Chapitre 10

Estimation

10.1 Sommes de variables aléatoires

Soient $X_1, \dots, X_n$ $n \geq 2$ variables aléatoires définies sur un espace probabilisé (Ω, P). On s'intéresse à la variable aléatoire $S_n = \sum_{k=1}^{n} X_k = X_1 + \cdots + X_n$. Précisément, on souhaiterait une formule simple permettant de calculer $E(S_n)$ et $V(S_n)$.

Proposition 1 (Espérance de la somme de n variables aléatoires)

Soient $X_1, \dots, X_n$ $n \geq 2$ variables aléatoires admettant une espérance. Alors, par *linéarité*,

$$E\left(\sum_{k=1}^{n} X_k\right) = \sum_{k=1}^{n} E(X_k).$$

Remarques :
- Plus simplement, $E(X_1 + \cdots + X_n) = E(X_1) + \cdots + E(X_n)$.
- On remarquera qu'il n'est fait mention d'*aucune* hypothèse d'indépendance.

On définit à présent la notion d'indépendance pour plus de deux variables aléatoires $X_1, \dots, X_n$.

Définition 2 (Indépendance mutuelle de variables aléatoires)

- Les variables aléatoires $X_1, \dots, X_n$ sont mutuellement indépendantes si et seulement si pour tout choix de n intervalles réels $I_1, \dots, I_n$, les événements $[X_1 \in I_1], \dots, [X_n \in I_n]$ sont mutuellement indépendants.
- Les variables aléatoires de la suite $(X_n)_{n \in \mathbb{N}^*}$ sont dites mutuellement indépendantes si et seulement si, pour tout entier $n \geq 1$, les variables aléatoires $X_1, \dots, X_n$ sont mutuellement indépendantes.

Remarque : En exercices, l'indépendance des variables aléatoires étudiées est toujours donnée en *hypothèse*.

On peut à présent donner une formule simple pour la variance d'une somme de variables aléatoires indépendantes.

Proposition 3 (Variance de la somme de n variables aléatoires indépendantes)

Soient $X_1, \dots, X_n$ $n \geq 2$ variables aléatoires indépendantes admettant une variance. Alors, par *indépendance*,

$$V\left(\sum_{k=1}^{n} X_k\right) = \sum_{k=1}^{n} V(X_k).$$

Remarque : Plus simplement, en cas d'indépendance, $V(X_1 + \cdots + X_n) = V(X_1) + \cdots + V(X_n)$.

Par exemple, si $X_1, \ldots, X_n$ sont n variables aléatoires indépendantes suivant toutes la loi de Bernoulli de paramètre $p \in [0,1]$, alors par linéarité, $E\left(\sum_{k=1}^{n} X_k\right) = \sum_{k=1}^{n} E(X_k) = \sum_{k=1}^{n} p = np$ et, par indépendance, $V\left(\sum_{k=1}^{n} X_k\right) = \sum_{k=1}^{n} V(X_k) = \sum_{k=1}^{n} p(1-p) = np(1-p)$. On retrouve l'espérance et la variance d'une variable aléatoire suivant la loi binomiale de paramètres n et p.

10.2 Inégalité de Markov, inégalité de Bienaymé-Tchebychev

On présente deux outils techniques qui seront d'un usage courant dans la section suivante.

Proposition 4 (Inégalité de Markov)

Soit X une variable aléatoire positive admettant une espérance. Alors,

$$\forall a > 0, P([X \geq a]) \leq \frac{E(X)}{a}.$$

Sous l'hypothèse que X admet un moment d'ordre 2, en appliquant l'inégalité de Markov à la variable aléatoire positive $(X - E(X))^2$ et au réel strictement positif $a = \varepsilon^2$, on obtient l'inégalité de Bienaymé-Tchebychev.

Théorème 5 (Inégalité de Bienaymé-Tchebychev)

Soit X une variable aléatoire admettant un moment d'ordre 2. Alors,

$$\forall \varepsilon > 0, P([|X - E(X)| \geq \varepsilon]) \leq \frac{V(X)}{\varepsilon^2}.$$

10.3 Estimation

$\Longrightarrow$ Position du problème : On considère un phénomène aléatoire qu'on modélise à l'aide d'une variable aléatoire X dont la loi dépend d'un paramètre θ inconnu. Le problème de l'estimation consiste alors à déterminer une valeur approchée du paramètre θ à partir d'un échantillon de données $x_1, \ldots, x_n$ obtenues en observant n fois le phénomène. On suppose que cet échantillon est la réalisation de n variables aléatoires $X_1, \ldots, X_n$ indépendantes et de même loi que X.

10.3.1 Estimation ponctuelle

Dans le cadre de notre programme, le paramètre considéré θ sera déterminé par la moyenne de la variable aléatoire X, c'est-à-dire que $\theta = E(X)$. Le résultat sur lequel se fonde notre méthode d'estimation est la loi faible des grands nombres.

Théorème 6 (Loi faible des grands nombres)

Soit $(X_n)_{n \in \mathbb{N}^*}$ une suite de variables aléatoires indépendantes admettant une même espérance m et une même variance. On note pour tout $n \in \mathbb{N}^*$, $\overline{X}_n = \dfrac{X_1 + \cdots + X_n}{n}$. Alors,

$$\forall \varepsilon > 0, \lim_{n \to +\infty} P([|\overline{X}_n - m| \geq \varepsilon]) = 0.$$

On retiendra donc que lorsque n est suffisamment grand, $\overline{X}_n \approx E(X)$ et donc $\theta \approx \overline{X}_n$.

Proposition 7 (Estimation du paramètre par la moyenne empirique)

La réalisation de $\overline{X}_n = \dfrac{X_1 + \cdots + X_n}{n}$ observée sur l'échantillon $x_1, \ldots, x_n$ est l'estimation du paramètre obtenue sur cet échantillon.

Traitons deux exemples concrets.

⋆ 1ier exemple : Une pièce a une probabilité $p_0 \in]0,1[$ de tomber sur « pile ». On lance 1000 fois cette pièce. On obtient 520 « pile » et 480 « face ». Donnons une estimation du paramètre p_0.
Soit X une variable aléatoire suivant la loi de Bernoulli de paramètre p_0. L'obtention de « pile » correspond à $[X = 1]$ et celle de « face » à $[X = 0]$. On dispose de 1000 réalisations $x_1, \ldots, x_{1000}$ dont 520 sont des 1 et 480 sont des 0. On suppose que cet échantillon est la réalisation de 1000 variables aléatoires $X_1, \ldots, X_{1000}$ indépendantes et de même loi que X.
La réalisation de $\overline{X}_{1000} = \dfrac{X_1 + \cdots + X_{1000}}{1000}$ sur cet échantillon vaut $\dfrac{520 \times 1 + 480 \times 0}{1000} = 0,52$. Une valeur approchée de p_0 est donc $0,52$.

⋆ 2ième exemple : Le nombre de coquilles dans une revue scientifique de 70 pages peut être modélisé par une variable aléatoire X suivant une loi de Poisson de paramètre λ inconnu, que l'on cherche à estimer. On dispose de 20 revues scientifiques dont on a compté le nombre d'erreurs que l'on a notées dans le tableau suivant.

x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
3	1	4	5	2	1	1	3	5	4
x_{11}	x_{12}	x_{13}	x_{14}	x_{15}	x_{16}	x_{17}	x_{18}	x_{19}	x_{20}
1	3	3	6	4	6	2	5	5	2

On suppose que cet échantillon est la réalisation de 20 variables aléatoires $X_1, \ldots, X_{20}$ indépendantes et de même loi que X.
La réalisation de $\overline{X}_{20} = \dfrac{X_1 + \cdots + X_{20}}{20}$ sur cet échantillon vaut $\dfrac{4 \times 1 + 3 \times 2 + 4 \times 3 + 3 \times 4 + 4 \times 5 + 2 \times 6}{20} = 3,3$. Une valeur approchée de λ est donc $3,3$.

10.3.2 Estimation par intervalle de confiance

La démarche consiste ici non plus à donner une estimation ponctuelle du paramètre θ mais à trouver un intervalle aléatoire, appelé intervalle de confiance, qui le contienne avec une probabilité minimale.

Définition 8 (Intervalle de confiance)

Soit X une variable aléatoire suivant une loi de paramètre θ inconnu. On appelle intervalle de confiance au niveau $1 - a$ ($a \in]0,1[$) tout intervalle I tel que :

$$P(\theta \in I) \geq 1 - a.$$

Traitons un exemple. Soit X une variable aléatoire admettant une espérance $E(X)$ inconnue et une variance connue. Soient $X_1, \ldots, X_n$ $n \geq 1$ variables aléatoires indépendantes ayant la même loi que X. Soit $\varepsilon > 0$. D'après l'inégalité de Bienaymé-Tchebychev,

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

Par linéarité, $E(\overline{X}_n) = E\left(\dfrac{1}{n}\sum_{k=1}^{n} X_k\right) = \dfrac{1}{n}E\left(\sum_{k=1}^{n} X_k\right) = \dfrac{1}{n}\sum_{k=1}^{n} E(X_k) = \dfrac{1}{n}\sum_{k=1}^{n} E(X) = \dfrac{1}{n} \times n \times E(X) = E(X)$.

Par indépendance,

$$V(\overline{X}_n) = V\left(\frac{1}{n}\sum_{k=1}^{n} X_k\right) = \frac{1}{n^2}V\left(\sum_{k=1}^{n} X_k\right) = \frac{1}{n^2}\sum_{k=1}^{n} V(X_k) = \frac{1}{n^2}\sum_{k=1}^{n} V(X) = \frac{1}{n^2} \times n \times V(X) = \frac{V(X)}{n}.$$

On a donc obtenu :

$$P(|\overline{X}_n - E(X)| \geq \varepsilon) \leq \frac{V(X)}{n\varepsilon^2}.$$

On en déduit que $P(|\overline{X}_n - E(X)| > \varepsilon) \leq \dfrac{V(X)}{n\varepsilon^2}$ (puisque $P(|\overline{X}_n - E(X)| > \varepsilon) \leq P(|\overline{X}_n - E(X)| \geq \varepsilon)$).

Par passage au complémentaire, $1 - P(|\overline{X}_n - E(X)| \leq \varepsilon) \leq \dfrac{V(X)}{n\varepsilon^2}$ et donc $P(|\overline{X}_n - E(X)| \leq \varepsilon) \geq 1 - \frac{V(X)}{n\varepsilon^2}$.

Soit $a \in]0,1[$. On pose $\varepsilon = \sqrt{\dfrac{V(X)}{na}}$ de sorte que $1 - \frac{V(X)}{n\varepsilon^2} = 1 - a$.

On a montré que : $P\left(|\overline{X}_n - E(X)| \leq \sqrt{\frac{V(X)}{na}}\right) \geq 1 - a$. On observe alors que :

$$\begin{aligned}
|\overline{X}_n - E(X)| \leq \sqrt{\frac{V(X)}{na}} &\iff -\sqrt{\frac{V(X)}{na}} \leq \overline{X}_n - E(X) \leq \sqrt{\frac{V(X)}{na}} \\
&\iff -\sqrt{\frac{V(X)}{na}} \leq E(X) - \overline{X}_n \leq \sqrt{\frac{V(X)}{na}} \\
&\iff \overline{X}_n - \sqrt{\frac{V(X)}{na}} \leq E(X) \leq \overline{X}_n + \sqrt{\frac{V(X)}{na}} \\
&\iff E(X) \in \left[\overline{X}_n - \sqrt{\frac{V(X)}{na}}, \overline{X}_n + \sqrt{\frac{V(X)}{na}}\right].
\end{aligned}$$

Finalement,

$$P\left(E(X) \in \left[\overline{X}_n - \sqrt{\frac{V(X)}{na}}, \overline{X}_n + \sqrt{\frac{V(X)}{na}}\right]\right) \geq 1 - a.$$

Autrement dit, la probabilité que l'intervalle $\left[\overline{X}_n - \sqrt{\frac{V(X)}{na}}, \overline{X}_n + \sqrt{\frac{V(X)}{na}}\right]$ contienne $E(X)$ est supérieure à $1 - a$.

On remarque que la précision augmente avec la taille de l'échantillon. Plus n est grand et plus l'intervalle de confiance est resserré autour de $E(X)$.

$\star$ 1^{ier} exemple : Supposons que X suit une loi de Bernoulli de paramètre p inconnu. Alors, $V(X) = p(1-p)$ est également inconnu et l'intervalle précédent n'est donc pas satisfaisant. On reprend donc les calculs à partir de l'inégalité : $P(|\overline{X}_n - E(X)| \geq \varepsilon) \leq \frac{V(X)}{n\varepsilon^2}$ qui se réécrit ici :

$$P(|\overline{X}_n - p| \geq \varepsilon) \leq \frac{p(1-p)}{n\varepsilon^2}.$$

On peut facilement démontrer (par exemple avec une étude de fonction) que : $\forall x \in [0,1], x(1-x) \leq \dfrac{1}{4}$, ce qui nous conduit à :

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

On démontre comme précédemment que : $P(|\overline{X}_n - p| \leq \varepsilon) \geq 1 - \frac{1}{4n\varepsilon^2}$.

Soit $a \in]0,1[$. On pose $\varepsilon = \dfrac{1}{\sqrt{4na}}$ de sorte que $1 - \frac{1}{4n\varepsilon^2} = 1 - a$.

On a montré que : $P\left(|\overline{X}_n - p| \leq \frac{1}{\sqrt{4na}}\right) \geq 1 - a$.

On observe alors que : $|\overline{X}_n - p| \leq \frac{1}{\sqrt{4na}} \iff p \in \left[\overline{X}_n - \frac{1}{\sqrt{4na}}, \overline{X}_n + \frac{1}{\sqrt{4na}}\right]$. Finalement,

$$P\left(p \in \left[\overline{X}_n - \frac{1}{\sqrt{4na}}, \overline{X}_n + \frac{1}{\sqrt{4na}}\right]\right) \geq 1 - a.$$

Autrement dit, la probabilité que l'intervalle $\left[\overline{X}_n - \frac{1}{\sqrt{4na}}, \overline{X}_n + \frac{1}{\sqrt{4na}}\right]$ contienne p est supérieure à $1 - a$.

• Reprenons notre lancer de pièce de la sous-section précédente et déterminons un intervalle de confiance de p_0 au seuil de confiance de 90%. Ici, $n = 1000$ et $a = 0,1$. Alors,

$$P\left(p_0 \in \left[\overline{X}_{1000} - \frac{1}{\sqrt{4 \times 1000 \times 0,1}}, \overline{X}_{1000} + \frac{1}{\sqrt{4 \times 1000 \times 0,1}}\right]\right) = P\left(p_0 \in \left[\overline{X}_{1000} - \frac{1}{20}, \overline{X}_{1000} + \frac{1}{20}\right]\right) \geq 0,9.$$

La réalisation de l'intervalle de confiance précédent est : $[0,52 - 1/20, 0,52 + 1/20] = [0,47, 0,57]$.
• Cherchons maintenant un intervalle de confiance de p_0 au seuil de confiance de 95%. Ici, $n = 1000$ et $a = 0,05$. Alors,

$$P\left(p_0 \in \left[\overline{X}_{1000} - \frac{1}{\sqrt{4 \times 1000 \times 0,05}}, \overline{X}_{1000} + \frac{1}{\sqrt{4 \times 1000 \times 0,05}}\right]\right) = P\left(p_0 \in \left[\overline{X}_{1000} - \frac{1}{10\sqrt{2}}, \overline{X}_{1000} + \frac{1}{10\sqrt{2}}\right]\right) \geq 0,95.$$

La réalisation de l'intervalle de confiance précédent est : $[0,52 - 1/(2\sqrt{10}), 0,52 + 1/(2\sqrt{10})] \approx [0,45, 0,60]$ (arrondi par excès).
On observe que la longueur de l'intervalle obtenu est plus grande que dans le cas précédent. C'est logique puisqu'on désire ici avoir un meilleur niveau de précision, ce qui est plus contraignant. On constate que dans les deux cas les intervalles obtenus ne sont pas très satisfaisants au regard du grand nombre d'observations (1000 !).

⋆ 2ième exemple : Supposons que X suit une loi normale de paramètres m et σ^2 avec m inconnu et $\sigma > 0$ connu. D'après le travail réalisé en amont, $\left[\overline{X}_n - \frac{\sigma}{\sqrt{na}}, \overline{X}_n + \frac{\sigma}{\sqrt{na}}\right]$ est un intervalle de confiance de m au niveau de confiance $1 - a$ avec $a \in]0,1[$. Notons qu'en pratique l'écart-type σ n'est pas connu et on doit donc utiliser l'écart-type de l'échantillon (écart-type empirique), ce qui ne sera pas étudié dans ce cours.