CM6

Théorèmes limites et tests statistiques

1 Théorèmes limites

1.1 Inégalité de Bienaymé-Tchebychev

Théorème 1 Inégalité de Bienaymé-Tchebychev
Soit X une variable aléatoire admettant une espérance et une variance.
Alors, pour tout $\varepsilon > 0$,

$$\mathbb{P}(|X - \mathbb{E}(X)| \geq \varepsilon) \leq \frac{V(X)}{\varepsilon^2}$$

$$\mathbb{P}(|X - \mathbb{E}(X)| < \varepsilon) \geq 1 - \frac{V(X)}{\varepsilon^2}.$$

Preuve. Comme d'habitude, il faut différencier les cas. Nous allons faire la démonstration dans le cas réel, en supposant que X admet une densité de probabilité f. Puisque X admet une variance,

$$\begin{aligned} V(X) &= \mathbb{E}\big((X - \mathbb{E}(X))^2\big) \\ &= \int_{-\infty}^{+\infty} (x - \mathbb{E}(X))^2 \, f(x) \, dx \\ &\geq \int_{|x-\mathbb{E}(X)|\geq\varepsilon} (x - \mathbb{E}(X))^2 \, f(x) \, dx \\ &\geq \int_{|x-\mathbb{E}(X)|\geq\varepsilon} \varepsilon^2 \, f(x) \, dx = \varepsilon^2 \, \mathbb{P}\big(|X - \mathbb{E}(X)| \geq \varepsilon\big). \end{aligned}$$

D'où la première inégalité. L'autre inégalité est une conséquence directe de la première :

$$\begin{aligned} \mathbb{P}(|X - \mathbb{E}(X)| < \varepsilon) &= 1 - \mathbb{P}(|X - \mathbb{E}(X)| \geq \varepsilon) \\ &\geq 1 - \frac{V(X)}{\varepsilon^2}. \end{aligned}$$

□

1.2 Lois des grands nombres

L'inégalité de Bienaymé-Tchebychev nous permet de trouver deux propositions intéressantes.

Proposition 2 Loi faible des grands nombres
Soit $(X_j)_{j\in\mathbb{N}^*}$ une suite de variables aléatoires ***indépendantes et de même loi****, admettant une espérance m et une variance σ^2. Soit $M_n = \frac{1}{n}\sum_{j=1}^{n} X_j$. Alors, M_n converge en probabilité vers m, c'est-à-dire, pour tout $\varepsilon > 0$,*

$$\lim_{n\mapsto+\infty} \mathbb{P}(|M_n - m| < \varepsilon) = 1.$$

Preuve. Calculons l'espérance et la variance de M_n :

$$\mathbb{E}(M_n) = \frac{1}{n}\sum_{j=1}^{n} \mathbb{E}(X_j) = \frac{1}{n}\sum_{j=1}^{n} m = m.$$

$$V(M_n) \underset{\perp}{=} \frac{1}{n^2}\sum_{j=1}^{n} V(X_j) = \frac{1}{n^2}\sum_{j=1}^{n} \sigma^2 = \frac{\sigma^2}{n}.$$

Alors, d'après l'inégalité de Bienaymé-Tchebychev, pour tout $\varepsilon > 0$, on a :

$$1 - \frac{V(M_n)}{\varepsilon^2} \leq \mathbb{P}(|M_n - \mathbb{E}(M_n)| < \varepsilon) \leq 1.$$

Ce qui peut encore s'écrire :

$$1 - \frac{\sigma^2}{n\varepsilon^2} \leq \mathbb{P}(|M_n - m| < \varepsilon) \leq 1.$$

Par passage à la limite ($n \longrightarrow +\infty$), on obtient :

$$1 = \lim_{n \mapsto +\infty} 1 - \frac{\sigma^2}{n\varepsilon^2} \leq \lim_{n \mapsto +\infty} \mathbb{P}(|M_n - m| < \varepsilon) \leq 1.$$

D'où le résultat de la proposition. □

On a en fait le résultat plus fort suivant.

Théorème 3 Loi forte des grands nombres
Soit $(X_j)_{j \in \mathbb{N}^*}$ une suite de variables aléatoires ***indépendantes et de même loi****, admettant une espérance m. Alors, M_n converge presque sûrement vers m, c'est-à-dire il existe un évènement A de probabilité 1 tel que*

$$\forall \omega \in A, \quad \lim_{n \mapsto +\infty} M_n(\omega) = m.$$

Si on répète une même expérience un grand nombre de fois de manière identique et que l'on regarde le nombre de fois où un résultat R apparaît, la loi forte des grands nombres montre que la fréquence empirique d'apparition de R tend vers la probabilité de R quand le nombre d'expériences tend vers l'infini. Par exemple, si on lance une pièce équilibrée un grand nombre de fois, la suite des fréquences relatives des piles que l'on obtient tend avec probabilité 1 vers 50.

1.3 Théorème central-limite

En reprenant la discution précédente, on peut se demander quelle est l'erreur faite quand on remplace directement la fréquence obtenue après n expériences par 50. Le théorème ci-dessous permettra de répondre d'une certaine manière à cette question.

Théorème 4 Théorème central-limite
Soit $(X_j)_{j \in \mathbb{N}^*}$ une suite de variables aléatoires ***indépendantes et de même loi*** *de carré intégrable (elles admettent donc une espérance et une variance).*
Soit $S_n = \sum_{j=1}^{n} X_j$. Alors, pour tout x et y tels que $x \leq y$,

$$\mathbb{P}\left(x \leq \frac{S_n - \mathbb{E}(S_n)}{\sqrt{V(S_n)}} \leq y\right) \longrightarrow \mathbb{P}(x \leq Z \leq y),$$

où Z désigne une variable aléatoire de loi gaussienne centrée réduite, donc

$$\mathbb{P}(x \leq Z \leq y) = \frac{1}{\sqrt{2\pi}} \int_x^y e^{-\frac{z^2}{2}} \mathrm{dz}.$$

On a un cas particulier de ce théorème, un corollaire quand $X_j \sim \mathcal{B}(p)$, c'est-à-dire quand $S_n \sim b(n,p)$.

Théorème 5 Théorème de De Moivre-Laplace
Si $S_n \sim b(n,p)$, alors quand $n \longrightarrow +\infty$,

$$\mathbb{P}\left(x \leq \frac{S_n - np}{\sqrt{np(1-p))}} \leq y\right) \longrightarrow \frac{1}{\sqrt{2\pi}} \int_x^y e^{-\frac{z^2}{2}} \mathrm{dz}.$$

2 Estimation statistique

Dans la démarche statistique, on cherche à traiter des données, en général une série d'observations $x_1, \cdots, x_n$. On suppose qu'elles correspondent à des réalisations de variables aléatoires $X_1, \cdots, X_n$, et on cherche une distribution théorique du vecteur $(X_k)_{1 \leq k \leq n}$ qui reflète correctement les propriétés des observations $(x_k)_{1 \leq k \leq n}$.

Concrètement, les valeurs x_k sont donc connues. Par exemple, x_k représente la durée de vie du moteur de la voiture numéro k que l'on a choisi d'observer. Si le fabriquant est un grand constructeur, on ne peut pas recenser la durée de vie de toutes les voitures fabriquées donc on ne considère qu'un échantillon de taille raisonnable. Le constructeur aimerait à partir de ces valeurs améliorer la fabrication des moteurs. Il serait alors utile de

connaitre la loi sous-jacente à la durée de vie d'un moteur. Cette loi est inconnue, mais à partir de l'échantillon $(x_k)_k$, on peut cependant estimer certaines valeurs, comme par exemple la durée de vie moyenne d'un moteur ; on parle alors de problèmes d'**estimation** et d'**intervalle de confiance**.

Dans toute la suite, on se restreint au cas le plus simple : on suppose que les variables aléatoires $(X_k)_{1\leq k\leq n}$ sont **indépendantes et de même loi** et que cette **loi appartient à une collection fixée** a priori, que l'on notera $\{\mathbb{P}_\theta, \theta \in \Theta\}$.

2.1 Estimation ponctuelle

On considère un échantillon $(X_k)_{1\leq k\leq n}$ tel que chaque X_k est à valeurs dans un ensemble E (par exemple, $E = \mathbb{R}$) et de même loi $\mathbb{P}_\theta$ qui dépend d'un paramètre inconnu $\theta \in \Theta$. Le but est d'estimer la valeur de θ à partir des valeurs des X_k.

Définition 6 *Un **estimateur** de θ est une variable aléatoire $\widehat{\theta_n}$ telle qu'il existe une fonction $F_n : E_n \to \Theta$ avec $\widehat{\theta_n} = Fn(X_1, X_2, \cdots, Xn)$.*

Remarque : Un estimateur est donc une fonction de l'échantillon. Il ne doit pas dépendre de θ, mais seulement des observations $(X_k)_{1\leq k\leq n}$.

Le problème est maintenant de choisir la fonction F_n de façon à estimer correctement θ. Pour tout θ dans Θ, on note $\mathbb{E}_\theta$ l'espérance par rapport à la loi $\mathbb{P}_\theta$.

Définition 7 *L'estimateur $\widehat{\theta_n}$ est un **estimateur consistant** de θ pour la valeur θ si $\widehat{\theta_n} \longrightarrow \theta$ presque sûrement par rapport à $\mathbb{P}_\theta$, quand n tend vers l'infini.*

La consistance est la propriété la plus importante d'un estimateur. Ainsi, pour de grands échantillons, l'approximation de θ par $\widehat{\theta_n}$ est correcte. Par exemple, la loi forte des grands nombres affirme que la moyenne M_n est un estimateur consistant de l'espérance.

Définition 8 *Le **biais d'un estimateur** $\widehat{\theta_n}$ de θ est*

$$B(\widehat{\theta_n}, \theta) := \mathbb{E}_\theta(\widehat{\theta_n}) - \theta.$$

*L'estimateur $\widehat{\theta_n}$ est dit **sans biais** si $B(\widehat{\theta_n}, \theta) = 0$ pour toute valeur de θ dans Θ, il est dit biaisé dans le cas contraire.*

Le biais $B(\widehat{\theta_n}, \theta)$ est donc un nombre, qui dépend de $\widehat{\theta_n}$ et de la fonction F_n définissant l'estimateur $\widehat{\theta_n}$.

Proposition 9 Moyenne empirique
La moyenne empirique, définie par $M_n := \frac{1}{n}\sum_{k=1}^n X_k$ est un estimateur sans biais et consistant de la moyenne $\mathbb{E}(X)$.

Preuve.

Proposition 10 Variance empirique

- *La variance empirique, définie par*

$$\hat{\sigma}_n^2 := \frac{1}{n}\sum_{k=1}^n X_k^2 - M_n^2 = \frac{1}{n}\sum_{k=1}^n (X_k - M_n)^2$$

est un estimateur consistant de la variance $\mathrm{Var}(X)$*, et biaisé.*

– *On définit un estimateur sans biais et consistant de la variance* $\mathrm{Var}(X)$ *en posant*

$$V_n := \frac{1}{n-1}\sum_{k=1}^{n}(X_k - M_n)^2.$$

Preuve.

2.2 Intervalle de confiance

Dans la section précédente on proposait une valeur unique $\widehat{\theta_n}$ pour estimer θ. On veut maintenant proposer un ensemble $I_n \subset \Theta$ aussi petit que possible, tel que θ appartienne souvent à I_n.

Comme précédemment, on ne dispose pour construire I_n que des observations $(X_k)_{1\leq k\leq n}$, que l'on suppose indépendantes et de même loi $\mathbb{P}_\theta$ pour une certaine valeur inconnue du paramètre $\theta \in \Theta$.

Définition 11 *Un* ***intervalle de confiance*** *au niveau a est un intervalle aléatoire I_n qui ne dépend que de l'échantillon $X_1, \cdots, X_n$, mais pas de θ, et tel que, pour tout θ dans Θ,*

$$\mathbb{P}_\theta(\theta \in I_n) \geq a.$$

Le nombre $1-a$ représente le taux d'erreur maximal que l"on accepte en prédisant que I_n contient θ.

Soit $\mathcal{E}$ une épreuve et A un événement lié à cette épreuve. Notons $p = \mathbb{P}(A)$. On ne connait pas p et on cherche à le déterminer. Pour cela, on répète l'épreuve n fois, n grand, et on note X_j la variable aléatoire qui vaut 1 si A est réalisé à la j^{eme} épreuve, 0 sinon. Alors $S_n = \sum_{j=1}^{n} X_j$ représente le nombre de fois où l'événement est réussi et $M_n = \frac{S_n}{n}$ sa fréquence de réussite. On a vu ci-dessus que M_n est un estimateur consistant et sans biais de p. On cherche un intervalle de confiance pour p.

2.2.1 Intérvalle de confiance via l'inégalité de Bienaymé-Tchebychev

Dans la pratique, on suppose qu'on a réalisé n fois l'expérience et que l'événement A a été réussi S fois. Il a donc une fréquence de réussite expérimentale de $\frac{S}{n}$. Par exemple, on considère l'épreuve "jeter une pièce" et A l'événement "obtenir face". On répète 100 fois l'expérience ($n = 100$) et on obtient 54 fois face ($S = 54$). Cherchons un intervalle de confiance pour $p = \mathbb{P}(A)$ au niveau de confiance $\alpha = 0,95$ grâce à l'inégalité de Bienaymé-Tchebychev.

On sait que $S_n \sim b(n,p)$ donc $\mathbb{E}(S_n) = np$ et $V(S_n) = np(1-p)$. Comme $M_n = \frac{S_n}{n}$, alors $\mathbb{E}(M_n) = p$ et $V(M_n) = \frac{p(1-p)}{n}$. D'après l'inégalité de Bienaymé-Tchebychev, dont les hypothèses sont ici respectées, pour tout $\varepsilon > 0$,

$$\mathbb{P}(|M_n - \mathbb{E}(M_n)| \leq \varepsilon) \geq 1 - \frac{V(M_n)}{\varepsilon^2}$$

donc

$$\mathbb{P}(|M_n - p| \leq \varepsilon) \geq 1 - \frac{p(1-p)}{n\varepsilon^2}.$$

Alors, on obtient :

$$\mathbb{P}(|M_n - p| \leq \varepsilon) \geq 1 - \frac{p(1-p)}{n\varepsilon^2} \geq 1 - \frac{1}{4n\varepsilon^2}.$$

On connaît n, la valeur expérimentale de M_n qui est $\frac{S}{n} = \frac{54}{100}$, il suffit de choisir ε tel que $1 - \frac{1}{4n\varepsilon^2} = \alpha$, c'est-à-dire $\varepsilon^2 = \frac{1}{4n(1-\alpha)}$.

Il suffit de prendre $\varepsilon = \frac{1}{2\sqrt{n(1-\alpha)}} = \frac{1}{2\sqrt{5}}$ pour que l'on ait :

$$\mathbb{P}\left(\left|\frac{S}{n} - p\right| \leq \varepsilon\right) \geq \alpha$$

c'est-à-dire :

$$\mathbb{P}\left(\frac{S}{n} - \varepsilon \leq p \leq \frac{S}{n} + \varepsilon\right) \geq \alpha.$$

Ce qui nous donne :

$$\mathbb{P}\left(\frac{54}{100} - \frac{1}{2\sqrt{5}} \leq p \leq \frac{54}{100} + \frac{1}{2\sqrt{5}}\right) \geq \alpha.$$

On a donc un intervalle de confiance pour p au niveau de confiance 0,95 :

$$\left[054 - \frac{1}{2\sqrt{5}}; 0,54 + \frac{1}{2\sqrt{5}}\right] \subset [0,31; 0,77].$$

On peut prendre, d'après un lemme précédent, n'importe quel intervalle qui contient $\left[0,54 - \frac{1}{2\sqrt{5}}; 0,54 + \frac{1}{2\sqrt{5}}\right]$, cela reste un intervalle de confiance au niveau 0,95.

2.2.2 Intervalle de confiance par le TCL

Reprenons l'exemple précédent. On sait que $S_n \sim b(n,p)$ et on cherche un intervalle de confiance pour p au niveau de confiance $\alpha = 0,95$. Les données expérimentales sont toujours $n = 100$ et $S = 54$. On peut appliquer le théorème de De Moivre-Laplace, et comme n est "grand", on estime que :

$$Z_n = \frac{S_n - np}{\sqrt{np(1-p)}} \simeq Y \sim \mathcal{N}(0,1).$$

En fait, Z_n ne suit pas vraiment la loi normale $\mathcal{N}(0,1)$, mais puisque n est grand, on estime que Z_n est très proche de la limite en loi $Y \sim \mathcal{N}(0,1)$ et qu'il y a égalité. Alors, pour tout $\varepsilon > 0$,

$$\mathbb{P}\left(\left|\frac{S_n - np}{\sqrt{np(1-p)}}\right| \leq \varepsilon\right) = 2\,\mathbb{P}(0 \leq Y \leq \varepsilon) = 2\int_0^{\varepsilon} \frac{1}{\sqrt{2\pi}}\, e^{-x^2/2}\, dx.$$

En utilisant la table de la loi normale, on choisit ε tel que $2\,\mathbb{P}(0 \leq Y \leq \varepsilon) \geq \alpha$. Ici, en prenant $\varepsilon = 1,96$, on obtient exactement $2\,\mathbb{P}(0 \leq Y \leq 1,96) = 0,95$. Donc :

$$\mathbb{P}(|S_n - np| \leq 1,96\sqrt{np(1-p)}) = 0,95.$$

En utilisant la majoration $p(1-p) \leq \frac{1}{4}$, on obtient :

$$\mathbb{P}\left(|S_n - np| \leq 1,96\sqrt{\frac{n}{4}}\right) \geq 0,95.$$

Ce qu'on peut encore écrire :

$$\mathbb{P}\left(\frac{S_n - 1,96\sqrt{\frac{n}{4}}}{n} \leq p \leq \frac{S_n + 1,96\sqrt{\frac{n}{4}}}{n}\right) \geq 0,95.$$

On obtient donc un intervalle de confiance pour p au niveau de confiance 0,95 :

$$\left[\frac{54 - 1,96 \times 5}{100}; \frac{54 + 1,96 \times 5}{100}\right] = [0,442; 0,638].$$

Remarque : En général, les résultats obtenus par la technique du TCL (ou de De Moivre-Laplace) sont meilleurs que ceux obtenus par l'inégalité de Bienaymé-Tchebychev. Ceci vient de ce que le TCL utilise des hypothèses plus contraignantes que l'inégalité de Bienaymé-Tchebychev. Les résultats sont donc meilleurs. Il faut toutefois se rappeler de la différence majeure entre les deux approches : l'inégalité de Bienaymé-Tchebychev donne une majoration d'une probabilité (c'est en fait une estimation d'une erreur d'approximation), le TCL donne seulement une approximation d'une probabilité.