Prise en main de SAS

Résumé

Ces tuteuriels proposent une introduction élémentaire à un usage classique du logiciel SAS pour lancer des analyses statistiques. L'objectif est volontairement restreint aux commandes et procédures de base disponibles dans la version de la licence académique de SAS correspondant également aux usages les plus fréquents dans les grandes entreprises. Les modules concernés sont : SAS de base, SAS/Stat, SAS/Graph. Le premier tuteuriel introduit le logiciel, son organisation, ses objets et propose une prise en main du mode interactif d'utilisation.

Plan des tuteuriels :

- *Prise en main*
- *Gestion des données*
- *Graphiques*
- *Macros-commandes*
- *Bases de données*

Les procédures du module SAS/STAT sont étudiées dans les cours de statistique afférents.

1 Introduction

1.1 Historique

Le système SAS, né au début des années 60, conserve, de son environnement initial de conception (IBM) les caractéristiques fondamentales : complexité (principe de compatibilité ascendante), lourdeur, coût et aussi puissance et efficacité. Il a acquis une position dominante jusqu'à la fin du siècle dernier dans beaucoup de secteurs d'activités. En France, les grandes entreprises de l'énergie et administrations : INSEE, EDF, GDF,..., toute l'industrie pharmaceutique l'avaient adopté ainsi que les entreprises du tertiaire impliqués dans la gestion volumineuse de bases clientèles (banques, assurances, marketing, VPC...). SAS, ne signifiant plus *Statistical Analysis Sytem*, devint un systèmes d'information global et le calcul statistique plus accessoire.

C'est tout le système d'information de l'entreprise qui peut être pris en charge, de la collecte, la gestion, la préparation des données, leur analyse à la modélisation et l'édition de tableaux de bords, rapports, page web...

Depuis le début du siècle, la domination de SAS rétrograde rapidement au profit des logiciels *open source*. Les raisons en sont notamment le coût de location prohibitif et les niveaux élevés, intégration, complexité de son organisation. Ce repli s'accélère avec le déluge des données massives qui impose une flexibilité technologique incompatible avec une politique intégrée (totalitaire ?) du système d'information, de la saisie à la décision.

1.2 Organisation

Le système SAS est un ensemble de modules logiciels exécutables par lots (en mode *batch*) oiu à travers différents types d'interfaces utilisateur, plus ou moins amicales, pilotées par menus, à partir d'un navigateur ou encore graphiques (*Enterprise Miner*).

Depuis la version 8, SAS propose des *solutions* : *analyse guidée des données*, *analyse marketing*, *Prévision de séries chronologiques... Analyse interactive des données* [1] associant une problématique et une interface spécifiques permettant un traitement de l'information sans écrire une ligne de programme. Il serait certes possible, en première approche, de se contenter de cette utilisation élémentaire mais l'usage montre que ces solutions sont nécessairement limitées et qu'un usage professionnel, associé à des contraintes nécessairement originales, rend incontournable l'usage d'une programmation basique utilisant des syntaxes complexes et hétérogènes d'un module ou d'une interface à l'autre.

1.3 Les modules et leur documentation

Toutes les documentations et des tutoriels sont disponibles en ligne. Des items sont spécifiques à la version de SAS utilisée (9.3), au système d'exploitation et à chacun des modules offerts à la location. Ceux les plus utilisés concernés par ce cours sont : Base SAS, SAS/STAT, SAS/GRAPH. Ce découpage est imposé par la politique commerciale proposant chaque module à une location annuelle.

1. Malheureusement le développement de ce module (*Insight*) est abandonné au profit de JMP à partir de la version 9.4.

Base SAS

C'est la documentation de base et le manuel de référence pour tous les traitements de gestion des donnés : l'*étape Data*, la syntaxe de ses commandes, la gestion des tables SAS, l'éditeur de texte des programmes. Cet item contient également la description des procédures élémentaires (Procedures Guide), du macro langage pour l'écriture de macro-commandes, des outils de production automatique des rapports et graphes (ODS) en html, des requêtes SQL de bases de données, de production de documents XML...

SAS Procedures

Fonctions et syntaxes de toutes les procédures à l'exception des procédures statistiques plus complexes : statistiques élémentaires, fréquences, graphiques basse résolution, impression, tri, tabulation, transposition,...

SAS/Stat Toutes les procédures statistiques et la grande variété de leurs options : tous les modèles de régression, les classifications, les durées de vie, la statistique non-paramétrique, les analyses multidimensionnelles...

SAS/Graph Software Détails des possibilités graphiques en haute résolution et de leurs innombrables options.

SAS/IML Le module de calcul matriciel interactif intégré est un langage interprété, comme Matlab ou R. Il traite des objets matriciels avec la syntaxe d'un langage évolué (PL1). Il est adapté à la mise en place rapide de méthodes originales construites à partir des opérateurs classiques de l'algèbre linéaire. Très rarement utilisée dans l'industrie, il n'est pas décrit dans ce document mais autorise une interface avec R pour manipuler les objets de ce langage.

2 Objets et utilisation de SAS

2.1 Table SAS

Après importation en provenance d'un fichiers ASCII ou d'une Base (SQL), les données sont gérées par SAS sous la forme d'un *SAS Data Set* nommé par la suite *Table SAS* qui est un fichier ou l'association de fichiers binaires contenant les données et leur descriptif (format, libellé).

Cet objet est de principe analogue au *data frame* de R ou Python : une liste de colonnes ou variables statistiques de types différents prenant leurs valeurs sur n lignes ou individus. Ce peut être aussi une matrice de distances ($n \times n$).

Attention à une différence notoire par rapport à R ou Python, une table SAS n'est qu'exceptionnellement chargée en mémoire en fonction des traitements exécutés. Ceci autorise donc la préparation de données massives plus volumineuse que la RAM de l'ordinateur. C'est un héritage ancien (données sur bande magnétique, cartes perforées) qui présente un intérêt actuel.

2.2 Programme SAS

Un *programme SAS* est un enchaînements d'*étapes* de gestion des données (Data Step) et d'appels de *procédures*, décrivant, dans une syntaxe souvent spécifique à chaque module, les traitements à réaliser sous le contrôle d'*options* prises par défaut ou explicitement définies. Les différentes étapes ou procédures communiquent entre elles exclusivement par l'intermédiaire de *tables SAS*, permanentes ou temporaires.

```
/* exemple de programme SAS */
/* Lecture, impression et tabulation de données. */
data Europe;
    infile "edc.fun.overseas";
    input date $ 1-7 dest $ 8-10 boarded 11-13;
proc print data = europe;
proc tabulate data = europe;
    class date dest;
    var boarded;
    table date, dest*boarded*sum;
run;
```

2.3 Fenêtres interactives

Les traitements opérationnels volumineux sont exécutés en différé (mode *batch*) ; initiation et exploration sont exécutées en mode interactif. Le lancement de SAS ouvre cinq fenêtres principales qu'il serait trop long de décrire

de façon exhaustive.

Chaque fenêtre contient une barre de menus déroulants contextuels :

Fichier pour lire ou écrire dans des fichiers extérieurs à SAS, importer ou exporter des données dans différents formats, quitter SAS (exit) en fermant toutes les fenêtres.

Édition pour gérer le texte (sélectionner, couper, copier, coller...),

Affichage pour rendre active une des fenêtres.

Outils pour accéder à des utilitaires de gestion de graphiques, de tables sas, de rapports, d'images, de textes et pour configurer les options personnalisant son environnement (couleurs, polices...).

Solutions pour exécuter les modules spécifiques (s'ils ont été payés !) pour la réalisation de tableaux de bord, le développement d'applications.

Fenêtre Pour changer l'organisation des fenêtre ou en sélectionner une spécifique.

Aide pour accéder à l'aide en ligne détaillée ainsi qu'à un tutoriel (*Getting started with SAS Software*).

Ces fenêtres sont :

Éditeur est un éditeur de texte rudimentaire. Il est vivement recommander, surtout sous Unix, d'ouvrir un éditeur fiable et d'y entrer le programme avant de le soumettre par simple copie de la souris (*submit clipboard*). Cela permet d'éviter des mauvaises surprises : caractères spéciaux cachés, crash de SAS...

- Passer alternativement du mode insertion au mode superposition : `<ctrl>x`.
- Insérer n lignes : taper `in` dans la zone des numéros de lignes.
- Supprimer une ligne : `d`, supprimer un block : entrer `dd` sur la première et `dd` sur la dernière ligne du bloc.
- Le menu spécifique **Exécuter** lance l'exécution (comme $< F3 >$) du programme de l'éditeur ou celui du tampon copié avec la souris et rappelle (comme $< F4 >$) le dernier programme exécuté.

Sortie affiche tous les résultats (texte) produits par l'exécution des différentes procédures. Les graphiques haute résolution apparaissent dans une fenêtre spécifique.

Journal affiche le compte rendu de la bonne exécution et les messages d'erreur. C'est la première fenêtre à consulter pour y découvrir les erreurs.

Explorateur affichage arborescent des librairies (répertoires) et tables gérées par SAS.

Résultas permet de gérer l'ensemble des résultats (textes et graphiques) de façon arborescente.

D'autres fenêtres s'ouvrent comme par exemple le navigateur par défaut pour afficher les résultats.

Attention aux courants d'air, il est important de gérer correctement la multitude des fenêtres qui remplissent l'écran et surtout de les refermer dans le bon ordre afin d'éviter de se retrouver bloqué, par exemple sur un *popup* qui attend une réponse.

2.4 Bibliothèques

Ce sont, du point de vue du système d'exploitation, les répertoires dans lesquels SAS gère les fichiers et *tables SAS* de façon temporaire, le temps d'une session ou exécution du programme, ou permanente.

SasUser : bibliothèque permanente créée par défaut. Les tables de cette bibliothèque sont nommées `sasuser.nomtab`.

Work : bibliothèque contenant les tables temporaires créées par défaut par les différentes étapes et procédures. Elles sont effacées à la fin de la session ou de l'exécution du programme. Les tables sont nommées `work.nomtab` ou plus simplement de façon implicite : `nomtab`.

Plutôt que d'utiliser toujours la librairie `SasUser`, il est possible de définir sa propre librairie (ou répertoire) de tables permanentes (commande globale `libname`).

3 Première exécution de SAS

3.1 Recommandations

- Créer un répertoire par cours où seront regroupés les jeux de données ainsi que les programme SAS, les fichiers graphiques, le rapport éditer par un traitement de texte.

- Sous Unix, définir ce répertoire comme répertoire courant avant de lancer SAS par la commande `sas &`. Il sera plus difficile de retrouver les fichiers sous Windows qui est lancé à partir du menu `Démarrer` ou de l'icône.
- Sous Unix et par sécurité il est préférable d'ouvrir un éditeur de texte (`kedit` ou ...) de votre préférence pour suppléer aux défauts de celui intégré à SAS. Il contiendra toutes les commandes exécutées et les commentaires associés. Le fichier créé sera l'annexe du rapport. En cas de problème, perte des résultats intermédiaires, mauvais choix stratégiques, plantage de SAS, il suffit de ré-exécuter le fichier pour revenir aux étapes antérieures.
- Ouvrir un traitement de texte afin d'y stoker les résultats pertinents au fur et à mesure de leur obtention sans oublier commentaires et légendes.

3.2 Les données

Une étude[2] réalisée entre 1961 et 1973 dans la maternité d'un hôpital d'Oakland (Californie) avait pour but de rechercher si certaines caractéristiques des parents avaient une influence sur le développement de l'enfant. Parmi les variables collectées, 19 variables décrites dans le tableau ci-dessous ont été observées sur 115 familles ou unités statistiques. Ces variables décrivent des informations médicales et socio-économiques concernant le bébé et ses parents au moment de la naissance puis dix ans plus tard. Ces données vont servir à illustrer la démarche classique d'une étude statistique.

Ces données permettent de se poser différentes questions de nature plutôt épidémiologique :

- Influence ou non de la consommation de cigarettes sur le sexe de l'enfant, sur son poids, sur sa taille,
- sur l'évolution du poids de la mère en 10 ans,
- sur les liaisons entre les caractéristiques des parents (poids, taille, rhésus) et celles de leur enfant,
- ...

Un scénario détaille l'analyse de ce jeu de données avec R. Il est en partie reprit pour illustrer l'usage des fonctions de SAS sur ces mêmes données.

2. J.L. Hodges, D. Krech et R. Crutchfield, *Statlab : an Empirical Introduction to Statistics*, 1975.

Code	Libellé	Unité ou modalités
ESx	sexe de l'enfant	M ou F
ERh	rhésus de l'enfant	Rh+ ou RH-
ET0	taille de l'enfant	à la naissance en cm
EP0	poids de l'enfant	à la naissance en kg
ET10	taille de l'enfant	à 10 ans en cm
EP10	poids de l'enfant	à 10 ans en kg
MRh	rhésus de la mère	Rh+ ou RH-
MA0	âge de la mère	à la naissance
MP0	poids de la mère	à la naissance
MCig0	consom. de cigarettes	0, 1 à 10, > 10
MT	taille de la mère	
MP10	poids de la mère	10 ans après
MCig10	consommation de cigarettes	10 ans après
PA0	âge du père	à la naissance
PCig0	consommation de cigarettes	à la naissance
PT	taille du père	
PP10	poids du père	10 ans après
RF0	revenus familiaux	à la naissance
RF10	revenus	10 ans après

TABLE 1 – Statlab : liste des variables

3.3 Créer une table SAS

Charger le fichier `statlab2.csv` du site `http://wikistat.fr/data` dans le répertoire courant. Visualiser le fichier et remarquer que la première ligne contient le nom des variables, chaque valeur est séparée par une ",", la marque décimale est le point ".". Entrer le programme suivant dans l'éditeur de texte :

```
proc import datafile="statlab2.csv"
     out=sasuser.statlab dbms=csv replace;
                  getnames=yes;
                  datarow=2;
run;
```

Sous Windows, compléter les chemin d'accès au répertoire contenant le fichier. Copier puis coller le texte dans l'éditeur de SAS ; `Exécuter`.

3.4 Exécution d'une procédure

Visualiser, vérifier, le contenu de la table.

```
proc print data=sasuser.statlab;run;
proc tabulate data=sasuser.statlab;
   class ESx ERh;
   var EP0 ET0;
   table ERh, ESx*EP0*mean;
run;
```

Commenter le résultat obtenu.

3.5 Gestion des fenêtres

Chercher le mode d'emploi de la procédure `tabulate`. Vérifier les contenus des fenêtres, les menus associés permettant de sélectionner du texte, d'effacer le contenu de la fenêtre `journal`. Utiliser l'explorateur pour retrouver la table créée et la visualiser :

`Explorateur>Bibliothèques>Sasuser>Stalab`

L'icône "Dossier" devient active et permet de remonter dans l'arborescence.

3.6 Édition des résultats

Contrôler la production du rapport dans un format de type traitement de texte.

```
options nonumber nodate;
title "Données stalab" ;
footnote "date du jour";
ods rtf body="rapport.rtf";
proc tabulate data=sasuser.statlab;
   class ESx ERh;
   var EP0 ET0;
   table ERh, ESx*EP0*mean;
run;
ods graphics on;
proc gchart data=sasuser.statlab;
pie MCig0;
run;
ods graphics off;
ods rtf close;
```

Contrôler dans la fenêtre `journal` la génération du ou des fichiers.