IMIS : Master 1
Université Paris Est – Marne la Vallée

TP 3. Etudes de variables quantitatives & qualitatives

1. Lecture d'un fichier, 'informat'

Sous SAS, il est possible de lire un fichier texte dont les lignes correspondent à un individu. La syntaxe est `INFILE`.

```
DATA tablecrée;
INFILE 'c:\....\donnée.txt';
INPUT description des variable (et 'informat');
RUN;
```

La descriptions des variables est une liste de noms pour ces variables,. On peut faire suivre le nom par un format de lecture (le 'informat') pour préciser à SAS comment la variable est écrite dans le fichier qu'il lit (c.f. la table des 'informats').
Certaines options de la commande `INFILE` sont très utiles :
`FIRSTOBS=n` numéro de la première ligne à lire
`LRECL=n` nombre maximal de caractères par ligne
`MISSOVER` s'il n'y a pas assez de données par lignes, dire à SAS de compléter par des données manquantes
`DLM=";"` délimitateur entre les données (ici, c'est ;)
`DSD` si SAS rencontres deux délimitateurs consécutifs, il doit comprendre que la donnée est manquante.

1. Lire dans une table le fichier tp1.txt créé (c.f. feuille TP1). Les variables étaient : Taille Poids Sexe Age. Pour quelles variables a-t-on précisé un 'informat'. Que fait SAS quand on ne lui précise pas ?

2. Lire dans une table le fichier tp3_donnée1.txt. Les variables sont : Taille Enfants Voiture Statut Prof.
   Pour être plus sûr que la lecture se déroule bien, on peut aussi indiquer à SAS où est située la position du début de la donnée dans chaque ligne. Exemple :
   `INPUT @1 taille @6 Enfant .......`

3. Il y a des données manquantes pour la variable 'Enfant'. Ecrire une procédure créant une nouvelle table dans laquelle ces données manquantes sont remplacées par 0.

4. Lire le fichier tp3_donnée2.txt. Les données sont : NOM CYL PUIS LON LAR POIDS VITESSE FINITION PRIX (rq. le code pour tabulation est "09"x).
   Refaire la même chose en ne retenant que les 4 premières lettres pour caractériser le véhicule.

2. Analyse de variables qualitatives

La procédure `FREQ` permet de connaître la répartition des modalité d'une variable. La syntaxe est :

```
PROC FREQ DATA=donnée;
TABLE variable;
RUN;
```

1. Sur les données du fichier tp3_donnée1.txt, afficher la fréquence de chacune des modalités de la variable 'statut' puis celle de 'Prof'.

2. Pour obtenir un tableau de dimension deux pour la répartition jointes des variables 'statut' et 'prof' il suffit d'utiliser :
   ```
   TABLE statut*prof;
   ```
   Que représente les deux dernières lignes du tableau obtenu ?

3. Il est possible de faire un test de χ^2 d'indépendance sur les deux variables en rajoutant l'option `/ CHISQ` juste après les variables. Que peut-on-dire des variables 'statut' et 'prof'.

4. Grâce a la commande `WEIGHT poids`, que l'on peut ajouter à presque toutes les fonctions statistiques de SAS, on peut pondérer les observations (la variable poids contient les effectifs par lesquels il faut pondérer).
   Le fichier 'tp3permis' contient les données de réussite au permis de conduire pour une population de 126 individus. On a retenu comme première variable le nombre de tentatives au permis (de une à trois) et comme deuxième variable le sexe des candidats. La troisième variable contient l'effectif. Charger le fichier dans une table (Attention : il faut commencer la lecture par la deuxième ligne du fichier).

5. Afficher la répartition des variables 'sexe' et 'tentative' et la répartition jointe (pondérer par les effectifs sinon SAS comprend qu'il n'y a que 6 individus !).

6. Tester l'indépendance.
   On peut afficher des graphes pour représenter ces tableaux d'effectifs. La procédure est `GCHART` avec la syntaxe :
   ```
   PROC GCHART DATA=donnée;
   instructions / options;
   RUN; QUIT;
   ```
   où les instructions sont `PIE variable` pour avoir une représentation en diagramme circulaire ; et `HBAR variable` et `VBAR variable` pour des diagrammes en bâtons.

7. Sur les données du fichier tp3_donnée1.txt, représenter par des diagrammes en bâtons puis circulaires la répartitions de la modalité 'statut'.

3. Analyse de variables quantitatives

Les procédures `MEANS` et `UNIVARIATE` permettent de calculer les moments empiriques d'un échantillon d'une variable.

1. Etudier la distribution de la variable Taille du fichier tp3_donnée1.txt en utilisant la commande `UNIVARIATE` sans options. Commenter le plus possible des sorties de SAS.
2. Tester la normalité de cette variable. Quelle est la p-valeur du test de Kolmogorov ? Qu'est ce que cela signifie ?
3. Effectuer un test de l'hypothèse que la taille moyenne est 1m75 (utiliser l'option `MU0=1.75`). Quel est le test utilisé ?

   Il est aussi possible d'avoir une estimation non-paramétrique de la densité d'une variable. Pour cela on utilise la commande `histogram` dans la procédure `UNIVARIATE`.
4. Tracer un histogramme de la loi de la variable taille. Une estimation plus précise de la densité de la loi est obtenue avec l'option `/ KERNEL`