Notes de cours
Statistique avec le logiciel R

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

Année 2014-2015

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 Rappel aux lois de probabilités et les statistiques descriptives

Nous prenons la loi uniforme comme exemple pour rappeler les notions des lois de probabilités et les statistiques descriptives.

	discrète	continue
Notation	Unif(n)	Unif$[a,b]$
$X \in$	$\{1,\dots,n\}$	$[a,b]$
Masse/Densité	$\mathbb{P}(X=i)=\dfrac{1}{n}$	$f(x)=\dfrac{1}{b-a}$
Fdr	$F(x)=\begin{cases}0, & x<1,\\ \dfrac{[x]}{n}, & 1\le x\le n,\\ 1, & x>n.\end{cases}$	$F(x)=\begin{cases}0, & x<a,\\ \dfrac{x-a}{b-a}, & a\le x<b,\\ 1, & x\ge b.\end{cases}$
$\mathbb{E}(X)$	$\dfrac{n+1}{2}$	$\dfrac{a+b}{2}$
$\mathrm{Var}(X)$	$\dfrac{n^2-1}{12}$	$\dfrac{(b-a)^2}{12}$
Médiane/0.5-quantile	$\dfrac{n+1}{2}$	$\dfrac{a+b}{2}$

1.6 Statistique d'ordre

La statistique d'ordre de rank k d'un échantillon est égal à la k-ème plus petite valeur. Étant donné un échantillon $X_1, X_2, \dots, X_n$, les statistiques d'ordres, notées $X_{(1)}, X_{(2)}, \dots, X_{(n)}$, sont donc obtenues par tri croissant. Si on suppose l'échantillon $X_1, X_2, \dots, X_n$ i.i.d. issu d'une loi de densité $f(x)$ et de fonction de répartition F, alors la densité de la k-ème statistique d'ordre est

$$f_{X_{(k)}}(x) = \frac{n!}{(k-1)!(n-k)!} F(x)^{k-1}(1-F(x))^{n-k} f(x).$$

Exercice

1. Montrer que les fonctions de répartition de $X_{(1)}$ et $X_{(n)}$ sont

$$F_{X_{(1)}} = 1-(1-F(x))^n \quad \text{et} \quad F_{X_{(n)}} = (F(x))^n.$$

2. Soit $X_1, X_2, \dots, X_n$ un n-échantillon i.i.d. de loi Unif$[a,b]$. Montrer que $X_{(1)}$ et $X_{(n)}$ sont respectivement les estimateurs du maximum de vraisemblance de a et b.

Dans la suite, on va présenter des applications de statistiques d'ordre afin de rappeler quelques sujets de statistique inférentielle : les estimations paramétriques, les tests statistiques, les convergences des suites de v.a., etc..

Les statistiques d'ordre peuvent être utilisées pour estimer les paramètres. Par exemple, soit $X_1, X_2, \ldots, X_n$ une suite de v.a. i.i.d. de loi gaussienne d'espérance μ et de variance σ^2 inconnus. On peut estimer μ et σ en utilisant les statistiques suivantes

$$\hat{\mu} = X_{(\lceil n\times 0.5\rceil)} \quad \text{et} \quad \hat{\sigma} = (X_{(\lceil n\times 0.84\rceil)} - X_{(\lceil n\times 0.16\rceil)})/2. \tag{1}$$

La figure 1.1 montre que la moyenne et la médiane convergent toutes deux vers l'espérance réelle pour la loi gaussienne, tandis que seule la médiane converge pour la loi de Cauchy.

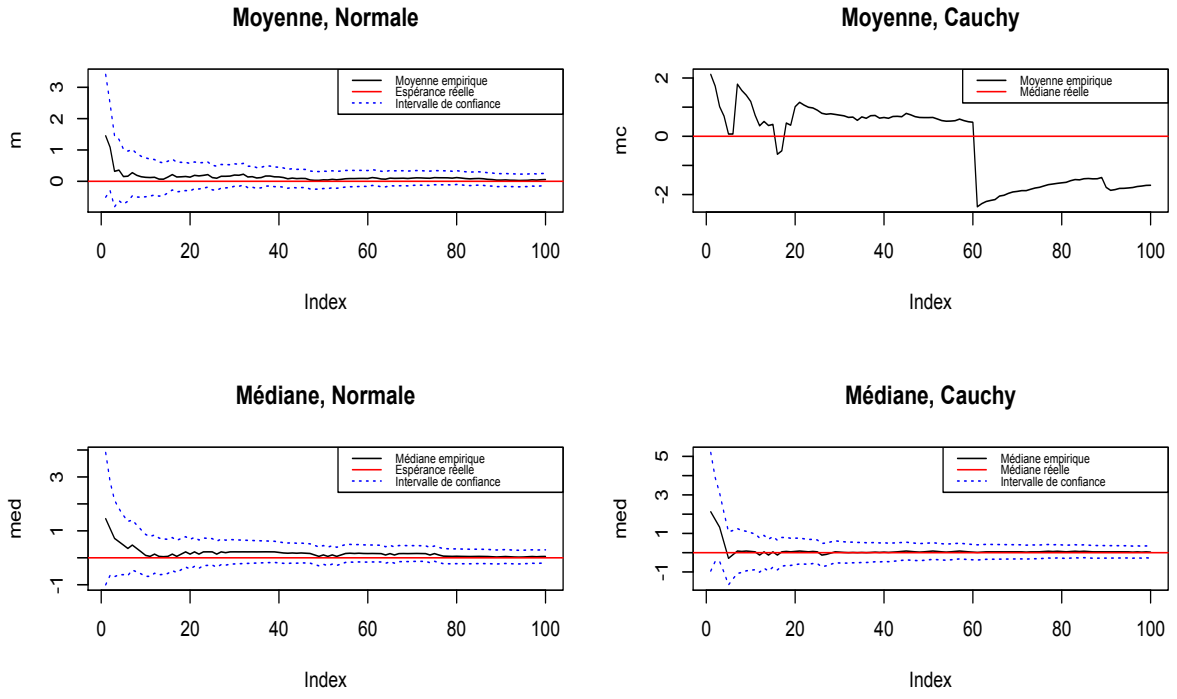

Figure 1.1 – Convergence de moyenne et médiane empiriques pour la loi gaussienne et la loi de Cauchy

Les intervalles de confiance de médiane dans la figure 1.1 sont calculés avec le théorème central limite suivant

$$\sqrt{n}(X_{(\lceil n\times p\rceil)} - q_p) \to \mathcal{N}\left(0, \frac{p(1-p)}{(f(F^{-1}(p)))^2}\right)$$

où q_p est le quantile d'ordre p réel, $f(x)$ et $F(x)$ sont la densité est la fonction de répartition de X.

Code R

```
n = 100;x = rnorm(n);m = cumsum(x)/(1:n);v = 1;q = qnorm(0.975)
ic1 = m+q*sqrt(v)/sqrt(1:n);ic2 = m-q*sqrt(v)/sqrt(1:n)
plot(m,type="l");abline(h=0,col="red")
lines(ic1,col="blue",lty="dotted");lines(ic2,col="blue",lty="dotted")

xc = rcauchy(n,0,1);mc = cumsum(xc)/(1:n)
plot(mc,type="l");abline(h=0,col="red")

med = rep(0,n);for (i in 1:n) med[i] = quantile(x[1:i],probs=0.5)
p = 0.5;v = p*(1-p)/(dnorm(qnorm(p)))^2
ic1 = med+q*sqrt(v)/sqrt(1:n);ic2 = med-q*sqrt(v)/sqrt(1:n)
plot(med,type="l");abline(h=0,col="red")
lines(ic1,col="blue",lty="dotted");lines(ic2,col="blue",lty="dotted")

for (i in 1:n) med[i] = quantile(xc[1:i],probs=0.5)
v = p*(1-p)/(dcauchy(qcauchy(p)))^2
# les calculs de ic1 et ic2 et le code de plot pour la loi de Cauchy
# sont identiques que ceux dans le cas gaussien
```

Exercice

1. Pourquoi la moyenne empirique de loi de Cauchy ne converge pas ?

2. Pourquoi $\hat{\sigma}$ défini dans (1) est un estimateur de σ ? Peut on l'utiliser pour estimer l'écart-type d'une loi uniforme ?

3. Proposer un estimateur d'écart-type d'une loi uniforme en utilisant les statistiques d'ordre et vérifier la convergence avec l'étude de simulation.

4. Écrire un programme en R pour illustrer la convergence d'estimateur $\hat{\sigma}$ défini dans (1).

Une deuxième utilisation de statistiques d'ordre est le diagramme quantile-quantile. C'est un outil graphique qui permet d'évaluer la pertinence de l'ajustement d'une loi de probabilités. À l'issue d'une enquête statistique, on soupçonne celle-ci de suivre une distribution *classique*, par exemple la loi gaussienne. À partir de la série statistique observée, on calcule alors un certain nombre de quantiles $q_{\frac{i}{n+1}}$. Si la série statistique suit bien la loi théorique choisie, on devrait avoir les quantiles observés $X_{(\lceil n\times\frac{i}{n+1}\rceil)}$ égaux aux quantile $q_{\frac{i}{n+1}} = F^{-1}(\frac{i}{n+1})$ associés au modèle théorique. On place alors le nuage de point $(X_{(\lceil n\times\frac{i}{n+1}\rceil)}, F^{-1}(\frac{i}{n+1}))$. En abscisse se trouvent donc les quantiles théoriques et en ordonnée les quantiles observés. Si la loi théorique choisie est pertinente, les points doivent se positionner suivant la première diagonale.

Code R

```
x = rnorm(100)
```

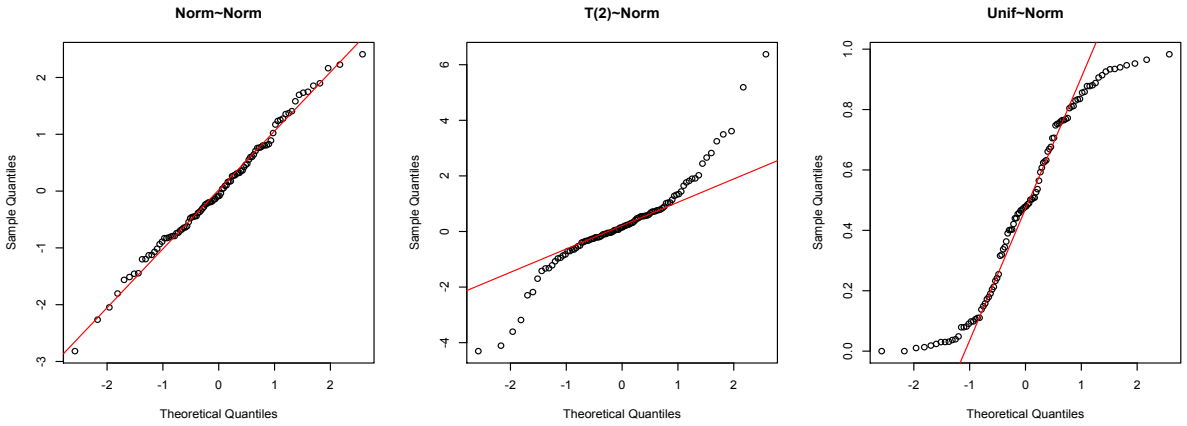

Figure 1.2 – Diagramme quantile-quantile des données issues de loi gaussienne, loi de Student et loi uniforme

```
qqnorm(x,main="Norm~Norm");qqline(x, col = 2)

y = rt(100,df=2)
qqnorm(y,main="T(2)~Norm");qqline(y, col = 2)

z = runif(100)
qqnorm(z,main="Unif~Norm");qqline(z, col = 2)

qqplot(x, rt(100,df=2));qqline(x, col = 2)
```

La troisième application de statistiques d'ordre est les tests d'adéquation. Par exemple, le test de normalité, test de Shapiro-Wilk, peut être interprété avec un diagramme quantile-quantile. Le test Shapiro-Wilk teste l'hypothèse nulle selon la quelle un échantillon est issu d'une loi normale. La statistique de test est

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

où $x_{(i)}$ désigne la i-ème statistique d'ordre, a_i sont des constantes générées à partir de la moyenne et de la matrice de variance co-variance des quantiles d'un échantillon de taille n suivant la loi normale. La statistique W peut être interprétée comme le coefficient de détermination entre la série des quantiles générés à partir de la loi normale et les quantiles empiriques obtenus à partir des données. Plus W est élevé, plus la compatibilité avec la loi normale est crédible.

Pour tester l'adéquation d'un échantillon à une loi qui n'est pas nécessairement normale, on peut utiliser le test de Kolmogorov-Smirnov (KS). Ce test repose sur les propriétés des fonctions de répartition empiriques. Soit $X_1, X_2, \ldots, X_n$ une suite de n v.a. i.i.d.. La fonction de répartition empirique de cet échantillon est définie par

$$\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{X_i \leq x\}}(x) = \begin{cases} 0, & x < X_{(1)}, \\ \frac{k}{n}, & X_{(k)} \leq x < X_{(k+1)}, \quad k = 1, \ldots, n-1, \\ 1, & x \geq X_{(n)}. \end{cases} \tag{2}$$

La figure 1.3 présente la fonction de répartition empirique d'un 10-échantillon de loi uniforme $[0,1]$. On a la convergence suivante

$$\sqrt{n}\sup_x |\hat{F}_n(x) - F(x)| \to K$$

où K est une v.a. de loi de Kolmogorov. Notons la statistique $D_n = \sup_x |\hat{F}_n(x) - F(x)|$. Si l'échantillon provient bien de la loi donnée $F(x)$, D_n devrait être petite. La région critique, rejet de l'adéquation, s'écrit donc

$$\sqrt{n}D_n > K_\alpha$$

où $\mathbb{P}\left(\sqrt{n}D_n < K_\alpha\right) = 1 - \alpha$.

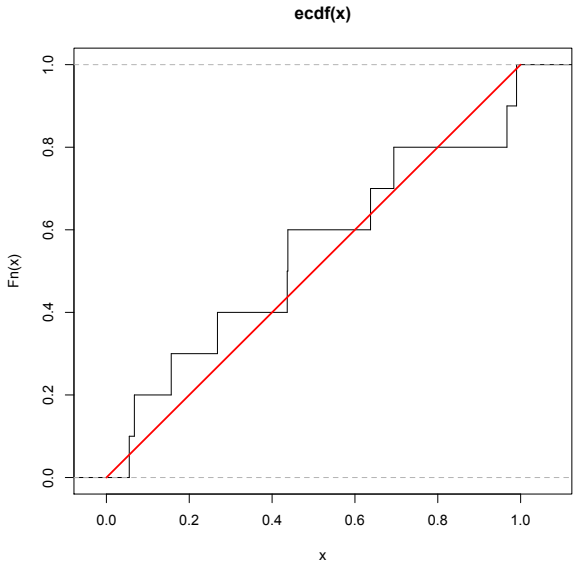

Figure 1.3 – Fonction de répartition empirique d'un 10-échantillon de loi uniforme $[0,1]$

Code R

```
x = runif(10);fn = ecdf(x);z = seq(0,1,0.01)
plot(fn,verticals = TRUE, do.points = FALSE)
lines(z,punif(z),col="red",lwd=2)

x = rnorm(50);y = runif(50);ks.test(x, y)
```

Exercice

1. Soit $X_1, X_2, \ldots, X_n$ une suite de n v.a. i.i.d.. La fonction de répartition de X_i est $F(x)$. Soit $\hat{F}_n(x)$ la fonction de répartition empirique définie par (2).
a) Calculer l'espérance et la variance de $\hat{F}_n(x)$ en fonction de $F(x)$.
b) Démontrer que pour tout $x \in \mathbf{R}$, $\hat{F}_n(x) \xrightarrow[n\to\infty]{P} F(x)$.
c) Montrer que pour tout $x \in \mathbf{R}$, $\hat{F}_n(x)$ vérifie un théorème de limite centrale que l'on établira.
d) À partir d'un contre-exemple, montrer que la convergence précédente n'a plus lieu en général.

2. Un vigneron veut savoir quelle est la contenance moyenne des bouteilles qu'il produit. Il effectue pour cela une mesure sur un échantillon de 10 bouteilles et obtient, en centilitres, les volumes suivants.

1	2	3	4	5	6	7	8	9	10
76	75	77	74	73	77	72	74	74	73

Il déduit que la moyenne empirique est 74.5. Supposons que la contenance d'une bouteille prise au hasard dans le production est distribuée selon une loi normale de moyenne μ et de variance σ^2. Alors peut on croire que $\mu = 75$?
a) $\sigma^2 = 1$.
b) σ^2 est inconnue.

1.7 Rappel aux tests statistiques

Un test statistique permet de décider entre deux hypothèses H_0 et H_1. Cette décision se fera à partir d'une réalisation d'un échantillon. On associe à un test un niveau α, ou risque de première espèce (généralement entre 1% et 10%). Une fois que le niveau α est fixé on peut en déduire la région de rejet.

1.7.1 Construction d'un test

1. Choix du niveau α

Un choix standard est $\alpha = 5\%$. Pour savoir plus sur cette norme officieuse, voir la section 2.4 "La canonisation du 5%" dans [1].

2. Choix des hypothèses H_0 et H_1

On veut tester l'appartenance du paramètre θ à un ensemble de valeurs Θ_0. On forme donc le test

$$H_0 : \theta \in \Theta_0 \quad \text{contre} \quad H_1 : \theta \in \Theta_1,$$

où $\Theta_0 \cap \Theta_1 = \emptyset$. Une hypothèse est dite simple si elle est associée à un singleton, c'est-à-dire $H_0 : \theta = \theta_0$ ou $H_1 : \theta = \theta_1$, sinon elle sera dite multiple, ou bien composite. Si H_1 est

multiple de la forme $\theta > \theta_0$ ou $\theta < \theta_0$, on parlera de test unilatéral. Si H_1 est multiple de la forme $\theta \neq \theta_0$ on parlera de test bilatéral.

3. Choix de la statistique de test

La statistique est une quantité calculée à partir d'un échantillon qui suit une loi connue sous l'hypothèse H_0. On note T "la statistique théorique" – la variable aléatoire qui suit la loi de statistique choisie, $\hat{T}$ "la statistique empirique" – la statistique calculée à partir de la réalisation de l'échantillon.

4. Détermination de la région de rejet

La région de rejet, notée W, est l'ensemble de valeur auquel la statistique choisie appartient sous l'hypothèse H_0 avec la probabilité égale à α, i.e.

$$\mathbb{P}\left(T \in W \,|\, H_0\right) = \alpha.$$

Autrement dit si H_0 est vraie la statistique T a peu de chance d'être dans W. Donc si la statistique calculée $\hat{T}$ fait partie de W, on a la raison suffisante pour rejeter H_0, d'où vient le nom "la région de rejet".

5. Conclusion pour la réalisation de l'échantillon

On calcule la statistique $\hat{T}$ et compare avec la région de rejet ; La décision est prise de façon suivante.

$$\begin{cases} \text{rejet de } H_0 & \text{si } \hat{T} \in W, \\ \text{acceptation de } H_0 & \text{si } \hat{T} \notin W. \end{cases}$$

Remarque 1. Un test a le plus souvent pour but de rejeter H_0 plutôt que de choisir entre deux hypothèses, puisque **accepter H_0 ne veut pas dire que cette hypothèse est vraie mais qu'il n'y pas de raison suffisante pour la rejeter. Rejeter H_0 est donc beaucoup riche en information que l'accepter**. C'est pour cela que dans le cas de rejet on dit que le test est significatif. Les deux hypothèses ne sont donc pas interchangeables.

Un exemple simple est présenté ci dessous pour illustrer la construction d'un test et en séduire l'intérêt de la notion p-value.

1.7.2 Exemple simple

On dispose d'un échantillon gaussien $X_1, \ldots, X_n$ de loi $\mathcal{N}(\mu, \sigma^2)$ avec μ inconnu et on veut réaliser le test suivant,

$$H_0 : \mu = m \quad \text{contre} \quad H_1 : \mu \neq m,$$

pour une valeur m donnée. On choisit intuitivement [1] la statistique de test $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$. On sait que sous H_0, $\bar{X}$ est distribuée suivant une loi gaussienne $\mathcal{N}(\mu, \sigma^2/n)$, ainsi $\frac{\sqrt{n}}{\sigma}(\bar{X} - m) \sim \mathcal{N}(0, 1)$. Soit un niveau $\alpha = 5\%$. On trouve dans la table de loi gaussienne centrée et réduite le quantile $q_{97.5\%} = 1.96$, i.e.

$$\mathbb{P}\left(\frac{\sqrt{n}}{\sigma}(\bar{X} - m) < 1.96 \,|\, H_0\right) = 0.975.$$

1. On peut bien sûr choisir une autre statistique, par exemple X_1. Qu'obtient-on ?

Puisque le test est bilatéral, on déduit

$$\mathbb{P}\,(\frac{\sqrt{n}}{\sigma}|\bar{X} - m| \geq 1.96 \,|\, H_0) = 0.05,$$

ainsi,

$$\mathbb{P}\,(\bar{X} \in (-\infty, m - 1.96\sigma/\sqrt{n}) \cup (m + 1.96\sigma/\sqrt{n}, \infty) \,|\, H_0) = 0.05.$$

La région de rejet est donc

$$W = (-\infty, m - 1.96\sigma/\sqrt{n}) \cup (m + 1.96\sigma/\sqrt{n}, \infty).$$

Notons $\hat{\bar{X}}$ la moyenne empirique calculée à partir d'une réalisation de l'échantillon. La décision est la suivante.

$$\begin{cases} \text{rejet de } \ H_0 & \text{si } \ \hat{\bar{X}} < m - 1.96\sigma/\sqrt{n} \ \text{ ou } \ \hat{\bar{X}} > m + 1.96\sigma/\sqrt{n}, \\ \text{acceptation de } \ H_0 & \text{si } \ m - 1.96\sigma/\sqrt{n} \leq \hat{\bar{X}} \leq m + 1.96\sigma/\sqrt{n}. \end{cases}$$

Cela veut dire qu'on rejettera l'hypothèse $H_0 : \mu = m$ si la moyenne observée est "loin" de m. Le niveau α est la probabilité de commettre une erreur qu'on rejette H_0 alors qu'elle est vraie. Autrement dit le risque de rejeter à tort l'hypothèse nulle est 5%.

Remarquons que la région de rejet est liée au choix du niveau α. Pour un niveau $\alpha = 10\%$ on obtient la région de rejet suivante

$$W = (-\infty, m - 1.64\sigma/\sqrt{n}) \cup (m + 1.64\sigma/\sqrt{n}, \infty),$$

qui inclut celle de $\alpha = 5\%$. Ainsi pour tous

$$\hat{\bar{X}} \in (m - 1.96\sigma/\sqrt{n}, m - 1.64\sigma/\sqrt{n}) \cup (m + 1.64\sigma/\sqrt{n}, m + 1.96\sigma/\sqrt{n}),$$

l'hypothèse H_0 sera rejetée si $\alpha = 10\%$, mais acceptée si $\alpha = 5\%$. C'est-à-dire, la décision du test dépend non seulement des observations, mais aussi du niveau α. Plus le niveau choisi est petit, plus il est difficile d'obtenir le résultat significatif.

Rappelons que le niveau α est en fait le risque seuil, en dessous duquel on est prêt à rejeter H_0. Un risque de 5% veut dire que dans 5% des cas quand H_0 est vraie, l'expérimentateur se trompera et la rejettera. Mais le choix du seuil à employer dépendra de la certitude désirée et de la vraisemblance des alternatives. On souhaite donc d'obtenir le résultat qui est indépendant du risque seuil et permet de choisir le risque seuil a posteriori.

1.7.3 La p-value

La p-value ou niveau de signification est la valeur critique de α qui fait basculer le résultat du test. Elle dépend uniquement de la réalisation de l'échantillon et permet de faire la décision avec le niveau α choisi arbitrairement. On a

$$\begin{cases} \text{rejet de } \ H_0 & \text{si } \ \alpha > \text{p-value}, \\ \text{acceptation de } \ H_0 & \text{si } \ \alpha < \text{p-value}. \end{cases}$$

La p-value est en fait la valeur la plus petite de α qui permet de rejeter H_0. Si la p-value est plus grande que α choisi a priori, le test est non concluant, ce qui revient à dire que

l'on ne peut rien affirmer. Prenons l'exemple du test bilatéral présenté précédemment, la p-value peur être calculée,

$$\text{p-value} = \mathbb{P}\left(\frac{\sqrt{n}}{\sigma}|\bar{X} - m| \geq \frac{\sqrt{n}}{\sigma}|\hat{\bar{X}} - m| \,\middle|\, H_0 \right).$$

C'est la probabilité, en supposant que H_0 est vraie, d'obtenir une valeur de la variable de décision $|\bar{X} - m|$ au moins aussi grande que la valeur de la statistique que l'on a obtenue avec notre échantillon $|\hat{\bar{X}} - m|$. Puisque sous H_0, on a $\frac{\sqrt{n}}{\sigma}(\bar{X} - m) \sim \mathcal{N}(0, 1)$. En notant $\mathcal{N}(0, 1)$ une variable aléatoire gaussienne standard, on a

$$\text{p-value} = \mathbb{P}\left(|\mathcal{N}(0, 1)| \geq \frac{\sqrt{n}}{\sigma}|\hat{\bar{X}} - m| \right).$$

Chapitre 2

Modèle linéaire

Notre but est d'étudier la relation de détermination de Y par k variables explicatives $X^{(1)}, \ldots, X^{(k)}$,

$$Y_i = \mu + \beta_1 X_i^{(1)} + \beta_2 X_i^{(2)} + \cdots + \beta_k X_i^{(k)} + \varepsilon_i, \quad i = 1, \ldots, n.$$

Il s'en suit l'écriture matricielle suivante

$$\begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} 1 & X_1^{(1)} & X_1^{(2)} & \cdots & X_1^{(k)} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & X_n^{(1)} & X_n^{(2)} & \cdots & X_n^{(k)} \end{pmatrix} \begin{pmatrix} \mu \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

ou encore

$$Y = X\theta + \varepsilon \quad \text{avec} \quad \theta = \begin{pmatrix} \mu \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix} \quad \text{et} \quad X = \begin{pmatrix} 1 & X_1^{(1)} & X_1^{(2)} & \cdots & X_1^{(k)} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & X_n^{(1)} & X_n^{(2)} & \cdots & X_n^{(k)} \end{pmatrix}.$$

Le vecteur aléatoire ε appelé erreur du modèle vérifie les 4 postulats suivants :

pour tous $i = 1, \ldots, n$,

1. $\mathbb{E}(\varepsilon_i) = 0$
2. $\mathrm{Var}(\varepsilon_i) = \sigma^2$
3. ε_i sont i.i.d. de loi gaussienne.
4. X_i sont déterministes.

2.1 Régression linéaire simple

Pour le modèle linéaire simple, i.e. $k = 1$, les estimateurs sont les suivants

$$\hat{\beta} = \frac{\mathrm{Cov}(Y, X)}{\mathrm{Var}(X)}, \quad \hat{\mu} = \bar{Y} - \hat{\beta}\bar{X}, \quad \hat{\sigma}^2 = \frac{\sum_{i=1}^n \hat{\varepsilon}_i^2}{n-2}$$

où $\mathrm{Cov}(Y, X) = \frac{1}{n}\sum_{i=1}^n (Y_i - \bar{Y})(X_i - \bar{X})$, $\mathrm{Var}(X) = \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2$, $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$ et $\bar{Y} = \frac{1}{n}\sum_{i=1}^n Y_i$.

On a les propriétés suivantes.

1. $\hat{\mu} \sim \mathcal{N}\left(\mu, \dfrac{\sigma^2 \sum X_i^2}{n^2 \mathrm{Var}\,(X)}\right)$
2. $\hat{\beta} \sim \mathcal{N}\left(\beta, \dfrac{\sigma^2}{n \mathrm{Var}\,(X)}\right)$
3. $\mathrm{Cov}\,(\hat{\mu}, \hat{\beta}) = -\dfrac{\sigma^2 \bar{X}}{n \mathrm{Var}\,(X)}$
4. $\dfrac{\hat{\sigma}^2(n-2)}{\sigma^2} \sim \chi^2(n-2)$
5. $(\hat{\mu}, \hat{\beta})$ et $\hat{\sigma}^2$ sont indépendantes.
6. $\mathrm{Var}\,(\hat{Y}_i) = \dfrac{\sigma^2}{n}\left(1 + \dfrac{(X_i - \bar{X})^2}{\mathrm{Var}\, X}\right)$
7. Sous l'hypothèse nulle : $\beta = 0$, la statistique $\hat{t} = \dfrac{\hat{\beta}}{\hat{\sigma}_{\hat{\beta}}}$ suit la loi $T(n-2)$, où $\hat{\sigma}^2_{\hat{\beta}} = \dfrac{\hat{\sigma}^2}{n \mathrm{Var}\, X}$.
8. Sous l'hypothèse nulle : $\beta = 0$, la statistique $\hat{F} = (n-2)\dfrac{\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}$ suit la loi $F(1, n-2)$.
9. L'intervalle de confiance au niveau $1-\alpha$ pour $\mathbb{E}\,(Y_i)$ est

$$\left[\hat{Y}_i - t_{n-2}(\alpha/2)\sqrt{\widehat{\mathrm{Var}}\,(\hat{Y}_i)}, \hat{Y}_i + t_{n-2}(\alpha/2)\sqrt{\widehat{\mathrm{Var}}\,(\hat{Y}_i)}\right].$$

10. Notons Y_i^0 la nouvelle observation indépendante des premières observations. L'intervalle de confiance au niveau $1-\alpha$ pour Y_i^0 est

$$\left[\hat{Y}_i - t_{n-2}(\alpha/2)\sqrt{\widehat{\mathrm{Var}}\,(\hat{Y}_i) + \hat{\sigma}^2}, \hat{Y}_i + t_{n-2}(\alpha/2)\sqrt{\widehat{\mathrm{Var}}\,(\hat{Y}_i) + \hat{\sigma}^2}\right]$$

où $\widehat{\mathrm{Var}}\,(\hat{Y}_i) = \dfrac{\hat{\sigma}^2}{n}\left(1 + \dfrac{(X_i - \bar{X})^2}{\mathrm{Var}\, X}\right)$ et $t_{n-2}(\alpha/2)$ est le $\alpha/2$-quantile d'une loi de Student à $n-2$ degrés de liberté.

Exercice

1. Montrer que $\dfrac{\hat{\sigma}^2(n-2)}{\sigma^2}$ suit la loi $\chi^2(n-2)$.
2. Montrer que $\dfrac{\hat{Y}_i - \mathbb{E}\,(Y_i)}{\sqrt{\widehat{\mathrm{Var}}\,(\hat{Y}_i)}}$ suit la loi $T(n-2)$.
3. Montrer que $\dfrac{Y_i^0 - \hat{Y}_i}{\sqrt{\widehat{\mathrm{Var}}\,(\hat{Y}_i) + \hat{\sigma}^2}}$ suit la loi $T(n-2)$.

2.2 Un jeu de données

On considère un jeu de données (Y, X) où Y représente la tension artérielle et X représente l'âge.

Y	114	124	143	158	166
X	35	45	55	65	75

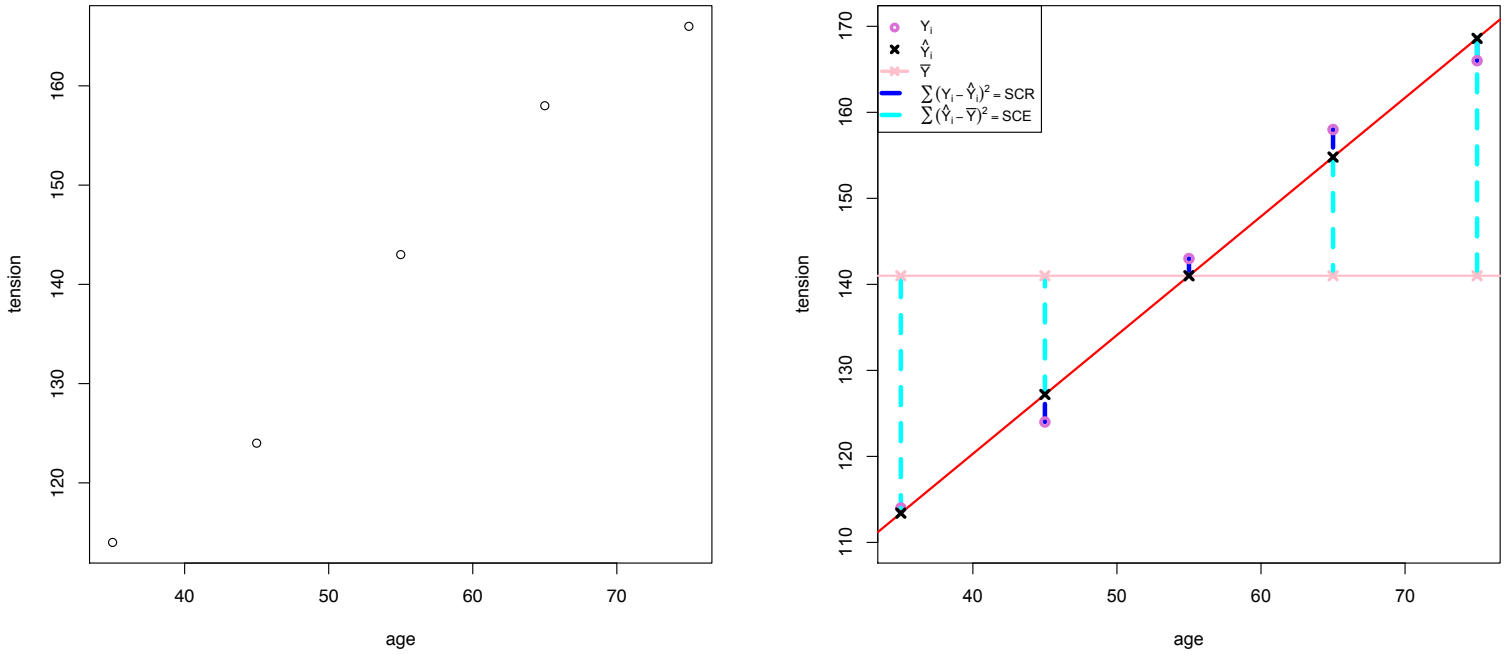

Figure 2.1 – Nuage de points des données `Tension ~ Age`, droite de régression, SCE et SCR

Commentaires du graphique :
– La tension artérielle augmente avec l'âge.
– Les points du graphe sont presque alignés.

Sorties de l'estimation des paramètres traitée par R

```
Call:
lm(formula = tension ~ age, data = Tens)

Residuals:
   1    2    3    4    5
 0.6 -3.2  2.0  3.2 -2.6

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
 (Intercept)  65.1000     5.8284   11.17 0.001538 **
 age           1.3800     0.1026   13.45 0.000889 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.246 on 3 degrees of freedom
Multiple R-squared: 0.9837,Adjusted R-squared: 0.9782
F-statistic: 180.8 on 1 and 3 DF,  p-value: 0.0008894
```

Les sorties de l'estimation des paramètres sont expliquées dans la table suivante.

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	$\hat{\mu}$	$\hat{\sigma}_{\hat{\mu}}$	$\hat{t}_\mu = \hat{\mu}/\hat{\sigma}_{\hat{\mu}}$	$\mathbb{P}\left(\lvert t_{n-2}\rvert > \lvert\hat{t}_\mu\rvert\right)$	très significative
age	$\hat{\beta}$	$\hat{\sigma}_{\hat{\beta}}$	$\hat{t}_\beta = \hat{\beta}/\hat{\sigma}_{\hat{\beta}}$	$\mathbb{P}\left(\lvert t_{n-2}\rvert > \lvert\hat{t}_\beta\rvert\right)$	hautement significative
Residual standard error			$\hat{\sigma}$		
Multiple R-squared	$R^2 = \frac{\text{SCE}}{\text{SCT}}$		Adjusted R-squared	$R^2_{\text{ajusté}} = 1 - \frac{\text{SCR}/n-k-1}{\text{SCT}/n-1}$	
F-statistic	$\hat{F} = (n-k-1)\frac{\text{SCE}}{\text{SCR}}$		p-value	$\mathbb{P}\left(F(k, n-k-1) > \hat{F}\right)$	

Sorties de l'analyse de la variance traitée par R

```
Analysis of Variance Table

Response: tension
          Df Sum Sq Mean Sq F value    Pr(>F)
age        1 1904.4 1904.40   180.8 0.0008894 ***
Residuals  3   31.6   10.53
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Les sorties de l'analyse de la variance sont expliquées dans la table suivante.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
age	1	SCE	$\text{SCE}/1$	$\hat{F}$	$\mathbb{P}\left(F(1, n-k-1) > \hat{F}\right)$
Residuals	$n-k-1$	SCR	$\text{SCR}/(n-k-1)$		

$\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 := \text{SCE}$
$\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 := \text{SCR}$
$\sum_{i=1}^{n}(Y_i - \bar{Y})^2 := \text{SCT}$

Équation d'analyse de variance : $\text{SCT} = \text{SCE} + \text{SCR}$

Code R

```
age=c(35,45,55,65,75);tension=c(114,124,143,158,166)
reg=lm(tension~age);summary(reg);anova(reg)
```

Dans les sorties d'exemple traité par R les valeurs de `Pr(>|t|)` et `Pr(>F)` donnent directement les résultats de test. C'est un indicateur de plausibilité de l'hypothèse H_0 qui prend une valeur comprise entre 0 et 1. Plus cette valeur est élevée, plus il sera raisonnable de retenir l'hypothèse H_0.

– significative dès lors que "$Pr(>|t|)" < 0.05$
– très significative dès lors que "$Pr(>|t|)" < 0.01$
– hautement significative dès lors que "$Pr(>|t|)"< 0.001$

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 $\alpha = \mathbb{P}\left(t_{n-k-1} > t_\alpha\right)$
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 $\alpha = \mathbb{P}\left(t_{n-k-1} < t_\alpha\right)$
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 $\alpha = \mathbb{P}\left(\lvert t_{n-k-1}\rvert > t_\alpha\right)$

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.

Remarque 2. Par symétrie de la loi de Student, on a

$$\alpha = \mathbb{P}\left(t_{n-k-1} > t_\alpha\right) = \mathbb{P}\left(t_{n-k-1} < -t_\alpha\right).$$

Ainsi dans le test unilatéral gauche, la région de rejet peut s'écrire $W = \left\{\hat{t} = \frac{\hat{\theta}-\theta_0}{\hat{\sigma}_{\hat{\theta}}} < -t_\alpha\right\}$ avec $\alpha = \mathbb{P}\left(t_{n-k-1} > t_\alpha\right)$.

Exercice

1. Comment calculer le coefficient de détermination à partir de la table d'analyse de la variance ? Vérifier R^2 en utilisant les sorties de l'appel `anova(reg)`.

2. Le coefficient β est il significativement égal à 1 ? supérieur à 1 ? inférieur à 1.5 ?

Avant toute analyse des estimations, il faut examiner les résidus $\hat{\varepsilon}_i$ pour voir si les présupposés de la régression sont vérifiés. On examine ici deux graphiques :

– graphique des résidus contre les valeurs ajustées. Ils peuvent montrer des erreurs dans la spécification de la moyenne ;

– Q-Q plot de normalité.

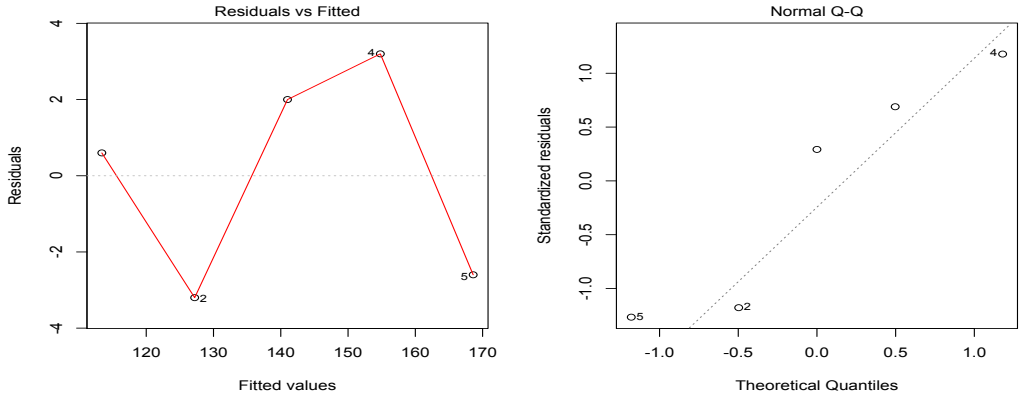

Figure 2.2 – Graphique des résidus contre les valeurs ajustées et Q-Q plot de normalité

À partir des paramètres estimés $\hat{\mu}$, $\hat{\beta}$ et $\hat{\sigma}$, on simule 100 jeux de données selon le modèle $Y_i = \hat{\mu} + \hat{\beta} X_i + \varepsilon_i$ où $\varepsilon_i \sim \mathcal{N}(0, \hat{\sigma}^2)$. Les données simulées sont considérées comme les observations obtenues par les expériences différentes. La méthode des moindres carrés est appliquée à chaque jeu pour obtenir les estimations différentes de μ et β. On obtient ainsi 100 droites de régression. On compare les intervalles de confiance obtenus à partir du jeu de données original et les résultats de simulation. La figure 2.3 présente les résultats de cette étude de simulation.

Code R

```
plot(age,tension,type="p");abline(reg,col="red")
mu=reg$coef[1];beta=reg$coef[2];sig=summary(reg)[[6]]

# simuler 100 jeux de données
for (i in 1:100){
      y=mu+beta*age+rnorm(5,0,sig)
      reg2=lm(y~age)
      points(age,y,col="green")
      abline(reg2,col="cyan")
}

points(age,tension,type="p");abline(reg,col="red",lwd=2)

# prediction
x = as.data.frame(cbind(tension,age))
p1=predict(reg,x,interval="confidence",level=0.8,se.fit=TRUE)
```

```
p2=predict(reg,x,interval="prediction",level=0.8,se.fit=TRUE)

# intervalle de confiance
points( p1$fit[,2] ~ age, type='l', lty="dotted")
points( p1$fit[,3] ~ age,  type='l', lty="dotted")

# intervalle de prédiction
points( p2$fit[,2] ~ age, type='l', lty="dashed" )
points( p2$fit[,3] ~ age,  type='l', lty="dashed" )

legend("topleft", c("données observées","données simulées","régression avec
les observées","régression avec les simulées","Bande de confiance", "Bande
de prédiction"), lwd=c(1,1,2,1,1,1), lty=c(-1,-1,1,1,3,2),col=c("black",
"green","red","cyan","black","black"),pch=c(1,1,NA,NA,NA,NA),cex=0.8)
```

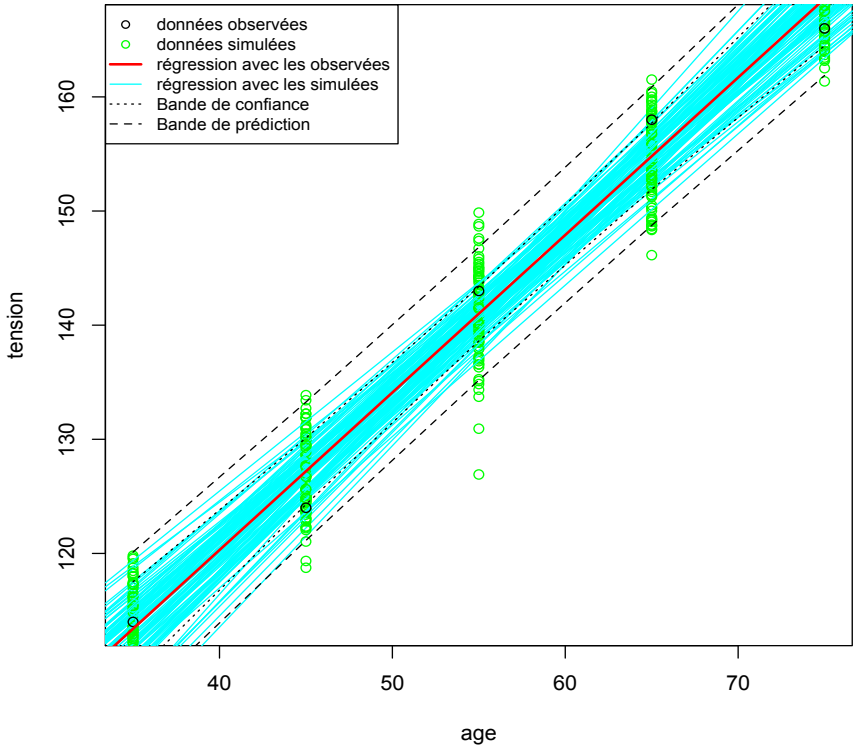

Figure 2.3 – Nuage de points, droites de régression et intervalles de confiance

2.3 Étude de cas : AirPassengers

Notons Y_i les nombres de passagers mensuels, t_i la variable de temps. On commence par les trois modèles suivants.

1. $Y_i = \mu + \beta t_i + \varepsilon_i$
2. $\log Y_i = \mu + \beta t_i + \varepsilon_i$
3. $\log Y_i = \mu + \beta_1 t_i + \beta_2 (t_i - \bar{t})^2 + \varepsilon_i$

Code R

```
data(AirPassengers); AP = AirPassengers; AP; plot(AP, ylab = "Passengers (1000's)")

# mod1
temps = time(AP); mod1=lm(AP~temps); summary(mod1)
shapiro.test(residuals(mod1))

par(mfcol=c(3,3))
plot(mod1,which=1)
plot(residuals(mod1),type="l");abline(h=0,col="red")
plot(mod1,which=2)

# mod2
mod2=lm(log(AP)~temps); summary(mod2); shapiro.test(residuals(mod2))

plot(mod2,which=1)
plot(residuals(mod2),type="l");abline(h=0,col="red")
plot(mod2,which=2)

# mod3
temps = time(AP); t=(temps-mean(temps))^2; mod3=lm(log(AP)~temps+t); summary(mod3)
shapiro.test(residuals(mod3))

plot(mod3,which=1)
plot(residuals(mod3),type="l");abline(h=0,col="red")
plot(mod3,which=2)
```

Les résultats de l'estimation et du test sont résumés dans le tableau suivant.

	`mod1`	`mod2`	`mod3`
Code en R	`AP~temps`	`log(AP)~temps`	`log(AP)~temps+(temps-mean(temps))^2`
Modèle	$Y_i = \mu + \beta t_i + \varepsilon_i$	$\log Y_i = \mu + \beta t_i + \varepsilon_i$	$\log Y_i = \mu + \beta_1 t_i + \beta_2 (t_i - \bar{t})^2 + \varepsilon_i$
Estimation	$\hat{\mu} = -62056, \hat{\beta} = 31.89$	$\hat{\mu} = -230, \hat{\beta} = 0.12$	$\hat{\mu} = -230, \hat{\beta}_1 = 0.12, \hat{\beta}_2 = -0.0032$
R^2	0.8536	0.9015(0.8535)	0.9074(0.8613)
$(\hat{Y}_i, \hat{\varepsilon}_i)$	non constante	parabolique	pas de remarque
QQ-plot	courbure évidente	presque aligné	presque aligné
Shapiro Test	10^{-5}	0.09	0.05

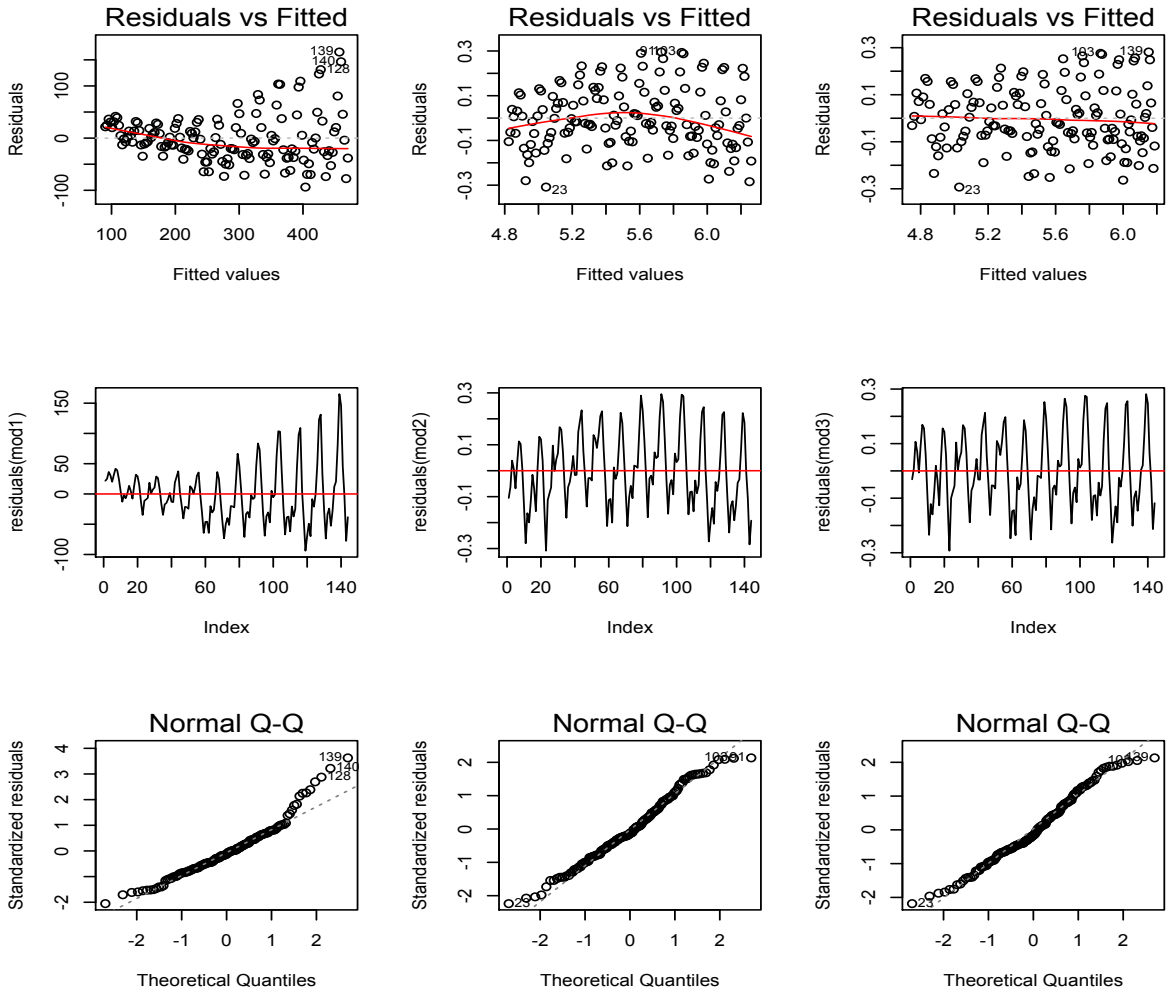

Figure 2.4 – Étude de résidus pour les trois modèles

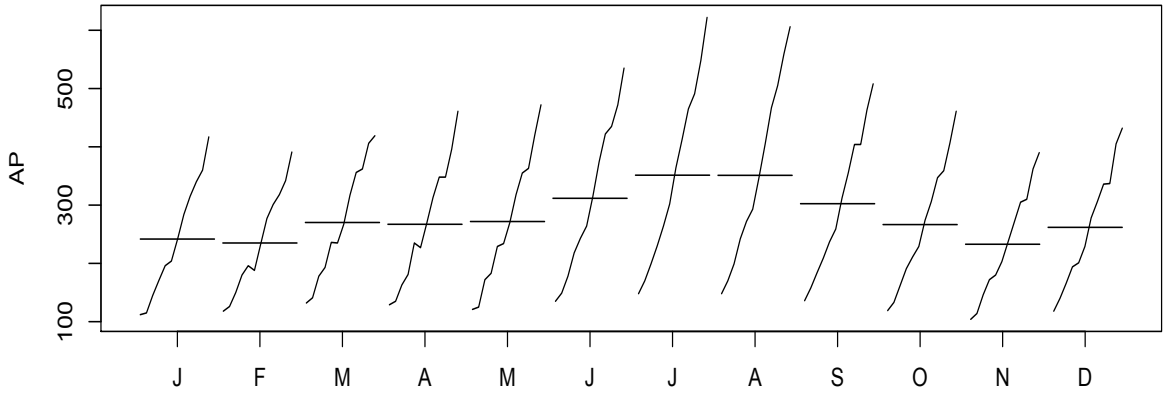

Figure 2.5 – Month plot du trafic

Pour comprendre la saisonnalité, nous examinons le month plot du trafic. La figure 2.5 montre un diagramme séquentiel pour chacune des saisons supposées de la série. S'il n'y avant pas d'effet saisonnier, ces 12 chronogrammes se ressembleraient. On note que les chronogrammes sont presque droits et parallèles. Juillet montre la plus forte activité. Janvier, février et novembre sont les mois de plus faible activité. D'après le month plot, on propose le modèle suivant,

$$\log Y_{i,j} = \mu_j + \beta_1 t_{i,j} + \beta_2 (t_{i,j} - \bar{t})^2 + \varepsilon_{i,j}, \quad i : \text{année}, \quad j : \text{mois}.$$

La forme matricielle du modèle est la suivante

$$\log \begin{pmatrix} Y_{1,1} \\ Y_{1,2} \\ \vdots \\ Y_{1,12} \\ \vdots \\ Y_{12,1} \\ Y_{12,2} \\ \vdots \\ Y_{12,12} \end{pmatrix} = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix} \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_{12} \end{pmatrix} + t\beta_1 + (t - \bar{t})^2 \beta_2 + \varepsilon.$$

2.4 Principales lois et propriété de vecteurs gaussiens utilisées en modèle linéaire

Dans le cadre du modèle linéaire, nous allons essentiellement utiliser des lois de probabilités suivantes.

Loi du χ^2 à n degrés de liberté
Soient $X_1, \ldots, X_n$ 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 $F(n_1, n_2)$.

Voici une propriété de vecteurs gaussiens que nous utilisons le plus.

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. $\sigma^2\chi^2(k_2)$).

Chapitre 3

Méthodes pour générer des données aléatoires

R peut générer des données aléatoires pour un grand nombre de fonctions de densité de probabilité. Les fonctions sont de la forme `rfunc(n, p1, p2, ...)` où `func` indique la loi de probabilité, `n` est le nombre de données générées et `p1, p2, ...` sont les valeurs des paramètres de la loi. Le tableau suivant donne la liste des lois de probabilité existantes en R.

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	

Code R

```
n = 10000; df = 10; r = rt(n, df)
par(mfrow=c(1,2)); hist(r, breaks=50, freq=F, main="df=10")
x = seq(-4, 4, 0.01); y = dt(x, df); lines(x, y, col="red")
```

```
df = 1; r = rt(n, df); h = hist(r[r>-4 & r<4], breaks=50, plot=F)
h$density = h$density*(pt(4,df)-pt(-4,df)); plot(h, freq=F, main="df=1")
y = dt(x, df); lines(x, y, col="red")
```

Pour simuler les échantillons des populations finies on peut utiliser la fonction `sample`.

Code R

```
sample(0:1, size=10, replace=T); sample(0:100, size=6, replace=F); sample(letters)
x = sample(1:3, size=100, replace=T, prob=c(0.2,0.3,0.5)); table(x)
```

3.1 Méthode de la transformée inverse

Soit F la fonction de répartition d'une variable aléatoire continue. Si la variable U a une loi uniforme sur $[0,1]$ alors la variable $F^{-1}(U)$ a comme loi F où F^{-1} est la fonction inverse de F, c.a.d. $F(F^{-1}(x)) = x$. En utilisant ce résultat, pour générer une donnée aléatoire de loi F, on génère d'abord une donnée de loi Unif$[0,1]$, ensuite on applique la fonction F^{-1} à la donnée simulée. Cette méthode est applicable seulement si la fonction inverse F^{-1} est calculable.

Exemple Loi exponentielle

Code R

```
n = 1000; r = rexp(n,1); x = seq(0,10,0.01); y = dexp(x,1)
par(mfrow=c(1,2))
hist(r,freq=F,main="rexp"); lines(x,y,col="red")
r = -log(runif(n))
hist(r,freq=F,main="inverse"); lines(x,y,col="red")
```

Exercice

En utilisant la méthode de la transformée inverse, générer les données aléatoires dont les densités de probabilité sont

1. $f(x) = 3x^2 \mathbb{1}_{(0,1)}(x)$;

2. $f(x) = x^{-2} \mathbb{1}_{[1,\infty)}(x)$.

3.2 Méthode de rejet

La méthode de rejet est utilisée pour engendrer indirectement une variable aléatoire X, de densité de probabilité f, lorsqu'on ne sait pas simuler directement la loi de densité

de probabilité f. Soit $g(x)$ la densité d'une variable aléatoire pour laquelle on connaît un algorithme de génération et qui, à une constante a près, majore $f(x)$ partout, c'est-à-dire

$$\exists a \in \mathbb{R}^+, \forall x, f(x) \leq a \times g(x).$$

Dans un premier temps, on tire un point (X, Y) distribué selon la loi $g(x)$ en abscisse et uniformément sur $[0, ag(X)]$ en ordonnée. Si $Y \leq f(X)$, on garde X, sinon on tire un autre point. On réitère jusqu'à ce que la condition $Y \leq f(X)$ soit remplie. L'algorithme consiste en trois étapes.

1. Boucler
   — Tirer X de densité $g(x)$;
   — Tirer U selon la loi uniforme $U[0, ag(X)]$, indépendamment de X ;
2. Tant que $f(X) < U$, reprendre en 1 ;
3. Accepter X comme un tirage aléatoire de densité de probabilité f.

Exemple Loi de Cauchy
On veut simuler les données de loi dont la densité est définie ci dessous

$$f(x) = \frac{1}{\pi(1+x^2)}.$$

La densité de l'enveloppe $g(x)$ est la suivante.

$$g(x) = \begin{cases} \dfrac{1}{\pi x^2}, & |x| > \frac{4}{\pi}, \\ \dfrac{\pi}{16}, & |x| \leq \frac{4}{\pi}. \end{cases}$$

Il est facile de vérifier que $f(x) \leq 16/\pi^2 \times g(x)$. On en déduit la fonction de répartition de la loi de densité $g(x)$ est

$$F_g(x) = \begin{cases} -\dfrac{1}{\pi x}, & x < -\dfrac{4}{\pi}, \\ \dfrac{1}{2} + \dfrac{\pi}{16}x, & |x| \leq \dfrac{4}{\pi}, \\ 1 - \dfrac{1}{\pi x}, & x > \dfrac{4}{\pi}. \end{cases}$$

Code R

```
qg = function(x){if(x < 1/4){y = -1/pi/x}else if
                ((x >= 1/4) & (x <= 3/4)){y = 16*(x-1/2)/pi}else
                {y = 1/pi/(1-x)}
                return(y)}
dg = function(x){if(abs(x) <= 4/pi){y = pi/16}else{y = 1/pi/x/x}
                return(y)}
n = 10000; r = rep(0,n); i = 1
```

```
while (i <= n){k = 1; y = 2; z=1;
               while (y > z & k < 10^6){u1 = runif(1); u2 = runif(1);
                                        r[i] = qg(u1)
                                        y = 16/pi/pi*dg(r[i])*u2
                                        z = 1/pi/(1+r[i]*r[i])
                                        k = k+1}
               if (k < 10^6){i=i+1}}
probf = function(x){if(x < -4/pi){y = -1/pi/x}else if
                    (abs(x) <= 4/pi){y = 1/2+pi*x/16}else
                    {y = 1-1/pi/x}
                    return(y)}

x = seq(-4, 4, 0.01); y = 1/pi/(1+x*x); h = hist(r[r>-4 & r<4], breaks=50, plot=F)
h$density = h$density*(probf(4)-probf(-4)); plot(h, freq=F); lines(x, y, col="red")
```

La figure 3.1 présente l'histogramme de 10000 données simulées de loi de Cauchy et la densité réelle.

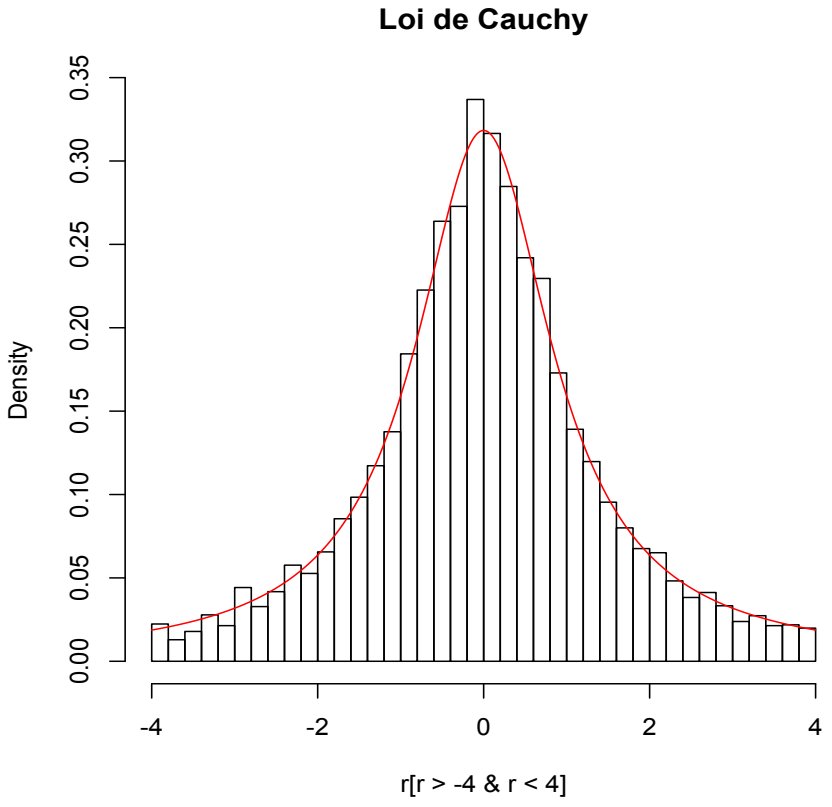

Figure 3.1 – Histogramme de données simulées et la densité réelle

Exercice

En utilisant la méthode de rejet, générer les données aléatoires dont la densité de probabilité est

$$f(x) = \frac{C}{1+|x|^{\alpha+1}},$$

où $C = \frac{1}{2D}$ et

$$D = \int_0^\infty \frac{1}{1+x^{\alpha+1}} dx.$$

La densité de l'enveloppe $g(x)$ est la suivante.

$$g(x) = \begin{cases} \dfrac{C}{|x|^{\alpha+1}}, & |x| > s \\ K, & |x| \leq s, \end{cases}$$

où

$$K = C\left(\frac{D\alpha}{1+\alpha}\right)^{\frac{\alpha+1}{\alpha}}, \quad s = \frac{1}{2C}\left(\frac{1+\alpha}{D^{\alpha+1}\alpha}\right)^{\frac{1}{\alpha}}.$$

La constante de majoration $a = C/K$. Les variables de loi $g(x)$ peuvent être simulées par la méthode de la transformée inverse.

La figure 3.2 présente les fonctions de densité de la loi de Cauchy généralisée définie dans l'exercice précédent avec α différent.

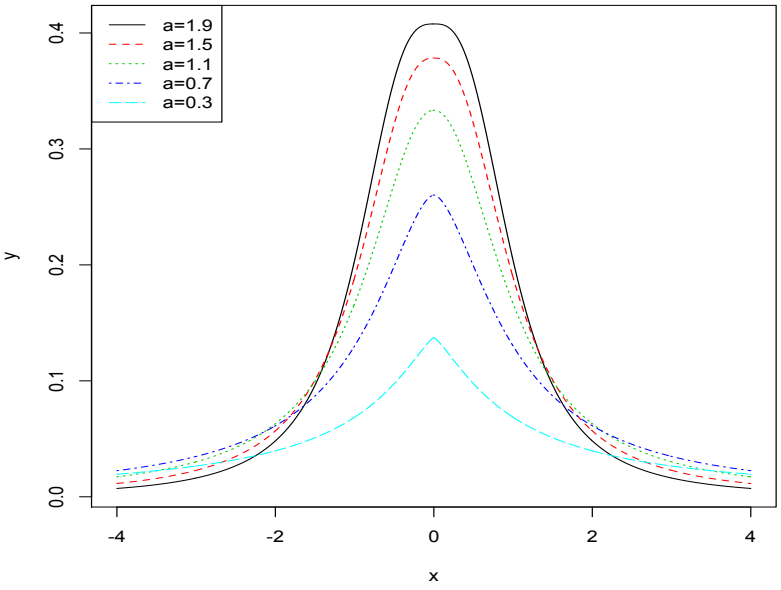

Figure 3.2 – Densité de la loi de Cauchy généralisée avec α différent

Bibliographie

[1] Jean-Marc Azaïs et Jean-Marc Bardet, *Le modèle linéaire par l'exemple : Régression, Analyse de la variance et Plans d'expérience illustrés avec R, SAS et Splus*. Dunod, 2006.