Statistiques non paramétriques

Rappels

Proposition (Inégalité de Hoffding). *Soient $Y_1, \dots, Y_n$ indépendants, $\forall i a_i \leq Y_i \leq b_i$ p.s.*

$$\mathbb{P}(\sum_{i=1}^{n}(Y_i - \mathbb{E}[Y_i]) > \lambda) \leq e^{-\frac{2\lambda^2}{\sum_{i=1}^{n}(b_i - a_i)}}$$

Proposition (Inégalité de Mill).

$$\mathbb{P}(|Z| > t) \leq \frac{1}{t}\sqrt{\frac{2}{\pi}}e^{-\frac{t^2}{2}}$$

Tests non paramètriques

Test du signe

Soient $U_1, \dots, U_n$ et $V_1, \dots, V_n$ deux échantillons de même tailles. $X_i = U_i - V_i$ et on suppose
$\forall x \in \mathbb{R}\mathbb{P}(U_i = x) = \mathbb{P}(V_j = x) = 0$ (Propriété de diffusivité)

Proposition. *Sous $H_0 = U$ et V ont même distribution la distribution des X_i est symétrique i.e*
$\mathbb{P}(X_i \leq 0) = \mathbb{P}(X_i \geq 0) = \frac{1}{2}$.
La statistique de test est :

$$h(X_1, \dots, X_n) = \sum_{i=1}^{n} \mathbb{1}_{X_i \leq 0} \sim \text{Bin}(n, \frac{1}{2})$$

Le test est alors : $\mathbb{P}(\phi(X) = 1) = \mathbb{P}(|\sum_{i=1}^{n} \mathbb{1}_{X_i \leq 0} - \frac{n}{2}| > k)$ où k le plus petit possible tel que $\mathbb{P}(\phi(X) = 1) \leq \alpha$

Test de Wilcoxon

Soit $R_{|X|}$ le vecteur rang associé à $|X|$.

$$W_n^+ = \sum_{i=1}^{n} R_{|X|}(i)\mathbb{1}_{X_i > 0}$$

Théorème. *Sous H_0 et l'hypothése de diffusivité :*

- *W_n^+ et W_n^- ont la même distribution et leur loi ne dépend pas de la loi de X*
- $\mathbb{E}[W_n^+] = \frac{n(n+1)}{4}$ *et* $\text{Var}[W_n^+] = \frac{n(n+1)(2n+1)}{24}$
- $\frac{W_n^+ - \mathbb{E}[W_n^+]}{\sqrt{\text{Var}[W_n^+]}} \to \mathcal{N}(0,1)$

Test de Mann-Whitney

Soient U et V deux échantillons de loi diffuses,
$|U| = n \neq p = |V|$, F la fonction de repartition de U et G celle de V.
Soit $H_0 : F = G$.
$U_1, \dots, U_n \to R_1, \dots, R_n$ vecteurs des rangs dans $(U_1, \dots, U_n, V_1, \dots, V_p)$ et $V_1, \dots, V_n \to S_1, \dots, S_n$

- $\Sigma_1 = R_1 + \cdots + R_n$
- $\Sigma_2 = S_1 + \cdots + S_n$
- $\frac{n(n+1)}{2} \leq \Sigma_1 \leq np + \frac{n(n+1)}{2}$
- $\frac{p(p+1)}{2} \leq \Sigma_2 \leq np + \frac{p(p+1)}{2}$
- Sous H_0, $\mathbb{E}[R_i] = \mathbb{E}[S_j] = \frac{n+p+1}{2}$
- Sous H_0, $\text{Var}[R_i] = \text{Var}[S_j] = \frac{(n+p)^2-1}{2}$
- Sous H_0, $\mathbb{E}[\Sigma_1] = \frac{n(n+p+1)}{2}$ et $\mathbb{E}[\Sigma_2] = \frac{p(n+p+1)}{2}$
- Sous H_0, $R_i \sim S_j \sim \mathcal{U}(1, \dots, n+p)$

Proposition. *Soient $W_X = \Sigma_1 - \frac{n(n+1)}{2}$ et $W_Y = \Sigma_2 - \frac{p(p+1)}{2}$*

1. *$W_Y =$ nombre de paires $(U_i, V_j)/U_i < V_j$*
2. *$W_X + W_Y = np$*
3. *Sous H_0 ; W_X et W_Y symetriques par rapport à $\frac{np}{2}$*
4. *Sous H_0 ; $W_X \sim W_Y$*

Théorème. *Les lois de W_X et W_Y ne dépendent pas des U_i et V_j sous H_0.*

$$\frac{W_X - \mathbb{E}[W_X]}{\sqrt{\text{Var}W_X}} \to \mathcal{N}(0,1)$$

Estimation de la fonction de répartition

Proposition. *Soit $F_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{X_i \leq x}$,*
$\forall x \in \mathbb{R}, nF_n(x) \sim \text{Bin}(n, F(x))$

$$\forall x \in \mathbb{R}, \sqrt{n}(F_n(x) - F(x)) \to \mathcal{N}(0, F(x)(1 - F(x)))$$

Théorème (Glivenko-Cantelli).

$$\lim_{n\to\infty} \|F_n - F\| = 0\text{p.s}$$

Définition (Inverse généralisée).

$$\forall q \in [0,1]; F^{(-1)}(q) = \inf\{x \in \mathbb{R}; F(x) \geq q\}$$

Proposition.

- *Si F inversible, $F^{(-1)} = F^{-1}$*
- *$F^{(-1)}$ croissante*
- *$\forall x \in \mathbb{R}; F(x) \geq q \Leftrightarrow x \geq F^{(-1)}(q)$*
- *Si $U \sim \mathcal{U}(0,1)$ alors $F^{(-1)}(u) \sim$ est de fonction de répartition F*
- *Si F admet F comme fonction de répartition alors $F(Z) \sim \mathcal{U}(0,1)$*

Estimation des quantiles

Définition (Quantile d'ordre β). *Pour $\beta \in [0,1]$ on appelle quantile d'ordre β : $q_\beta = F^{(-1)(\beta)}$*

Définition (Quantile empirique). $\hat{q}_{\beta;n} = X_{\lceil n\beta \rceil}$ *où* $\lceil u \rceil = \min\{n \in \mathbb{N}; n \geq u\}$
On remarque que $\hat{q}_{\beta;n} = F_n^{(-1)}(\beta)$

Théorème. *Soit $\beta \in]0,1[$. On suppose que F est strictement croissante au voisinage de q_β alors : $\hat{q}_{n,\beta} \to q_\beta$ p.s.*

Test d'ajustement à une loi ou famille de loi

Test de Kolmogorov

$h_n(X; F_0) = \|F_n - F_0\|_\infty$
Objectif : test de $F = F_0$ contre $F \neq F_0$

Théorème.

- *Si F_0 continue, $\exists \xi_{n,\alpha}$ ne dépendant que de n et α tel que sous H_0, $\mathbb{P}(h_n(X, F_0) \geq \xi_{n,\alpha} = \alpha$*
- *Si F_0 pas continu sous H_0, il y a inégalité.*

Soit la bande de confiance :

$$B(n,\alpha) = \{G c.d.f \mid \|F_n - G\|_\infty \leq \xi_{n,\alpha}\}$$

Théorème.

$$\mathbb{P}(F \in B(n,\alpha)) \geq 1 - \alpha$$

Ajustement à la famille exponentielle

Soit $h_n(X, F_0) = \|F_n - F_{\hat{\lambda}}\|_\infty$ avec $\hat{\lambda_n} = \frac{1}{\bar{X_n}}$ l'e.m.v.

Théorème. *Sous H_0 la loi de $h_n(X)$ est libre de λ.*

Test d'homogénéité de Kolmogorov-Smirnov

On observe $X = (X_1, \dots, X_n)$ et $Y = (Y_1, \dots, Y_m)$ de c.d.f F et G. Soit $H_0 = \{F = G\}$ et $H_1 = \bar{H_0}$
Soit $h_{n,m}(X, Y) = \|F_n - G_m\|_\infty$

Théorème. *Sous H_0 ; $H_{n,m}(X)$ ne dépend pas de F et G si elles sont continues.*

Estimateurs à noyau

Risque quadratique ponctuel

On appelle éspace de Hölder :
$\Sigma(\beta, L) = \left\{f | |f^{\lceil\beta\rceil}(x) - f^{\lceil\beta\rceil}(y)| < L|x-y|^{\lceil\beta\rceil - \beta}\right\}$
Soit $R(\widehat{(f)}_n, f) = \mathbb{E}[\|\widehat{(f)}_n(x) - f(x)\|]$

Définition. *On appelle estimateur à noyau de f :*

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

avec K tel que $\|K\| = 1$ et K symmetrique.

Proposition. *Si f est bornée alors* $\text{Var}(\hat{f}_n(x)) \leq \frac{\|f\|_\infty \|K\|_2}{nh}$

Définition. *Soit $l \in \mathbb{N}^*$. On dit que K est d'ordre l si :*

- *$u \to u^j K(u) \in \mathcal{L}^1 \ \forall j \in [|1, l|]$*
- *$\int_{-\infty}^{+\infty} u^j K(u)\,\text{d}u = 0$*

Proposition. *Si $f \in \Sigma(\beta, L)$ avec $\beta > 0$, $L > 0$ et K noyau d'ordre $l = \lfloor\beta\rfloor$ tel que $\int_{-nfty}^{+\infty} |u|^\beta |K(u)|\,\text{d}u < \infty$ alors :*

$$|\mathbb{E}[\hat{f}_n(x) - f(x)]| \leq \frac{Lh^\beta}{l!}\int_{-\infty}^{+\infty} |u|^\beta |K(u)|\,\text{d}u$$

Théorème. *Soit K noyau d'ordre $\lceil\beta\rceil$ tel que $\|K\|_{L^2} < \infty$ et $\|u^\beta K(u)\|_{L^1} < \infty$. Alors en choisissant $h = Cn^{-\frac{1}{2\beta-1}}$ on obtient $R(\hat{f}_n(x_0) = \sup_{f\in\Sigma(\beta,L)} \mathbb{E}[|\hat{f}_n(x_0) - f(x_0)|^2] \leq Cn^{-\frac{2\beta}{2\beta-1}}$*

Proposition. *Soit (Φ_n) bases des polynomes de Legendres. $K : u \to \sum_{n=0}^{l} \Phi_n(0)\Phi_n(u)\mathbb{1}_{|u|\leq 1}$ noyau d'ordre l*

Théorème. *Soit*

$$\hat{R} = \|\hat{f}_n\|_2^2 - \frac{2}{n(n-1)} \sum_{i=1}^{n} \sum_{j=1, j\neq i}^{n} \frac{1}{h} K(\frac{X_i - X_j}{h})$$

$$\mathbb{E}[\hat{R}] = R - \|\hat{f}\|_2^2$$

Pour choisir h on minimise $\hat{R}$.

Régression non paramétrique

Théorème (Estimateur de Nadaraya-Watson)**.** *On pose :*

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

$$\hat{f}(x, y) = \frac{1}{nh^2} \sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right) K\left(\frac{Y_i - x}{h}\right)$$

$$\hat{r}(x) = \int y \frac{\hat{f}(x, y)}{\hat{f}_X(x)} \, \mathrm{d}y$$

Proposition. *Si K d'ordre 1 :*

$$\hat{r}(x) = \frac{\sum_{i=1}^{n} Y_i K\left(\frac{X_i - x}{h}\right)}{\sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right)}$$