Analyse exploratoire des données

Introduction à R pour la recherche biomédicale

http://www.aliquote.org/cours/2012_biomed

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).

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` (Murrell, 2005; Sarkar, 2008).

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. C'est une des études qui sert de base aux traitements statistiques présentés dans Hosmer & Lemeshow (1989).

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"))
})
```

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 $1.5 \times$ IQR des quartiles supérieurs et inférieurs. 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)
```

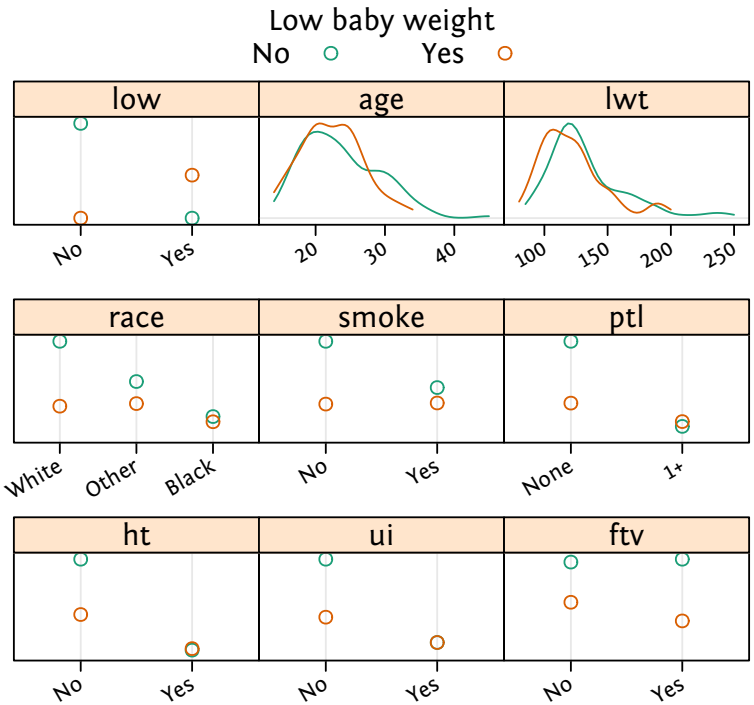
On peut aussi chercher des "patterns" multivariés (p.ex., des individus avec des valeurs sytématiquement élevées ou basses).

```
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")
```

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.
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.formul` 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`.

Description by low birth weight status (`low`)

		No	Yes
		$N = 130$	$N = 59$
age	years	19.0 **23.0** 28.0	19.5 **22.0** 25.0
lwt	pounds	113.0 **123.5** 147.0	104.0 **120.0** 130.0
race : White		56% (73)	39% (23)
Black		12% (15)	19% (11)
Other		32% (42)	42% (25)
smoke : Yes		34% (44)	51% (30)
ptl : 1+		9% (12)	31% (18)
ht : Yes		4% (5)	12% (7)
ui : Yes		11% (14)	24% (14)
ftv : Yes		51% (66)	39% (23)

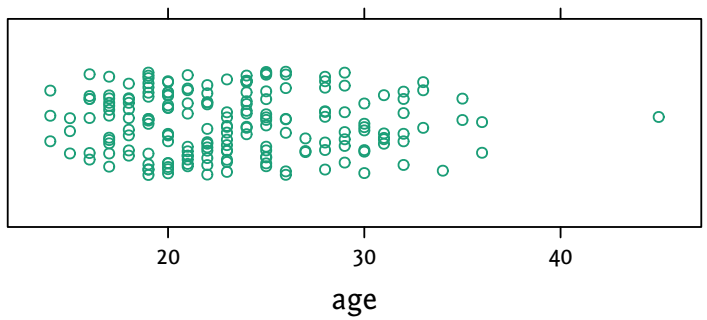
$a\ b\ c$ represent the three quartiles.

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)
```

Diagramme de dispersion (1D)



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)
```

Autres possibilités : Estimateurs robustes

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

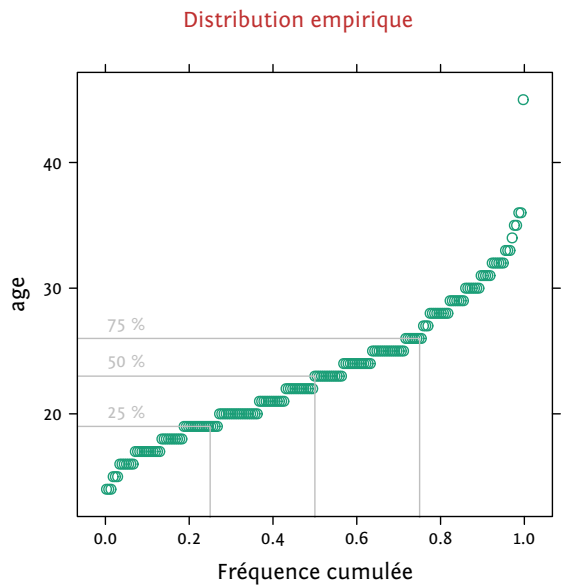
Fonction de répartition

On peut représenter la distribution des effectifs cumulés à l'aide d'un diagramme quantile-quantile.

```
qqmath(~ age, data=birthwt, dist=qunif)
```

L'idée de base consistant à "trier" les données permet en outre d'identifier facilement n'importe quel quantile. Un algorithme pour calculer la valeur médiane est indiqué ci-dessous.

```
med <- function(x) {
  odd.even <- length(x) %% 2
  if (odd.even == 0)
    (sort(x)[length(x)/2] +
      sort(x)[1+length(x)/2]) / 2
  else
    sort(x)[ceiling(length(x)/2)]
}
```



On pourrait également comparer cette distribution empirique à une distribution théorique, par exemple une distribution gaussienne.

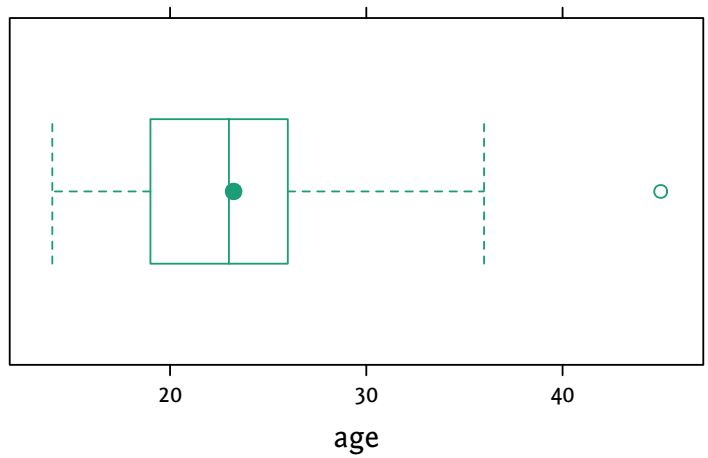
Approche graphique

On peut visualiser les cinq indicateurs renvoyés par la fonction `summary` à l'aide d'une "boite à moustaches" (Wickham & Stryjewski, sub.).

```
bwplot(~ age, data=birthwt)
```

Dans la figure ci-dessous, la moyenne est indiquée par un point, l'étendue des données est représentée par les "moustaches", et 50 % des observations sont comprises entre les 1er et 3ème quartiles figurés par la "boîte".

Boîte à moustaches

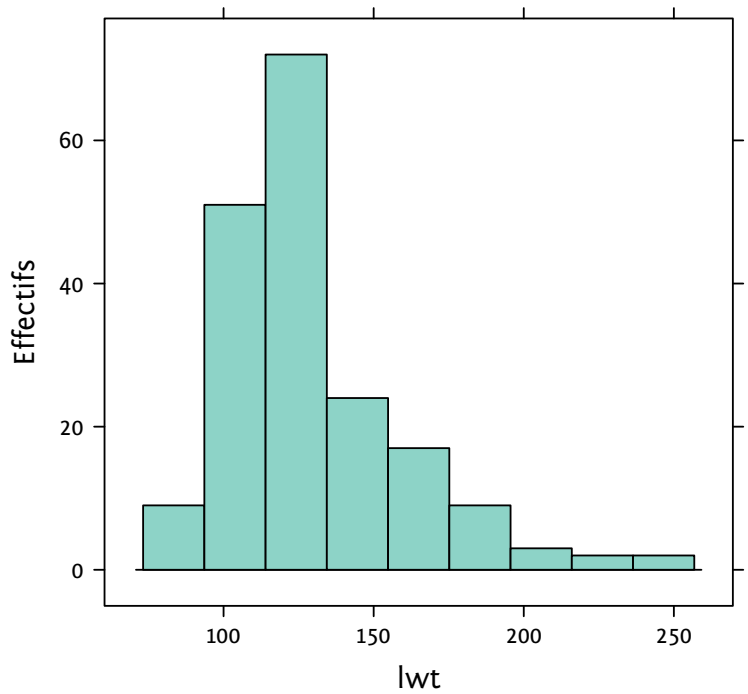


Histogramme

Un histogramme permet de représenter la répartition des valeurs d'une variable continue en classes. Ici, on voit que la distribution du poids des mères est asymétrique (asymétrie > 0, indépendant de l'unité de mesure).

```
histogram(~ lwt, data=birthwt, type="count")
library(e1071)
skewness(birthwt$lwt)
```

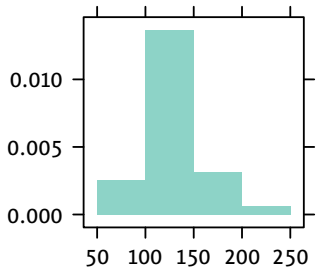
Histogramme (options par défaut)



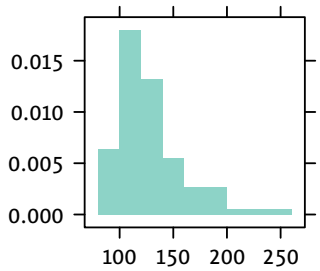
Histogramme (variations)

Le paramètre `breaks` permet de changer le nombre d'intervalles construits. Par défaut, la méthode utilisée est la méthode de Sturges (Sturges, 1926). L'ajout d'une loi de densité gaussienne dont les paramètres sont estimés* à partir de l'échantillon est également possible.

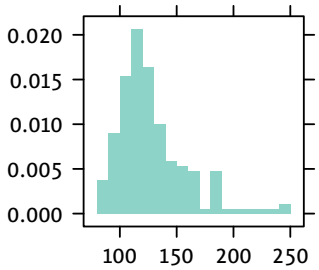
5 intervalles



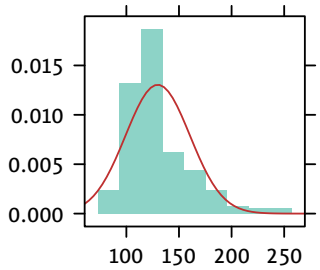
10 intervalles



15 intervalles



N(129.8;30.6) superposée

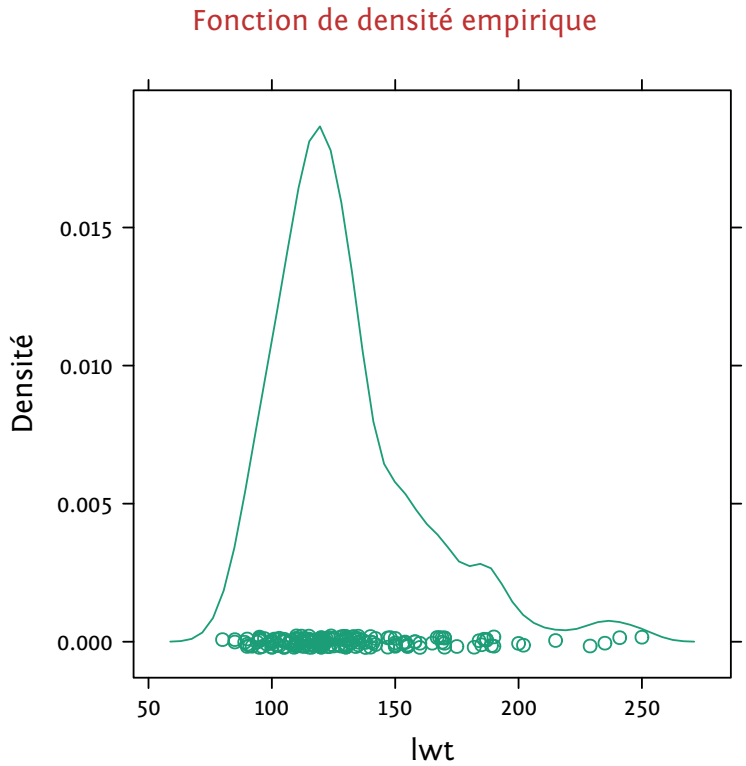


*Si l'on estime les paramètres d'une distribution empirique, les p-valeurs d'un test d'adéquation à une loi connue sont biaisées.

Densité empirique

Pour pallier à l'arbitraire du choix du nombre d'intervalles, on peut préférer représenter la fonction de densité empirique (Silverman, 1986; Venables & Ripley, 2002). Il reste toutefois à définir la largeur de la fenêtre de lissage associée à la fonction noyau de lissage (voir `help(bw.nrd0)`).

```
densityplot(~ lwt, data=birthwt)
```



Résumé numérique bivarié

Dans le cas de deux variables continues, on peut s'intéresser à la distribution jointe ou quantifier une certaine mesure de leur covariation, p.ex.

```
with(birthwt, cor(lwt, bwt))
with(birthwt, cor(lwt, bwt, method="spearman"))
```

Lorsque l'on croise une variable numérique avec une variable catégorielle, on peut produire des résumés numériques séparés pour chaque niveau de la variable de classification.

```
with(birthwt, by(age, low, summary))
with(birthwt, tapply(lwt, race,
                     function(x) c(mean(x), sd(x))))
```

Des fonctionnalités étendues sont disponibles dans `Hmisc`. La commande ci-dessus se résume à

```
summary(race ~ lwt, data=birthwt, method="reverse")
```

ou encore, si l'on souhaite afficher moyenne et écart-type,

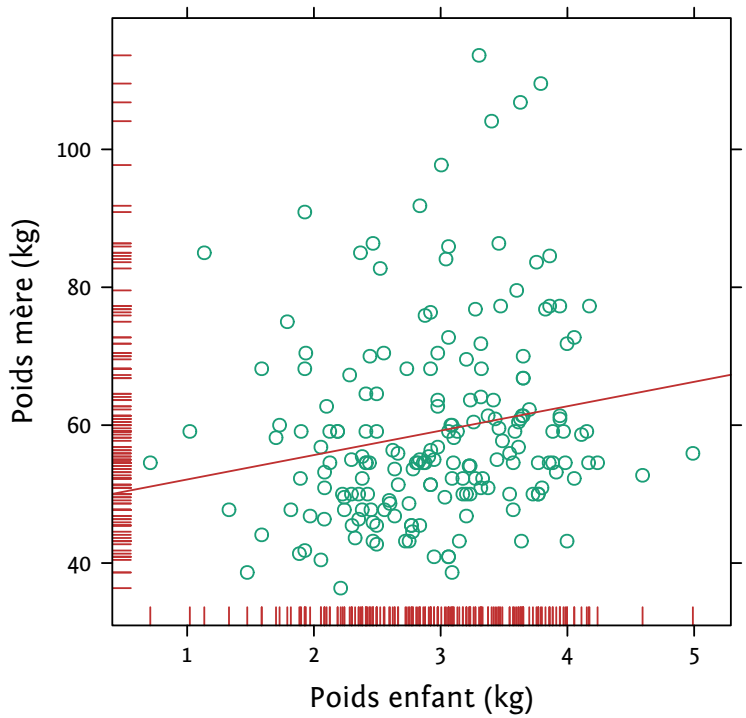
```
print(summary(race ~ lwt, birthwt, method="reverse"),
      prmsd=TRUE)
```

Diagramme de dispersion

Pour deux variables numériques, un diagramme de dispersion permet de résumer rapidement la distribution conjointe, ici dans les mêmes unités* :

```
xyplot(lwt/2.2 ~ bwt/1000, data=birthwt, type=c("p","r"))
```

La concentration des points autour de la droite de régression est un bon indicateur de la “consistance” de l’hypothèse de linéarité entre x et y.

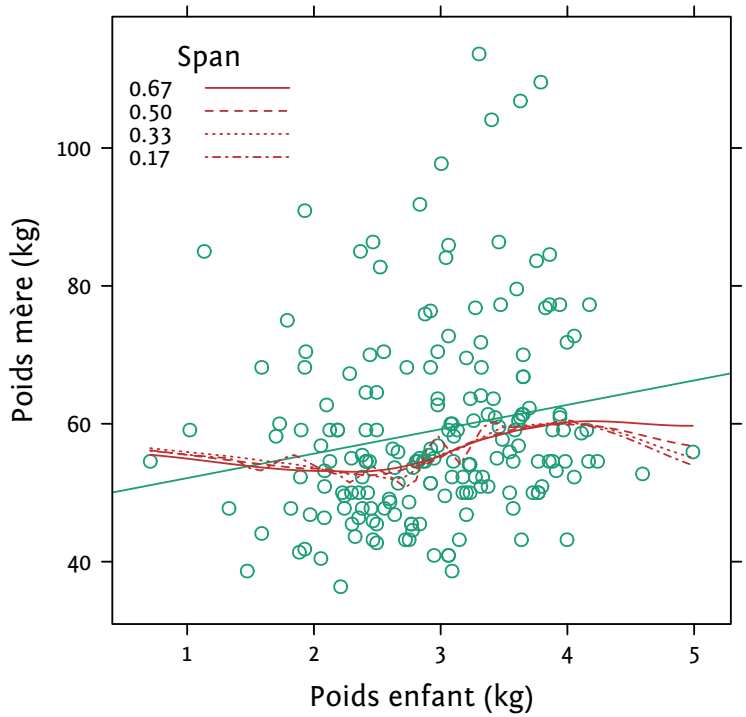


*Il existe d’autres moyens d’utiliser des transformations sur les unités de mesure en lien avec les graphiques `lattice`.

Scatterplot smoother

Supposer la linéarité n'est pas toujours la meilleure des options, et peut masquer certaines distorsions locales dans la relation entre les deux variables. Pour cette raison, on peut préférer utiliser une courbe plus flexible de type `loess` (Cleveland, 1979).

Pour cela, on remplacera avantageusement `type=c("p","r")` (points + droite de régression) par `type=c("p","smooth")` (points + lowess).

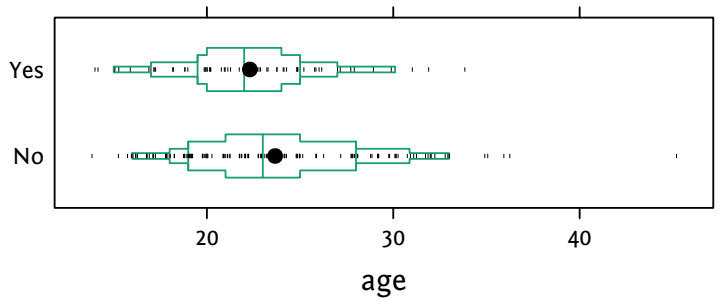


Distributions conditionnelles

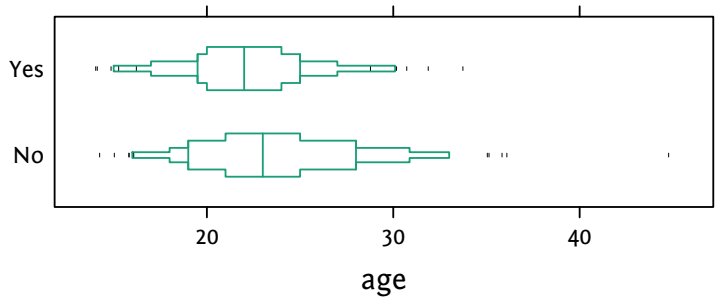
Les diagrammes de type boîte à moustaches et densité empirique peuvent être conditionnés sur un facteur, ici l'âge de la mère en fonction de l'indicateur de poids à la naissance :

```
bwplot(low ~ age, data=birthwt, panel=Hmisc::panel.bpplot)
```

Quantiles par défaut : c(.05,.125,.25,.375)



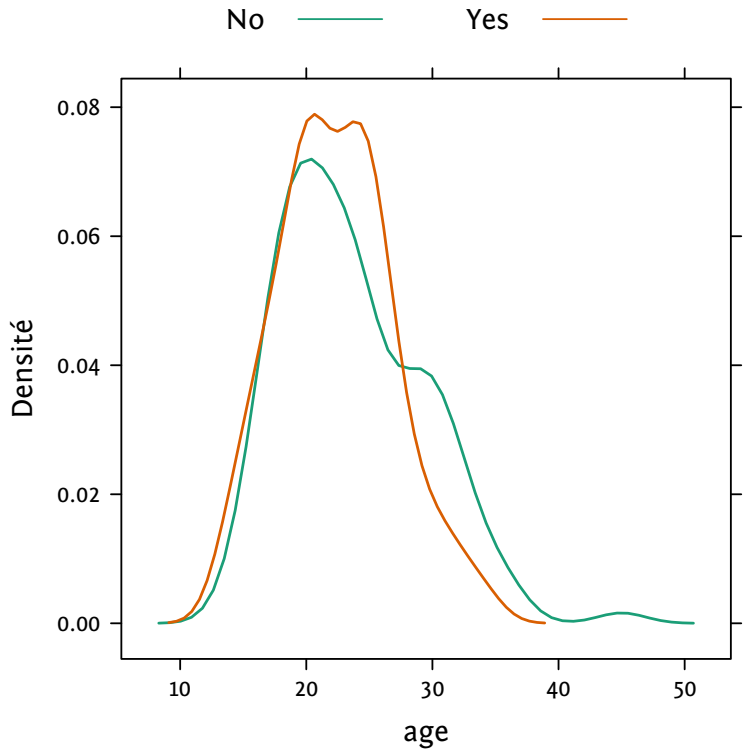
Cinq observations extrêmes affichées



Distributions conditionnelles (2)

L'avantage des densités non-paramétriques est que l'on peut visualiser directement les variations dans l'allure des distributions conditionnelles (tendance centrale et forme de la distribution).

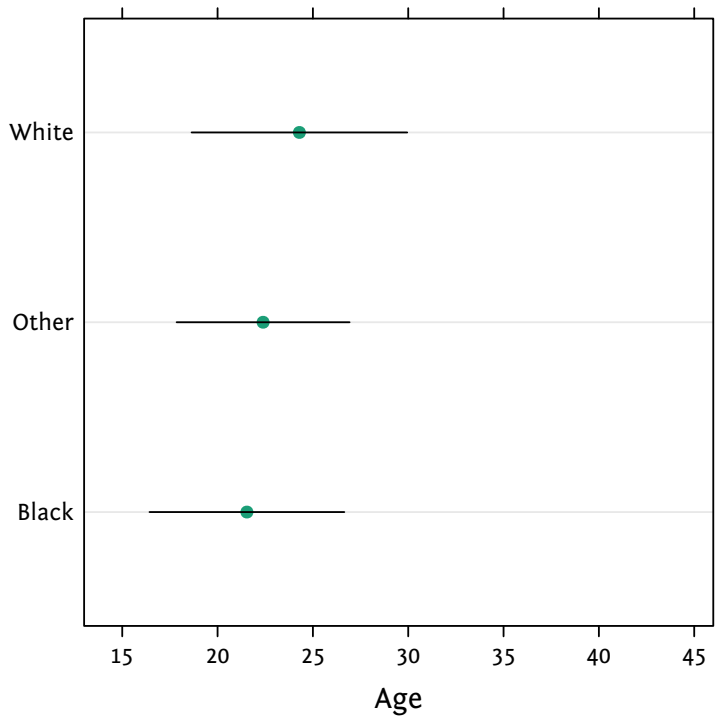
```
densityplot(~ age, data=birthwt, groups=low,
            auto.key=list(columns=2), plot.points=FALSE)
```



Mesures aggrégées

Plutôt que des diagrammes en barres (barchart), on préférera les diagrammes de type dotplot (Cleveland, 1985) pour représenter les données aggrégées, telles que des moyennes ou des proportions*.

```
age.by.race <- aggregate(age ~ race, data=birthwt, FUN=mean)
dotplot(race ~ age, data=age.by.race)
```



*En plus simple, grâce à Hmisc : `plot(summary(race ~ age, birthwt, method="reverse"))`.

Croiser plus de deux variables

Les graphiques en trellis (Becker, Cleveland, & Shyu, 1996; Cleveland, 1993) offrent une structure de graphique simple et efficace pour représenter des données multidimensionnelles. En particulier, ils introduisent la notion de facettes pour représenter des distributions conditionnelles.
La notation utilisée est la notation par formule :

```
y ~ x | a       # y en fonction de x condit. à a
y ~ x | a + b   # y en fonction de x condit. à a et b
```

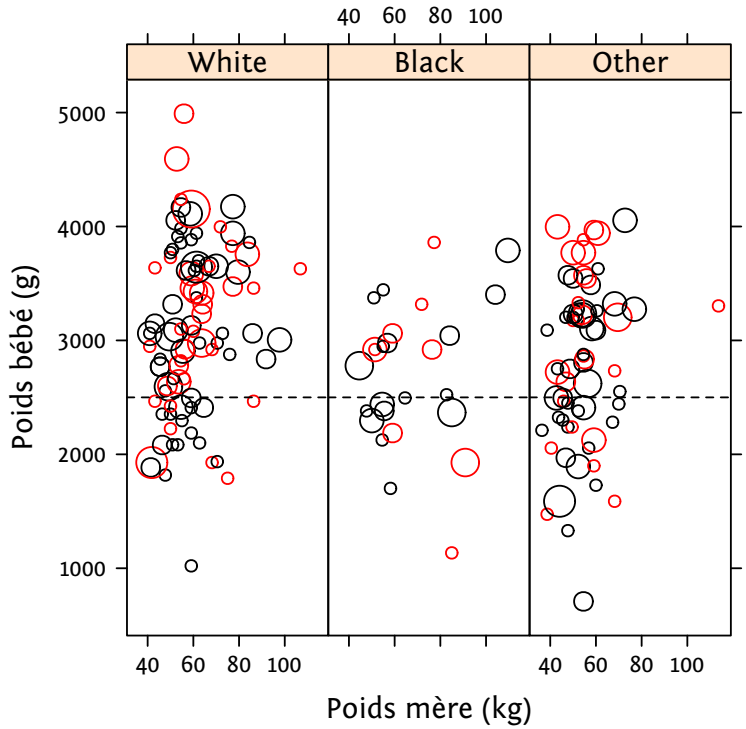
Les variables continues peuvent être “catégorisées” à l’aide de shingles. En transformant les variables numériques en facteurs, il devient possible de représenter un plus grand nombre de croisement de variables, tout en tenant compte de la nature “continue” des données.

Quelques exemples :

```
xyplot(bwt/1000 ~ lwt/2.2 | smoke, data=birthwt,
       groups=ptl>0)
bwplot(age ~ low | smoke + race, data=birthwt)
```

Illustrations

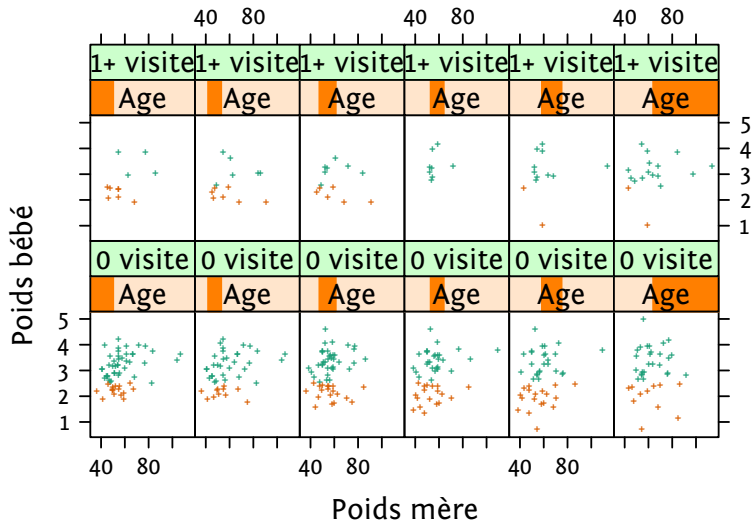
```
xyplot(bwt ~ lwt/2.2 | race, data=birthwt, layout=c(3,1),
       cex=sqrt(birthwt$ftv+.5), col=birthwt$smoke,
       panel=function(...) {
         panel.xyplot(...)
         panel.abline(h=2500, lty=2)})
```



*Ce n'est pas toujours une bonne idée d'utiliser `cex` ou `col` directement avec l'interface `lattice`.

Illustrations (2)

```
Age <- equal.count(birthwt$age)
ftvc <- factor(birthwt$ftv>1,
               labels=c("0 visite","1+ visite"))
xyplot(bwt/1000 ~ lwt/2.2 | Age + ftvc, data=birthwt,
       groups=low, pch="+")
```



*Il est possible d'utiliser des variables externes au `data.frame`.

Ce qu'il faut retenir

- Il est important de vérifier le codage des variables, et de recoder en fonction des besoins de l'analyse ou de la visualisation.
- On caractérise d'abord les distributions univariées (toutes !) avant de passer aux visualisations/modèles multivariés. Cela permet de détecter les éventuelles valeurs aberrantes, la mauvaise représentation de certaines modalités d'une variable catégorielle, ou l'existence d'asymétrie dans les distributions.
- L'interface `lattice` utilise la même notation par formule que les fonctions R pour la modélisation statistique.
- La visualisation des données catégorielles n'a pas vraiment été abordée, mais de nombreux outils sont disponibles dans les packages `vcd` et `vcdExtra` (Friendly, 2011; Meyer, Zeilis, & Hornik, 2006).

Index

Bibliographie

Becker, R. A., Cleveland, W. S., & Shyu, M. J. (1996). The Visual Design and Control of Trellis Display. *Journal of Computational and Statistical Graphics*, *5*, 123–155.

Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, *74*, 829–836.

Cleveland, W. S. (1985). *The Elements of Graphing Data*. Monterey, CA: Wadsworth.

Cleveland, W. S. (1993). *Visualizing Data*. Hobart Press.

Friendly, M. (2011). Tutorial: Working with categorical data with R and the vcd package. Retrieved from http://www.datavis.ca/courses/VCD

Hosmer, D., & Lemeshow, S. (1989). *Applied Logistic Regression*. New York: Wiley.

Meyer, D., Zeilis, A., & Hornik, K. (2006). The Strucplot Framework: Visualizing Multi-way Contingency Tables with vcd. *Journal of Statistical Software*, *17*.

Murrell, P. (2005). *R Graphics*. Chapman & Hall/CRC.

Sarkar, D. (2008). *Lattice, Multivariate Data Visualization with R*. Springer.

Silverman, B. W. (1986). *Density estimation*. Chapman and Hall.

Sturges, H. A. (1926). The choice of a class interval. *Journal of the American Statistical Association*, 65–66.

Tukey, J. W. (1977). *Exploratory Data Analysis*. Addison-Wesley.

Venables, W. N., & Ripley, B. D. (2002). *Modern Applied Statistics with S*. Springer.

Wickham, H., & Stryjewski, L. (sub.). 40 years of boxplots. *The American Statistican*.