Université Joseph Fourier
L2/STA230

TP 1 : Description d'un jeu de données

Cecile.Amblard@imag.fr

L'objectif de ce TP est de décrire un jeu de données : le lire sous forme de tableau, extraire une partie des individus ou une partie des variables caractéristiques, construire un tableau de distribution, faire un histogramme ou diagramme en bâtons et calculer les caractéristiques centrales usuelles (mode, médiane, moyenne) ainsi que l'écart type : la dispersion moyenne des données autour de la moyenne.
Les fonctions d'Open Office à utiliser sont : NB.SI(), SOMMEPROD(), MEDIANE(), MODE(), MOYENNE, VAR(), VAR.P().....

Nous nous intéressons au jeu de données suivant : BB.csv. Ce jeu provient d'une étude réalisée à Baystate Medical Center, Springfield en 1986. L'étude s'intéresse aux facteurs de risque maternels pouvant conduire à la naissance d'un bébé de faible poids, c'est à dire inférieur à 2,5 Kg. Le jeu de données est un tableau de 189 lignes et 10 colonnes : chaque couple "mère-bébé" est décrit par 10 variables :

- low : indicateur binaire du poids du bébé inférieur à 2,5 Kg,
- age : age de la mère (en années)
- lwt : poids de la mère (en livres) lors du dernier mois avant la naissance,
- race : race de la mère (1= blanc, 2= noir, 3= autre)
- smoke : status Fumeur-Non fumeur pendant la grosesse,
- ptl : nombre d'accouchements prématuré déjà vécus,
- ht : historique d'hypertention,
- ui : présence d'infection urinaire,
- ftv : nombre de visites pendant le premier trimestre,
- bwt : poids du bébé (en grammes) à la naissance.

1. Lancez Open Office Calc.
2. A partir du menu *Fichier*, ouvrez le fichier de données "BB.csv". *Vous devez sélectionner comme Type de fichier le type Texte (*.csv, *.txt) et choisir comme séparateur la virgule.*
3. Enregistrez ce document sous le nom "TP1.ods".
4. **Nous nous intéressons à la variable discrète : age**
   (a) Nommez le vecteur des ages : **ages**. *Sélectionnez le vecteur des ages et remplacez en haut à gauche l'adresse de type C2 :C190 par le nom* **ages**.
   (b) Changez de nom ce vecteur. *Utilisez le menu Insertion > Nom> Definir, supprimez le nom ages puis renommez votre plage de données.*
   (c) Quel est l'age minimal et l'age maximal ? *Utilisez MIN() et MAX()*
   (d) Construisez le tableau de distribution donnant la répartition des ages des mères. Les colonnes du tableau doivent être les ages, les effectifs, les effectifs cumulés,

les fréquences, les fréquences cumulées. *Pour les effectifs, utilisez la fonction NB.SI(). En utilisant Format>Cellule, affichez les fréquences et fréquences cumulées en pourcentage avec 2 chiffres significatifs après la virgule. En aucun cas vous ne devez multiplier par 100 les proportions obtenues.*

(e) Réalisez l'histogramme des fréquences (diagramme en bâtons) associé à ce tableau de distribution. *il est possible de changer la largeur des bâtons en sélectionnant sur le graphique à la souris la série de données puis en agrandissant l'espace entre 2 rectangles en utilisant Formater la série> Option> Espacement*

(f) Calculez et interprétez les caractéristiques centrales : modes, médiane, moyenne, à partir des données brutes (*MEDIANE(), MOYENNE()*) et à partir du tableau de distribution, (*SOMMEPROD()*).

(g) Calculez la variance et l'écart type à partir des données brutes *VAR.P()* et du tableau de distribution *SOMMEPOD()*. Interprétez l'écart type.

5. **Nous nous intéressons à la variable continue : le poids des bébés**

(a) Nommez le vecteur des poids : **w**.

(b) Construisez le tableau de distribution donnant la répartition des poids des bébés. Chaque modalité poids est ici un intervalle d'amplitude fixe 400g. L'intervalle est matérialisé dans le tableur sous forme de 2 cellules : la valeur min et la valeur max de l'intervalle. *Vous pouvez calculer tout d'abord les effectifs cumulées en utilisant la fonction NB.SI() avec comme critère "< poids max de l'intervalle", puis en déduire les autres colonnes du tableau. Attention à l'ecriture du critère NB.SI(w;"<"&B23)*

(c) Réalisez l'histogramme des fréquences associé à ce tableau de distribution.

(d) Calculez et interprétez les caractéristiques centrales : modes, médiane, moyenne, à partir des données brutes (*MEDIANE(), MOYENNE()*) et à partir du tableau de distribution, (*interpolation linéaire et SOMMEPROD() avec comme arguments le vecteur des centres de classe et le vecteur des fréquences.*) Comparez les valeurs obtenues.

(e) Calculez la variance et l'écart type à partir des données brutes *VAR.P()* et du tableau de distribution *SOMMEPOD()*. Interprétez l'écart type.

(f) Construisez un autre tableau de distribution donnant la répartition des poids des bébés. Chaque modalité poids est ici un intervalle d'amplitude tel que chaque inervalle de poids doit contenir 10% des bébés. *La colonne des fréquences est un vecteur de 10 valeurs 10 %. Vous pouvez utiliser la fonction CENTILE() pour calculer les bornes des classes ou bien des interpolations linéaires en utilisant les fréquences cumulées du premier tableau. N 'oubliez pas la colonne des densités.*

(g) Pourquoi n'est il pas possible de représenter l'histogramme des poids à partir de ce tableau avec Open Offfice ?

(h) Calculez et interprétez les caractéristiques centrales : modes, médiane, moyenne, à partir de ce tableau de distribution. Comparez les valeurs obtenues avec celles calculées sur le premier tableau.