Séminaire d'introduction à R

séance 3 - Statistique descriptive

Joël Gombin

CURAPP - UPJV

24 février 2012

Plan de la séance

Plan de la séance

Plan de la séance

Plan

Introduction

Les statistiques descriptives concernent les méthodes permettant de décrire, de diverses manières, les données qu'on étudie.
Dans une large mesure, la commande `summary()` fournit l'essentiel de la statistique descriptive...
L'information peut être résumée numériquement mais aussi graphiquement.

Plan

Analyser une variable

Analyser une seule variable peut sembler trivial, mais permet en fait souvent de découvrir beaucoup de choses sur nos données.
Les méthodes de résumé de l'information varient selon le type de variable.

Tendances centrales

On cherche ici à résumer une variable par une seule valeur, en s'intéressant plutôt à ce qui est commun qu'à ce qui diffère entre les observations. (Voir aussi `weighted.mean`)

```
mean(mini_picardie$abstentionspci95)
```

```
## [1] 15.47
```

```
mean(mini_picardie$abstentionspci95, trim = 0.1,
    na.rm = TRUE)
```

```
## [1] 15.43
```

```
median(mini_picardie$abstentionspci95, na.rm = TRUE)
```

```
## [1] 15.38
```

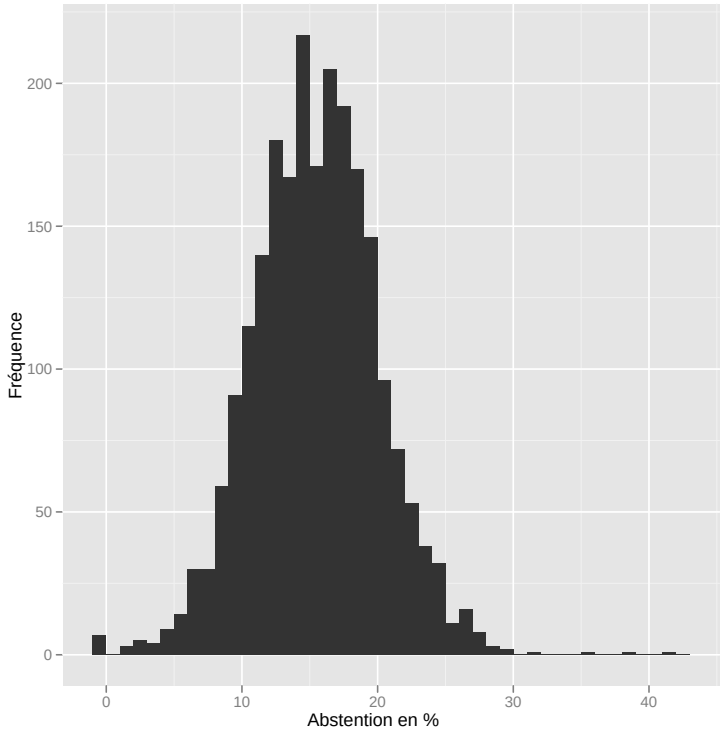
Dispersion

On s'intéresse plutôt ici aux variations au sein des observations.

```
range(mini_picardie$abstentionspci95, na.rm = TRUE)

## [1]  0.00 41.51


quantile(mini_picardie$abstentionspci95, probs = seq(0,
    1, 0.25), na.rm = TRUE)

##    0%   25%   50%   75%  100%
##  0.00 12.42 15.38 18.49 41.51


sd(mini_picardie$abstentionspci95, na.rm = TRUE)

## [1] 4.604
```

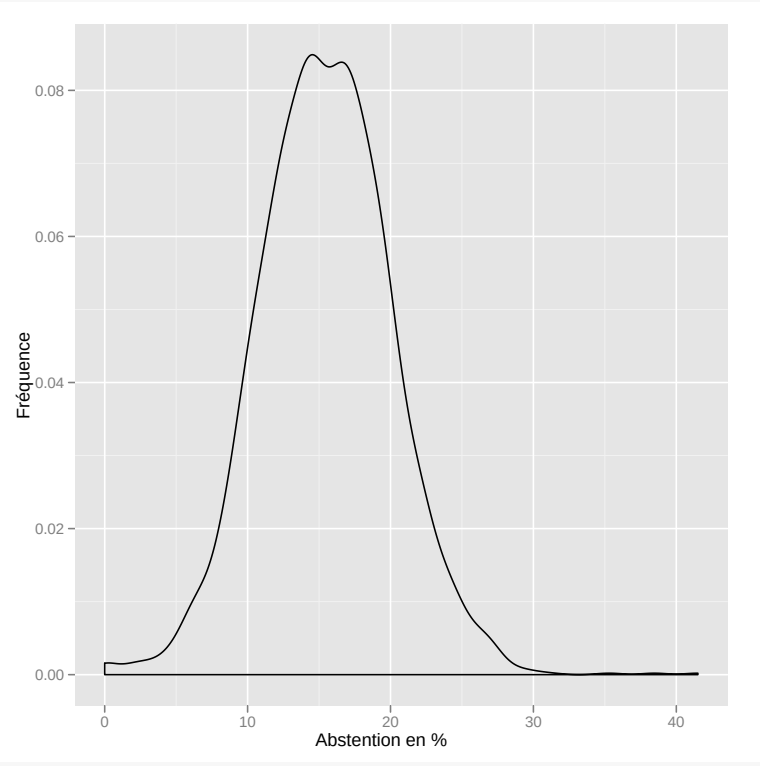
L'histogramme

```
require(ggplot2)
m <- ggplot(mini_picardie, aes(x=abstentionspci95))
m + geom_histogram(binwidth = 1) +
scale_x_continuous("Abstention en %") +
scale_y_continuous("Fréquence")
```



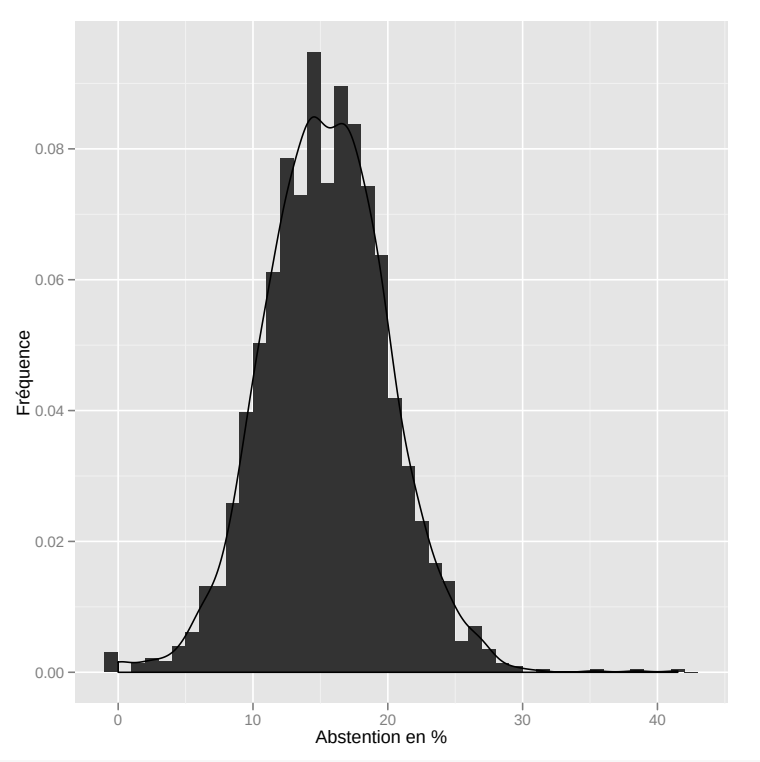
La courbe de densité

```
m <- ggplot(mini_picardie, aes(x=abstentionspci95))
m + geom_density() + scale_x_continuous("Abstention en %")
+ scale_y_continuous("Fréquence")
```



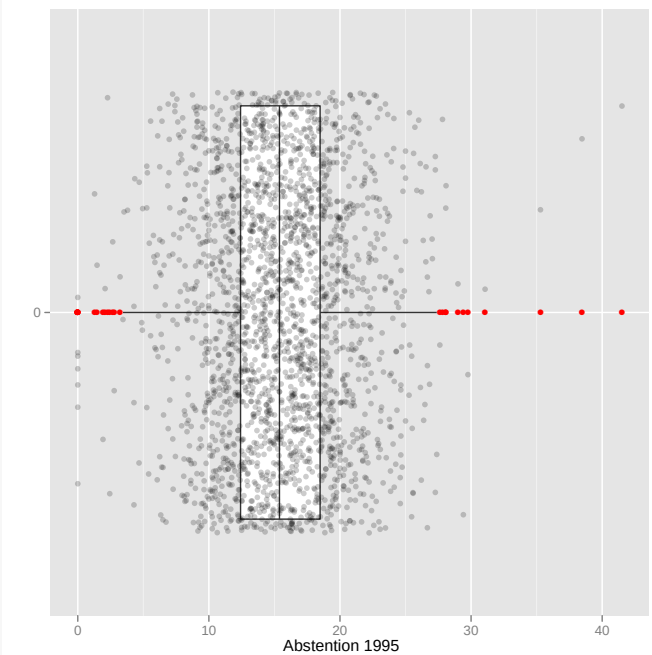
On peut cumuler les deux...

```
m + geom_histogram(aes(y=..density..), binwidth=1) +
geom_density() + scale_x_continuous("Abstention en %") +
scale_y_continuous("Fréquence")
```



Représenter les points et leur dispersion

```
qplot(x=factor(0), y=abstentionspci95, data=mini_picardie,
geom="boxplot", xlab="", ylab="Abstention 1995",
outlier.colour="red") + coord_flip() +
geom_jitter(alpha=0.2)
```

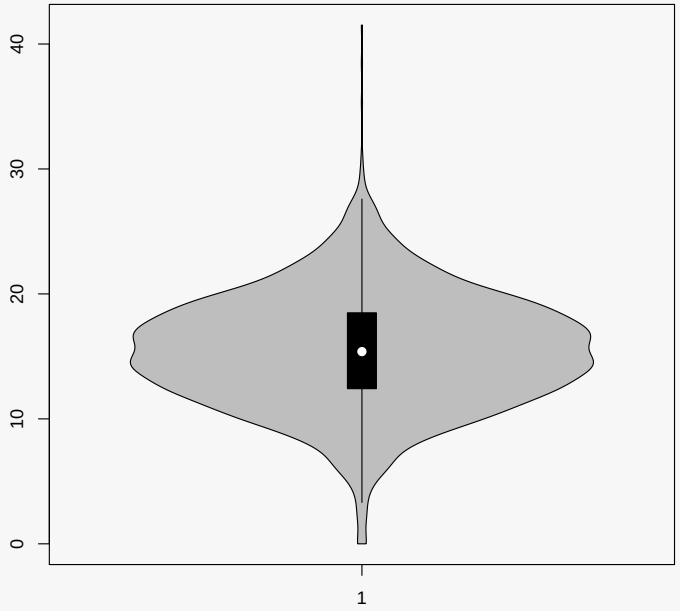


Les mesures représentées sont la médiane, les 1er et 3ème quartiles, et les lignes représentent 1,5 fois l'écart interquartile. La commande `boxplot` permet de contrôler ce paramètre.

Représenter les points et leur dispersion

Le graphe en violon combine un boxplot et un histogramme plus ou moins lissé.

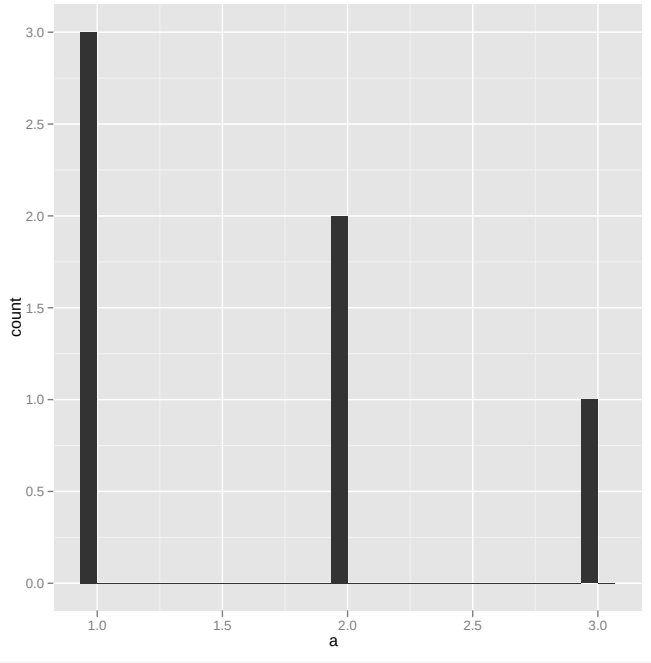
```
require(wvioplot)
wvioplot(mini_picardie$abstentionspci95, adjust=1)
```



Analyser une variable ordinale

Pour une variable ordinale (exemple : échelles d'attitude), on utilisera essentiellement la médiane et la distance interquartile. La représentation graphique passe par un `barplot` :

```
a <- c(1,2,1,1,2,3)
qplot(a, geom="bar")
```



Analyser une variable qualitative

Le résumé de l'information qu'on peut obtenir est assez pauvre : mode (classe la plus nombreuse) et tri à plat.

```
require(prettyR)
a <- as.factor(c("A", "B",
"A", "A", "B", "C"))
Mode(a)

## [1] "A"

table(a)

## a
## A B C
## 3 2 1
```

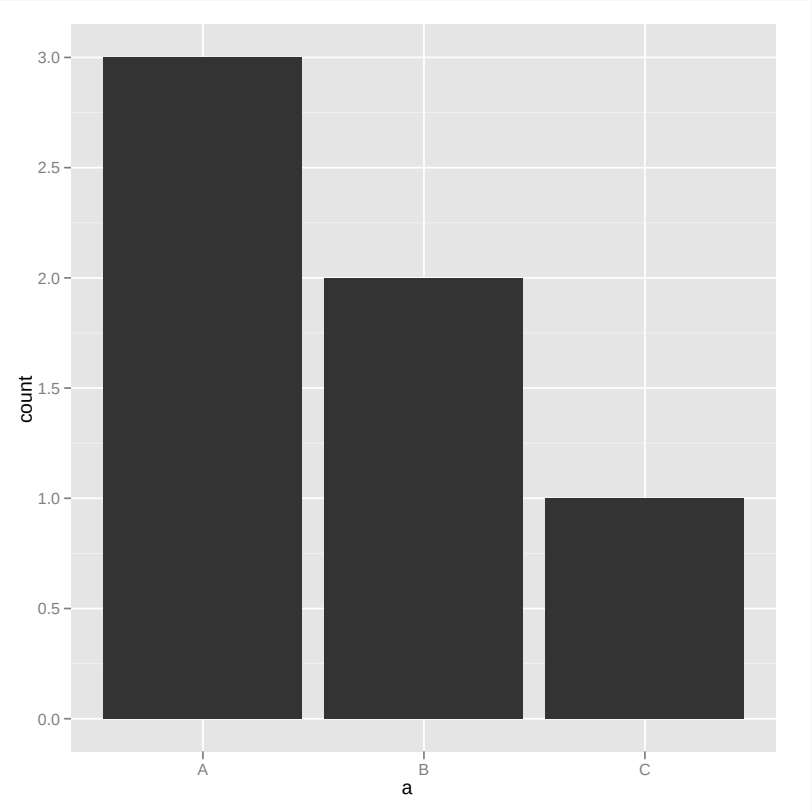
	A	B	C	All
n	3	2	1	6

```
require(rgrs)
cprop(table(a), percent =
T)

##          NA
## a         tab    Ensemble
##   A        50.0%  50.0%
##   B        33.3%  33.3%
##   C        16.7%  16.7%
##   Total   100.0% 100.0%
```

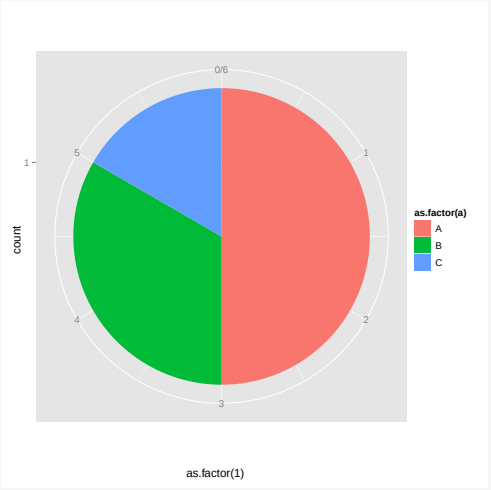
Analyser une variable qualitative

Le résumé graphique sera un barplot ou un camembert.

```
qplot(a, geom = "bar")
```




```
c <- as.data.frame(a)
ggplot(data = c, aes(x =
as.factor(1), fill =
as.factor(a))) +
    geom_bar(width = 1, xlab
= "", ylab = "") +
coord_polar(theta = "y")
```



Plan

Comparer deux variables quantitatives

On peut largement réutiliser les outils (numériques et graphiques) précédents. Pour savoir dans quelle mesure deux moyennes sont différentes de manière significative, on utilise le test de Student (*t-test*).

```
t.test(mini_picardie$abstentionspci95, mini_picardie$AbsIns)

##
##  Welch Two Sample t-test
##
## data:  mini_picardie$abstentionspci95 and mini_picardie$AbsIns
## t = 16.68, df = 4574, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal t
## 95 percent confidence interval:
##  1.974 2.499
## sample estimates:
## mean of x mean of y
##     15.47     13.23
##
```

Mesurer la relation entre deux variables

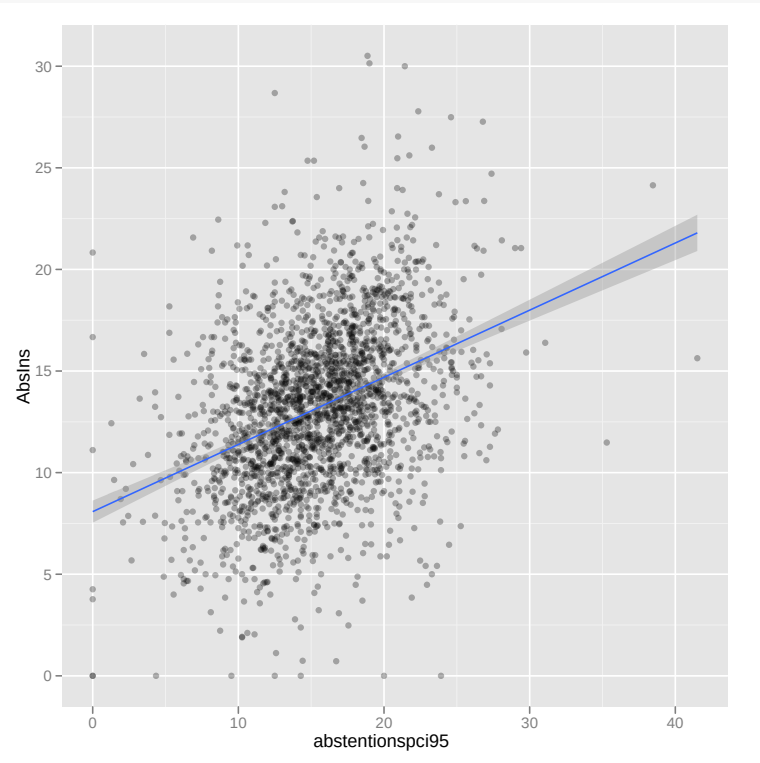
On utilise le coefficient de corrélation de Bravais-Pearson R.

```
cor.test(mini_picardie$abstentionspci95[-1699],
    mini_picardie$AbsIns[-1699])

##
##  Pearson's product-moment correlation
##
## data:  mini_picardie$abstentionspci95[-1699] and mini_picardie$
## t = 19.2, df = 2287, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.3367 0.4073
## sample estimates:
##    cor
## 0.3725
##
```

Visualiser la relation entre deux variables

On représente le nuage de points

```
qplot(abstentionspci95, AbsIns,
data=mini_picardie[-1699,], alpha=I(0.3)) +
geom_smooth(method="lm")
```



Analyser deux variables ordinales

La mesure numérique de la relation entre deux variables ordinales (deux échelles d'attitude, par exemple) se fait généralement au moyen du coefficient de corrélation de Spearman.

```
a <- c(1, 2, 1, 1, 1, 2, 3)
b <- c(10, 15, 9, 7, 10, 18, 25)
cor.test(a, b, method = "spearman")

## Warning message: Cannot compute exact p-values with ties

##
##  Spearman's rank correlation rho
##
## data:  a and b
## S = 5.346, p-value = 0.005134
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##    rho
## 0.9045
##
```

Analyser deux variables qualitatives

On utilise le tableau croisé et le χ^2.

```
mini_picardie$codedpt <- factor(mini_picardie$codedpt)
table(mini_picardie$codedpt, mini_picardie$type_urbain)

##
##      Commune monopolarisée Commune multipolarisée
##   2                    279                    193
##   60                   335                    196
##   80                   241                     21
##
##      Espace à dominante rurale Pole urbain
##   2                         305          39
##   60                        116          46
##   80                        499          20

chisq.test(table(mini_picardie$codedpt,
mini_picardie$type_urbain))

##
##  Pearson's Chi-squared test
##
## data:  table(mini_picardie$codedpt, mini_picardie$type_urbain)
## X-squared = 410.3, df = 6, p-value < 2.2e-16
##
```

Analyser numériquement une variable quantitative et une variable qualitative

On commence par faire un tri à plat :

```
tapply(mini_picardie$abstentionspci95,
mini_picardie$type_urbain,
    mean)

##     Commune monopolarisée    Commune multipolarisée
##                     15.38                     16.27
## Espace à dominante rurale               Pole urbain
##                     14.92                     17.72

tapply(mini_picardie$abstentionspci95,
mini_picardie$type_urbain,
    sd)

##     Commune monopolarisée    Commune multipolarisée
##                     4.229                     4.638
## Espace à dominante rurale               Pole urbain
##                     4.876                     3.816
```

Analyser numériquement une variable quantitative et une variable qualitative

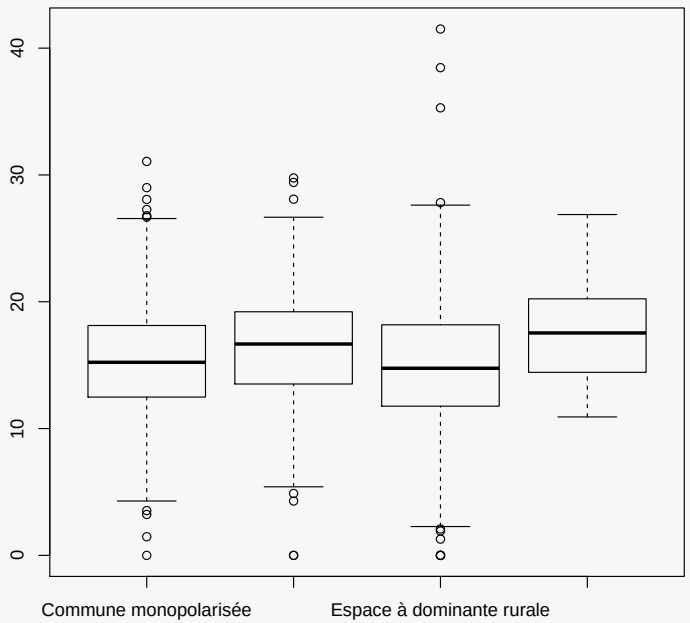
On peut ensuite faire une analyse de variance (ANOVA) :

```
anova(lm(abstentionspci95 ~ type_urbain, data =
mini_picardie))

## Analysis of Variance Table
##
## Response: abstentionspci95
##                Df Sum Sq Mean Sq F value  Pr(>F)
## type_urbain     3   1075     358    17.3 4.3e-11 ***
## Residuals    2286  47436      21
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Analyser graphiquement une variable quantitative et une variable qualitative

Le plus indiqué semble de comparer les boxplots :

```
boxplot(abstentionspci95 ~ type_urbain, data = mini_picardie)
```



Quelques ressources somplémentaires

Le plus utile ici est sans doute le travail de Julien Barnier, *R pour les sociologues (et assimilés)*, qui accompagne son package `rgrs`.
On trouvera par ailleurs en ligne de très nombreuses ressources concernant l'utilisation de R pour la statistique descriptive.