

STATISTIQUES INFÉRENTELLES

ESTIMATION

I Vocabulaire de base

Def:

- Procédement d'échantillon = "sondage"
- Population, on appelle individu chaque élément de \mathcal{P} .
- Le tirage d'un individu est dit probabiliste si la probabilité qu'un individu de \mathcal{P} d'être tiré est connue.
- Le tirage est dit au hasard si chaque individu a la même probabilité d'être tiré.
- Un tirage au hasard est dit non exhaustif.
Parce que l'effectif de \mathcal{P} ne varie pas (ou très peu) au cours des tirages: la probabilité d'un individu d'être tiré reste constante.
 \hookrightarrow Dans la suite on se place dans ce cas.

Cadre probabiliste.

- Exp aléatoire $e =$ Tirage au hasard d'1 individu de \mathcal{F}
Notons w_i l'évén. élémentaire : "On obtient l'indiv i ".
- On définit X Pa va associée au caract C
 $(\Omega, \mathcal{F}, P) \rightarrow \Omega(X) = E$
 $X :$
 $w_i \mapsto c_i = x_i$
- Hypothèse : On connaît Pa Poi P_θ mais pas Pa param. Θ .
- Obs d'un échantillon de X de taille n :
 - c.a $e_n =$ Tirage au hasard nrm exhaustif de n indiv. de \mathcal{P}
 - On dit que les x_i (x_1, \dots, x_n) forment un échantillon indépendant et uniformément distribué (iid) de X
 $(x_1, \dots, x_n) = (X_1(w), \dots, X_n(w))$ pour un exp. w .
 - On cherche G à partir de (x_1, \dots, x_n) .
- On parle de modèle statistique pour la famille $(\Omega, \mathcal{F}, X, E, (P_\theta; \theta \in \Theta))$ définie

Critères

Def :

Soit (X_1, \dots, X_n) un échantillon iid de X , $(\alpha_1, \dots, \alpha_n)$ une réalisa^o de celles-ci. On appelle estimateur d'ordre n de θ , une statistique de (X_1, \dots, X_n) : $T_n = f(X_1, \dots, X_n)$ dont la réalisatio^o observée $f(\alpha_1, \dots, \alpha_n)$.

Def :

Une statistique de X est une vu $f(X_1, \dots, X_n)$ où f est une fct mesurable de n variables et (X_1, \dots, X_n) est un échantillon de X .

Props :

- T_n est un estimateur convergent cl^e θ Puisque la suite de var $(T_n)_{n \geq 1} \xrightarrow{P} \theta$ Puisque $n \rightarrow +\infty$.
- T_n est un estimateur fortement consistant cl^e θ Puisque la suite de var $(T_n)_{n \geq 1} \xrightarrow{P.S} \theta$ Puisque $n \rightarrow +\infty$.
- T_n est un estimateur sans biais de θ si $E(T_n) = \theta$, sinon on note $b_n(\theta) = (E(T_n) - \theta)$ le biais.
 - Ou asymptotiquement sans biais si $E(T_n) \xrightarrow{n \rightarrow +\infty} \theta$

Théorème :

Soit T_n un estimateur de θ :

(i) Si $E(T_n) \xrightarrow{n \rightarrow +\infty} \theta$ et $\text{Var}(T_n) \xrightarrow{n \rightarrow +\infty} 0$, alors

T_n est convergent.

(ii) Si T_n est sans biais et si $\text{Var}(T_n) \xrightarrow{n \rightarrow +\infty} 0$,
alors T_n est convergent.

Déf:

Si T_n et S_n sont deux estimateurs sans biais de θ
et si $\text{Var}(T_n) \leq \text{Var}(S_n)$, alors on dit que T_n est
plus efficace de S_n . IP aura la préférence.

Déf:

On appelle moyenne de l'échantillon (X_1, \dots, X_n) et on
note $\bar{X}_n : \Omega^n \rightarrow \mathbb{R}$ la r.v. : $\bar{X} = \frac{X_1 + \dots + X_n}{n}$

Prop:

- $E(\bar{X}) = m$, $\text{Var}(\bar{X}) = \frac{\sigma^2}{n}$ en notant $E(X) = m$

et $\text{Var}(X) = \sigma^2$

- \bar{X} est un estimateur sans biais et convergent de $\theta = E(X)$.

De plus si $E(X) \neq 0$ c'est le θ efficace des estimateurs sans
biais. Précis.

Dcf.

On appelle variance empirique de l'échantillon (X_1, \dots, X_n) et on note $S^2 : \Omega^n \rightarrow \mathbb{R}$ la va : $S^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2$

Prop:

$$\cdot \mathbb{E}(S^2) = \frac{n-1}{n} \sigma^2$$

$$\cdot \mathbb{V}(S^2) = \frac{\mu_4 - \sigma^4}{n} - 2 \frac{(\mu_2 - 2\sigma)^2}{n^2} + \frac{(\mu_2 - 3\sigma^2)}{n^3}$$

Prop: (2^{nde} df)

(i) Notons $\bar{S}^2: \mathbb{R}^n \rightarrow \mathbb{R}$ la v.a df pur:

$$\bar{S}^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$$

C'est un estimation sans biais et cvg de $\Theta = \text{Var}(X)$

(ii) Si de plus on connaît $E(X)$, alors $\bar{T}^2 = \frac{1}{n} \sum_{i=1}^n (X_i - E(X))^2$

$$\bar{T}^2 = \frac{1}{n} \sum_{i=1}^n (X_i - E(X))^2$$

est aussi un estimation sans biais et cvg de $\Theta = \text{Var}(X)$
ip est plus efficace que \bar{S}^2

II. ESTIMATION DES PARAMÈTRES

Définition :

Si deux statistiques S_n, T_n sont tq $P(S_n \leq \theta | t_n) \geq 1-\alpha$

$[S_n, T_n]$ un intervalle de confiance de Θ au risque α .

($1-\alpha$ est appelé niveau ou degré de confiance de l'intervalle).

« Estimation d'une moyenne

•

• Cas d'une loi normale

▷ Hyp: $\mathcal{L}(X) = \mathcal{N}(\mu, \sigma)$

▷ Prop: $\mathcal{L}(\bar{X}) = \mathcal{N}(\mu, \frac{\sigma}{\sqrt{n}})$

▷ Donc $\mathcal{L}\left(\frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}\right) = \mathcal{N}(0, 1)$

2 cas à distinguer, si σ est connu ou non.

▷ Si σ est connue.

Def:

Le quantile d'ordre a d'une var T est la valeur z_q tq $P(T < z_q) = a$

Prop:

Intervalle de confiance de m au risque α dans le cas d'une va $X \sim \mathcal{N}(\mu, \sigma)$ avec m inconnue et σ connue:

$$\left[\bar{X} - z \frac{\sigma}{\sqrt{n}}, \bar{X} + z \frac{\sigma}{\sqrt{n}} \right]$$

▷ Si σ est inconnue:

On l'estime via $S^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$

ou $\bar{S}^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$

On estime donc σ par $\sqrt{S^2}$ ou $\sqrt{\frac{S^2 n}{n-1}}$

Prop :

Si $X \sim \mathcal{N}(m, \sigma)$ avec m, σ inconnues

Alors $\mathcal{Z} \left(\frac{\bar{X} - m}{S/\sqrt{n-1}} \right)$ est une loi de Student à $n-1$

degrés de liberté S_{n-1} .

Loi de Student / Student - Fisher

Soit $n > 0$ un entier. Une var X suit la loi de Student à n degrés de liberté si elle a pour densité de probabilité :

$$P(x) = \frac{1}{\sqrt{n}\Gamma(\frac{n+1}{2})} \frac{\Gamma(\frac{n+1}{2})}{\Gamma(\frac{n}{2})} \frac{1}{(1 + \frac{x^2}{n})^{\frac{n+1}{2}}}$$

où pour $a > 0$, $\Gamma(a) = \int_0^{+\infty} u^{a-1} \exp(-u) du$ (Fct gamma)

On note $\mathcal{Z}(x) = S_n$.

On approche la loi de Student avec $n > 30$ par $\mathcal{N}(0, 1)$

Prop:

• $E(X) = 0$ pour $n > 1$.

• $E(X^2) = \frac{n}{n-2}$ pour $n > 2$.

• On note $t_n \in t_n(\beta)$ le quantile d'ordre $1 - \frac{\beta}{2}$ de la loi de Student à n d.d.P. Pour $0 < \beta < 1$, soit $t_n(\beta)$ def. par $\mathbb{P}(Y > t_n(\beta)) = \beta$ si $y \in S_n$.

Prop:

Intervalle de confiance de m au risque α dans P_c cas d'un Poi $\mathcal{N}(m, \sigma)$ avec m, σ inconnus :

$$\left[\bar{X} - t_{n-1} \frac{\tilde{s}}{\sqrt{n}}, \bar{X} + t_{n-1} \frac{\tilde{s}}{\sqrt{n}} \right]$$

ou

$$\left[\bar{X} - t_{n-1} \frac{s}{\sqrt{n-1}}, \bar{X} + t_{n-1} \frac{s}{\sqrt{n-1}} \right]$$

Rq : si n grand $t_n \approx z$.

Cas d'une loi quelconque et des grands échantillons

- $\mathcal{L}(X)$ inconnue $\Rightarrow \mathcal{L}(\bar{X})$ inconnue à priori.
- Si n est gd, on utilise P_c TCL.
- Pour $n > 30$:

$$\mathcal{L}\left(\frac{\bar{X} - m}{\sigma / \sqrt{n}}\right) = \mathcal{N}(0, 1)$$

Prop:

Intervalle de confiance approché de m au risque α , par gd échantillle.

$$\left[\bar{X} - 2 \frac{\tilde{s}}{\sqrt{n}}, \bar{X} + 2 \frac{\tilde{s}}{\sqrt{n}} \right]$$

G

On veut estimer quelle proportion p d'une pop satisfait une propriété. On a $\hat{p} \sim \text{Bin}(n, p)$.
Donc $\bar{I}(\hat{p}) = p$. Donc on doit estimer la moy.

Prop:

On estime p par la proportion d'individus satisfaisant M dans l'échantillon k/n .

Intervalle de confiance approché pour n g.d. de pr.

P : ob

$$\left[\frac{k}{n} - 2 \sqrt{\frac{k(n-k)}{n^3}}, \frac{k}{n} + 2 \sqrt{\frac{k(n-k)}{n^3}} \right]$$