TP1 : Analyse univariée

Durée : 6h

L'objectif de ce TP est d'exposer les outils élémentaires permettant de présenter de façon synthétique et claire une série de données brute, et d'en résumer les principales caractéristiques. Ce TP est illustré grâce langage de programmation et environnement statistique  .

Quelques mots sur R

R *est un langage et un environnement logiciel open source pour les calculs statistiques et graphiques. R devient incontournable dans le traitement exploratoire et statistique des données.*

Site Internet : Statistiques avec R
http://zoonek2.free.fr/UNIX/48_R_2004/all.html

Les commandes nécessaires aux traitements des données seront précisées au fur-et-à-mesure des exercices. Voici quelques commandes de base :

- `# pour écrire un commentaire`
- `help(`*`fonction`*`)  # pour obtenir l'aide sur une fonction`
- `ls()  # permet d'afficher tous les objets de la session de travail`
- `rm()  # efface les objets de la session de travail`
- `q()  # permet de quitter R`

La session de travail avec tous les objets qui auront été définis dedans (matrices, dataframes, fonctions, …) peut être sauvegardée puis rechargée à chaque utilisation :

- `getwd()  # affiche le répertoire courant`
- `setwd("K:/ING1/GI/Statistiques Descriptives")  # permet de fixer le chemin d'accès au répertoire. attention d'utiliser / et non \`
- `save.image(file="nom.Rdata") # sauvegarde la session de travail dans le fichier nom.RData du répertoire courant`
- `load("nom.Rdata") # recharge la session de travail`

Exercice 1 — *Type de variables*

Pour chaque variable, déterminer sa nature et préciser son échelle de mesure. Il est possible de proposer plusieurs échelles de mesure afin de changer la nature de la variable.

	Qualitatif			Quantitatif	
	nominal	dichotomique	ordinal	discret	continu
Nb d'enfants dans un foyer					
Age					
Département					
Note sous forme A, B, …					
Date de naissance					
Classe d'âge					
Note sur 20					
Abstentionniste					
Sexe					

Exercice 2 — *Questions de réflexion*

1) Normalement, on ne fait des calculs de moyennes que sur les caractères quantitatifs. Si on code un caractère dichotomique avec les codes 0 pour l'absence de la caractéristique et 1 pour la présence de la caractéristique. Expliquer pourquoi dans ce cas précis, la moyenne des codes sur une série donnée a vraiment un sens.
2) Comment peut-on transformer un caractère quantitatif en caractère ordinal ? On différenciera les caractères discrets des caractères continus. Quelle peut être la conséquence néfaste de cette transformation.
3) Pourquoi est-il a priori incorrect de transformer un caractère ordinal en un caractère quantitatif discret ?
4) Pour une variable quantitative, comment se situe la moyenne par rapport à la médiane lorsque le coefficient d'asymétrie (skewness) est négatif ?
5) Démontrer qu'une série centrée-réduite est de moyenne nulle et de variance égale à 1.

Exercice 3 — *Série synthétique*

Pour la série de valeurs observées ci-dessous,

4	0	2	2	8	5

1) Calculer la moyenne, la médiane, le mode, les quartiles, l'étendue inter-quartiles, l'écart-type et le coefficient de variation. Les représenter graphiquement.

```
> x <- c(4,0,2,2,8,5)        # création d'un vecteur
> sort(x)                    # tri par valeurs croissantes
> median(x)                  # médiane
> mean(x)                    # moyenne
> quantile(x)                # quantiles
> Q1 <- quantile(x)[2]
> Q3 <- quantile(x)[4]
> etendue <- Q3-Q1           # étendue inter-quartiles
> sd(x)                      # écart-type(standard deviation)
> var(x)                     # variance
> boxplot(x)                 # trace la boite de Tuckey
> points(mean(x))            # ajoute la moyenne

# il est possible de modifier les paramètres graphiques dans
toutes les fonctions graphiques : col="red",lwd=5 …
```

2) Refaire la même chose en ajoutant 10 à la série. Conclusion.
3) Refaire la même chose en multipliant la série par 2. Conclusion.
4) Refaire la même chose en ajoutant la valeur supplémentaire 25 à la série. Conclusion.

```
> y <- x+10     # 10 s'ajoute à chaque composante de x
> z <- x*2      # chaque composante de x est multipliée par 2
> w <- c(x,25)  # concaténation
```

Exercice 4 — *Elèves*

Le fichier ElevesData.csv contient les résultats d'un sondage d'une classe de 58 élèves concernant des caractères physiques.

1) Importer le fichier ElevesData.csv dans votre session de travail R.

L'importation des données des données avec R se fait sous forme d'un objet appelé ***data.frame****. Il s'agit d'un tableau avec en ligne les observations (individus) et en colonne les variables observées. A la différence d'une matrice, un data.frame est un objet dont on peut extraire des renseignements (nom et type des variables, résumé,…).*

```
> tab <- read.table("ElevesData.csv",header=TRUE,sep=";",dec=".")
                                                  # importe le fichier
> names(tab)     # affiche le nom des variables
> tab$TAILLE     # affiche uniquement la variable TAILLE
> dim(tab)       # donne la dimension du tableau (nrow(tab) et ncol(tab)
                   pour le nb de lignes ou de colonnes)
> tab[,1:3]      # affiche les 3eres colonnes
> tab[10:12,]    # affiche les lignes 10, 11, 12
```

2) Déterminer la nature de chaque caractère et faire la représentation graphique des effectifs appropriés.

```
> ######
> str(tab)       # affiche la nature des variables
> summary(tab)   # donne un résumé numérique des variables

> #### Variables qualitatives nominales
> tab$SEXE <- as.factor(tab$SEXE)    # transforme la variable SEXE en
                                       variable qualitative
> effSEXE <- table(tab$SEXE)         # calcul des effectifs
> pie(effSEXE,labels=c("H","F"))     # trace le diagramme circulaire
> title("Répartition H/F")           # titre du graphique
> effYEUX <- table(tab$YEUX)
> barplot(as.matrix(effYEUX),col=heat.colors(3),legend.text=TRUE)
                                                    # diagramme en barre
> #### Variables ordinales ou discrètes
> effPOINTURE <- table(tab$POINTURE)
> barplot(effPOINTURE, main="Distribution Pointures") # diagramme en
                                                                bâtons
> #### Variables continues
> hist(tab$POIDS,breaks=5,freq=FALSE,xlab="POIDS",main="Histogramme")
```

3) Transformer le caractère taille en caractère qualitatif en considérant le découpage suivant : <150cm = petit, >180 = grand, sinon moyen. Faire sa représentation graphique et comparer avec la précédente.

```
> vect <- hist(tab$TAILLE,breaks=c(131,149,180,198))$counts
> barplot(vect,names.arg=c("petits","moyens","grands"))
```

4) Créer un tableau croisé des effectifs des couleurs des yeux chez les hommes et chez les femmes. Peut-on comparer ?

```
> tableau <- table(tab$SEXE,tab$YEUX)
> addmargins(tableau)     # ajoute les totaux en ligne et colonne
> prop.table(tableau)     # calcule les pourcentages
```

Exercice 5 — *Salaires en France*

Le fichier SalairesData.xls donne un récapitulatif de l'évolution des salaires en France entre 1950 et 2006. La première partie concerne tous les salariés, la deuxième et la troisième concernent les hommes puis les femmes.

1) La valeur de la case B12 signifie que : « En 1955, le salaire moyen était de 700 euros pour l'ensemble des salariés ». Rédigez des phrases similaires pour chacune des cases : C25, D111 et E178.

2) Quel type d'information apportent les colonnes F, G et H. Complétez le titre en F4.
3) Peut-on retrouver le salaire moyen de l'ensemble des salariés en 1999 (B56) à partir du salaire moyen pour les hommes (B114) et de celui des femmes (B172)? Sinon, quelle est l'information manquante?
4) Proposez un autre indicateur de dispersion. Est-il possible de le calculer?

Exercice 6 — *Dépenses éducation USA*

Le fichier EducationUSAData.csv présente les dépenses moyennes (en $) par élève et les résultats d'évaluation nationale pour chacun des états unis d'Amérique.

1) Calculer la médiane, la moyenne et l'écart-type de chacune des deux séries (dépenses par élève et résultats d'évaluation). Faire une boite de Tukey pour chacune des séries.
2) Peut-on comparer les degrés de dispersion de chacune des séries à partir de leur écart-type.
3) Calculer les rapports écart-type sur moyenne et donner une première impression sur la corrélation entre dépenses et résultats. Illustrer votre conclusion à l'aide d'un graphique.

Exercice 7 — *Salariés d'une entreprise*

Le fichier SalariesData.csv présente les salariés d'une entreprise ayant 3 sites (A, B et C). On y indique leur sexe, leur salaire annuel (en milliers d'euros), leur catégorie (cadre supérieur, moyen ou ouvrier employé), leur âge et leur site.

1) Calculer le salaire moyen et le salaire médian. Que signifie la différence entre ces deux valeurs?
2) Calculer le 1^{er} et le 3eme quartile, ainsi que le 1^{er} et le $9^{ème}$ déciles de la série des salaires.
3) Lesquels faudrait-il mettre en avant lors d'une négociation salariale, et pourquoi?
4) Etablir un tableau avec effectifs ou fréquences pour la série des catégories et celle des établissements. Faire une représentation graphique adaptée.
5) Représenter les variables discrètes salaire et âge? Cette représentation vous semble t'elle adaptée?
6) Effectuer un regroupement par classe des salaires puis des âges (pas plus de 5 classes). Recalculer la moyenne et comparer avec la série d'origine.
7) Faire une étude comparative des âges de chaque catégorie de salariés (CS, CM, OE), en termes de tendances centrales et mesures de dispersion. Faire une représentation graphique de cette comparaison.