Laboratoire de Statistique et Probabilités — U.M.R. 5583 (C.N.R.S. - U.P.S. - INSAT)

T.P. SAS numéro 01

Après quelques généralités sur le logiciel SAS, l'objet de ce T.P. est de réaliser les premières manipulations avec ce logiciel. Avant cela, il faut mettre en place l'environnement UNIX nécessaire à l'utilisation de SAS dans des conditions commodes.

Généralités sur le logiciel SAS

Le logiciel SAS est de conception américaine : il est developpé et commercialisé par la société SAS-Institute, située à Cary, en Caroline du nord. Écrit en langage C, SAS (*Statistical Analysis System*) est, à l'origine, un logiciel de statistique polyvalent, c'est-à-dire susceptible de traiter pratiquement tous les domaines de la statistique. Il est assez ancien (ses débuts remontent aux années 1970) et est constamment enrichi de nouvelles méthodes. Par suite, il est très volumineux et souvent redondant : le même problème statistique peut être traité par différents modules du logiciel (avec souvent des présentations différentes !). Aujourd'hui, SAS est devenu un véritable système de gestion de l'information plutôt qu'un simple logiciel de statistique.

La "doc" (documentation) papier de SAS est "monstrueuse" : plusieurs centaines de volumes, certains dépassant les 1000 pages... Bien entendu, elle n'existe qu'en anglais (on imagine le coût d'une traduction...). Toutefois, des cours polycopiés synthétiques existent en français. Par ailleurs, cette doc est aujourd'hui partiellement en ligne.

Les 6 volumes les plus utiles de la doc, pour un utilisateur de base de SAS, sont les suivants :

- *SAS language* : ce volume donne une description générale du logiciel, ainsi que des informations sur les commandes SAS, les sous-programmes standards, le principe général des *macros*...
- *SAS procedures guide* : donne une notice détaillée sur toutes les procédures statistiques de base contenues dans SAS ;
- *SAS/STAT user's guide, volumes 1 & 2* : donnent également une notice détaillée sur les procédures statistiques plus avancées ;
- *SAS/GRAPH software, volumes 1 & 2* : précisent la façon d'écrire des procédures SAS pour obtenir des graphiques élaborés.

Toutefois, la documentation la plus accessible dans le cadre de ces T.P., et à laquelle on se réfèrera en permanence dans toute la suite, est le cours polycopié suivant

SAS sous Unix : Logiciel hermétique pour système ouvert (version de septembre 2001)

corédigé par J.M. Azaïs, P. Besse, H. Cardot, V. Couallier et A. Croquette (Laboratoire de Statistique et Probabilités, Université Paul Sabatier, Toulouse). Les renvois aux pages seront relatifs à ce cours polycopié. On peut se le procurer à l'adresse suivante

http ://www.lsp.ups-tlse.fr/

rubrique "Doc. pédagogiques", Deuxième cycle.

Deux cours de statistique descriptive multidimensionnelle, disponibles sur le même site internet, seront également utiles dans les séances de T.P.

Signalons, pour terminer, 3 ouvrages (en anglais) présentant, à différents niveaux, l'usage du logiciel SAS dans le traitement statistique de données.

- *How SAS works*, de P.A. Herzberg, Springer, 1990 (pour un niveau élémentaire).
- *Applied statistics and the SAS programming language*, de R.P. Cody et J.K. Smith, third edition, Prentice Hall, 1991 (pour un niveau plus avancé).
- *A handbook of statistical analyses using SAS*, de B.S. Everitt et G. Der, Chapman and Hall, 1996 (également pour un niveau plus avancé).

Connexion sous UNIX

Depuis un terminal X, se connecter au C.I.C.T. (Centre Interuniversitaire de Calcul de Toulouse) avec la machine "ondine". Le système d'exploitation utilisé est UNIX (on trouvera une présentation résumée de ce système en annexe A du cours polycopié). Afin de définir un environnement commode pour UNIX et pour SAS, taper les commandes suivantes :

```
cp ~besse/.login ~
cp ~besse/.tcshrc ~
```

Une déconnexion-reconnexion permet alors d'activer les fichiers `.login` et `.tcshrc` qui mettent en place un environnement adéquat.

Il est ensuite important de créer un répertoire spécial (utiliser la commande `mkdir` d'UNIX) sous lequel on se placera pour appeler SAS (on pourra, par exemple, l'appeler `tpsas`).

Les fenêtres principales de SAS

Entrer dans SAS en faisant la commande `sas &` (la version actuelle est la version 8.2). Une fois connecté à SAS, plusieurs fenêtres s'ouvrent à l'écran (pages 11–12). Les 3 plus utiles sont :

- *SAS : Program Editor* : c'est l'éditeur de texte de SAS, dans lequel on doit entrer tout programme à exécuter ; des rudiments sur sa manipulation se trouvent page 12 ;
- *SAS : Log* : fenêtre dans laquelle s'affichent, au cours de l'exécution d'un programme, le programme lui-même, séquence par séquence (en noir) et les commentaires du système SAS sur ce programme (en bleu) ; le cas échéant, s'affichent également ici un message d'avertissement, lorsqu'un problème non fatal est détecté (précédé de *warning*, en vert) ou un message d'erreur, lorqu'une erreur fatale est détectée (précédé de *error*, en rouge) ; (noter que *to log* signifie, en anglais, enregistrer, noter sur un registre) ;
- *SAS : Output* : fenêtre dans laquelle s'affichent tous les résultats obtenus à l'issue d'un programme (lorqu'il a marché !).

Ces 3 fenêtres possèdent sensiblement les mêmes "menus déroulants" sur leur partie haute. Il est vivement conseillé de les disposer de façon commode à l'écran (en gardant également accessible la fenêtre UNIX).

On quitte SAS en se plaçant dans la fenêtre *program editor* et en utilisant le menu déroulant `File/Exit`.

Les 2 possibilités d'utilisation de SAS en mode interactif

Il existe diverses façons de faire du traitement de données avec SAS (pages 11-13). Les 2 façons de le faire en mode interactif sont indiquées ci-dessous.

- *Programmation SAS.* Cela consiste à écrire un programme SAS et à lancer son exécution par le menu déroulant `Run/Submit`. C'est essentiellement de cette façon que nous procèderons dans le cadre de ces T.P. Un programme SAS est une succession de procédures, chacune réalisant un traitement statistique homogène ou un graphique.
- *SAS/INSIGHT* (menu déroulant `Solutions/Analysis/Interactive Data Analysis`). Permet un traitement interactif immédiat et puissant des données ; de nombreuses méthodes sont disponibles et on peut réaliser des graphiques très élaborés. SAS/INSIGHT sera abordé dans les dernières séances de T.P.

Notion de table SAS

Un fichier de données ne peut être reconnu, lu et traité par SAS que s'il est dans un format spécifique. De même, les fichiers produits en sortie d'une procédure SAS seront dans ce format spécifique. Nous appellerons *table SAS* (traduction officielle de *SAS data set*) un tel fichier mis dans un tel format (page 9). Nous verrons dans le T.P. 02 qu'il existe des tables SAS temporaires et d'autres permanentes.

Articulation d'un programme SAS

Un programme SAS comprend des entrées-sorties (lectures et écritures de données) et des enchaînements de procédures.

Gestion des données avec SAS

Il existe 2 possibilités pour lire un fichier de données dans un programme SAS. Tout d'abord, il est possible de lire directement les données, en les incluant dans le programme, au moyen de la commande `cards` (voir l'exercice 01.1). Cette façon de procéder, peu commode, n'est pas recommandée. Ensuite, on peut lire des données préalablement enregistrées dans un fichier ASCII (extérieur à SAS), au moyen de la commande `infile` (voir l'exercice 01.2). Dans un cas comme dans l'autre, une déclaration des variables est obligatoire, au moyen de la commande `input`. Enfin, noter que les données ne seront réellement utilisables qu'une fois transformées en table SAS, ce qui se fait par la commande `data`, suivie du nom que l'on souhaite donner à cette table ; la commande `data` doit être placée en début de séquence, avant même la lecture des données. Ainsi, une séquence de lecture des données se présente en général de la façon suivante :

```
data <nom de la table SAS>;
infile '<nom du fichier des données>';
input <liste des variables>;
run;
```

Noter que, dans la liste des variables, le séparateur est un blanc.

Par ailleurs, chaque procédure SAS réalisant un traitement statistique produit un certain nombre de résultats. Ces résultats sont soit affichés dans la fenêtre *output*, soit enregistrés dans une table SAS particulière. Dans ce dernier cas, la procédure `print` permet d'afficher le contenu de cette table SAS dans la fenêtre *output*. On peut ensuite archiver le contenu de la fenêtre *output* dans un fichier ASCII en utilisant le menu déroulant `File/Save as`.

Les procédures SAS

Un programme SAS est en fait un enchaînement de procédures, chacune réalisant un traitement homogène. Les procédures de base sont répertoriées dans le volume *SAS procedures guide*. En dehors de la procédure `print`, déjà citée, les principales procédures sont les suivantes (pages 33–36) :

- *means, univariate* : servent à la description élémentaire de variables quantitatives (nombre d'observations, minimum, maximum, moyenne, écart-type...) ; *univariate* est plus élaborée ;
- *freq* : sert à la description élémentaire de variables qualitatives (effectifs, fréquences... ; permet aussi de croiser 2 ou plusieurs variables, de déterminer des profils, de calculer des khi-deux...) ;
- *plot* : réalise le nuage de points relatif à 2 variables quantitatives ;
- *chart* : réalise différents graphiques pour une variable qualitative ;
- *sort* : range le fichier selon les valeurs croissantes ou décroissantes d'une variable quantitative spécifiée par `by` ;
- *rank* : calcule une variable "rang" pour chaque variable quantitative déclarée (déclaration obligatoire) ;
- *standard* : permet de déterminer les valeurs centrées et réduites associées à une variable quantitative donnée (nécessite les options *mean = 0* et *std = 1*) ;
- *corr* : permet de calculer la matrice des corrélations (ainsi que la matrice des variances-covariances) d'un ensemble de variables quantitatives.

Exemple : `proc univariate; run;`

Options, title, footnote et commentaires

Diverses commandes générales peuvent être rajoutées au début d'un programme SAS.

- *Options* : il est ici question des options générales, à mettre en début de programme (nous verrons dans le T.P. 03 des options de procédure et des options de commande) ; on les déclare avec la commande `options` (le s est facultatif) ; citons en 3 : `pagesize`, qui spécifie le nombre de lignes dans une page de sortie (*output*) ; `linesize`, qui spécifie le nombre de caractères par ligne ; `nodate`, qui supprime l'impression de la date dans les sorties.
- *Title (titre)* : permet de placer un titre en haut de chaque page des sorties ; exemple : `title 'ceci est un titre';`
- *Footnote (pied-de-page)* : permet de placer un titre en bas de chaque page des sorties ; permet également une meilleure mise-en-page des sorties SAS sur une imprimante ; exemple : `footnote 'ceci apparaitra en bas de page';`

– *Commentaires* : des commentaires peuvent être insérés n'importe où dans un programme SAS ; ils doivent être placés de la façon suivante :
`/* ceci est un commentaire */` (surtout commode au sein d'une ligne de commande)
`* ceci en est`
`un autre;`

Exercice 01.1

Reproduire et faire marcher le programme SAS suivant (noter que la commande `input` précède ici la commande `cards`) :

```
data donnees;
input V1 V2;
cards;
5 6
2 8
3 10
12 6
7 4
;
run;
proc print;
run;
```

À cette occasion, utiliser les fonctionnalités de la fenêtre *programm editor* (voir page 12).

Exercice 01.2

– Sous UNIX (pas dans SAS), copier les fichiers
`~baccini/tpsas/exolim/exo01_2.sas`
`~baccini/tpsas/exolim/data/notes.don`
`~baccini/tpsas/exolim/data/notes.txt`
et contrôler leur contenu.
– Dans la fenêtre *program editor*, faire afficher, puis exécuter le programme `exo01_2.sas` (noter que, dans ce cas, la commande `input` doit suivre la commande `infile`).

Pour en savoir plus

À l'issue de ce premier T.P., il est vivement conseillé de se reporter au cours polycopié pour bien s'approprier les premières manipulations de SAS. Pour cela, nous conseillons de lire attentivement les chapitres 1 et 2, ainsi que l'annexe A. Il est également conseillé, en dehors des séances de T.P., de s'entraîner à la manipulation de SAS.