

La proc GCHART expliquée à ma fille

Produire des diagrammes en bâtons ne semble pas une tâche ardue. Les options disponibles dans la procédure GCHART de SAS ne sont d'ailleurs pas si nombreuses. Mais pour obtenir quelques graphiques complexes, il vaut souvent mieux réfléchir à l'organisation de la table SAS en entrée, que chercher une option qui n'existe peut-être pas. Quelques exemples pour illustrer cette manière de penser.

La syntaxe de base : bâtons, groupes et sous-groupes

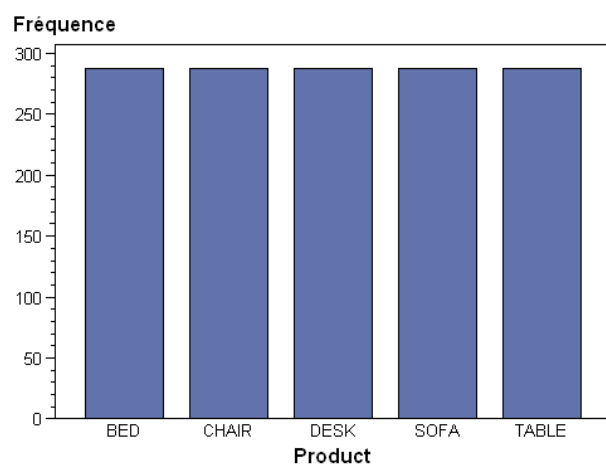
La procédure GCHART définit 4 rôles pour des variables présentes dans la table SAS lue :

- la variable de bâtons, citée dans l'instruction VBAR ou HBAR (selon que les bâtons du graphique à produire seront horizontaux ou verticaux) directement après ce mot-clé ;
- la variable de groupes, citée après le slash des options et GROUP= ; on produit des bâtons juxtaposés (côte à côte) avec un bâton pour chaque valeur de la variable de groupe et de la variable de bâtons ;
- la variable de sous-groupes qui est citée après l'option SUBGROUP= ; on empile alors les bâtons avec des zones colorées différemment pour chaque valeur de la variable de sous-groupe ;
- la variable statistique, citée après l'option SUMVAR= (qui ne signifie pas qu'on somme les valeurs, mais que c'est la variable résumée [summarized]) qui permet, avec l'option TYPE=SUM ou TYPE=MEAN, de donner le mode de calcul des hauteurs de bâtons.

```
PROC GCHART DATA = tableLue ;
  VBAR | HBAR batons / GROUP = groupes
          SUBGROUP = sous_groupes
          SUMVAR = statistique
          TYPE = MEAN | SUM | FREQ | PCT
          DISCRETE ;
RUN ; QUIT ;
```

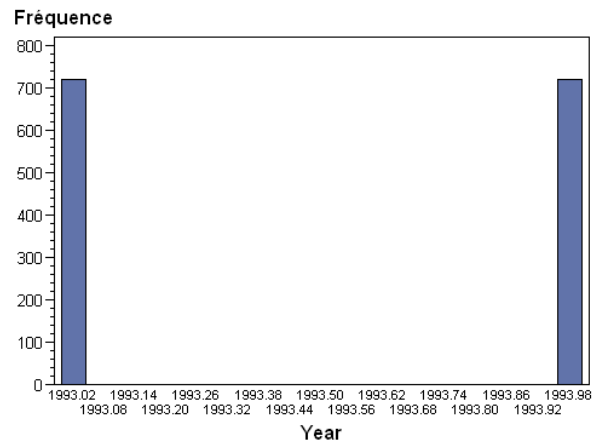
```
PROC GCHART
  DATA = sashelp.prdsale ;
  VBAR product ;
RUN ; QUIT ;
```

Par défaut, avec une variable de type caractère en bâtons, et aucune option, on obtient un bâton par valeur distincte de cette variable, et des hauteurs proportionnelles aux fréquences.



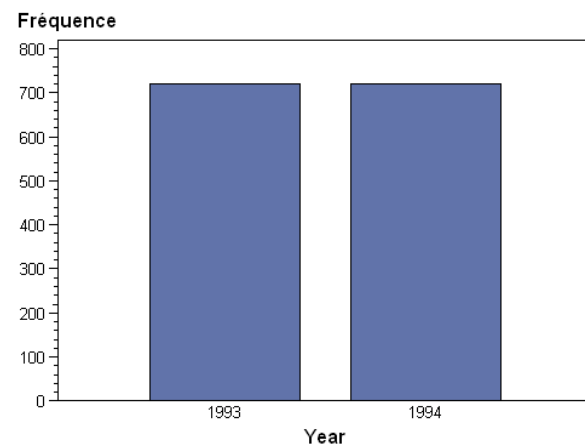
```
PROC GCHART
  DATA = sashelp.prdsale ;
  VBAR year ;
RUN ; QUIT ;
```

Avec une variable bâtons de type numérique, le choix fait automatiquement par la procédure GCHART est discutable. La variable a ses valeurs découpées en tranches (ce qui est utile pour produire un histogramme, mais parfois curieux comme dans cet exemple).



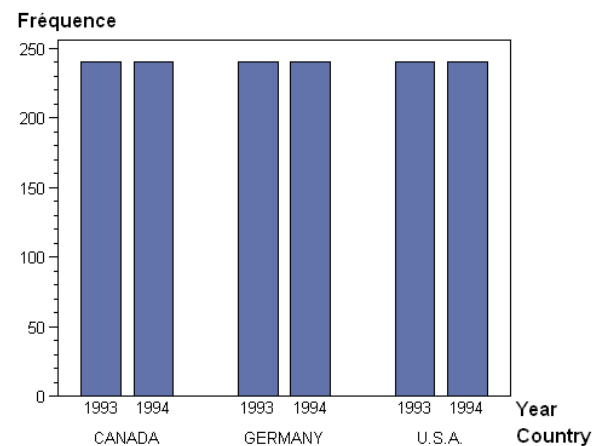
```
PROC GCHART
  DATA = sashelp.prdsale ;
  VBAR year / DISCRETE ;
RUN ; QUIT ;
```

Dans le cas où on veut éviter ce comportement, on ajoutera l'option DISCRETE pour avoir un bâton par valeur de la variable, quel que soit son type.



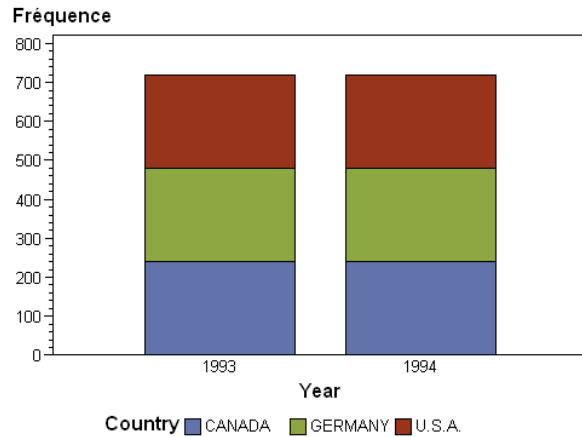
```
PROC GCHART
  DATA = sashelp.prdsale ;
  VBAR year / DISCRETE
  GROUP = country ;
RUN ; QUIT ;
```

L'option GROUP permet d'obtenir plusieurs séries de bâtons. Si on souhaite comparer le cumul des bâtons, cette présentation n'est pas la mieux adaptée (SUBGROUP est préférable).



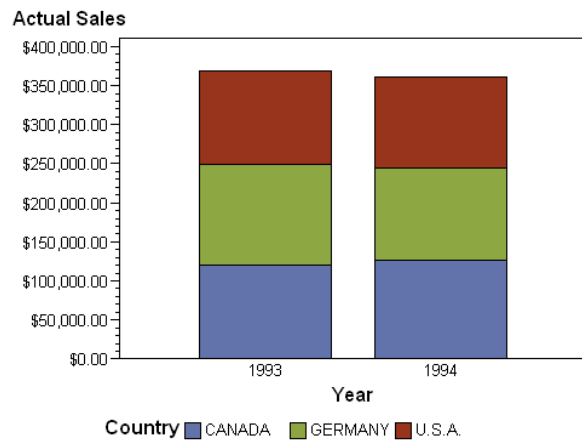
```
PROC GCHART
  DATA = sashelp.prdsale ;
  VBAR year / DISCRETE
      SUBGROUP = country ;
RUN ; QUIT ;
```

Avec l'option SUBGROUP, le graphique est automatiquement enrichi d'une légende. Celle-ci peut être masquée en ajoutant l'option NOLEGEND à l'instruction HBAR ou VBAR.



```
PROC GCHART
  DATA = sashelp.prdsale ;
  VBAR year / DISCRETE
      SUBGROUP = country
      SUMVAR = sales
      TYPE = SUM ;
RUN ; QUIT ;
```

Les options SUMVAR et TYPE permettent de sortir des graphiques n'affichant que des fréquences.



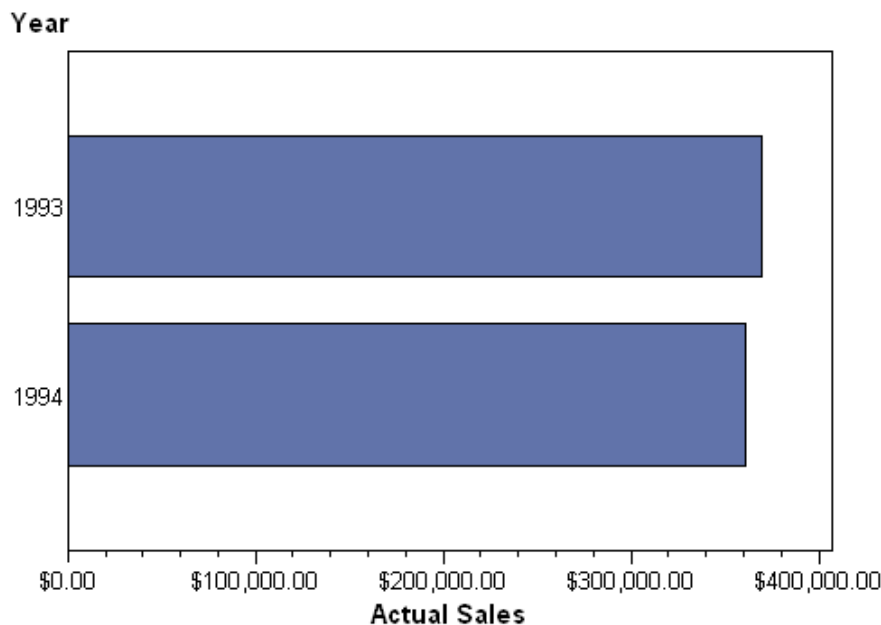
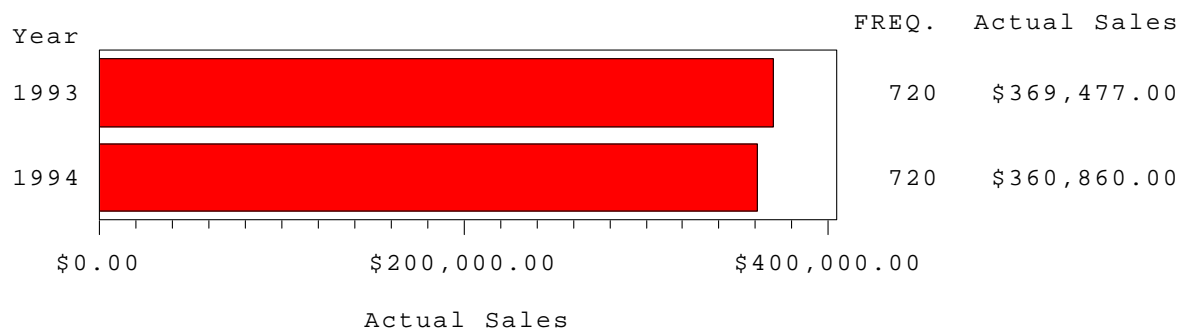
On peut enfin cumuler toutes ces options pour représenter 4 variables simultanément sur le graphique (bâtons, groupes, sous-groupes et statistiques). On est alors à la limite d'avoir une représentation incompréhensible des données.

<pre>PROC GCHART DATA = sashelp.prdsale ; VBAR year / DISCRETE SUBGROUP = country SUMVAR = actual TYPE = SUM GROUP = product ; RUN ; QUIT ;</pre>	<pre>PROC GCHART DATA = sashelp.prdsale ; VBAR product / DISCRETE SUBGROUP = country SUMVAR = actual TYPE = SUM GROUP = year ; RUN ; QUIT ;</pre>

Il est à noter que l'aspect du graphique, si on utilise l'instruction HBAR au lieu de VBAR, peut être assez différente. En fonction du type (driver) graphique utilisé, on verra éventuellement s'afficher des statistiques à droite des bâtons.

```
PROC GCHART DATA = sashelp.prdsale ;
  HBAR year / DISCRETE
                SUMVAR = actual TYPE = SUM ;
RUN ; QUIT ;
```

Les deux graphiques ci-dessous sont générés par le même programme (ci-dessus), mais dans le premier cas on utilise le driver EMF, dans le second, ACTXIMG (image ActiveX), pour envoyer le graphique dans un document Word via l'ODS RTF.



Pour ne pas voir, de manière certaine, les statistiques s'afficher, on peut ajouter l'option NOSTATS à l'instruction HBAR.

Exemple n°1 : représenter plusieurs variables

Imaginons que vos données soient organisées ainsi :

	Actual Sales	Predicted Sales	Country	Region	Division	Product type	Product	Quarter	Year	Month
1	\$925.00	\$850.00	CANADA	EAST	EDUCATION	FURNITURE	SOFA	1	1993	Jan
2	\$999.00	\$297.00	CANADA	EAST	EDUCATION	FURNITURE	SOFA	1	1993	Feb
3	\$608.00	\$846.00	CANADA	EAST	EDUCATION	FURNITURE	SOFA	1	1993	Mar
4	\$642.00	\$533.00	CANADA	EAST	EDUCATION	FURNITURE	SOFA	2	1993	Apr
5	\$656.00	\$646.00	CANADA	EAST	EDUCATION	FURNITURE	SOFA	2	1993	May
6	\$948.00	\$486.00	CANADA	EAST	EDUCATION	FURNITURE	SOFA	2	1993	Jun
7	\$612.00	\$717.00	CANADA	EAST	EDUCATION	FURNITURE	SOFA	3	1993	Jul
8	\$114.00	\$564.00	CANADA	EAST	EDUCATION	FURNITURE	SOFA	3	1993	Aug
9	\$685.00	\$230.00	CANADA	EAST	EDUCATION	FURNITURE	SOFA	3	1993	Sep
10	\$657.00	\$494.00	CANADA	EAST	EDUCATION	FURNITURE	SOFA	4	1993	Oct
11	\$608.00	\$903.00	CANADA	EAST	EDUCATION	FURNITURE	SOFA	4	1993	Nov
12	\$353.00	\$266.00	CANADA	EAST	EDUCATION	FURNITURE	SOFA	4	1993	Dec
13	\$107.00	\$190.00	CANADA	EAST	EDUCATION	FURNITURE	SOFA	1	1994	Jan
14	\$354.00	\$139.00	CANADA	EAST	EDUCATION	FURNITURE	SOFA	1	1994	Feb
15	\$101.00	\$217.00	CANADA	EAST	EDUCATION	FURNITURE	SOFA	1	1994	Mar
16	\$553.00	\$560.00	CANADA	EAST	EDUCATION	FURNITURE	SOFA	2	1994	Apr
17	\$877.00	\$148.00	CANADA	EAST	EDUCATION	FURNITURE	SOFA	2	1994	May
18	\$431.00	\$762.00	CANADA	EAST	EDUCATION	FURNITURE	SOFA	2	1994	Jun
19	\$511.00	\$457.00	CANADA	EAST	EDUCATION	FURNITURE	SOFA	3	1994	Jul
20	\$157.00	\$522.00	CANADA	EAST	EDUCATION	FURNITURE	SOFA	3	1994	Aug

Vous souhaitez représenter, sur des bâtons juxtaposés, les ventes réelles (1^{ère} colonne, Actual sales) et les prédictions (2^e colonne, Predicted sales), par pays (Country).

En l'état, vos données ne peuvent pas se prêter à un tel graphique : l'option SUMVAR n'accepte qu'une seule variable par graphique. Il faut donc transformer les données au préalable pour avoir sur des observations différentes ce qui se trouve actuellement dans des variables différentes. Transformer ce qui est en colonnes en lignes, c'est un travail de transposition : nous utiliserons donc la procédure TRANSPOSE.

```
PROC SORT DATA = sashelp.prdsale
  OUT = work.ventes ;
  BY country region product year month division ;
RUN ;
PROC TRANSPOSE DATA = work.ventes OUT = work.ventes ;
  BY country region product year month division ;
  VAR actual predict ;
RUN ;
```

L'aspect des données devient alors celui-ci. On fera bien attention de ne pas oublier de variables dans le BY de manière à n'obtenir qu'un pivot partiel des données (cf. le document « La proc TRANSPOSE expliquée à ma fille » disponible dans la même rubrique).

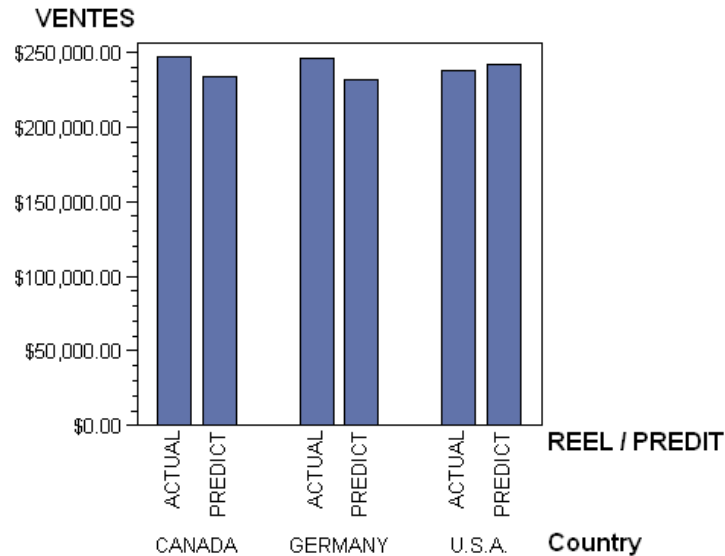
	Country	Region	Product	Year	Month	Division	NOM DE L'ANCIENNE VARIABLE	LIBELLE DE L'ANCIENNE VARIABLE	COL1
1	CANADA	EAST	BED	1993	Jan	CONSUMER	ACTUAL	Actual Sales	\$284.00
2	CANADA	EAST	BED	1993	Jan	CONSUMER	PREDICT	Predicted Sales	\$414.00
3	CANADA	EAST	BED	1993	Jan	EDUCATION	ACTUAL	Actual Sales	\$220.00
4	CANADA	EAST	BED	1993	Jan	EDUCATION	PREDICT	Predicted Sales	\$585.00
5	CANADA	EAST	BED	1993	Feb	CONSUMER	ACTUAL	Actual Sales	\$705.00
6	CANADA	EAST	BED	1993	Feb	CONSUMER	PREDICT	Predicted Sales	\$770.00
7	CANADA	EAST	BED	1993	Feb	EDUCATION	ACTUAL	Actual Sales	\$444.00
8	CANADA	EAST	BED	1993	Feb	EDUCATION	PREDICT	Predicted Sales	\$267.00
9	CANADA	EAST	BED	1993	Mar	CONSUMER	ACTUAL	Actual Sales	\$737.00
10	CANADA	EAST	BED	1993	Mar	CONSUMER	PREDICT	Predicted Sales	\$679.00
11	CANADA	EAST	BED	1993	Mar	EDUCATION	ACTUAL	Actual Sales	\$178.00
12	CANADA	EAST	BED	1993	Mar	EDUCATION	PREDICT	Predicted Sales	\$487.00

Il ne reste alors « plus qu'à » construire la procédure GCHART avec une option GROUP, plus une instruction LABEL pour gérer les intitulés correctement, et le tour est joué.

```

PROC GCHART DATA = work.ventes ;
  VBAR _name_ / DISCRETE GROUP = country
      SUMVAR = coll TYPE = SUM ;
  LABEL coll = "VENTES" _name_ = "REEL / PREDIT" ;
RUN ; QUIT ;

```

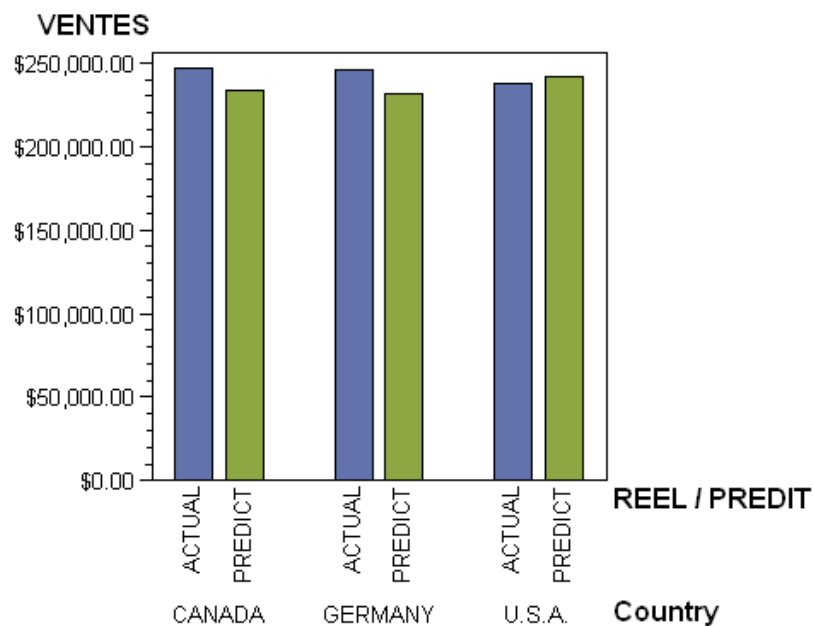


Si on souhaite en plus une couleur différente entre les prédictions et la réalité, il n'y a qu'à ajouter l'option SUBGROUP avec la variable _NAME_ (en plus du reste) pour obtenir deux couleurs ; l'option NOLEGEND masquera une légende redondante.

```

PROC GCHART DATA = work.ventes ;
  VBAR _name_ / DISCRETE GROUP = country
      SUMVAR = coll TYPE = SUM
      SUBGROUP = _NAME_ NOLEGEND ;
  LABEL coll = "VENTES" _name_ = "REEL / PREDIT" ;
RUN ; QUIT ;

```



Exemple n°2 : pyramide des âges

A partir du nombre de Français par sexe et par tranche d'âge, on souhaite produire une pyramide des âges. Cet exemple illustre la possibilité de faire des bâtons dans deux directions différentes. On y voit également la modification d'aspect qu'apporte simplement une instruction PATTERN pour les couleurs des barres, et une instruction AXIS pour les intitulés et les graduations d'axes.

Voici la forme initiale des données (source : site Internet de l'Insee, recensement 1999) :

VIEWTABLE: Work.Recensement			
	age	F	H
1	0 à 4 ans	1457010	1529915
2	5 à 9 ans	1770850	1858444
3	10 à 14 ans	1873426	1959694
4	15 à 19 ans	1921897	2010204
5	20 à 24 ans	1834936	1876612
6	25 à 29 ans	2086403	2091373
7	30 à 34 ans	2128908	2110021
8	35 à 39 ans	2189076	2150411
9	40 à 44 ans	2148598	2096099
10	45 à 49 ans	2126074	2095742
11	50 à 54 ans	1984086	1985266
12	55 à 59 ans	1385938	1371578
13	60 à 64 ans	1412859	1313366
14	65 à 69 ans	1484985	1272868
15	70 à 74 ans	1403212	1085926
16	75 à 79 ans	1290561	876122
17	80 à 84 ans	581025	333659
18	85 à 89 ans	636632	285186
19	90 à 94 ans	294221	96168
20	95 à 99 ans	80455	19289
21	100 ans ou plus	10117	1476

Pour pouvoir réaliser notre pyramide, nous allons transposer les données, de manière à avoir une seule variable pour les populations d'hommes et de femmes, et deux séries d'observations selon les sexes.

L'astuce pour avoir deux séries de barres dans des directions opposées est de donner un signe négatif à la population d'un des deux sexes. Pour pouvoir faire toutes ces opérations d'un seul coup, nous allons transposer à l'aide d'une étape DATA au lieu d'une procédure TRANSPOSE.

```
DATA work.pyramide (DROP = f h) ;
  SET work.recensement ;
  sexe = "F" ;
  pop = -1 * f ;
  OUTPUT ;
  sexe = "H" ;
  pop = h ;
  OUTPUT ;
RUN ;
```

VIEWTABLE: Work.Pyramide			
	age	sexe	pop
1	0 à 4 ans	F	-1457010
2	0 à 4 ans	H	1529915
3	5 à 9 ans	F	-1770850
4	5 à 9 ans	H	1858444
5	10 à 14 ans	F	-1873426
6	10 à 14 ans	H	1959694
7	15 à 19 ans	F	-1921897
8	15 à 19 ans	H	2010204
9	20 à 24 ans	F	-1834936

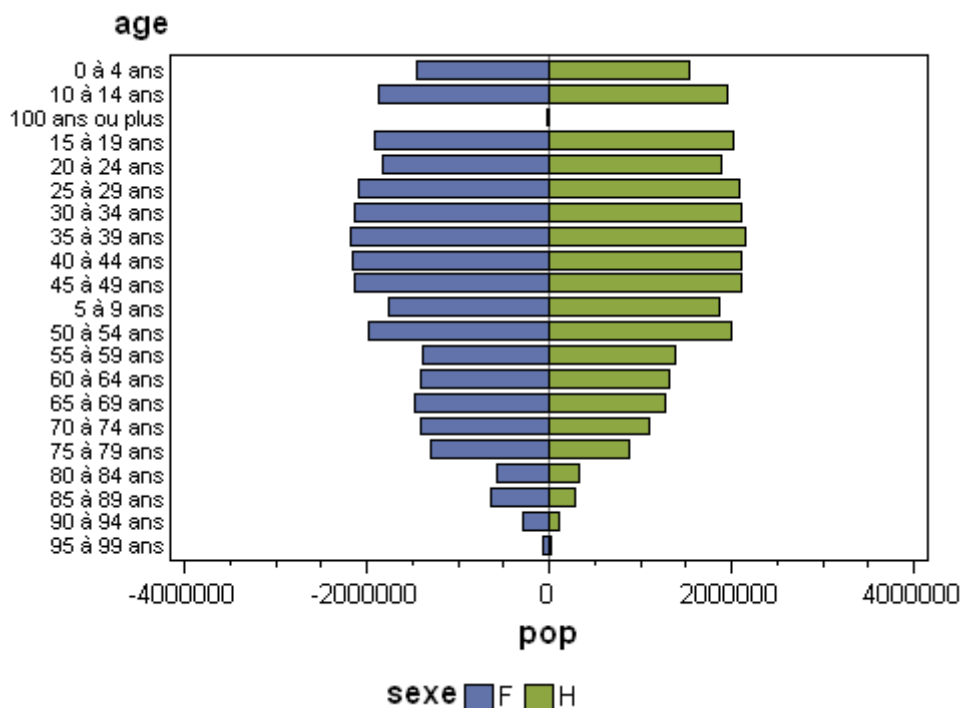
On ajoutera à cette table une variable RANG qui nous permettra d'ordonner comme on le souhaite les tranches d'âges (dans une pyramide, par convention, les plus âgés sont en haut du graphique). On transforme donc légèrement le programme ci-dessus :

```
DATA work.pyramide
      (DROP = f h) ;
SET work.recensement
      NOBS = n ;
rang = n - _N_ ;
sexe = "F" ;
pop = -1 * f ;
OUTPUT ;
sexe = "H" ;
pop = h ;
OUTPUT ;
RUN ;
```

	age	rang	sexe	pop
1	0 à 4 ans	20	F	-1457010
2	0 à 4 ans	20	H	1529915
3	5 à 9 ans	19	F	-1770850
4	5 à 9 ans	19	H	1858444
5	10 à 14 ans	18	F	-1873426
6	10 à 14 ans	18	H	1959694
7	15 à 19 ans	17	F	-1921897
8	15 à 19 ans	17	H	2010204
9	20 à 24 ans	16	F	-1834936
10	20 à 24 ans	16	H	1876612
11	25 à 29 ans	15	F	-2086403

On produit ensuite une première version de la pyramide des âges : POP est la variable statistique, AGE est la variable de bâtons, et un sous-groupe par SEXE est demandé. On obtient alors ceci.

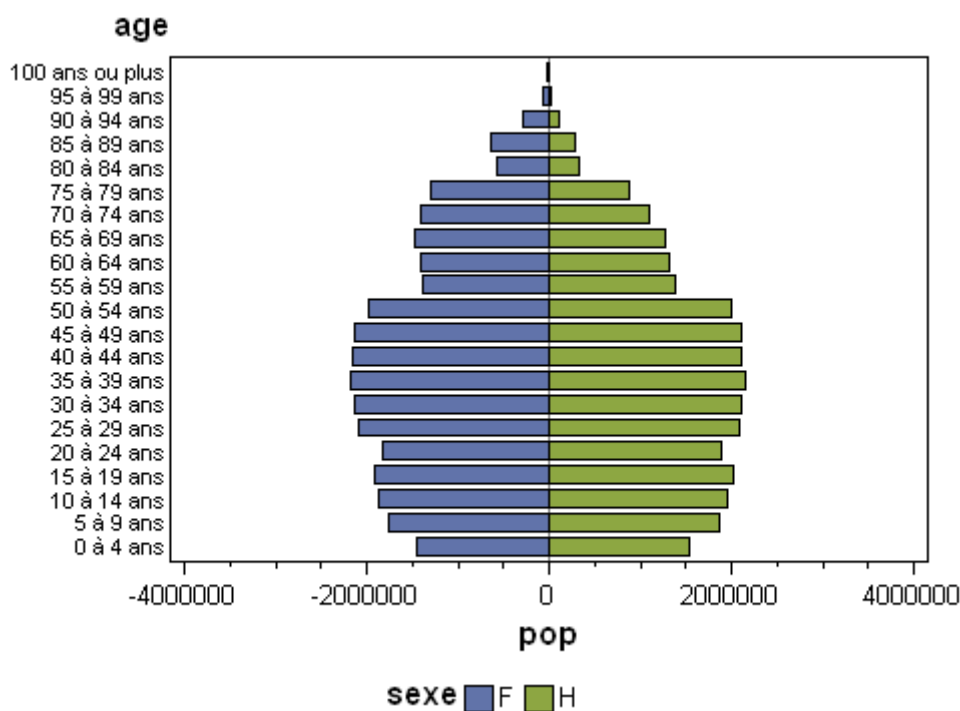
```
PROC GCHART DATA = work.pyramide ;
  HBAR age / DISCRETE NOSTATS
      SUMVAR = pop TYPE = SUM
      SUBGROUP = sexe ;
RUN ; QUIT ;
```



Par rapport à ce premier résultat, il faut avant tout ordonner les bâtons comme on le souhaite. On dispose pour cela de l'option MIDPOINTS, derrière laquelle on énumère, entre guillemets, toutes les valeurs de AGE que l'on veut voir s'afficher, dans cet ordre. Par

exemple, MIDPOINTS = "U.S.A." "CANADA" n'affiche que ces deux valeurs dans le graphique, et dans cet ordre. Il faudrait donc écrire, en ce qui nous concerne, MIDPOINTS = "100 ans ou plus" "95 à 99 ans" ... "5 à 9 ans" "0 à 4 ans". Comme il y a ainsi 20 valeurs à énumérer, il est plus simple (et moins risqué) de le générer par programme avec une macro-variable.

```
PROC SQL NOPRINT ;
  SELECT DISTINCT QUOTE(STRIP(age))
    INTO :valeurs SEPARATED BY " "
  FROM work.pyramide
  ORDER BY rang
  ;
QUIT ;
PROC GCHART DATA = work.pyramide ;
  HBAR age / DISCRETE NOSTATS
    SUMVAR = pop TYPE = SUM
    SUBGROUP = sexe
    MIDPOINTS = &valeurs ;
RUN ; QUIT ;
```



Restent alors des modifications d'ordre cosmétique : choix des couleurs, élimination des graduations sur l'axe horizontal (celui des statistiques), changement de label sur ce même axe.

```
PATTERN1 C = ROSE ;
PATTERN2 C = CYAN ;
AXIS1 LABEL = (J=C "Femmes" "Hommes")
  MINOR=NONE MAJOR=NONE VALUE=NONE ;
```

On ajoute des espaces entre Femmes et Hommes dans le label de l'axe horizontal pour retrouver ces mots de part et d'autre du point 0. On aura ainsi des libellés à peu près centrés pour les deux zones du graphique.

Les options MINOR, MAJOR et VALUE = NONE suppriment graduations et affichage de valeurs.

Quant aux deux instructions PATTERN, elles permettent de jouer sur la couleur des bâtons. L'option COUTLINE dans la procédure GCHART fournira pour sa part le moyen de changer la couleur du trait entourant les bâtons.

Si PATTERN1 et PATTERN2 sont automatiquement associés aux deux premières zones à colorier dans le graphique, AXIS1 n'est associé à l'axe horizontal qu'en ajoutant l'option RAXIS = axis1 dans l'instruction HBAR. On distingue 3 axes dans un graphique en bâtons :

- RAXIS (Response Axis) est l'axe des statistiques ;
- MAXIS (Midpoints Axis) est celui des valeurs de bâtons ;
- GAXIS (Group Axis) est celui des valeurs de groupes.

```
PROC GCHART DATA = work.pyramide ;
  HBAR age / DISCRETE NOSTATS
      SUMVAR = pop TYPE = SUM
      SUBGROUP = sexe NOLEGEND
      MIDPOINTS = &valeurs
      COUTLINE = GRAY
      RAXIS = axis1 ;
RUN ; QUIT ;
```

