécifié'; /*Nous créons un
                                                 nouveau format*/
run;

 /* Nous allons utiliser le Format Fsexe. dans le PRINT ci-dessous  */

proc print data=moi.stid193;
 var groupe sexe;
 format sexe fsexe.; /*  Ne pas oublier le . a la fin du nom de format */
run;
```

Nous obtenons :

OBS	GROUPE	SEXE
1	A	masculin
2	A	masculin
3	A	masculin
4	A	Féminin
5	A	Féminin
6	A	Féminin
7	A	Féminin
8	A	Féminin
9	A	masculin

Remarque :

Dans l'exemple ci-dessus, le format a été utilisé ponctuellement dans la procédure PRINT. Si vous voulez « l'attacher » à la variable SEXE, il vous suffit de l'utiliser dans une étape DATA (cf. plus bas)

Nous pouvons appliquer le Format précédemment défini pour tout autre variable.

Ainsi

```
proc print data=moi.stid193;
 var groupe nbfs;
 format nbfs fsexe.;
run;
```

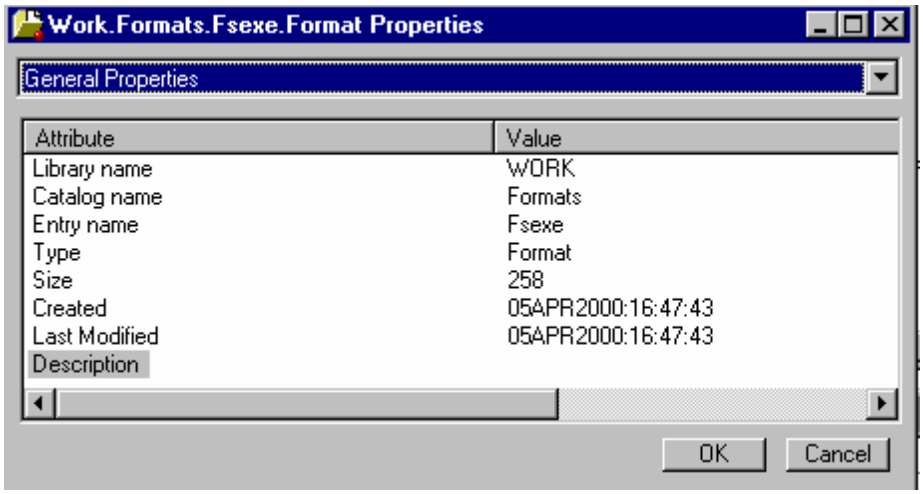
va donner :

OBS	GROUPE	NBFS
1	A	Non spécifié
2	A	Non spécifié
3	A	Féminin
4	A	Non spécifié
5	A	masculin
6	A	masculin
7	A	Féminin
8	A	masculin

Ce qui est complètement idiot je vous l'accorde !

4. Visualisation des formats utilisateurs

Nous pouvons vérifier les formats en cours en allant dans L'EXPLORER et en cliquant sur Work. Enfin, dans cette bibliothèque, il suffit de double cliquer sur le catalogue nommé FORMATS qui contient tous les formats (temporaires) créés par Proc format. Vous sélectionnez FSEXE et choisissez PROPERTIES dans le menu contextuel :



5. Exercices

I)

```
proc format;
 value niveau 0-<8='Faible' 8-12='Moyen' 12<-20='Bon';
run;
proc print data=pub.stid193;
 format mat hg fra niveau.;
run;
```

Que fait ce programme ?

II) Créez un format remplaçant donnant le nom du département de chaque individu et ce, pour l'Isère et les départements voisins de l'Isère. Les individus issus d'un autre département seront qualifiés «Extérieurs ». Vous pourrez utiliser la variable CP qui contient le code postal de chaque individu de STID193.

6. Format permanent *Library=* ; puis *Libname library 'nom de bibliothèque' ;*

Tous les formats précédents ont été crées dans la bibliothèque WORK. Ils seront donc perdus dès que SAS sera stoppé. Pour créer un format permanent, il faut utiliser l'option LIBRARY=

De plus, il vous est possible de spécifier un format lors d'une étape DATA. Dans ce cas, le format restera TOUJOURS attaché à la variable. Ceci est utile lorsque vous donnez un caractère définitif à un format, mais cela impose à votre format d'être toujours accessible.[120]

Exemple

```
libname moi 'i :\SAS';
Proc format LIBRARY=MOI ;
 value  fsexe
 1='masculin' 2='Féminin' other='Non spécifié'; /*Nous créons un
nouveau format*/
run;
```

Le format fsexe. sera ajouté au catalogue FORMATS de la bibliothèque MOI. Je pourrais donc y accéder. [121]

Je vais maintenant utiliser ce format pour créer un fichier de données nommé MOI.ESSAI dont la variable SEXE sera formatée en utilisant « fsexe. ».

```
libname library 'i :\SAS'; /*pour que SAS aille y lire le format fsexe.*/

data moi.essai;
 set pub.stid193;
 keep groupe ordre sexe;
 format sexe fsexe.;
run;
```

Les instructions précédentes font que le format fsexe. sera toujours attaché à la variable sexe ci-dessus. Pour appliquer un format juste une fois, indiquez-le uniquement dans le PROC PRINT par exemple.

```
proc print data=moi.essai; /*Cet affichage impose au format fsexe d'être disponible*/
run;
```

[120] Il devra donc être dans WORK ou dans LIBRARY

[121] En déclarant comme bibliothèque LIBRARY le répertoire contenant le format voulu cf. ci dessous

Remarques importantes:

A chaque fois que vous voudrez afficher le fichier MOI.ESSAI, le format fsexe. DEVRA ETRE DISPONIBLE. (c'est à dire, dans WORK.FORMATS, ou LIBRARY.FORMATS avec le libname adéquat.)
Autrement dit, si vous quittez SAS et que vous voulez afficher MOI. ESSAI, vous devrez taper :

```
libname library 'i :\STID9799\LOGICIEL\...';
proc print data=moi.essai;
run;
```

Lorsque SAS rencontre MOI.ESSAI, il recherche le format fsexe. (car ce format est attaché à la variable sexe). Il commence la recherche dans WORK.FORMATS puis dans LIBRARY.FORMATS. S'il ne le trouve pas, un message d'erreur marque l'interruption de la procédure.

7. Masques d'affichage (picture)

a) Exemple n°1 : Numéros de téléphone

L'exemple le plus classique d'utilisation concerne les n° de téléphone. Il semble que 04/76-82-56-41 soit plus lisible que 0476825641.

Pour ce faire, nous utilisons un masque d'affichage :

```
proc format;
 picture telfax other='99/99-99-99-99';
run;

data carnet;
 input nom $ tel;
 format tel telfax.;


 cards;
 Marie 0525002165
 Eric  0125486565
 Anne  0145479589
 Etienne 565412
;
run;

proc print;
run;
```

Other implique que tous les numéros sont concernés par cette instruction

Le masque d'affichage est attaché au fichier de données. Chaque fois que l'on voudra afficher le fichier de données, le masque devra être présent dans FORMATS

Va donner :

```
                    The SAS System          14:26 Sunday, May 18, 1997

                              OBS    NOM             TEL

                               1     Marie       05/25-00-21-65
                               2     Eric        01/25-48-65-65
                               3     Anne        01/45-47-95-89
                               4     Etienne     00/00-56-54-12
```

Les « 9 » dans PICTURE indiquent que les emplacements des chiffres quitte à remplacer par « 0 » ce qui manque.

Si l'on place des « 0 », PICTURE laissera des blancs s'il manque des chiffres.

Ainsi : `picture telfax other='00/00-00-00-00';`

donnera :

OBS	NOM	TEL
1	Marie	5/25-00-21-65
2	Eric	1/25-48-65-65
3	Anne	1/45-47-95-89
4	Etienne	56-54-12

b) Exemple n°2 : Nombres au format européen (options PREFIX, MULT de Picture)

Nous allons mettre des nombres à 2 décimales sous forme européenne.

Ainsi 14526.26 devient 14.526,26 ce qui est plus lisible.

Pour cela , il faut convertir le nombre en nombre entier avant de lui affecter un masque. On multiplie donc le nombre initial par 100 [122] et on lui applique un masque « 000.000,00 » ou « 000.000.000,00 » si l'on s'attend à de gros chiffres.

Il faut ensuite tenir compte du signe car picture l'enlève. C'est pour cela que nous avons 2 « picture » un pour les négatif (avec un PREFIX='-') et un pour les positifs (avec un PREFIX= '+' Par défaut PREFIX='').

Nous avons donc

```
proc format;
 picture nombfran low-<0='000.000.009,00 FF' (prefix='-' mult=100)
                 0-high='000.000.009,00 FF' (mult=100 prefix='+');
run;

data essai;
 input nom $ nombre;
 format nombre nombfran.;
 cards;

 Marie 2502165.23
 Eric  0125.25
 Anne  -014595.89
 Etienne 26565412
;
run;

proc print;
run;
```

[122] Nous utilisons l'option MULT= de picture qui permet de préciser un coefficient multiplicatif au nombre avant de l'afficher

OBS	NOM	NOMBRE
1	Marie	+2.502.165,23 FF
2	Eric	+125,25 FF
3	Anne	-14.595,89 FF
4	Etienne	+26.565.412,00 FF

Génial non ?

Modifiez le format précédent pour qu'en plus SAS fasse la conversion en Kilo francs puis en Euro. Attention, l'option DIV n'existe pas, il faut vous contenter de MULT...
(19.5 FF doit devenir 3,00 Euro...)

L'énorme intérêt de ce système « Picture » est que le fichier de données n'est pas modifié. Seule l'affichage change. Si par exemple le cours de l'Euro change, il suffit de changer le multiplicateur pour que tout soit à jour.

Remarque : Picture possède une option FILL=' ' qui permet de spécifier le caractère de remplissage des masques lorsqu'il n'y a pas de chiffres. Par défaut c'est un blanc.

c) Exemple n°3 : Dates : Option DATATYPE de PICTURE

Les options DATE, TIME et DATETIME de (DATATYPE) servent à indiquer à SAS que le masque indiqué correspond à une Date, heure ou les deux.

%y indique l'année sur deux caractères %Y indique l'année sur 4 caractères %m indique le mois %d indique le jour du mois.

Exemple :

```
proc format;
picture nouvdate (default=15)                    Spécifie la taille (en caractères)
 other='%d-%m %Y' (datatype=date);               Notre nouveau format
 run;

 data ech;
 set moi.stid193 (obs=10);
 keep groupe ordre sexe date;
 format date nouvdate.;
 run;

 proc print data=ech noobs;
 run;
```

Va donner :

GROUPE	ORDRE	SEXE	DATE		
	A		1	1	21-10 1973
	A		2	1	8-12 1974
	A		3	1	15-8 1972
	A		4	2	10-11 1972
	A		5	2	30-11 1974
	A		6	2	11-2 1974
	A		7	2	14-9 1974
	A		8	2	22-11 1973
	A		9	1	8-6 1974
	A		10	1	15-6 1973

8. Informat (INVALUE)

Dernier point, vous pouvez définir vos propres formats d'entrée.

Un exemple simple va vous permettre de comprendre la portée de ce concept.

J'ai un fichier d'élèves avec leurs 3 notes sous forme inhabituelle. Chaque élève a été noté par une lettre N,MM, M, AB,B,TB. La correspondance avec notre système de notes est la suivante : N(6/20) ; MM(8/20), M(10/20), AB(12/20) , B(14/20) TB(16/20)

Nous voulons importer ce fichier et convertir les notes numériquement pour pouvoir calculer les moyennes correspondantes.

J'ai le fichier suivant, je veux calculer la moyenne par élève...

```
Jean-pierre AB B M
Herve N MM MM
Anestis TB TB TB
Bernard TB N TB
```

```
proc format;
  invalue fnote 'N'=6 'MM'=8 'M'=10
                'AB'=12 'B'=14 'TB'=16
                'ABS'=.  ;
 run;

data stid;
 input nom $ @;

 do i=1 to 3;
  input note : fnote. @;
  output;
 end;
keep nom note;
cards;
 Jean-pierre AB B M
 Herve N MM MM
 Anestis TB TB TB
 Bernard TB N TB
run;

proc print data=stid;
run;
```

Nous définissons l'Informat Fnote. ici

Note manquante remplacée par un point.

Le caractère @ permet d'attendre la fin de la ligne avant de sauter à la ligne suivante. Nous avons encore trois notes à lire avant de changer de ligne.

Nous les lisons l'une après l'autre.

Pour ôter la variable i du fichier de données

On obtient :

OBS	NOM	NOTE
1	Jean-pie	12
2	Jean-pie	14
3	Jean-pie	10
4	Herve	6
5	Herve	8
6	Herve	8
7	Anestis	16
8	Anestis	16
9	Anestis	16
10	Bernard	16
11	Bernard	6
12	Bernard	16

```
proc sort data=stid;

by nom;
run;

proc means data=stid;
by nom;
run;
```

Nous trions le fichier pour utiliser le BY dans la procédure MEANS

Ce qui donne :

```
Analysis Variable : NOTE

------------------------------------------ NOM=Anestis ------------------------------------------
                N          Mean       Std Dev       Minimum       Maximum
               ------------------------------------------------------------
                3    16.0000000             0    16.0000000    16.0000000
               ------------------------------------------------------------

------------------------------------------ NOM=Bernard ------------------------------------------
                N          Mean       Std Dev       Minimum       Maximum
               ------------------------------------------------------------
                3    12.6666667     5.7735027     6.0000000    16.0000000
               ------------------------------------------------------------

------------------------------------------ NOM=Herve ------------------------------------------
                N          Mean       Std Dev       Minimum       Maximum
               ------------------------------------------------------------
                3     7.3333333     1.1547005     6.0000000     8.0000000
               ------------------------------------------------------------

------------------------------------------ NOM=Jean-pie ------------------------------------------
                N          Mean       Std Dev       Minimum       Maximum
               ------------------------------------------------------------
                3    12.0000000     2.0000000    10.0000000    14.0000000
               ------------------------------------------------------------
```

9. Compléments

Je vous encourage à consulter le SAS *Procedures guide* pour des compléments sur cette procédure FORMAT.

B. TRANSPOSE (Transposer un fichier)

Son nom est très explicite, elle permet en effet de transposer un fichier de données. Les colonnes deviennent les lignes et inversement. Elle permet aussi de réarranger de façon plus complexe certains fichiers.

Voici un petit exemple tiré du SAS *Procedures guide* :

```
DATA A;
   INPUT A B C;
   CARDS;
1 2 3
4 5 6
;
PROC PRINT;
TITLE 'Le fichier original...';
 RUN;

PROC TRANSPOSE DATA=A;
RUN;

PROC PRINT;
   TITLE 'Le fichier transposé';
RUN;
```

Ce qui donne

Le fichier original...

OBS	A	B	C
1	1	2	3
2	4	5	6

Le fichier transposé

OBS	_NAME_	COL1	COL2
1	A	1	4
2	B	2	5
3	C	3	6

SAS a créé trois variables _NAME_,COL1 et COL2 qui contiennent les « lignes » du fichier précédent.
La variable _NAME_ contient tous les noms de variables.

Syntaxe simplifiée

```
PROC TRANSPOSE (options);
  VAR variables;
  BY variables;
  ID variable ;
  COPY variables :
RUN ;
```

Les variables à transposer

Permet la création de sous-groupes... **le fichier devra avoir été trié avant**
Les variables BY ne sont pas transposées.

Colonne contenant les noms des futures colonnes

variables à recopier dans le fichier final **sans les transposer.**

Les principales options étant :

DATA= et OUT= pour spécifier les fichiers de données d'entrée et de sortie.

Exemples

utilisation de id.

On modifie légèrement le programme précédent :

```
DATA A;
   INPUT A B C D $;
   CARDS;
1 2 3 X
4 5 6 Y
;
PROC PRINT;
TITLE 'Le fichier original...';
 RUN;

PROC TRANSPOSE DATA=A;
ID D;
RUN;
PROC PRINT;
   TITLE 'Le fichier transposé';
RUN;
```

```
                          Le fichier original...

                         OBS    A    B    C    D

                          1     1    2    3    X
                          2     4    5    6    Y

                          Le fichier transposé

                         OBS    _NAME_    X    Y

                          1       A       1    4
                          2       B       2    5
                          3       C       3    6
```

Utilisation de COPY

En ajoutant COPY C ; au dessous du ID D ; on obtient :

```
                          Le fichier original...

                       OBS    A    B    C    D
                        1     1    2    3    X
                        2     4    5    6    Y

                          Le fichier transposé

                     OBS    C    _NAME_    X    Y
                      1     3      A       1    4
                      2     6      B       2    5
```

C. CONTENTS (Inventaire d'une bibliothèque)

Cette procédure est très utile pour connaître les fichiers figurant dans votre bibliothèque ou pour avoir des informations précises sur un fichier en particulier.

Syntaxe simplifiée

```
PROC CONTENTS options;
RUN;
```

Les options étant:

DATA= *nom de fichier ou bilbiothèque._all_*

Exemples

```
proc contents data= moi.stid;
run;
```

pour obtenir des informations sur le fichier STID de la bibliothèque MOI.

```
libname MONREP 'G:\STID9597\TOTO';
proc contents data=monrep._all_;
run;
```

pour obtenir des informations sur TOUS les fichiers SAS situés sur mon répertoire.

OUT= *fichier SAS*
Pour sauvegarder les informations affichées dans la fenêtre OUTPUT.

DIRECTORY
Affiche un catalogue résumé des fichiers de la bibliothèque.

MEMTYPE=
Spécifie le(s) type(s) des fichiers SAS à cataloguer.
DATA (pour les fichiers de données, c'est notre cas !)
CATALOG pour les fichiers de type "catalog"
PROGRAM pour les programmes compilés.
...
ALL tous les types.

POSITION
Pour afficher les variables du fichier de données dans leur ordre d'apparition sur ce même fichier (et non plus dans l'ordre alphabétique).

SHORT
Affiche seulement une liste des variables du fichier de données.

NOPRINT
Aucun affichage. Cette option n'est utile que si elle est combinée avec OUT=

Exemple

```
proc contents data=pub.stid193 directory;
run;
```

va donner:

Un résumé des fichiers se trouvant dans la bibliothèque SASUSER:

```
                          CONTENTS PROCEDURE

                                -----Directory-----

                           Libref:        SASUSER
                           Engine:        V608
                           Physical Name: C:\SAS\SASUSER

                           #  Name      Memtype  Indexes
                          -------------------------------
                           1  ACP       DATA
                           2  AGENTS    CATALOG
                           3  AGENTS    DATA
                           4  ARB2TYPE  DATA
                           5  ARB3TYPE  DATA
                           6  BUILD     CATALOG
                           7  BUILD     DATA
                           8  CLASS     CATALOG
                           9  CLASS     DATA
                          ..................
                          36  STID193N  DATA
                          37  VENEER    DATA
```

Des informations complètes sur le fichier STID193

```
                          CONTENTS PROCEDURE

Data Set Name: SASUSER.STID193                    Observations:         106
Member Type:   DATA                               Variables:            9
Engine:        V608                               Indexes:              0
Created:       9:25 Wednesday, March 29, 1995     Observation Length:   60
Last Modified: 9:25 Wednesday, March 29, 1995     Deleted Observations: 0
Protection:                                       Compressed:           NO
Data Set Type:                                    Sorted:               YES
Label:

                -----Engine/Host Dependent Information-----

                    Data Set Page Size:       4096
                    Number of Data Set Pages: 2
                    File Format:              607
                    First Data Page:          1
                    Max Obs per Page:         67
                    Obs in First Data Page:   44
```

Une liste des variables du fichier STID193:

Type indique le type des variables:
Char pour les variables caractères (alphanumériques)
Num pour les variables numériques

Len indique la longueur de chaque variable

Pos sa position dans le fichier.

CONTENTS PROCEDURE

-----Alphabetic List of Variables and Attributes-----

#	Variable	Type	Len	Pos
3	BAC	Char	3	9
4	FRA	Num	8	12
1	GROUPE	Char	1	0
5	HG	Num	8	20
6	MAT	Num	8	28
9	NBFS	Num	8	52
8	POIDS	Num	8	44
2	SEXE	Num	8	1
7	TAILLE	Num	8	36

Des informations sur un éventuel tri du fichier:

CONTENTS PROCEDURE

-----Sort Information-----

Sortedby: SEXE
Validated: YES
Character Set: ANSI

Ici, nous voyons que le fichier STID193 a été trié par rapport à la variable SEXE. Nous avons donc perdu l'ordre original de ce fichier.

D. **DATASETS** (gestion de bibliothèques, de fichiers de données**)**

Cette procédure vous permet de gérer vos fichiers de données. Vous pouvez copier tout ou partie d'un fichier de données, en supprimer, visualiser leur contenu, éditer le contenu d'une bibliothèque.

Vous pouvez également modifier les noms des variables d'un fichier, leur format etc.

Syntaxe simplifiée

`PROC DATASETS library=`*nom de bibliothèque* *<options>*`;`

Certaines options étant:

KILL (Entraîne la destruction de tous les fichiers de la bibliothèque spécifiée)
NOLIST (Supprime l'affichage du nom des fichiers contenu dans la bibliothèque spécifiée)

*Attention, dans ce qui suit, les fichiers spécifiés sont contenus dans la bibliothèque spécifiée derrière l'instruction « library »***. Leur nom n'est donc pas composé avec celui de la bibliothèque à laquelle ils appartiennent**. *(exemple MOI.TOTO est à remplacer par TOTO)*

SOUS COMMANDES de la procédure DATASETS

1. Concaténation de fichiers

APPEND BASE=*fich1* DATA=*fich2* <Force>;

Pour ajouter fich2 à la suite de fich1. Le résultat est dans fich1. L'option « force » oblige Sas a concaténer les deux fichiers même s'ils n'ont pas les mêmes variables.

Exemple:

Voici le contenu du fichier work.hommes:

OBS	NAME	AGE	HEIGHT	WEIGHT
1	Mary	15	66.5	112.0
2	Sharon	15	62.5	112.5

Voici le contenu du fichier work.femmes:

OBS	NAME	AGE	HEIGHT	WEIGHT
1	Guido	15	67.0	133
2	William	15	66.5	112
3	Philip	16	72.0	150

Tapons le programme suivant:

```
proc datasets library=work;
 append base=femmes data=hommes;
run;
quit;
```

Le fichier work.femmes contiendra maintenant:

OBS	NAME	AGE	HEIGHT	WEIGHT
1	Mary	15	66.5	112.0
2	Sharon	15	62.5	112.5
3	Guido	15	67.0	133.0
4	William	15	66.5	112.0
5	Philip	16	72.0	150.0

Concaténation d'une partie d'un fichier

Il suffit d'ajouter la commande Where comme le montre l'exemple suivant où nous souhaitons ajouter au fichier femmes uniquement les individus du fichier hommes dont l'âge est égal à 15.

```
proc datasets library=work;
 append base=femmes data=hommes(where=(age=15));
run;
quit;
```

Le fichier work.femmes contiendra maintenant

OBS	NAME	AGE	HEIGHT	WEIGHT
1	Mary	15	66.5	112.0
2	Sharon	15	62.5	112.5

3	Guido	15	67.0	133.0
4	William	15	66.5	112.0

2. Changement de nom d'un fichier

CHANGE *nom actuel du fichier=nouveau-nom...*;

Pour renommer un (ou plusieurs) fichier(s) de données.

Exemple:

Pour renommer le fichier work.femmes en work.ensemble, nous entrons le programme suivant:

```
proc datasets library=work;
 change femmes=ensemble;
run;
quit;
```

EXCHANGE *nom=autre-nom1 etc...;*

Pour échanger les noms de 2 ou plusieurs fichiers. Ici le fichier « nom1 » sera renommé « autre-nom1 » et le fichier « autre-nom1 » sera renommé « nom1 ».

3. Inventaire d'une bibliothèque, informations sur un fichier

CONTENTS ;

Référez vous au paragraphe PROC CONTENTS de ce document.

4. Suppression de fichiers

DELETE *fichiers*;

Pour supprimer les fichiers spécifiés.

Exemple:

```
proc datasets library=work;
 delete femmes;
run;
quit;
```

Programme détruisant le fichier work.femmes.

5. Copie de fichiers

COPY OUT=*biblio1* <IN=*biblio2*> <Move>;

Copie tout ou partie des fichiers de biblio2 dans la biblio1. Si IN n'est pas spécifiée, c'est la bibliothèque de la procédure DATASETS qui est retenue. Pour sélectionner les fichiers qui doivent être copiés voir SELECT et EXCLUDE.
Si l'option MOVE est spécifiée, les fichiers copiés seront détruit de biblio2.

EXCLUDE *fichiers;*

Cette instruction ne peut être utilisée qu'après un COPY. Vous spécifiés ici les fichiers qui ne doivent pas être concernés par la commande COPY.

Exemple:
```
Proc datasets library=sasuser;
copy out=moi;
exclude toto stid92;
Run;
```

Ce programme va copier tous les fichiers de la bibliothèque SASUSER dans la bibliothèque MOI à l'exception des fichiers SASUSER.MOI, SASUSER.TOTO. Si un Move est spécifié dans l'instruction Copy, tous les fichiers copiés seront détruits dans la bibliothèque SASUSER.

SELECT *fichiers;*

Cette instruction ne peut être utilisée qu'après un COPY. Vous spécifiés ici les fichiers qui doivent être concernés par la commande COPY.

Exemplen°1:

```
Proc datasets library=sasuser;
copy out=moi;
select toto stid92;
Run;
```

Ce programme va copier seulement les fichiers SASUSER.STID92, SASUSER.TOTO dans la bibliothèque MOI. Si un Move est spécifié dans l'instruction Copy, ces deux fichiers seront détruits dans la bibliothèque SASUSER.

Exemple n°2:

Ce programme permet de copier sur la disquette (A:) les fichiers CLASS, STID193 et ACP de la bibliothèque SASUSER.

```
libname moi 'A:';
proc datasets library=sasuser;
 copy out=moi in=sasuser;
 select class stid193 acp;
run;
quit;
```

6. Modifications sur les variables d'un fichier (format, nom...)

MODIFY *fichier;*

Pour modifier, entre autres, les noms des variables d'un fichier de données, modifier les formats, les étiquettes etc... Le fichier doit être situé dans la bibliothèque indiquée dans la procédure DATASETS.

Les instructions suivantes ne peuvent êtres utilisées qu'après une instruction MODIFY

LABEL='*étiquette du fichier de données';*

Exemple:

```
Proc datasets library=sasuser;
  modify stid193 (label='Questionnaire STID93');
run;
quit;
```

va modifier l'étiquette du fichier STID193 de la bibliothèque SASUSER (Notez la présence des parenthèses) Pour voir le résultat, tapez le programme suivant et consultez l'Output.

```
Proc datasets library=sasuser;
 contents data=stid193;
run;
quit;
```

RENAME *variable1=nouveau nom variable2=nouveau nom ...;*

Pour renommer certaines variables du fichier de données spécifié dans la commande Modify.

Exemple:

```
Proc datasets library=sasuser;
  modify stid193;
  rename bac=seriebac;
run;
quit;
```

Va renommer la variable BAC en SERIEBAC.

Pour voir le résultat, tapez le programme suivant et consultez l'Output.

```
Proc datasets library=sasuser;
 contents data=stid193;
run;
quit;
```

FORMAT *liste de variables format1 liste de variables format2 ...*

Exemple:

```
Proc datasets library=sasuser;
  modify stid193;
  format fra hg mat 4.1 taille 6.2;
run;
quit;
```

Dans cet exemple, les notes de Français, Math et Histoire-géo auront un format 4.1 et la taille un format 6.2. (Le premier chiffre indique la taille

maximale du nombre et le deuxième, le nombre de décimale(s)). Pour avoir plus d'information sur les formats disponibles, allez voir en annexe.

Pour visualiser le résultat faites un PROC PRINT et vous obtenez:

OBS	FRA	HG	MAT	TAILLE
103	8.0	12.0	17.0	165.00
104	11.0	6.0	12.0	160.00
105	10.0	12.0	18.0	167.00
106	6.0	6.0	15.0	168.00

Ici, on voit que la variable taille est codée sur 6 caractères dont 2 décimales.

7. Réparer des fichiers endommagés par une panne système....

REPAIR *nom de fichier;*

Cette instruction a pour but de réparer les fichiers endommagés par une panne du système.

VII. Une autre façon d'utiliser SAS: SAS / ASSIST

A. Présentation

SAS/ASSIST constitue une autre façon d'utiliser SAS. Ce module vous permet, à l'aide de quelques clics de souris, d'obtenir rapidement des analyses statistiques simples ainsi que les programmes correspondants.[123] Notez qu'un nouveau module SAS : SAS *Enterprise Guide* très amusant à utiliser fait un peu la même chose que SAS ASSIST. L'utilitaire *Graph n Go* est un SAS ASSIST spécialisé dans les graphiques.

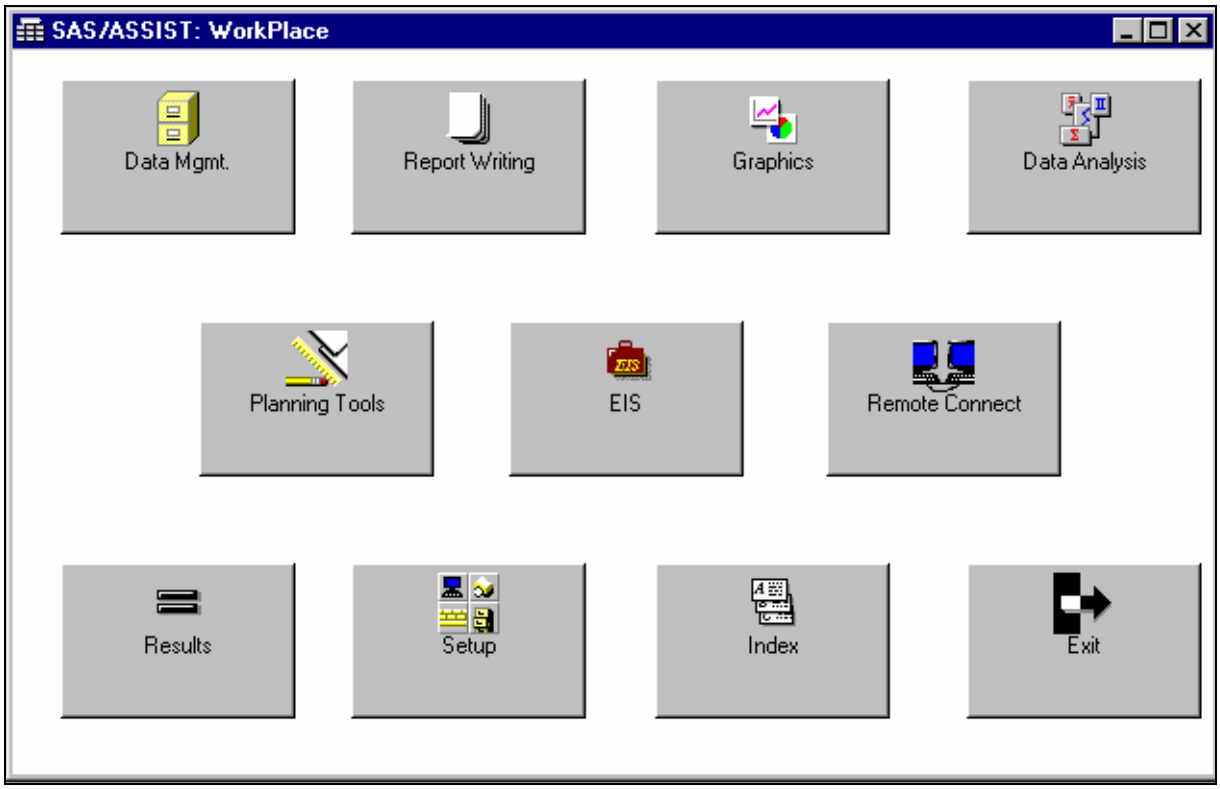
Si cette façon d'utiliser SAS s'avère trop limitée, elle contitue néanmoins une alternative qu'il ne faut pas négliger pour une utilisation occasionelle de SAS par exemple.

[123] Ceci peut être très intéressant pour retrouver la syntaxe d'une procédure.

B. Comment lancer SAS/ASSIST ?

Pour lancer SAS ASSIST vous pouvez aller dans le menu SOLUTIONS/ASSIST. Choisissez ensuite WORK PLACE.[124]

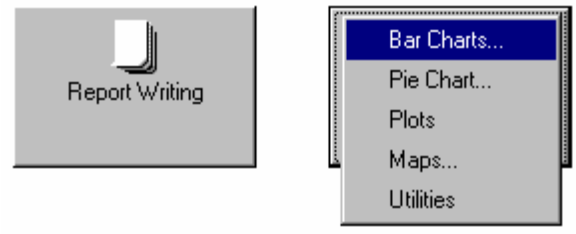
Vous voyez alors l'écran suivant:



C. Exemple d'utilisation de SAS/ASSIST:

Diagramme à bandes des moyennes des notes de maths selon les groupes.

Dans le menu précédent, cliquez sur GRAPHICS. Puis sur BAR CHART :



Remplissez la boite de dialogue comme suit :

[124] L'autre option permet de retrouver un SAS ASSIST conforme à la V 6.12.

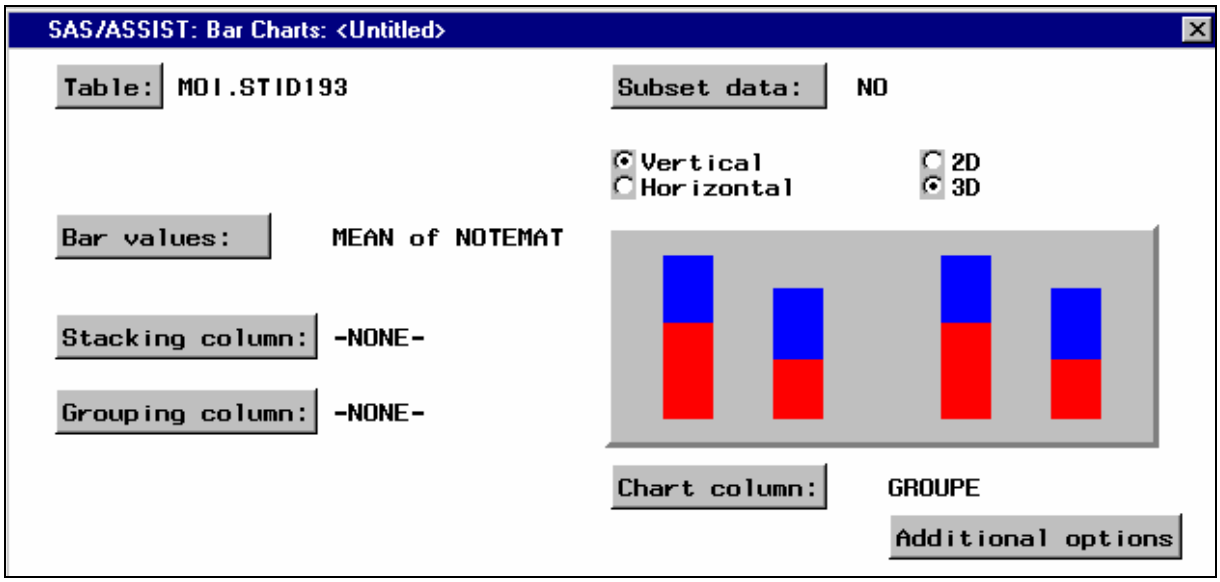


Table: SAS demande ici le fichier de données SAS à considérer. Cliquez sur ce bouton et choisissez le fichier SAS MOI.STID193.

Bar value: Par défaut SAS va représenter les effectifs. Si vous souhaitez avoir les pourcentages cliquez sur ce bouton et choisissez ce que vous voulez.

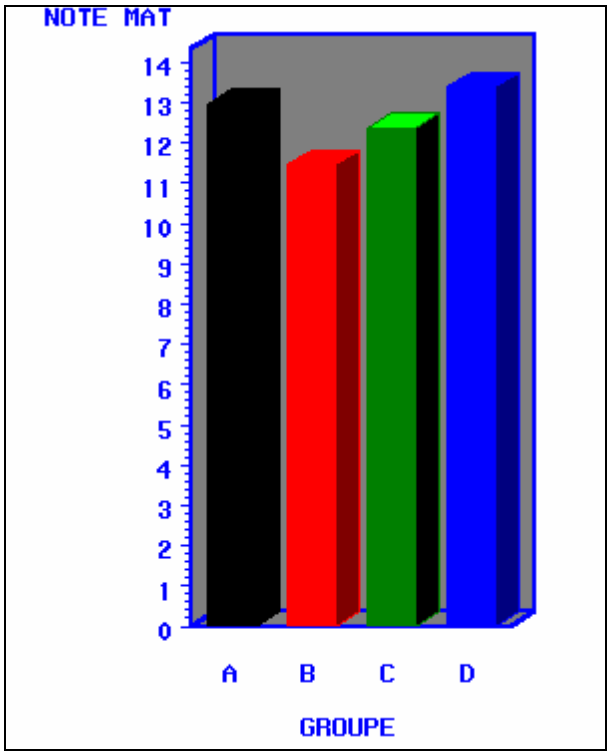
Subset data: Correspond à l'option **BY** des procédures SAS. Si vous souhaitez diviser votre population en distinguant selon le sexe, vous pouvez indiquer ici la variable sexe. ATTENTION, le fichier devra avoir été trié selon la variable (sexe ici) si nous voulons que cela fonctionne. Pour notre exemple, nous ne divisons pas en sous-populations.

Chart column: Quelle est la variable à représenter ? Ici c'est le bac. Cliquez sur le bouton et choisissez la variable BAC.

Additional options: Si vous souhaitez modifier les couleurs par défaut du graphique, vous pouvez aller dans ce menu et les changer à votre guise.

Cliquez ensuite sur le bouton SUBMIT pour lancer la procédure SAS. Devant vos yeux ébahis, un "magnifique graphique" apparaît. [125]

[125] Il est possible de le transférer sous Word (Avec la souris, délimitez la zone qui vous intéresse et fait un Edit/ Copy to paste buffer, basculez sous Word et collez) mais la copie se fait en mode Bit-map. Cette image sera très encombrante, très longue à imprimer...



D. Comment obtenir les instructions SAS qui ont donné le résultat précédent

Fermez le graphique. Vous pouvez récupérer le programme qui vous a permis d'obtenir ce résultat. Allez dans l'editeur de programmes, la touche F4 permet de rappeler le programme :

```
/*------------------------------------------------------------------*
  | Summary:                                                         |
  |    Creating a grouped and stacked  bar chart using the table     |
  |    MOI.STID193 and charting the columns                          |
  |    GROUPE along the vertical axis.                               |
  | Generated: 02MAY2000 13:50:33                                    |
  *------------------------------------------------------------------*/
 /*------------------------------------------------------------------*
  | The GOPTIONS statement allows you to have more control over the  |
  | final appearance of your output such as fonts, colors, text      |
  | height and so on.  The output device and destination is also     |
  | specified in the goptions statement.                             |
  *------------------------------------------------------------------*/

goptions reset=(axis, legend, pattern, symbol, title, footnote) norotate
         hpos=0 vpos=0 htext= ftext= ctext= target= gaccess= gsfmode= ;
goptions device=WIN  ctext=blue
         graphrc interpol=join;

 /*------------------------------------------------------------------*
  | PATTERN statements allow you to define colors and patterns in    |
  | the chart, map or plot that you are creating.  SAS/GRAPH uses    |
  | any pattern statements that you specify.  If more are needed,    |
  | default PATTERN statements are used.                             |
  *------------------------------------------------------------------*/
pattern1 value=SOLID;
 /*------------------------------------------------------------------*
  | AXIS statements allow you to supply information on how your      |
  | vertical and horizontal axes will appear on the graph.           |
  *------------------------------------------------------------------*/
axis1
   color=blue
   width=2.0
   ;
```

```
axis2
   color=blue
   width=2.0
   ;
axis3
   color=blue
   width=2.0
   ;
 /*------------------------------------------------------------------*
  | This section produces the actual bar chart and contains the      |
  | options that directly relate to the data and the axis area.      |
  *------------------------------------------------------------------*/

proc gchart data=MOI.STID193;
   VBAR3D GROUPE /


      maxis=axis1
      raxis=axis2
      frame
      type=MEAN
      sumvar=NOTEMAT
     patternid=midpoint
      ;
run; quit;
```

Exercices :

I) Effectuez des statistiques élémentaires sur les individus de STID193 pour les variables : Taille, Poids. (Mean, Max, Min etc.) Récupérez le programme correspondant.

II) Faites un TTest tester la différence de moyenne entre les tailles des hommes et celles des femmes des STID de france (en supposant que les STID193 sont un échantillon aléatoire représentatif de la population des STID de france)

Dans SAS/ASSIST choisissez DATA ANALYSIS, ANOVA puis TTEST. Choisissez ensuite la bonne option. (paired=appariée cf arbres debout et abattus de la fiche de tests Minitab)

Active data set: Donnez le fichier STID193.

"Dependent variable": est la variable sur laquelle nous faisons le test: Taille.
Quant à la variable de "classification" c'est le sexe.
Le résultat s'affiche. Interpréter le résultat.[126] Choisissez File/End pour revenir à SAS ASSIST.
Si vous le souhaitez, vous pouvez récupérer le programme SAS par la touche RESULTS.

[126] Vous pouvez consultez le paragraphe TTEST de ce document pour interpréter cette sortie.

VIII. PETIT DICTIONNAIRE ANGLAIS-FRANCAIS

pour l'utilisation de SAS

Anglais-US	Français
Analysis	Analyse
To average	Faire la moyenne
Axis	Axes
To cancel	Annuler
Chi square	χ^2
To clear	Effacer
Column	Colonne
Cumulative function	Fonction de répartition
Cut	Couper
Data	Donnée
Decreasing	Décroissant
Density	Densité
Discard graph	Effacer le graphe
Drop	Enlever
Eigen value	Valeur propre
Eigen vector	Vecteur propre
factorial design	Plan d'experience (factoriel)
experiment	
exponential smoothing	Lissage exponentiel
File	Fichier
Fill	Remplissage
Fonts	Polices de caractères
Frequency	**Effectif**
Hide	Cacher
Increasing	Croissant
Keep	Garder
Label	Etiquette (légende)
Management (data)	Gestion(de données)
Marked	Sélectionné (e)
Mean	Moyenne (arith. en gén.)
Median	Médiane
Merge	Fusionner
Model	Modèle

Anglais-US	Français
Near	Près (de)
Nearest	Le plus près
None	Aucun
To paste	Coller
Percent	Pourcent %
Popup menu	Menu contextuel
Random	Aléatoire
Remaining	Restant (es)
Recall	Rappeler
Rename	Renommer
Row	Ligne (worksheet)
Run	lg sas: exécuter
Setup	Configuration
Slice	Portion (de graphique camenbert)
Small	Petit
Smallest	Le(la,les) plus petit(es)
Smoothing	Lissage
Solid	Plein (remplissage)
Sort	Trier
Sum of Squares (USS ou CSS)	Somme des carrés
Range	Etendue
Rank	Rang
Time serie	Série chronologique
To remove	Enlever, ôter
To round	Arrondir
Ties	Exaequo
Toolbar	Barre d'outils
Toolbox	Boîte d'outils
Unmark	Ne plus sélectionner
Values	Valeurs
Window	Fenêtre
Worksheet	Feuille de données

IX. BIBLIOGRAPHIE COMMENTEE

Cette bibliographie n'est pas exhaustive. Elle contient les livres qui vous sont accessibles pour approfondir le cours de Statistiques et pour l'utilisation de SAS.

Ouvrages statistiques

Théorie et méthodes statistiques tomes 1 et 2, Pierre Dagnélie, Presses agronomiques de Gembloux 1992. *Ces ouvrages couvrent le cours du DUT STID, ils comportent beaucoup d'exemples. Très utile pour approfondir le cours vu en STID. Ils sont aussi assez cher (250FF par tome environ)*

Analyse statistique à plusieurs variables, Pierre Dagnélie, Presses agronomiques de Gembloux 1975. *Idem*

Méthodes Statistiques en Gestion, Michel Tenenhaus, Dunod Entreprise 1994. *Ce livre est essentiellement pratique. Cet ouvrage ne comporte pratiquement pas de démonstrations mathématiques. Il couvre un large éventail de notions (Tests, ANOVA, Régression linéaire, , Séries temporelles, ACP, AFC, Classification hiérarchique) et est illustré par de nombreux exemples numériques entièrement traités sur SAS, Statgraphics...*

Analyses factorielles simples et multiples, Brigitte Escofier, Jérôme Pages, Dunod 1990 *C'est un ouvrage très complet sur les analyses factorielles. A utiliser pour approfondir les notions vues en STID.*

Régression non linéaire et application, Anestis Antoniadis, Jacques Berruyer, René Carmona, Economica 1992. *Ouvrage qui permet de*

prolonger votre cours sur la régression non linéaire. Il apporte également des compléments intéressants sur la régression linéaire.

Probabilités, Analyse des données et Statistique, Saporta, Technip 1990. *Ouvrage théorique de référence en probabilité et statistique.*

Documentation SAS

Les trois ouvrages suivants sont indispensables pour l'utilisation du module SAS de base.

SAS Companion for the Microsoft Windows environment Version 6 première édition. *Ouvrage indispensable pour utiliser SAS dans l'environnement Windows PC.*

SAS Language version 6, première édition. *Comme son nom l'indique vous trouverez tout ce qui concerne le langage SAS. Il doit être accompagné de l'ouvrage suivant pour utiliser les procédures SAS.*

SAS Procedures Guide version 6, 3ème édition. *Attention, vous ne trouverez dans cet ouvrage que les procédures du module SAS de base. Pour les procédures purement statistiques, il faut vous référer aux ouvrages suivants. Pour la procédure SQL, il existe un ouvrage spécifique cf. ci-dessous.*

SAS guide to the SQL procedure Version 6 First Edition. *Cet ouvrage très clair décrit les possibilité (nombreuses) de la procédure SQL. Elle constitue, à elle seule, une autre façon d'utiliser SAS grâce au langage SQL.*

SAS guide to Macro Processing Version 6 Second Edition. *Cet ouvrage, très clair, permet d'utiliser les Macros du langage SAS. C'est un ouvrage indispensable pour qui utilise le langage SAS couramment.*

SAS technical reports P222 et P242 Release 6.07 et 6.08. *Ils décrivent les changements et les améliorations pour la version 6.07et 6.08 du module SAS de base (PROC Contents, Datasets, SQL...)*

Module STAT

SAS/STAT User's Guide Volumes 1 et 2 version 6, 4ème édition *Ces 2 volumineux ouvrages décrivent les procédures du module STAT (REG, ANOVA, GLM, PRINCOMP, TTEST et cie)*

SAS technical report P229 Release 6.07 *Décrit les changements pour le module STAT version 6.07.*

Module SAS/ACCESS

SAS/ACCESS Interface to PC File Formats Usage and reference, Version 6 first edition.*Comme son nom l'indique, elle décrit l'utilisation des procédures du module SAS/ACESS permettant de lire et de transférer les fichiers au format Dbase.*

Module SAS/FSP

SAS/FSP Software Usage and Référence, Version 6 First edition *Il décrit les procédures FSVIEW, FSBROWSE... C'est un ouvrage très clair et accessible.*

Module SAS/AF, langage SCL

SAS Screen Control Language : *Référence Version 6, Second Edition*
pour accéder à la syntaxe d'une commande SCL. Tout y est.

SAS Screen Control Language : *Usage Version 6, First Edition*
pour à des exemples de programmation SCL. Manipulation des fichiers de données en SCL etc.

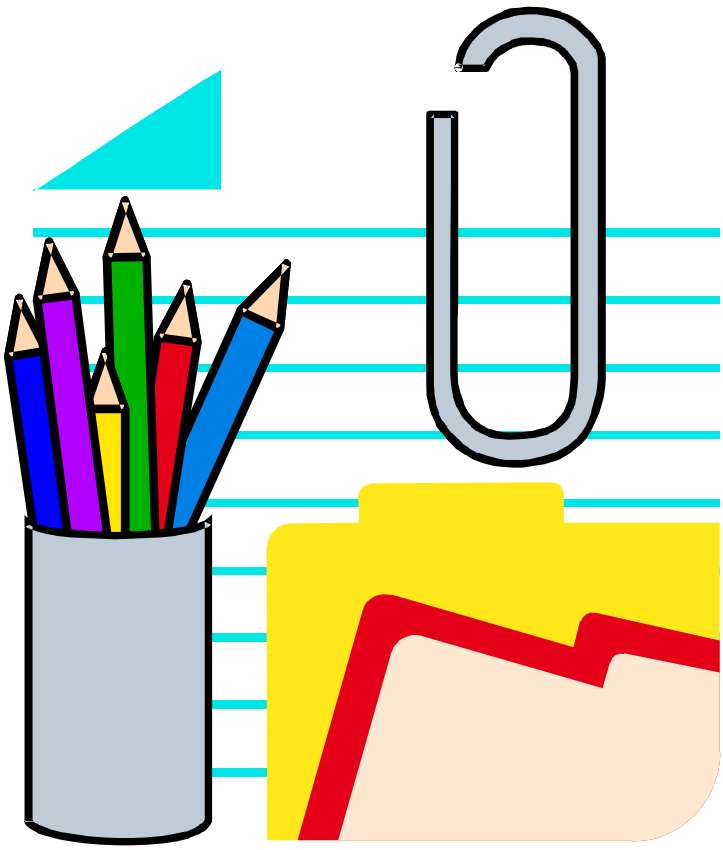
SAS/AF Software : Frame Entry : *Usage and Référence, Version 6, First Edition.*
Pour avoir des détails sur les objets accessibles dans les entrées de type FRAME, les méthodes...

Module SAS/ASSIST

Introduction au Système SAS avec le logiciel SAS/ASSIST Version 6, deuxième édition.

Doing more with SAS/ASSIST software Version 6, First edition.
Ces deux ouvrages expliquent l'utilisation de SAS grâce au module SAS/ASSIST. Ils sont très conviviaux. Le premier est en français et le deuxième en anglais.

X. ANNEXES



A. Raccourcis clavier

La touche F9 lance l'affichage des commandes associées aux touches de fonction :

Vous pouvez les modifier en tapant simplement en face de la combinaison de touches associée la commande souhaitée.

	Touche	Commande	Signification
	F1	help	Aide
	F2	reshow	
	F3	end	Fermeture de la fenêtre en cours (OUTPUT)
	F4	recall	Rappel du dernier programme tapé (program EDITOR)
	F5	pgm	Fenêtre program editor active
	F6	log	Fenêtre log active
	F7	output	Fenêtre output active
	F8	zoom off ; submit	Compilation du programme
	F9	keys	Cet écran
	F11	command bar	Barre de commande
	F12		
SHF	F1	subtop	Compilation de la première ligne du program editor
SHF	F2		
SHF	F3		
SHF	F6		
SHF	F7	left	A gauche
SHF	F8	right	A droite
SHF	F9		
SHF	F10	wpopup	Menu contextuel, clic droit de la souris
SHF	F11		
SHF	F12		
CTL	A		
CTL	B	libname	Affichage des bibliothèques en cours
CTL	D	dir	Affichage du contenu de la bibliothèque en cours
CTL	E	clear	Nettoyage de la fenêtre active
CTL	F	footnote	Affichage de la fenêtre de footnote qui permet de définir les notes de bas de page à afficher à chaque nouvelle page.
CTL	G		
CTL	H	help	Aide
CTL	I	options	Liste des options du système SAS
CTL	J		
CTL	K	cut	Couper
CTL	L	log	Fenêtre LOG active.
CTL	M	mark	Marquer
CTL	Q	filename	Liste des fileref
CTL	R	rfind	
CTL	T	title	Liste des titres à afficher sur chaque page
CTL	U	unmark	Annuler la marque
CTL	W	access	Fenêtre ACCESS

B. OPERATEURS ET FONCTIONS

1. Les opérateurs

a) arithmétiques

Syntaxe	Nom	Exemple
*	Multiplication	AIRE=LON*LAR
/	Division	TACM=TAM/100
+	Addition	TOT=PRIX1+PRIX2
-	Soustraction	MONTANT=PRIX-REMISE
**	Puissance	CUBE=ABSCISSE**3
><	Maximum	MAX=MATH1><MATH2
<>	Minimum	MIN=MATH1<>MATH2

b) de comparaisons

Syntaxe	Signification
= ou EQ	égal à (EQual)
^= ou NE	différent de (Not Equal)
> ou GT	supérieur strictement à (Greater Than)
< ou LT	inférieur strictement à (Less Than)
>= ou GE	supérieur ou égal à (Greater or Equal)
<= ou LE	inférieur ou égal à (Less or Equal)

Exemples

IF NOM='MARTIN' THEN... (sélectionne tous les individus dont la variable NOM est égale à MARTIN)

IF NOTMAT >=10 THEN... (sélectionne tous les individus qui ont la moyenne en maths)

c) logiques

Syntaxe	Signification
AND	et
OR	ou
NOT	non

d) divers

IN permet de simplifier des tests: `IF num IN (1,3,5) THEN ...`remplace la suite d'instructions: IF num=1 THEN... IF num=3 THEN... etc...

: (deux points) Comparaison de chaines de caractères.
Si `:` suit l'opérateur, la comparaison entre deux chaines de caractères va s'effectuer sur les deux chaines dont on ne gardera que le même nombre de caractère. (Ce nombre étant le minimum entre les deux longueurs). Par exemple:

`IF NOM='M' THEN` sélectionne tous les individus dont le nom est égal à M
`IF NOM=:'M' THEN` sélectionne tous les individus dont le NOM commence par M
`IF NOM=:'MA' THEN`... sélectionne tous les individus dont le NOM commence par MA.

2. Les fonctions

a) Statistiques usuelles

Nom	Signification
CSS	Somme des carrés corrigée $\sum (x_i - \bar{x})^2$
CV	Coefficient de variation (STD/MEAN)
KURTOSIS	Kurtosis (Ce coefficient mesure l'aplatissement d'une distribution. Un nombre négatif indique généralement une distribution plus pointue qu'une distribution normale; inversement, un nombre positif indique une distribution plus aplatie qu'une distribution gaussienne) Formule: $\frac{n(n+1)}{(n-1)(n-2)(n-3)} \sum z_i^4 - 3\frac{(n-1)^2}{(n-2)(n-3)}$ avec $z_i = \frac{x_i - \bar{x}}{s}$ (variable centrée réduite)
MAX	La plus grande valeur
MIN	La plus petite valeur
MEAN	La moyenne arithmétique
N	Nombre de non-manquants
NMISS	Nombre de manquants
RANGE	L'étendue (MAX-MIN)
SKEWNESS	Skewness (Ce coefficient mesure l'asymétrie d'une distribution. Une valeur plus grande ou plus petite que zéro indique une distribution disymétrique) Formule: $\frac{n}{(n-1)(n-2)} \sum z_i^3$ avec $z_i = \frac{x_i - \bar{x}}{s}$ (variable centrée réduite)
STD	Déviation standard $\sqrt{VAR}$
STDERR	Erreur standard $STD / \sqrt{n}$
SUM	Somme
USS	Somme des carrés $\sum x_i^2$
VAR	Variance $1/(n-1)\sum (x_i - \bar{x})^2 = CSS/(n-1)$

Exemple d'utilisation

```
data moy;
  set sasuser.stid193;
  keep hg math fra moy ;
  moy=mean(hg,mat,fra);
run;

proc print;
var moy;
run;
```

Un nouveau tableau de données WORK.MOY est créé...
à partir du fichier STID193 de SASUSER...
en ne conservant que les variables hg, math, fra et moy
cette dernière étant la moyenne des trois notes

Affiche le dernier fichier de données créé...
Uniquement avec la variable Moy

b) Probabilistes

Fonctions de répartition	Signification	Fonction Inverse
POISSON(λ,m)	$P(X \leq m)$ où X suit une loi de Poisson de paramètre λ	
PROBBETA(x,a,b)	$P(X \leq x)$ où X suit une loi Beta(a,b)	BETAINV(p,a,b)
PROBBNML(p,n,m)	$P(X \leq m)$ où X suit une loi binomiale B(n,p).	
PROBCHI(x,ddl)	$P(X \leq x)$ où X suit une loi du χ^2 à ddl degrés de liberté	CINV(p,ddl)
PROBF(x,nuddl,deddl)	$P(X \leq x)$ où X suit une loi de Ficher-Snedeccor à *nuddl* degrés de liberté pour le numérateur et *deddl* pour le dénominateur.	FINV(p,nuddl,deddl)
PROBGAM(x,a)	$P(X \leq x)$ où X suit une loi Gamma de paramètres (a,1)	GAMINV(p,a)
PROBNORM(x)	$P(X \leq x)$ où X suit une loi normale N(0,1)	PROBIT(p)
PROBT(x,ddl)	$P(X \leq x)$ où X suit une loi de Student à ddl degrés de liberté.	TINV(p,ddl)

Exemple n°1:

Je veux calculer le fractile d'ordre 0.95 d'une loi du χ^2 à 20 degrés de liberté. (ou encore x? tel que $P(X \leq x) = 0.95$ où X est une VA suivant un χ^2 à 20 ddl)

```
data _null_;
x=CINV(0.95,20);
put x;
run;
```

null pour que les données ne soient pas conservées en mémoire.
calcule la valeur voulue.
et l'affiche dans la fenêtre LOG.
lance l'exécution.

Exemple n°2:

Je veux calculer P(X=3) où X suit une loi binomiale B(10,0.57). En remarquant que $P(X \leq 2) + P(X = 3) = P(X \leq 3)$(loi discrète),cette valeur est en fait $P(X \leq 3) - P(X \leq 2)$. D'où:

```
data _null_;
x=PROBBNML(0.57,10,3)-PROBBNML(0.57,10,2);
put x;
run;
```

Génération de nombres aléatoires	**Génère une réalisation d'une VA suivant une ...**
RANBIN(seed,n,p)	loi binomiale B(n,p)
RANCAU(seed)	loi de cauchy de paramètres 0,1.
RANEXP(seed)	loi exponentielle de paramètre 1
RANGAM(seed,a)	loi Gamma de paramètres a,1.
RANNOR(seed) (ou NORMAL)	loi normale N(0,1)
RANPOI(seed,m)	loi de Poisson de paramètre m
RANUNI(seed) (ou UNIFORM)	loi uniforme U[0,1]

Note: seed (graine en anglais) vous permet d'obtenir les mêmes séquences de nombres aléatoires (avec la même graine initiale). Si vous n'uilisez pas cette possibilité, **choisissez pour seed la valeur 0** (la graine est déterminée à partir de l'heure et de la date) Voir l'exemple suivant.

Exemple: Générez 1000 réalisation d'une $N(2,3^2)$.

```
DATA alean23;
 DO I=1 TO 100;
   x=RANNOR(12)*3+2;
   OUTPUT;
 END;
 KEEP x;
RUN;
PROC PRINT;
RUN;
```

Un fichier temporaire WORK.ALEAN23 est créé.
Nous allons répéter 100 fois un groupe d'instruction (FOR du Pascal)
x est une réalisation d'une $N(2,3^2)$
Nous plaçons cette valeur dans le fichier
La boucle est terminée.
Nous ne conservons que la variable x dans le fichier(i est inutile)

Remarque: Si vous exécutez 2 fois cette séquence, vous obtiendrez la même série de nombre (car la graine est imposée à 12. Par contre, si vous mettez 0, les deux séquences obtenues seront différentes. Essayez !

c) Mathématiques

Syntaxe	Signification
ABS(x)	Valeur absolue de x
ARCOS(x)	Arccosinus(x)
ARSIN(x)	Arcsinus(x)
ATAN(x)	Arctangente(x)
CEIL(x)	Retourne le plus petit entier supérieur ou égal à l'argument (Ceil(3.2)vaut 4)
COS(x)	Cosinus
COSH(x)	Cosinus hyperbolique
DIGAMMA(x)	La dérivée du logarithme de la fonction Gamma
EXP(x)	Exponentielle de x
FLOOR(x)	Retourne le plus grand entier inférieur ou égal à l'argument (Floor(3.2) vaut 3)
GAMMA(x)	Fonction Gamma
INT(x)	Partie entière de x
MAX(x1,x2,...,xn)	Retourne la plus grande valeur
MIN(x1,x2,...,xn)	Retourne la plus petite valeur
MOD(x,y)	Reste dans la division de x par y
LGAMMA(x)	Lorarithme népérien de la fonction Gamma
LOG(x)	Logarithme népérien de x
LOG10(x)	Logarithme décimal de x
ROUND(x,y)	arrondi x à y près ex: Round(223.456,0.01) donne 223.46
SIGN(x)	Signe de x (-1 si négatif, 1 si positif, 0 si nul)
SIN(x)	Sinus(x)
SINH(x)	Sinus hyperbolique
SQRT(x)	Racine carrée
TAN(x)	Tangente(x)
TANH(x)	Tangente hyperbolique
TRIGAMMA(x)	La dérivée de DIGAMMA.

d) Date et heure

Syntaxe	Signification
DAY(date)	Retourne le jour d'une date donnée
MONTH(date)	Retourne le mois d'une date donnée
YEAR(date)	Retourne l'année d'une date donnée
WEEKDAY(date)	Retourne le jour de la semaine d'une date
MDY(month,day, year)	Retourne une date SAS correspondant au jour/mois/année cités en argument.
MINUTE(time)	Retourne les minutes d'une heure donnée
HOUR(time)	Retourne les heures d'une heure donnée
SECOND(time)	Retourne les secondes d'une heure donnée
HMS(hour,minute ,second)	Retourne une heure SAS correspodant aux heures minutes et secondes citées en argument.
TODAY( )	Retourne la date courante. Ne rien mettre entre les parenthèses.
TIME( )	Retourne l'heure courante

Voir l'annexe sur les Formats-Informats pour la manipulation des dates.

C. Format et Informat

1. Formats

SAS dispose d'environ 70 formats d'affichage pour les variables. (nombre, date, heure, chaîne de caractère...Si vous ne trouvez pas votre bonheur, vous pouvez toujours définir vos propres formats, informats, masques d'affichage en utilisant PROC FORMAT (cf. ce document)

Pour modifier le format d'une variable, il y a plusieurs possibilités. Si votre fichier de données existe déjà, vous pouvez utiliser la procédure DATASETS ou PRINT (sous commande Format). Si vous êtes en train de créer votre fichier de données, vous pouvez spécifier directement le format dans la commande DATA servant à créer ce fichier. Le format sera alors toujours attaché à la variable.

Pour connaître les formats en cours des variables de vos fichiers, il suffit d'exécuter une procédure Contents (Datasets).

Exemple simple:

```
data work.essai;
 input x;
 format x 7.1;
 cards;
 1528
 1255.325
 15255.3
 1228.62
 1148.5
run;
Proc print;
run;
```

donnera la sortie:

```
OBS          X
 1      1528.0
 2      1255.3
 3     15255.3
 4      1228.6
 5      1148.5
```

Le format (7.1) spécifié signifie que les nombre doivent être représentés sur 7 caractère dont une décimale.

Un **format 10.4** donnerait:

```
OBS              X
 1      1528.0000
 2      1255.3250
 3     15255.3000
 4      1228.6200
 5      1148.5000
```

Liste de quelques formats

Signification des symboles utilisés

« w » est un nombre entier donnant la largeur maximale de la représentation du nombre. Si le nombre est omis, juste la place nécessaire sera allouée.

« d » est un nombre entier égal au nombre de décimales (pour une variable numérique)

Le format de l'exemple précédent a pour abréviation: « **w.d** »

Formats de nombres:

Format (syntaxe)	Valeur non formattée	Format (pour notre exemple)	Valeur formattée	Remarque
w.d	125.236	6.2	125.24	C'est le format standard.
BESTw.		best3.		Sas cherche le meilleur format.
BINARYw.	123	binary.	01111011	Passage en binaire.
COMMAw.d	23451.23	comma10.2	23,451.23	Séparateur des milliers (version US)
COMMAXw.d	23451.23	commax10.2	23.451,23	Séparateur des milliers (version française !)
DOLLARw.d	1254.71	dollar10.2	$1,254.71	
DOLLARXw.d	1254.71	dollarx10.2	$1.254,71	
Ew.	1257	e10.	1.257E+03	Notation scientifique
FRACTw.	0.666666666667 0.2784	fract4. fract.	2/3 174/625	Convertion sous forme fractionnaire
HEXw.	88	hex8.	00000058	Conversion en base 16 (hexadécimal)
OCTALw.	3592	octal6.	007010	Conversion en base 8 (octal)
PERCENTw.d	0.1 1.2 -0.05	percent10. percent10. percent10.	10% 120% (5%)	Convertit en %, le nombres négatifs sont entre parenthèses.
ROMANw.	1992	roman10.	MCMXCII	Convertit en chiffres romains !
WORDFw.	2	wordf15.	two	Convertit les nombres en lettres

Exemple:

Si vous voulez convertir 1995 en chiffres romains, il suffit de taper le programme suivant et de regarder la LOG

```
data essai;
x=1995;
put x roman10.;
run;
```

Vous pouvez alors voir que 1995=MCMXCV !!!

Et en lettres:

```
data essai;
x=1995;
put x wordf60.;
run;
```

on obtient alors:

1995= one thousand nine hundred ninety-five and 00/100 !!!

Formats de dates, heures...

Préliminaires sur les dates, heures et « dateheure » SAS.

Dates

Les dates sous SAS sont codées sous forme numériques. Le 0 correspond au 1/1/1960 le 1 au 2/1/1960 etc... Les nombres négatifs correspondent aux dates antérieures au 1/1/1960.

Ainsi le 26/7/1989 est codé 10799 par SAS. (Il y a donc 10799 jours entre le 1/1/1960 et le 26/7/1989 !)
Il est naturellement possible d'afficher les dates convenablement grâce qux formats qui suivent.

Les dates sur deux chiffres sont codées à partir de 1900. 56 signifie 1956. Pour modifier cela, en rapport avec l'an 2000, voir l'option YEARCUTOF de l'instruction OPTIONS. CF. cette annexe.

Heures

Les heures sous SAS sont codées de la même façon en secondes à partir de minuit. Ainsi 9H30 et codé 34200. (Il y a 34200 secondes entre minuit et 9H30 !)

Date et heure

Enfin SAS possède un système spécial de codage d'une date et d'une heure dans un même nombre. C'est le nombre de secondes entre le 1/1/1960 minuit et votre date et heure. Ainsi, le 5/6/1989 9H30 est codé 928661400.

Format (syntaxe)	Valeur non formattée	Format (pour notre exemple)	Valeur formattée	Remarque
DATEw.	10847	date5. date7. date9.	12SEP 12SEP89 12SEP1989	
DATETIMEw.d	10847	datetime7. datetime12. datetime18.	12SEP89 12SEP89:03 12SEP89:03:19:43	Les chiffres apparaissant après la date sont les heures minutes secondes.
DAYw.	10919	day2.	23	C'est le jour du mois
DDMMYYw.	11316 11316 11316	ddmmyy5. ddmmyy6. ddmmyy8.	25/12 251290 25/12/90	Format européen classique.
DDMMYYxw.	11316 11316	ddmmyyp8. ddmmyyd8.	25.12.90 25-12-90	x peut prendre les valeurs : B pour un blanc. C pour une virgule D pour un tiret N pour rien P pour un pont. S pour une barre de fraction.
DOWNAMEw.	10621 10621	downame.	Sunday	Retourne le jour de la semaine
HHMMw.d	46796	hhmm.	13:00	
HOURw.d	41400 41400 41400	hour4.1	11.5	Ecrit l'heure et la fraction décimale de l'heure. (11.5=11H30)
MMDDYYw. Et MMDDYYxw.	10847 10847 10847	mmddyy5. mmddyy6. Mmddyy8.	09/12 091289 09/12/89	Attention, le mois est avant le jour ! (format courant américain) Le x peut prendre les valeurs : B pour un blanc. C pour une virgule D pour un tiret N pour rien P pour un pont. S pour une barre de fraction.
MMSSw.d	4530	mmss.	75:30	Convertit une variable horaire en minute et seconde après minuit.
MMYYxw.	29may1989	mmyy5. mmyyc7. mmyyd7. mmyyp6. mmyys7.	05M89 05:1989 05-1989 05.89 05/1989	Formate une date en mois et année séparées par un caractère.
MONNAMEw.	10919	monname9.	November	Affiche le mois en lettre d'une date.
MONTHw.	10919	month.	11	Affiche le mois en chiffre d'une date.
MONYYw.	10750	monyy7.	JUN1989	Affiche le mois et l'année d'une date.
QTRw.	10741	qtr.	2	Affiche le n° du trimestre correspondant à la date en question.
QTRRw.	10897	qtr.	IV	Idem mais en chiffres

				romains !
TIMEw.d	59083	time.	16:24:43	
WEEKDATEw.	10848 10848 10848 10848	weekdate3. weekdate9. weekdate15. weekdate17.	Wed Wednesday Wed, Sep 13, 89 Wed, Sep 13, 1989	
WEEKDATXw.	10869	weekdatx.	Wednesday, 4 October 1989	
WEEKDAYw.	10621	weekday.	1	Jour de la semaine d'une date (1=dimanche, 2=lundi etc...)
WORDDATEw.	11212	worddate3. worddate9. worddate12. worddate18.	Sep September Sep 12, 1990 September 12, 1990	
YEARw.	11212	year4.	1990	
YYMMxw.	11212	yymmc7.	1990:09	
YYMMDDw. Et YYMMDDxw.	11212	yymmdd8.	90-09-12	 x peut prendre les valeurs : B pour un blanc. C pour une virgule D pour un tiret N pour rien P pour un pont. S pour une barre de fraction.
YYMONw.	10621	yymon7.	1989JAN	
YYQxw.	11212	yyqc6.	1990:3	Année suivie du numéro de trimestre
YYQRxw.	11212	yyqrc8.	1990:III	Idem mais le numéro de trimestre est en chiffres romains.

Exemples d'utilisation:

Avec la date...

Une réunion est prévue le 27 janvier 1995 et je souhaite envoyer un courrier 45 jours avant. A quelle date cela correspond-il ?

```
data _null_;
date='27JAN1995'D;
envoi=date-45;
put envoi DDMMYY8.;
run;
```

Le D indique à SAS que la chaine de caractères est une date.

Donnera dans la Log: 13/12/94.

avec l'heure...

Ici on va retire 10 minutes à 10:03 et afficher le résultat.

```
data _null_;
heure='10:03'T;
envoi=heure-600;
put envoi TIME10.;
run;
```

On initialise l'heure. Le T signale à SAS qu'il va lire une heure.

On retire 10 minutes (=600 secondes).

On obtient alors: 09:53:00.

avec la date et l'heure en même temps

Supposons que l'on veuille connaître le nombre de minutes entre le 1/5/1966 4H et le 27/04/1995 09:26. (Problème complètement stupide je vous l'accorde !)

```
data _null_;
date1='01may1966:04:00'DT;
date2='27apr1995:09:27'DT;
diff=(date2-date1)/60;
put diff;
run;
```

DT indique à SAS que l'on va lire une Date+Heure

Avril=April in English !

La différence est en secondes...

(On trouve 15247047 minutes !)

Exercices

I) Calculez votre age en années puis en heures (si vous connaissez votre heure de naissance.

II) Quel jour (de la semaine) êtes-vous nés ?

Complément

Il est possible de créer ses propres formats et de les appliquer ensuite à des variables adéquates. Pour en savoir plus, allez consulter le paragraphe PROC FORMAT de ce document.

2. Les Informats

Ils sont utilisés pour lire des données qui ont un format spécial dans un fichier texte externe. Vous pouvez en définir de nouveaux avec la procédure FORMAT.

Exemple

Le fichier STID93 contient une variable « date de naissance » que SAS considère (pour l'instant) comme chaîne de caractères. Aucun calcul n'est donc possible dessus.

Grâce à un Informat, nous pouvons préciser à SAS ce qu'il va lire (une date) et dans quel ordre seront les éléments qui la constitue.

Voici un exemple en supposant les données incluses dans le programme.

```
data work.age;

 input datnaiss ddmmyy10. ;

 aujour=today();
 agejour=aujour-datnaiss;

 agean=int(agejour/365.25);


 format datnaiss date.;
 format aujour date.;
  cards;
01/05/1966
23/10/1975
23/05/1944
;
run;
proc print;
run;
```

Nous indiquons à SAS qu'il va lire une date dans les 10 premiers caractères avec dans l'ordre le jour, le mois, l'année.

Donne le jour d'aujourd'hui.

Nous calculons l'age en jours.

Nous calculons l'age en années.
Nous prenons la partie entière

Nous donnons des formats d'affichages...
Les dates apparaitront en clair.

Nous obtenons l'affichage:

OBS	DATNAISS	AUJOUR	AGEJOUR	AGEAN
1	01MAY66	27APR95	10588	29
2	23OCT75	27APR95	7126	20
3	23MAY44	27APR95	18601	51

Avec une lecture en colonnes

Nous supposons que le fichier STID193 (format texte) contient la date de naissance des individus sur les colonnes 8 à 17 sous la forme jj/mm/aaaa. (exemple 20/06/1975.)

Le programme suivant réalise l'importation des dates.

```
data work.naiss;
infile 'z:\public\stid193.txt' missover firstobs=2;

input @8 datnaiss ddmmyy10.;
```

C'est la ligne clé ! Le @8 demande à SAS de lire à partir de la colonne 8. Le ddmmyy10. signale que la date est sur 10 caractères dans l'ordre jour mois an.

```
format datnaiss date.;
```

Pour permettre l'affichage 'en clair' de la date dans le Proc Print ci-dessous.

```
run;
proc print;
run;
```

L' affichage donne:

OBS	DATNAISS
1	21OCT73
2	08DEC74
3	15AUG72
4	10NOV72

Quelques Informats

Syntaxe	Valeur à lire	Informat utilisé pour la lire	Résultat de la lecture	Remarques
$w.	Marie-Christine	$10.	Marie-Chri	Lit une chaîne de longueur w. (8 par défaut)
$HEXw.	6C6C	$hex4.	11	Convertit une chaîne hexadécimale en chaîne de caractères.
$COMMAw.d	$1,000,000 (500)	comma10. comma10.	1000000 -500	Enlève les , () $ des nombres.
COMMAXw.d.	$1.000.000	commax10.	1000000	Idem en changeant les rôles des , et les .
DATEw.	1jan1990 01 jan 90 1 jan 90 1-jan-1990	date10.	10958 10958 10958 10958	
DATETIMEw.	23dec89:10:03:17	datetime20.	946029797	
DDMMYYw.	231090 23/10/90 23 10 90	ddmmyy8.	11253 11253 11253	Lecture de dates (cf exemple ci-dessus)
Ew.d	1.257E3	e7.	1257	Lecture de nombres en notation scientifique.
HEXw.	88F	hex3.	2191	
MMDDYYw.	010190 1/1/90 01 1 90	mmddyy8.	10958 10958 10958	
MONYYw.	jun89	monyy5.	10744	
PERCENTw.d	1% (20%)	percent3. percent5.	0.01 -0.2	Conversion de %
TIMEw.	14:22:25	time8.	51745	Lecture d'heures minutes secondes.
YYMMDDw.	900101 90-01-01	yymmdd8.	10958 10958	

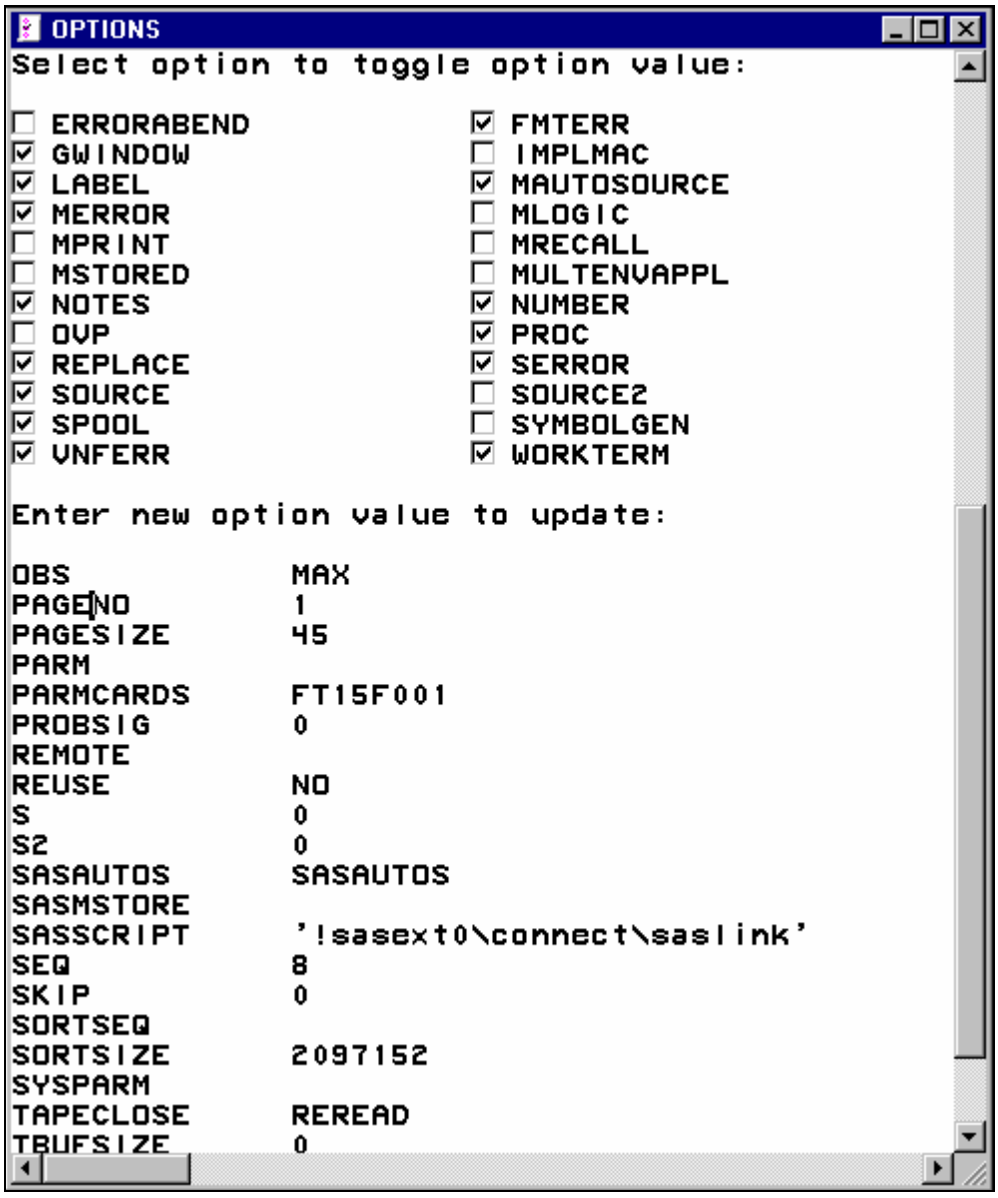
D. Commande ou fenêtre OPTIONS en langage SAS

Cette instruction ou fenêtre permet de gérer l'affichage dans la fenêtre OUTPUT, d'afficher le contenu des variables macros ou des programmes et bien d'autres choses encore.

1. La fenêtre d'options

Elle est activée en faisant un CTRL I ou en allant dans GLOBAL/OPTIONS/GLOBAL OPTIONS.

Vous obtenez ce qui suit :



Vous pouvez modifier les paramètres comme vous le souhaitez.

2. L'instruction

Exemples :
Dans le program Editor lancez un

```
Options linesize=70 pagesize=50 ;
```

Ceci vous permet de changer à 70 le nombre de caractères par lignes dans la fenêtre OUTPUT et à 50 le nombre de lignes par page.

```
Options nodate nonumber ;
```

Va vous permettre de supprimer l'affichage de la date et des numéros de page dans la fenêtre OUTPUT. Un « `Options date number ;` » remettra tout comme avant !

Voici quelques options parmi les principales.

	Nom	Signification
	AUTOEXEC=	Spécifie le nom du fichier à exécuter au lancement de SAS
	CENTER/ NOCENTER	Centre ou non les sorties dans l'OUTPUT
	CONFIG=	Spécifie le nom du fichier de configuration à prendre en compte sous SAS
	DATE/ NODATE	Affiche ou non la date dans l'OUTPUT
	ECHOAUTO/NOECHOAUTO	Affiche ou non l'autoexec dans la LOG lorsqu'il est exécuté.
	ERRORS=n	N est le nombre d'erreurs à afficher complêtement dans la LOG.
	FIRSTOBS=n	Obligera SAS à ne lire qu'à partir de la nieme observation dans les fichiers de données… Elle est initialisée à 1 par défaut.
	FMTERR/NOFMTERR	Spécifie si SAS doit s'arreter ou non lorsqu'il rencontre un format de variable inconnu. Dans le cas ou NOFMTERR est activée, le format défaillant est remplacé par w. ou $w.
	FORMDLIM=	Spécifie un caractère à utiliser lors des sauts de pages de SAS dans l'OUTPUT.
	FULLSTIMER/ NOFULLSTIMER	Affiche ou non des statistiques internes de SAS sur la performance du système (occupation mémoire etc) lors de l'exécution du code
	GWINDOW/ NOGWINDOW	Affiche ou non les graphiques haute résolution dans le display manager.
	IMPLMAC/ NOIMPLMAC	Autorise le compilateur SAS à vérifier s'il rencontre des macros, ce qui ralentit le temps d'exécution.
	INITCMD	Supprime l'affichage des fenêtres LOG, OUTPUT et PROGRAM EDITOR lors du lancement d'une

		application AF
	INVALIDDATA=’caractère’	Spécifie la modalité prise par défaut lorsque INPUT lit une « mauvaise » valeur. Par défaut c’est un point « . »
	LABEL / NOLABEL	Spécifie si oui ou non les procédures SAS peuvent utiliser ou non les labels
	*_LAST_*=nom de fichier SAS	Spécifie le nom du dernier fichier créé.
	LINESIZE=	Nombre de caractères par ligne.
	LOG=destination / NOLOG	Choisit une destination pour le contenu de la fenêtre LOG ou supprime la fenêtre LOG
	MACRO / NOMACRO	Spécifie si oui ou non le langage MACRO est disponible ou non.
	MAPS= bibliothèque	Bibliothèque SAS contenant les cartes du module GRAPH
	MAUTOSOURCE / NOMAUTOSOURCE	Spécifie si oui ou non l’autocall est disponible pour les macros.
	MERROR / NOMERROR	Spécifie si le système renvoie une erreur s’il rencontre un nom de macro inconnu.
	MISSING=’caractère’	Spécifie le caractère pour les valeurs manquantes. C’est un point « . » par défaut.
	MLOGIC / NOMLOGIC	Spécifie si oui ou non, le processeur de macro « trace » l’éxécution de celle ci quant aux conditions %IF etc.
	MPRINT / NOMPRINT	Affiche ou non les instructions lors de l’exécution d’une macro.
	MSGCASE / NOMSGCASE	Spécifie si les messages affichés dans les notes, warnings sont en majuscules ou non.
	MSGLEVEL= N ou I	Affiche moins ou plus d’informations dans la fenêtre LOG. N par défaut.
	MSTORED / NOMSTORED	Autorise ou non SAS à utiliser des macros compilées.
	NEWS=fichier	Contient un fichier à mettre dans la fenêtre LOG.
	NOTES / NONOTES	Spécifie si les notes sont affichées dans la LOG ou non.
	NUMBER / NONUMBER	Spécifie si SAS affiche les numéro de page dans la fenêtre OUTPUT ou non. Voir aussi PAGENO= et DATE
	OBS= numéro	Spécifie quelle observation SAS utilisera en dernier. A combiner avec FIRSTOBS=
	OVP / NOOVP	Spécifie si SAS souligne les erreurs dans la LOG ou non.
	PAGENO=	Spécifie le prochain numéro de page à indiquer dans l’OUPUT. voir aussi NUMBER
	PAGESIZE=n	Spécifie le nombre de lignes dans une page.
	RSASUSER / NORSASUSER	Spécifie si SASUSER est en lecture seule ou non.
	SASAUTOS=	Spécifie les bibliothèques autocall.
	SASMSTORE=	Spécifie le libref d’une bibliothèque SAS contenant le SASMACR catalogue.
	YEARCUTOFF=nombre	Spécifie comment SAS doit comprendre les années à deux chiffres. Par exemple avec YEARCUTOFF=1950, 33 sera 2033 mais 55 sera 1955

E. Echange dynamique de données SAS-EXCEL :Liaisons DDE

Les liaisons DDE de Windows permettent de transférer des données entre 2 applications Windows. L'application qui reçoit les données est l'application client, celle qui les envoie est l'application serveur.

Les liaisons DDE peuvent être permanentes (hotlinks) ou temporaires (coldlinks). Les deux exemples suivants sont des « coldlinks », les données ne sont transférées qu'une seule fois et la liaison est coupée.

1. Voyons un exemple de transfert SAS vers Excel

On souhaite transférer les variables Groupe, Sexe, Taille, Poids du fichier de données SAS STID193 vers les 4 premières colonnes de la feuille 1 d'Excel.

Lançons Excel et SAS.
Tapons le programme suivant dans le Program Editor :

```
Libname moi 'g :\stid9597\public\logiciel' ;
filename extrait dde 'excel|feuil1!l1c1:l106c4';
```

pour accéder au répertoire PUBLIC

Altgr 6 — Point d'exclamation.

Mise en place de la liaison DDE :
`excel:` c'est l'application concernée (Application)
`feuil1` : c'est le n° de la feuille (Topic)
`l1c1 :l106` :c4 : c'est la zone concernée par le recueil des données (Item). Il y a 106 individus dans le fichier STID193, on réserve donc 106 lignes.

```
data _null_;
  file extrait;
  set moi.stid193;
  put groupe sexe taille poids;

run;
```

Création d'un fichier de données SAS fictif
Pour orienter les PUT vers le fichier extrait
Fichier de données dont sont issues les donées
Ecriture des variables dans le fichier « extrait » donc sous Excel car se fichier est « lié » à Excel par la liaison DDE précédente.

Voici comment se présentent les première lignes de la feuille 1 d'Excel :

A	1	12	68
A	1	13	61
A	1	11.5	68
A	2	15	54
A	2	10	50
A	2	18	58
A	2	15	58

Génial non ?

Remarques : Pour terminer le travail, il faut remplacer les « . » par des « , » dans les données pour qu'Excel les reconnaissent comme des données numériques (cf. le 11.5 de la note de maths). Il faut également faire attention

aux manquants que SAS remplace par des points « . » alors qu'Excel les reconnaît par une cellule vide.

Attention : Si vous avez donné un nom à votre fichier Excel, vous devez remettre ce nom dans la définition de la liaison DDE à la rubrique TOPIC. [127]

Ainsi, si votre fichier est nommé ESSDDE.XLS, l'instruction filename sera du type :

```
filename extrait dde 'excel|essdde.xls!l1c1:l106c4';
```

2. Transfert Excel vers SAS

Lancez Excel et mettez les données suivantes dans les 3 premières colonnes de la feuille1 :

Paul	12	-8.2
Géronimo	27.5	7.582
Charles-Edouard	32	123.2513
Henri	-52152.2215	258.32

Remarque : Nous avons mis un point « . » comme séparateur décimal (SHIFT ;) sinon SAS ne reconnaîtra pas ces variables numériques.

Nous allons transférer ces données sous SAS en tapant le programme suivant :

```
filename essai dde 'excel|feuil1!l1c1:l4c3';
```

Altgr 6 — Point d'exclamation. — Mise en place de la liaison DDE :
`excel:` c'est l'application concernée (Application)
`feuil1` : c'est le n° de la feuille (Topic)
`l1c1 :l4` :c3 : c'est la zone concernée par le recueil des données (Item)

```
data moi.donnee;
 infile essai;

 input nom $ var1 var2;

run;

proc print data=moi.donnee;
run;
```

Création d'un fichier de données SAS
A partir du fichier (logique) « essai »qui est en fait composé des 3 premières colonnes et des 4 premières lignes d'excel.
On importe la première colonne sous la variable alpha ($) « nom », puis les deux autres (numériques) sous les noms « var1 » et « var2 ».
Un Proc Print pour afficher le fichier ainsi importé.

Attention : Si vous avez donné un nom à votre fichier Excel, vous devez remettre ce nom dans la définition de la liaison DDE à la rubrique TOPIC.[128]
Ainsi, si votre fichier Excel est nommé ESSDDE.XLS, l'instruction filename sera du type :

```
filename essai dde 'excel|essdde.xls!l1c1:l4c3';
```

Nous obtenons :

---

127 Sinon SAS ne le retrouvera pas (physical file doesn't exist...

128 SAS est intelligent ! En effet, quand vous lancez Excel, celui ci charge le classeur1 (class1) par défaut et si vous chargez un autre fichier nommé, vous vous retrouvez avec deux classeurs. SAS veut donc que vous nommiez votre classeur pour savoir ou faire le transfert !

```
The SAS System          09:46 Thursday, May 8, 1997

                      OBS    NOM              VAR1        VAR2

                       1     Paul            12.00       -8.200
                       2     Géronimo        27.50        7.582
                       3     Charles-        32.00      123.251
                       4     Henri        -52152.22      258.320
```

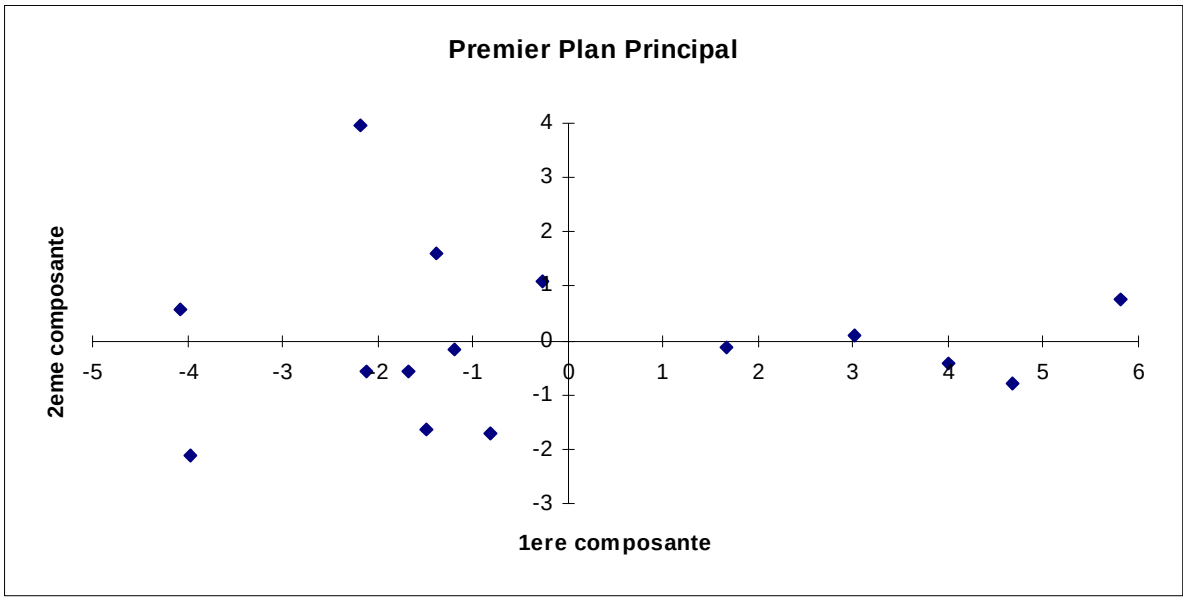
3. Applications

a) Tracé du premier plan principal sous Excel dans une ACP faite sous SAS

Nous allons charger le fichier ACP que vous connaissez bien, effectuer l'ACP sur les 12 variables de températures, calculer les qualités des différentsindividus (cf. chapitre ACP de ce document) et importer sous Excel les projections des villes sur le premier plan principal.

```
proc princomp data=moi.acp out=work.essai;
var jan fev mar avr mai jun jui aou sep oct nov dec;
run;

/* Calcul de la qualité de représentation des individus*/

proc standard data=work.essai out=work.essai mean=0 std=1;
var jan fev mar avr mai jun jui aou sep oct nov dec;
run;

data work.essai2;
 set work.essai;
 denom= USS(jan,fev,mar,avr,mai,jun,jui,aou,sep,oct,nov,dec);
 numer=USS(prin1,prin2);
 qual=numer/denom;
run;
proc print data=work.essai2;
run;

/* transfert des données NOM PRIN1 PRIN2 QUAL sous Excel*/

filename extrait dde 'excel|feuil1!l2c1:l16c4';
data _null_;
  file extrait;
  set work.essai2;
  put nom prin1 prin2 qual;
run;
```

Nous pouvons ensuite tracer le nuage de points très simplement.



Remarque : Il serait intéressant d'étiquetter les points ici, mais Excel ne met que la première série de données comme étiquette. Il faut ensuite les modifier une à une avec les noms des villes !

b) Récupération des résidus et des mesures d'influence sous Excel dans le cas d'une Regression linéaire multiple

Dans la pratique, il est souvent commode d'utiliser SAS pour effectuer les calculs et d'Excel pour les représentations graphiques.[129]

Prenez le fichier REGDDE.XLS, changez toutes les « , » en « . » pour importer ce fichier sous SAS.

Ecrivez ensuite le programme d'importation en utilisant la liaison DDE (attention ne transférez pas les noms des colonnes)

Effectuez ensuite une regression linéaire de Y en (X1,X2) et récupérez dans un fichier de données SAS le DFITS, la distance de COOK, les résidus, les Y chapeaux.

Tranférez ces nouvelles variables sous Excel dans des colonnes libres du fichier REGDDE.XLS. Changez les points en virgules. Voici les premières lignes de ce que vous devez obtenir :

X1	X2	Y	Ychapeau	Residu	Stdres	Cook	Dfits
0,4227	8,6895	37,09	-1474,69379	1511,78379	2,8356555	0,2201091	0,84345645
1,2497	99,1988	311,66	-974,307121	1285,96712	2,35805921	0,11871351	0,61154592
1,5177	63,0564	204,51	-1117,94958	1322,45958	2,40291472	0,08337637	0,51308017
4,852	94,8181	327,7	-656,520754	984,220754	1,77362171	0,05888386	0,42522574

[129] SAS possède bien des graphiques Haute résolution (GPLOT, GCHART etc...) mais leur utilisation est très lourde et les graphiques obtenus sont médiocres. De plus, ils sont en mode Bitmap ce qui alourdit les copier-coller, les impressions. Il est alors préférable de passer par PAINT pour sauvegarder ces fichiers en BMP puis les intégrer dans un document (Insérer Image) .

Faites les graphiques des valeurs résiduelles. Que pensez-vous de ce modèle ?

Pour des informations complémentaires, consultez l'indispensable « *SAS Companion for the Microsoft Windows Environment* »

F. Quelques procédures usuelles

Les procédures en gras sont détaillées dans ce document.

Nom	Module	Aperçu des possibilités
ACCESS	SAS/PC File format	Conversions de fichiers externes au format SAS. La commande FILE/IMPORT permet un accès automatique à cette procédure.
ANOVA	SAS / Stat	Analyse de variance à un ou plusieurs critères avec des données équilibrées.
APPEND	SAS / Base	Ajout d'observations contenues dans un fichier SAS à la suite d'un autre fichier SAS
CATALOG	SAS / Base	Gérer les « catalogues » SAS
CANDISC	SAS/Stat	Analyse discriminante canonique.
CHART	SAS / Base	Effectuer des diagrammes à bandes, histogrammes etc... en mode texte.
CLUSTER	SAS/Stat	Classification ascendante hiérarchique.
COMPARE	SAS / Base	Comparer de fichiers SAS
CONTENTS	SAS / Base	Lister le contenu d'une bibliothèque, les attributs d'un fichier ...
COPY	SAS / Base	Copier tout ou partie de fichiers de données SAS
CORR	SAS / Base	Calculer des coefficients de corrélation (Pearson, Spearman...)
CORRESP	SAS/Stat	Analyse factorielle des correspondances (simples et multiples)
DISCRIM	SAS/Stat	Analyse discriminante bayésienne.
DATASETS	SAS / Base	Lister, copier, supprimer, renommer des fichiers de données SAS. (les possibilités des procédures Append, Contents et Copy y sont incluses)
DBLOAD	SAS/PC File format	Conversion d'un fichier SAS au format DBASE... FILE/EXPORT per met de faire ce travail en étant assisté.
FORMAT	SAS / Base	Définir des « formats » et des « informats » pour les variables numériques et alphanumériques
FORECAST	SAS/ETS	Prévision pour les séries temporelles. La commande FORECAST est présentée dans ce document.
FREQ	SAS / Base	Tris à plat, tri croisés pour des variables qualitatives. Tests du Chi 2...
FSEDIT	SAS/FSP	Création de masque de saisie pour créer de nouveaux fichiers de données.
FSVIEW	SAS/FSP	Visualisation / correction de fichiers de façon interactive.
GLM	SAS / Stat	Analyse de variance à un ou plusieurs critères avec des données équilibrées ou non. (Il est préférable d'utiliser Anova avec des données équilibrées)
LOGISTIC	SAS/Stat	Régression logistique.
MEANS	SAS / Base	Produire des statistiques élémentaires sur une variable numérique. (Univariate est plus puissante)
NPAR1WAY	SAS / Stat	Effectuer des tests non paramétriques (Mann-Whitney ...)
OPTIONS	SAS / Base	Lister les options en cours de SAS.
PLOT	SAS / Base	Créer des nuages de points en mode caractère. (GPLOT pour la haute résolution)
PRINCOMP	SAS / Stat	Effectuer des analyses en composantes principales. (ACP)
PRINT	SAS / Base	Affichage de tout ou partie d'un fichier.
RANK	SAS / Base	Calculer les rangs des observations d'une ou plusieurs variables numériques.
REG	SAS / Stat	Effectuer des régressions. De nombreuses options sont disponibles. (PROC LOGISTIC pour la régression logistique)
SORT	SAS / Base	Trier un fichier selon un ou plusieurs critères. (C'est un préliminaire indispensable pour l'utilisation de BY avec d'autres procédures)

STANDARD	SAS / Base	Centrer réduire (entre autres) des variables quantitatives.
SUMMARY	SAS / Base	Similaire à Means à qq détails près...utilisée pour récupérer le nombre d'individus, les moyennes varainces etc. pour refaire d'autres calculs derrière.
TIMEPLOT	SAS / Base	Représenter graphiquement des séries chronologiques en mode caractère.
TRANSPOSE	SAS / Base	Pour « transposer » la matrice des données. Les variables deviennent les observations et vice versa.
TTEST	SAS / Stat	Tester l'égalité de deux moyennes (Student), tester l'égalité de deux variances (Fisher). (Pour comparer une moyenne à un standard, utilisez Univariate)
UNIVARIATE	SAS / Stat	Résumé statistique pour une variable quantitative. Test de normalité (Shapiro Wilk) et test de nullité de moyenne (Student) inclus.

Remarque: Pour connaître la syntaxe complète d'une procédure, consultez l'aide en ligne.

G. Execution d'un FICHIER DE COMMANDES SAS depuis le DOS

Il vous est possible de lancer SAS directement depuis le DOS en utilisant l'instruction suivante:

```
win c:\sas\sas.exe -config c:\sas\config.sas
```

Si vous souhaitez en plus que SAS execute un de vos programmes SAS construit antérieurement *monprog.sas* par exemple, il faut introduire l'instruction *-sysin*:

```
win c:\sas\sas.exe -config c:\sas\config.sas -sysin
c:\sas\monprog.exe
```

Vous verez alors Windows se lancer et une petite fenêtre s'ouvrir signalant que SAS execute votre programme et que le contenu de la fenêtre LOG sera dans le fichier *monprog.log* et celui de la fenêtre OUTPUT dans le fichier *monprog.lst*.

Il vous est possible de modifier la largeur de la page de la fenêtre OUTPUT grâce à *-linesize* ou *-ls*. Exemple `-ls 70` à la suite des instructions précédentes imposera à SAS 70 caractères dans la fenêtre OUTPUT.

De même *-pagesize*, ou *-ps* modifie le nombre de ligne par page.

Pour avoir la liste des options, vous pouvez consulter le *SAS Companion for the Microsoft Windows Environment* pp159ss

H. Importation de fichiers ayant un format connu PROC IMPORT

C'est l'approche langage SAS de l'importation de fichier EXCEL, DBASE, CSV etc. effectuée avec FILE/IMPORT.

Pour importer les fichier de type EXCEL, DBASE etc., vous devez disposer du module ACCESS To Pc File Formats.

Syntaxe simplifiée

```
PROC IMPORT DATAFILE='chemin et nom du fichier de données'
            OUT=nom de fichier SAS
            DBMS=type de fichier <REPLACE> ;
            <options selon type du fichier> ;
RUN ;
```

DBMS peut prendre les valeurs : XLS, CSV, TXT, DLM

REPLACE : remplace un fichier existant. Si REPLACE n'est pas spécifié, IMPORT n'écrasera pas un fichier existant.

Les *options selon type du fichier* étant :

GETNAMES=Yes ou No
: Si la première ligne de votre fichier de données comporte les noms des variables ou non.

RANGE=
: Délimite la zone à importer. Tout sera importé RANGE est omis.

SHEET=
: Identifie la feuille à importer.

Exemple

```
proc import datafile='D:\data\bsclient.csv'
out=work.bsclient dbms=CSV replace;
            getnames=yes;
run;
```

Va importer le fichier BSCLIENT au format CSV sous SAS.

I. Exportation de fichiers PROC EXPORT

C'est la réplique exacte de PROC IMPORT mais pour l'exportation de fichiers. Ce que nous allons voir fait la même chose que PROC EXPORT mais en utilisant le langage SAS. L'import export de fichiers peut donc être géré avec les MACROS ou les programmes en SCL.

Syntaxe simplifiée

```
PROC EXPORT DATAFILE= nom de fichier SAS
            OUTFILE= 'chemin et nom du fichier de données'
            DBMS=type de fichier <REPLACE> ;
            <options selon type du fichier> ;
RUN ;
```

DBMS peut prendre les valeurs : XLS, CSV, TXT, DLM

REPLACE : remplace un fichier existant. Si REPLACE n'est pas spécifié, EXPORT n'écrasera pas un fichier existant.

Les *options selon type du fichier* étant :

DELIMITER='*caractère*'
Identifie le délimiteur qui va séparer les colonnes de données. Par défaut c'est un espace.

Exemple

```
proc export data=moi.bordeaux outfile='D:\data\bordeaux.csv'
DBMS=CSV replace;
run;
```

Va exporter la table SAS BORDEAUX en un fichier Bordeaux.Csv dont voici les 3 premières lignes :

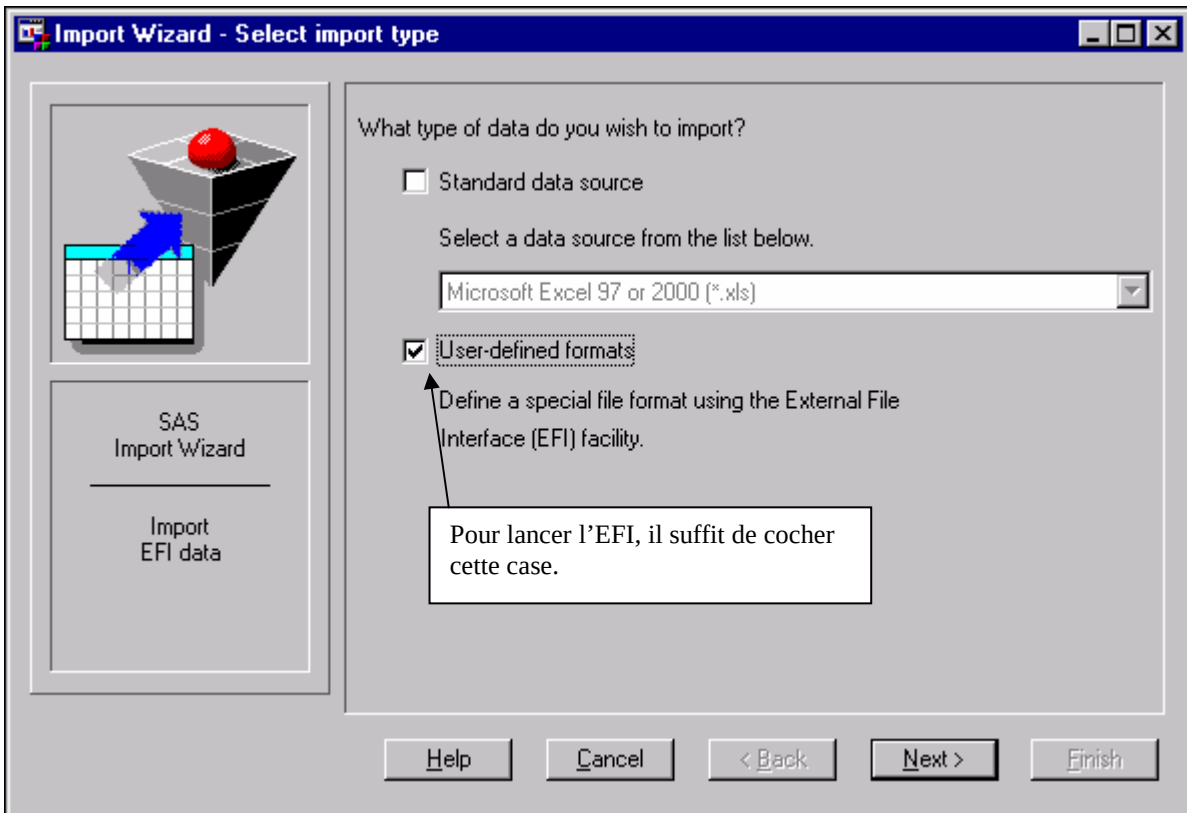
```
ANNEE,X1,X2,X3,X4,QUAL
1924,3064,1201,10,361,2
1925,3000,1053,11,338,3
```

J. Complément : Données importées d'un fichier texte ASCII externe

Pour importer des fichiers ASCII particuliers, vous disposez de l'EFI ou de l'instruction INFILE.

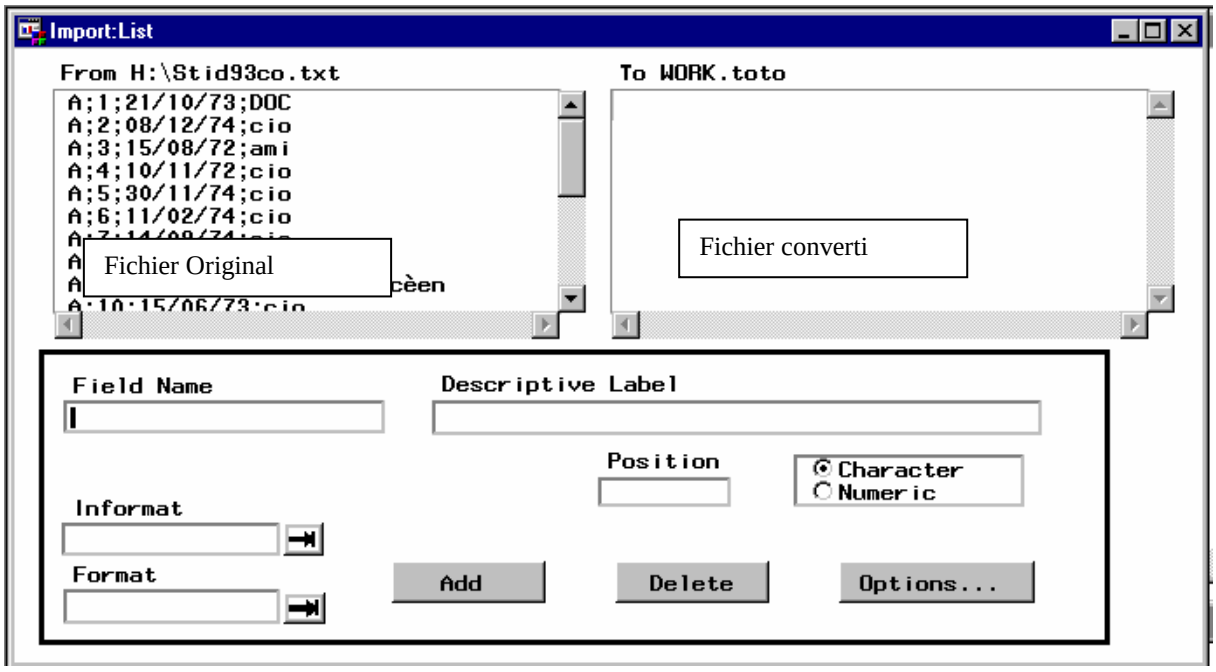
1. L'EFI

- Nous allons importer le fichier texte : STID193CO
- Faites un file/Import. Cocher la case EFI. Validez.



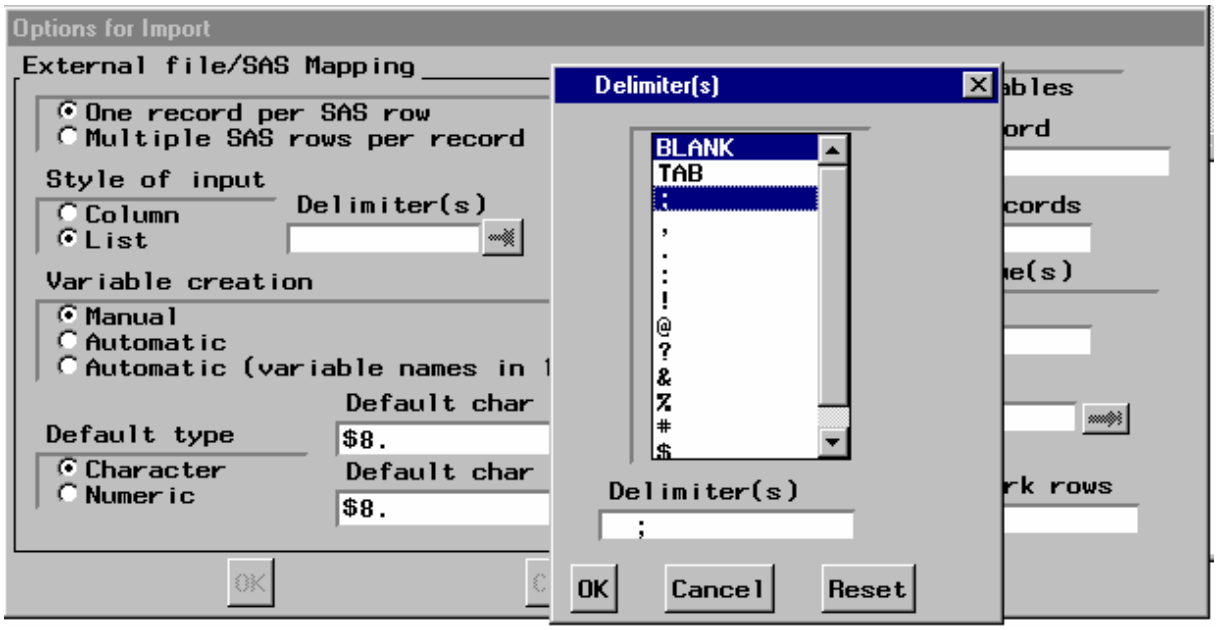
- Sélectionnez le fichier STID193CO.TXT dans PUBLIC. Validez.
- Indiquez un fichier de sortie SAS à votre convenance. Validez

La fenêtre suivante s'affiche :

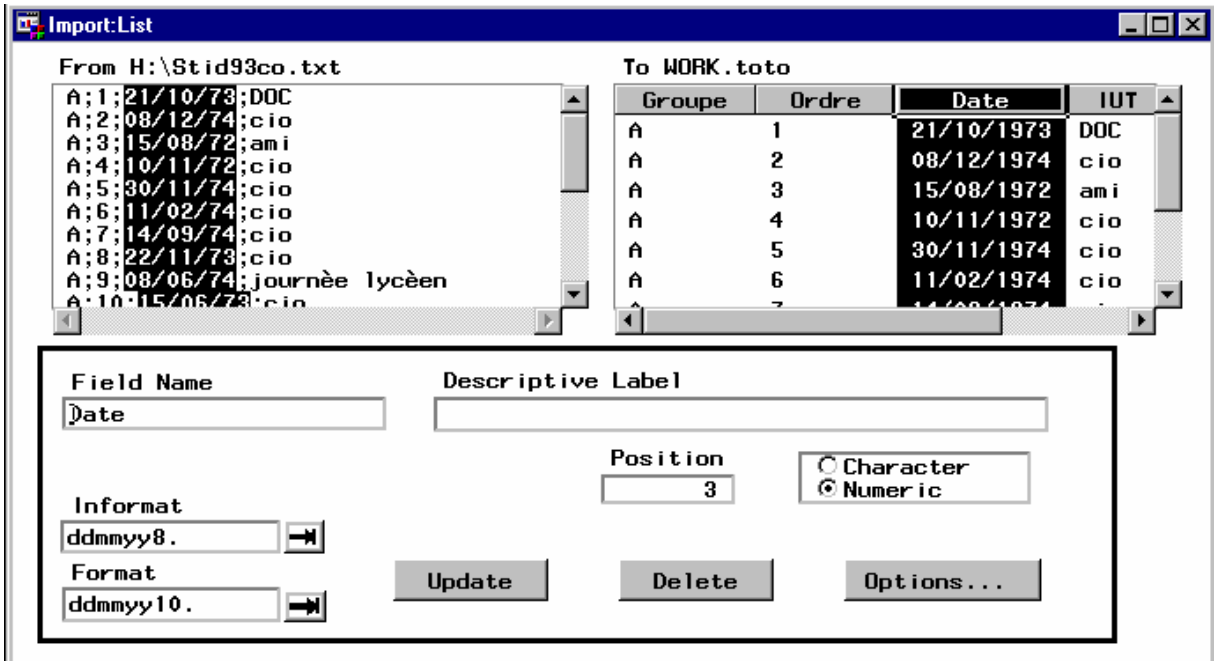


Nous voyons que le « ; » sert de délimiteur entre les variables. Nous allons l'indiquer à SAS en cliquant sur le bouton OPTIONS.

Choisissez ; comme delimiteur de fichier :



Validez.



Ensuite, cliquez sur une colonne du fichier TEXTE, donnez lui un Nom (nom de variable) dans Field Name et cliquez sur ADD.

Faites ceci pour le GROUPE et l'ORDRE.
Pour la DATE, mettez la en numérique, donnez lui comme INFORMAT DDMMYY8. Et comme FORMAT ce que vous voulez. Cf. paragraphe sur les formats.

Faites un File/Save pour entérinner vos changements.

La LOG doit contenir :

```
NOTE: WORK.STID193CO was successfully created.
```

2. Instruction INFILE : Syntaxe simplifiée

Si vous avez un fichier récalcitrant, mettez le en type texte et utilisez les instructions suivantes.

Le seul changement par rapport à ce qui précède est l'instruction INFILE qui donne l'adresse du fichier texte à importer suivi d'options. Bien entendu CARDS a disparu !

```
DATA nom fichier SAS ;
 INFILE 'nom du fichier texte' options ;
 INPUT var1 type1 var2 type2 .... var n type n ;
 instructions complémentaires
RUN ;
```

Exemple

Supposons que vos données soient dans un fichier texte (ASCII) externe. Par exemple, voici le contenu du fichier ENTREE.TXT:

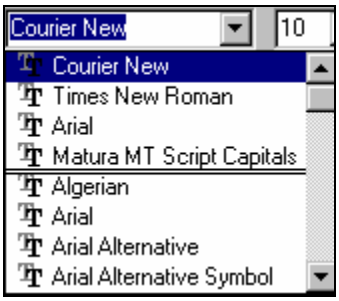
```
ERIC M 190 78
MARC M 178.5 75
CLAIRE F . 59
AURELIE F 168 62
JEAN M 168 59
```

(Note: Les conventions sont les mêmes que précédemment: un espace au moins entre chaque donnée, un point pour les données manquantes)

Editez le fichier Entree.txt(G:\STID9799\PUBLIC\LOGICIEL) sous Word et vérifiez que les données sont correctes. Fermez le document. (SAS ne lit pas des fichiers ouverts par une autre application)

Remarque: Sous Word, soyez en Police Courier New et faites afficher en clair les espaces et les tabulations (OUTILS/OPTION/AFFICHER/Tabulations Espaces)

ATTENTION: Les tabulations faites sous Word ne sont pas considérées par SAS comme des espaces. Ainsi une tabulation est considérée comme un seul caractère par SAS alors qu'elle occupe la place de plusieurs sous Word.




```
DATA WORK.ENTREE;
INFILE 'Z:\PUBLIC\LOGICIEL\ENTREE.TXT' ;
INPUT NOM $ SEXE $ ETC...;
RUN ;

PROC PRINT DATA=WORK.ENTREE;
RUN;
PROC MEANS DATA=WORK.ENTREE;
CLASS SEXE;
RUN;
```

Pour visualiser le fichier importé

Pour faire quelques statistiques... en distinguant selon le sexe.

Tapez le programme SAS et vérifiez que l'importation s'est bien déroulée.

3. Données ou fichiers inhabituels

Dans le fichier précédent, les observations étaient sur une ligne. (Il y avait un individu par ligne). SAS permet également d'importer des fichiers textes lorsqu'il y a une observations sur plusieurs lignes ou plusieurs observations sur une ligne. Il est aussi possible de changer de délimiteur (l'espace par défaut).

a) Cas où il y a plusieurs observations par ligne

Le fichier TAILLE.TXT contient les données suivantes:

```
Julien 185 Paul 192 Pierre 187 Andre 167
Fiona 154 Eric 185 Juliette 166
Robert 167 Vivien 170 Maxime 196
```

Pour l'importer correctement, il suffit d'inclure un « @@ » dans le input qui permet de garder la ligne en mémoire tant que tout n'a pas été lu.

Voici le programme SAS correspondant:

```
DATA TAILLE;
INFILE 'Z:\PUBLIC\LOGICIEL\TAILLE.TXT';
INPUT NOM $ TAILLE @@;
RUN;
```

b) Cas où il y a une observation sur plusieurs lignes

Voici le contenu du LOGICIEL.TXT:

```
EXCEL
48
WORD
35
MINITAB
10
SAS
10
SPSS
9
```

Pour l'importer correctement, il faut donner un « / » dans l'instruction INPUT pour préciser la fin de chaque ligne:

```
DATA WORK.LOGICIEL;
INFILE 'Z:\PUBLIC\LOGICIEL\LOGICIEL.TXT';
INPUT NOM $ / NOMBRE ;
RUN;
```

c) Changement de délimiteur

Jusqu'ici, les variables étaient séparées par des espaces dans le fichier à importer. Si ce n'est pas le cas, il faut utiliser l'option `dlm=''` derrière `INFILE`.

Voici le contenu du fichier TAILLEV.TXT:

```
JULIEN;185;PAUL;192;PIERRE;187;ANDRE;167
FIONA;154;ERIC;185;JULIETTE;166
ROBERT;167;VIVIEN;170;MAXIME;196
```

Pour l'importer correctement, il faut donner un `dlm=';'` dans l'instruction `Infile` pour préciser à SAS le délimiteur utilisé.

```
DATA WORK.TAILLEV;
INFILE 'Z:\PUBLIC\LOGICIEL\TAILLEV.TXT' DLM=';';
INPUT NOM $ TAILLE @@;
RUN;
```

d) Plusieurs lignes de titre...(FIRSTOBS)

Si votre fichier texte commence par une ligne de titre, vous pouvez demander à SAS de lire à partir de la deuxième grace à l'option FIRSTOBS de INFILE:

Exemple : `INFILE 'I:\STID\ERIC\ESS.TXT' MISSOVER FIRSTOBS=2;`

e) Lignes très longues (plus de 132 caractères) Infile/ lrecl.

SAS fixe par défaut la largeur maximale de lecture à 132 caractères. Si votre fichier fait 500 colonnes de large, il faut ajouter sur Infile l'option LRECL=500.

```
INFILE 'G:\STID\ERIC\ESS.TXT' MISSOVER FIRSTOBS=2 LRECL=500;
```

f) Saut à la ligne Option Flowover et Missover de Infile.

Par défaut, SAS va à la ligne suivante s'il n'a pas saisi toutes les données sur la ligne courante que sont INPUT demandait. (option FLOWOVER)

Il est beaucoup plus habituel de considéré comme manquantes les données ne figurant pas sur la ligne en question (plutôt que d'aller en chercher sur la ligne suivante ! ! !). (option Missover)

```
INFILE 'G:\STID\ERIC\ESS.TXT' MISSOVER FIRSTOBS=2;
```

Exercice récapitulatif.

Nous avions demandé aux STID193 de nous donner leur date de naissance. Voici le fichier STID93CO.TXT

```
          A;1;21/10/73;DOC
          A;2;08/12/74;cio
          A;3;15/08/72;ami
          A;4;10/11/72;cio
          A;5;30/11/74;cio
          A;6;11/02/74;cio
          A;7;14/09/74;cio
...
          B;26;18/12/73;cio
          B;27;15/03/73;cio
          B;28;08/08/69;livre ètudiant
          C;1;25/07/73;
          C;2;28/10/72;
          C;3;30/04/74;
```

Il contient, pour les STID193 : leur groupe, leur numéro d'ordre dans le groupe, leur date de naissance et la façon dont il ont connu l'IUT.

Ecrivez le programme SAS réalisant son importation en considérant dans un premier temps la variable date comme du texte ($)et lancez le.

Il y a une difficulté ici. Remarquez les individus du groupe C, ils n'ont pas renseigné la dernière variable. SAS, par défaut veut mettre quelque chose dans la dernière variable !(option Flowover de Infile), il va donc lire ce qui lui manque sur le l'individu du groupe C à la ligne suivante ce qui fausse tout !

Il faut donc demander à SAS de considérer comme manquant la dernière variable lorsqu'il n'y a rien sur la ligne. Pour cela, vous ajouterez l'option MISSOVER dans l'instruction INFILE.

Faites un PROC PRINT pour vous assurer que le fichier comporte bien 106 personnes.

Compléter le programme précédent pour importer la date correctement et calculez l'âge des STID193 à la date d'aujourd'hui.

4. Lecture des données par colonnes dans un fichier ASCII externe.

Cette méthode d'importation n'est à utiliser que **lorsque File/Import ne fonctionne pas et lorsque vous souhaitez importer une partie du fichier ASCII original**. Pour importer tout le fichier, utilisez l'instruction Input classique.

a) Instruction INPUT (lecture en colonne)

Supposons que vous ayez un fichier 'G:\STID9597\MONREP\ESS.TXT' sous la forme suivante: (l'encadré indique les n° de colonne: 1, 5, 10, 15, 20 et 25)

```
          1    1    2    2
1...5....0....5....0....5
A  180     68      69700
B  168     61.5    26000
A  178     68.5
A  175.5   78.5    74100
```

← première ligne du fichier

Vous avez ici 3 individus sur lesquels agissent quatre variables.

- Colonne 1: variable Bac
- Colonnes 4 à 8: variable taille
- Colonnes 12 à 15: variable poids (il faut prendre le plus large possible !)
- etc.

Pour mettre ce fichier au format SAS sous le nom WORK.ESSAI en ne conservant que les variables bac, poids, et taille, il faudra taper:

```
DATA WORK.ESSAI;
        infile 'G:\STID9597\MONREP\ESS.TXT' Missover;
        input BAC $ 1 TAILLE 4-8 POIDS 12-15;
run;
```

Application:

Que faut-il ajouter au programme précédent pour charger en plus la variable code postal ?

Remarque: L'option MISSOVER de INFILE indique à SAS de compléter par des valeurs manquantes toute ligne incomplète. (Sinon SAS va lire à la ligne suivante ce qui lui manque (Flowover))

Mise en pratique

*Editez le fichier ESS.TXT (**au format texte DOS**)avec un éditeur de texte (Word en police courier new pour avoir des colonnes bien alignées avec affichage des espaces et des tabulations (OUTILS/Options/Afficher /Tabulations, espaces. Vérifiez les données (et leurs positions (colonnes)) et fermez le fichier.*

Importez ce fichier grâce à INPUT (en colonnes) dans un fichier SAS: Work.Essai. Vérifiez l'importation grâce à Proc Print.

Si l'importation précédente a réussi, ajoutez les lignes suivantes à votre fichier ESS.TXT et importez-le à nouveau:

```
B  185.5   79.5    7100
A  175.5   78
A  175.5   78.5    74100
```

Il se peut que les 4è et 5è lignes ne soient pas lues correctement sous SAS. Comprenez pourquoi et rectifiez le problème (sous Word)

Répéter l'opération précédente jusqu'à ce que vous ayez obtenu un succès complet.

b) Données générées dans un programme SAS (simulations)

DO/END

(1) Instruction DO/END

(=FOR du Pascal)

Syntaxe:

```
DO variable=début TO fin [BY incrément];
    instruction1;
    instruction2;
...;
END;
```

Exemple:

```
DATA  _NULL_;          On ne fera pas de sauvegarde des variables qui suivent...
 DO I=1 TO 10;         On commence la boucle.
    X=RANUNI(0);       Nous calculons une réalisation d'une VA suivant une U[0,1]
    PUT X;             Nous l'affichons à l'écran (fenêtre LOG)
 END;                  La boucle est terminée.
RUN;
```

Ce programme génère 10 nombres (réalisations d'une loi uniforme [0,1]) et les affiche dans la fenêtre LOG. (Pour plus de détails sur RANUNI allez dans l'annexe sur les fonctions de SAS)

Si vous voulez mettre cet échantillon dans un fichier de données SAS work.ech, nous avons le programme suivant :

```
DATA WORK.ECH;
 DO I=1 TO 10;
 X=RANUNI(0);
 OUTPUT ;          Envoie les variables présentes dans le fichier work.ech.
 END;
RUN;

PROC PRINT DATA=WORK.ECH;
RUN;
```

Faites un PROC PRINT ; RUN ; quel est l'inconvénient du programme précédent ?

Pour y remédier, nous utilisons une option, que vous verrez plus tard, qui permet de supprimer certaines variables.

Nous changeons la première ligne en :

```
data work.ech (drop=i);
```

(nous supprimons « i » du fichier final)

Effectuez cette modification et visualisez le résultat.

(2) Instruction DO WHILE / END

Les instructions sont répétées tant que la condition est vraie.

Cette fois, nous allons générer des nombres suivant une N(0,1) tant qu'ils sont inférieurs ou égaux à 2 et stopper dès que cette condition n'est plus remplie.

```
DATA WORK.EXEMPLE;
        X=0;                          Initialise x à 0
        DO WHILE (x<=2);              Commence la boucle...(on en sort dès que x>2)
          X=RANNOR(0);                x est une réalisation d'une N(0,1)
          OUTPUT;                     Nous l'inscrivons dans le fichier exemple
        END;                          La boucle est terminée
RUN;                                  La séquence DATA aussi.
```

Remarque : Vous découvrez ici une possibilité de l'instruction OUTPUT. Elle permet d'inscrire chaque valeur valide de x à la suite du fichier de donnée ouvert. Supprimez-là et le fichier exemple ne contiendra qu'une seule observation ! (La dernière !)

(3) Instruction DO UNTIL/ END

Les instructions sont répétées jusqu'à ce que la condition soit vraie.

```
DATA WORK.EXEMPLE;
        X=0;                          Initialise x à 0
        DO UNTIL (x>2);               Commence la boucle...(on en sort dès que x>2)
          X=RANNOR(0);                x est une réalisation d'une N(0,1)
          OUTPUT;                     Nous l'inscrivons dans le fichier exemple
        END;                          La boucle est terminée
RUN;                                  La séquence DATA aussi.
```

Ce programme fait la même chose que le précédent.

(4) Instruction GO TO

Un petit exemple vaut mieux qu'un long discours...

```
DATA _NULL_;                          On ne conservera pas ces données !
  DO I=1 TO 100 BY 2;                 i va de 1 à 100 de 2 en 2
     PUT I=;                          on affiche i dans la LOG
     IF I=11 THEN GO TO FIN;          on sort dès que i est égal à 11
  END;                                fin de la boucle
FIN: PUT 'TERMINE';                   on affiche le message dans la LOG
RUN;
```

K. Utilisateurs du système SAS en France au 1.1.1996

Liste partielle des utilisateurs du système SAS en France au 1.1.1996[130]

[130] SAS refuse de communiquer toute mise à jour de cette liste. Dommage…

L. INDEX

D

E

F

G

H

I

J

K

L

M

N

O

P

Q

R

S

T

U

V

W

Y