Feuille de TP
Mme Malot

1 Une introduction au logiciel SAS

1. Entrer le programme suivant dans l'éditeur de programme

```
data tp;
input numero taille poids sexe $ sexecode;
cards;
1 174 65 m 1
2 169 56 f 2
3 166 48 f 2
4 181 80 m 1
5 168 53 f 2
6 176 76 m 1
7 190 77 m 1
8 159 70 f 2
9 162 60 f 2
10 164 51 f 2
11 160 73 f 2
;
run;
```

2. L'exécuter et afficher les données

3. La table ainsi créer est stockée dans le répertoire de travail work. Si vous fermez le logiciel et que vous le rouvrez, qu'est-il advenu de la table tp?

4. Si vous voulez créer une table de façon permanente, il vous faut la stocker dans un de vos répertoires. Utilisez l'instruction libname pour créer une librairie associée à votre répertoire et stockée y votre table.

5. Le nom des variables ne peut contenir ni d'espace ni plus de 8 caractères. Pour avoir des noms plus précis à l'affichage, utiliser l'instruction label de sorte de voir poids de l'élève et taille de l'élève en lieu et place de poids et taille.

6. Dans la table tp, la variable sexecode donne le sexe de l'individu avec le code 1 pour masculin et 2 pour féminin. Comment faire apparâitre ces mots à la place des chiffres?

7. On souhaite importer les données contenues dans le fichier ozone.xls. En quoi le programme suivant vous aide moyennant des modifications?

```
proc import out=malib.ozone
datafile=" chemin ozone.xls"
DBMS=XLS REPLACE;
SHEET="tp1";
GETNAMES=YES;
MIXED=YES;
run;
```

8. Appliquer la procédure proc contents à la base nouvellement créée.

9. Créer la base de données nombres de façon permanente qui ressemble à:

x	y
5	5
2	-3
4.5	10
3.2	1
2	0

10. Taper et commenter les instructions suivantes :

```
data malib.calcul;
set malib.nombres;
a=x+y; b=x-y; c=x*y; e=min(x,y); f=max(x,y);
g=x/y; h=abs(y); i=exp(x); j=int(x); k=log(y);
l=log10(x); m=sign(y); n=sqrt(x);
run;
```

11. Que fait le programme suivant?

```
data malib.compt;
do i=1 to 100 by 1;
x=rand('binomial',0.4,20);
y=1+x;
x=x-1;
output malib.compt;
end;
run;
```

12. Comment faire pour simuler deux suites aléatoires où $X_1, \ldots, X_n$ sont des variables aléatoires indépendantes uniformes sur $[0, 1]$ et $Y_0 = 0$, $Y_k = Y_{k-1} + X_{k-1}$.

13. Reprendre la table tp et la trier suivant la taille des individus.

14. Reprendre encore une fois la table tp et la trier selon le sexe et utiliser la commande proc print avec l'argument by sexe pour avoir la liste des garçons et la liste des filles.

15. Utiliser maintenant l'instruction keep pour conserver les variables numero, taille, poids et sexecode. Créer ainsi une table tp2.

16. Utiliser maintenant l'instruction drop pour supprimer les variables taille et poids. Créer ainsi une table tp3.

17. Reconstituer la table tp à partir de tp2 et tp3.

18. On souhaite ajouter à la table tp la partie suivante :

numero	sexecode
12	1
13	1
14	2

19. Que fait :

```
data malib.tp6;
set malib.tp;
where sexecode=1;
run;
```

20. Et ceci?

```
data malib.tp7;
set malib.tp;
if sexe='f' then delete;
run;
```

21. Taper et commenter :

```
data malib.garcons malib.filles;
set malib.tp;
if sexecode=1 then output malib.garcons;
if sexecode=2 then output malib.filles;
run;

data malib.garcons2 malib.filles2;
set malib.tp;
select(sexecode);
when(1) output malib.garcons2;
otherwise output malib.filles2;
end;
run;
```

2 Etude d'une variable quantitative

- Exercice 1:

  1. Reprendre la base ozone et créer une variable t qui correspond au numéro de l'observation. Utiliser pour cela t=_n_.
  2. A l'aide de la procédure gplot tracer l'évolution de la variable maxO3.
  3. Obtenir les statistiques les plus courantes sur la variable maxO3.
  4. Tracer un histogramme de maxO3 à l'aide de gchart.
  5. La procédure UNIVARIATE permet d'obtenir un grand nombre de statistiques. Quelle est la valeur de la moyenne, de la variance, de l'écart-type, du maximum, du minimum, de l'étendue, du coefficient de variation, du coefficient d'asymétrie, du coefficient d'applatissement pour maxO3? Quelles sont également la médiane, les centiles et les valeurs extrêmes?
  6. A l'aide de la procédure UNIVARIATE, obtenir le graphe steam and leaf, boxplot, un histogramme avec la superposition des modélisations qui vous semblent envisageables et une qqplot contre la loi qui vous semble la plus réaliste.
  7. Faire une analyse statistique sur la variable T12 qui n'est autre que la température à midi.
  8. Tracer le nuage de points d'abscisse T12 et d'ordonnée maxO3 avec la procédure gplot. Qu'en pensez vous?

- Exercice 2: On s'intéresse aux données contenues dans le fichier salaires.xls qui contient le salaire, le sexe, la catégorie socio-professionnelle (CSP) et le nombre de jours d'absence pour chaque salarié d'une entreprise.

  1. Importer ce jeu de données sous SAS.
  2. En utilisant la procédure gchart avec l'option by ou class, faire une représentation graphique adaptée de la variabloe CSP par sexe.
  3. La variable d'intérêt est CSP. A l'aide de la procédure means obtenir les moyennes et écarts-types pour les hommes, les femmes et pour tous. Combien y a t'il d'hommes et de femmes? Stocker les statistiques dans une table de sortie. Comparer les moyennes par sexe et analyser?
  4. Calculer les variances inter et intra groupes.
  5. La variables d'intérêt est maintenant le salaire. Donner les moyennes et les écarts-types des salaires par sexe. Tracer un histogramme des salaires par sexe.
  6. Faire une décomposition de la variance pour étudier les disparités des salaires entre les hommes et les femmes. Quelle est votre conclusion?
  7. Donner la somme de tous les salaires et le nombre de salariés.
  8. On s'intéresse maintenant aux inégalités de répartition des salaires pour l'ensemble des salariés. Tracer la courbe de Lorentz et calculer l'indice de Gini. Pour cela :
     - Classer les salariés par salaire croissant
     - calculer une variable correspondant aux salaires cumulés
     - associer à chaque individu la part P_2 que représente la masse salariale de l'ensemble des personnes gagnant moins que lui par rapport à l'ensemble de la masse salariale

- créer une variable P_1 associant à chaque individu la proportion de salariés gagnant moins que lui
- tracer la courbe de lorentz : P_2 en fonction de P_1. Superposer la 1ère bissectrice et commenter.
- calculer l'indice de Gini qui est égal à 2 fois l'aire entre la courbe de Lorentz et la 1ère bissectrice.
- Créer un découpage des salaires suivant les quartiles : on crée une variable auxiliaire groupe qui vaut 1 si l'individu gagne moins que le 1er quartile et ainsi de suite.
- A l'aide de la procédure means, stocker dans une table de sortie nommée sorite, l'effectif, la moyenne et la variance de chacun des groupes de salaires.