Analyse exploratoire des données

Objectifs

Au travers de l'analyse exploratoire des données, on cherche essentiellement à résumer la distribution de chaque variable (**approche univariée**) ainsi que les relations entre les variables (**approche bivariée** essentiellement), dont les caractéristiques pourraient suggérer un recodage ou une transformation des mesures (Tukey, 1977).

Objectifs (suite)

Plutôt que de modéliser directement les données, on s'attachera donc dans un premier temps à les **décrire** à l'aide de résumés numériques et graphiques. L'idée est de caractériser la **forme** d'une distribution et d'identifier les éventuelles **valeurs influentes**.

On profitera de cette approche pour présenter les principales fonctionnalités graphiques de R, en particulier l'interface `lattice` (Sarkar, 2008 ; Murrel, 2005).

Un jeu de données d'exemple

The low birth weight study

Il s'agit d'une étude prospective visant à identifier les facteurs de risque associés à la naissance de bébés dont le poids est inférieur à la norme (2,5 kg). Les données proviennent de 189 femmes, dont 59 ont accouché d'un enfant en sous-poids. Parmi les variables d'intérêt figurent l'âge de la mère, le poids de la mère lors des dernières menstruations, l'ethnicité de la mère et le nombre de visites médicales durant le premier trimestre de grossesse (Hosmer, 1989).

La base de données `birthwt`

Elle est disponible sous R dans le package `MASS` :

```
data(birthwt, package="MASS")
```

Traitement préalable

Les données fournies dans R nécessitent quelques recodages:

```
str(birthwt)
summary(birthwt)
birthwt <- within(birthwt, {
  low <- factor(low, labels=c("No","Yes"))
  race <- factor(race, labels=c("White","Black","Other"))
  smoke <- factor(smoke, labels=c("No","Yes"))
  ui <- factor(ui, labels=c("No","Yes"))
  ht <- factor(ht, labels=c("No","Yes"))
})
```

Traitement préalable (suite)

Il est intéressant d'ajouter les **unités de mesure**, pour s'en souvenir lorsqu'on en a besoin (le poids de la mère est en pounds, celui des bébés en kg !).

```
library(Hmisc)
birthwt <- within(birthwt, {
  units(age) <- "years"
  units(lwt) <- "pounds"
})
```

Valeurs extrêmes, atypiques ou outliers

filter.r

Il n'y a pas vraiment de consensus sur la dénomination correcte des valeurs qui "ne ressemblent pas" à la majorité des valeurs observées.

Une **valeur extrême** ou atypique est souvent assimilée à une valeur qui est située à plus de 1.5 x IQR des quartiles supérieurs et inférieurs.

Valeurs extrêmes, atypiques ou outliers (suite)

Un **outlier** est une valeur susceptible d'influencer les résultats obtenus par un modèle statistique.

```
idx <- sapply(birthwt, is.numeric)
bwt <- apply(birthwt[,idx], 2, scale)
boxplot(bwt)
```

Valeurs extrêmes, atypiques ou outliers (suite)

On peut aussi chercher des "patterns" multivariés (p.ex., des individus avec des valeurs sytématiquement élevées ou basses).

```
parallel(bwt, groups=birthwt$low, horiz=FALSE)
idx <- apply(bwt, 2, filter.perc, cutoff=c(.01,.99),
             collate=TRUE)
my.col <- as.numeric(1:nrow(bwt) %in% unique(unlist(idx)))+1
splom(~ bwt, pch=19, col=my.col, alpha=.5, cex=.6)
```

Résumé de la structure de données

Un aperçu synthétique des données, stratifié par groupe de poids des bébés, peut être obtenu comme suit :

```
summary(low ~ ., data=birthwt[,-10], method="reverse")
library(latticeExtra)
marginal.plot(birthwt, data=birthwt, groups=low)
```

Synthèse numérique et transformation

On peut aussi recoder certaines des variables discrètes (`ftv` et `ptl`) en variables binaires, pour la description ou pour la modélisation :

```
bwt.df <- transform(birthwt[,-10],
                    ftv=factor(ftv>0, lab=c("No","Yes")),
                    ptl=factor(ptl>0, lab=c("None","1+")))
summary(low ~ ., data=bwt.df, method="reverse")
```

Synthèse numérique et transformation (suite)

Il n'est pas nécessaire de stratifier pour résumer les données, mais dans le cas présent il est intéressant de vérifier la **distribution** des variables pour les deux groupes de bébés.

En ajoutant l'option `overall=TRUE` dans l'expression ci-dessus, on obtient également le résumé numérique sur l'ensemble de l'échantillon.

Synthèse numérique et transformation (suite)

Concernant les unités de mesure, on peut convertir "à la volée" lors de l'appel à des fonctions comme `mean` ou `summary.formula` (ci-dessus) :

```
mean(birthwt$lwt/2.2)  # poids de la mère en kg
summary(lwt/2.2 ~ low + race, data=birthwt)
```

Synthèse numérique et transformation (2)

D'autres types de résumés numériques peuvent être produits avec `summary.formula`, en particulier des tableaux croisés ou des descriptions stratifiées.

Pour plus d'informations, consulter l'excellent guide `Hmisc` :

http://biostat.mc.vanderbilt.edu/wiki/Main/Hmisc

Qu'est-ce qu'une distribution?

Considérons l'âge des mères, variable numérique souvent assimilée à une variable continue mais qui ici prend plutôt des valeurs discrètes. La plupart des valeurs prises par la variable age semble se concentrer autour de la tranche 20-25 ans.

```
stripplot(~ age, data=birthwt, jitter.data=TRUE,
          amount=.3, aspect=.3, cex=.6)
```

Résumé numérique

Mesures de **tendance centrale** (moyenne, médiane) associées à des mesures de **dispersion relative** (écart-type, IQR).

```
# Tukey's five-point summary
summary(birthwt$age)
quantile(birthwt$age, probs=c(.1, .25, .5, .75, .9))
desc <- function(x, dig=2)
 round(c(ety=sd(x), iqr=IQR(x), "max-min"=diff(range(x))),
          digits=dig)
desc(birthwt$age)
```

Résumé numérique (suite)

Autres possibilités : **Estimateurs robustes**

- Moyennée tronquée : `mean(birthwt$age, trim=.025)`
- Déviation médiane absolue : `mad(birthwt$age)`