Analyse de données
M1 Statistique et économétrie - 2012
V. Monbet
Exploration des données - Analyse factorielle

Les objectifs de ce TD sont

1. de revoir le cours de statistique exploratoire,
2. d'apprendre à mener une analyse descriptive simple sous SAS et sous R pour un ensemble d'observations de variables aléatoires quantitatives,

1 Questions de cours

1. Dans un boxplot (boîte à moustaches), est-il possible que la moyenne et la médiane soit en dehors des barres correspondant aux quartiles ?
2. Quel poucentage de données s'attend-on à voir en dehors des barres extrêmes si on suppose que les données sont distribuées suivant une loi de Gauss de moyenne 0 et de variance σ^2.
   (a) Répondre à la question par des arguments théoriques.
   (b) Proposer une programme de simulation permettant de vérifier le résultat pour différentes tailles d'échantillon.
   On utilisera le logiciel R. La fonction `rnorm` permet de générer un échantillon distribué suivant une loi de Gauss. La commande `?rnorm` permet d'obtenir l'aide en ligne.
3. Est-il possible que les 5 nombres résumé soient égaux ? Si oui sous quelle(s) condition(s) ?
4. Vrai ou faux : la hauteur des barres d'un histogramme est égale à la fréquence relative avec laquelle une observation tombe dans l'interval correspondant.
5. Les données suivantes représentent la taille de 13 étudiants de master :

   1.72, 1.83, 1.74, 1.79, 1.94, 1.81, 1.66, 1.60, 1.78, 1.77, 1.85, 1.70, 1.76.

   (a) Trouver les cinq nombres résumé.

(b) Construire une boite à moustaches.

(c) Tracer un histogramme pour ce jeu de données.

2 Les données

Les données (fichier `depart.dat` sur la page du cours `http://perso.univ-rennes1.fr/valerie.monbet/Cours_AD/AD.html`) proviennent du Groupe d'Etude et de Reflexion Inter-régional (GERI). Elles portent sur 4 grands thèmes : la démographie, l'emploi, la fiscalité directe locale, la criminalité. Les indicateurs sont mesurés sur l'ensemble des départements français métropolitains ainsi que la Corse pendant l'année 1990, ils sont, pour la plupart, des taux calculés relativement à la population totale du département concerné. On observe les variables suivantes :

- numéro de département,
- code du département,
- code de la région,
- `URBR` indicateur de concentration de la population mesurant le caractère urbain ou rural du département,
- `TXCR` taux de croissance de la population sur la période intercensitaire 1983-1990,
- `JEUN` part des 0-19 ans dans la population totale,
- `AGE` part des plus de 65 ans dans la population totale,
- `FE90` taux de fécondité (pour 1000) égal au nombre de naissances rapporté au nombre de femmes fécondes (15 à 49 ans) en moyenne triennale,
- `ETRA` part des étrangers dans la population totale,
- `CHOM` taux de chomage,
- `CRIM` taux de criminanité : nombre de délits par habitant,
- `FISC` produit, en francs constants 1990 et par habitant des quatre taxes locales (professionelle, habitation, foncier bati, foncier non bati).

On observe également les parts de chaque profession en catégorie socioprofessionnelle (PCS) dans la population active occupée du département :

- `AGRI` : agriculteurs,
- `ARTI` : artisans,
- `CADR` : cadres supérieurs,
- `EMPL` : employés,
- `OUVR` : ouvriers,
- `PROF` : professions intermédiaires.

3 Exploration avec le logiciel SAS

3.1 Lecture des données

Utiliser, par exemple, les instructions suivantes pour lire le fichier de données.

```
libname TPexplor '~/AnalyseDonnees/TP1' ;
data TPexplor.depart ;
infile '~/AnalyseDonnees/TP1/depart.dat' ;
input num $ depart $ region $ txcr etra urbr jeun age chom agri
    arti cadr empl ouvr prof fisc crim fe90 ;
run;
```

3.2 Analyse interactive des données

En utilisant l'outil d'analyse interactive les données (ou SAS Insight)

- Visualiser les distributions des différentes variables (histogrammes, boites à moustaches).
- Tracer des nuages de points des variables deux à deux (scatter plot).
- Etudier, rapidement, la relation linéaire entre la variable criminalité et les autres (fit).

Choisir dans le menu déroulant `Solution` la ligne `Analyse Interactive des données` puis sélectionner les variables à étudier et utiliser ensuite le menu `Analyze`)

3.3 PROC UNIVARIATE et PROC CORR

Répondre de nouveau aux questions précédentes à l'aide des procédures
`PROC UNIVARIATE` et `PROC CORR`.
Dans la procedure `UNIVARIATE` l'option `KERNEL` permet d'obtenir le graphe de l'estimateur à noyau, l'option `K=` permet de choisir le noyau et l'option `C=` la largeur de fenêtre standardisée. Pour `C=`, on peut choisir une valeur à la main ou choisir `C=MISE`. Dans le premier cas, on trace l'estimation pour plusieurs valeurs de C et on retient celle qui nous semble la plus raisonnable. Dans le second cas, la largeur de fenêtre est choisie telle que celle ci minimise le critère AMISE (approximate mean integrated square error) pour une loi de Gauss ayant la moyenne et la variance estimées dans l'échantillon.
Tester les deux approches et commenter.
Rq : le plus souvent on combine ces deux approches, la seconde donnant une valeur intiale cohérente pour la première.

```
proc univariate data = TPexplor.depart plots ;
var crim empl ouvr fe90 jeun ;
run ;
```

```
proc univariate data = TPexplor.depart plots ;
histogram crim empl ouvr fe90 jeun ;
run ;


proc univariate data = TPexplor.depart plots ;
histogram crim  / KERNEL (K=NORMAL C=.3);
run;
```

On peut aussi utiliser cette fonction pour ajuster des modèles de loi. Les paramètres sont alors estimés par maximum de vraisemblance.

```
proc univariate data = TPexplor.depart plots ;
histogram crim  / GAMMA (THETA=EST ALPHA=EST SIGMA=EST);
run;


ods graphics on;
title 'Données des crimes';
proc corr data=TPexplor.depart plots/*=matrix(histogram)*/;
  var crim empl ouvr fe90 jeun ;
run;
ods graphics off;
```

4 Exploration avec le logiciel R

1. Répondre aux mêmes questions que précédemment en utilisant le logiciel R et en vous aidant des commandes ci-dessous. On rappelle que la commande `?` suivie du nom d'une fonction permet d'ouvrir l'aide en ligne pour cette fonction.
   On observe qur les figures que, par défaut, l'abscisse des histogrammes est `Frequency`. Est-ce vraiment une fréquence ou un effectif ? Xomment passe-t'on en fréquence ?
2. Peut-on améliorer les estimations par histogramme en faisant varier la largeur de bande ? Si oui, proposer une solution.
3. Dans les commandes ci-dessous, la fonction `density` calcule les estimations par les estimateurs à noyau. L'option `bw=` permet de choisir la largeur de fenêtre. Le choix proposé ci-dessous est vraisemblablement peu pertinent.
   (a) En vous aidant du cours, proposer un meilleur choix. Mettez le en oeuvre.
   (b) Choisit-on la même largeur de fenêtre pour toutes les variables ? Pourquoi ?

```
dep <- read.table("~/AnalyseDonnees/TP1/depart_names.dat",header=TRUE)
str(dep)
dep$num <- factor(dep$num)
```

```
summary(dep)
# Histogrammes
list_var = c("txcr","etra","urbr","jeun","age",
"chom","agri","arti","empl","ouvr","prof","fisc","crim","fe90")
par(mfrow=c(4,4)) # On divise la fenêtre graphique en 16 espaces
for (k in (1:length(list_var))) {
   hist(dep[,k+3],xlab="",main=list_var[k])}
# Estimateurs à noyau
X11() # On ouvre une nouvelle fenêtre graphique
   par(mfrow=c(4,4))
for (k in (1:length(list_var))) {
   d = density(dep[,k+3],bw=1)
   plot(d$x,d$y,xlab="",main=list_var[k],type="l")}
# Scatter plots
pairs(dep[,4:14],pch=16)
```