Chap IV - Estimation par intervalles de confiance

1 Intervalle de confiance pour l'espérance d'un n-échantillon gaussien

1.1 Cas où la variance est connue

Soit $X_1, \cdots, X_n$ un n-échantillon de loi $\mathcal{N}(\mu, \sigma_0^2)$ pour lequel on dispose des réalisations (observations) $x_1, \cdots, x_n$; μ est inconnu et σ_0^2 est connue. On cherche un intervalle

$$I_{n,\alpha} = [A_{n,\alpha}; B_{n,\alpha}],$$

où $A_{n,\alpha}$ et $B_{n,\alpha}$ sont deux variables aléatoires fonctions de $X_1, \cdots, X_n$, tel que

$$\mathbb{P}(\mu \in I_{n,\alpha}) = \mathbb{P}(A_{n,\alpha} \leq \mu \leq B_{n,\alpha}) = 1 - \alpha.$$

L'espérance μ est naturellement estimée par l'estimateur sans biais et consistant

$$\overline{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i.$$

Comme moyenne de n variables aléatoires indépendantes de loi $\mathcal{N}(\mu, \sigma_0^2)$,

$$\overline{X}_n \sim \mathcal{N}\left(\mu, \frac{\sigma_0^2}{n}\right) \text{ et donc } \sqrt{n}\frac{\overline{X}_n - \mu}{\sigma_0} \sim \mathcal{N}(0,1).$$

Considérons u_α tel que si $Z \sim \mathcal{N}(0,1)$, alors $\mathbb{P}(-u_\alpha \leq Z \leq u_\alpha) = 1 - \alpha$, c'est-à-dire tel que

$$\mathbb{P}(Z \leq u_\alpha) = 1 - \frac{\alpha}{2}.$$

Pour ce u_α, lu dans la table de la loi $\mathcal{N}(0,1)$, on a

$$\mathbb{P}\left(-u_\alpha \leq \sqrt{n}\frac{\overline{X}_n - \mu}{\sigma_0} \leq u_\alpha\right) = 1 - \alpha.$$

Comme

$$\left\{-u_\alpha \leq \sqrt{n}\frac{\overline{X}_n - \mu}{\sigma_0} \leq u_\alpha\right\} \iff \left\{\overline{X}_n - u_\alpha\frac{\sigma_0}{\sqrt{n}} \leq \mu \leq \overline{X}_n + u_\alpha\frac{\sigma_0}{\sqrt{n}}\right\},$$

on en déduit que

$$\mathbb{P}\left(\overline{X}_n - u_\alpha \frac{\sigma_0}{\sqrt{n}} \leq \mu \leq \overline{X}_n + u_\alpha \frac{\sigma_0}{\sqrt{n}}\right) = 1 - \alpha.$$

L'intervalle de confiance pour μ de niveau de confiance $1-\alpha$ est donc

$$I_{n,\alpha} = \left[\overline{X}_n - u_\alpha \frac{\sigma_0}{\sqrt{n}} \ ; \ \overline{X}_n + u_\alpha \frac{\sigma_0}{\sqrt{n}}\right],$$

où u_α est tel que $\mathbb{P}(Z \leq u_\alpha) = 1 - \alpha/2$. Le paramètre μ appartient à l'intervalle aléatoire $I_{n,\alpha}$ avec une probabilité de $1-\alpha$.

Pour un échantillon de taille n donné pour lequel on dispose des réalisations $x_1, \cdots, x_n$, on calcule $\overline{x}_n$. L'intervalle de confiance observé associé de niveau de confiance $1-\alpha$ est

$$\left[\overline{x}_n - u_\alpha \ \frac{\sigma_0}{\sqrt{n}} \ ; \ \overline{x}_n + u_\alpha \ \frac{\sigma_0}{\sqrt{n}}\right]$$

(qui contiendra ou non μ). Si l'on répétait un grand nombre de fois l'expérience consistant à prendre un échantillon de taille n et à lui faire correspondre l'intervalle de confiance observé pour μ, dans $(1-\alpha)100\%$ des cas, l'intervalle couvrirait la vraie valeur de μ.

1.2 Cas où la variance est inconnue

Soit $X_1, \cdots, X_n$ un n-échantillon de loi $\mathcal{N}(\mu, \sigma^2)$ pour lequel on dispose des réalisations (observations) $x_1, \cdots, x_n$; μ et σ^2 sont inconnus. On cherche un intervalle

$$I_{n,\alpha} = [A_{n,\alpha}; B_{n,\alpha}],$$

où $A_{n,\alpha}$ et $B_{n,\alpha}$ sont deux variables aléatoires fonctions de $X_1, \cdots, X_n$, tel que

$$\mathbb{P}(\mu \in I_{n,\alpha}) = \mathbb{P}(A_{n,\alpha} \leq \mu \leq B_{n,\alpha}) = 1 - \alpha.$$

L'espérance μ est naturellement estimée par l'estimateur sans biais et consistant

$$\overline{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i.$$

Comme moyenne de n variables aléatoires indépendantes de loi $\mathcal{N}(\mu, \sigma^2)$,

$$\overline{X}_n \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right) \text{ et donc } \sqrt{n}\frac{\overline{X}_n - \mu}{\sigma} \sim \mathcal{N}(0,1).$$

On ne peut cependant pas construire un intervalle de confiance à partir de ce résultat : l'intervalle dépendrait de σ^2 qui est inconnu. On estime donc σ^2 par l'estimateur sans biais :

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

On considère la statistique

$$\sqrt{n}\frac{\overline{X}_n - \mu}{S_n}.$$

Cette statistique ne suit pas une loi gaussienne : sa loi est appelée loi de Student à $n-1$ degrés de liberté. On note

$$\sqrt{n}\frac{\overline{X}_n - \mu}{S_n} \sim \mathrm{St}(n-1).$$

Cette loi est centrée et sa fonction de densité est, comme pour la loi gaussienne, une fonction paire. Elle admet donc les mêmes propriétés de symétrie que la loi gaussienne. Sa fonction de répartition est aussi tabulée. Considérons t_α tel que si $T \sim \mathrm{St}(n-1)$, alors

$$\mathbb{P}(-t_\alpha \leq T \leq t_\alpha) = 1-\alpha,$$

c'est-à-dire tel que

$$\mathbb{P}(T \leq t_\alpha) = 1 - \frac{\alpha}{2}.$$

Pour ce t_α, lu dans la table de la loi de Student à $n-1$ degrés de liberté, on a

$$\mathbb{P}\left(-t_\alpha \leq \sqrt{n}\frac{\overline{X}_n - \mu}{S_n} \leq t_\alpha\right) = 1-\alpha.$$

Comme

$$\left\{-t_\alpha \leq \sqrt{n}\frac{\overline{X}_n - \mu}{S_n} \leq t_\alpha\right\} \iff \left\{\overline{X}_n - t_\alpha\frac{S_n}{\sqrt{n}} \leq \mu \leq \overline{X}_n + t_\alpha\frac{S_n}{\sqrt{n}}\right\},$$

on en déduit que

$$\mathbb{P}\left(\overline{X}_n - t_\alpha\frac{S_n}{\sqrt{n}} \leq \mu \leq \overline{X}_n + t_\alpha\frac{S_n}{\sqrt{n}}\right) = 1-\alpha.$$

L'intervalle de confiance pour μ de niveau de confiance $1-\alpha$ est donc

$$I_{n,\alpha} = \left[\overline{X}_n - t_\alpha\frac{S_n}{\sqrt{n}} \ ; \ \overline{X}_n + t_\alpha\frac{S_n}{\sqrt{n}}\right],$$

où t_α est tel que $\mathbb{P}(T \leq t_\alpha) = 1-\alpha/2$. Le paramètre μ appartient à l'intervalle aléatoire $I_{n,\alpha}$ avec une probabilité de $1-\alpha$.

Pour un échantillon de taille n donné pour lequel on dispose des réalisations $x_1, \cdots, x_n$, on calcule $\overline{x}_n$ et s_n les réalisations correspondantes de $\overline{X}_n$ et S_n. L'intervalle de confiance observé associé de niveau de confiance $1-\alpha$ est

$$\left[\overline{x}_n - t_\alpha\ \frac{s_n}{\sqrt{n}} \ ; \ \overline{x}_n + t_\alpha\ \frac{s_n}{\sqrt{n}}\right]$$

(qui contiendra ou non μ). Si l'on répétait un grand nombre de fois l'expérience consistant à prendre un échantillon de taille n et à lui faire correspondre l'intervalle de confiance observé pour μ, dans $(1-\alpha)100\%$ des cas, l'intervalle couvrirait la vraie valeur de μ.

2 Intervalle de confiance de l'espérance d'un grand échantillon de loi quelconque

2.1 Cas où la variance est connue

Soit $X_1, \cdots, X_n$ un n-échantillon de loi d'espérance $\mu = \mathbb{E}(X_1)$ et de variance $\sigma_0^2 = \mathrm{Var}(X_1)$; μ est inconnu et σ_0^2 est connue. L'estimateur (sans biais et consistant) de μ est

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

On suppose que n est grand de sorte que le théorème central limite s'applique. La fonction de répartition F_n de

$$\sqrt{n}\frac{\overline{X}_n - \mu}{\sigma_0}$$

tend, quand n tend vers l'infini, vers la fonction de répartition Φ de la loi $\mathcal{N}(0,1)$: pour tout $t \in \mathbb{R}$, $F_n(t) \to \Phi(t)$. Ceci implique que pour n assez grand, alors pour tout a et b réels,

$$\mathbb{P}\left(a \leq \sqrt{n}\frac{\overline{X}_n - \mu}{\sigma_0} \leq b\right) \text{ est proche de la probabilité } \mathbb{P}(a \leq Z \leq b)$$

où $Z \sim \mathcal{N}(0,1)$.

On cherche un intervalle $I_{n,\alpha} = [A_{n,\alpha}; B_{n,\alpha}]$, où $A_{n,\alpha}$ et $B_{n,\alpha}$ sont deux variables aléatoires fonctions de $X_1, \cdots, X_n$, tel que

$$\mathbb{P}(\mu \in I_{n,\alpha}) \text{ est proche de } 1-\alpha \text{ si } n \text{ est grand,}$$

plus exactement tel que

$$\lim_{n\to\infty} \mathbb{P}(\mu \in I_{n,\alpha}) = 1 - \alpha.$$

Soit u_α tel que si $Z \sim \mathcal{N}(0,1)$, alors $\mathbb{P}(-u_\alpha \leq Z \leq u_\alpha) = 1-\alpha$, c'est-à-dire tel que $\mathbb{P}(Z \leq u_\alpha) = 1 - \alpha/2$. Pour ce u_α, lu dans la table de la loi $\mathcal{N}(0,1)$,

$$\mathbb{P}\left(-u_\alpha \leq \sqrt{n}\frac{\overline{X}_n - \mu}{\sigma_0} \leq u_\alpha\right) \text{ est proche de } 1-\alpha \text{ si } n \text{ est grand.}$$

On en déduit que

$$\mathbb{P}\left(\overline{X}_n - u_\alpha\frac{\sigma_0}{\sqrt{n}} \leq \mu \leq \overline{X}_n + u_\alpha\frac{\sigma_0}{\sqrt{n}}\right) \text{ est proche de } 1-\alpha \text{ si } n \text{ est grand.}$$

Plus exactement :

$$\lim_{n\to\infty} \mathbb{P}\left(\overline{X}_n - u_\alpha\frac{\sigma_0}{\sqrt{n}} \leq \mu \leq \overline{X}_n + u_\alpha\frac{\sigma_0}{\sqrt{n}}\right) = 1 - \alpha.$$

L'intervalle de confiance pour μ de niveau de confiance asymptotique $1-\alpha$ est donc

$$I_{n,\alpha} = \left[\overline{X}_n - u_\alpha\frac{\sigma_0}{\sqrt{n}} \ ; \ \overline{X}_n + u_\alpha\frac{\sigma_0}{\sqrt{n}}\right], \text{ où } u_\alpha \text{ est tel que } \mathbb{P}(Z \leq u_\alpha) = 1 - \alpha/2.$$

2.2 Cas où la variance est inconnue

Soit $X_1, \cdots, X_n$ un n-échantillon de loi d'espérance μ et de variance σ^2 où μ et σ^2 sont inconnus. Soit $\hat{\sigma}_n^2$ un estimateur consistant de σ^2. On peut alors remplacer σ^2 par $\hat{\sigma}_n^2$ dans le Théorème Central Limite et on obtient que la fonction de répartition F_n de

$$\sqrt{n}\frac{\overline{X}_n - \mu}{\hat{\sigma}_n}$$

tend, quand n tend vers l'infini, vers la fonction de répartition Φ de la loi $\mathcal{N}(0,1)$: pour tout $t \in \mathbb{R}$, $F_n(t) \to \Phi(t)$.

Ceci implique que pour n assez grand, alors pour tout a et b réels,

$$\mathbb{P}\left(a \leq \sqrt{n}\frac{\overline{X}_n - \mu}{\hat{\sigma}_n} \leq b\right) \text{ est proche de } \mathbb{P}(a \leq Z \leq b)$$

où $Z \sim \mathcal{N}(0,1)$. Soit u_α tel que si $Z \sim \mathcal{N}(0,1)$, alors $\mathbb{P}(-u_\alpha \leq Z \leq u_\alpha) = 1-\alpha$, c'est-à-dire tel que $\mathbb{P}(Z \leq u_\alpha) = 1 - \alpha/2$. Pour ce u_α, lu dans la table de la loi $\mathcal{N}(0,1)$, la probabilité

$$\mathbb{P}\left(-u_\alpha \leq \sqrt{n}\frac{\overline{X}_n - \mu}{\hat{\sigma}_n} \leq u_\alpha\right) \text{ est proche de } 1-\alpha.$$

On en déduit que la probabilité

$$\mathbb{P}\left(\overline{X}_n - u_\alpha\frac{\hat{\sigma}_n}{\sqrt{n}} \leq \mu \leq \overline{X}_n + u_\alpha\frac{\hat{\sigma}_n}{\sqrt{n}}\right) \text{ est proche de } 1-\alpha.$$

Plus exactement

$$\lim_{n\to\infty} \mathbb{P}\left(\overline{X}_n - u_\alpha\frac{\hat{\sigma}_n}{\sqrt{n}} \leq \mu \leq \overline{X}_n + u_\alpha\frac{\hat{\sigma}_n}{\sqrt{n}}\right) = 1-\alpha.$$

Un intervalle de confiance pour μ de niveau de confiance asymptotique $1-\alpha$ est donc

$$I_{n,\alpha} = \left[\overline{X}_n - u_\alpha\frac{\hat{\sigma}_n}{\sqrt{n}} \ ; \ \overline{X}_n + u_\alpha\frac{\hat{\sigma}_n}{\sqrt{n}}\right], \text{ où } u_\alpha \text{ est tel que } \mathbb{P}(Z \leq u_\alpha) = 1 - \alpha/2.$$

En général, on choisit comme estimateur de σ^2, $\hat{\sigma}_n^2 = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n}(X_i - \overline{X}_n)^2$. Mais on peut aussi choisir tout autre estimateur consistant de σ^2.

3 Intervalle de confiance pour une probabilité dans le cas d'un grand échantillon

Soient $X_1, \cdots, X_n$ n variables aléatoires i.i.d. de loi $\mathcal{B}(p)$ avec p inconnue et soient $x_1, \cdots, x_n$ les n observations associées. Nous avons vu que $\overline{X}_n$ est un bon estimateur de

p puisqu'il est sans biais et qu'il converge en moyenne quadratique. Nous allons utiliser le théorème central limite pour obtenir un intervalle de confiance pour p.

On suppose dans cette section que n est grand. On cherche un intervalle $I_{n,\alpha} = [A_{n,\alpha}; B_{n,\alpha}]$, où $A_{n,\alpha}$ et $B_{n,\alpha}$ sont deux variables aléatoires fonctions de $X_1, \cdots, X_n$, tel que

$$\lim_{n\to\infty} \mathbb{P}(p \in I_{n,\alpha}) = 1 - \alpha.$$

Comme $\mathbb{E}(X_1) = p$, c'est un cas particulier de l'estimation par intervalle d'une espérance. D'après le Théoème Central Limite, la fonction de répartition de

$$\sqrt{n}\frac{\overline{X}_n - p}{\sqrt{p(1-p)}}$$

tend, quand n tend vers l'infini, vers la fonction de répartition de la loi $\mathcal{N}(0,1)$. Ici

$$\mathrm{Var}(X_1) = p(1-p)$$

est inconnue et peut être estimée de manière consistante par $\overline{X}_n(1-\overline{X}_n)$ puisque la probabilité p est estimée par l'estimateur sans biais et consistant $\overline{X}_n$.
On obtient que la fonction de répartition F_n de

$$\sqrt{n}\frac{\overline{X}_n - p}{\sqrt{\overline{X}_n(1-\overline{X}_n)}}$$

tend, quand n tend vers l'infini, vers la fonction de répartition Φ de la loi $\mathcal{N}(0,1)$: pour tout $t \in \mathbb{R}$, $F_n(t) \to \Phi(t)$. Soit u_α tel que si $Z \sim \mathcal{N}(0,1)$, alors $\mathbb{P}(-u_\alpha \leq Z \leq u_\alpha) = 1-\alpha$, c'est-à-dire tel que $\mathbb{P}(Z \leq u_\alpha) = 1 - \alpha/2$. Pour ce u_α, lu dans la table de la loi $\mathcal{N}(0,1)$, la probabilité

$$\mathbb{P}\left(-u_\alpha \leq \sqrt{n}\frac{\overline{X}_n - p}{\sqrt{\overline{X}_n(1-\overline{X}_n)}} \leq u_\alpha\right) \text{ est proche de } 1-\alpha.$$

On en déduit que la probabilité

$$\mathbb{P}\left(\overline{X}_n - u_\alpha\sqrt{\frac{\overline{X}_n(1-\overline{X}_n)}{n}} \leq p \leq \overline{X}_n + u_\alpha\sqrt{\frac{\overline{X}_n(1-\overline{X}_n)}{n}}\right) \text{ est proche de } 1-\alpha$$

ou bien plus précisément

$$\lim_{n\to\infty} \mathbb{P}\left(\overline{X}_n - u_\alpha\sqrt{\frac{\overline{X}_n(1-\overline{X}_n)}{n}} \leq p \leq \overline{X}_n + u_\alpha\sqrt{\frac{\overline{X}_n(1-\overline{X}_n)}{n}}\right) = 1-\alpha.$$

L'intervalle de confiance asymptotique pour p de niveau de confiance $1-\alpha$ est donc

$$I_{n,\alpha} = \left[\overline{X}_n - u_\alpha\sqrt{\frac{\overline{X}_n(1-\overline{X}_n)}{n}} \ ; \ \overline{X}_n + u_\alpha\sqrt{\frac{\overline{X}_n(1-\overline{X}_n)}{n}}\right],$$

où u_α est tel que $\mathbb{P}(Z \leq u_\alpha) = 1 - \alpha/2$.

4 Cas général

Soit $X_1, X_2, \cdots, X_n$, un n-échantillon de variables aléatoires de loi P_θ, dépendant d'un paramètre θ inconnu. On va chercher un intervalle $I_{n,\alpha} = [A_{n,\alpha}, B_{n,\alpha}]$, où $A_{n,\alpha}$ et $B_{n,\alpha}$ sont deux variables aléatoires fonctions de $X_1, \cdots, X_n$, tel que

$$\mathbb{P}(\theta \in [A_{n,\alpha}; B_{n,\alpha}]) = \mathbb{P}(\theta \in I_{n,\alpha}) = 1 - \alpha.$$

Parfois on ne peut pas avoir exactement l'égalité. On cherche donc un intervalle $I_{n,\alpha}$ tel que

$$\mathbb{P}(\theta \in I_{n,\alpha}) \geq 1 - \alpha$$

ou bien tel que

$$\lim_{n\to\infty} \mathbb{P}(\theta \in I_{n,\alpha}) = 1 - \alpha.$$

On a les définitions suivantes :

Définition 1 *Soit $\alpha \in]0, 1[$. On appelle intervalle de confiance du paramètre θ de niveau de confiance $1 - \alpha$ (ou de risque α) un intervalle (aléatoire) $I_{n,\alpha}$ tel que*

$$\mathbb{P}(\theta \in I_{n,\alpha}) \geq 1 - \alpha.$$

Définition 2 *Soit $\alpha \in]0, 1[$. On appelle intervalle de confiance du paramètre θ de niveau de confiance asymptotique $1 - \alpha$ (ou de risque asymptotique α) un intervalle (aléatoire) $I_{n,\alpha}$ tel que*

$$\lim_{n\to\infty} \mathbb{P}(\theta \in I_{n,\alpha}) = 1 - \alpha.$$

Plus le niveau de confiance $1 - \alpha$ est grand, plus l'intervalle de confiance est de grande amplitude. Autrement dit, si on ne s'autorise que peu d'erreurs, l'intervalle de confiance sera moins précis.
Plus la taille de l'échantillon augmente, plus l'observation contient de l'information, plus l'amplitude de l'intervalle de confiance est faible (à niveau de confiance égal).