RÉSUMÉ DE COURS : ESTIMATION

En statistiques, on connait bien souvent le type de loi suivie par un phénomène aléatoire (ne serait-ce que par des observations pratiques). Par contre, on ne connait pas forcément le(s) paramètre(s) de la loi en question. Dans ce chapitre, on se propose d'estimer des quantités inconnues liées à une variable aléatoire suivant une loi d'un type connu (comme par exemple le(s) paramètre(s) de la loi).

A partir de maintenant, on considère une variable aléatoire X qui suit une loi d'un type connu (par exemple, la loi de Poisson ou la loi normale). On suppose que la loi en question dépend d'un paramètre θ appartenant à une partie Θ de $\mathbb{R}^k$ (c'est-à-dire un paramètre scalaire ou vectoriel). Cette loi sera notée μ_θ par la suite. A titre d'exemple, on prendra $k = 1$ pour la loi de Poisson (un seul paramètre réel λ) et $k = 2$ pour la loi normale (deux paramètres réels m et σ^2). Etant donnée une fonction $g : \Theta \longrightarrow \mathbb{R}$, on se propose dans ce chapitre de donner une estimation de la valeur de $g(\theta)$. Dans la plupart des cas, la quantité à estimer est notée θ, mais on la notera en général $g(\theta)$, et ce conformément au programme.

1 Echantillonnage

Définition 1.1. *Soit X une variable aléatoire définie sur un espace probabilisé $(\Omega, \mathcal{A}, P)$. On appelle n-échantillon de la loi de X tout n-uplet $(X_1, ..., X_n)$ de variables aléatoires définies sur $(\Omega, \mathcal{A}, P)$, mutuellement indépendantes et suivant la même loi que X.*

A noter que la loi $\mathcal{L}$ commune à X et aux X_k est appelée *la loi parente de l'échantillon.* Dans la littérature mathématique, on parle d'échantillon *i.i.d.*, c'est-à-dire indépendant et identiquement distribué, ce qui signifie que les X_k sont indépendantes et suivent toutes la même loi.

Définition 1.2. *Soit $(X_1, ..., X_n)$ un n-échantillon de la loi de X. On appelle réalisation de cet échantillon tout n-uplet $(x_1, ..., x_n)$ pour lequel il existe un élément $\omega \in \Omega$ tel que $(x_1, ..., x_n) = (X_1(\omega), ..., X_n(\omega))$.*

Dans la pratique, pour réaliser un n-échantillon de la loi de X, on effectue n expériences aléatoires indépendantes et identiques, chacune d'entre elles étant associée à une variable aléatoire X_k suivant la même loi que X. Dans ce cas, la réalisation du n-échantillon est le n-uplet $(x_1, ..., x_n)$, où x_k est la valeur prise par la variable aléatoire X_k pour tout $k \in \{1, ..., n\}$. *A partir de maintenant, toutes les variables aléatoires $X_1, ..., X_n$ seront supposées définies sur un même espace probabilisable $(\Omega, \mathcal{A})$ muni d'une probabilité P_θ, qui dépend a priori de θ mais que l'on notera P pour des raisons de simplicité.*

2 Estimateurs et estimation ponctuelle

A partir de maintenant, on se donne un n-échantillon $(X_1, ..., X_n)$ de la loi de X, laquelle dépend d'un paramètre $\theta \in \Theta$. Cette loi sera notée μ_θ par la suite.

Définition 2.1. *On appelle estimateur de $g(\theta)$ toute variable aléatoire de la forme $T_n = \varphi_n(X_1, ..., X_n)$, où $(X_1, ..., X_n)$ est un n-échantillon de la loi de X et où φ_n est une fonction de $\mathbb{R}^n$ dans $\mathbb{R}$.*

En d'autres termes, on voit qu'un estimateur de $g(\theta)$ est une variable aléatoire fonction de $(X_1, ..., X_n)$ et indépendante de θ. A noter qu'ici toutes les fonctions φ_n ne conviennent pas forcément. En effet, seules les fonctions φ_n telles que $T_n = \varphi_n(X_1, ..., X_n)$ soit une variable aléatoire conviennent.

Définition 2.2. *Soit $T_n = \varphi_n(X_1, ..., X_n)$ un estimateur de $g(\theta)$. Une estimation de $g(\theta)$ est une réalisation de T_n, c'est-à-dire une valeur de la forme $T_n(\omega) = \varphi_n(X_1(\omega), ..., X_n(\omega))$, où $\omega \in \Omega$.*

D'un point de vue pratique, une estimation de $g(\theta)$ est la valeur que le statisticien accordera à $g(\theta)$ après avoir réalisé un n-échantillon. Un premier exemple d'estimateur (pour l'espérance) est donné par la :

Définition 2.3. *Soit $(X_1, ..., X_n)$ un n-échantillon de la loi de X, laquelle admet une espérance. On appelle moyenne empirique associée à $(X_1, ..., X_n)$ la variable aléatoire :*

$$\overline{X_n} = \frac{X_1 + ... + X_n}{n}.$$

Un deuxième exemple d'estimateur (cette fois-ci pour la variance) est donnée par la :

Définition 2.4. *Soit $(X_1, ..., X_n)$ un n-échantillon de la loi de X, laquelle admet une variance. On appelle variance empirique associée à $(X_1, ..., X_n)$ la variable aléatoire :*

$$\overline{S_n}^2 = \frac{1}{n}\sum_{k=1}^{n}\left(X_k - \overline{X_n}\right)^2.$$

En particulier, l'écart-type empirique est défini comme la racine carrée de la variance empirique.

Définition 2.5. *Soit T_n un estimateur de $g(\theta)$ admettant une espérance pour tout $\theta \in \Theta$. On dit que T_n est sans biais si $E(T_n) = g(\theta)$ pour tout $\theta \in \Theta$. Dans le cas contraire, on dit que T_n est biaisé.*

A noter que l'on écrit souvent l'espérance de T_n sous la forme $E_\theta(T_n)$ (et non $E(T_n)$) dans les sujets de concours, pour indiquer la dépendance vis-à-vis de θ. Cette notation est purement formelle et n'affecte en rien les calculs que l'on fera avec l'espérance, le paramètre θ ne jouant ici aucun rôle. En d'autres termes, un estimateur T_n est sans biais si la valeur moyenne prise par T_n correspond à la valeur du paramètre recherché. A titre d'exemple, étant donné un n-échantillon $(X_1, ..., X_n)$ de la loi de X, dont l'espérance existe et vaut m, on montre facilement par linéarité de l'espérance que :

la moyenne empirique $\overline{X_n} = \dfrac{X_1 + ... + X_n}{n}$ est un estimateur sans biais de l'espérance de X.

En revanche, on peut vérifier par le calcul que la variance empirique $\overline{S_n}^2$ n'est pas un estimateur sans biais de la variance. Par contre, l'estimateur de la variance donnée par :

$$\widehat{S_n}^2 = \frac{n}{n-1}\overline{S_n}^2 = \frac{1}{n-1}\sum_{k=1}^{n}\left(X_k - \overline{X_n}\right)^2.$$

est bien sans biais. Cet estimateur est appelé *la variance empirique corrigée*, et son absence de biais explique pourquoi on l'utilise bien souvent dans des calculs statistiques à la place de la variance empirique.

Définition 2.6. *On dit que $(T_n)_{n\geq 1}$ est une suite d'estimateurs de $g(\theta)$ si, pour tout $n \in \mathbb{N}^*$, $(X_1, ..., X_n)$ est un n-échantillon de la loi de X et si T_n est un estimateur de $g(\theta)$ défini à l'aide de $X_1, ..., X_n$.*

Par abus de notation et de langage, on écrit souvent T_n au lieu de $(T_n)_{n\geq 1}$, et on dit simplement que T_n est un estimateur au lieu de dire que c'est une suite d'estimateurs.

Définition 2.7. *On dit qu'une suite d'estimateurs $(T_n)_{n\geq 1}$ de $g(\theta)$ est convergente si, pour tout $\theta \in \Theta$, la suite $(T_n)_{n\geq 1}$ converge en probabilité vers $g(\theta)$.*

Toujours par abus de langage, on dit souvent que T_n *est un estimateur convergent de* $g(\theta)$. D'après les résultats sur la convergence en probabilité, on a le :

Théorème 2.8. *Si $(T_n)_{n\geq 1}$ est une suite convergente d'estimateurs de $g(\theta)$ et si $f : \mathbb{R} \longrightarrow \mathbb{R}$ est continue, alors $(f(T_n))_{n\geq 1}$ est une suite convergente d'estimateurs de $f(g(\theta))$.*

A l'aide de l'inégalité de Markov, on obtient le :

Théorème 2.9. *Soit $(T_n)_{n\geq 1}$ une suite d'estimateurs de $g(\theta)$, admettant un moment d'ordre 2. Si $\lim\limits_{n\to+\infty} E(T_n) = g(\theta)$ et si $\lim\limits_{n\to+\infty} V(T_n) = 0$, alors $(T_n)_{n\geq 1}$ est convergente.*

En particulier, on voit que, si T_n est un estimateur sans biais de $g(\theta)$ et si $\lim\limits_{n\to+\infty} V(T_n) = 0$, alors T_n est convergent. Dans la démonstration du théorème 2.9, on montre que, pour tout $\varepsilon > 0$:

$$P([T_n - g(\theta)| \geq \varepsilon]) \leq \frac{E((T_n - g(\theta))^2)}{\varepsilon^2} = \frac{(E(T_n) - g(\theta))^2 + V(T_n)}{\varepsilon^2}.$$

En particulier, on voit que, plus l'expression $E((T_n - g(\theta))^2)$ tend vite vers 0, plus l'estimateur T_n converge vite. La quantité $E((T_n - g(\theta))^2)$ permet donc de mesurer si un estimateur est de bonne qualité ou non. C'est ce que l'on appelle le "risque quadratique de l'estimateur", notion qui n'est plus au programme. Dès lors, on retiendra que, plus cette quantité tend vite vers 0, meilleur est l'estimateur. En particulier, on voit que, si T_n est un estimateur sans biais de $g(\theta)$, alors on a $E((T_n - g(\theta))^2) = V(T_n)$, et donc la qualité de T_n en tant qu'estimateur de $g(\theta)$ sera d'autant meilleure que sa variance est petite, voire qu'elle converge vite vers 0. A noter enfin qu'un exemple très classique d'estimateur convergent (dont l'énoncé et la démonstration sont à connaître par cœur, et à reproduire à chaque fois) est donné par le :

Théorème 2.10. *Soit $(X_1, ..., X_n)$ un n-échantillon de la loi de X, laquelle admet une variance. Alors la moyenne empirique $\overline{X}_n$ est un estimateur sans biais et convergent de l'espérance de X.*

3 Estimation par intervalle de confiance

Au paragraphe précédent, on a vu des critères pour juger de la qualité d'un estimateur T_n de $g(\theta)$, comme l'absence de biais ou la convergence. Cependant, même si l'estimateur est de bonne qualité, on ne peut pas être assuré que la valeur $T_n(\omega)$ obtenue lors d'une expérience aléatoire soit une bonne approximation de $g(\theta)$ pour tout $\omega \in \Omega$. La seule chose que l'on puisse dire est qu'il y a de grandes chances que $T_n(\omega)$ soit une bonne approximation de $g(\theta)$. Evidemment, les notions de "grandes chances" et de "bonne approximation" ont besoin d'être précisées. C'est pourquoi on introduit la notion d'estimation par intervalle de confiance. Le principe de cette estimation consiste à trouver un intervalle qui contienne $g(\theta)$ avec une probabilité minimale donnée au départ (on parle alors de niveau de confiance).

Définition 3.1. *Soit $\alpha \in]0,1[$, et soient U_n et V_n deux estimateurs définis à partir d'un n-échantillon $(X_1, ..., X_n)$ de la loi de X, tels que $P([U_n \leq V_n]) = 1$ pour tout $\theta \in \Theta$. On dit que $[U_n, V_n]$ est un intervalle de confiance de $g(\theta)$ au niveau de confiance $1-\alpha$ si, pour tout $\theta \in \Theta$, on a : $P([U_n \leq g(\theta) \leq V_n]) \geq 1-\alpha$.*

A noter que, dans cette définition, le réel α est appelé *le risque*. A noter aussi que les bornes d'un intervalle de confiance sont des variables aléatoires ! Une estimation de l'intervalle de confiance est alors la réalisation de l'intervalle de confiance, c'est-à-dire un intervalle de la forme $[U_n(\omega), V_n(\omega)]$, où $\omega \in \Omega$. On parle aussi d'intervalle de confiance observé ou de fourchette. Par définition d'un intervalle de confiance, l'ensemble des $\omega \in \Omega$ tels que $[U_n(\omega), V_n(\omega)]$ contient $g(\theta)$ est de probabilité $\geq 1-\alpha$.

A présent, on va voir comment obtenir un intervalle de confiance à l'aide de l'inégalité de Markov. Pour ce faire, supposons que l'on recherche un intervalle de confiance pour un estimateur T_n de $g(\theta)$, admettant une variance. D'après l'inégalité de Markov, on sait que, pour tout réel $\gamma > 0$:

$$P([|T_n - g(\theta)| \leq \gamma]) = 1 - P([|T_n - g(\theta)| > \gamma]) \geq 1 - \frac{E((T_n - g(\theta))^2)}{\gamma^2}.$$

Dès lors, on voit que $P([|T_n - g(\theta)| \leq \gamma]) \geq 1-\alpha$ si $1 - \frac{E((T_n-g(\theta))^2)}{\gamma^2} \geq 1-\alpha$, c'est-à-dire si :

$$\gamma \geq \sqrt{\frac{E((T_n - g(\theta))^2)}{\alpha}}.$$

En particulier, si l'on choisit $\gamma = \sqrt{\frac{E((T_n-g(\theta))^2)}{\alpha}}$, alors on trouve que :

$$P\left(\left[T_n - \sqrt{\frac{E((T_n - g(\theta))^2)}{\alpha}} \leq g(\theta) \leq T_n + \sqrt{\frac{E((T_n - g(\theta))^2)}{\alpha}}\right]\right) \geq 1-\alpha.$$

En d'autres termes, un intervalle de confiance $g(\theta)$ au niveau de confiance $1-\alpha$ est donné par :

$$\boxed{[U_n, V_n] = \left[T_n - \sqrt{\frac{E((T_n - g(\theta))^2)}{\alpha}}, T_n + \sqrt{\frac{E((T_n - g(\theta))^2)}{\alpha}}\right].}$$

Si maintenant T_n est un estimateur sans biais de $g(\theta)$ et l'on dispose d'un majorant M_n de $V(T_n)$ indépendant du paramètre θ et facile à calculer, alors on voit que $E((T_n - g(\theta))^2) = V(T_n) \leq M_n$. Dès lors, on obtient par croissance des probabilités que :

$$P\left(\left[T_n - \sqrt{\frac{M_n}{\alpha}} \leq g(\theta) \leq T_n + \sqrt{\frac{M_n}{\alpha}}\right]\right) \geq P\left(\left[|T_n - g(\theta)| \leq \sqrt{\frac{E((T_n - g(\theta))^2)}{\alpha}}\right]\right) \geq 1-\alpha.$$

En particulier, un intervalle de confiance de $g(\theta)$ au niveau de confiance $1-\alpha$ est donné par :

$$\boxed{[U_n, V_n] = \left[T_n - \sqrt{\frac{M_n}{\alpha}}, T_n + \sqrt{\frac{M_n}{\alpha}}\right].}$$

A titre d'exemple, plaçons-nous dans le cas d'une suite (X_n) de variables de Bernoulli indépendantes, de même paramètre p inconnu. Si l'on utilise la moyenne empirique $\overline{X_n}$ pour estimer la valeur de p, alors on voit que cet estimateur est sans biais et que sa variance est majorée par $\frac{1}{4n}$, et ce car $p(1-p) \leq \frac{1}{4}$ pour tout $p \in]0,1[$. Par conséquent, on en déduit le :

Théorème 3.2. *Soit $(X_n)_{n\geq 1}$ une suite de variables de Bernoulli indépendantes, de même paramètre p. Alors un intervalle de confiance de p au niveau de confiance $1-\alpha$ est donné par :*

$$[U_n, V_n] = \left[\overline{X_n} - \frac{1}{2\sqrt{n\alpha}}, \overline{X_n} + \frac{1}{2\sqrt{n\alpha}}\right].$$

En résumé, l'inégalité de Markov (ou de Bienaymé-Tchebychev dans le cas d'un estimateur sans biais de $g(\theta)$) nous permet d'obtenir un intervalle de confiance pour un estimateur quelconque de $g(\theta)$. Par contre, cet intervalle de confiance n'est pas forcément le plus efficace, c'est-à-dire avec le plus petit écart possible. C'est pourquoi il peut être préférable de rechercher un intervalle de confiance plus efficace, c'est-à-dire dont l'écart est plus petit. Ceci nous amène à la :

Définition 3.3. *Soit $\alpha \in]0,1[$, et soient U_n et V_n deux estimateurs définis à partir d'un n-échantillon $(X_1, ..., X_n)$ de la loi de X, tels que $P([U_n \leq V_n]) = 1$ pour tout $\theta \in \Theta$. On dit que $[U_n, V_n]$ est un intervalle de confiance asymptotique de $g(\theta)$ au niveau de confiance $1-\alpha$ si, pour tout $\theta \in \Theta$, il existe une suite (α_n) d'éléments de $[0,1]$, de limite α, telle que, pour tout $n \geq 1$, on a : $P([U_n \leq g(\theta) \leq V_n]) \geq 1 - \alpha_n$.*

En général, on utilisera un intervalle de confiance asymptotique si l'on dispose d'une bonne approximation de la loi de l'estimateur T_n de $g(\theta)$. En effet, ceci permet de faire des calculs plus précis et on obtient ainsi des intervalles de confiance plus efficaces, alors que les inégalités de Markov et de Bienaymé-Tchebychev sont grossières et donnent du coup des intervalles de confiance moins efficaces, car plus grands !

A titre d'exemple, considérons une suite (X_n) de variables de Bernoulli indépendantes, de même paramètre p inconnu à estimer. Pour n assez grand, c'est-à-dire si $n \geq 30$, on sait que la loi de $\overline{X_n}^*$ est bien approchée par la loi normale $\mathcal{N}(0,1)$ d'après le théorème limite central. Si l'on pose $\sigma^2 = p(1-p)$, alors on trouve que, pour tout $\gamma > 0$:

$$P\left(\left[\left|\overline{X_n} - p\right| \leq \gamma\right]\right) = P\left(\left[|\overline{X_n}^*| \leq \frac{\gamma}{\frac{\sigma}{\sqrt{n}}}\right]\right) = P\left(\left[-\frac{\gamma}{\frac{\sigma}{\sqrt{n}}} \leq \overline{X_n}^* \leq \frac{\gamma}{\frac{\sigma}{\sqrt{n}}}\right]\right).$$

Dès lors, on obtient en utilisant l'approximation par la loi normale centrée réduite que :

$$P\left(\left[\left|\overline{X_n} - p\right| \leq \gamma\right]\right) \simeq \Phi\left(\frac{\gamma}{\frac{\sigma}{\sqrt{n}}}\right) - \Phi\left(-\frac{\gamma}{\frac{\sigma}{\sqrt{n}}}\right) = 2\Phi\left(\frac{\gamma}{\frac{\sigma}{\sqrt{n}}}\right) - 1.$$

En particulier, on trouve avec cette approximation que $P\left(\left[\left|\overline{X_n} - p\right| \leq \gamma\right]\right) \geq 1 - \alpha$ si :

$$\gamma \geq \frac{\sigma}{\sqrt{n}} \Phi^{-1}\left(1 - \frac{\alpha}{2}\right).$$

Comme $\sigma^2 = p(1-p) \leq \frac{1}{4}$ pour tout $p \in]0,1[$, il s'ensuit que $P\left(\left[\left|\overline{X_n} - p\right| \leq \gamma\right]\right) \geq 1 - \alpha$ si :

$$\gamma \geq \frac{1}{2\sqrt{n}} \Phi^{-1}\left(1 - \frac{\alpha}{2}\right).$$

Par conséquent, on en déduit le :

Théorème 3.4. *Soit $(X_n)_{n\geq 1}$ une suite de variables de Bernoulli indépendantes, de même paramètre p. Alors un intervalle de confiance asymptotique de p au niveau de confiance $1-\alpha$ est donné par :*

$$[U_n, V_n] = \left[\overline{X_n} - \frac{1}{2\sqrt{n}} \Phi^{-1}\left(1 - \frac{\alpha}{2}\right), \overline{X_n} + \frac{1}{2\sqrt{n}} \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)\right].$$

A noter qu'avec le théorème 3.4, on obtient un intervalle de confiance avec une précision de $0,01$ et un niveau de confiance de $0,99$ pour $n \geq 16577$. A titre de comparaison, on obtient avec le théorème 3.2 un tel intervalle de confiance pour $n \geq 250000$!