STATISTIQUE INFERENTIELLE

M. BOUTAHAR

Definition 1

On appelle **modèle statistique** une famille d'espaces de probabilités $(E, \mathcal{E}, P_\theta, \theta \in \Theta)$ où Θ est l'espace des paramètres.

Dans la pratique, $E = \mathbb{R}$ ou $E = \mathbb{R}^p$, on notera X l'application $(\Omega, \mathcal{F}, P) \to (E, \mathcal{E}, P_\theta)$. En général $X = (X_1, ..., X_n)$ et $X(\omega), \omega \in \Omega$, est l'observation ou le résultat de l'expérience, P_θ est la loi inconnue de X, c'est à dire que P_θ est la mesure image de P par l'application X .

Definition 2

Soit Q_θ une loi de probabilité sur un espace $(E_1, \mathcal{E}_1)$. On appelle **échantillon** de taille n de loi Q_θ tout vecteur $X = (X_1, ..., X_n)$, où les $(X_i)_{1 \leq i \leq n}$ sont i.i.d. de loi Q_θ.

Le modèle statistique correspondant est $E = E_1^n, \mathcal{E} = \mathcal{E}_1^{\otimes\, n}$, $P_\theta = Q_\theta^n$.

On dit que le modèle est **paramétrique** s'il existe un entier k, tel que $\Theta \subset \mathbb{R}^k$, non paramétrique dans le cas contraire.

Definition 3

Le modèle statistique $(E, \mathcal{E}, P_\theta, \theta \in \Theta)$ est dit **dominé** s'il existe une mesure $\sigma-$finie μ qui domine $P_\theta, \forall \theta$, la mesure μ est appelée mesure dominante, et il existe alors des densités $f_\theta, \theta \in \Theta$ telle que

$$P_\theta = f_\theta.\mu, \text{ i.e. } P_\theta(A) = \int_A f_\theta(x)\mu(dx), \forall A \in \mathcal{E}. \quad (1)$$

Remarque. On dit que μ domine P_θ (ou encore P_θ est absolument continue par rapport à μ) si $\forall A \in \mathcal{E}, \mu(A) = 0 \implies P_\theta(A) = 0$, l'égalité (1) n'est autre que le théorème de *Radon Nikodym*.

Contenu

Definition 4

On appelle **statistique**, qu'on note $T(X)$ ou T, toute fonction mesurable de l'observation X, à valeurs dans $\mathbb{R}^d$, cela veut dire aussi tout vecteur aléatoire défini sur $(E, \mathcal{E})$.

Definition 5

Pour un modèle statistique $(E, \mathcal{E}, P_\theta, \theta \in \Theta)$ dominé par μ, une statistique T est dite **exhaustive** (ou **suffisante**) si l'on peut trouver une version indépendante de θ , définie $\mu.p.s.$, de $E_\theta(\mathbf{1}_A \mid T)$ $\forall A \in \mathcal{E}$, on note alors $E(\mathbf{1}_A \mid T)$ cette version.

Dans ce cas, il existe une version indépendante de θ de $E_\theta(Y \mid T)$ noté $E(Y \mid T), \forall Y \geq 0$ ou P_θ-intégrable $\forall \theta$.

Contenu

Definition 6

Soit $(E, \mathcal{E}, \mu)$ un espace mesuré, avec μ $\sigma - finie$,
$T : E \to \mathbb{R}^d$ une application mesurable, $\Theta \subset \mathbb{R}^d$ et
$\rho : E \to \mathbb{R}^d$ une application injective telle que:
i) $\overline{\overset{\circ}{\rho(\Theta)}} \neq \emptyset$, (l'intérieur de $\rho(\Theta)$ est non vide)
ii) $\Phi(\theta) = \log(\int_E exp(< \rho(\theta), T(x) >) d\mu(x)) < \infty, \forall \theta \in \Theta$,
alors le modèle statistique $(E, \mathcal{E}, P_\theta, \theta \in \Theta)$ dominé par μ, et où

$$dP_\theta(x) = exp(< \rho(\theta), T(x) > - \Phi(\theta)) d\mu(x),$$

est appelé modèle **exponentiel** , et $(P_\theta, \theta \in \Theta)$ est une famille exponentielle de probabilités.

Remarque. Si $\rho(\theta) = \theta$, on dira que le modèle est canonique.

Theorem 7

Soit $(E, \mathcal{E}, P_\theta, \theta \in \Theta)$ *un modèle exponentiel canonique, avec* $dP_\theta(x) = exp(<\theta, T(x)> -\Phi(\theta))d\mu(x)$, *alors* Φ *est de classe* C^∞ *sur* $\overset{\circ}{\Theta}$, *et* $\forall \theta \in \Theta$, *le moment de tout ordre de* T *est fini :* $E(||T||^n) < \infty, \forall n \in \mathbb{N}$, *de plus*

$$E_\theta(T) = \nabla\Phi(\theta), \mathrm{cov}\ _\theta(T) = (\frac{\partial^2}{\partial\theta_i\partial\theta_j}\Phi(\theta))_{1\leq i,j\leq d}, \qquad (2)$$

La démonstration de ce théorème repose sur le théorème suivant, et ce dernier est une conséquence directe du lemme 10 ci-dessous.

Soit $X = {}^t(X_1, ..., X_d)$ un v.a. de dimension d défini sur

$(\Omega, \mathcal{F}, P)$, on définit la transformée de Laplace de X par

$$\Psi(s) = E(e^{<s,X>}) = \int_{\mathbb{R}^d} e^{<s,x>} dP_X(x)$$

où P_X est la mesure image de P par X.

Theorem 8

On suppose l'hypothèse suivante :

$$\exists r > 0 \text{ tel que } \Psi(s) < \infty, \forall \, ||s|| \leq r, \tag{3}$$

alors Ψ *est de classe* C^∞ *sur un voisinage de 0, les v.a.r.* $X_i, i = 1, ..., d$ *possèdent des moments finis de tous ordre, de plus*

$$E(X) = \nabla\Psi(0), E(X_i X_j) = \frac{\partial^2}{\partial s_i \partial s_j}\Psi(0). \tag{4}$$

Lemme 9

Sous l'hypothèse (3) on a:
i) $E(|< s, X >|^k) < \infty, \forall ||s|| \leq r, \forall k \in \mathbb{N},$
ii) $\Psi(s) = \sum_0^\infty \frac{E((<s,X>)^k)}{k!}, \forall ||s|| \leq r.$

Démonstration

$$e^{|<s,X>|} \leq e^{<s,X>} + e^{-<s,X>},$$

donc l'hypothèse (3) implique que $E(e^{|<s,X>|}) < \infty$, si $||s|| \leq r$. La preuve du lemme s'obtient alors en utilisant les développements en séries de $e^{<s,X>}$ et $e^{-<s,X>}$ et le théorème de convergence dominée de Lebesgue . $\square$

En utilisant encore une fois le théorème de convergence dominée de Lebesgue et le i) du lemme ci-dessus on montre le lemme suivant:

Lemme 10

On suppose l'hypothèse (3), alors Ψ *admet des dérivées de tous ordres en tout* $s \in \mathbb{R}^d$, *et* $||s|| \leq r$, *et on a :*
$\forall k = (k_1, ..., k_d) \in \mathbb{N}^d$, *en notant* $|k| = \sum_1^d k_i$,

$$\frac{\partial^{|k|}\Psi(s)}{\partial^{k_1} s_1 \ldots \partial^{k_d} s_d} |_{s=0} = E\left(X_1^{k_1} \ldots X_d^{k_d}\right).$$

Démonstration du théorème 7

$$E_\theta\left(e^{<s,T>}\right) = e^{\Phi(s+\theta)-\Phi(\theta)};$$

donc cette quantité est finie dès que $\theta \in \overset{\circ}{\Theta}$ et que $||s||$ est très petit, par ailleurs si on note $T = {}^t(T_1, ..., T_d)$:

$$\begin{aligned} E_\theta(T_k) &= \frac{\partial}{\partial s_k} e^{\Phi(s+\theta)-\Phi(\theta)} \big|_{s=0} \\ &= \frac{\partial \Phi(\theta)}{\partial \theta_k}, \end{aligned} \tag{5}$$

$$\begin{aligned} E_\theta(T_iT_j) &= \frac{\partial^2}{\partial s_i \partial s_j} e^{\Phi(s+\theta)-\Phi(\theta)}\,\Big|_{s=0} \\ &= \frac{\partial \Phi(\theta)}{\partial \theta_i}\frac{\partial \Phi(\theta)}{\partial \theta_j} + \frac{\partial^2 \Phi(\theta)}{\partial \theta_i \partial \theta_j}. \end{aligned} \tag{6}$$

On déduit alors le résultat (2) en appliquant (5)-(6). □

Soit $(E, \mathcal{F}, P_\theta, \theta \in \Theta)$ un modèle statistique. Supposons que l'on veut estimer une fonction de θ, soit $g(\theta)$ avec $g : \Theta \to \mathbb{R}$, en se basant sur la suite $(X_i)_{1<i<n}$. Posons $X = {}^t(X_1, ..., X_n)$.

Contenu

Definition 11

On appelle **Estimateur** de $g(\theta)$ toute statistique $T(X)$.

Definition 12

On définit le biais et l'erreur quadratique de l'estimateur $T(X)$ respectivement par

$$Biais_T(\theta) = E_\theta(T(X)) - g(\theta)$$

et

$$EQ_T(\theta) = E_\theta(|T(X) - g(\theta)|^2).$$

Remarque.

Ces deux quantités permettent de mesurer la qualité de l'estimateur $T(X)$; on cherche souvent à construire des estimateurs ayant un biais nulle (on dit alors que l'estimateur est sans biais) et une erreur quadratique la plus petite possible mais généralement non nulle (voir plus loin). On utilise aussi comme critère de qualité de l'estimateur sa variance en utilisant la relation :

$$EQ_T(\theta) = var_\theta(T(X)) + Biais_T^2(\theta).$$

Definition 13

Soient $T_1(X)$ et $T_2(X)$ deux estimateurs de $g(\theta)$, si $EQ_{T_1(X)}(\theta) \leq EQ_{T_2(X)}(\theta), \forall\theta \in \Theta$, on dit alors que $T_1(X)$ est **préférable** à $T_2(X)$, si l'inégalité est stricte, on dira que $T_1(X)$ est **strictement préférable** à $T_2(X)$.

Definition 14

On dit que l'estimateur $T(X)$ est **admissible** s'il n'existe pas d'estimateur strictement préférable à $T(X)$.

Theorem 15

(Rao-Blackwell) . Soit $(E, \mathcal{E}, P_\theta, \theta \in \Theta)$ un modèle statistique dominé par μ, et T une statistique exhaustive.
Si U est un estimateur de $g(\theta)$ et U est de carré intégrable, $(E_\theta(U^2) < \infty), \forall \theta \in \Theta$; alors $E(U \mid T)$ est un estimateur strictement préférable à U, sauf si U est presque sûrement égal à une fonction de T.

Démonstration

Posons $\widehat{U} = E(U \mid T)$, d'après les propriétés de l'espérance conditionnelle on a :

$$E(\widehat{U}) = E(U) \text{ et } \widehat{U} - E(\widehat{U}) \text{ est } \sigma(T)\text{-mesurable },$$

donc

$$Biais_{\widehat{U}}(\theta) = Biais_U(\theta) \text{ et } E\left((\widehat{U} - U)(\widehat{U} - E(\widehat{U}))\right) = 0; \quad (7)$$

$$\begin{aligned}
\operatorname{var}(U) &= E\left((U - EU)^2\right) \\
&= E\left((U - E\widehat{U})^2\right) \\
&= E\left((U - \widehat{U} + \widehat{U} - E\widehat{U})^2\right) \\
&= E\left((U - \widehat{U})^2\right) + E\left((\widehat{U} - E\widehat{U})^2\right) \\
&+ 2E\left((U - \widehat{U})(\widehat{U} - E\widehat{U})\right) \\
&= \operatorname{var}(\widehat{U}) + E\left((U - \widehat{U})^2\right) \text{ d'après (7)}
\end{aligned}$$

donc

$$EQ_U(\theta) = EQ_{\widehat{U}}(\theta) + E\left((U - \widehat{U})^2\right),$$

par conséquent

$$EQ_{\widehat{U}}(\theta) < EQ_U(\theta) \text{ sauf si } \ \widehat{U} = U \text{ p.s.}\square$$

Definition 16

On dit que l'estimateur $T(X)$ de $g(\theta)$ est un estimateur **uniformément de variance minimum sans biais (U.V.M.B.)** si $T(X)$ est sans biais et pour tout estimateur $U(X)$ sans biais de $g(\theta)$ tel que $\mathrm{var}_{\theta}(U(X)) < \infty, \forall\theta \in \Theta$, on a

$$\mathrm{var}_{\theta}(T(X)) \leq \mathrm{var}_{\theta}(U(X)), \forall\theta \in \Theta.$$

Definition 17

Soit $(E, \mathcal{E}, P_\theta, \theta \in \Theta)$ un modèle statistique dominé. Une statistique à valeurs dans $(E, \mathcal{E})$ est dite **complète** (resp. **b-complète**) si toute v.a.r. Y définie sur $(E, \mathcal{E})$ telle que $Y(T)$ soit P_θ-intégrable $\forall\theta \in \Theta$ (resp. bornée) et P_θ-intégrale nulle $\forall\theta \in \Theta$, vérifie $Y(T) = 0, P_\theta.p.s$. $\forall\theta \in \Theta$.

Theorem 18

*(Lehmann-Schéffé) Si T est une statistique exhaustive et complète , et U est un estimateur sans biais de $g(\theta)$ tel que $E_\theta(U^2) < \infty, \forall\theta \in \Theta$, alors $\widehat{U} = E(U \mid T)$ est l'unique estimateur **U.V.M.B.** de $g(\theta)$.*

Démonstration

Un utilisant le théorème de Rao-Blackwell , $\widehat{U}$ est toujours préférable à U, et comme U est sans biais, $\widehat{U}$ est sans biais aussi et donc $\operatorname{var}_{\theta}(\widehat{U}) \leq \operatorname{var}_{\theta}(U), \forall \theta \in \Theta$. Il suffit alors de montrer que $\widehat{U}$ ne dépend pas de U.
Soit $\widehat{U}_1 = g_1(T), \widehat{U}_2 = g_2(T)$, deux estimateurs sans biais de $g(\theta)$ et de carré intégrable, donc

$$E_{\theta}(g_1(T) - g_2(T)) = 0, \forall \theta \in \Theta;$$

comme T est complète alors
$g_1(T) - g_2(T) = 0 \;\; P_{\theta}.p.s \;\; \forall \theta \in \Theta$ et par conséquent
$\widehat{U}_1 = \widehat{U}_2 \;\; P_{\theta}.p.s. \forall \theta \in \Theta. \square$

Remarque. Le problème qui se pose au statisticien est de trouver une statistique T exhaustive et complète, une fois c'est le cas on peut construire de deux manières un estimateur **U.V.M.B.** de $g(\theta)$:

- On construit un estimateur sans biais de $g(\theta)$ de la forme $g(T)$ c'est alors un estimateur **U.V.M.B.** de $g(\theta)$,
- On construit un estimateur sans biais de carré intégrable U de $g(\theta)$, alors $E(U \mid T)$ est un estimateur **U.V.M.B.** de $g(\theta)$.

Theorem 19

Soit $(E, \mathcal{E}, P_\theta, \theta \in \Theta)$ *un modèle exponentiel :* $dP_\theta(x) = exp(< \rho(\theta), T(x) > -\Phi(\theta))d\mu(x)$, *alors* T *est une statistique exhaustive complète.*

Contenu

Un modèle statistique est dit linéaire (ou modèle de régression linéaire) si les variables X_t, $t = 1, ..., n$ qu'on observe vérifient l'équation de régression suivante:

$$X_t = \theta_1 Z_{t,1} + ... + \theta_k Z_{t,p} + u_t, t = 1, ..., n. \tag{8}$$

- Les $Z_{t,k}$ sont les variables explicatives et sont connues,
- $\theta = {}^t(\theta_1, ..., \theta_p)$ est le vecteur des paramètres inconnus,
- u_t est un bruit blanc c'est à dire une suite de variables

aléatoires réelles telles que:

- **(H.1)**: $E(u_t) = 0, \ 1 \leq t \leq n,$

- **(H.2)**: $E(u_t u_s) = \begin{cases} \sigma^2 & \text{si } t = s \\ 0 & \text{si } t \neq s \end{cases}$, σ^2 inconnue.

Remarques .

1. Souvent on suppose que le bruit blanc (u_t) est gaussien, (cette hypothèse est très utile lorsqu' on veut estimer le paramètre θ par la méthode du maximum de vraisemblance).
2. L'hypothèse que (u_t) est un bruit blanc est très restrictive, souvent on est amené à supposer des hypothèses moins fortes.

On peut écrire le modèle de régression (8) sous la forme vectorielle suivante:

$$X = \mathbf{Z}\theta + \sigma U, \tag{9}$$

Z est une matrice de taille(n,p) connue, U est un v.a. centré réduit. $\theta = {}^t(\theta_1, ..., \theta_k) \in \mathbb{R}^p$, et $\sigma \in \mathbb{R}_+^*$ sont inconnus.

- On suppose que $rang(Z) = p$ (cette hypothèse traduit le fait que les variables explicatives $Z_{t,j}$ sont linéairement indépendantes et donc il n y'a pas de redondance dans le modèle de régression (8)).
- Le modèle statistique associé est $(E, \mathcal{E}, P_{\theta,\sigma}, (\theta, \sigma) \in \mathbb{R}^k \times \mathbb{R}_+^*)$, $(E, \mathcal{E}) = (\mathbb{R}^n, \mathcal{B}_n)$ et $P_{\theta,\sigma}$ est la mesure image de la loi de U par l'application $u \to \mathbf{Z}\theta + \sigma u$.

Contenu

On considère le modèle de régression (9), l'estimateur des moindres carrés $\widehat{\theta}$ de θ est solution du problème de minimisation suivante:

$$\widehat{\theta} = \arg\min_{\theta \in \Theta} ||X - \mathbf{Z}\theta||^2 ,$$

et dont la solution est donnée par

$$\widehat{\theta} = ({}^t\mathbf{Z}\mathbf{Z})^{-1} \, {}^t\mathbf{Z}X \tag{10}$$

Theorem 20

i. L'estimateur des moindres carrés est un estimateur sans biais de θ, qui est de variance minimale dans la classe des estimateurs linéaires en X (de la forme BX) sans biais, i.e. pour tout $\alpha \in \mathbb{R}^p$, $< \alpha, \widehat{\theta} >$ est un estimateur sans biais de $< \alpha, \theta >$, de variance inférieure à celle de tout estimateur sans biais de la forme $< \beta, X >$ de $< \alpha, \theta >$.
ii.

$$\widehat{\sigma}^2 = \frac{\left\|X - \widehat{X}\right\|^2}{n - p} \tag{11}$$

est un estimateur sans biais de la variance inconnue σ^2, avec $\widehat{X} = \mathbf{Z}\widehat{\theta}$

Remarque. $\widetilde{X} = X - \widehat{X}$ est le vecteur des résidus de la régression linéaire.

Démonstration

i) Vu que $E(X) = \mathbf{Z}\theta$, alors
$E(\widehat{\theta}) = ({}^t\mathbf{Z}\mathbf{Z})^{-1}\,{}^t\mathbf{Z}E(X) = ({}^t\mathbf{Z}\mathbf{Z})^{-1}\,{}^t\mathbf{Z}\mathbf{Z}\theta = \theta$, donc $\widehat{\theta}$ est sans biais.

Soit $< \beta, X >$ un estimateur sans biais de $< \alpha, \theta >$, alors $< \beta, \mathbf{Z}\theta > = < \alpha, \theta >, \forall \theta \in \mathbb{R}^k$, ce qui implique que ${}^t\mathbf{Z}\beta = \alpha$.

$$\mathrm{var}_{\theta,\sigma}(< \beta, X >) = E(< \beta, U >^2) = \sigma^2 \|\beta\|^2,$$

$$\mathrm{var}_{\theta,\sigma}(<\alpha,\widehat{\theta}>) = \sigma^2 \left\|\mathbf{Z}({}^t\mathbf{Z}\mathbf{Z})^{-1}\alpha\right\|^2 .$$

Mais ${}^t\mathbf{Z}\beta = \alpha$ implique que $\mathbf{Z}({}^t\mathbf{Z}\mathbf{Z})^{-1}\alpha = \mathbf{Z}({}^t\mathbf{Z}\mathbf{Z})^{-1}\,{}^t\mathbf{Z}\beta$. Posons $P = \mathbf{Z}({}^t\mathbf{Z}\mathbf{Z})^{-1}\,{}^t\mathbf{Z}$, on a $P^2 = P$ et ${}^tP = P$ donc P est une projection orthogonale sur $Im(P)$ donc $\|P\|^2 \leq 1$, et par conséquent

$$\sigma^2 \left\|\mathbf{Z}({}^t\mathbf{Z}\mathbf{Z})^{-1}\alpha\right\|^2 \leq \sigma^2 \left\|\beta\right\|^2 .$$

ii) $P\mathbf{Z} = \mathbf{Z}$ ce qui implique que $Im(\mathbf{Z}) \subset Im(P)$, mais $\dim(Im(P)) = rang(P) \leq rang(\mathbf{Z}) = \dim(Im(\mathbf{Z}))$ donc $Im(\mathbf{Z}) = Im(P)$, donc $(\mathbf{I} - P)U =$ la projection orthogonale de

U sur un sous espace vectoriel de dimension $n-p$ par conséquent

$$E_{\theta,\sigma}(\left|\left|X-\widehat{X}\right|\right|^2) = \sigma^2 E(||(\mathbf{I}-P)U||^2) = (n-p)\sigma^2.\square.$$

Corollaire 21

Si le modèle est gaussien, i.e. que le v.a. U est gaussien, alors $\forall\alpha \in \mathbb{R}^d$, $<\alpha, \widehat{\theta}>$ est un estimateur ***U.V.M.B.*** *de $<\alpha,\theta>$ et $\widehat{\sigma}^2$ est aussi un estimateur* ***U.V.M.B.*** *de σ^2.*

Démonstration. $U \rightsquigarrow N(0,\sigma^2\mathbf{I})$ donc

$$dP_{\theta,\sigma^2}(x) = (2\pi\sigma^2)^{-n/2} exp(-\frac{1}{2\sigma^2}\left|\left|x-\mathbf{Z}\theta\right|\right|^2)dx,$$

que l'on peut écrire

$$dP_{\theta,\sigma^2}(x) = exp(< \rho(\theta, \sigma^2), T(x) > -\Phi(\theta))d\mu(dx),$$

avec

$$\rho(\theta, \sigma^2) = \begin{pmatrix} \theta/\sigma^2 \\ -1/(2\sigma^2) \end{pmatrix}, T(x) = \begin{pmatrix} {}^tZx \\ ||x||^2 \end{pmatrix},$$

donc, et tenant compte du théorème 19, on déduit que T est une statistique exhaustive et complète. Par conséquent d'après le théorème de Lehmann-Scheffé 18, il suffit alors de montrer que les deux estimateurs $< \alpha, \widehat{\theta} >$ et $\widehat{\sigma}^2$ sont des fonctions de T.

Posons $T(X) = \begin{pmatrix} T_1 \\ T_2 \end{pmatrix}$, avec $T_1 = {}^t\mathbf{Z}X$, $\quad T_2 = ||X||^2$, on a alors

$$\begin{aligned}
\widehat{\sigma}^2 &= \frac{1}{n-p}(\left|\left|X - \widehat{X}\right|\right|^2) \\
&= \frac{1}{n-p}(||X||^2 - \left|\left|\widehat{X}\right|\right|^2) \\
&= \frac{1}{n-p}(T_2 - \left|\left|\mathbf{Z}({}^t\mathbf{Z}\mathbf{Z})^{-1}T_1\right|\right|^2).
\end{aligned}$$

et $\widehat{\theta} = ({}^t\mathbf{Z}\mathbf{Z})^{-1}T_1$.

Theorem 22

i) $E(\widehat{\theta}) = \theta$ *(estimateur sans biais)*

ii) $Var(\widehat{\theta}) = \sigma^2 \, ({}^t\mathbf{Z}\mathbf{Z})^{-1}$ *(matrice de covariance).*

- Hypothèse de normalité (conséquence)

En plus des hypothèses **(H.1)** et **(H.2)** on ajoute une hypothèse **(H.3)** de normalité du vecteur aléatoire u :

$$u \sim N(0, \sigma^2 \mathbf{I}_n).$$

Theorem 23

i) $\widehat{\theta} \sim N\left(\theta, \sigma^2 \; ({}^t\mathbf{Z}\mathbf{Z})^{-1}\right)$,

ii) $\frac{(n-p)\widehat{\sigma}^2}{\sigma^2} \sim \chi^2(n-p)$,

iii) $\widehat{\theta}$ *et* $\widehat{\sigma}^2$ *sont des statistiques indépendantes.*

Corollary 24

$$\frac{\widehat{\theta}_j - \theta_j}{\sqrt{\widehat{\sigma}^2 a_{jj}}} \sim t(n-p), \;\; 1 \leq j \leq p;$$

a_{jj} *est le* j^{ieme} *élément diagonal de* $({}^t\mathbf{Z}\mathbf{Z})^{-1}$.

- **Construction des tests d'hypothèses au seuil α:**

$$H_0 : \theta_j = 0 \quad \overrightarrow{\text{région critique}} \quad \frac{\left|\widehat{\theta}_j\right|}{\sqrt{\widehat{\sigma}^2 a_{jj}}} \geq t_{1-\alpha/2}(n-p) \quad (12)$$

$$H_0 : \theta_j \leq 0 \quad \overrightarrow{\text{région critique}} \quad \frac{\widehat{\theta}_j}{\sqrt{\widehat{\sigma}^2 a_{jj}}} \geq t_{1-\alpha}(n-p),$$

où $t_{1-\alpha}(n-p)$ désigne le quantile d'ordre $1-\alpha$ de la loi de Student à $(n-p)$ degrés de liberté.

- **Construction de région de confiance (au niveau α) pour $\theta_j, (1 \leq j \leq p)$:**

$$\widehat{\theta}_j \pm \sqrt{\widehat{\sigma}^2 a_{jj}} \;\; t_{1-\alpha/2}(n-p)$$

- Prévision

Comment, à l'instant n, prévoir la réalisation de X_{n+1} disposant de $Z_{n+1} = {}^t(Z_{n+1,1}, \ldots, Z_{n+1,p})$, c'est à dire d'une réalisation supplémentaire de chacune des p variables explicatives. Les hypothèses sont **(H.1)**, **(H.2)** et

$$\begin{cases} X_{n+1} = {}^tZ_{n+1}\theta + u_{n+1}, \\ E(u_{n+1}) = 0, \\ E(u_{n+1}u_t') = 0, \quad 1 \leq t \leq n. \end{cases}$$

On a deux incertitudes:

i) Celle due à u_{n+1} (perturbation aléatoire),

ii) Celle due au fait que θ est inconnu.

Prévision optimale

$$\widehat{X}_{n+1} = {}^t Z_{n+1}\widehat{\theta}.$$

L'intervalle de confiance au niveau $(1-\alpha)$ est donné par

$$\widehat{X}_{n+1} \pm \widehat{\sigma} t_{1-\alpha/2}(n-p)\sqrt{1 + {}^t Z_{n+1}\left({}^t\mathbf{ZZ}\right)^{-1} Z_{n+1}}.$$

- Exemples de modélisation. 1. Modèle linéaire simple.

Commandes sur R

data(cars), plot(cars)

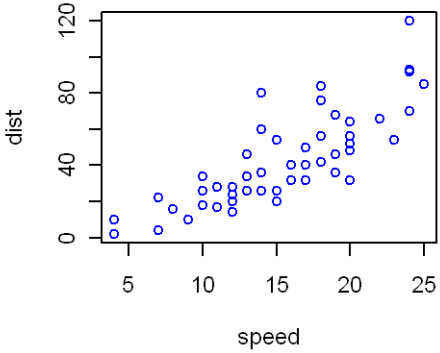


Figure: Distance de freinage

- On cherche la relation qui lie la variable dist (distance de freinage) et la variable speed (vitesse).
- On propose le modèle $dist_t = \theta_1 + \theta_2 speed_t + u_t$,

Commandes sur R

fit < − lm(cars$dist ~ cars$speed), summary(fit)

Call:lm(formula = cars$dist ~cars$speed)

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	*
cars$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: 1.490e-12

- Interprétation des résultats:

Residuals: En premier lieu R affiche les statistiques descriptives des résidus de la régression.

Coefficients: Les estimations $\widehat{\theta}_1$ =-17.5791, $\widehat{\theta}_2$ =3.9324.

Std. Error désigne l'écart type estimé.

t. value est la statistique de Student.

$Pr(>|t|)$ est la p-value associé à la statistique de Student t value, une valeur plus petite que 0.01 nous conduit au rejet de H_0 c'est à dire que le paramètre est significatif. Pour notre modèle le paramètre θ_1 n'est pas significatif alors que θ_2 l'est.

Residual standard error $= \sqrt{\frac{||\widehat{u}||^2}{n-p}}$.

Multiple R-Squared:

Definition 25

Lorsqu'il y'a une constante dans le modèle de régression multiple, on appelle coefficient de détermination le scalaire

$$R^2 = \frac{\left|\left|\widehat{X} - \overline{X}\delta_n\right|\right|^2}{\left|\left|X - \overline{X}\delta_n\right|\right|^2}$$

où $\overline{X} = \frac{1}{n}\sum_{t=1}^{n} X_t$ et $\delta_n = {}^t(1, ..., 1)$, vecteur $(n, 1)$.

Plus R^2 est proche de 1, plus l'ajustement est meilleur.

Adjusted R-squared: $R_a^2 = 1 - (1 - R^2)\frac{n-1}{n-p-1}$, où p est le nombre de paramètres sans compter la constante. R_a^2 ne croit que si la nouvelle variable explicative ajoutée améliore l'ajustement, elle peut être négative, et $R_a^2 \leq R^2$.

F-statistic désigne la statistique de Fisher qui correspond à l'hypothèse nulle $H_0 : \theta_1 = ... = \theta_p = 0$; elle est donnée par $F = \frac{||\widetilde{u}||^2 - ||\widehat{u}||^2}{||\widehat{u}||^2}$ avec $\widetilde{u}$: résidus sous H_0 et $\widehat{u}$: résidus sous H_1.
On a sous H_0 la statistique F suit une loi de Fisher $F(p, n-p)$.

Commandes sur R

fit1 <– lm(cars$dist ~cars$speed-1), summary(fit1)

Call: lm(formula = cars$dist ~cars$speed -1)

Residuals:

Min	*1Q*	*Median*	*3Q*	*Max*
-26.183	*-12.637*	*-5.455*	*4.590*	*50.181*

Coefficients:

Estimate Std. Error t value $Pr(> |t|)$

cars$speed 2.9091 0.1414 20.58 $<$ *2e-16 ****
*Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1*

Residual standard error: 16.26 on 49 degrees of freedom

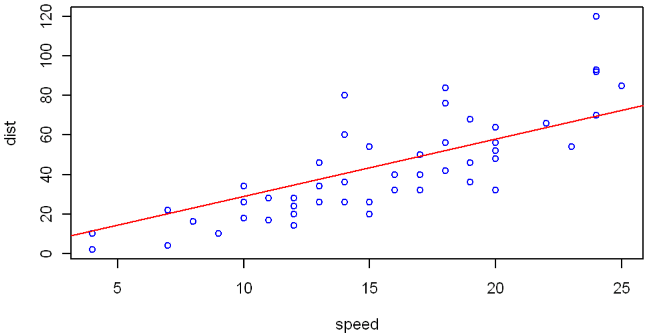
Multiple R-Squared: 0.8963, Adjusted R-squared: 0.8942

F-statistic: 423.5 on 1 and 49 DF, p-value: $< 2.2e - 16$.

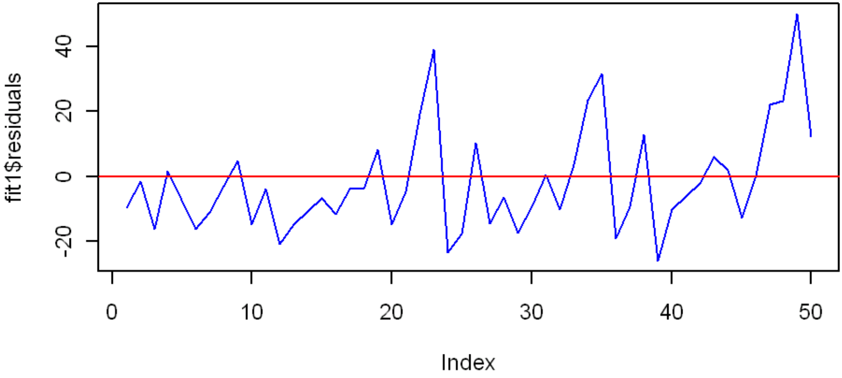
Le R^2 s'approche de 1, donc l'ajustement est meilleur.

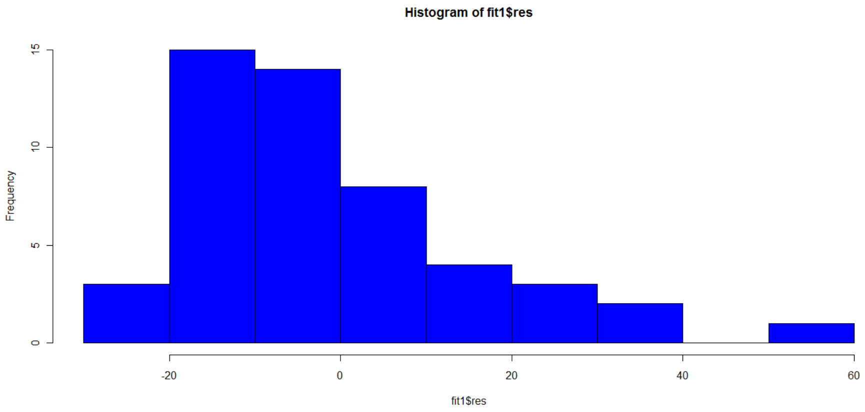
Commandes sur R

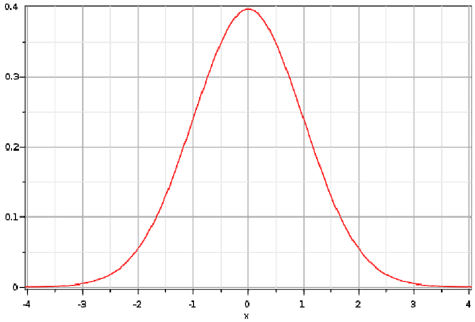
plot(cars,col="blue"), z $< -$ lm(dist ~ speed, data = cars), abline(z,col="red")



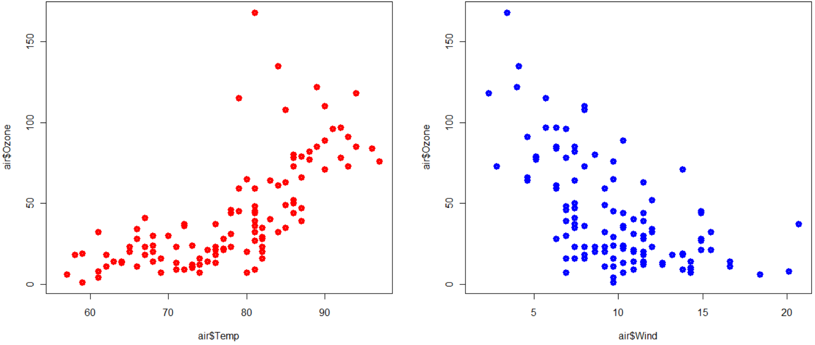
Analyse des résidus de la régression

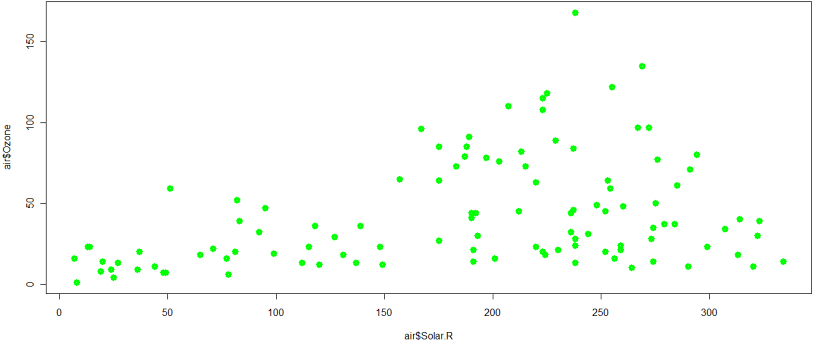

fit1$residuals
-20
0
20
40
0
10
20
30
40
50
Index


Histogram of fit1$res
Frequency
0
5
10
15
-20
0
20
40
60
fit1$res


0.4
0.3
0.2
0.1
0
-4
-3
-2
-1
0
1
2
3
4
x

Exemple 2. Régression multiple: On considère l'évolution de l'ozone en fonction de la température, du vent et de la radiation solaire.


air$Ozone
0
50
100
150
60
70
80
90
air$Temp
air$Ozone
0
50
100
150
5
10
15
20
air$Wind


air$Ozone
0
50
100
150
0
50
100
150
200
250
300
air$Solar.R

Commandes sur R

$fit3 <- lm(air\$Ozone \sim air\$Temp + air\$wind + air\$Solar.R)$,
$summary(fit3)$

Call: lm(formula = air$Ozone   air$Temp + air$Wind + air$Solar.R)

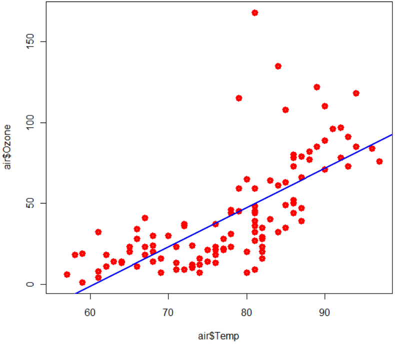
Residuals:

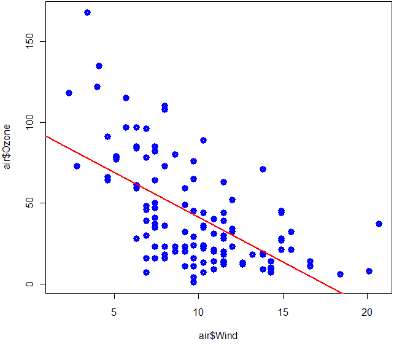
Min	1Q	Median	3Q	Max
-40.485	-14.219	-3.551	10.097	95.619

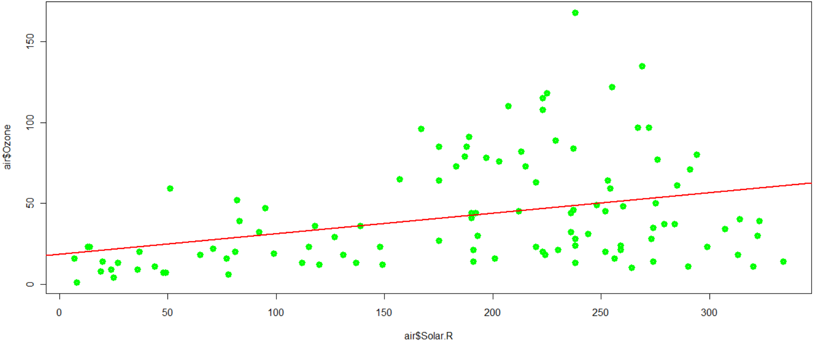
Coefficients:

	Estimate	Std. Error	t value	$Pr(>\mid t \mid)$

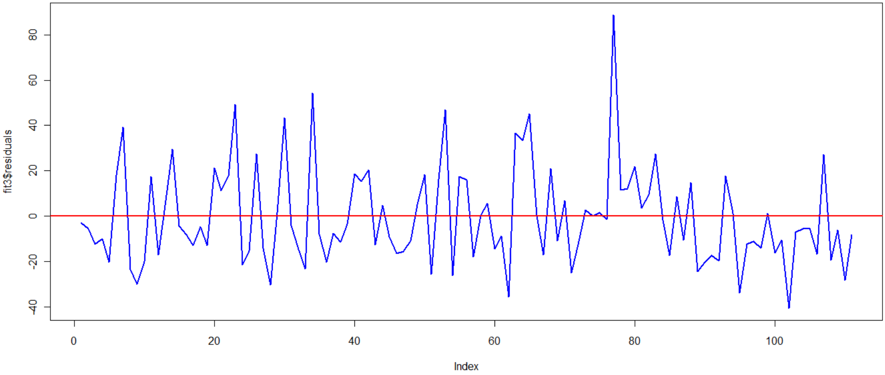
(Intercept) -64.34208 23.05472 -2.791 0.00623 **
air$Temp 1.65209 0.25353 6.516 2.42e-09 ***
air$Wind -3.33359 0.65441 -5.094 1.52e-06
***
air$Solar.R 0.05982 0.02319 2.580 0.01124 *
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 21.18 on 107 degrees of freedom (42 observations deleted due to missingness)
Multiple R-squared: 0.6059, Adjusted R-squared: 0.5948
F-statistic: 54.83 on 3 and 107 DF, p-value: $< 2.2e - 16$

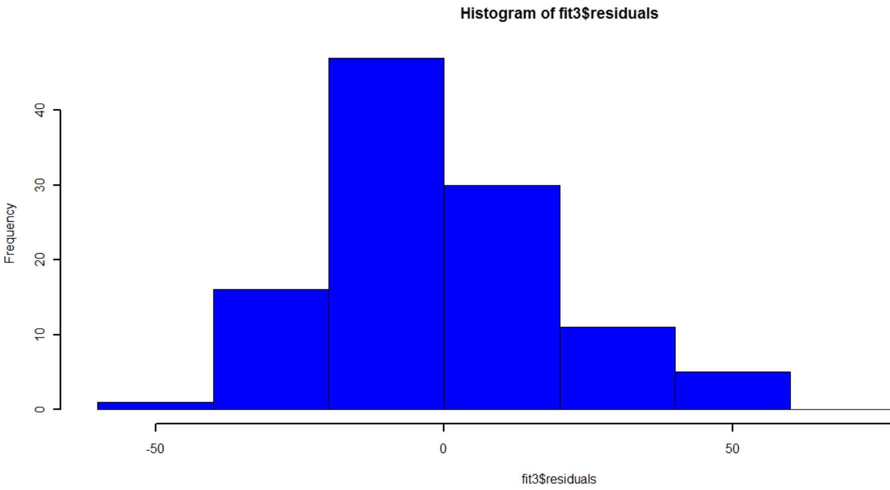

air$Ozone
air$Temp
0
50
100
150
60
70
80
90


air$Ozone
air$Wind
0
50
100
150
5
10
15
20



Les résidus de la régression sont donnés par


fit3$residuals
-40
-20
0
20
40
60
80
0
20
40
60
80
100
Index


Histogram of fit3$residuals
Frequency
0
10
20
30
40
-50
0
50
fit3$residuals

Conclusion: L'ajustement est assez bon.

Contenu

Definition 26

Soit $X = (X_1, ..., X_n)$ un échantillon de taille n suivant une loi $Q \in \mathcal{Q}$, avec $\mathcal{Q}$ est une famille de lois de probabilités sur un espace $(E, \mathcal{E})$. On définit la **loi empirique** de X par

$$P_n = \frac{1}{n} \sum_{1}^{n} \delta_{X_k} \qquad (13)$$

où δ_{X_k} est la mesure de Dirac au point X_k.

Definition 27

Soit $X = (X_1, ..., X_n)$ un échantillon de taille n suivant une loi $Q \in \mathcal{Q}$, et de loi empirique P_n. Soit $\mathcal{P}_n$ l'ensemble des mesures de probabilités sur $(E, \mathcal{E})$ de la forme $\frac{1}{n}\sum_1^n \delta_{x_k}, x_k \in E, k = 1, ..., n$. Soit F une fonctionnelle définie sur $\mathcal{Q} \cap \mathcal{P}_n$.
L'estimateur **empirique** de $F(Q)$ est la v.a.r. $F(P_n)$.

- Estimateur d'une loi à support fini On suppose qu'il existe $\{x_1, ..., x_r\} \in \mathcal{E}$ tel que $\forall Q \in \mathcal{Q}, \sum_1^r Q(\{x_k\}) = 1$ et on veut estimer $q_k = F_k(Q) = Q(x_k) = P(X_1 = x_k)$.

L'estimateur empirique de q_k n'est autre que la fréquence relative empirique

$$F_k(P_n) = \frac{1}{n}\sum_{j=1}^{n} \mathbf{1}_{\{X_j = x_k\}},$$

- Méthode des moments

On suppose que $(E, \mathcal{E}) = (\mathbb{R}, \mathcal{B})$ et que X_i a son moment d'ordre k fini, et on souhaite l'estimer.
$m_k = F_k(Q) = E(X_1^k) = \int x^k dQ(x)$, donc l'estimateur empirique de m_k est donné par

$$\widehat{m}_k = F_k(P_n) = \frac{1}{n}\sum_{j=1}^{n} X_j^k.$$

En particulier les estimateurs empiriques de la moyenne m_1 et la variance $\sigma^2 = m_2 - m_1^2$ de X_1 sont respectivement

$\widehat{m}_1 = \overline{X}_n = \frac{1}{n}\sum_{j=1}^{n} X_j$ et

$\widehat{\sigma}^2 = \widehat{m}_2 - \widehat{m}_1^2 = \frac{1}{n}\sum_{j=1}^{n} X_j^2 - (\overline{X}_n)^2 = \frac{1}{n}\sum_{j=1}^{n}(X_j - \overline{X}_n)^2.$

- Estimateur du support d'une loi

On suppose que $(E, \mathcal{E}) = (\mathbb{R}, \mathcal{B})$ et qu'il existe deux réels finis a et b tels que l'intervalle $[a, b]$ charge la loi Q i.e

$Q[a,b]) = 1$. On définit

$$F_m(Q) = \sup \left\{ \alpha \in \mathbb{R}, \text{tel que} \quad Q([\alpha, +\infty[)=1 \right\},$$

$$F_M(Q) = \inf \left\{ \beta \in \mathbb{R}, \ \text{tel que} \ \ Q(\,]-\infty, \beta[\,) = 1 \right\},$$

les estimateurs empiriques de a et b sont donnés par :

$$\widehat{a} = F_m(P_n) = inf(X_k, k = 1, ..., n)$$

et

$$\widehat{b} = F_M(P_n) = sup(X_k, k = 1, ..., n).$$

- Estimateur des quantiles d'une loi

Definition 28

Soit $X = (X_1, ..., X_n)$ un échantillon de taille n, le v.a. $\widetilde{X} = (X_{(1)}, ..., X_{(n)})$ obtenu en rangeant les X_i par ordre croissant est appelé la **statistique d'ordre** de X et $X_{(i)}$ est la ième statistique d'ordre.

Definition 29

Soit Q une loi de probabilité sur $(\mathbb{R}, \mathcal{B})$ et F sa fonction de répartition.
Soit $\alpha \in]0,1[$, alors $\exists x_\alpha \in \mathbb{R}$ tel que $F(x_\alpha) \leq \alpha \leq F(x_\alpha^+)$; x_α est appelé **quantile** d'ordre α, un quantile d'ordre $1/2$ est appelé **médiane**.
On appelle **intervalle des quantiles** d'ordre α :
$I_\alpha = \left\{x \in \mathbb{R},\ \text{tel que } \mathrm{F(x)} \leq \alpha \leq \mathrm{F(x^+)}\right\}$.

Notons $q_\alpha(Q)$ le milieu de l'intervalle I_α, on définit alors les estimateurs empiriques des quantiles d'ordre α par :

i) Si $\alpha = 1/2$ (estimation de la médiane),

$$q_{1/2}(P_n) = \begin{cases} X_{(k+1)} & \text{si} \quad n = 2k+1 \\ (X_{(k)} + X_{(k+1)})/2 & \text{si } n = 2k \end{cases}$$

ii)Si α n'est pas un multiple de $1/n$,

$$q_\alpha(P_n) = (X_{([n\alpha])} + X_{([n\alpha]+1)})/2.$$

Contenu

Definition 30

Soit $(E, \mathcal{E}, P_\theta, \theta \in \Theta)$ un modèle statistique dominé par une mesure μ.
Une **vraisemblance** est une fonction $(\theta, x) \to L(\theta, x)$ telle que $\forall \theta \in \Theta, L(\theta) = L(\theta, x)$ soit une version de $f_\theta = \frac{dP_\theta}{d\mu}$.

Definition 31

Soit $(E, \mathcal{E}, P_\theta, \theta \in \Theta)$ un modèle paramètrique ($\Theta \in \mathbb{R}^d$) dominé par une mesure μ, et T une statistique exhaustive à valeurs dans $(E, \mathcal{E})$.
On suppose donné une application $L : \Theta \times E \to \mathbb{R}$, telle que $L(\theta, T)$ est une vraisemblance.
L'estimateur du **maximum de vraisemblance** de θ est donné par $\widehat{\theta}(T)$ avec $\widehat{\theta} : E \to \Theta$ vérifiant:

$$\forall t \in E, L(\widehat{\theta}(t), t) = \sup_{\theta \in \Theta} L(\theta, t).$$

Soit $X = (X_1, ..., X_n)$ un échantillon de taille n suivant une loi $Q_\theta = f_\theta.\mu$, avec $f_\theta > 0$ $\mu.p.p.$, alors $L(\theta, X) = \prod_1^n f_\theta(X_i)$ est une vraisemblance.

L'estimateur du maximum de vraisemblance $\widehat{\theta}(X)$ est tel que

$$L(\widehat{\theta}(x), x) = \sup_{\theta \in \Theta} \prod_{i=1}^{n} f_\theta(x_i), \quad \mu^n p.p.$$

Remarques

- Souvent il est beaucoup plus pratique, surtout pour les modèles exponentiels, de maximiser la fonction $\log L(\theta, x)$ au lieu de $L(\theta, x)$.
- Si Θ est un ouvert de $\mathbb{R}^d$, et $\theta :\to f_\theta(x_i)$ est différentiable sur $\Theta, \forall x_i \in support(\mu)$ alors une condition nécessaire pour que $\log L(\theta, x)$ soit maximale est que

$$\sum_{i=1}^{n} \nabla_\theta \log f_\theta(x_i) \Big|_{\theta=\widehat{\theta}(x)} = 0.$$

Theorem 32

Soit $(E, \mathcal{E}, P_\theta, \theta \in \Theta)$ un modèle exponentiel canonique dominé par une mesure μ, $dP_\theta = exp(< \theta, T > -\Phi(\theta))d\mu$. On suppose en plus que $\forall \theta \in \Theta$, $\mathrm{var}_\theta(T)$ est définie positive, si $\widehat{\theta}(t)$ vérifie

$$E_{\widehat{\theta}(t)}(T) = t, \forall t \in support\ \mu T^{-1}, \tag{14}$$

alors $\widehat{\theta}(T)$ est l'unique estimateur du maximum de vraisemblance de θ.

Démonstration. Dans ce cas $L(\theta, t) = exp(< \theta, t > -\Phi(\theta))$, donc $\widehat{\theta}(t)$ est solution de $\frac{d\partial}{\partial\theta} \log L(\theta, t) = 0$ ou encore $t - \nabla\Phi(\theta) = 0$; de plus d'après le théorème 7 $\nabla\Phi(\theta) = E_\theta(T)$, c'est à dire $E_\theta(T) = t$, donc si $\widehat{\theta}(t)$ vérifie l'équation (14) il réalise le maximum de la vraisemblance et donc $\widehat{\theta}(T)$ est un estimateur du maximum de la vraisemblance de θ.
Son unicité vient du fait que la fonction $\theta \rightarrow < \theta, t > -\Phi(\theta)$ possède un Hessien définie négative (donc elle est strictement concave), car d'après le théorème 7

$$-\frac{\partial^2\Phi(\theta)}{\partial\theta_i\partial\theta_j} = -\operatorname{var}_\theta(T) < 0.\square$$

Exemple. On considère le modèle de régression définit par (8) avec (u_t) un bruit blanc gaussien.

Dans ce cas la log-vraisemblance est donnée par

$$\log L(\theta, \sigma^2, X) = -\frac{n}{2}\log \sigma^2 - \frac{n}{2}\log 2\pi - \frac{1}{2\sigma^2}\left\|X - \mathbf{Z}\theta\right\|^2,$$

$$\frac{\partial \log L(\theta, \sigma^2, X)}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\left\|X - \mathbf{Z}\theta\right\|^2 = 0,$$

$$\frac{\partial \log L(\theta, \sigma^2, X)}{\partial \theta} = -\frac{1}{2\sigma^2}(-2\,{}^t\mathbf{Z}X + 2\,{}^t\mathbf{Z}\mathbf{Z}\theta) = 0;$$

et donc

$$\widehat{\theta}_m = ({}^t\mathbf{Z}\mathbf{Z})^{-1}\,{}^t\mathbf{Z}X, \widehat{\sigma}_m^2 = \frac{1}{n}\left|\left|X - \mathbf{Z}\widehat{\theta}_m\right|\right|^2$$

Remarque. L'estimateur du maximum de vraisemblance $\widehat{\theta}_m$ de θ coïncide avec l'estimateur des moindres carrés $\widehat{\theta}$ donné par l'équation (10), ceci est du au fait que le modèle de régression est linéaire en le paramètre θ; cependant les estimateurs de la variance σ^2 ne coïncident pas.

Contenu

On suppose que Θ est un ouvert de $\mathbb{R}^d$, et que $L(\theta) \overset{p.s}{=} \frac{dP_\theta}{d\mu}$ est une vraisemblance.

Definition 33

On appelle **Score** la quantité

$$S(\theta) = \frac{\partial \log L(\theta)}{\partial \theta}. \tag{15}$$

Definition 34

On appelle **matrice d'information de Fisher** la quantité

$$I(\theta) = E_\theta\left(S(\theta)\,{}^tS(\theta)\right) = E_\theta\left(\frac{\partial \log L(\theta)}{\partial \theta_i}\frac{\partial \log L(\theta)}{\partial \theta_j}\right)_{1\leq i,j\leq d} \tag{16}$$

La matrice $I(\theta)$ est bien définie si l'application $\theta \to L(\theta)$ est presque sûrement différentiable et que le v.a. $S(\theta)$ est de carré intégrable, de plus $I(\theta)$ est une matrice positive (c'est le moment d'ordre 2 de $S(\theta)$).

Theorem 35

(Borne de Cramer-Rao) On suppose que:
i) $L(\theta) > 0 \;\; P.p.s.$
ii) $\theta \to L(\theta)$ *est presque sûrement différentiable sur* Θ *et que le vecteur Score* $S(\theta)$ *est centré et de carré intégrable pour* P_θ.
iii) $I(\theta)$ *est inversible.*
Alors pour toute variable aléatoire T *de carré* P_θ*-intégrable vérifiant*

$$\nabla E(L(\theta)T) = E(\nabla L(\theta)T),$$

on a

$$var_\theta(T) \geq \, {}^t(\nabla E_\theta(T)) I^{-1}(\theta)(\nabla E_\theta(T)). \tag{17}$$

Démonstration.

$$\begin{aligned}
\nabla E_\theta(T) &= E(\nabla L(\theta)T) \\
&= E_\theta(S(\theta)T) \text{ car } \nabla L(\theta) = L(\theta)\nabla \log L(\theta) \\
&= E_\theta\left((S(\theta)(T - E_\theta T)\right) \text{ car } E_\theta\left(S(\theta)\right) = 0;
\end{aligned}$$

Soit $v \in \mathbb{R}^d$,

$$\begin{aligned}
< v, \nabla E_\theta(T) >^2 &= \left(E_\theta\left(< v, S(\theta) > (T - E_\theta T)\right)\right)^2 \\
&\leq E_\theta\left(< v, S(\theta) >^2\right) E_\theta\left((T - E_\theta T)^2\right),
\end{aligned}$$

d'après l'inégalité de Cauchy-Schwartz, par conséquent

$$\begin{aligned} \text{var}_{\theta}(T) &\geq \frac{< v, \nabla E_\theta(T) >^2}{E_\theta\left(< v, S(\theta) >^2\right)} \\ &= \frac{< v, \nabla E_\theta(T) >^2}{< v, I(\theta) v >}, \end{aligned}$$

le résultat (17) s'obtient alors en prenant
$v = I^{-1}(\theta)\nabla E_\theta(T).\square$

Corollaire 36

Si T est un estimateur sans biais de $g(\theta) \in \mathbb{R}$ et si g est différentiable, alors

$$\operatorname{var}_{\theta}(T) \geq {}^{t}(\nabla g(\theta)) I^{-1}(\theta)(\nabla g(\theta)).$$

Definition 37

Si l'estimateur T sans biais de $g(\theta) \in \mathbb{R}$ atteint la borne de Cramer-Rao (i.e il y a égalité dans (17), alors il est dit **efficace** .

Contenu

Le but de ce paragraphe est de proposer une région qui contient le paramètre inconnu $g(\theta)$ avec une probabilité fixée à l'avance, la construction de cette région est basée sur un estimateur $T(X)$ de $g(\theta)$ et dépend de la loi de cet estimateur.

Definition 38

On appelle **région de confiance de niveau** $\alpha, 0 < \alpha < 1$ pour $g(\theta)$ une application mesurable $R : E \to P(g(\Theta))$, telle que

$$\forall \theta \in \Theta, P(\omega, g(\theta) \in R(X(\omega))) \geq 1 - \alpha \qquad (18)$$

$(1 - \alpha)$ est le **niveau de confiance** de la région R.

Remarque. En pratique la région R est la plus petite région telle qu'on a égalité dans (18),

Definition 39

On appelle **pivot** pour $g(\theta)$ toute application $\Pi : E \times g(\Theta) \to \mathbb{R}^p$ telle que:
i) $\forall \theta \in \Theta$ fixé, l'application : $x \to \Pi(x, g(\theta))$ est mesurable,
ii) la loi de probabilité de $\Pi(X, g(\theta))$ ne dépend pas de θ.

Généralement c'est à l'aide d'un pivot qu'on peut construire une région de confiance, nous allons illustrer ceci à l'aide de plusieurs exemples.

Intervalle de confiance pour la moyenne d'un échantillon gaussien

On suppose que $(X_i)_{1\leq i\leq n}$ est une suite i.i.d. $N(m,\sigma^2)$.

Cas σ^2 connue.

On a $\theta = m$,

$$\Pi(X,\theta) = \frac{m-\overline{X}_n}{\sigma/\sqrt{n}} \rightsquigarrow N(0,1),$$

où $\overline{X}_n = \frac{1}{n}\sum_1^n X_k$.

$$a \leq \Pi(X,\theta) \leq b \Longleftrightarrow \overline{X}_n + \frac{\sigma}{\sqrt{n}}a \leq m \leq \overline{X}_n + \frac{\sigma}{\sqrt{n}}b;$$

donc R est telle que $-a = b = u_{\alpha/2}$ et

$$P(|N(0,1)| \leq u_{\alpha/2}) = 1 - \alpha;$$

donc l'intervalle de confiance pour m est donné par

$$I_m = [\overline{X}_n - \frac{\sigma}{\sqrt{n}} u_{\alpha/2}, \overline{X}_n + \frac{\sigma}{\sqrt{n}} u_{\alpha/2}]. \tag{19}$$

Cas σ^2 inconnue. On a

$$\theta = (m, \sigma^2), \Pi(X, \theta) = \frac{m - \overline{X}_n}{\sqrt{S_n^2/(n-1)}}; S_n^2 = \frac{1}{n}\sum_{k=1}^{n}(X_k - \overline{X}_n)^2;$$

alors $\Pi(X, \theta) = \frac{(m-\overline{X}_n)\sqrt{n}/\sigma}{\sqrt{nS_n^2/(n-1)\sigma^2}} \rightsquigarrow t(n-1)$, (loi de Student à (n-1) degrés de liberté); donc l'intervalle de confiance pour m est donné par

$$I_m = [\overline{X}_n - \frac{\widetilde{S}_n}{\sqrt{n}} t_{\alpha/2}, \overline{X}_n + \frac{\widetilde{S}_n}{\sqrt{n}} t_{\alpha/2}], \widetilde{S}_n^2 = nS_n^2/(n-1).$$

avec $P(|t(n-1)| \leq t_{\alpha/2}) = 1 - \alpha$.

Intervalle de confiance pour la variance d'un échantillon gaussien

$\theta = (m, \sigma^2)$,

$$\Pi(X, \theta) = \frac{1}{\sigma^2} \sum_{k=1}^{n} (X_k - \overline{X}_n)^2 \rightsquigarrow \mathcal{X}^2(n-1);$$

$$0 < x_1 \leq \Pi(X, \theta) \leq x_2$$

$$\Longleftrightarrow \frac{1}{x_2} \sum_{k=1}^{n} (X_k - \overline{X}_n)^2 \leq \sigma^2 \leq \frac{1}{x_1} \sum_{k=1}^{n} (X_k - \overline{X}_n)^2;$$

donc l'intervalle de confiance pour σ^2 est donné par

$$I_{\sigma^2} = [\frac{1}{x_2}\sum_{k=1}^{n}(X_k - \overline{X}_n)^2, \frac{1}{x_1}\sum_{k=1}^{n}(X_k - \overline{X}_n)^2], \qquad (20)$$

x_1 et x_2 sont tels que $\frac{1}{x_1} - \frac{1}{x_2}$ soit minimale et $P(x_1 \leq \Pi(X,\theta) \leq x_2) = 1 - \alpha$.

Intervalle de confiance pour les paramètres de régression

On considère le modèle de régression (8), et on suppose que le bruit (u_t) est gaussien $N(0, \sigma^2)$; on considère les estimateurs des moindres carrés de θ et σ^2 définis par (10) et (11) ; et on veut construire un intervalle de confiance pour les paramètres $\theta_j, 1 \leq j \leq k$.

Theorem 40

l'estimateur des moindres carrés et tel que

$$\frac{\widehat{\theta}_j - \theta_j}{\sqrt{\widehat{\sigma}^2 z_{j,j}}} \rightsquigarrow t(n-p), \forall 1 \leq j \leq p.$$

où $z_{j,j}$ est le j ème élément diagonal de la matrice $({}^t\mathbf{Z}\mathbf{Z})^{-1}$.

Pour la démonstration de ce théorème on a besoin du lemme suivant

Lemme 41

Soit $u \rightsquigarrow N(0, I_n)$ et H un projecteur de $\mathbb{R}^n$ sur un sous-espace E de $\mathbb{R}^n$ de dimension $r \leq n$. Alors

$$||Hu||^2 \rightsquigarrow \mathcal{X}^2(r).$$

Démonstration

Soit $B = (e_1, ..., e_n)$ la base canonique de $\mathbb{R}^n$, soit $(f_1, ..., f_r)$ une base orthonormale de E. Alors on peut toujours compléter cette dernière base de telle sorte que
$B_1 = (f_1, ..., f_r, f_{r+1}, ..., f_n)$ soit une base orthonormale de $\mathbb{R}^n$; soit P la matrice de passage de B à B_1,

alors $u = \sum_1^n u_i e_i = \sum_1^n v_i f_i$ avec
$v = P^{-1}u, v = {}^t(v_1, ..., v_n), u = {}^t(u_1, ..., u_n)$, donc
$Hu = \sum_1^r v_i f_i$ ce qui implique que $||Hu||^2 = \sum_1^r v_i^2$, mais
comme P est orthonormale (i.e. $P^{-1} = {}^tP$) on a
$v \rightsquigarrow N(0, I_n)$.$\square$

Démonstration du théorème 40

$\widehat{\theta} - \theta \rightsquigarrow N(0, \sigma^2({}^tZZ)^{-1})$ (car c'est un transformé linéaire de $U \rightsquigarrow N(0, I_n)$), donc

$$\frac{\widehat{\theta}_j - \theta_j}{\sqrt{\sigma^2 z_{j,j}}} \rightsquigarrow N(0,1).$$

$P = \mathbf{Z}({}^t\mathbf{Z}\mathbf{Z})^{-1}\ {}^t\mathbf{Z}$, $P^\perp = I_n - P$, alors

$$\frac{\widehat{\sigma}^2(n-p)}{\sigma^2} = \left|\left|P^\perp U\right|\right|^2,$$

comme $P^\perp$ est un projecteur sur un sous-espace de dimension $(n-p)$ on a alors

$$\widehat{\sigma}^2(n-p)/\sigma^2 \rightsquigarrow \mathcal{X}^2(n-p). \quad (21)$$

Le théorème est prouvé si on montre que $\widehat{\theta}$ et $\widehat{\sigma}^2$ sont indépendants, en effet $\widehat{\sigma}^2$ est une fonction mesurable de $P^{\perp}U$, $\widehat{\theta} = \theta + ({}^tZZ)^{-1}\, {}^tZPU$ est une fonction mesurable de PU, il suffit alors de montrer que $P^{\perp}U$ et PU sont deux v.a. indépendants ; or ${}^t({}^t(P^{\perp}U), {}^t(PU))$ est gaussien et $\mathrm{cov}(P^{\perp}U, PU) = \sigma^2 P^{\perp}P = 0.\square$

Le théorème ci-dessus implique que l'intervalle de confiance pour θ_j est donné par

$$I_{\theta_j} = [\widehat{\theta}_j - t_{\alpha/2}\sqrt{z_{j,j}}\widehat{\sigma}, \widehat{\theta}_j + t_{\alpha/2}\sqrt{z_{j,j}}\widehat{\sigma}].$$

Soient $(X_i)_{1\leq i\leq n}$ une suite de v.a. à valeurs dans $(E, \mathcal{E}, P_\theta, \theta \in \Theta)$, et $T_n = T(X_1, ..., X_n)$ un estimateur de $g(\theta)$.

Le but de ce chapitre est de décrire le comportement asymptotique de T_n lorsque la taille n tends vers $+\infty$.

Contenu

Definition 42

On dit que T_n est un estimateur **fortement consistant** si T_n converge presque sûrement vers $g(\theta)$:

$$T_n \overset{p.s}{\rightarrow} g(\theta) \tag{22}$$

Definition 43

On dit que T_n est un estimateur **faiblement consistant** si T_n converge en probabilité vers $g(\theta)$:

$$T_n \overset{P}{\rightarrow} g(\theta) \tag{23}$$

Remarque . Souvent on dit que T_n est consistant s'il est faiblement consistant.

Theorem 44

Si T_n est un estimateur, posons $b_n = E(T_n) - g(\theta)$, (i.e. le biais de T_n), $v_n = \mathrm{var}(T_n)$; on suppose que
i) $b_n \to 0$ (c'est à dire T_n est asymptotiquement sans biais),
ii) $v_n \to 0$.
Alors T_n est consistant.

Démonstration

Soit $EQ_{T_n} = E(|T_n - g(\theta)|^2)$. On a $EQ_{T_n} = v_n + b_n^2$; donc

$$EQ_{T_n} \to 0, \text{ donc } T_n \overset{L^2(\Omega,P)}{\to} g(\theta),$$

et par conséquent

$$T_n \overset{P}{\to} g(\theta)$$

(car la convergence en moyenne quadratique entraîne la convergence en probabilité, par l'inégalité de Bienaymé-Tchebychev).

Theorem 45

Soit $T_n(X_1, ..., X_n)$ une suite d'estimateurs U.V.M.B. de $g(\theta)$, et $(X_i)_{1\leq i\leq n}$ est une suite i.i.d. Alors T_n est consistant.

Démonstration. Vu que T_n est sans biais, il suffit de montrer, d'après le théorème 44, que la variance de T_n converge vers 0. Posons $U_n = \frac{1}{n}\sum_1^n T_1(X_k)$;

$$\begin{aligned} var_\theta(U_n) &= \frac{1}{n}\mathrm{var}(T_1(X_1)); \text{ car les } X_k \text{ sont i.i.d.} \\ &\geq \mathrm{var}(T_n) \text{ car } T_n \text{ est U.V.M.B. et } U_n \text{ est sans biais.} \end{aligned}$$

□

On suppose que (X_i) est une suite i.i.d suivant une loi $N(5, 1)$

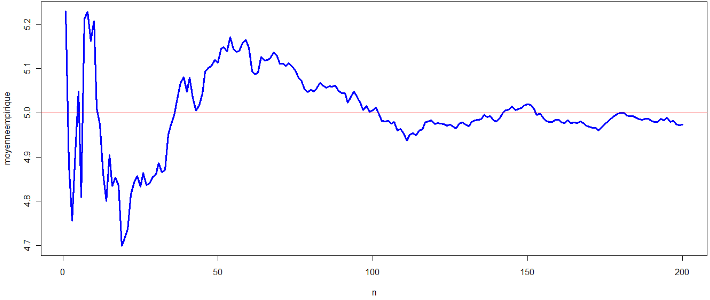


Figure: Evolution de la moyenne empirique

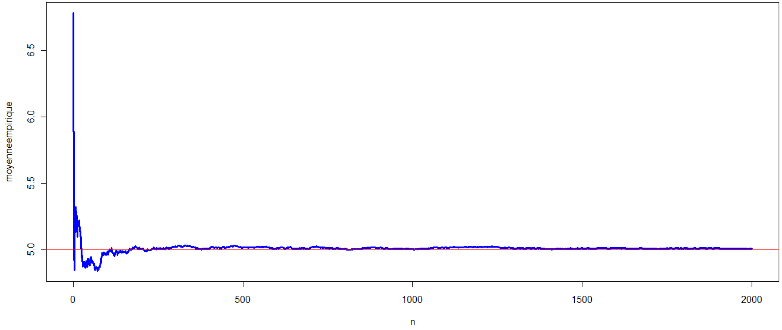


Figure: Evolution de la moyenne empirique

Contenu

On s'intéresse ici à étudier la distribution asymptotique de l'estimateur T_n lorsque n tends vers l'infini, le but de ceci et de faire de l'inférence statistique à savoir la construction d'intervalle de confiance pour les paramètres à estimer et tester la significativité des paramètres dans les modèles paramétriques. Soit $T_n(X_1, ..., X_n)$ une suite d'estimateurs de $g(\theta)$.

Definition 46

Supposons qu'il existe deux suites réelles $m_n(\theta)$ et $\sigma_n(\theta) > 0$ telles que

$$\frac{T_n - m_n(\theta)}{\sigma_n(\theta)} \xrightarrow{\mathcal{L}} N(0,1); \qquad (24)$$

on dit que T_n est **asymptotiquement normal.**

- $\lim_{n\to\infty} m_n(\theta)$ est l'espérance asymptotique de T_n,
- $\lim_{n\to\infty} \sigma_n^2(\theta)$ est la variance asymptotique de T_n.

On suppose que $(X_i)_{1 \leq i \leq n}$ est une suite i.i.d., Θ est un ouvert de $\mathbb{R}$, et $E = \mathbb{R}^{\mathrm{k}}$.

Posons

$$\widetilde{P}_\theta = P_\theta^{\mathbb{N}},$$

$\widetilde{P}_\theta$ est la probabilité image de P par le vecteur aléatoire de dimension infinie $(X_1, X_2, ...)$, elle est définie sur $E^{\mathbb{N}}$.

Theorem 47

On suppose que:
i) $(\mathbb{R}^k, \mathcal{B}_k, P_\theta, \theta \in \Theta)$ *est dominé par* μ *avec* $dP_\theta = f_\theta . d\mu$,
ii) $f_\theta > 0 \;\; \mu.p.s.$
iii) l'application $\theta \to \log f_\theta(x)$ *est différentiable sur* $\Theta \;\; \mu.p.s.$
iv) $\forall \theta \in \Theta$, *il existe un voisinage* V *de* θ *tel que*
$\forall \theta' \in V, \log f_{\theta'}$ *est* P_θ*-intégrable. Alors il existe une suite* $\widehat{\theta}_n$ *telle que à partir d'un certain rang* $n_0(x)$,$\widehat{\theta}_n$ *soit une solution de l'équation de la vraisemblance :*

$$\sum_1^n \frac{\partial}{\partial \theta} \log f_\theta(X_i) = 0 \text{ et } \widehat{\theta}_n \overset{\widetilde{P}_\theta . p.s}{\to} \theta.$$

Theorem 48

Si en plus des hypothèses du théorème 47 on suppose que
i) l'application $\theta \to \log f_\theta(x)$ *est deux fois continûment différentiable sur* Θ, μ.*p.s. et* $\theta \to \frac{\partial^2}{\partial\theta^2} \log f_\theta(x)$ *est continue en* θ *uniformément par rapport à* x.

ii) $E_\theta(\frac{\partial}{\partial\theta} \log f_\theta(X_1)) = 0$, $E_\theta(\frac{\partial^2}{\partial\theta^2} f_\theta(X_1)/f_\theta(X_1)) = 0$,
iii) $0 < I_1(\theta) = E_\theta((\frac{\partial}{\partial\theta} \log f_\theta(X_1))^2) < \infty$.
Alors

$$\sqrt{nI_1(\theta)}(\widehat{\theta}_n(X) - \theta) \xrightarrow{\mathcal{L}} N(0,1). \tag{25}$$

Démonstration. Posons $l(\theta, x) = \log f_\theta(x)$.

En appliquant la formule de Taylor

$$\begin{aligned}
0 &= \sum_1^n \frac{\partial}{\partial\theta} l(\widehat{\theta}_n(X), X_i) = \sum_1^n \frac{\partial}{\partial\theta} l(\theta, X_i) \\
&+ (\widehat{\theta}_n(X) - \theta) \sum_1^n \frac{\partial^2}{\partial\theta^2} l(\theta, X_i) \\
& \quad +(\widehat{\theta}_n(X) - \theta) \left[\sum_1^n \frac{\partial^2}{\partial\theta^2} l(\overline{\theta}_n, X_i) - \sum_1^n \frac{\partial^2}{\partial\theta^2} l(\theta, X_i) \right],
\end{aligned}$$

avec $\overline{\theta}_n = \lambda\widehat{\theta}_n + (1-\lambda)\theta, \lambda \in]0, 1[$ donc

$$\sqrt{n}(\widehat{\theta}_n(X) - \theta) \quad = \quad \frac{\frac{1}{\sqrt{n}} \sum_1^n \frac{\partial}{\partial \theta} l(\theta, X_i)}{D(n, X)}. \tag{26}$$

$$D(n, X) = \frac{1}{n} \sum_1^n \frac{\partial^2}{\partial \theta^2} l(\theta, X_i) + \frac{1}{n} \left[\sum_1^n \frac{\partial^2}{\partial \theta^2} l(\overline{\theta}_n, X_i) - \frac{\partial^2}{\partial \theta^2} l(\theta, X_i) \right].$$

La suite $(\frac{\partial}{\partial \theta} l(\theta, X_i))$ est une suite i.i.d. centrée et de variance $E_\theta((\frac{\partial}{\partial \theta} \log f_\theta(X_1))^2) = I_1(\theta)$, donc d'après le théorème central limite on a

$$\frac{1}{\sqrt{n}}\sum_{1}^{n}\frac{\partial}{\partial\theta}l(\theta,X_i)\xrightarrow{\mathcal{L}}N(0,I_1(\theta)). \tag{27}$$

La suite $(\frac{\partial^2}{\partial\theta^2}l(\theta,X_i))$ est une suite i.i.d. d'espérance mathématique, en utilisant l'hypothèse ii), :

$$E_\theta(\frac{\partial^2}{\partial\theta^2}l(\theta,X_i))=-I_1(\theta);$$

donc d'après la loi des grands nombres

$$\frac{1}{n}\sum_{1}^{n}\frac{\partial^2}{\partial\theta^2}l(\theta,X_i)\xrightarrow{\widetilde{P}_\theta}-I_1(\theta).$$

L'hypothèse i) entraîne qu'il existe une fonction $d(x)$ avec $d(x) \to 0$ si $x \to 0$, telle que

$$\sum_{1}^{n} \frac{\partial^2}{\partial\theta^2} l(\overline{\theta}_n, X_i) - \frac{\partial^2}{\partial\theta^2} l(\theta, X_i) \leq d(\overline{\theta}_n - \theta),$$

et comme $\overline{\theta}_n - \theta \to 0 \ \widetilde{P}_\theta.p.s.$ on déduit que

$$\sum_{1}^{n} \frac{\partial^2}{\partial\theta^2} l(\overline{\theta}_n, X_i) - \frac{\partial^2}{\partial\theta^2} l(\theta, X_i) \overset{\widetilde{P}_\theta.p.s}{\longrightarrow} 0$$

La convergence (25) découle de (26)-(27) et du théorème 52 $\square$.

Contenu

Theorem 49

(version scalaire) Soit $(X_i)_{1 \leq i \leq n}$ une suite de variables aléatoires i.i.d. de moyenne m et de variance σ^2. Alors

$$\frac{1}{\sqrt{n}}(\sum_{i=1}^{n} X_i - nm) \xrightarrow{\mathcal{L}} N(0, \sigma^2).$$

A l'aide du théorème de Cramer-Wold:

$$(X_n) \xrightarrow{\mathcal{L}} X \Longleftrightarrow < u, X_n > \xrightarrow{\mathcal{L}} < u, X >, \forall u \in \mathbb{R}^d,$$

on déduit le théorème suivant :

Theorem 50

(version vectorielle) Soit $(X_i)_{1\leq i\leq n}$ une suite de vecteurs aléatoires i.i.d. de moyenne m et de matrice de covariance Σ. Alors

$$\frac{1}{\sqrt{n}}\left(\sum_{i=1}^{n} X_i - nm\right) \xrightarrow{\mathcal{L}} N(0,\Sigma).$$

Theorem 51

(changement de variables) Soit $(\theta_i)_{1\leq i\leq n}$ une suite de vecteurs aléatoires, et θ un vecteur aléatoire tels que

$$\sqrt{n}(\theta_n - \theta) \xrightarrow{\mathcal{L}} N(0, \Sigma). \tag{28}$$

Soit $g = {}^t(g_1, ..., g_q)$ une application de $\mathbb{R}^p$ dans $\mathbb{R}^q$ telle que ses dérivées partielles existent et continues. Alors

$$\sqrt{n}(g(\theta_n) - g(\theta)) \xrightarrow{\mathcal{L}} N(0, G(\theta)\Sigma\, {}^tG(\theta)),$$

où $G(\theta) = (G_{i,j})_{1\leq i\leq q, 1\leq j\leq p}$ est une matrice (q,p) avec $G_{i,j} = \frac{\partial g_i(\theta)}{\partial \theta_j}$.

Démonstration

En utilisant le développement de Taylor pour la fonction g on obtient

$$g(\theta_n) = g(\theta) + G(\overline{\theta}_n)(\theta_n - \theta),$$

où $\overline{\theta}_n = \lambda_i \theta_n(i) + (1 - \lambda_i)\theta(i), \lambda_i \in [0, 1], 1 \leq i \leq p$, avec $\theta(i)$ désignant la ième composante du vecteur θ, or la convergence (28) entraîne que $\theta_n \xrightarrow{P} \theta$ et donc $\overline{\theta}_n \xrightarrow{P} \theta$.

Par conséquent la continuité de G implique que $G(\overline{\theta}_n) \xrightarrow{P} G(\theta)$. La démonstration s'achève en écrivant

$$\sqrt{n}(g(\theta_n) - g(\theta)) = G(\overline{\theta}_n)\sqrt{n}(\theta_n - \theta),$$

et en utilisant le théorème suivant (voir cours licence).

Theorem 52

Si $X_n \xrightarrow{P} c$ *et* $Y_n \xrightarrow{\mathcal{L}} Y$. *Alors*

$$X_n Y_n \xrightarrow{\mathcal{L}} cY.$$

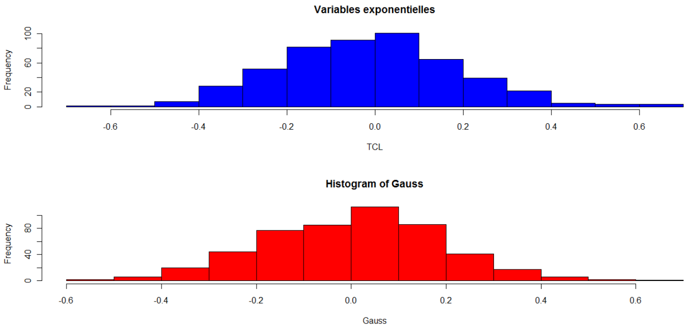


Figure: Théorème central limite

Tester une hypothèse revient à décider, en se basant sur l'observation $x = X(\omega)$, si celle-ci est vraie ou fausse, de tels problèmes sont nombreux et se posent dans la vie courante dans plusieurs domaines, par exemple:

- On veut tester l'efficacité d'un vaccin ou un médicament.
- Un patron d'usine de fabrication de pneus veut savoir si ses machines sont bien réglées.
- L'ajout d'un métal augmente ou non la résistance de l'acier fabriqué.

- Il se pose alors le problème de formulation statistique du test et de la décision.
- La formulation du test consiste à décider d'une hypothèse H_0 appelée **hypothèse nulle** contre une **hypothèse alternative** H_1.
- On se donne alors un modèle statistique $(E, \mathcal{E}, P_\theta, \theta \in \Theta)$ et une partition de l'espace des paramètres $\Theta = \Theta_0 \cup \Theta_1$ telle que $H_i = \{\theta \in \Theta_i\}$, $i = 0, 1$.

Definition 53

On appelle **région critique** d'un test le sous ensemble de E des observations qui mènent à rejeter H_0.

Definition 54

On appelle **fonction test** une application mesurable ϕ de $(E, \mathcal{E})$ dans $([0,1], \mathcal{B}_{[0,1]})$. Un test basé sur ϕ consiste à :

1) Rejeter H_0 si $\phi(x) = 1$
2) Accepter H_0 si $\phi(x) = 0$
3) Faire un tirage au sort selon une Bernoulli $B(\phi(x))$ et qui consiste à accepter H_0 avec une probabilité $1 - \phi(x)$ et refuser H_0 avec une probabilité $\phi(x)$.

Remarque 55

Le point 3) ne se présente que dans le cas discret.

Definition 56

Un test ϕ_1 est dit **préférable** (resp. **strictement préférable**) à un test ϕ_2 si

$$E_\theta(\phi_1(X)) \leq E_\theta(\phi_2(X)) \ \forall\theta \in \Theta_0,$$

et

$$E_\theta(\phi_1(X)) \geq E_\theta(\phi_2(X)) \ \forall\theta \in \Theta_1, \tag{29}$$

(resp. les deux inégalités sont stricts).
Si aucun test n'est strictement préférable à ϕ_1, on dit que celui-ci est **admissible**.

		hypothèse vraie	
		H_0	H_1
hypothèse acceptée	H_0	Décision correcte	Erreur de type II
	H_1	Erreur de type I	Décision correcte

Neyman et Pearson se sont intéressé aux tests dont la probabilité d'erreur de type I (accepter à tort H_1) est en dessous d'un certain seuil fixé auparavant et ensuite chercher parmi ceux-ci un test qui minimise la probabilité d'erreur de type II (accepter à tort H_0)

Definition 57

La **taille** d' un test ϕ est donnée par

$$\alpha_\phi = \sup_{\theta \in \Theta_0} E_\theta(\phi(X))$$

Le test est dit de **niveau** α si sa taille est inférieure ou égale à α.
La **puissance** du test ϕ est la fonction : $\Theta_1 \to \mathbb{R}$, définie par $\beta_\phi(\theta) = E_\theta(\phi(X))$.
On dit que ϕ_1 est plus puissant que ϕ_2 si $\beta_{\phi_1}(\theta) \geq \beta_{\phi_2}(\theta), \forall \theta \in \Theta_1$.
Le test ϕ est dit de **puissance maximum au niveau** α s'il est de niveau α, et s'il est plus puissant que tout test de niveau α.

Definition 58

Supposons que la région critique du test ϕ de niveau α est donnée par $W = \{x, T(x) > c\}$, soit F_T la fonction de répartition, sous H_0, de $T(X)$.
On définit la **probabilité critique ou " p-value "** du test ϕ par

$$p.c. \quad = \quad 1 - F_T(c). \tag{30}$$

Cela veut dire que si on choisit un niveau α plus grand que la probabilité critique alors l'hypothèse H_0 sera rejetée.

Remarque. Soit W la région critique du test ϕ c'est dire que

$$W = \{x \in E, x \text{ conduit au rejet de } H_0\},$$

ou encore $\{$ rejeter $H_0\} = \mathbf{1}_W(.)$, donc

$$\begin{aligned}
\phi(x) &= P(\text{ rejeter } H_0 \mid X = x) \\
&= P(\mathbf{1}_W(.) \mid X = x) \\
&= E(\mathbf{1}_W(X) \mid X = x)
\end{aligned}$$

donc

$$\begin{aligned}
\phi(X) &= E(\mathbf{1}_W(X) \mid X) \\
&= \mathbf{1}_W(X)
\end{aligned}$$

ceci implique que

$$\begin{aligned} E_\theta(\phi(X)) &= \int_E \mathbf{1}_W(x) dP_\theta(x) \\ &= P_\theta(W) \\ &= P_\theta(\text{ rejeter } H_0). \end{aligned}$$

En général on choisit α petit (0,01 ou 0,05), ce choix nous donne la probabilité maximale de rejeter à tort l'hypothèse H_0. En effet, si on pose $H_0 = \{$ les machines sont bien réglées$\}$, avec α petit la probabilité de rejeter H_0 sachant qu'elle est vraie est petite donc ceci minimise le risque qui est d'arrêter à tort les machines.

Contenu

Definition 59

Pour tester H_0 contre H_1 un test basé sur la région critique $W = \{x, \lambda(x) < k\}$, où

$$\lambda(x) = \frac{\sup_{\theta\in\Theta_0} f(\theta, x)}{\sup_{\theta\in\Theta} f(\theta, x)} \tag{31}$$

est appelé test de **rapport de vraisemblance**.

Remarque. Plus le rapport est petit plus on aura tendance à rejeter H_0.

Definition 60

On considère le modèle de régression (9). On dit que le paramètre θ satisfait une **contrainte linéaire** s'il existe une matrice $\mathbf{R}$ de taille (r, p) de rang $r \leq p$, et un vecteur c tels que

$$\mathbf{R}\theta = c \quad (32)$$

On veut tester $H_0 : \mathbf{R}\theta = c$ contre $H_1 : \mathbf{R}\theta \neq c$; accepter H_0 conduit souvent à une réduction du nombre de paramètres à estimer. De tel test sert aussi à tester la significativité des variables explicatives $z_{t,i}$ dans le modèle de régression donné par (8).

Theorem 61

(Test de Fisher) On suppose que le bruit u_t dans le modèle (8) est gaussien. Alors
i. Le test du rapport de vraisemblance de niveau α pour tester $H_0 : \mathbf{R}\theta = c$ contre $H_1 : \mathbf{R}\theta = c$; est basé sur la région critique $\{F(x) > f_0\}$, où

$$F(X) = \frac{\left|\left|X - \mathbf{Z}\widetilde{\theta}\right|\right|^2 - \left|\left|X - \mathbf{Z}\widehat{\theta}\right|\right|^2}{\left|\left|X - \mathbf{Z}\widehat{\theta}\right|\right|^2};$$

avec $\widehat{\theta}$ et $\widetilde{\theta}$ sont les estimateurs du maximum de vraisemblance sans la contrainte et sous la contrainte (32).
ii. Sous H_0, la v.a.r. $(n-p)F(X)/r$ suit une loi Fisher

Démonstration. i. On a

$$\lambda(x) = \frac{\sup_{(\theta,\sigma^2)\in\Theta_0} f((\theta,\sigma^2),x)}{\sup_{(\theta,\sigma^2)\in\Theta} f((\theta,\sigma^2),x)} = \frac{f((\widetilde{\theta},\widetilde{\sigma}^2),x)}{f((\widehat{\theta},\widehat{\sigma}^2),x)},$$

comme

$$\widehat{\sigma}^2 = \frac{1}{n}\left|\left|X - \mathbf{Z}\widehat{\theta}\right|\right|^2, \text{ et } \widetilde{\sigma}^2 = \frac{1}{n}\left|\left|X - \mathbf{Z}\widetilde{\theta}\right|\right|^2,$$

on en déduit alors que

$$\lambda(x) = (\widehat{\sigma}^2/\widetilde{\sigma}^2)^{n/2};$$

donc

$$
\begin{aligned}
\lambda(x) < k \quad & \Longleftrightarrow \quad \widetilde{\sigma}^2 / \widehat{\sigma}^2 > k_1 = k^{2/n} \\
& \Longleftrightarrow \quad \frac{\left\| X - \mathbf{Z}\widetilde{\theta} \right\|^2}{\left\| X - \mathbf{Z}\widehat{\theta} \right\|^2} > k_1 \\
& \Longleftrightarrow \quad \frac{\left\| X - \mathbf{Z}\widetilde{\theta} \right\|^2 - \left\| X - \mathbf{Z}\widehat{\theta} \right\|^2}{\left\| X - \mathbf{Z}\widehat{\theta} \right\|^2} > k_1 - 1.
\end{aligned}
$$

ii. Notons $\widehat{U} = X - \mathbf{Z}\widehat{\theta}, \widetilde{U} = X - \mathbf{Z}\widetilde{\theta}$. Comme les estimateurs des moindres carrés et le maximum de vraisemblance sont les mêmes, alors $\widehat{U} = X - \mathbf{Z}\widehat{\theta}$ et $\widetilde{U} = X - \mathbf{Z}\widetilde{\theta}$ sont respectivement les résidus dans le modèle non contraint et contraint. De plus on montre que (voir exercices)

$$\widetilde{\theta} = \widehat{\theta} - ({}^t\mathbf{Z}\mathbf{Z})^{-1}\,{}^t\mathbf{R}(\mathbf{R}({}^t\mathbf{Z}\mathbf{Z})^{-1}\,{}^t\mathbf{R})^{-1}(c - \mathbf{R}\widehat{\theta})$$

donc $\widetilde{U} - \widehat{U} = QU$ où

$Q = \mathbf{Z}({}^t\mathbf{Z}\mathbf{Z})^{-1}\,{}^t\mathbf{R}(\mathbf{R}({}^t\mathbf{Z}\mathbf{Z})^{-1}\,{}^t\mathbf{R})^{-1}\mathbf{R}({}^t\mathbf{Z}\mathbf{Z})^{-1}\,{}^t\mathbf{Z}$, donc $\widetilde{U}-\widehat{U}\perp\widehat{U}$, ce qui implique que $\left|\left|\widetilde{U}\right|\right|^2-\left|\left|\widehat{U}\right|\right|^2=\left|\left|\widetilde{U}-\widehat{U}\right|\right|^2$, et $\widetilde{U}-\widehat{U}$ et $\widehat{U}$ sont indépendants; de plus $Q^2=Q$ donc Q est un projecteur sur un espace de dimension r et par conséquent, d'après le lemme 41,

$$\frac{1}{\sigma^2}\left|\left|\widetilde{U}-\widehat{U}\right|\right|^2\rightsquigarrow\mathcal{X}^2(r),$$

et d'après (21)

$$\frac{1}{\sigma^2}\left|\left|\widehat{U}\right|\right|^2\rightsquigarrow\mathcal{X}^2(n-k).\square$$

Contenu

- Le but du test d'adéquation de $\mathcal{X}^2$ est de décider si l'échantillon observé $x = (X_1(\omega), ..., X_n(\omega))$ peut être considéré comme issu d'une loi bien spécifiée P_0.
- Effectuer le test $H_0 : P = P_0$ contre $H_1 : P \neq P_0$, ou plus généralement $H_0 : P \in \{P_\theta, \theta \in \Theta\}$ contre $H_1 : P \notin \{P_\theta, \theta \in \Theta\}$.
- Supposons que $card(E) < \infty$; c'est à dire que $E = \{x_1, ..., x_d\}$ et que l'on veut tester $H_0 : P = P_0$ avec $P_0(x_j) = p_j, j = 1, ..., d$ contre $H_1 : P \neq P_0$.

- Pour effectuer ce test on introduit la mesure de divergence (ou distance de Khi-deux) entre la loi théorique P_0 et la loi empirique P_n (définie par (13)):

$$D(P_n, P_0) = \sum_{j=1}^{d} \frac{n}{p_j}\left(\frac{N_j}{n} - p_j\right)^2; \tag{33}$$

où $N_j = \sum_{k=1}^{n} \mathbf{1}_{\{X_k = x_j\}}$ est la fréquence absolue empirique.

- Si H_0 est vraie alors par la loi des grands nombres $\frac{N_j}{n}$ s'approche de p_j et donc $D(P_n, P_0)$ est très petite.

- La région critique est de la forme $W = \{D(P_n, P_0) \geq k\}$,

k est tel que le test soit de niveau α.

Theorem 62

On suppose que $p_j > 0$, alors sous H_0

$$D(P_n, P_0) \xrightarrow{\mathcal{L}} \mathcal{X}^2(d-1).$$

Démonstration. Soit $(e_1, ..., e_d)$ la base canonique de $\mathbb{R}^d$ et posons $Y_m = \sum_{j=1}^d e_j \mathbf{1}_{\{X_m = x_j\}}$; les v.a. sont i.i.d. de moyenne $E(Y_1) = {}^t(p_1, ..., p_d)$ et de matrice de covariance $\operatorname{cov}(Y_1)_{i,j} = (p_i \delta_i^j - p_i p_j)$, de plus $\sum_1^n Y_m = {}^t(N_1, ..., N_d)$, et $Z_n = {}^t(\frac{1}{\sqrt{n}}(N_1 - np_1), ..., \frac{1}{\sqrt{n}}(N_d - np_d)) = \frac{1}{\sqrt{n}}(\sum_1^n Y_m - nE(Y_1))$; d'après le théorème central limite on a

$$Z_n \xrightarrow{\mathcal{L}} Z \rightsquigarrow N(0, \operatorname{cov}(Y_1));$$

de plus $D(P_n, P_0) = f(Z_n)$ avec $f(z_1, ..., z_d) = \sum_1^d z_j^2 / p_j$ donc continue et par conséquent $D(P_n, P_0) \xrightarrow{\mathcal{L}} f(Z)$.

Posons $u = {}^t(\sqrt{p_1}, ..., \sqrt{p_d})$, et soient $v_1, ..., v_{d-1}$ des vecteurs tels que $(v_1, ..., v_{d-1}, u)$ soit une base orthonormale de $\mathbb{R}^d$, soit H la matrice (d, d) définie par $H = \begin{pmatrix} {}^t v_1 \\ \vdots \\ {}^t v_{d-1} \\ {}^t u \end{pmatrix}$, alors $H\,{}^tH = I_d$ (H est orthogonale) et $Hu = e_d$.

Posons $\zeta = H^{\,t}(Z_1/\sqrt{p_1}, ..., Z_d/\sqrt{p_d})$, il est facile de vérifier que

$$\zeta \rightsquigarrow N(0, H(\delta_i^j - \sqrt{p_i p_j})^{\,t}H) = N(0, \begin{pmatrix} I_{d-1} & 0 \\ 0 & 0 \end{pmatrix}),$$

ceci entraîne que $\zeta_i \rightsquigarrow N(0,1)$, les ζ_i sont indépendantes pour $i = 1, ..., d-1$, et $\zeta_d = 0$ est dégénérée.

Finalement

$$f(Z) = ||\,^t(Z_1/\sqrt{p_1}, ..., Z_d/\sqrt{p_d})||^2 = ||\zeta||^2 = \sum_1^{d-1} \zeta_i^2 \rightsquigarrow \mathcal{X}^2(d-1). \square$$

Exemples 1. On a lancé 100 fois une pièce de monnaie, on a obtenu 40 fois pile. Peut-on dire que la pièce est truquée ?

Commandes sur R

piece=c(40,60), H0=p(1/2,1/2), chisq.test(x=piece, y=NULL, p=H0)

Chi-squared test for given probabilities

data: piece

X-squared = 4, df = 1, p-value = 0.0455.

Donc à 5% on rejette H0, et donc la pièce est truquée.

Exemples 2. On a lancé 100 fois un dé, on a obtenu :

face	1	2	3	4	5	6
Nombre de faces	25	16	12	15	13	19

Peut-on dire que le dé est pipé?

Commandes sur R

de=c(25,16,12,15,13,19), H0=p(1/6,1/6, 1/6,1/6,1/6,1/6),
chisq.test(x=piece, y=NULL, p=H0)

Chi-squared test for given probabilities

data: de X-squared = 6.8, df = 5, p-value = 0.2359 Donc à 5% on ne rejette pas H0, et donc le dé n'est pas pipé.

Contenu

- Soit P une loi de probabilité d'un couple aléatoire (X, Y) à valeurs dans $\mathbb{R}^2$, on veut tester l'indépendance de X et Y en se basant sur un échantillon $(X_i, Y_i)_{1 \leq i \leq n}$.
- Ceci revient à tester $H_0 : P(dx, dy) = P_1(dx)P_2(dy)$ contre $H_1 : P(dx, dy) \neq P_1(dx)P_2(dy)$; où P_1 et P_2 sont respectivement les lois marginales de X et Y.

- On construit alors une partition uniforme de l'ensemble des valeurs prises par X : $A_1, ..., A_r$, et de l'ensemble des valeurs prises par Y : $B_1, ..., B_s$, (uniforme veut dire que $P_1(A_j) \approx 1/r, P_2(B_j) \approx 1/s$). • On définit alors

$$N_{i,j} = \sum_{k=1}^{n} \mathbf{1}_{\left\{(X_k, Y_k) \in A_i \times B_j\right\}}.$$

- On pose

$$d_n = \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{(N_{i,j} - np_{i,j})^2}{np_{i,j}};$$

comme dans le théorème 62.

- On montre que

- En pratique les probabilités marginales sont inconnues et donc on les estime par les estimateurs empiriques

$$\widehat{p}_{i,\bullet} = \frac{1}{n}\sum_{j=1}^{s} N_{i,j}, \widehat{p}_{\bullet,j} = \frac{1}{n}\sum_{i=1}^{r} N_{i,j}.$$

- On montre alors que

$$\widehat{d}_n = \sum_{i=1}^{r}\sum_{j=1}^{s} \frac{(N_{i,j} - n\widehat{p}_{i,\bullet}\ \widehat{p}_{\bullet,j})^2}{n\ \widehat{p}_{i,\bullet}\ \widehat{p}_{\bullet,j}} \xrightarrow{\mathcal{L}} \mathcal{X}^2((r-1)(s-1)).$$

Exemple. Dans un atelier les 157 pannes de machines dénombrés au cours d'un trimestre se répartissent comme suit, selon l'indentité des machines et les équipes aux postes de travail au moment des pannes:

Machines / *Postes de travail*	M1	M2	M3	M4
Equipe1 (matin)	10	6	12	13
Equipe2 (soir)	10	12	19	21
Equipe3 (nuit)	13	10	13	18

Tester l'hypothèse selon laquelle le nombre de pannes des machines ne dépend pas de l'équipe.

Commandes sur R

```
atelier= matrix(c(10,10,13,6,12,10,12,19,13,13,21,18), 3,4,
dimnames=list( c("Equipe 1", "Equipe 2", "Equipe 3"),c(
"Machine 1", "Machine 2", "Machine 3", "Machine 4") ))
chisq.test(atelier)
```

Pearson's Chi-squared test

data: atelier

X-squared = 2.0081, df = 6, p-value = 0.919

Conclusion : On ne rejette pas l'hypothèse d'indépendance entre les variables "équipe" et "nombre de pannes des machines".

Contenu

- Le test de Kolmogorov-Smirnov est un test non paramétrique qui porte sur la fonction de répartition de la loi testée.
- C'est à dire qu'en se basant sur un échantillon $(X_i)_{1\leq i\leq n}$ de fonction de répartition inconnue F, on teste $H_0 : F = F_0$ contre $H_1 : F \neq F_0$.

Definition 63

On définit la fonction de **répartition empirique** de $(X_i)_{1\leq i\leq n}$, l'application $F_n^* : \mathbb{R} \to [0,1]$ définie par

$$F_n^*(x) = \frac{1}{n}\sum_{j=1}^{n} \mathbf{1}_{\{X_j\leq x\}}.$$

- On définit les statistiques de Kolmogorov-Smirnov suivantes :

$$K_n = \sup_{x\in\mathbb{R}} \left| F_n^*(x) - F_0(x) \right|, K_n^+ = \sup_{x\in\mathbb{R}} (F_n^*(x) - F_0(x)),$$

$$K_n^- = \sup_{x\in\mathbb{R}} (F_0(x) - F_n^*(x)).$$

Theorem 64

(Glivenko-Cantelli)
On suppose que $(X_i)_{1\leq i\leq n}$ est une suite i.i.d. de fonction de répartition $F(x)$, alors sous H_0, on a

$$K_n \stackrel{p.s.}{\rightarrow} 0.$$

Démonstration. Soit $x \in \mathbb{R}$ fixé, alors $Y_j = \mathbf{1}_{\{X_j<x\}}$ et $Z_j = \mathbf{1}_{\{X_j\leq x\}}, j \geq 1$ sont deux suites de variables aléatoires distribuées selon une loi de Bernoulli de paramètre $p = E(\mathbf{1}_{\{X_j<x\}}) = P(X_j < x) = F_0(x)$ et $p = F_0(x^+)$ respectivement sous H_0.

Par application de la loi forte des grands nombres $F_n^*(x) = \frac{1}{n}\sum_1^n Y_j \overset{p.s.}{\to} F_0(x)$, et $F_n^*(x^+) = \frac{1}{n}\sum_1^n Z_j \overset{p.s.}{\to} F_0(x^+)$.

Soit $N \in \mathbb{N}$, posons $x_{k,N} = \sup\left\{x \in \mathbb{R};\ \text{tel que } F_0(x) \leq \frac{k}{N}\right\}, 1 \leq k \leq N-1, x_{0,N} = -\infty, x_{N,N} = +\infty.$

- Soit $k \leq N$; tel que $]x_{k,N}, x_{k+1,N}] \neq \emptyset$, et soit $x \in]x_{k,N}, x_{k+1,N}]$, alors

$$\begin{aligned} F_n^*(x) - F_0(x) &\leq F_n^*(x_{k+1,N}) - F_0(x_{k+1,N}) \\ &+ F_0(x_{k+1,N}) - F_n^*(x_{k,N}^+) \\ &\leq F_n^*(x_{k+1,N}) - F_0(x_{k+1,N}) + \frac{1}{N}. \end{aligned}$$

- On montre aussi que

$$F_n^*(x) - F_0(x) \geq F_n^*(x_{k,N}^+) - F_0(x_{k,N}^+) - \frac{1}{N};$$

ceci implique que

$$\begin{aligned} K_n = \sup_{x \in \mathbb{R}} |F_n^*(x) - F_0(x)| &\leq \max_k |F_n^*(x) - F_0(x)| \\ &+ \max_k \left| F_n^*(x_{k,N}^+) - F_0(x_{k,N}^+) \right| + \frac{1}{N}; \end{aligned}$$

et donc

$$\lim_{n \to +\infty} \sup_{x \in \mathbb{R}} |F_n^*(x) - F_0(x)| \leq \frac{1}{N}.\Box$$

- Les deux théorèmes suivants sont admis (voir la démonstration dans Van Der Waerden).

Theorem 65

On suppose que $(X_i)_{1\leq i\leq n}$, est une suite i.i.d. de fonction de répartition $F(x)$, que $F_0(x)$ est continue, alors sous H_0

$$\sqrt{n}K_n \xrightarrow{\mathcal{L}} Z;$$

la fonction de répartition de Z est donnée par

$$F_Z(x) = 1 - \sum_{k=1}^{\infty}(-1)^{k-1}e^{-2k^2x^2}.$$

Theorem 66

On suppose que $(X_i)_{1\leq i\leq n}$, est une suite i.i.d. de fonction de répartition $F(x)$, que $F_0(x)$ est continue, alors sous H_0

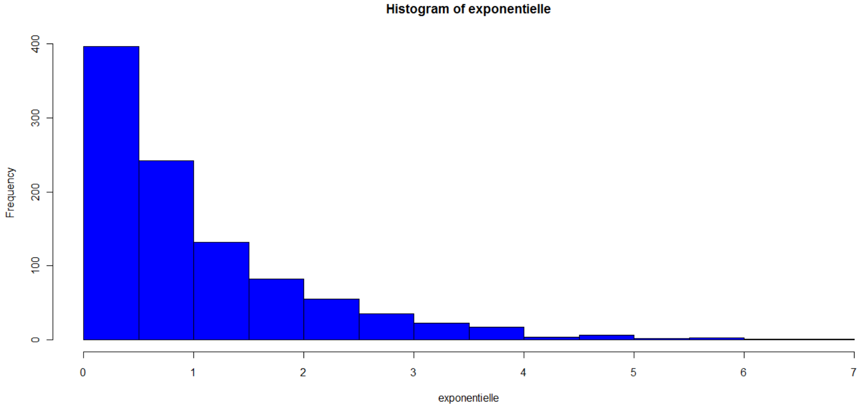
$$\sqrt{n}K_n^+ \xrightarrow{\mathcal{L}} Z^+;$$

la fonction de répartition de Z^+ est donnée par
$F_{Z^+}(x) = 1 - e^{-2x^2}$.

Exemple 1. On génére une variable selon une loi exponentielle de paramètre 1, et on teste sa normalité.

Commandes sur R

exponentielle=rexp(1000), libray(nortest), hist(exponentielle, col="blue"), lillie.test(exponentielle)


Histogram of exponentielle
Frequency
0
100
200
300
400
0
1
2
3
4
5
6
7
exponentielle

Lilliefors (Kolmogorov-Smirnov) normality test

data: exponentielle

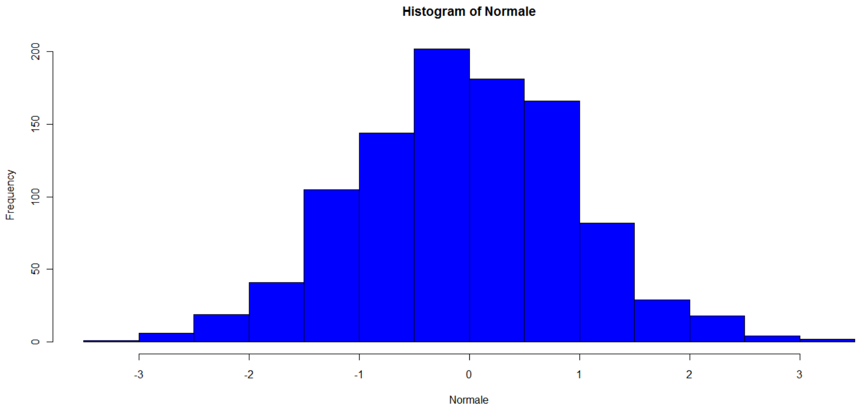
D = 0.1552, p-value < 2.2e-16

Conclusion : On rejette fortement la normalité.

Exemple 2. On génére une variable selon une loi Gauss de moyenne 0 et de variance 1, et on teste sa normalité.

Commandes sur R

Normale=rnorm(1000), libray(nortest), hist(Normale, col="blue"), lillie.test(Normale)


Histogram of Normale
Frequency
0
50
100
150
200
-3
-2
-1
0
1
2
3
Normale

Lilliefors (Kolmogorov-Smirnov) normality test

data: Normale

D = 0.0167, p-value = 0.7127

Conclusion : On ne rejette pas la normalité.

Contenu

- Ils permettent de comparer deux échantillons $(X_i)_{1\leq i\leq n}$ et $(Y_i)_{1\leq i\leq m}$, et décider s'ils proviennent de la même loi ou pas, donc on effectue le test $H_0 : F = G$ contre $H_1 : F \neq G$, avec F est la fonction de répartition de $(X_i)_{1\leq i\leq n}$ et G celle de $(Y_i)_{1\leq i\leq m}$.

Contenu

- Il est basé sur la statistique

$$D(F_n^*, G_m^*) = sup_{x\in\mathbb{R}} \left|F_n^*(x) - G_m^*(x)\right|.$$

Theorem 67

On suppose que $(X_i)_{1\leq i\leq n}$, et $(Y_i)_{1\leq i\leq m}$ sont deux suites i.i.d. et sont indépendantes. Alors sous H_0

$$\sqrt{\frac{nm}{n+m}}D(F_n^*, G_m^*) \quad \xrightarrow[n,m\to+\infty]{\mathcal{L}} \quad D;$$

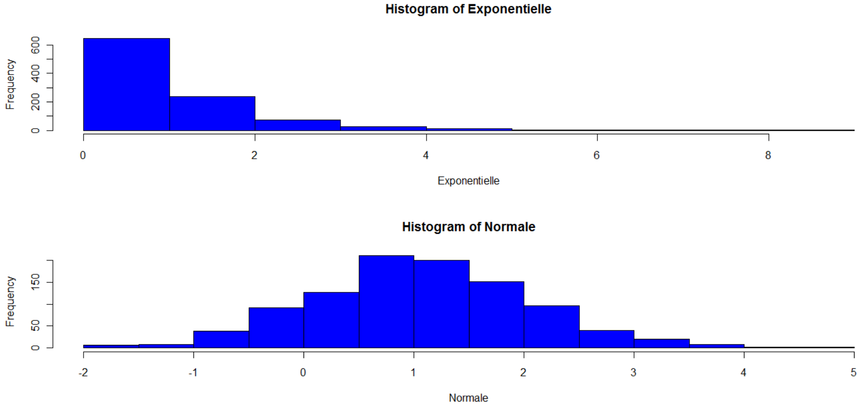
la fonction de répartition de D est donnée par

$$F_D(x) = 1 - 2\sum^{\infty}(-1)^{k-1}e^{-2k^2x^2}.$$

Exemple 1. On génére deux échantillons. La premier suit une loi exponentielle de moyenne 1. Le deuxième suit une loi Gauss de moyenne 1 et de variance 1, et on teste si les deux échantillons sont issus de la même loi

Commandes sur R

Expontielle= rexp(1000), Normale=rnorm(1000,1,1), ks.test(Normale)


Histogram of Exponentielle
Frequency
0
200
400
600
0
2
4
6
8
Exponentielle
Histogram of Normale
Frequency
0
50
150
-2
-1
0
1
2
3
4
5
Normale

Two-sample Kolmogorov-Smirnov test

data: Normale and Exponentielle

D = 0.526, p-value $<$ 2.2e-16

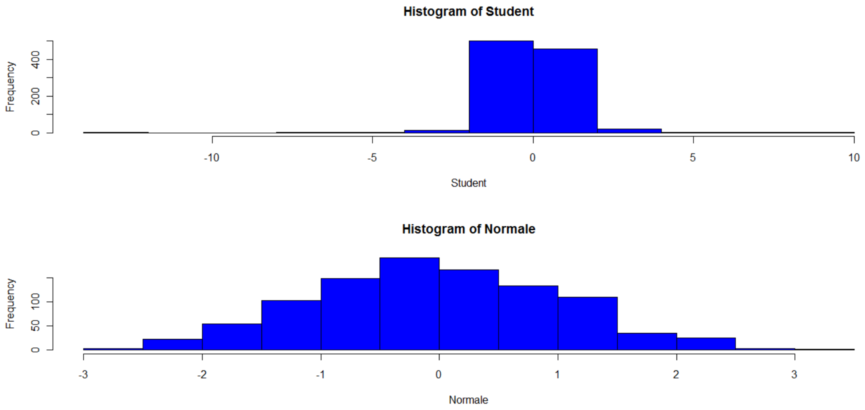
alternative hypothesis: two-sided

Conclusion: Les deux échantillons ne sont pas issu de la même loi.

Exemple 2. On génére deux échantillons. La premier suit une loi Student degré de liberté 2. Le deuxième suit une loi Gauss de moyenne 1 et de variance 1, et on teste si les deux échantillons sont issus de la même loi

Commandes sur R

Student= rt(1000,2), Normale=rnorm(1000), ks.test(Normale)


Histogram of Student
Frequency
0
200
400
-10
-5
0
5
10
Student
Histogram of Normale
Frequency
0
50
100
-3
-2
-1
0
1
2
3
Normale

Two-sample Kolmogorov-Smirnov test

data: Normale and Student

D = 0.199, p-value < 2.2e-16

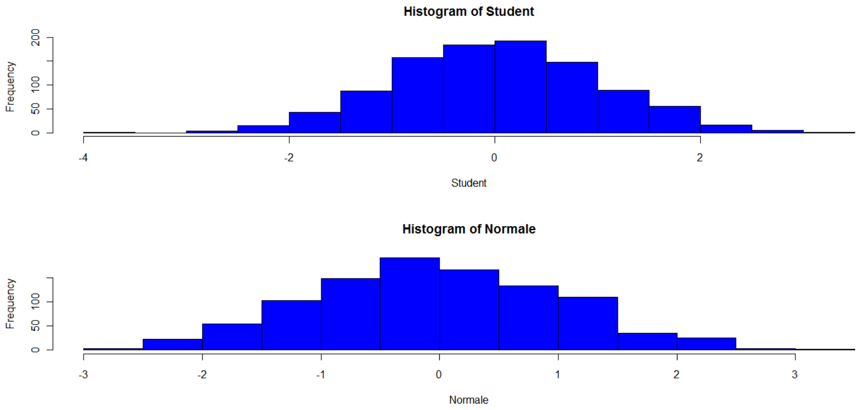
alternative hypothesis: two-sided

Conclusion: Les deux échantillons ne sont pas issus de la même loi.

Exemple 3. On génére deux échantillons. La premier suit une loi Student de degré de liberté 100. Le deuxième suit une loi Gauss de moyenne 1 et de variance 1, et on teste si les deux échantillons sont issus de la même loi

Commandes sur R

Student= rt(1000,100), Normale=rnorm(1000),
ks.test(Normale)


Histogram of Student
Frequency
0
50
100
200
-4
-2
0
2
Student
Histogram of Normale
Frequency
0
50
100
-3
-2
-1
0
1
2
3
Normale

Two-sample Kolmogorov-Smirnov test

data: Normale and Student

D = 0.047, p-value = 0.2193

alternative hypothesis: two-sided

Conclusion: Les deux échantillons sont issus de la même loi.

Contenu

- Il est basé sur la statistique

$$U_{n,m} = \sum_{i=1}^{n} \sum_{j=1}^{m} \mathbf{1}_{\{X_i \leq Y_j\}}.$$

Theorem 68

Sous les mêmes hypothèses que le théorème 67, on a sous H_0

$$\frac{U_{n,m} - \frac{nm}{2}}{\sqrt{nm(n+m+1)/12}} \underset{n,m \to +\infty}{\overset{\mathcal{L}}{\to}} \qquad U \rightsquigarrow N(0,1).$$

Pour les 3 exemples précédents on obtient:

1. " Mann Whitney Wilcoxon Test"

data: Normale and Exponentielle

"Alternative Hypothesis: the samples do not have the same distribution"

"pvalue:" 0.0009879936.

2. " Mann Whitney Wilcoxon Test"

data: Normale and Student

"Alternative Hypothesis: the samples do not have the same distribution"

"pvalue:" 0.851282.

3. " Mann Whitney Wilcoxon Test"

data: Normale and Student

"Alternative Hypothesis: the samples do not have the same distribution"

"pvalue:" 0.9129897.