

Les formats expliqués à ma fille

Pour certains, le format, ce n'est qu'une cochonnerie qui complique passablement le travail sur les dates, ou sur les nombres, en n'affichant pas les données telles qu'elles sont stockées. On sait moins que, dans SAS, les formats sont également un formidable outil de recodage de variables.

Quel est le rôle du format ?

Le format est un moyen de présenter (à l'affichage principalement) les données différemment de la façon dont elles sont physiquement stockées. L'exemple le plus frappant est celui des dates : stockées par SAS comme un nombre de jours depuis le 1^{er} janvier 1960, un format les affichera en jours, mois et années.

Dans une procédure statistique (aussi bien basique, comme MEANS ou FREQ, que complexe comme REG, GLM, LOGISTIC et autres FORECAST et LIFEREG), l'utilisation d'un format peut éviter le recodage d'une variable. Ainsi, le traitement par code activité (NAF) d'une entreprise peut être fait par chacune des 696 activités (sans format), ou par regroupement en 222, 60, 31 ou 17 groupes d'activités, uniquement par l'application du format adéquat. De même, un traitement par date pourra se faire au jour près (sans format), ou par semaine, par mois, par trimestre ou par année, en fonction du format utilisé.

Les formats fournis par SAS

Sans nécessiter d'intervention de l'utilisateur, une multitude de formats sont déjà créés dans SAS. Le tableau ci-dessous en récapitule une partie, parmi les plus utiles. La lecture de la documentation de SAS complétera la liste.

Dans tous ces formats, le nombre qui *précède* immédiatement le point indique le nombre de caractères utilisés pour l'affichage. Dans le cas de données numériques, ce nombre de caractères inclut le séparateur décimal, un éventuel signe moins, un éventuel séparateur de milliers, etc.

Si un nombre est situé *après* le point, il indique le nombre de décimales à afficher.

Données « caractère »		Dates SAS (nombres de jours depuis le 01/01/1960)	
\$2.	Ab	DDMMYY10.	24/11/2006
\$UPCASE2.	AB	MMYY57.	11/2006
Données numériques		YEAR4.	2006
5.	12345	FRADFWDX.	24 novembre 2006
7.2	1234.56	FRADFWKX.	vendredi 24 novembre 2006
NUMX7.2	1234,56	FRADFMN.	novembre
NLNUM10.2	1 234,56	FRADFDWN.	Vendredi

Création de formats personnalisés

La procédure FORMAT de SAS permet la création de formats selon vos besoins. Cette création peut se faire en énumérant les correspondances entre valeurs stockées et valeurs affichées (voir les deux exemples ci-dessous) ou en lisant une table SAS à la structure bien

déterminée, contenant toutes les informations pour créer un format (exemple de l'option CNTLIN à la fin de cette partie).

Le 1^{er} exemple crée un format associant leur nom aux numéros de département d'Ile de France. On supposera que les numéros de département ont été stockés dans une variable de type caractère : il faut alors que le nom du format commence par un \$.

Les autres contraintes sur le nom du format sont : pas plus de 8 caractères (32 en version 9), signe \$ compris, ne commençant ni ne se terminant par un chiffre, ne contenant que lettres non accentuées, chiffres et underscore _.

```
PROC FORMAT ;
  VALUE $idf
    "75" = "PARIS"
    "77" = "SEINE ET MARNE"
    "78" = "YVELINES"
    "91" = "ESSONNE"
    "92" = "HAUTS DE SEINE"
    "93" = "SEINE SAINT-DENIS"
    "94" = "VAL DE MARNE"
    "95" = "VAL D'OISE"
  ;
RUN ;
```

Imaginons maintenant que l'on souhaite créer des tranches d'effectifs d'entreprises. On partira alors d'une variable numérique (donc le nom du format doit impérativement ne pas commencer par un \$). Le signe < situé immédiatement avant et/ou immédiatement après le tiret indique que l'on souhaite exclure la borne de gauche et/ou de droite de l'intervalle.

Le mot-clé HIGH désigne la plus grande valeur rencontrée. On pourrait aussi utiliser le mot-clé LOW en lieu et place du zéro dans la 1^{ère} tranche.

```
PROC FORMAT ;
  VALUE effectif
    0 -< 20 = "Très petite entreprise"
    20 -< 250 = "PME"
    250 - HIGH = "Grande entreprise"
  ;
RUN ;
```

Un troisième cas consiste à utiliser une table SAS contenant la définition d'un ou de plusieurs formats. Une telle table doit impérativement contenir les variables suivantes :

- FMTNAME qui contient le nom du format à créer (sans \$ liminaire même pour un format caractère) ;
- TYPE qui vaut C pour un format caractère ou N pour un format numérique ;
- START qui contient la valeur telle qu'elle est stockée par SAS ;
- LABEL qui contient la valeur telle qu'elle sera affichée à travers.

Dans notre exemple, on utilise une table appelée CODESNAF (qui peut être construite aisément à partir du fichier Excel définissant les codes NAF que propose l'Insee en libre téléchargement sur <http://www.insee.fr>) qui se présente ainsi. Elle contient les définitions (avec libellés en clair) des codes activités des entreprises en 696 postes, ou 222, 60, 31 ou 17 regroupements.

VIEWTABLE: Work.Codesnaf				
	label	start	fmtname	type
1	Agriculture, chasse, sylviculture	01.1A	NAF17F	C
2	Agriculture, chasse, sylviculture	01.1C	NAF17F	C
3	Agriculture, chasse, sylviculture	01.1D	NAF17F	C
4	Agriculture, chasse, sylviculture	01.1F	NAF17F	C
5	Agriculture, chasse, sylviculture	01.1G	NAF17F	C
6	Agriculture, chasse, sylviculture	01.2A	NAF17F	C
7	Agriculture, chasse, sylviculture	01.2C	NAF17F	C
8	Agriculture, chasse, sylviculture	01.2E	NAF17F	C
9	Agriculture, chasse, sylviculture	01.2G	NAF17F	C
10	Agriculture, chasse, sylviculture	01.2J	NAF17F	C
11	Agriculture, chasse, sylviculture	01.3Z	NAF17F	C
12	Agriculture, chasse, sylviculture	01.4A	NAF17F	C
13	Agriculture, chasse, sylviculture	01.4B	NAF17F	C
14	Agriculture, chasse, sylviculture	01.4D	NAF17F	C
15	Agriculture, chasse, sylviculture	01.5Z	NAF17F	C
16	Agriculture, chasse, sylviculture	02.0A	NAF17F	C
17	Agriculture, chasse, sylviculture	02.0B	NAF17F	C
18	Agriculture, chasse, sylviculture	02.0D	NAF17F	C
19	Pêche, aquaculture	05.0A	NAF17F	C
20	Pêche, aquaculture	05.0C	NAF17F	C
21	Industries extractives	10.1Z	NAF17F	C
22	Industries extractives	10.2Z	NAF17F	C
23	Industries extractives	10.3Z	NAF17F	C
24	Industries extractives	11.1Z	NAF17F	C

Le simple programme ci-dessous permet de créer le format \$NAF17F.

```
PROC FORMAT CNTLIN = work.codesNAF ;
RUN ;
```

Utiliser un format dans une procédure statistique

Comme on l'a dit, le format est un moyen d'éviter les recodages de variables dans une procédure statistique.

```
PROC FORMAT ;
  VALUE ages
    LOW - 13 = "13 ans et moins"
    14 - HIGH = "14 ans et plus" ;
RUN ;
PROC MEANS DATA = sashelp.class MEAN MEDIAN ;
  VAR weight ;
  CLASS age ;
  FORMAT age ages. ;
RUN ;
```

Age	Obs	Moyenne	Médiane
13 ans et moins	10	87.35	84.25
14 ans et plus	9	114.11	112.00

Utiliser un format dans la procédure SQL

Par rapport aux autres procédures de SAS, la procédure SQL ne tient pas compte d'un format quand on groupe les statistiques selon une variable formatée. Il faut alors créer une nouvelle variable qui a comme valeur le résultat de l'application « en dur » du format, avec une fonction PUT. La syntaxe est PUT(nomVariable, nomFormat.).

```
PROC SQL ;
  SELECT PUT(date, YEAR4.) AS annee,
         COUNT(*) AS nb
  FROM sashelp.air
  GROUP BY annee
  ;
QUIT ;
```

annee	nb
1949	12
1950	12
1951	12
1952	12
1953	12
1954	12
1955	12
1956	12
1957	12
1958	12
1959	12
1960	12