Notes de cours
Statistique avec le logiciel R

Shuyan LIU
Shuyan.Liu@univ-paris1.fr
http ://samm.univ-paris1.fr/Shuyan-LIU-Enseignement

Année 2013-2014

Chapitre 1

Introduction

L'objectif de ce cours est de mettre en évidence les liens entre toutes les notions en statistique rencontrées dans les années antérieures par les étudiants : statistiques descriptives, variables aléatoires et statistique mathématique. Le logiciel R est utilisé pour illustrer les applications des outils de statistique.

1.1 Plan de cours

- Statistiques descriptives
- Graphiques sous R : personnalisation des graphes
- Modèle linéaire : régression simple et multiple, ANOVA sous R
- Estimations ponctuelles et par intervalles de confiance
- Tests paramétriques
- Tests non paramétriques
- Outils R pour les valeurs extrêmes

1.2 Vocabulaire

Une statistique	La statistique / Les statistiques
- un nombre calculé à partir des observations - le résultat de l'application de méthode	- un domaine - la collecte des données - le traitement (descriptive) - l'interprétation (exploratoire) - la prévision et la décision (décisionnelle)

1.3 Statistique et probabilité

Les probabilités sont l'outil de base du statisticien, car le "hasard" intervient à plusieurs niveaux en statistique : la répartition des données, le bruit, etc..

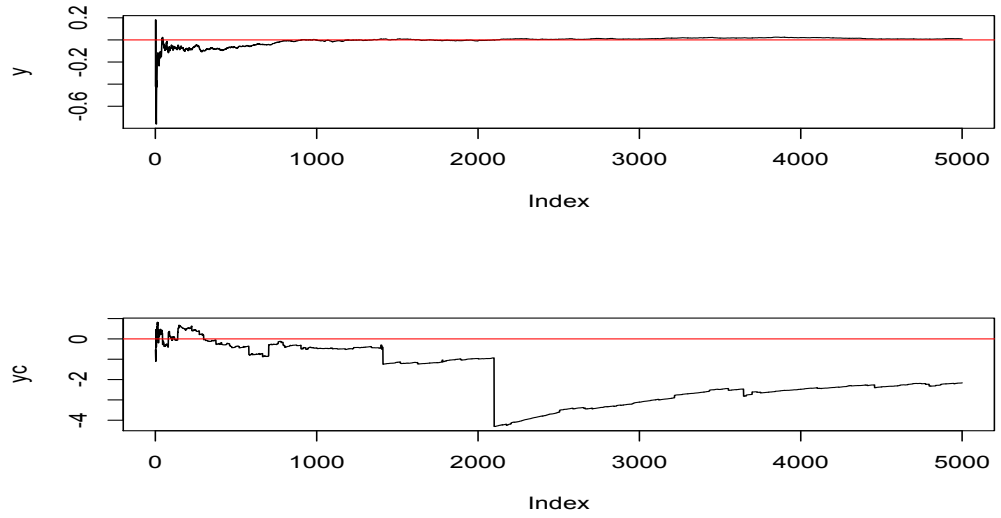
1.4 Les avantages de R

– S-PLUS
– méthodes récentes
– multi-plateforme
– gratuit
Installation : http ://www.r-project.org

1.5 Quelques exemples

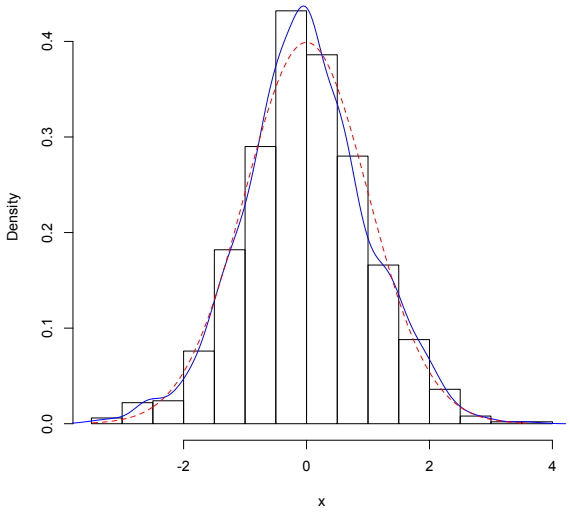
1. Illustration de la convergence p.s.

$$\frac{1}{n}\sum_{i=1}^{n} X_i \to \mathbb{E}\,(X_1)$$



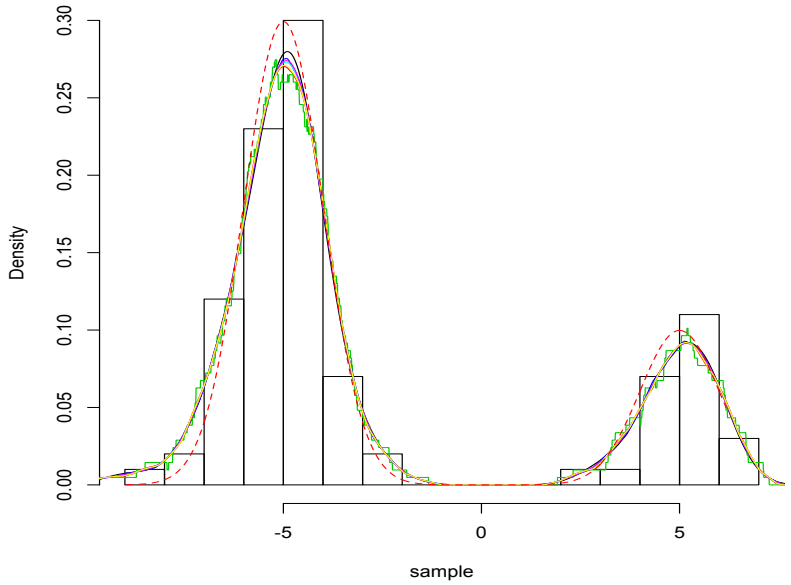
2. Estimation de la densité par l'estimateur à noyau

$$\hat{f}(x) = \frac{1}{nh}\sum_{i=1}^{n} K(\frac{x - x_i}{h})$$



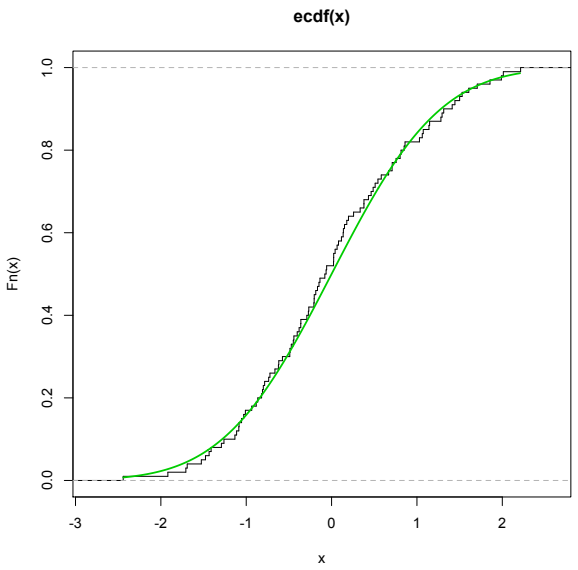
3. Mélange de lois

$$f(x) = \frac{1}{4}\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}(x-5)^2} + \frac{3}{4}\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}(x+5)^2}$$



4. Fonction de répartition empirique

$$\hat{F}(t) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{(\infty,t)}(x_i) \to F(t)$$



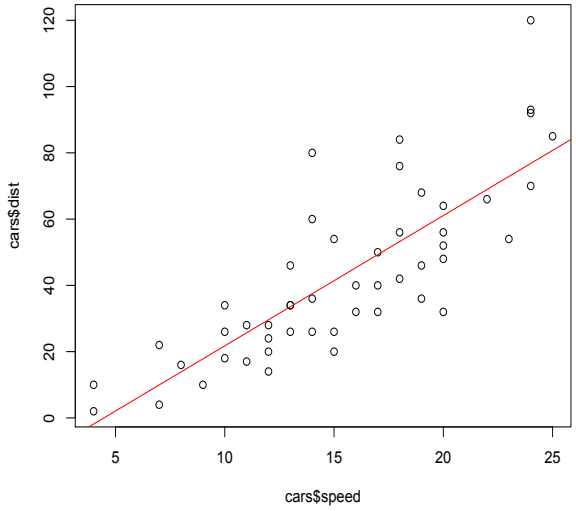
5. Régression linéaire

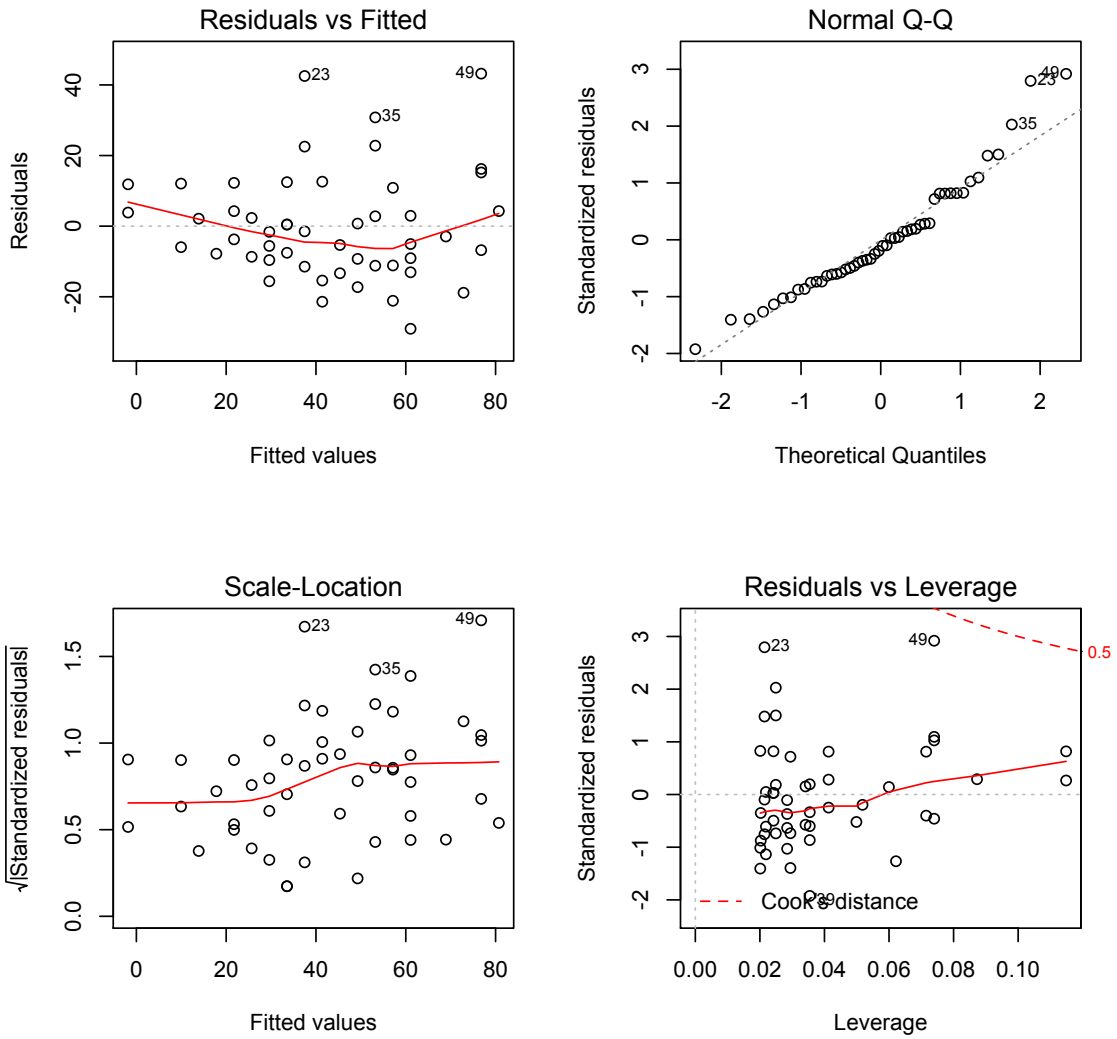
$$y_i = a + bx_i + \varepsilon_i$$

On considère un jeu de donnés `cars` fourni par R.

```
  speed dist
1     4    2
2     4   10
```

```
3       7    4
4       7   22
5       8   16
6       9   10
7      10   18
8      10   26
9      10   34
10     11   17
11     11   28
12     12   14
...

Call:
lm(formula = dist ~ speed, data = cars)

Residuals:
    Min      1Q  Median      3Q     Max
-29.069  -9.525  -2.272   9.215  43.201

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -17.5791     6.7584  -2.601   0.0123 *
speed         3.9324     0.4155   9.464 1.49e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 15.38 on 48 degrees of freedom
Multiple R-squared:  0.6511,Adjusted R-squared:  0.6438
F-statistic: 89.57 on 1 and 48 DF,  p-value: 1.49e-12
```




Residuals vs Fitted
Residuals
Fitted values
23
49
35
Normal Q-Q
Standardized residuals
Theoretical Quantiles
Scale-Location
√|Standardized residuals|
Residuals vs Leverage
Standardized residuals
Leverage
Cook's distance
0.5

Chapitre 2

Statistique descriptive et première session de R

Nous présentons dans ce chapitre les objets et les commandes élémentaires du logiciel R. Dans la suite, sont écrites à gauche des commandes à taper, et à droite des commentaires sur ces commandes ou des questions relatives.

Le symbole > apparaît automatiquement en début de chaque ligne de commandes.

Le symbole + apparaît en début de ligne si la précédente est incomplète.

Le symbole # permet d'insérer un commentaire.

Une petite astuce très utile lorsque vous tapez des commandes directement dans la console : en utilisant les flèches Haut et Bas du clavier, vous pouvez naviguer dans l'historique des commandes tapées précédemment, que vous pouvez alors facilement réexécuter ou modifier.

2.1 Les commandes élémentaires à connaître

`help()`	
`help.start()`	L'aide au format "html" (web)
`q()`	Quitter R
`example(plot)`	
`demo()`	

2.2 Premiers pas

`x=c(1,4,9)`	La fonction $c()$ concatène des scalaires ou des vecteurs.
`y=c(x,2,3)`	
`c(1:4)`	
`seq(10,100,10)`	Le premier terme est 10, le dernier est ≤ 100 et le pas est 10.
`x=rep(0,10)`	On crée un vecteur constitué de 0 répété 10 fois.
`rep(y,10)`	
`x=rnorm(20)`	On simule 20 v.a. i.i.d. suivant la loi normale standard.
`y=rexp(20)`	On simule 20 v.a. i.i.d. suivant la loi exponentielle d'espérance 1.
`median(x)`	
`mean(x)`	
`var(x)`	
`sd(x)`	
`summary(x)`	
`sum(x)`	
`length(x)`	
`plot(x)`	Pour tracer la première "courbe"
`lines(x)`	Pour ajouter une ligne
`points(y)`	Pour ajouter un nuage de points
`hist(x)`	
`boxplot(x)`	
`barplot(x)`	

Exercice

Pour une loi normale d'espérance 5 et de variance 2 :

1. Faire une représentation graphique de sa fonction de répartition et de sa densité sur $[0, 10]$.

2. Calculer la probabilité des évènements : $X \leq 0$, $X \leq 5$, $-1 < X \leq 3$ et $X > 10$.

3. Calculer entre quelles valeurs 95% des tirages de X sont compris.

2.3 Langage de programmation

1. Les différents objets : variable, vecteur, matrice, liste, table, fonction

a. list : Les listes sont très pratiques pour retourner des objets complexes notamment en sortie de fonctions.

```
data=c(1,2,3,4)
name="ABC"
maliste=list(donnees=data,nom=name)
```

`list` permet de combiner les objets de différents types et différents tailles dans une même structure.

b. data.frame : Les `data.frame` sont très utiles pour stocker des données.

```
before=c(1,2,3,4)
after=c(4,5,6,7)
type=c("A","B","C","D")
x=data.frame(before,after,type)
```
`data.frame` permet de combiner des objets "homogènes" dans une même structure.

```
typeof(x$type)
x=data.frame(before,after,I(type))
```
`I` permet de garder le type initial.

c. les données avec labels : `factor`

```
factor(c(1,3,2,2,1),levels=1:3)
```

d. matrice

```
matrix(1:12,nrow=3,ncol=4)
matrix(1:12,nrow=3,byrow=TRUE)
```
Par défaut la matrice est remplie par colonne.

```
cbind(1:3,4:6,7:9)
rbind(1:3,4:6,7:9)
```

e. fonction

```
masomme=function(x,y){s=x+y
s}
```

2. Importer et enregistrer les données : fichiers plats, `read.table`, `write.table`, `save`

```
y=data.frame(a=I("a"),b=pi)
write.table(y,file="y.csv",sep=",",col.names=T)
y2=read.table("y.csv")
y2=read.table("y.csv",sep=',',header=T)
write.table(x,file="x.txt",sep=" ",col.names=T)
write.table(x,file="x.txt",sep=" ",col.names=T,row.names=F)
save(x,y,file="xy.RData")
load("xy.RData")
```

Exercice

1. Créez deux vecteurs de dimensions quelconques et insérez le second vecteur entre les 2ème et 3ème éléments du premier vecteur.

2. Saisissez deux matrices 3×3 A et B et calculez la somme et le produit de ces matrices et enfin inversez la matrice A si possible.

3. Soient `x` un vecteur avec n "levels" et `y` un vecteur numérique de dimension n, que se passe-t-il si l'on tape `y[x]` ?

2.4 Statistique descriptive

1. variable : quantitative, qualitative, modalité, effectif, tableaux de contingence
Exemple : `chickwts` –masses de poulets en fonction de leur alimentation

Nourriture	Traduction
casein	caséine
horsebean	fève
linseed	graine de lin
meatmeal	farine animale
soybean	soja
sunflower	tournesol

```
chickwts                                  Données fournies par R
unique(chickwts$feed)                     Modalité de la variable "feed"
w=cut(chickwts$weight,3)                  Découper et transformer un vecteur en facteur.
table(w,chickwts$feed)                    Table de contingence
summary(chickwts)
plot(weight~feed,data=chickwts)
```

2. paramètre : position (min, max, moyenne, médiane, quantile), dispersion (étendue, variance, écart-type)

3. lois de probabilités

nom de loi	nom en R	paramètres		
beta	beta	shape1	shape2	ncp
binomial	binom	size	prob	
Cauchy	cauchy	location	scale	
chi-squared	chisq	df	ncp	
exponential	exp	rate		
F	f	df1	df2	ncp
gamma	gamma	shape	scale	
geometric	geom	prob		
hypergeometric	hyper	m	n	k
log-normal	lnorm	meanlog	sdlog	
logistic	logis	location	scale	
negative binomial	nbinom	size	prob	
normal	norm	mean	sd	
Poisson	pois	lambda		
Student	t	df	ncp	
uniform	unif	min	max	
Weibull	weibull	shape	scale	
Wicoxon	wilcox	m	n	

Exercice

1. Tirez au hasard 100 nombres x_i dans l'intervalle $[0, 1]$ avec une probabilité uniforme

puis donnez le nombre de x_i supérieurs à 0.5.

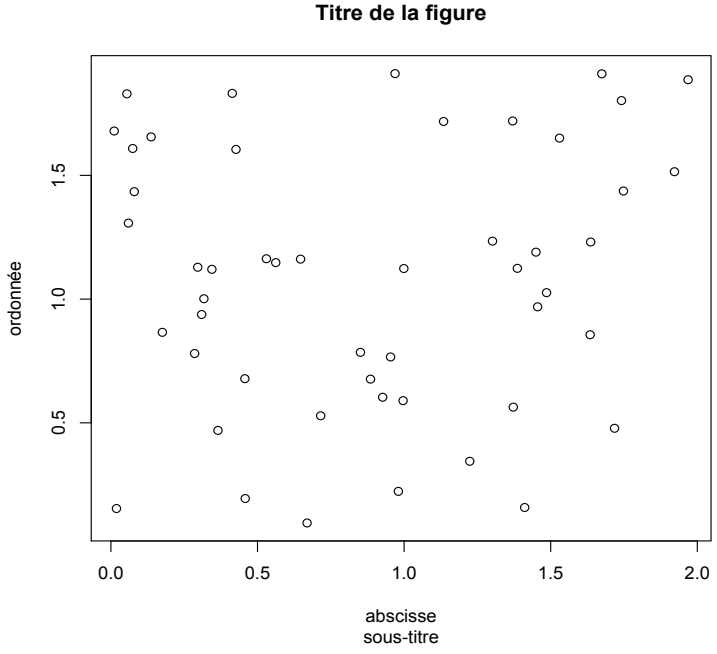
2. Tirez 100 couples de points (x, y) aléatoirement dans le carré $[0, 1] \times [0, 1]$. Calculez le centre de gravité du nuage de points, représentez le nuage de points obtenus dans une fenêtre graphique, puis y ajouter en rouge le centre de gravité.

3. Un bulbe fleurit avec une probabilité égale à 0.7. On plante 100 bulbes dans un pré carré de 11 mètres de côté en respectant un espace d'environ 1 mètre entre chaque bulbe. Simulez la floraison de ces 100 bulbes. Combien de bulbes ont fleuri ? Sauvegardez les données dans un unique fichier texte.

2.5 Représentations graphiques

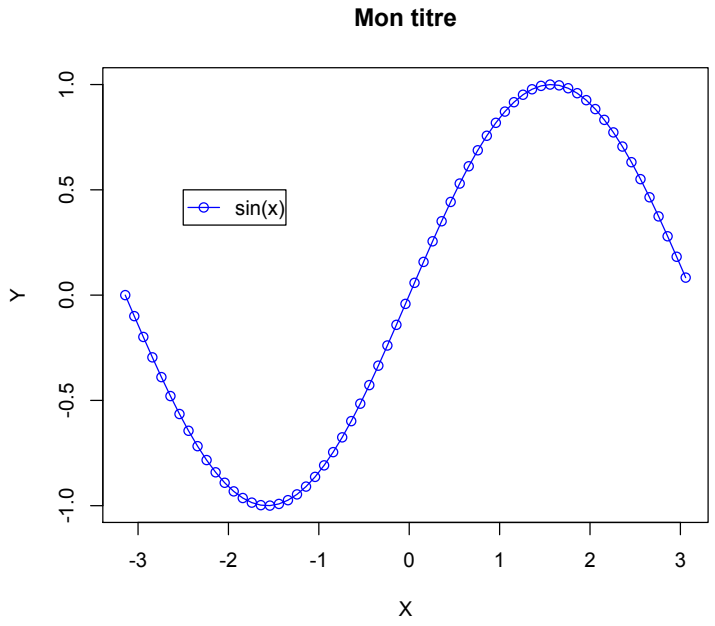
1. Un graphique standard

```
plot(x,y,main="Titre de la figure",sub="sous-titre",
xlab="abscisse",ylab="ordonnée")
```

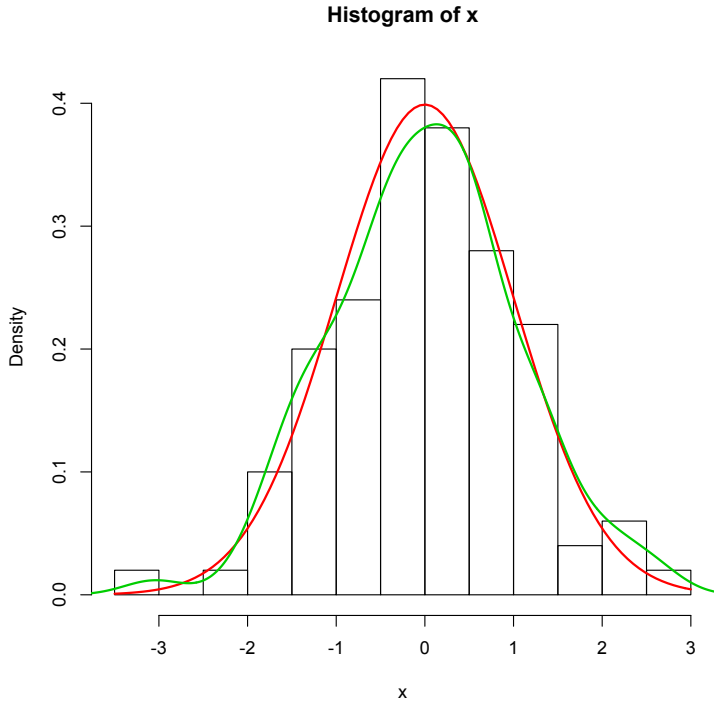


2. Il est possible d'ajouter une légende, du texte, une ligne ...

```
legend(-2.5,0.5,"sin(x)",col=4,pch=1,lty=1)
```

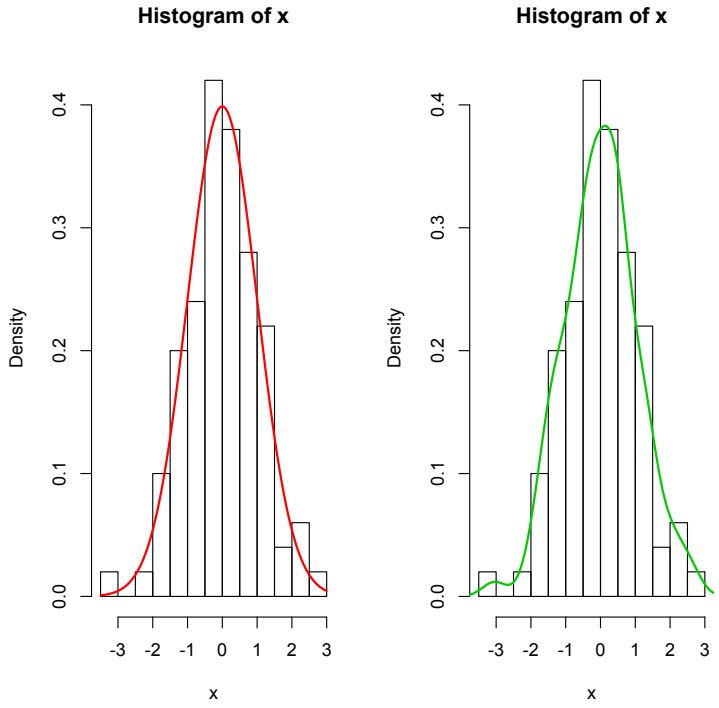


3. Superposer des courbes



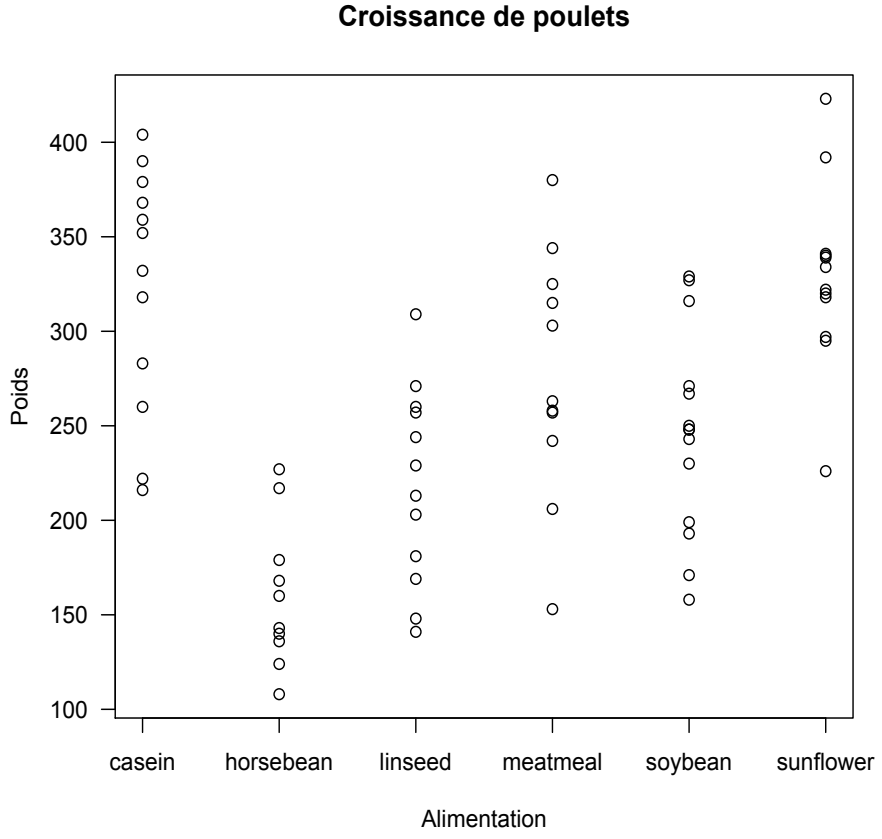
4. Mettre côte à côte plusieurs graphiques

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



Exercice

En utilisant les données disponibles `chickwts` dans R, tracer la figure suivante.



Chapitre 3

Modèle linéaire

3.1 Régression linéaire simple

4 postulats du modèle
pour tous $i = 1, \dots, n$,
– $\mathbb{E}(\varepsilon_i) = 0$
– $\mathrm{Var}(\varepsilon_i) = \sigma^2$
– ε_i sont i.i.d. de loi gaussienne.
– X_i sont déterministes.

Les estimateurs

$$\hat{\beta} = \frac{\sum_{i=1}^n (Y_i - \bar{Y})(X_i - \bar{X})}{\sum_{i=1}^n (X_i - \bar{X})^2}, \quad \hat{\mu} = \bar{Y} - \hat{\beta}\bar{X},$$

où $\bar{X} := \frac{1}{n}\sum_{i=1}^n X_i$ et $\bar{Y} := \frac{1}{n}\sum_{i=1}^n Y_i$.

Quelques notations

$$\hat{Y}_i = \hat{\mu} + \hat{\beta} X_i, \qquad \hat{\varepsilon}_i = Y_i - \hat{Y}_i,$$

$$\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^n (Y_i - \hat{Y}_i)^2 = \frac{\mathrm{SCR}}{n-2}, \quad \hat{\sigma}^2_{\hat{\beta}} = \frac{\hat{\sigma}^2}{\sum_{i=1}^n (X_i - \bar{X})^2} = \frac{\mathrm{SCR}}{(n-2)\sum_{i=1}^n (X_i - \bar{X})^2},$$

$$\hat{t} = \frac{\hat{\beta}}{\hat{\sigma}_{\hat{\beta}}}, \qquad \hat{F} = (n-2)\frac{\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2} = (n-2)\frac{\mathrm{SCE}}{\mathrm{SCR}}.$$

Loi du χ^2 à n degrés de liberté
Soient $X_1, \dots, X_n$ n v. a. indép. de loi $\mathcal{N}(0,1)$, alors

$$S = X_1^2 + \cdots + X_n^2$$

suit une loi du χ^2 à n degrés de liberté, notée $\chi^2(n)$.

Loi de Student à n degrés de liberté
La loi de Student à n degrés de liberté, notée $T(n)$, est la loi du quotient

$$T = \frac{N}{\sqrt{S/n}}$$

où N suit une loi $\mathcal{N}(0,1)$ et S suit une loi $\chi^2(n)$, N et S étant deux v. a. indép..

Loi de Fisher à n_1 et n_2 degrés de liberté
Soient S_1 et S_2 deux v. a. indép. de loi respectives $\chi^2(n_1)$ et $\chi^2(n_2)$. Alors le quotient

$$F = \frac{S_1/n_1}{S_2/n_2}$$

suit une loi de Fisher à n_1 et n_2 degrés de liberté, notée $F(n_1, n_2)$.

Théorème de Cochran
Soient E_1 et E_2 deux sous-espaces vectoriels orthogonaux de $E = \mathbb{R}^d$ de dimensions respectives k_1 et k_2 et soit Y un vecteur aléatoire de $\mathbb{R}^d$ de loi normale centrée isotrope de variance σ^2. Alors $P_{E_1}(Y)$ et $P_{E_2}(Y)$ sont deux v. a. gaussienne centrées indépendantes et $\|P_{E_1}(Y)\|^2$ (resp. $\|P_{E_2}(Y)\|^2$) est une loi $\sigma^2\chi^2(k_1)$ (resp. $\sigma^2\chi^2(k_2)$).

Exercice 1
Montrer les propriétés des estimateurs suivantes.
1. $\mathbb{E}(\hat{\beta}) = \beta$
2. $\mathbb{E}(\hat{\mu}) = \mu$
3. $\mathrm{Var}(\hat{\beta}) = \dfrac{\sigma^2}{\sum_{i=1}^n (X_i - \bar{X})^2}$
(Indication : $\sum_{i=1}^n (Y_i - \bar{Y})(X_i - \bar{X}) = \sum_{i=1}^n (X_i Y_i - \bar{X}\bar{Y})$)

Exercice 2
Montrer que $\mathrm{SCR} = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$ est une v. a. de loi $\sigma^2\chi^2(n-2)$.

Exercice 3
Montrer que sous l'hypothèse nulle : $\beta = 0$, $\hat{t}$ suit la loi $T(n-2)$.

Exercice 4
Montrer que sous l'hypothèse nulle : $\beta = 0$, $\hat{F}$ suit la loi $F(1, n-2)$.

3.2 Régression linéaire multiple

– $\hat{\theta} = (X'X)^{-1}X'Y$
– $\mathbb{E}(\hat{\theta}) = \theta$
– $\mathrm{Var}(\hat{\theta}) = \sigma^2(X'X)^{-1}$
– $\mathrm{SCR}(\hat{\theta}) = \|Y - \hat{Y}\|^2$ est une variable aléatoire indépendante de $\hat{\theta}$ et suit une loi $\sigma^2\chi^2(n-k-1)$.

Les 4 formules fondamentales sont issues de la minimisation en θ de la somme des carrés résiduelle (SCR), somme qui peut s'écrire matriciellement sous la forme

$$\mathrm{SCR}\,(\theta) = \|Y - X\theta\|^2 = (Y - X\theta)'(Y - X\theta).$$

3.3 Autres utilisations du test de Student

Il peut être intéressant de tester la position des paramètres par rapport à des valeurs particulières, ce qu'autorise le test de Student, comme l'indique le tableau ci dessous.

test unilatéral droit	$H_0 : \theta = \theta_0$ contre $H_1 : \theta > \theta_0$ $W = \left\{\hat{t} = \frac{\hat{\theta}-\theta_0}{\hat{\sigma}_{\hat{\theta}}} > t_\alpha\right\}$ avec $\alpha = \mathbb{P}\,(t_{n-k-1} > t_\alpha)$
test unilatéral gauche	$H_0 : \theta = \theta_0$ contre $H_1 : \theta < \theta_0$ $W = \left\{\hat{t} = \frac{\hat{\theta}-\theta_0}{\hat{\sigma}_{\hat{\theta}}} < t_\alpha\right\}$ avec $\alpha = \mathbb{P}\,(t_{n-k-1} < t_\alpha)$
test bilatéral	$H_0 : \theta = \theta_0$ contre $H_1 : \theta \neq \theta_0$ $W = \left\{\hat{t} = \frac{\lvert\hat{\theta}-\theta_0\rvert}{\hat{\sigma}_{\hat{\theta}}} > t_\alpha\right\}$ avec $\alpha = \mathbb{P}\,(\lvert t_{n-k-1}\rvert > t_\alpha)$

La notation θ désigne le coefficient β_j de la variable $X^{(j)}$ ou le terme constant μ du modèle linéaire estimé par la MMCO.

3.4 Étude de cas : CHAMPA

$$\ln Y = \mu + \beta_1 \ln X^{(1)} + \beta_2 \ln X^{(2)} + \beta_3 \ln X^{(3)}$$

Y : la consommation de champagne en volume par personne
$X^{(1)}$: le revenu par personne
$X^{(2)}$: le prix du champagne en euro constants
$X^{(3)}$: le prix des liqueurs et apéritifs en euro constants

	Y	X1	X2	X3
1	42.53	57.21	76.60	73.60
2	38.73	59.11	80.65	72.91
3	40.00	61.47	86.76	66.97
4	45.39	64.03	85.35	63.20
5	51.74	67.60	84.11	55.08
6	65.39	71.71	81.74	59.20
7	72.38	75.50	80.95	65.60
8	59.04	76.16	91.73	59.54
9	61.26	78.01	96.46	56.80
10	75.55	81.91	95.41	61.25
11	82.53	86.64	96.16	73.02
12	90.47	93.04	98.40	88.91

```
13 100.00 100.00 100.00 100.00
14 110.47 105.36  99.30 111.42
15 127.93 109.76  97.84 123.31
16 139.04 115.03  95.96 136.91
17 143.80 120.53 104.27 150.62
```

Dans une première étape, tester le modèle log-linéaire suivant.

$$\ln Y = \mu + \beta_1 \ln X^{(1)} + \beta_2 \ln X^{(2)} + \beta_3 \ln X^{(3)}$$

1. Le modèle est il globalement significatif ?
2. Quelle est la signification économique des coefficients ? Les coefficients ont-ils un signe conforme à vos attentes ?
3. Une variable ne doit-elle pas être éliminée ? Laquelle ? Pourquoi ?

Dans une seconde étape, tester le modèle log-linéaire suivant.

$$\ln Y = \mu + \beta_1 \ln X^{(1)} + \beta_2 \ln X^{(2)}$$

4. Ce modèle est il statistiquement satisfaisant ?
5. Testez l'hypothèse $\beta_1 > 1$.
6. Testez l'hypothèse « la baisse des prix du champagne entraîne un progrès de la consommation en volume ». Peut on dire la même chose de la consommation en valeur ?

Réalisez une régression ascendante.

7. Présentez de manière synthétique les résultats obtenus.
8. La régression ascendante donne t il les mêmes résultats que la régression descendante ? Qu'en concluez-vous ?

Chapitre 4

Les lois à queue épaisse

4.1 Lois α-stables

Il existe plusieurs définitions équivalentes de lois stables ainsi que plusieurs paramétrisations. Nous présentons ici trois définitions équivalentes.

Définition 4.1.1 *Un v.a. X dans $\mathbb{R}^d$ a une* loi stable *si pour tout $a, b \in \mathbb{R}_+$ il existe $c \in \mathbb{R}_+$ et un vecteur $D \in \mathbb{R}^d$ tels que*

$$aX_1 + bX_2 \stackrel{\mathcal{L}}{=} cX + D, \tag{1}$$

où X_1 et X_2 sont des copies indépendantes de X.

Le nombre c dans (1) vérifie $c^\alpha = a^\alpha + b^\alpha$ pour un nombre $0 < \alpha \leqslant 2$. La loi de X est dite *α-stable*. Si $D = 0$ on dit que X est *strictement α-stable*. Remarquons que si X est de loi gaussienne $\mathcal{N}(\mu, \sigma^2)$, alors les termes de gauche dans (1) sont respectivement $\mathcal{N}(a\mu, (a\sigma)^2)$ et $\mathcal{N}(b\mu, (b\sigma)^2)$, et le terme de droite est $\mathcal{N}(c\mu + D, (c\sigma)^2)$. Si on prend $c^2 = a^2 + b^2$ et $D = (a + b - c)\mu$, l'égalité (1) est établie, c'est-à-dire que les lois normales sont 2-stables.

La difficulté technique dans l'étude de lois α-stables est que, sauf dans les cas particuliers ($\alpha = 0.5, 1$ et 2), il n'y a pas de forme explicite pour la densité. Les seules informations utilisables pour les v.a. α-stables sont leur fonction caractéristique.

Définition 4.1.2 *La fonction caractéristique d'un v.a. α-stable ($0 < \alpha < 2$) dans $\mathbb{R}^d$ s'exprime par l'expression suivante :*

$$\phi_{\alpha,\sigma}(t) = \exp\left(-\frac{1}{C_\alpha}\int_{S^{d-1}} \psi_\alpha(\langle t, s\rangle)\sigma(ds) + i\langle \delta, t\rangle\right), \tag{2}$$

où $S^{d-1} = \{x \in \mathbb{R}^d \,|\, \|x\| = 1\}$, σ est une mesure finie sur la sphère unité S^{d-1}, δ est un vecteur dans $\mathbb{R}^d$, $\langle t, s\rangle$ représente le produit scalaire,

$$C_\alpha = \begin{cases} \frac{1-\alpha}{\Gamma(2-\alpha)\cos(\pi\alpha/2)}, & \alpha \neq 1, \\ \frac{2}{\pi}, & \alpha = 1, \end{cases} \tag{3}$$

et

$$\psi_\alpha(x) = \begin{cases} |x|^\alpha \left(1 - i\, sign(x) \tan\frac{\pi\alpha}{2}\right), & \alpha \neq 1, \\ |x| \left(1 + i\frac{\pi}{2}\, sign(x) \ln|x|\right), & \alpha = 1. \end{cases}$$

Cette définition montre que la loi stable dans $\mathbb{R}^d$ est spécifiée par un nombre α entre 0 et 2, *indice de stabilité*, une mesure finie σ sur S^{d-1} dite *mesure spectrale* et un vecteur δ dans $\mathbb{R}^d$.

Dans le cas unidimensionnel la sphère unité ne contient que deux points, i.e. $S^0 = \{-1, 1\}$. La mesure spectrale σ se réduit à deux valeurs $\sigma(-1)$ et $\sigma(1)$. La fonction caractéristique peut être écrite comme suit

$$\phi_{\alpha,\sigma}(t) = \begin{cases} \exp\left(-\frac{\gamma}{C_\alpha}|t|^\alpha \left(1 - i\beta \operatorname{sign}(t) \tan\frac{\pi\alpha}{2}\right) + i\delta t\right), & \alpha \neq 1, \\ \exp\left(-\frac{\gamma}{C_\alpha}|t| \left(1 + i\beta\frac{\pi}{2} \operatorname{sign}(t) \ln|t|\right) + i\delta t\right), & \alpha = 1, \end{cases}$$

où $\gamma = \sigma(1) + \sigma(-1)$ et $\beta = (\sigma(1) - \sigma(-1))/\gamma$. La loi stable unidimensionnelle est déterminée par quatre paramètres : $\alpha \in (0, 2]$, $\beta \in [-1, 1]$, $\gamma > 0$ et $\delta \in \mathbb{R}^1$. Le paramètre β contrôle l'asymétrie. La loi est dite *biaisée totalement vers la droite* si $\beta = 1$ et *biaisée totalement vers la gauche* si $\beta = -1$. Si $\beta = 0$ la densité est symétrique par rapport à δ. Les paramètres γ et δ sont les paramètres d'échelle et de position.

Nous utilisons la notation $\mathcal{S}_1(\alpha, \beta, \gamma, \delta)$ pour la loi α-stable unidimensionnelle. Le fait que X a la loi $\mathcal{S}_1(\alpha, \beta, \gamma, \delta)$ est noté par l'écriture "$X \sim \mathcal{S}_1(\alpha, \beta, \gamma, \delta)$". La Figure 4.1 présente l'influence des paramètres sur la forme de la densité d'une loi stable.

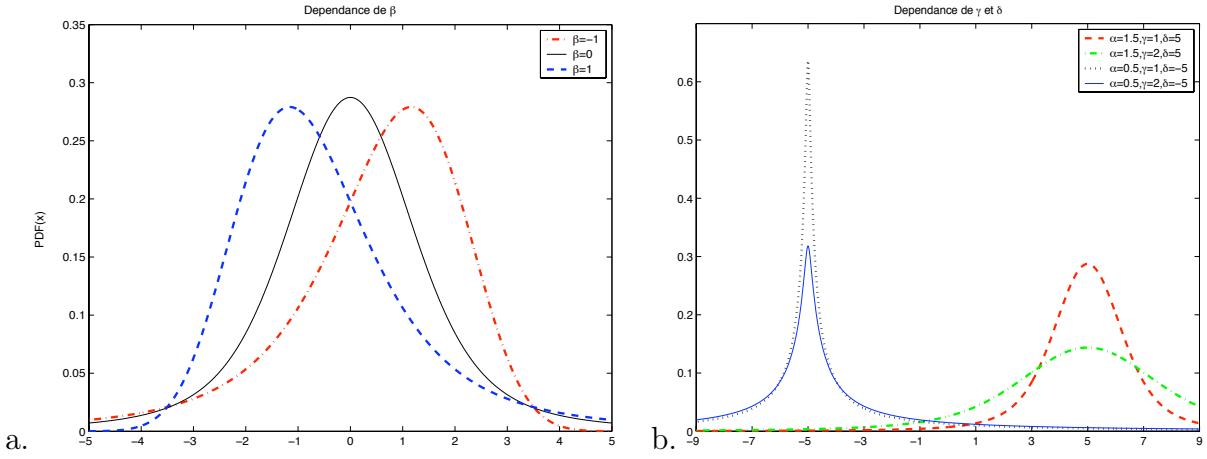


Figure 4.1 – Influence des paramètres sur la forme de la loi stable. Les densités de la loi $\mathcal{S}_1(1.5, (\beta, 1), 0)$ $\beta = -1, 0, 1$ (a) et de la loi $\mathcal{S}_1(\alpha, (0, \gamma), \delta)$ (b).

Propriété Si X a la loi α-stable avec $0 < \alpha < 2$ alors

$$\begin{aligned} \mathbf{E}\|X\|^p &< \infty \quad \text{pour tout} \quad 0 < p < \alpha, \\ \mathbf{E}\|X\|^p &= \infty \quad \text{pour tout} \quad p \geqslant \alpha. \end{aligned}$$

La troisième définition montre que les lois α-stables sont la seule limite possible non-triviale des sommes normalisées des v.a. i.i.d..

Définition 4.1.3 *Un v.a. X dans $\mathbb{R}^d$ a une* loi stable *s'il possède un* domaine d'attraction*, i.e. s'il existe une suite de v.a. i.i.d. $Y_1, Y_2, \ldots$ dans $\mathbb{R}^d$, une suite de nombres positifs $\{b_n\}$ et une suite de vecteurs réels $\{a_n\}$ telles que*

$$\frac{Y_1 + Y_2 + \cdots + Y_n}{b_n} + a_n \Rightarrow X, \tag{4}$$

où $\Rightarrow$ représente la convergence faible.

Si X est une variable aléatoire gaussienne et si les Y_i sont des variables aléatoires i.i.d. de variance finie, alors (4) est le théorème central limite ordinaire. En général, $b_n = n^{1/\alpha}L(n)$ où $L(x)$ est une fonction à variation lente.

Si $X \sim \mathcal{S}_1(\alpha, \beta, \gamma, \delta)$, alors on a

$$\begin{cases} \lim\limits_{x\to\infty} x^\alpha \mathbf{P}\{X > x\} &= \frac{(1+\beta)}{2}\gamma, \\ \lim\limits_{x\to\infty} x^\alpha \mathbf{P}\{X < -x\} &= \frac{(1-\beta)}{2}\gamma. \end{cases} \tag{5}$$

La Figure 4.2 présente le log-log graphe de la queue de lois stables.

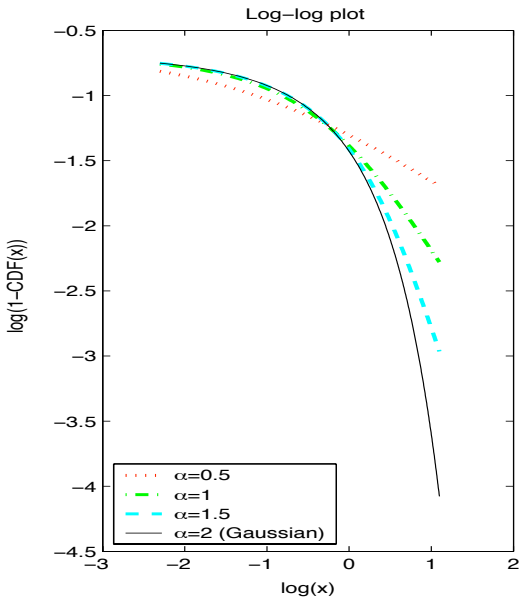


Figure 4.2 – Log-log graphe de $\mathbf{P}\{X > x\}$ où $X \sim \mathcal{S}_1(\alpha, 0, 1, 0)$, $\alpha = 0.5, 1, 1.5$ et 2.

4.2 Lois de valeurs extrêmes

Soit $(X_n, n \geq 1)$ une suite de variables aléatoires i.i.d.. Notons leurs maximum $M_n = \max(X_i, \ldots, X_n)$. Supposons qu'il existe une suite $((a_n, b_n), n \geq 1)$ telle que $a_n > 0$ et la suite $((M_n - b_n)/a_n, n \geq 1)$ converge en loi vers une limite non triviale. Alors à une

translation et un changement d'échelle près la fonction de répartition de la limite est de la forme suivante :

$$\text{Loi de Weibull} \quad \Psi_\alpha(x) = \begin{cases} e^{-(-x)^\alpha}, & x \leq 0 \\ 1, & x > 0 \end{cases} \quad \text{et } \alpha > 0.$$

$$\text{Loi de Gumbel} \qquad \Lambda(x) = e^{-e^{-x}}, x \in \mathbb{R}.$$

$$\text{Loi de Fréchet} \quad \Phi_\alpha(x) = \begin{cases} 0, & x \leq 0 \\ e^{-x^{-\alpha}}, & x > 0 \end{cases} \quad \text{et } \alpha > 0.$$

Exemples de convergence du maximum renormalisé

– Si $X_1 \sim \text{Unif}[0,1]$, alors la suite $(n(M_n - 1), n \geq 1)$ converge en loi vers une loi de Weibull.

– Si $X_1 \sim \text{Exp}(1)$, alors la suite $(M_n - \log(n), n \geq 1)$ converge en loi vers une loi de Gumbel.

Définition 4.2.1 *La loi $\mathcal{L}_0$ est dite* max-stable si pour tout $n \geq 2$, $(X_1, \ldots, X_n)$ étant des variables aléatoires indépendantes de loi $\mathcal{L}_0$, il existe $a_n > 0$ et $b_n \in \mathbb{R}$, tels que $(\max(X_1, \ldots, X_n) - b_n)/a_n$ suit la loi $\mathcal{L}_0$.

L'exercice suivant permet de vérifier que les lois de Weibull, Gumbel et Fréchet sont max-stables.

Exercice

Soit $(X_n, n \geq 1)$ une suite de variables aléatoires indépendantes, de même loi que X. On pose $M_n = \max(X_i, \ldots, X_n)$. Montrer que si X suit la loi de

– Weibull de paramètre α, alors M_n a même loi que $n^{-1/\alpha}X$;

– Gumbel, alors M_n a même loi que $X + \log(n)$;

– Fréchet de paramètre α, alors M_n a même loi que $n^{1/\alpha}X$.

4.3 Estimateur par paquet

On divise l'échantillon $\xi_1, \xi_2, \ldots, \xi_N$ en n groupes disjoints $G_{m,1}, \ldots, G_{m,n}$, chacun contient m éléments. En pratique on choisit d'abord m et on pose $n = [N/m]$ ($[a]$ représente la partie entière d'un nombre a positif). Quand N tend vers l'infini on a $nm = [N/m]m \sim N$. De plus on demande $n, m \to \infty$ quand $N \to \infty$.

On estime d'abord le paramètre α. Posons

$$M_{m,i}^{(1)} = \max\{\|\xi\| \mid \xi \in G_{m,i}\}, \ i = 1, \ldots, n, \tag{6}$$

c'est-à-dire $M_{m,i}^{(1)}$ est la plus grande norme dans le groupe $G_{m,i}$. Notons $\xi_{m,i}$ tel que

$$\|\xi_{m,i}\| = M_{m,i}^{(1)}, \tag{7}$$

et

$$M_{m,i}^{(2)} = \max\{\|\xi\| \mid \xi \in G_{m,i} \backslash \{\xi_{m,i}\}\}, \ i = 1, \ldots, n, \tag{8}$$

$M_{m,i}^{(2)}$ est alors la deuxième grande norme dans le même groupe.

On a pour chaque i

$$\left(\frac{M_{m,i}^{(1)}}{b_m}, \frac{M_{m,i}^{(2)}}{b_m}\right) \Rightarrow c(\Gamma_1^{-1/\alpha}, \Gamma_2^{-1/\alpha}) \quad \text{quand } m \to \infty, \tag{9}$$

où $b_m = m^{1/\alpha} L(m)$ et $c = \sigma(S)^{1/\alpha}$. Ici $\Gamma_i = \sum_{j=1}^{i} \lambda_j$, et $\lambda_1, \lambda_2, \ldots$ sont des variables aléatoires i.i.d. suivant la loi exponentielle standard, i.e. $\mathbf{E}(\lambda_i) = 1$. En posant

$$\varkappa_{m,i} = \frac{M_{m,i}^{(2)}}{M_{m,i}^{(1)}}, \qquad S_n = \sum_{i=1}^{n} \varkappa_{m,i} \tag{10}$$

on construit un estimateur du paramètre α comme suit,

$$\hat{\alpha}_N = \frac{S_n}{n - S_n}. \tag{11}$$

On peut prouver que cet estimateur est consistant, c'est-à-dire

$$\hat{\alpha}_N \xrightarrow[N\to\infty]{p.s.} \alpha. \tag{12}$$

Sous les conditions convenables sur m et n, on a la normalité asymptotique de l'estimateur, c'est-à-dire

$$\frac{\sqrt{n}\left(\frac{1}{n}S_n - \frac{\alpha}{\alpha+1}\right)}{\left(\frac{1}{n}\sum_{i=1}^{n} \varkappa_{m,i}^2 - \left(\frac{1}{n}S_n\right)^2\right)^{1/2}} \Rightarrow \mathcal{N}(0,1). \tag{13}$$

Dans la suite nous supposons que $\sigma(S) = 1$. Notons

$$\theta_{m,i} = \frac{\xi_{m,i}}{\|\xi_{m,i}\|}, \ i = 1, \ldots, n, \tag{14}$$

où $\xi_{m,i}$ est défini par (7). Les e.a. $\theta_{m,1}, \ldots, \theta_{m,n}$ sont i.i.d. à valeurs dans S.

On peut prouver que pour chaque i,

$$\theta_{m,i} \Rightarrow \sigma \text{ quand } m \to \infty. \tag{15}$$

Sous les conditions convenables sur m et n, on a aussi la normalité asymptotique de l'estimateur de β.

Bibliographie

[1] Peter Dalgaard, *Introductory Statistics with R*. Springer, 2002.

[2] Maria L. Rizzo, *Statistical Computing with R*. Chapman & Hall/CRC, 2008.