SAS - Compléments

(notes de cours)

SAS est un ensemble logiciel d'une grande richesse. On présente ici quelques commandes de base de statistique descriptive univariée et bivariée.

Cet exposé ne prétend nullement à l'exhaustivité, on se souviendra par ailleurs qu'il est fréquent que plusieurs commandes SAS permettent de demander un même calcul.

Etape DATA

On connaît déjà le format libre, qui exige seulement que les observations associées à un individu se succèdent sur une ligne, séparées par une ou plusieurs espaces, les données manquantes éventuelles étant signalées par un point isolé (caractère retenu en SAS par défaut).

Il est également possible d'indiquer dans la commande **INPUT** les positions sur la ligne des variables à lire.

```
DATA ecole;
        INFILE 'monfich.don';
        INPUT nom $ 1-10 age 11-12 poids 13-17 sexe $ 25;
```

cela est nécessaire en particulier lorsque les données à lire ne sont pas séparées, mais consécutives. Il faut évidemment posséder alors la description exacte du fichier, c'est à dire les positions en colonne de chaque variable.

Procédure FREQ

La procédure **FREQ** permet d'opérer des tris à plat, des tris croisés, voire des tris d'ordre supérieur, sur des variables alphanumériques aussi bien que numériques.

La syntaxe est la suivante:

```
PROC FREQ;
        TABLES demandes [/ options];
```

```
PROC FREQ;
        TABLES sexe csp age;
```

demande les tris à plat des variables indiquées.

PROC FREQ;
TABLES sexe sexe*csp / missing;

demande un tri à plat et un tri croisé, une donnée manquante étant considérée comme un item particulier.

PROC FREQ;
TABLES csp*region / chisq;
TITLE 'Tableau CSP x Région, avec Chi-deux';

demande un tri croisé avec calcul du chi-deux associé au test d'indépendance et indique un titre.

PROC FREQ;
TABLES sexe*csp / list;
BY region;

demande un tri croisé avec présentation des résultats en liste et non dans un tableau, et ceci séparément pour chaque région.

On rappelle que l'emploi de **BY** avec n'importe quelle procédure exige que les données soient préalablement triées selon la ou les variables invoquées par cette demande, on opère donc d'abord une **PROC SORT** si nécessaire.

Procédure TABULATE

La procédure **TABULATE**, aux très nombreuses possibilités, permet de calculer des statistiques variées associées au découpage d'une population selon une, deux ou trois variables de classification.

La syntaxe est la suivante:

PROC TABULATE [options];
CLASS variables1;
VAR variables2;
TABLE demandes [/ options];

où variables1 désigne la ou les variables selon lesquelles est effectuée la classification de la population étudiée, et variables2, la ou les variables quantitatives pour lesquelles sont calculées des statistiques.

Parmi les options de la commande **PROC TABULATE** elle-même, les plus utiles sont **NOSEPS**, qui limite le nombre de lignes de séparation des tableaux édités, et **FORMAT**, qui précise les dimensions d'affichage des nombres.

PROC TABULATE NOSEPS FORMAT = f8.;

permet d'afficher à l'économie des entiers n'excédant pas huit chiffres.

PROC TABULATE FORMAT = f10.2;

affiche les nombres sur dix positions avec deux décimales.

On considère un exemple élémentaire aisé à comprendre dont les variables sont introduites par le début de programme SAS suivant :

```
DATA exemple;
        INPUT sexe $ region $ salaire;
        [...]
PROC TABULATE;
        CLASS sexe region;
        VAR salaire;
```

On retiendra pour les demandes de tables qu'une demande telle que **A*B** produit une présentation en arborescence, tandis qu'une demande telle que **A, B** produit une sortie croisée, avec A en ligne et B en colonne (et **A, B, C** en trois dimension, A pour la page, B pour la ligne et C pour la colonne). Ces règles peuvent se combiner, par exemple en des demandes telles que **A, B*C** etc.

Par défaut, ce sont les comptages qui sont opérés pour les variables de classification (demande implicite **N**), et les sommations pour les variables quantitatives (demande implicite **SUM**).

TABLE sexe region salaire;

donne les tris à plat des variables sexe et région, et la somme de tous les salaires.

TABLE region, sexe;

donne le tri croisé région par sexe en tableau.

TABLE region*sexe;

donne le même tri croisé présenté en arborescence.

TABLE region, salaire;
TABLE region*sexe, salaire;
TABLE region*sexe*salaire;
TABLE region*salaire;
TABLE region, sexe*salaire;

donnent la somme des salaires selon les divers découpages et présentations demandés.

Une variable de classification implicite **ALL** permet d'ajouter des calculs marginaux à des demandes croisées.

TABLE region, (sexe **ALL**); **TABLE** region*sexe, salaire **ALL**;

Les demandes **MEAN** et **STD** permettent de faire calculer la moyenne et l'écart-type pour les groupes sur lesquels est calculée la somme **SUM** par défaut.

TABLE region, salaire***MEAN**;
TABLE region*salaire*(**SUM MEAN**);
TABLE region, sexe*salaire*(**MEAN STD**);

Les demandes **PCTN** et **PCTSUM** (**PCT** pour *pourcentage*) permettent enfin de faire calculer les fréquences ou les fractions d'une somme selon la syntaxe illustrée par les exemples suivants :

TABLE region, sexe***PCTN**<sexe>;

calcule les fréquences conditionnelles en colonnes (conditionnement par le sexe).

TABLE region, sexe***PCTN**<region>;

calcule selon le conditionnement contraire.

TABLE region, sexe***PCTN**<region*sexe>;

calcule les fréquences associées à tous les couples d'items.

TABLE region, salaire*(**SUM PCTSUM**<region>);

calcule les parts du total des salaires par région.

Procédure SUMMARY

La commande **SUMMARY** permet de faire calculer des statistiques usuelles sur des variables quantitatives, les résultats sont envoyés dans un "fichier SAS" dont on peut demander l'affichage.

Sa syntaxe élémentaire est la suivante :

PROC SUMMARY;
CLASS variables;
VAR variables;
OUTPUT OUT = fichier SAS des résultats [statistiques];

par défaut les statistiques évaluées sont le nombre d'observations, le minimum, le maximum, la moyenne et l'écart type.

PROC SUMMARY;
CLASS sexe region;
VAR salaire age;
OUTPUT OUT = result;
PROC PRINT DATA = result;

calcule les statistiques précédentes pour toute la population, pour les groupes découpés par les variables sexe et région, et enfin par le croisement de ces deux variables de classification. L'affichage de ces résultats est ensuite demandé par la **PROC PRINT**.

(23 janvier 2004)

----===ooOoo===----