01.12.2025

lundi, 1 décembre 2025 08:13

Rappel : Estimation paramétrique : $\Theta \subset \mathbb{R}^m$ ens. de paramètres. $X_1, \dots, X_n$ échantillon de loi $\mathbb{P}_\theta$ $\theta \in \Theta$, $(\mathbb{P}_\theta)_{\theta \in \Theta}$ une familles de lois.

Intervalles de confiance : Soit $f(\theta)$ une quantité à estimer, $\hat{f}_n(X_1, \dots, X_n)$ un estimateur de $f(\theta)$, $[\hat{f}_n - a, b + \hat{f}_n]$ est un intervalle de confiance pour $f(\theta)$ au niveau $1-\alpha$ si

$$\mathbb{P}\left(f(\theta) \in [\hat{f}_n - a, \hat{f}_n + b] \right) \overset{(\geq)}{=} 1 - \alpha .$$

$\rightsquigarrow$ $f(\theta)$ est dans l'intervalle $[\hat{f}_n - a, \hat{f}_n + b]$ proba $1-\alpha$.

On peut plus généralement prendre un intervalle $I(X_1, \dots, X_n)$ et demander

$$\inf_{\theta \in \Theta} \mathbb{P}\left(f(\theta) \in I(X_1, \dots, X_n) \right) \overset{(\geq)}{=} 1 - \alpha .$$

Test d'hypothèses :

Exemple : 2 fabriquants de tasseaux.

offre 1 : on ne connait pas la [illegible]

offre 1 : on ne connait pas la qualité, mais l'offre est peu chère.

offre 2 : On sait qualité ok, mais plus cher.

Question : faut-il changer de fournisseur ?

On estimer la qualité des tuyaux de 1 par rapport à la qualité de 2.

On prend 500 tuyaux de chaque fabriquant et on les teste.

Fabriquant 1 : $X_1, \dots, X_{500} \to \begin{cases} 0 \text{ si tuyau } i \text{ survit} \\ 1 \text{ si } - \text{ " } - \text{ casse.} \end{cases}$

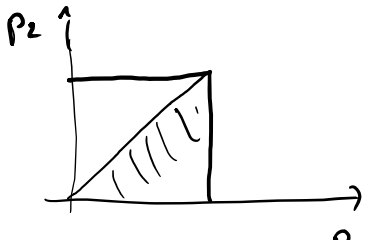
" " 2 : $Y_1, \dots, Y_{500}$ " " "

En proba : . $X_1, \dots, X_{500}$ est un échantillon de loi $\text{Bern}(p_1)$

. $Y_1, \dots, Y_{500}$ " " " loi $\text{Bern}(p_2)$

. $\Theta = [0,1]^2$, $\theta = (p_1, p_2)$

But : déterminer si $p_1 \geq p_2$ ou non.



$\Theta_0 = \{(x,y) \in \Theta : \; x > y\}$

$\Theta_1 = \Theta \setminus \Theta_0 = \{(x,y) \in \Theta : \; x \leq y\}$.

Θ_0 : fab 1 est moins sûre que 2

Θ_1 : fab 2 est au moins aussi sûre que 1

Procédure de teste : $\hat{\Theta} = (\overline{X}_{500}, \overline{Y}_{500})$ est un est. de (p_1, p_2)

↳ sur une réalisation $x_1, \dots, x_{500}, y_1, \dots, y_{500}$, on décide de dire que 1 est ok si $\overline{x}_{500} \leq \overline{y}_{500}$.

En faisant ceci, on prend 2 risques :

- Erreur de type 1 : on dit que 1 ok alors que non. ↝ dire $\Theta \in \Theta_1$ alors que $\Theta \in \Theta_0$
- Erreur de type 2 : on dit que 1 pas ok alors que si : ↝ dire $\Theta \in \Theta_0$ alors que $\Theta \in \Theta_1$.

Mathématiquement, type 1 et type 2 sont semblaire.

Dans la pratique, non : dans cmd.

Dans la pratique, non : dans espl :

erreur de type 1 : ⇒ on achète des fusées moins sûrs

erreur de type 2 : ⇒ on rate une occasion d'économie.

Principe général : On pose les problèmes de sorte à ce que les erreurs de type 1 soit les plus graves.

Cadre abstrait : Donnée : $\circ$ $X_1, \dots, X_n$ un échantillon $: \Omega \to \mathbb{R}^d$ de loi $\mathbb{P}_\theta$, $\theta \in \Theta$.

$\circ$ $\Theta_0 \subset \Theta$, $\Theta_1 = \Theta \setminus \Theta_0$.

But : déterminer si $\theta \in \Theta_0$ ou $\theta \in \Theta_1$.

Déf : l'hypothèse "$\theta \in \Theta_0$" est appelée hypothèse nulle, on la note H_0. L'hypothèse "$\theta \in \Theta_1$" est appelée hypothèse alternative.

Déf : Un événement D est appelé région de rejet si D ne dépend que de $X_1, \dots, X_n$ et

si D ne dépend que de $X_1, \dots, X_n$ et pas de θ.

Très souvent, D sera donné par :

i) On prend une statistique $T : (\mathbb{R}^d)^n \to \mathbb{R}$

ii) On prend comme région de rejet

$$\{ T(X_1, \dots, X_n) \geq C \} \quad \text{pour un } C \in \mathbb{R}.$$

Déf: Soit D une région de rejet, la procédure de teste associée à D est :

1. On rejette H_0 si D se produit (p.ex: si $T(X_1, \dots, X_n) \geq C$)

2. On ne rejette pas H_0 si D ne se produit pas.

Les erreurs de prédiction sont de deux types:

- type I : on rejette H_0 alors qu'elle était correcte.

- type II : on ne rejette pas H_0 alors qu'elle n'était pas correcte.

Déf: Soit $\alpha \in [0,1]$. On dit que la procédure de test associée à la région de rejet

test associée à la région de rejet $D(X_1, \dots, X_n)$ a un <u>niveau de risque α</u> si

$$\sup_{\theta \in \Theta_0} \mathbb{P}\left(D(X_1, \dots, X_n) \right) = \alpha$$

⚠ $X_i \sim \mathbb{P}_\theta$. Alternativement, on dit que la procédure de test a un niveau de confiance $1-\alpha$.

Déf: La <u>puissance</u> d'un test associé à D est

$$\inf_{\theta \in \Theta_1} \mathbb{P}\left(D(X_1, \dots, X_n) \right) =: 1 - \beta$$

β est la probabilité d'erreur de type II.

↳ $\beta = \sup_{\theta \in \Theta_1} \mathbb{P}\left((D(X_1, \dots, X_n))^c \right)$.

<u>Exemple</u> (Expl 4.5.1) On veut faire un teste d'hypothèse sur l'exemple "US v.s. Brit. TV".

<u>Setup</u>: • $X_1, \dots, X_{100}$ un échantillon de temps TV US.

• $Y_1, \dots, Y_{100}$ —"— Brit.

<u>Supposition</u>: $X_1 \sim \mathcal{N}(\mu_X, 223)$

Supposition: $X_1 \sim \mathcal{N}(\mu_X, 223)$

$Y_1 \sim \mathcal{N}(\mu_Y, 2028)$

$\Theta = \mathbb{R}^2$, $\theta = (\mu_X, \mu_Y)$.

On veut tester "$\mu_X = \mu_Y$" $\leftrightarrow$ H_0 est $\Theta_0 = \{(\mu,\mu) : \mathbb{R}\}$.

On va utiliser la region de rejet :

$$D(X_1, \dots X_{100}, Y_1, \dots Y_{100}) = \left\{ |\overline{X}_{100} - \overline{Y}_{100}| \geq \delta \right\}$$

Ici: Statistique de teste $T(X_1, \dots, X_{100}, Y_1, \dots, Y_{100})$

$$= |\overline{X}_{100} - \overline{Y}_{100}|.$$

On cherche δ de telle sorte à avoir un risque $\alpha = 0.05$.

$$\alpha = \sup_{\theta \in \Theta_0} \mathbb{P}\left(D(X_1, \dots, X_{100}, Y_1, \dots, Y_{100}) \right)$$

$$= \sup_{\substack{\mu_X, \mu_Y \in \mathbb{R}: \\ \mu_X = \mu_Y}} \mathbb{P}\left(|\overline{X}_{100} - \overline{Y}_{100}| \geq \delta \right)$$

On sait $\overline{X}_{100} \sim \mathcal{N}\left(\mu_X, \frac{223}{10}\right)$

$\overline{Y}_{100} \sim \mathcal{N}\left(\mu_Y, \frac{2028}{10}\right)$ et $\overline{Y}$ et $\overline{X}$ indép.

$\rightsquigarrow$ $\overline{X}_{100} - \overline{Y}_{100} \sim \mathcal{N}(\mu_X - \mu_Y, 225.1)$

Sous H_0, $\mu_X = \mu_Y \Rightarrow \bar{X}_{100} - \bar{Y}_{100} \sim \mathcal{N}(0, 225.1)$

Pour trouver δ, on pose

$$0.05 = \mathbb{P}\left(|\mathcal{N}(0, 225.1)| \geq \delta \right)$$

$$\overset{\text{sym } \mathcal{N}(0,\cdot)}{=} 2\,\mathbb{P}\left(\mathcal{N}(0, 225.1) \geq \delta \right)$$

$$= 2\,\mathbb{P}\left(\mathcal{N}(0,1) \geq \frac{\delta}{\sqrt{225.1}} \right)$$

$$= 2\left(0.5 - \mathbb{P}\left(\mathcal{N}(0,1) \in \left[0, \delta/\sqrt{225.1}\right] \right)\right)$$

↳ $$0.475 = \mathbb{P}\left(\mathcal{N}(0,1) \in \left[0, \delta/\sqrt{225.1}\right] \right)$$

on regarde les tables gaussiennes, et on trouve que $\frac{\delta}{\sqrt{225.1}} = 1.96$ marche

↳ $\delta = 1.96 \cdot \sqrt{225.1} \approx 29.4.$

on rejette H_0 si $|\bar{X}_{100} - \bar{Y}_{100}| \geq 29.4$

pour avoir un test avec risque 0.05.

X et Y indép

X et -Y sont indép

$\mathrm{Var}(X-Y) = \mathrm{Var}(X) + \mathrm{Var}(-Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$