Leçon 448 : Exemples d'utilisation d'intervalles de fluctuation et d'intervalles de confiance

Intervalle de fluctuation

Soit X une variable aléatoire réelle, admettant une espérance finie $\mathbb{E}(X)$. Soit $s \in]0,1[$ un nombre appelé *seuil*. Le nombre $\alpha = 1 - s$ est appelé *risque*. On s'intéresse à la situation où

$$0 < \alpha \ll s < 1 \ .$$

On appelle *intervalle de fluctuation* un intervalle $I = [a,b] \subset \mathbb{R}$ tel que

$$\mathbb{P}\big\{X \in I\big\} = s \ .$$

On choisit d'habitude l'intervalle I symétrique autour de l'espérance $\mathbb{E}(X)$, donc de la forme

$$I = \big[\mathbb{E}(X) - c, \mathbb{E}(x) + c\big] \ .$$

Exemple 1. On suppose que X suit une loi normale centrée réduite. On a donc $\mathbb{E}(X) = 0$ et

$$\mathbb{P}\big\{X \in [-c,c]\big\} = \int_{-c}^{c} \frac{\mathrm{e}^{-x^2/2}}{\sqrt{2\pi}} \,\mathrm{d}x \ .$$

Soit

$$\Phi(t) = \mathbb{P}\big\{X \leqslant t\big\} = \int_{-\infty}^{t} \frac{\mathrm{e}^{-x^2/2}}{\sqrt{2\pi}} \,\mathrm{d}x$$

la fonction de répartition de X. Comme $\Phi(-c) = 1 - \Phi(c)$, on obtient

$$\mathbb{P}\big\{X \in [-c,c]\big\} = \Phi(c) - \Phi(-c) = 2\Phi(c) - 1 \ .$$

L'intervalle de fluctuation correspondant au seuil $s = 1 - \alpha$ est donc égal à $[-c,c]$ où c vérifie l'équation

$$2\Phi(c) - 1 = s \qquad \Rightarrow \qquad \Phi(c) = \frac{1+s}{2} = 1 - \frac{\alpha}{2} \ .$$

Par approximation numérique de Φ, on trouve, pour $s = 0{,}95$,

$$c = \Phi^{-1}(0{,}975) = 1{,}96\ldots \ .$$

L'intervalle de fluctuation correspondant au seuil $s = 0{,}95$ est donc approximativement $[-1{,}96; 1{,}96]$.

Approximation d'un intervalle de fluctuation

1. Par l'inégalité de Bienaymé–Tchebychev : pour tout $a > 0$,

$$\mathbb{P}\big\{\big|X - \mathbb{E}(X)\big| \geqslant a\big\} \leqslant \frac{1}{a^2} \operatorname{Var}(X) \ .$$

   Ceci permet de majorer l'intervalle de fluctuation cherché.

2. Par le théorème central limite : si $(X_i)_{i\geqslant 1}$ est une suite i.i.d. de variables aléatoires d'espérance μ et de variance σ^2, et $S_n = \sum_{i=1}^n X_i$, alors la *variable pivotale*

$$\widehat{S}_n = \frac{S_n - n\mu}{\sqrt{n}\sigma}$$

converge en loi, lorsque $n \to \infty$, vers une variable aléatoire de loi normale centrée réduite. Ceci fournit une approximation de l'intervalle de fluctuation pour n grand.

Exercice 1. On suppose que X suit une loi binomiale de paramètres (n, p). Déterminer un intervalle de fluctuation approximatif au seuil s pour les variables X et X/n,

1. en utilisant l'inégalité de Bienaymé–Tchebychev ;
2. en utilisant le théorème central limite.

Comparer les résultats. Que peut-on en conclure ?

Application : Sachant que 80% des Françaises et Français croient en une théorie du complot, déterminer un intervalle de fluctuation au seuil $0{,}95$ pour le nombre de personnes dans un échantillon de 1000 français croyant en une théorie du complot.

Exercice 2. La compagnie aérienne Air Metik constate qu'en moyenne 10% des personnes ayant acheté un billet pour un vol Paris–Marseille, assuré par un Airbus d'une capacité de 300 places, ne se présentent pas à l'embarquement.

1. Donner un intervalle de fluctuation, au seuil de $0{,}95$, du nombre de personnes personnes parmi 300 ayant acheté un billet qui se présentent à l'embarquement.
2. Estimer le nombre de places que la compagnie peut vendre pour que la probabilité que le vol soit « surbooké » soit inférieure à 5%.

Estimateur

L'intervalle de confiance apparaît dans la situation inverse de celle considérée jusqu'à présent : on voudrait estimer un ou plusieurs paramètres de la loi d'une variable X à partir d'un nombre fini d'observations $X_1, \ldots, X_n$, des variables i.i.d. de même loi que X.

Exemple 2. On effectue un sondage dans un échantillon de taille n d'une population de taille N, avec N très grand devant n. Si le sondage n'offre que deux réponses (« Oui » on « Non »), on modélise la situation par n variables i.i.d., de loi de Bernoulli de paramètre p, égal à la proportion d'individus de la population qui auraient répondu « Oui ». Il s'agit donc d'estimer p à partir de n variables i.i.d. de loi de Bernoulli de paramètre p.

Soit X une variable aléatoire réelle, suivant une loi dépendant d'un paramètre θ. Cette loi est supposée appartenir à une famille donnée $\{\mathbb{P}_\theta : \theta \in \Theta\}$, mais la valeur de θ est inconnue.

On dispose de n échantillons $X_1, \dots, X_n$ de variables i.i.d. de loi $\mathbb{P}_\theta$. Un *estimateur* de θ est une fonction $f : \mathbb{R}^n \to \Theta$ telle que $\hat{\theta} = f(X_1, \dots, X_n)$ donne une estimation du paramètre θ.

Exemple 3. On suppose que $\theta = \mathbb{E}(X)$ est l'espérance de X. On appelle *moyenne empirique* l'estimateur

$$\overline{X}_n = \frac{1}{n}(X_1 + \cdots + X_n) \, .$$

Qualité d'un estimateur

On dispose de plusieurs mesures de la qualité d'un estimateur $\hat{\theta} = f(X_1, \dots, X_n)$:

1. On dit que $\hat{\theta}$ est un estimateur *sans biais* si $\mathbb{E}(f(X_1, \dots, X_n)) = \theta$.
2. Le *risque quadratique* de $\hat{\theta}$ est la quantité

$$R_\theta(f) = \mathbb{E}\big((f(X_1, \dots, X_n) - \theta)^2\big) \, .$$

3. L'estimateur est *consistant* si

$$\lim_{n \to \infty} \mathbb{P}\big\{|f(X_1, \dots, X_n) - \theta| > \varepsilon\big\} = 0$$

pour tout $\varepsilon > 0$ et tout $\theta \in \Theta$, c'est-à-dire que $f(X_1, \dots, X_n)$ converge vers θ en probabilité.

Exercice 3. On suppose que X admet une variance finie. Montrer que la moyenne empirique $\overline{X}_n$ est un estimateur sans biais et consistant de l'espérance $\theta = \mathbb{E}(X)$, et calculer son risque quadratique.

Intervalle de confiance

Un *intervalle de confiance* au seuil $s \in]0, 1[$ est un intervalle $I \subset \mathbb{R}$, dépendant des observations $X_1, \dots, X_n$, tel que

$$\mathbb{P}\big\{\theta \in I\big\} \geqslant s \, .$$

Exemple 4. On suppose que la variable aléatoire X suit une loi de Bernoulli de paramètre inconnu $p \in]0, 1[$. Son espérance μ est égale à p, et sa variance σ^2 vaut

$p(1-p)$. On dispose de n observations $X_1, \dots, X_n$ de variables i.i.d. suivant cette loi. Notons que $\mathbb{E}(X) = p$ et que l'estimateur de moyenne empirique

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

suit une loi binomiale de paramètres (n, p). Par le théorème central limite, la variable pivotale

$$\widehat{S}_n = \frac{n\overline{X}_n - n\mu}{\sqrt{n}\sigma} = \sqrt{n} \frac{\overline{X}_n - p}{\sqrt{p(1-p)}}$$

suit une loi proche, pour n grand, d'une loi normale centrée réduite. Comme on ne connaît pas p, on invoque la consistance de l'estimateur $\overline{X}_n$ pour approcher p par $\overline{X}_n$. On applique alors le résultat de l'exemple 1 pour obtenir l'intervalle de confiance (approximatif) au seuil s de

$$I(s) = \left[\overline{X}_n - c(s) \sqrt{\frac{\overline{X}_n(1-\overline{X}_n)}{n}}, \overline{X}_n + c(s) \sqrt{\frac{\overline{X}_n(1-\overline{X}_n)}{n}} \right],$$

où $c(s)$ est l'unique solution de $2\Phi(c) = s + 1$.

Exercice 4. Dans un sondage effectué auprès de 1000 personnes en France, 800 ont indiqué croire à au moins une théorie du complot.

1. Déterminer un intervalle de confiance, au seuil $s = 0{,}95$, pour la proportion des Français(es) croyant à au moins une théorie du complot.
2. Combien de personnes aurait-il fallu interroger pour obtenir un intervalle de confiance au seuil $s = 0{,}95$ de taille $0{,}02$?