Estimation et Intervalles de confiance

Chap. 4 du polycopié

Chargés de cours
V. Léger & F. Leblanc (resp. UE)

Données : $x_1, ..., x_n$
Modèle : n-échantillon $X_1, ..., X_n$ de X de loi dépendant d'un ou deux paramètres.
dans ce cours :

- Modèle Normal $\mathcal{N}(\mu, \sigma^2)$:
  - μ inconnu et σ connu
  - μ inconnu et σ inconnu
- Modèle de Bernoulli $\mathcal{B}(p)$ avec p inconnu

Rem : Les résultats obtenus dans le modèle normal s'étendront aux autres modèles pour n suffisamment grand grâce au TCL.

Le problème : Proposer un ou plusieurs estimateurs du ou des paramètres inconnus du modèle et en établir les propriétés pour choisir le meilleur.

Estimateur : Un estimateur de θ est une fonction connue de $X_1, ..., X_n$: $T(X_1, ..., X_n) = T_n$ (aussi noté $\hat{\theta}(X_1, ..., X_n)$).

Exemples d'estimateurs

- pour $\mu = E(X)$: $\bar{X}_n$, $min(X_i)$, X_1, $(X_1 + X_2)/2$ ...
- pour $\sigma^2 = V(X)$: S_n^2, $S_n'^2$, $X_1^2 - \bar{X}_n^2$ ...

Lequel choisir ?

Qualités attendues d'un estimateur :

- **Sans biais** : $E(T_n) = \theta$ (retrouve en moyenne θ)
- **De variabilité décroissante** :

$$V(T_n) \to 0 \text{ si } n \to \infty$$

- **Convergent** : sans biais et de variance tendant vers zéro avec n
- **Vocabulaire** : T_n calculé avec les données $x_1, \ldots, x_2$ produit une valeur appelée estimation de θ
- **Notations** : $\hat{\theta}$ l'estimateur de θ appliqué aux données sera parfois noté $\hat{\theta}_{calc}$ (ou plus généralement une statistique T appliquée aux données sera notée T_{calc})

Dans **tous les modèles** où la variable X est tq $\mu = E(X) < \infty$ et $\sigma^2 = V(X) < \infty$ les meilleurs estimateurs de μ et σ^2 sont :

$$\hat{\mu} = \bar{X}_n \qquad \text{et} \qquad \widehat{\sigma^2} = S_n'^2$$

Un premier argument pour estimer μ par $\bar{X}_n$ est la loi des grands nombres : $\bar{X}_n \to \mu$ (on peut justifier $S_n'^2$ de manière analogue). De plus ils sont sans biais et de variance qui tend vers 0 avec n, en effet $\forall n$:

$$E(\bar{X}_n) = \mu \quad \text{et} \quad V(\bar{X}_n) = \frac{\sigma^2}{n}$$

$$E(S_n'^2) = \sigma^2 \quad \text{et} \quad V(S_n'^2) = \mathcal{O}\left(\frac{1}{n}\right)$$

Intervalles de confiance

Définition : On dira que $[T_1, T_2]$ est un intervalle de niveau de confiance $1-\alpha$ pour θ si T_1 et T_2 **ne dépendent que des** X_i et éventuellement d'un paramètre connu et est tel que :

$$P(\theta \in [T_1, T_2]) = 1 - \alpha$$

Pour construire des intervalles de confiance dans le cas gaussien on utilise le résultat suivant :

Lois des estimateurs :

si X de loi $\mathcal{N}(\mu, \sigma^2)$

- $\bar{X}_n$ suit une loi normale $\mathcal{N}(\mu, \sigma^2/n)$
- $nS_n^2/\sigma^2 = (n-1)S_n'^2/\sigma^2$ suit une loi $\mathcal{X}_{n-1}^2$

Intervalle de confiance pour μ :

- **si σ^2 connue :**

$$IC(\bar{X}_n, \alpha, \sigma^2) = \left[\bar{X}_n - \frac{\sigma}{\sqrt{n}} u_{1-\alpha/2}, \bar{X}_n + \frac{\sigma}{\sqrt{n}} u_{1-\alpha/2}\right]$$

  où u_p est le quantile d'ordre p d'une $\mathcal{N}(0,1)$.

- **si σ^2 inconnue :**

$$IC(\bar{X}_n, \alpha) = \left[\bar{X}_n - \frac{S'_n}{\sqrt{n}} t_{n-1,1-\alpha/2}, \bar{X}_n + \frac{S'_n}{\sqrt{n}} t_{n-1,1-\alpha/2}\right]$$

  où $t_{n-1,p}$ est le quantile d'ordre p d'une $\mathcal{T}_{n-1}$.
  R : `t.test(x,conf.level=0.90)` fournit un IC symétrique de niveau de conf. 90%

Intervalle de confiance pour σ^2 avec μ inconnu :

$$IC(S_n^2, \alpha) = \left[\frac{nS_n^2}{z_{n-1,1-\alpha/2}}, \frac{nS_n^2}{z_{n-1,\alpha/2}} \right]$$

Comme nS_n^2/σ^2 est une variable $\mathcal{X}^2_{n-1}$ en notant $z_{n-1,p}$ son quantile d'ordre p on a :

$$P(z_{n-1,\alpha/2} < nS_n^2/\sigma^2 < z_{n-1,1-\alpha/2}) = 1 - \alpha$$

d'où

$$P(\sigma^2 \in IC(S^2, \alpha)) = 1 - \alpha$$

Si X de loi $\mathcal{B}(p)$, $\mu = E(X) = p$ et $\sigma^2 = p(1-p)$.
$\bar{X}_n$ représente la fréquence d'apparition du 1 dans l'échantillon notée F_n:

$$\bar{X}_n = F_n = \frac{1}{n}\sum_{i=1}^{n} X_i \quad \text{et} \quad S_n^2 = \bar{X}_n(1-\bar{X}_n) = F_n(1-F_n)$$

car

$$X_i \in \{0,1\} \implies \sum_{i=1}^{n} X_i^2 = \sum_{i=1}^{n} X_i$$

Rem : si n est suffisamment grand $S_n'^2 \approx S_n^2$ puisque $(n-1)/n \approx 1$

Modèle de Bernoulli pour p :

$$IC(F_n,\alpha)=\left[F_n-\frac{\sqrt{F_n(1-F_n)}}{\sqrt{n}}u_{1-\alpha/2},F_n+\frac{\sqrt{F_n(1-F_n)}}{\sqrt{n}}u_{1-\alpha/2}\right]$$

satisfait $P(p\in I(p,\alpha))\approx 1-\alpha$ à condition que $np>10$ et $n(1-p)>10$. IC de niveau asymptotique (ou approx.) $1-\alpha$.

La preuve s'appuie sur le TCL qui permet de montrer que pour X de loi $\mathcal{B}(p)$: F_n suit approx. une $\mathcal{N}(p,p(1-p)/n)$ et sur le Théorème de Slutsky pour établir ensuite que $\sqrt{n}(F_n-p)/\sqrt{F_n(1-F_n)}$ suit approx. une loi normale centrée réduite (admis).