2. Estimation

Soit $(X_1, ..., X_n)$ un échantillon de taille n d'une v.a. X. On suppose que la loi de X dépend d'un paramètre $\theta \in \mathbb{R}$. On cherche à estimer ce paramètre grâce à l'échantillon.

2.1. **Estimateurs.**

Définition 2.1. On appelle *estimateur du paramètre* θ toute statistique T de l'échantillon $(X_1, ..., X_n)$ ne dépendant pas fonctionnellement du paramètre θ lui-même.

Si T est intégrable, le *biais de l'estimateur* T est l'écart moyen entre T et θ, soit $\mathbf{E}[T - \theta] = \mathbf{E}[T] - \theta$. Un *estimateur sans biais de* θ est un estimateur dont le biais est nul.

Si T est de carré intégrable, la *précision de l'estimateur* T est l'écart quadratique moyen entre T et θ, soit $\mathbf{E}\left[(T-\theta)^2\right] = \mathbf{Var}[T] + (\mathbf{E}[T] - \theta)^2$.

Si T et T' sont deux estimateurs d'un même paramètre pour une même v.a. X, on dit que T *est un meilleur estimateur que* T' si sa précision est inférieure à celle de T'.

Définition 2.2. Lorsque l'estimateur T_n dépend de la taille n de l'échantillon, on dira qu'il est, lorsque les moments ci-dessous sont bien définis,

(i) *asymptotiquement sans biais pour* θ si $\lim_{n\to+\infty} \mathbf{E}[T_n] = \theta$;
(ii) *convergent* si $\lim_{n\to+\infty} \mathbf{E}\left[(T_n - \theta)^2\right] = 0$.

Remarque. Un estimateur est convergent si et seulement si il est asymptotiquement sans biais et si sa variance converge vers 0 quand n tend vers l'infini.

Proposition 2.1. Soit $(X_1, ..., X_n)$ un échantillon d'une v.a. X.

(1) Si X est une v.a. intégrable alors $\overline{X}$ est un estimateur sans biais de $\mathbf{E}[X]$. Si de plus X est de carré intégrable alors $\overline{X}$ est convergent.
(2) Si X est de carré intégrable alors Σ^2 est un estimateur asymptotiquement sans biais de $\mathbf{Var}[X]$. Si de plus X^4 est intégrable, alors Σ^2 est convergent.

Remarque. Lorsque X est de carré intégrable, on préfère à Σ^2, comme estimateur de la variance, l'estimateur sans biais

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

2.2. **Deux méthodes pour déterminer un estimateur.**

On suppose que la loi de la variable parente X est connue mais que cette loi dépend d'un paramètre θ inconnu. On dispose d'un échantillon $(X_1, ..., X_n)$ de la variable aléatoire X et on cherche des estimateurs du paramètre θ.

Méthode des moments. On suppose que la variable X est intégrable et que l'espérance de X s'écrit comme une fonction de θ, soit $\mathbf{E}[X] = g(\theta)$. Un estimateur obtenu par la *méthode des moments* est une statistique T de l'échantillon $(X_1, ..., X_n)$ telle que $\overline{X} = g(T)$.

Si X est de carré intégrable, cette méthode peut permettre de déterminer également des estimateurs vectoriels lorsque l'on a deux paramètres à estimer : soient θ_1, θ_2 ces paramètres. On suppose $espX = g_1(\theta_1, \theta_2)$ et $\mathbf{Var}[X] = g_2(\theta_1, \theta_2)$. On cherche alors des estimateurs T_1 et T_2 tels que $\overline{X} = g_1(T_1, T_2)$ et $S^2 = g_2(T_1, T_2)$.

Méthode du maximum de vraisemblance. La vraisemblance d'un échantillon $(X_1, ..., X_n)$ de la v.a. X est la famille, indicée par $\theta \in \Theta$, des lois du vecteur $(X_1, ..., X_n)$ sous l'hypothèse que X suit la loi $\mathcal{L}(\theta)$. Plus précisément, on a les définitions suivantes :

Cas discret : on suppose que, pour tout $\theta \in \Theta$, X suit une loi discrète prenant ses valeurs dans un ensemble dénombrable D_θ. La *vraisemblance de l'échantillon* $(X_1, ..., X_n)$ est la famille, indicée par $\theta \in \Theta$, des applications, $L(\cdot;\theta) : D_\theta^n \longrightarrow [0,1]$ définies, pour tout $(x_1, ..., x_n) \in D_\theta^n$, sous l'hypothèse que X suit la loi $\mathcal{L}(\theta)$, par $L(x_1, ..., x_n; \theta) = \mathbf{P}[X = x_1]...\mathbf{P}[X = x_n]$.

Cas absolument continu : on suppose que, pour tout $\theta \in \Theta$, X suit une loi de densité f_θ. La *vraisemblance de l'échantillon* $(X_1, ..., X_n)$ est la famille, indicée par $\theta \in \Theta$, des applications $L(\cdot;\theta) : \mathbb{R}^n \longrightarrow \mathbb{R}_+$ définies, pour tout $(x_1, ..., x_n) \in \mathbb{R}^n$, par $L(x_1, ..., x_n; \theta) = f_\theta(x_1)...f_\theta(x_n)$.

La *méthode du maximum de vraisemblance* consiste à déterminer, pour tout $(x_1, ..., x_n)$, un paramètre réalisant le suprémum de $L(x_1, ..., x_n; \theta)$ sur tous les paramètres possibles $\theta \in \Theta$. Puisqu'un tel paramètre dépend *a priori* de $(x_1, ..., x_n)$, on peut le noter $g(x_1, ..., x_n)$. Un *estimateur du maximum de vraisemblance* est alors un estimateur T s'écrivant $T = g(X_1, ..., X_n)$, où l'on admet que g est mesurable.

2.3. **Estimation par intervalles de confiance.** On suppose que l'on dispose d'un échantillon $(X_1, ..., X_n)$ d'une v.a. X dont on connaît la loi mais pas un ou plusieurs paramètres. Un sl intervalle de confiance à un seuil de risque fixé α pour un paramètre réel est un intervalle dans lequel on sait que se trouve le paramètre avec probabilité au moins égale à $1 - \alpha$, connaissant une réalisation $(x_1, ..., x_n)$ de l'échantillon.

Cas des variables gaussiennes. On suppose que la variable parente suit une loi gaussienne de paramètres m et σ^2 dont on cherche des intervalles de confiance I au seuil de risque donné α.

Estimation de la moyenne lorsque σ est connu. Pour tout $\beta \in]0,1[$, on note u_β le réel tel que $\phi(u_\beta) = \mathbf{P}[X^* \leq u_\beta] = 1 - \beta$ si $X^* \rightsquigarrow \mathcal{N}(0,1)$. Pour une réalisation $\overline{x}$ de $\overline{X}$,

$$I = \left\{ m \in \mathbb{R} \,\middle/\, -u_{\alpha/2} < \frac{\overline{x} - m}{\sigma/\sqrt{n}} < u_{\alpha/2} \right\} = \left] \overline{x} - \frac{\sigma}{\sqrt{n}} u_{\alpha/2}, \overline{x} + \frac{\sigma}{\sqrt{n}} u_{\alpha/2} \right[.$$

Estimation de la moyenne lorsque σ est inconnu. Pour tout $\beta \in]0,1[$, on note t_β le réel tel que $\mathbf{P}[T_{n-1} \leq t_\beta] = 1 - \beta$ pour une v.a. T_n suivant une loi de Student à $n-1$ degrés de liberté. Pour des réalisations $\overline{x}$ et s^2 de $\overline{X}$ et S^2,

$$I = \left] \overline{x} - \frac{s}{\sqrt{n}} t_{\alpha/2}, \overline{x} + \frac{s}{\sqrt{n}} t_{\alpha/2} \right[.$$

Estimation de la variance lorsque m est connu. On utilise l'estimateur

$$T = \frac{1}{n} \sum_{i=1}^{n} (X_i - m)^2.$$

On note x_β le réel tel que $\mathbf{P}\left[\chi_n^2 \leq x_\beta\right] = 1 - \beta$ pour tout $\beta \in]0,1[$. Pour une réalisation $\overline{t}$ de T,

$$I = \left] \frac{n\overline{t}}{x_{\alpha/2}}, \frac{n\overline{t}}{x_{1-\alpha/2}} \right[.$$

Estimation de la variance lorsque m est inconnu. On note x_β le réel tel que $\mathbf{P}\left[\chi_{n-1}^2 \leq x_\beta\right] = 1 - \beta$ pour tout $\beta \in]0,1[$. Pour une réalisation s^2 de S^2,

$$I = \left] \frac{(n-1)s^2}{x_{\alpha/2}}, \frac{(n-1)s^2}{x_{1-\alpha/2}} \right[.$$

Cas de variables non gaussiennes de carré intégrable. Par le théorème limite central, les intervalles de confiances du cas gaussien pour la moyenne sont encore valables. Par contre, ça n'est pas le cas pour les estimations de la variance.