Chapitre XX

Théorèmes limites en probabilités

Table des matières

1 Approximation d'une loi hypergéométrique par une loi binomiale

Heuristique : On considère une urne qui contient a boules blanches et b boules noires. On note $N = a + b$ le nombre total de boules et $p = \dfrac{a}{N}$ la proportion de boules blanches et $q = \dfrac{b}{N}$ celle de boules noires. On tire successivement *sans remise* n boules de l'urne ($n \in \mathbb{N}^*$) et on note X la variable aléatoire égale au nombre de boules blanches tirées. On reconnaît alors une situation de loi usuelle : X suit la loi hypergéométrique $\mathcal{H}(N, n, p)$. On a donc, pour tout $k \in [\![0, n]\!]$:

$$P(X = k) = \frac{C_{Np}^{k}\, C_{Nq}^{n-k}}{C_N^n}.$$

Si on suppose maintenant que le nombre N total de boules est « très grand » devant le nombre n de boules tirées, alors on peut considérer que la proportion de boules blanches dans l'urne est quasiment inchangée après le tirage d'une des n boules.

Si l'on effectue l'approximation consistant à supposer que la proportion de boules blanches ne change pas au cours des différents tirages (elle est donc constamment égale à p), on peut aussi bien considérer que les tirages sont faits *avec remise*. Mais si l'on considère que le tirage est *avec remise*, alors on reconnaît une autre situation de loi usuelle : la loi $\mathcal{B}(n, p)$. On a donc, pour tout $k \in [\![0, n]\!]$:

$$P(X = k) \approx C_n^k\, p^k q^{n-k}.$$

Cette présentation heuristique et quelque peu imprécise (e.g. le sens de $\approx$ n'est pas clair) débouche sur un résultat mathématique qui s'énonce de façon rigoureuse comme suit.

Théorème 1 : Soient $a, b \in \mathbb{N}^*$ et soit $n \in \mathbb{N}^*$. On pose $N = a + b$, $p = \dfrac{a}{N}$ et $q = 1 - p$. Pour toute suite $(X_m)_{m \in \mathbb{N}^*}$ de variables aléatoires telle que :

$$\forall\, m \in \mathbb{N}^* \qquad X_m \sim \mathcal{H}(mN, n, p)$$

on a :
$$\lim_{m\to+\infty} P(X_m = k) = C_n^k\, p^k q^{n-k}$$
pour tout $k \in [\![0,n]\!]$. On dit que la suite de variables aléatoires $(X_m)_{m\in\mathbb{N}^*}$ converge en loi vers une variable aléatoire qui suit la loi $\mathcal{B}(n,p)$.

Remarque : Le premier paramètre de la loi hypergéométrique de X_m tend vers $+\infty$ quand m tend vers $+\infty$, alors que n reste fixe. Ceci est à rapprocher de l'expression « le nombre N total de boules est « très grand » devant le nombre n de boules tirées » utilisée dans l'heuristique.

Une condition pratique d'approximation : Soient $N, n \in \mathbb{N}$ et soit $p \in [0,1]$ tel que $Np \in \mathbb{N}$. On pose $q = 1-p$. Soit X une variable aléatoire qui suit la loi $\mathcal{H}(N,n,p)$. On admet que si
$$\boxed{N > 10n}$$
alors l'approximation
$$P(X=k) \approx C_n^k\, p^k q^{n-k}$$
est satisfaisante, pour tout $k \in [\![0,n]\!]$. La condition $N > 10n$ quantifie l'idée suivante : « N est très grand devant n », introduite dans l'heuristique.

Exemple : Un étang contient deux espèces de poissons : des carpes et des brochets. On compte $a = 200$ brochets et $b = 800$ carpes dans l'étang. On pêche 4 poissons et on note X la variable aléatoire égale au nombre de brochets obtenus.

On reconnaît une situation de loi usuelle : $X \sim \mathcal{H}(N,n,p)$ avec $N = a+b = 1000$, $n = 4$ et $p = \dfrac{a}{N} = \dfrac{1}{5}$.

On s'intéresse à la probabilité $P(X \geq 1)$, i.e. à la probabilité de pêcher au moins un brochet. Notons que l'on a $P(X \geq 1) = 1 - P(X=0)$.

- Calcul exact

  On a
$$P(X=0) = \frac{C_{200}^0\, C_{800}^4}{C_{1000}^4} = \frac{1 \times \dfrac{800\times 799\times 798\times 797}{4!}}{\dfrac{1000\times 999\times 998\times 997}{4!}} = \frac{800\times 799\times 798\times 797}{1000\times 999\times 998\times 997}$$
  et donc :
$$P(X \geq 1) = 1 - \frac{800\times 799\times 798\times 797}{1000\times 999\times 998\times 997} \simeq 0,5910155545.$$

- Calcul approché à l'aide de l'approximation de la loi de X par une loi binomiale (la loi $\mathcal{B}(n,p)$)

  On calcule $\dfrac{n}{N} = \dfrac{4}{1000} = \dfrac{1}{250} < \dfrac{1}{10}$. La condition d'approximation de la loi de X par la loi $\mathcal{B}(n,p)$ donnée ci-dessus est donc satisfaite.

  On a donc :
$$P(X \geq 1) = 1 - P(X=0) \approx 1 - C_4^0 \left(\frac{1}{5}\right)^0 \left(\frac{4}{5}\right)^4 = 1 - \left(\frac{4}{5}\right)^4 = 0,5904.$$

2 Approximation d'une loi binomiale par une loi de Poisson

Théorème 2 : Soit $\lambda > 0$, soit $N \in \mathbb{N}^*$ et soit $(p_n)_{n\in\mathbb{N}^*}$ une suite d'éléments de $]0,1[$ telle que $np_n \underset{n\to+\infty}{\to} \lambda$ (e.g. la suite $(p_n)_{n\in\mathbb{N}^*}$ définie par $p_n = \dfrac{\lambda}{n}$ pour tout $n \in \mathbb{N}^*$). Pour toute suite $(X_n)_{n\in\mathbb{N}^*}$ de variables aléatoires telle que :
$$\forall\, n \in \mathbb{N}^* \qquad X_n \sim \mathcal{B}(n, p_n)$$
on a :
$$\lim_{n\to+\infty} P(X_n = k) = e^{-\lambda}\, \frac{\lambda^k}{k!}$$

pour tout $k \in \mathbb{N}$. On dit que la suite de variables aléatoires $(X_n)_{n\in\mathbb{N}^*}$ converge en loi vers une variable aléatoire qui suit la loi $\mathcal{P}(\lambda)$.

Preuve : Soit $k \in \mathbb{N}$ et soit $n \in \mathbb{N}^*$ tel que $k \leq n$.

$$\begin{aligned} P(X_n = k) = C_n^k\, p_n^k\, (1-p_n)^{n-k} &= \frac{n(n-1)(n-2)\dots(n-k+1)}{k!}\, p_n^k\, (1-p_n)^{n-k} \\ &= \frac{n(n-1)(n-2)\dots(n-k+1)}{k!}\, \frac{n^k}{n^k}\, p_n^k\, (1-p_n)^{n-k} \\ &= \frac{1}{k!}\, \frac{n(n-1)(n-2)\dots(n-k+1)}{n^k}\, (np_n)^k\, (1-p_n)^{n-k}. \end{aligned}$$

On a donc :

$$P(X_n = k) = \frac{1}{k!}\, \frac{n(n-1)(n-2)\dots(n-k+1)}{n^k}\, (np_n)^k\, (1-p_n)^{n-k}. \tag{1}$$

On a :

$$\frac{\overbrace{n(n-1)(n-2)\dots(n-k+1)}^{k \text{ termes}}}{n^k} = \frac{n}{n} \times \frac{n-1}{n} \times \frac{n-2}{n} \times \dots \times \frac{n-k+1}{n}$$

et donc

$$\lim_{n\to+\infty} \frac{n(n-1)(n-2)\dots(n-k+1)}{n^k} = 1. \tag{2}$$

Comme la fonction $x \mapsto x^k$ est continue sur $\mathbb{R}$ (en particulier en λ) et comme $\lim\limits_{n\to+\infty} np_n = \lambda$, on a :

$$\lim_{n\to+\infty} (np_n)^k = \lambda^k. \tag{3}$$

Passons à l'étude du comportement asymptotique de la suite $((1-p_n)^{n-k})_{n\in\mathbb{N}^*}$. On remarque tout d'abord que :

$$\forall\, n \in \mathbb{N}^* \qquad (1-p_n)^{n-k} = e^{(n-k)\ln(1-p_n)}. \tag{4}$$

Ensuite, de $p_n = \dfrac{1}{n} \times np_n$ pour tout $n \in \mathbb{N}^*$ et de $\lim\limits_{n\to+\infty} np_n = \lambda$, on déduit que la suite $(p_n)_{n\in\mathbb{N}^*}$ converge et que :

$$\lim_{n\to+\infty} p_n = 0. \tag{5}$$

Comme $\lim\limits_{n\to+\infty} p_n = 0$ et $\ln(1+x) \underset{x\to 0}{\sim} x$, on a :

$$\ln(1-p_n) \underset{n\to+\infty}{\sim} -p_n$$

et par suite :

$$(n-k)\ln(1-p_n) \underset{n\to+\infty}{\sim} -(n-k)p_n = -np_n + kp_n. \tag{6}$$

De $\lim\limits_{n\to+\infty} np_n = \lambda$ et de (5), on déduit que :

$$\lim_{n\to+\infty} -np_n + kp_n = -\lambda. \tag{7}$$

De (6) et (7), on tire alors :

$$\lim_{n\to+\infty} (n-k)\ln(1-p_n) = -\lambda. \tag{8}$$

La fonction exponentielle étant continue sur $\mathbb{R}$ (en particulier en $-\lambda$) on a alors :

$$\lim_{n\to+\infty} \underbrace{e^{(n-k)\ln(1-p_n)}}_{=(1-p_n)^{n-k} \text{ (cf. (4))}} = e^{-\lambda}. \tag{9}$$

Enfin, en rassemblant les résultats (1), (2), (3) et (9), on obtient :

$$P(X_n = k) \underset{n\to+\infty}{\to} \frac{1}{k!} \times 1 \times \lambda^k \times e^{-\lambda} = e^{-\lambda}\,\frac{\lambda^k}{k!}.$$

Q.E.D.

Une condition pratique d'approximation : Soient $n \in \mathbb{N}^*$ et soit $p \in [0,1]$. On pose $\lambda = np$. Soit X une variable aléatoire qui suit la loi $\mathcal{B}(n,p)$. On admet que si

$$\boxed{n \geq 30 \quad ; \quad p \leq 0,1 \quad ; \quad \lambda = np \leq 10}$$

alors l'approximation

$$P(X = k) \approx e^{-\lambda}\,\frac{\lambda^k}{k!}$$

est satisfaisante, pour tout $k \in [\![0, n]\!]$. Les conditions $n \geq 30$, $p \leq 0,1$, $np \leq 10$ quantifient les idées suivantes : « n est grand », « p est petit » et « np n'est pas trop grand ».

Exemple : Une boîte contient 1000 jetons dont 10 sont blancs et tous les autres noirs. On effectue 500 tirages au hasard avec remise et on note X la variable aléatoire égale au nombre de jetons blancs tirés.

On reconnaît alors une situation de loi usuelle : X suit la loi binomiale $\mathcal{B}(n,p)$, avec $n = 500$ et $p = \dfrac{10}{1000} = 0,01$.

On s'intéresse à la probabilité de $P(X = 0)$, i.e. à la probabilité de ne tirer aucun jeton blanc.

- Calcul exact

  On a
  $$P(X = 0) = C_{500}^0\, 0,01^0\, 0,99^{500} = 0,99^{500} \simeq 0,00657048.$$

- Calcul approché à l'aide de l'approximation de la loi de X par une loi de Poisson (la loi $\mathcal{P}(\lambda)$ avec $\lambda = np$)

  On a $n = 500 \geq 30$, $p = 0,01 < 0,1$ et $\lambda = np = 5 \leq 10$. La condition d'approximation de la loi de X par la loi $\mathcal{P}(\lambda)$ donnée ci-dessus est donc satisfaite.

  On a donc :
  $$P(X = 0) \approx e^{-5}\,\frac{5^0}{0!} = e^{-5} \simeq 0,00673794.$$

3 Inégalité de Bienaymé-Tchebychev

Heuristique : Soit X une variable aléatoire admettant une variance (et donc une espérance). L'inégalité de Bienaymé-Tchebychev (qui n'est pas, à proprement parler, un théorème limite) permet de donner un sens quantitatif à la remarque suivante.

> Plus l'écart-type (ou la variance) est faible, plus la distribution de probabilités est concentrée autour de l'espérance mathématique.

Théorème 3 (inégalité de Bienaymé-Tchebychev)

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

Exemple : On jette un dé cubique équilibré dont 5 faces sont blanches et une face rouge. On jette le dé jusqu'à l'obtention de la face rouge et on note X la variable aléatoire égale au nombre de lancers effectués.

On reconnaît un schéma de Bernoulli que l'on répète de façons indépendantes jusqu'à l'obtention d'un succès. Un succès dans le schéma de Bernoulli correspond à l'obtention de la face rouge lors d'un lancer et la probabilité de succès est $\dfrac{1}{6}$. X suit donc une loi géométrique de paramètre $p = \dfrac{1}{6}$.

Par conséquent X admet une variance (et donc une espérance) et on a :

$$E(X) = \frac{1}{p} = 6 \qquad \text{et} \qquad V(X) = \frac{q}{p^2} = 30.$$

On s'intéresse à l'événement « s'écarter au moins de 10 de la valeur moyenne », i.e. à l'événement

$$[|X-6| \geq 10].$$

Notons que :

$$\begin{aligned} [|X-6| \geq 10] &= [X-6 \geq 10] \cup [X-6 \leq -10] \qquad (\forall\, x \in \mathbb{R} \quad \forall\, A \in \mathbb{R}^+ \quad |x| \geq A \Longleftrightarrow (x \geq A \text{ ou } x \leq -A)) \\ &= [X \geq 16] \cup \underbrace{[X \leq -4]}_{\emptyset \text{ car } X(\Omega)=\mathbb{N}^*} \\ &= [X \geq 16]. \end{aligned}$$

Donc l'événement $[|X-6| \geq 10]$ s'écrit aussi « effectuer au moins 16 lancers avant d'avoir la face rouge ». D'après l'inégalité de Bienaymé-Tchebychev (appliquée avec $\varepsilon = 10$) on a la majoration suivante :

$$P(\text{« effectuer au moins 16 lancers avant d'avoir la face rouge »}) = P(|X-6| \geq 10) \leq \frac{30}{100} = 0,3.$$

4 La loi faible des grands nombres – Cas particulier

Heuristique : On dispose d'une pièce équilibrée. On la lance n fois ($n \in \mathbb{N}^*$) et pour tout $i \in [\![1,n]\!]$, on note X_i la variable aléatoire égale à 1 si l'on a obtenu PILE au i-ème lancer et égale à 0 sinon.

La variable aléatoire $S_n = \sum_{i=1}^{n} X_i = X_1 + X_2 + \ldots + X_n$ est égale au nombre de PILE obtenus lors des n lancers.
La variable aléatoire

$$\overline{X_n} = \frac{1}{n}\, S_n = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{X_1 + X_2 + \ldots + X_n}{n}$$

donne la fréquence d'apparition de PILE lors des n lancers.

En effectuant des simulations à l'aide d'un ordinateur, on peut constater que, pour « n grand », les valeurs prises par $\overline{X_n}$ sont voisines de $\frac{1}{2}$, mais que de légères fluctuations apparaissent. Aussi est-il assez tentant de dire que « $(\overline{X}_n)$ converge vers $\frac{1}{2}$ ». Pour donner un sens précis à cette intuition, il convient tout d'abord de préciser ce que signifie converger pour une suite de variables aléatoires.

Une définition possible de convergence est la suivante. On dit que $(\overline{X}_n)$ *converge en probabilité vers* $\frac{1}{2}$ si pour tout $\varepsilon > 0$ on a :

$$(*) \qquad\qquad P\left(\left|\overline{X_n} - \frac{1}{2}\right| \geq \varepsilon\right) \underset{n\to+\infty}{\to} 0.$$

On démontrera cette convergence en probabilité (résultat appelé « loi faible des grands nombres ») dans la partie suivante, dans une plus grande généralité.

Soit $\varepsilon > 0$. Comme l'événement contraire de $\left[\left|\overline{X_n} - \frac{1}{2}\right| \geq \varepsilon\right]$ est

$$\overline{\left[\left|\overline{X_n} - \frac{1}{2}\right| \geq \varepsilon\right]} = \left[\left|\overline{X_n} - \frac{1}{2}\right| < \varepsilon\right] = \left[\frac{1}{2} - \varepsilon < \overline{X_n} < \frac{1}{2} + \varepsilon\right]$$

l'assertion $(*)$ est équivalente à

$$P\left(\frac{1}{2} - \varepsilon \leq \overline{X_n} \leq \frac{1}{2} + \varepsilon\right) \underset{n\to+\infty}{\to} 1.$$

On a donc en particulier pour $\varepsilon = 10^{-6}$ (par exemple) :

$$P\left(0,499999 \leq \overline{X_n} \leq 0,500001\right) \underset{n\to+\infty}{\to} 1.$$

Remarque : Il faut bien comprendre ce que signifie la convergence en probabilité de $\overline{X_n}$ vers $\frac{1}{2}$: il est toujours *possible* qu'un écart ε soit dépassé pour de très grandes valeurs de n, mais c'est « de plus en plus » *improbable.*

5 La loi faible des grands nombres – Cas général

Contexte : Dans toute cette partie, on considère une suite de variables aléatoires $X_1, X_2, X_3, \ldots, X_n, \ldots$ deux à deux indépendantes et de même loi. Pour tout $n \in \mathbb{N}^*$, on introduit deux nouvelles variables aléatoires :

$$S_n = \sum_{i=1}^{n} X_i = X_1 + X_2 + \ldots + X_n \qquad \text{et} \qquad \overline{X_n} = \frac{1}{n}\, S_n = \frac{1}{n} \sum_{i=1}^{n} X_i = \frac{X_1 + X_2 + \ldots + X_n}{n}.$$

On suppose que X_1 admet une variance (et donc une espérance). On pose

$$\mu = E(X_1)\ ;\ \sigma = \sigma(X_1) \text{ (écart type de } X_1\text{)}.$$

On a donc $V(X_1) = \sigma^2$. Comme les variables aléatoires $X_1, X_2, X_3, \ldots, X_n, \ldots$ ont la même loi, les variables $X_2, X_3, \ldots, X_n, \ldots$ possèdent elles aussi une variance (et donc une espérance) et on a

$$\forall\, i \in \mathbb{N}^{\geq 2} \qquad E(X_i) = \mu\ ;\ \sigma(X_i) = \sigma\ ;\ V(X_i) = \sigma^2. \tag{10}$$

Lemme 4 (espérances de S_n et de $\overline{X_n}$) : On a :

$$E(S_n) = n\,\mu \qquad \text{et} \qquad E(\overline{X_n}) = \mu.$$

Preuve : Soit $n \in \mathbb{N}^*$. On a :

$$\begin{aligned}
E(S_n) &= E(X_1 + X_2 + \ldots + X_n) \\
&= E(X_1) + E(X_2) + \ldots + E(X_n) \qquad \text{(linéarité de l'espérance)} \\
&= \underbrace{\mu + \mu + \ldots + \mu}_{n \text{ termes}} \qquad \text{(cf. (10))} \\
&= n\,\mu
\end{aligned}$$

d'où

$$\begin{aligned}
E(\overline{X_n}) &= E\left(\frac{1}{n}\, S_n\right) \\
&= \frac{1}{n}\, E(S_n) \qquad \text{(cf. formule } E(aX+b) = aE(X) + b\text{)} \\
&= \frac{1}{n}\, n\,\mu \qquad \text{(car } E(S_n) = \mu\text{)} \\
&= \mu.
\end{aligned}$$

Q.E.D.

Lemme 5 (variances de S_n et de $\overline{X_n}$) : On a :

$$V(S_n) = n\,\sigma^2 \qquad \text{et} \qquad V(\overline{X_n}) = \frac{\sigma^2}{n} \quad \text{(donc } \sigma(\overline{X_n}) = \frac{\sigma}{\sqrt{n}}\text{)}.$$

Preuve : Soit $n \in \mathbb{N}^*$. On a :

$$\begin{aligned}
V(S_n) &= V(X_1 + X_2 + \ldots + X_n) \\
&= V(X_1) + V(X_2) + \ldots + V(X_n) \qquad (X_1, X_2, \ldots, X_n \text{ sont deux à deux indépendantes}) \\
&= \underbrace{\sigma^2 + \sigma^2 + \ldots + \sigma^2}_{n \text{ termes}} \qquad \text{(cf. (10))} \\
&= n\,\sigma^2
\end{aligned}$$

d'où

$$\begin{aligned}
V(\overline{X_n}) &= V\left(\frac{1}{n}\,S_n\right) \\
&= \frac{1}{n^2}\,V(S_n) \qquad \text{(cf. formule } V(aX+b)=a^2V(X)\text{)} \\
&= \frac{1}{n^2}\,n\,\sigma^2 \qquad \text{(car } V(S_n)=n\sigma^2\text{)} \\
&= \frac{\sigma^2}{n}.
\end{aligned}$$

Q.E.D.

Théorème 6 (loi faible des grands nombres) : Dans le contexte précédent, i.e. si $X_1, X_2, X_3, \ldots, X_n, \ldots$ est une suite de variables aléatoires deux à deux indépendantes, de même loi, admettant une variance (et donc une espérance), alors on a pour tout $\varepsilon > 0$:

$$P\left(\left|\overline{X_n} - \mu\right| \geq \varepsilon\right) \underset{n\to+\infty}{\to} 0$$

où μ désigne l'espérance commune des variables aléatoires $X_1, X_2, X_3, \ldots, X_n, \ldots$ et où $\overline{X_n} = \dfrac{X_1 + X_2 + \ldots + X_n}{n}$ pour tout $n \in \mathbb{N}^*$. On dit que la suite de variables aléatoires $(X_n)_{n\in\mathbb{N}^*}$ converge en probabilité vers μ.

Preuve : Soit $\varepsilon > 0$. Soit $n \in \mathbb{N}^*$. On applique l'inégalité de Bienaymé-Tchebychev à la variable $\overline{X_n}$ pour obtenir :

$$P\left(\left|\overline{X_n} - E(\overline{X_n})\right| \geq \varepsilon\right) \leq \frac{V(\overline{X_n})}{\varepsilon^2}. \tag{11}$$

À l'aide des deux lemmes précédents, l'inégalité (11) se réécrit :

$$P\left(\left|\overline{X_n} - \mu)\right| \geq \varepsilon\right) \leq \frac{\sigma^2}{n\,\varepsilon^2}.$$

De cette inégalité et du fait qu'une probabilité est positive ou nulle, on déduit :

$$0 \leq P\left(\left|\overline{X_n} - \mu)\right| \geq \varepsilon\right) \leq \frac{\sigma^2}{n\,\varepsilon^2}.$$

De cette dernière inégalité, du fait que $\lim\limits_{n\to+\infty} \dfrac{1}{n} = 0$ et du théorème d'encadrement, on déduit que :

$$P\left(\left|\overline{X_n} - \mu\right| \geq \varepsilon\right) \underset{n\to+\infty}{\to} 0.$$

Q.E.D.

Remarque : Dans le cas où toutes les variables aléatoires X_i $(i \in \mathbb{N}^*)$ suivent la loi $\mathcal{B}\left(\dfrac{1}{2}\right)$, la loi faible des grands nombres donne le résultat $(*)$ énoncé dans la partie 4.

Corollaire 7 (application de la loi faible des grands nombres à la loi binomiale) : Soit $p \in]0,1[$ et soit $(Y_n)_{n\in\mathbb{N}^*}$ une suite de variables aléatoires telle que :

$$\forall\, n \in \mathbb{N}^* \qquad Y_n \sim \mathcal{B}(n,p).$$

On a pour tout $\varepsilon > 0$:

$$P\left(\left|\frac{1}{n}Y_n - p\right| \geq \varepsilon\right) \underset{n\to+\infty}{\to} 0.$$

Preuve : On introduit une suite de variables aléatoires $X_1, X_2, X_3, \ldots, X_n, \ldots$ mutuellement indépendantes, qui suivent toutes la loi $\mathcal{B}(p)$.
Soit $n \in \mathbb{N}^*$. On pose $S_n = X_1 + X_2 + \ldots + X_n$. Comme les variables aléatoires $X_1, X_2, X_3, \ldots, X_n$ sont

mutuellement indépendantes, on sait que $S_n \sim \mathcal{B}(n,p)$ (cf. cours sur les couples de variables aléatoires discrètes). La variable aléatoire S_n ayant même loi que Y_n, on a :

$$\forall \varepsilon > 0 \qquad P\left(\left|\frac{1}{n}Y_n - p\right| \geq \varepsilon\right) = P\left(\left|\frac{1}{n}S_n - p\right| \geq \varepsilon\right). \tag{12}$$

Soit $\varepsilon > 0$. On rappelle que l'espérance d'une variable aléatoire suivant la loi $\mathcal{B}(p)$ est p. La loi faible des grands nombres, appliquée à la suite de variables aléatoires $X_1, X_2, X_3, \ldots, X_n, \ldots$ donne :

$$P\left(\left|\frac{1}{n}S_n - p\right| \geq \varepsilon\right) = P\left(\left|\overline{X_n} - \mu\right| \geq \varepsilon\right) \underset{n\to+\infty}{\to} 0. \tag{13}$$

De (12) et (13), on déduit alors que :

$$P\left(\left|\frac{1}{n}Y_n - p\right| \geq \varepsilon\right) \underset{n\to+\infty}{\to} 0.$$

Q.E.D.

Application de la loi faible des grands nombres (un premier pas vers les statistiques) [1]

On s'intéresse au nombre de véhicules se présentant à un poste de péage donné entre 18h et 22h le dimanche. On a effectué un comptage durant n semaines consécutives. Le nombre obtenu pour le i-ème dimanche est noté x_i ($i \in [\![1, n]\!]$). On souhaite faire des prévisions à partir de ces n données :

$$x_1, x_2, \ldots, x_n.$$

On introduit pour cela la variable aléatoire X égale au nombre de véhicules passant au poste de péage entre 18h et 22h le dimanche. Comme il s'agit de compter un flux pendant un temps donné, on peut supposer que X suit une loi de Poisson. On suppose donc qu'il existe $\lambda > 0$ tel que $X \sim \mathcal{P}(\lambda)$.

On explique maintenant comment estimer λ à l'aide des nombres $x_1, x_2, \ldots, x_n$. Comme $X \sim \mathcal{P}(\lambda)$, on a l'identité $E(X) = \lambda$. Le paramètre λ cherché est donc égal à l'espérance de X.

On introduit n variables aléatoires $X_1, X_2, \ldots, X_n$ deux à deux indépendantes, de même loi que X, i.e. suivant toutes la loi de Poisson de paramètre λ. D'après la loi faible des grands nombres, on sait que la probabilité que

$$\overline{X_n} = \frac{X_1 + X_2 + \ldots + X_n}{n}$$

soit « proche » de λ (l'espérance commune des variables aléatoires $X_1, X_2, \ldots, X_n$) est forte (i.e. voisine de 1), si « n est très grand ». En considérant que $(x_1, x_2, \ldots, x_n)$ est une valeur possible du n-uplet de variables aléatoires $(X_1, X_2, \ldots, X_n)$ (i.e. qu'il existe $\omega \in \Omega$ tel que $x_i = X_i(\omega)$ pour tout $i \in [\![1, n]\!]$), alors on obtient que

$$\frac{x_1 + x_2 + \ldots + x_n}{n}$$

est « très probablement voisin de λ, si n est très grand » [2]. Dans la suite, on note λ_{emp} (valeur empirique de λ) le nombre

$$\lambda_{emp} = \frac{x_1 + x_2 + \ldots + x_n}{n}.$$

L'étude précedente met en exergue un lien, assez intuitif, entre la moyenne empirique λ_{emp} et l'espérance mathématique de X, aussi appelée parfois valeur moyenne de X.

On peut, par exemple, utiliser ce qui précède pour estimer la probabilité qu'il y ait moins de 5 voitures qui se présentent au péage considéré entre 18h et 22h le dimanche à venir :

$$P(X \leq 5) = e^{-\lambda} \sum_{k=0}^{5} \frac{\lambda^k}{k!} \simeq e^{-\lambda_{emp}} \sum_{k=0}^{5} \frac{\lambda_{emp}^k}{k!} = e^{-\lambda_{emp}} \left(1 + \lambda_{emp} + \frac{\lambda_{emp}^2}{2} + \frac{\lambda_{emp}^3}{6} + \frac{\lambda_{emp}^4}{24} + \frac{\lambda_{emp}^5}{120}\right).$$

1. Cette application peut être omise lors de la première lecture.
2. On pourrait appliquer l'inégalité de Bienaymé-Tchebychev pour donner un sens précis à cette expression.

6 Le théorème de la limite centrée

Heuristique : On considère une suite de variables aléatoires $X_1, X_2, X_3, \ldots, X_n, \ldots$ mutuellement (et donc deux à deux) indépendantes et de même loi. Pour tout $n \in \mathbb{N}^*$, on introduit la variable aléatoire S_n définie par :

$$S_n = \sum_{i=1}^{n} X_i = X_1 + X_2 + \ldots + X_n.$$

On se propose ici d'étudier « le comportement asymptotique » de la suite de variables aléatoires $(S_n)_{n \in \mathbb{N}^*}$. Commençons par étudier les espérances et les variances des variables aléatoires S_n $(n \in \mathbb{N}^*)$.

On suppose que X_1 admet une variance (et donc une espérance). On pose

$$\mu = E(X_1) \; ; \; \sigma = \sigma(X_1) \text{ (écart type de } X_1).$$

On a donc $V(X_1) = \sigma^2$. Comme les variables aléatoires $X_1, X_2, X_3, \ldots, X_n, \ldots$ ont la même loi, les variables $X_2, X_3, \ldots, X_n, \ldots$ possèdent elles aussi une variance (et donc une espérance) et on a

$$\forall\, i \in \mathbb{N}^{\geq 2} \qquad E(X_i) = \mu \; ; \; \sigma(X_i) = \sigma \; ; \; V(X_i) = \sigma^2. \tag{14}$$

Ce contexte est voisin de celui envisagé dans la partie 5. La différence réside dans l'hypothèse d'indépendance, plus forte ici. Les lemmes 4 et 5 donnent :

$$\forall\, n \in \mathbb{N}^* \qquad E(S_n) = n\,\mu \; ; \; V(S_n) = n\,\sigma^2.$$

On remarque que si μ est non nul, alors la suite numérique $(E(S_n))_{n \in \mathbb{N}^*}$ diverge vers un infini et que si σ est non nul, ce que l'on supposera désormais, alors la suite numérique $(V(S_n))_{n \in \mathbb{N}^*}$ diverge vers $+\infty$. Ces deux faits nous incitent à penser qu'il y a plutôt « divergence » que « convergence » de la suite de variables aléatoires $(S_n)_{n \in \mathbb{N}^*}$.

Pour parer à cela, on va appliquer une transformation affine à chacune des variables aléatoires S_n $(n \in \mathbb{N}^*)$ de façon à ce que leur espérance soit nulle (on dira qu'elles sont centrées) et à ce que leur variance soit égale à 1 (on dira qu'elles sont réduites). Pour tout $n \in \mathbb{N}^*$, on pose :

$$S_n^* = \frac{S_n - n\,\mu}{\sqrt{n}\,\sigma}.$$

En utilisant les résultats précédents sur la variance et l'espérance de S_n $(n \in \mathbb{N}^*)$ et les formules $E(aX + b) = aE(X) + b$ et $V(aX + b) = a^2 V(X)$ on vérifie que :

$$\forall\, n \in \mathbb{N}^* \qquad E(S_n^*) = 0 \; ; \; V(S_n^*) = 1.$$

On s'intéresse donc plutôt à présent au comportement asymptotique de la suite de variables aléatoires $(S_n^*)_{n \in \mathbb{N}^*}$ qu'à celui de la suite de variables aléatoires $(S_n)_{n \in \mathbb{N}^*}$.

Pour comprendre ce qui se passe considérons un cas particulier : celui où les X_n suivent la loi de Bernoulli de paramètre $\frac{1}{2}$. On peut se représenter la suite $X_1, X_2, \ldots, X_n, \ldots$ comme la suite des résultats d'un jeu de PILE ou FACE répété indéfiniment avec une pièce équilibrée. Pour visualiser « le comportement asymptotique » de la suite de variables aléatoires $S_1^*, S_2^*, \ldots, S_n^*, \ldots$ on peut représenter les lois de celles-ci à l'aide de diagrammes en barres.

La vidéo de la partie *Le théorème central* de la page web, réalisée par Jean-Pierre Kahane (professeur à l'Université Paris Sud et membre de l'Académie des Sciences), d'adresse

http://images.math.cnrs.fr/La-courbe-verte-en-cloche.html

nous montre une courbe, celle de la densité de la loi normale centrée réduite $\mathcal{N}(0,1)$, i.e. de la fonction

$$\varphi \colon t \mapsto \frac{1}{\sqrt{2\pi}}\, e^{-\frac{t^2}{2}}$$

et l'évolution du diagramme en barres de la loi de S_n^* pour n variant de 1 à 32.

Il semble, d'après cette vidéo, que la somme des aires des barres représentant la loi de S_n^* tende vers l'aire sous la courbe représentant la fonction φ quand n tend vers $+\infty$. On peut en fait observer ce phénomène au-dessus de chaque intervalle I d'extrémités $a, b \in \mathbb{R}$: la somme des aires des barres dont la base est contenue dans I tend vers l'aire sous la courbe représentant la fonction φ au dessus de I quand n tend vers $+\infty$.

Autrement dit, on peut conjecturer que :

$$\forall\, a, b \in \mathbb{R} \text{ tels que } a < b \qquad P(a \leq S_n^* \leq b) \underset{n \to +\infty}{\to} \int_a^b \varphi(t)\, dt = \frac{1}{\sqrt{2\pi}} \int_a^b e^{-\frac{t^2}{2}}\, dt \tag{15}$$

en s'appuyant sur l'interprétation géométrique de l'intégrale. Cette conjecture peut être reformulée de plusieurs manières. On en expose deux ci-dessous.

- Soit Y une variable aléatoire de loi $\mathcal{N}(0,1)$. Alors (15) se reformule comme suit :

$$\forall\, a, b \in \mathbb{R} \text{ tels que } a < b \qquad P(a \leq S_n^* \leq b) \underset{n \to +\infty}{\to} P(a \leq Y \leq b) \tag{16}$$

- Soit Φ la fonction de répartition de Y, i.e. la fonction

$$\Phi\colon \mathbb{R} \to \mathbb{R}\ ;\ x \mapsto \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}}\, dt.$$

  Bien que l'on ne sache pas calculer explicitement les valeurs de cette fonction, on dispose d'une table de valeurs (approchées) de Φ (cf. page 3 de la feuille d'exercices n˚19). Alors (15) de réécrit

$$\forall\, a, b \in \mathbb{R} \text{ tels que } a < b \qquad P(a \leq S_n^* \leq b) \underset{n \to +\infty}{\to} \Phi(b) - \Phi(a) \tag{17}$$

  en appliquant la relation de Chasles pour les intégrales généralisées.

Les conjectures (15), (16) et (17) sont vraies et sont non seulement valables pour l'exemple considéré ici (le nombre de PILE obtenus dans un jeu de PILE ou FACE indéfiniment répété, avec une pièce équilibrée), mais aussi dans une grande généralité. C'est le contenu du théorème suivant.

Théorème 8 (de la limite centrée) : Soit $(X_n)_{n \in \mathbb{N}^*}$ une suite de variables aléatoires mutuellement indépendantes et de même loi. On suppose que ces variables possèdent une variance (donc une espérance). On note μ leur espérance commune et σ leur écart-type commun, supposé non nul. Pour tout $n \in \mathbb{N}^*$, on introduit la variable aléatoire S_n définie par :

$$S_n = \sum_{i=1}^{n} X_i = X_1 + X_2 + \ldots + X_n$$

et la variable centrée [3] réduite [4] associée :

$$S_n^* = \frac{S_n - n\, \mu}{\sqrt{n}\, \sigma}.$$

On note Φ la fonction de répartition d'une variable aléatoire suivant la loi normale centrée réduite, i.e. la fonction Φ définie par :

$$\Phi\colon \mathbb{R} \to \mathbb{R}\ ;\ x \mapsto \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}}\, dt.$$

Alors pour tout $x \in \mathbb{R}$, on a :

$$F_{S_n^*}(x) = P(S_n^* \leq x) \underset{n \to +\infty}{\to} \Phi(x).$$

On dit que la suite de variables aléatoires $(S_n^*)_{n \in \mathbb{N}^*}$ converge en loi vers la loi $\mathcal{N}(0,1)$.

Corollaire 9 : On conserve les notations et les hypothèses de l'énoncé du théorème de la limite centrée. Soient $a, b \in \mathbb{R}$ tels que $a < b$.

1. $P(S_n^* \leq b) \underset{n \to +\infty}{\to} \Phi(b)$

2. $P(S_n^* < b) \underset{n \to +\infty}{\to} \Phi(b)$

3. i.e. d'espérance nulle
4. i.e. de variance 1

3. $P(a \leq S_n^*) \underset{n \to +\infty}{\to} 1 - \Phi(a)$

4. $P(a < S_n^*) \underset{n \to +\infty}{\to} 1 - \Phi(a)$

5. $P(a \leq S_n^* \leq b) \underset{n \to +\infty}{\to} \Phi(b) - \Phi(a)$

6. $P(a < S_n^* \leq b) \underset{n \to +\infty}{\to} \Phi(b) - \Phi(a)$

7. $P(a \leq S_n^* < b) \underset{n \to +\infty}{\to} \Phi(b) - \Phi(a)$

8. $P(a < S_n^* < b) \underset{n \to +\infty}{\to} \Phi(b) - \Phi(a)$

Éléments de preuve : La propriété 1 est une simple reformulation du théorème de la limite centrée. La propriété 4 se déduit de 1. par passage au complémentaire. La propriété 6 se déduit de 1. et du fait que :

$$]-\infty, b\,] =]-\infty, a] \underset{\text{disjointe}}{\cup}]a, b].$$

Les autres propriétés sont admises.

Remarque : Soit $a > 0$. Comme $\Phi(-a) = 1 - \Phi(a)$ (cf. exercice 264 de la feuille d'exercices n°9), on a

$$P(-a \leq S_n^* \leq a) \underset{n \to +\infty}{\to} \Phi(a) - \Phi(-a) = 2\Phi(a) - 1.$$

Corollaire 10 (application du théorème de la limite centrée à la loi binomiale) : Soient $p \in]0, 1[$ et $q = 1 - p$. Soit $(Y_n)_{n \in \mathbb{N}^*}$ une suite de variables aléatoires telle que :

$$\forall\, n \in \mathbb{N}^* \qquad Y_n \sim \mathcal{B}(n, p).$$

On pose pour tout $n \in \mathbb{N}^*$:

$$T_n = \frac{Y_n - n\,p}{\sqrt{n\,p\,q}}.$$

Alors pour tout $x \in \mathbb{R}$, on a :

$$F_{T_n}(x) = P(T_n \leq x) \underset{n \to +\infty}{\to} \Phi(x)$$

et les propriétés du corollaire 9 obtenues en remplaçant partout S_n^* par T_n restent vraies. On dit que la suite de variables aléatoires $(T_n)_{n \in \mathbb{N}^*}$ converge en loi vers la loi $\mathcal{N}(0, 1)$.

Preuve : On procède de façon analogue à ce que l'on a fait pour montrer le corollaire 7. On introduit une suite de variables aléatoires $X_1, X_2, X_3, \ldots, X_n, \ldots$ mutuellement indépendantes, qui suivent toutes la loi $\mathcal{B}(p)$.
Soit $n \in \mathbb{N}^*$. On pose $S_n = X_1 + X_2 + \ldots + X_n$. Comme les variables aléatoires $X_1, X_2, X_3, \ldots, X_n$ sont mutuellement indépendantes, on sait que $S_n \sim \mathcal{B}(n, p)$ (cf. cours sur les couples de variables aléatoires discrètes). La variable aléatoire S_n ayant même loi que Y_n, on a :

$$\forall\, x \in \mathbb{R} \qquad P\left(T_n \leq x\right) = P\left(\frac{Y_n - n\,p}{\sqrt{n\,p\,q}} \leq x\right) = P\left(\frac{S_n - n\,p}{\sqrt{n\,p\,q}} \leq x\right). \tag{18}$$

Soit $x \in \mathbb{R}$. On rappelle que l'espérance d'une variable aléatoire suivant la loi $\mathcal{B}(p)$ est p et que sa variance est $p\,q$. La variable aléatoire S_n^* centrée réduite associée à S_n dans l'énoncé du théorème de la limite centrée est donc :

$$S_n^* = \frac{S_n - n\,E(X_1)}{\sqrt{n}\,\sqrt{V(X_1)}} = \frac{S_n - n\,p}{\sqrt{n\,p\,q}}. \tag{19}$$

D'après (18) et (19), on a donc :

$$F_{T_n}(x) = P\left(T_n \leq x\right) = P\left(S_n^* \leq x\right) = F_{S_n^*}(x). \tag{20}$$

De (20) et du théorème de la limite centrée, on déduit alors :

$$F_{T_n}(x) = P(T_n \leq x) \underset{n \to +\infty}{\to} \Phi(x)$$

Q.E.D.

Exemple : Un joueur lance une pièce équilibrée. Lorsqu'il obtient PILE, il gagne 1€ et lorsqu'il obtient FACE, il perd 1 €. On se propose d'estimer le nombre maximal de lancers à effectuer pour que ce joueur ait une probabilité de plus de 95% de perdre au plus 20€, en utilisant une approximation par la loi normale.

Soit $n \in \mathbb{N}^*$. On suppose qu'on a effectué n lancers. On note Y_n la variable aléatoire égale au nombre de PILE obtenus. On reconnaît une situation de loi usuelle : $Y_n \sim \mathcal{B}(n,p)$. Soit G_n le gain algébrique du joueur. On a la relation :

$$G_n = 1\ Y_n - 1\ (n - Y_n) = 2Y_n - n$$

car $n - Y_n$ est le nombre de parties perdues.

On cherche donc $n \in \mathbb{N}^*$ tel que l'inégalité

$$(I)\ :\ P(G_n \geq -20) \geq 0,95$$

soit valide. Notons que la variable aléatoire T_n introduite dans le corollaire 10, s'écrit ici :

$$T_n = \frac{Y_n - \dfrac{n}{2}}{\sqrt{\dfrac{n}{4}}}$$

car $p = q = \dfrac{1}{2}$.

L'idée consiste à exprimer l'événement $[G_n \geq -20]$ en fonction de T_n pour pouvoir ensuite appliquer le corollaire 10 et ainsi donner une valeur approchée de la probabilité $P(G_n \geq -20)$.

On a :

$$[G_n \geq -20] = [2Y_n - n \geq -20] = \left[Y_n - \frac{n}{2} \geq -10\right] = \left[\frac{Y_n - \dfrac{n}{2}}{\sqrt{\dfrac{n}{4}}} \geq -10\ \frac{1}{\sqrt{\dfrac{n}{4}}}\right] = \left[T_n \geq -\frac{20}{\sqrt{n}}\right]$$

et donc :

$$P(G_n \geq -20) = P\left(T_n \geq -\frac{20}{\sqrt{n}}\right) = 1 - P\left(T_n < -\frac{20}{\sqrt{n}}\right). \tag{21}$$

D'après le corollaire 10, on a :

$$P\left(T_n < -\frac{20}{\sqrt{n}}\right) \underset{n\to+\infty}{\to} \Phi\left(-\frac{20}{\sqrt{n}}\right). \tag{22}$$

Comme l'énoncé nous y invite, on interprète (22) par une approximation :

$$P\left(T_n < -\frac{20}{\sqrt{n}}\right) \approx \Phi\left(-\frac{20}{\sqrt{n}}\right).$$

Grâce à cette approximation et à (21), (I) se réécrit :

$$1 - \Phi\left(-\frac{20}{\sqrt{n}}\right) \geq 0,95.$$

ou encore :

$$(I')\ :\ \Phi\left(\frac{20}{\sqrt{n}}\right) \geq 0,05$$

car $\Phi(-x) = 1 - \Phi(x)$ pour tout $x \in \mathbb{R}$ (cf. exercice 264 de la feuille d'exercices n°19).

La fonction Φ est dérivable (donc continue) sur $\mathbb{R}$ et pour tout $x \in \mathbb{R}$ on a $\Phi'(x) = \dfrac{1}{\sqrt{2\pi}}\ e^{-\frac{x^2}{2}} > 0$. La fonction Φ est donc strictement croissante sur $\mathbb{R}$. De plus, puisque Φ est une fonction de répartition, on a :

$$\lim_{x\to-\infty} \Phi(x) = 0 \qquad \text{et} \qquad \lim_{x\to+\infty} \Phi(x) = 1.$$

D'après le théorème de la bijection, la fonction Φ réalise donc une bijection de $\mathbb{R}$ sur $]0,1[$ et sa bijection réciproque $\Phi^{-1}\colon]0,1[\to \mathbb{R}$ est strictement croissante sur $]0,1[$. En conséquence :

$$(I') \Longleftrightarrow \frac{20}{\sqrt{n}} \geq \Phi^{-1}(0,95).$$

Or d'après la table de valeurs (approchées) de Φ (cf. page 3 de la feuille d'exercices n°19), on a :

$$\Phi^{-1}(0,95) \approx 1,65.$$

En conséquence, une valeur approchée du nombre n cherché est donnée par la plus grande des solutions de l'inéquation :

$$\frac{20}{\sqrt{n}} \geq 1,65.$$

On trouve ainsi : $n \approx 146$.

Remarque conclusive : L'approximation de la loi de T_n par la loi normale nous a permis de résoudre, en utilisant la table de valeurs (approchées) de Φ, l'inéquation

$$P(G_n \geq -20) \geq 0,95$$

d'inconnue $n \in \mathbb{N}^*$ initiale, peu évidente à résoudre de façon exacte.

Corollaire 11 (application du théorème de la limite centrée à la loi de Poisson) : Soit $\lambda > 0$. Soit $(Y_n)_{n\in\mathbb{N}^*}$ une suite de variables aléatoires telle que :

$$\forall\, n \in \mathbb{N}^* \qquad Y_n \sim \mathcal{P}(n\,\lambda).$$

On pose pour tout $n \in \mathbb{N}^*$:

$$T_n = \frac{Y_n - n\,\lambda}{\sqrt{n\,\lambda}}.$$

Alors pour tout $x \in \mathbb{R}$, on a :

$$F_{T_n}(x) = P(T_n \leq x) \underset{n\to+\infty}{\to} \Phi(x)$$

et les propriétés du corollaire 9 obtenues en remplaçant partout S_n^* par T_n restent vraies. On dit que la suite de variables aléatoires $(T_n)_{n\in\mathbb{N}^*}$ converge en loi vers la loi $\mathcal{N}(0,1)$.

Preuve : Analogue à celle du corollaire 10, laissée en exercice.

7 Deux exercices

Exercice 1 : On réalise 400 fois une expérience de Bernoulli (i.e. une expérience avec pour résultat 0 ou 1) dont la probabilité de succès est $\frac{4}{5}$ (i.e. telle que la probabilité d'obtenir 1 est $\frac{4}{5}$). On suppose que les 400 expériences sont deux à deux indépendantes. On note X le nombre de succès obtenus.

1. Quel est la loi de X ?
2. Donner l'espérance et la variance de X.
3. Donner un minorant de la probabilité $P(300 < X < 400)$.

Exercice 2 : Soit $n \in \mathbb{N}^*$ et soit X une variable aléatoire suivant la loi $\mathcal{B}\left(n, \frac{1}{3}\right)$.

1. Préciser l'espérance et la variance de $F = \frac{X}{n}$.
2. Déterminer $n \in \mathbb{N}^*$ tel que
$$P(|F - E(F)| < 10^{-2}) \geq 0,98$$
de deux manières :
   (a) en utilisant l'inégalité de Bienaymé-Tchebychev ;
   (b) en utilisant une approximation par une loi normale.
3. Comparer les résultats obtenus en 2.(a) et 2.(b).