Les données

Les 3 modes d'importations possibles sur R

Importer les données depuis un url

```
X <- read.table('http://pbil.univ-lyon1.fr/R/donnees/essence.txt', h=T, dec=',')
```

```
X <- read.csv( "https://pages.isfa.fr/data/liais54c7cd.csv",  h=TRUE, dec = "," )
```

args(read.csv) **ou** help("read.csv") permettent d'obtenir les informations sur la fonction

```
dim(X)
```

Importer les données depuis un espace local

```
X <- read.table("essence.txt", header=TRUE, dec=',')
```

Importer les données depuis la library

```
library(ade4)
```

```
data(banque)
```

Variables et Description générale

Relation entre terminologie et noms des objets sur R

tableau	data frame
variable qualitative	factor
modalité	level
variable quantitative	numeric, integer

str(pb) permet vérifier la nature de la table et corriger si besoin.

Ex : A variable quantitative → A = as.numeric(A) B variable qualitative → Y = as.factor(B)

Variable quantitative

Une série statistique associée à une variable quantitative A est une liste de valeurs mesurées sur n individus. À chaque individu i est associée la valeur xi.

Paramètres descriptifs

Variance descriptive : vardes <- function(x) var(x)*(length(x)-1)/length(x)

vardes(A)

Variance estimée : var(A)

Ecart-type descriptif : ecartype <- function(x) sqrt(vardes(x))

ecartype(A)

Paramètres de position : summary(A)

Représentation

Histogramme : hist(A)

Le graphe de Cleveland : dotchart(A)

La boite à moustache : bp = boxplot(A) → déterminer les quartiles avec bp$stats

1	2	3	4	5
Valeur min	Q1	médiane	Q3	Valeur max

Variable Qualitative

Paramètres statistiques

Fréquences absolues : summary(as.factor(A))

Fréquences relatives : summary(A)/ length(A)

Représentation :

pie(summary(A))

barplot(summary(A))

Les tables

Création d'une table à partir de 2 variables quantitatives

X <- matrix(c(258, 210, 117, 38, 73, 82, 433, 181), byrow=TRUE, ncol=2)

rownames(X) <- c("piétons", "bicyclettes", "cycles<50", "cycles>50")

colnames(X) <- c("jour", "nuit")

X = as.data.frame(X)

attach(X)

	jour	nuit
piétons	258	210
bicyclettes	117	38
cycles<50	73	82
cycles>50	433	181

Création d'une table entre 2 variables qualitatives

```
Y <- matrix(c("c", "b", "a", "c", "b", "b", "b", "a", "a", "c", "x", "z", "y", "x", "x", "z", "x", "y", "y", "y"), byrow=F, ncol=2)

colnames(Y) <- c("Variable1","Variable2")

Y = as.data.frame(Y)

attach(Y)
```

	Variable1	Variable
1	c	x
2	b	z
3	a	y
4	c	x
5	b	x
6	b	z
7	b	x
8	a	y
9	a	y
10	c	y

Création d'une table de contingence entre 2 variables qualitatives

```
T <- table(Variable1, Variable2)
```

```
> table(Variable1, Variable2)
         Variable2
Variable1 x y z
        a 0 3 0
        b 2 0 2
        c 2 1 0
```

TD 1134 Visualiser les données

> Importer des données et lire les données

read.csv(file ou URL, header = { TRUE / FALSE } , sep = "α", dec = ".", ...)

- header : 1ère ligne = nom des colonnes
- sep : espace par défaut

read.table(file.txt ou URL, header = { TRUE / FALSE } , sep = "α", ...)

names (file) : renvoie noms des colonnes d'un dataframe

→ selection d'un sous-ensemble de données :

File [..,..] :

File [File $ Colonne 1 == "α",] : donne toutes les lignes dont la donnée de la colonne 1 est égale à α

File [File $ Colonne 1 == "α", 2] : donne tous les éléments de la colonne 2 qui correspondent à une ligne dont la donné de la colonne 1 est α

File [File $ Colonne 1 == "α" c(1,4)]

> Représentations

→ HISTOGRAMME : hist ( )

hist (vecteur, main = "...", col = "...", border = "...", xlab = "...", ylas = "...", xlim = , ylim = , breaks)

- xlim / ylim : range of values des axes
- breaks : largeur des barres ie découpage des valeurs

→ GRAPHE DE CLEVELAND : dotchart ( )

dotchart (vecteur, main = "...", pch = 20)

dotchart (sort(vecteur), main = "...", pch = 20)
↳ ordonné

par afficher plusieurs plot ensemble : par(...)

1) ① ② 2) ③ ④

3) ① ② 1) ④ 2) ④ ③
③ ④ ② ①

par(mfrow) : 1 2 / 3 4

par(mfrow = c(1,2)) 1
c(2,1) 2
c(2,2) 3

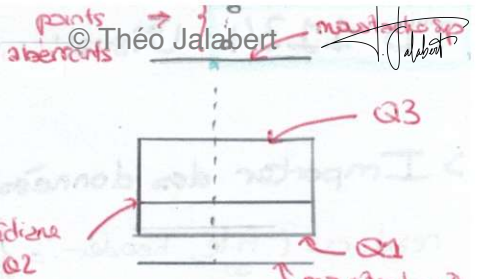
par(mfcol) : 1 3 / 2 4

par(mfcol) = c(2,1) 1
c(2,2) 2

→ BOÎTE À MOUSTACHE : boxplot()

boxplot(vecteur, col = "...", main = "...", horizontal = TRUE)

↳ par placer la boîte à moustache horizontalement
↳ respecte son sens de lecture de gauche à droite



→ portion centrale ⇒ répartition uniforme des valeurs
→ ⊕ haut / ⊕ bas ⇒ tassement des valeurs

$val\,max\,(m_{sup}) = Q_3 + 1,5 \times (Q_3 - Q_1)$
$val\,min\,(m_{inf}) = Q_1 - 1,5 \times (Q_3 - Q_1)$

→ REPRÉSENTATION EN CAMEMBERT pie()

pie(summary(vecteur), main = "...", col = "..." / c = ("...", "...", "..."))

→ REPRESENTATION EN BÂTONS barplot()

barplot({ summary(vecteur) / summary(vecteur)/length(vecteur) / vecteur }, main = "...", col = "...", las = { 1 / 2 })

↳ fréquences absolues ou relatives
↳ label horizontal
↳ label vertical

> RELATION ENTRE DEUX VARIABLES

Croisement	Paramètre	Graphique
Quantitatif × Quantitatif	Covariance coefficient de corrélation coeff de détermination	nuage de points
Qualitatif × Qualitatif	Chi-Deux Coefficient de Cramer	mosaïque représentation en ballons
Quantitatif × Qualitatif	Rapport de corrélation	repréz inter et intra-grapes boîtes à moustaches

$cov(X,Y) = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$

$r = \frac{cov(X,Y)}{S_x S_y}$ et r^2

cor(v1, v2)

→ NUAGE DE POINTS plot()

plot(vecteur_abs, vecteur_ord, pch = 20, main = "...", xlab, ylab, type = "n")

↳ donne le cadre du graphe mais vide (graphe à ajouter à la ligne suivante)

text(vect_abs, vect_ord, vecteur nom)

attach(dataframe) : permet de travailler directement (en variables) du dataframe en

→ Valeur du Chi-Deux de contingence

On construit une table de contingence (table théorique) par définir le lien entre deux variables : hypothèse d'indépendance entre les deux variables ie equi proba : $\frac{n_i n_j}{n}$

Chi-Deux de contingence : compare $EO = n_{ij}$ (effectifs observés) avec $ET = \frac{n_i n_j}{n}$ (effectifs théoriques)

On a $\chi^2 = \sum \frac{(EO - ET)^2}{ET}$ ⇒

- si $\chi^2 = 0$: indépendance des deux variables
- si χ^2 est petit : EO et ET presque identiques
  ↳ variables peu liées entre elles
- si χ^2 est grand : EO et ET différents
  ↳ variables liées entre elles

CodeR : `chisq.test(v1, v2)`

→ REPRÉSENTATION EN BALLONS balloonplot()

```
library(gplots)
balloonplot(table(v1, v2), main = "...")
```

→ REPRÉSENTATION EN MOSAÏQUE mosaicplot()

se base sur les écarts entre EO et ET : $\frac{(EO - ET)}{\sqrt{ET}}$

```
mosaicplot(table(v1, v2), main = "...", shade = TRUE)
```

↳ permet de mettre en relief les différences

- bleu : les EO sont ⊕ grands que ET sous l'absence de lien entre les deux variables
- rouge : les EO sont ⊕ petits que les ET sous l'absence de lien entre les deux variables

→ Quantitatif x Qualitatif

> Variation :

- variance descriptive mesurée sur un groupe de n individus

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n}$$

- variance estimée de la population à partir d'un échantillon de n individus

$$\hat{\sigma}^2 = \frac{n}{n-1} s^2 \quad \text{ou encore} \quad \hat{\sigma}^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$$

On préfère travailler sur la **variation totale** i.e sur la somme des carrés des écarts à la moyenne (SCE)

$$SCE_T = \sum_{i=1}^{n} (x_i - \bar{x})^2$$

```
SCE ← function(x) sum((x-mean(x))^2)
```

Variation inter-groupe : carré des écarts entre la moyenne du groupe et la moyenne globale

$$SCE_B = \sum_{k=1}^{n} n_k (\bar{x}_k - \bar{x})^2$$

Exemple : crimes dans les Etats Américains

- n_k : nbr d'états ayant été échantillonnés l'année k
- $\bar{x}_k$: nbr moyen de crimes pour l'année k

```
SCEB ← function(x, gpe){
    moyenne ← tapply(x, gpe, mean)
    effectifs ← tapply(x, gpe, length)
    res ← (sum(effectifs * (moyenne - mean(x))^2))
    return(res)
    }
```

Rapport de corrélation : $\eta^2 = \frac{SCE_B}{SCE_T}$ =>
- proche de 0 : variables pas liées
- proche de 1 : variables liées

```
eta2 ← function(x, gpe) { res ← SCEB(x, gpe) / SCE T(x) ; return(res)}
```

(à finir)

(Voir le cas des graphes avec les modalités d'une variables qualitatives : boîte à moustache, graphe de Cleveland)