T.D. V - Estimation

I - Construction d'estimateurs

Solution de l'exercice 1.

1. D'après la définition, $Y_n = \sum_{i=1}^{n} X_i$ est une somme de n variables aléatoires indépendantes de mêmes lois de Bernoulli de paramètre p. Ainsi, Y_n suit une loi binomiale, $Y_n \hookrightarrow \mathscr{B}(n,p)$.

2. Comme $Y_n \hookrightarrow \mathscr{B}(n,p)$, alors

$$\begin{aligned}\mathbf{E}\left[Y_n\right] &= np,\\ \mathbf{V}\left(Y_n\right) &= np(1-p).\end{aligned}$$

D'après la formule de Kœnig-Huygens,

$$\begin{aligned}\mathbf{V}\left(Y_n\right) &= \mathbf{E}\left[Y_n^2\right] - \mathbf{E}\left[Y_n\right]^2\\ \mathbf{E}\left[Y_n^2\right] &= \mathbf{V}\left(Y_n\right) + \mathbf{E}\left[Y_n\right]^2\\ &= np(1-p) + (np)^2\\ &= np(1-p+np)\\ &= np(1+(n-1)p).\end{aligned}$$

3. En utlisant la question précédente et la linéarité de l'espérance,

$$\begin{aligned}\mathbf{E}\left[Y_n^2\right] &= np(1+(n-1)p)\\ \mathbf{E}\left[n^2\overline{X}_n^2\right] &= np(1+(n-1)p)\\ \mathbf{E}\left[\overline{X}_n^2\right] &= \frac{p(1+(n-1)p)}{n}.\end{aligned}$$

Ainsi, si $n \neq 1$, alors $\mathbf{E}\left[\overline{X_n}^2\right] \neq p$ et $\overline{X}_n^2$ n'est pas un estimateur sans biais de p^2.

Pour obtenir un estimateur sans biais de p^2, on utilise le fait que $\mathbf{E}\left[X_n\right] = p$. Ainsi,

$$\begin{aligned}\mathbf{E}\left[\overline{X}_n^2\right] &= \frac{p}{n} + \frac{n-1}{n}p^2\\ \mathbf{E}\left[\overline{X}_n^2\right] &= \frac{\mathbf{E}\left[\overline{X}_n\right]}{n} + \frac{n-1}{n}p^2\\ \mathbf{E}\left[\overline{X}_n^2\right] - \frac{1}{n}\mathbf{E}\left[\overline{X}_n\right] &= \frac{n-1}{n}p^2\\ \frac{n}{n-1}\mathbf{E}\left[\overline{X}_n^2 - \frac{1}{n}\overline{X}_n\right] &= p^2\\ \mathbf{E}\left[\frac{n}{n-1}\overline{X}_n^2 - \frac{1}{n-1}\overline{X}_n\right] &= p^2.\end{aligned}$$

Ainsi, $\frac{n}{n-1}\overline{X}_n^2 - \frac{1}{n-1}\overline{X}_n$ est un estimateur sans biais de p^2.

Remarque. Comme X_i est à valeurs dans $\{0,1\}$, alors $X_i^2 = X_i$. Ainsi,

$$\begin{aligned}\frac{n}{n-1}\overline{X}_n^2 - \frac{1}{n-1}\overline{X}_n &= \frac{n}{n-1} \times \left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)^2 - \frac{1}{(n-1)n}\sum_{i=1}^{n} X_i\\ &= \frac{1}{(n-1)n}\left[\sum_{i=1}^{n} X_i^2 + \sum_{1\leqslant i\neq j\leqslant n} X_iX_j - \sum_{i=1}^{n} X_i\right]\\ &= \frac{1}{n(n-1)}\sum_{1\leqslant i\neq j\leqslant n} X_iX_j\\ &= \frac{2}{n(n-1)}\sum_{1\leqslant i< j\leqslant n} X_iX_j\\ &= \frac{1}{\binom{n}{2}}\sum_{1\leqslant i< j\leqslant n} X_iX_j.\end{aligned}$$

Ainsi, un estimateur sans biais de p^2 est $\frac{1}{\binom{n}{2}}\sum_{1\leqslant i<j\leqslant n} X_iX_j$ (et il est aisé de vérifier directement que l'espérance de cet estimateur est bien p^2).

2e méthode. En utilisant la variance empirique, $\overline{X}_n - \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2$ est un estimateur sans biais de $p - p(1-p) = p^2$

De plus, comme $X_i(\Omega) = \{0, 1\}$, alors $X_i^2 = X_i$ et

$$\begin{aligned}
\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2 &= \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i^2 - 2X_iX_n + \overline{X}_n^2\right)\\
&= \frac{1}{n-1}\sum_{i=1}^{n}X_i - \frac{2}{n-1}\overline{X}_n\sum_{i=1}^{n}X_i + \frac{n}{n-1}\overline{X}_n^2\\
&= \frac{n}{n-1}\overline{X}_n - \frac{2n}{n-1}\overline{X}_n^2 + \frac{n}{n-1}\overline{X}_n^2\\
&= \frac{n}{n-1}\overline{X}_n - \frac{n}{n-1}\overline{X}_n^2\\
\overline{X}_n - \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2 &= \left(1 - \frac{n}{n-1}\right)\overline{X}_n + \frac{n}{n-1}\overline{X}_n^2\\
&= \frac{n}{n-1}\overline{X}_n^2 - \frac{1}{n-1}\overline{X}_n,
\end{aligned}$$

et on retrouve l'estimateur précédent. □

Solution de l'exercice 2.

1. Comme $\overline{X}_n$ est une fonction de $X_1, \ldots, X_n$, alors $\overline{X}_n$ est bien un estimateur. De plus,

$$\mathbf{E}\left[\overline{X}_n\right] = \mathbf{E}\left[\frac{1}{n}\sum_{i=1}^{n}X_i\right] = \frac{1}{n}\sum_{i=1}^{n}\mathbf{E}\left[X_i\right] = \frac{1}{n}\sum_{i=1}^{n}m = m.$$

Ainsi, $\overline{X}_n$ est un estimateur sans biais de m.

2. Comme les variables aléatoires $X_1, \ldots, X_n$ sont indépendantes, d'après la propriété de la variance,

$$\begin{aligned}
\mathbf{V}\left(\sum_{i=1}^{n}X_i\right) &= \sum_{i=1}^{n}\mathbf{V}\left(X_i\right) = \sum_{i=1}^{n}\sigma^2 = n\sigma^2\\
\mathbf{V}\left(\overline{X}_n\right) &= \mathbf{V}\left(\frac{1}{n}\sum_{i=1}^{n}X_i\right) = \frac{1}{n^2}\mathbf{V}\left(\sum_{i=1}^{n}X_i\right) = \frac{\sigma^2}{n}.
\end{aligned}$$

3. En utilisant une identité remarquable, puis les propriétés de la somme,

$$\begin{aligned}
ns_n &= \sum_{i=1}^{n}(X_i - \overline{X}_n)^2\\
&= \sum_{i=1}^{n}\left(X_i^2 - 2X_i\overline{X}_n + \overline{X}_n^2\right)\\
&= \sum_{i=1}^{n}X_i^2 - 2\overline{X}_n\sum_{i=1}^{n}X_i + \sum_{i=1}^{n}\overline{X}_n^2\\
&= \sum_{i=1}^{n}X_i^2 - 2n\overline{X}_n^2 + n\overline{X}_n^2\\
s_n &= \frac{1}{n}\sum_{i=1}^{n}X_i^2 - \overline{X}_n^2.
\end{aligned}$$

Ainsi, en utilisant la linéarité de l'espérance et les questions précédentes,

$$\begin{aligned}
\mathbf{E}\left[s_n\right] &= \frac{1}{n}\sum_{i=1}^{n}\mathbf{E}\left[X_i^2\right] - \mathbf{E}\left[\overline{X}_n^2\right]\\
&= \frac{1}{n}\sum_{i=1}^{n}\left(\mathbf{V}\left(X_i\right) + \mathbf{E}\left[X_i\right]^2\right) - \left(\mathbf{V}\left(\overline{X}_n\right) + \mathbf{E}\left[\overline{X}_n\right]^2\right)\\
&= \frac{1}{n}\sum_{i=1}^{n}(\sigma^2 + m^2) - \left(\frac{\sigma^2}{n} + m^2\right)\\
&= \sigma^2 + m^2 - \left(\frac{\sigma^2}{n} + m^2\right) = \frac{n-1}{n}\sigma^2.
\end{aligned}$$

4. D'après la question précédente, dès que $n \neq 1$, la variable aléatoire s_n n'est pas un estimateur sans biais de σ^2.

5. S_n est une fonction de $X_1, \ldots, X_n$ donc c'est un estimateur.
D'après la question précédente, $S_n = \frac{n}{n-1}s_n$. Ainsi,

$$\mathbf{E}\left[S_n\right] = \frac{n}{n-1}\mathbf{E}\left[s_n\right] = \frac{n}{n-1} \times \frac{n-1}{n}\sigma^2 = \sigma^2.$$

$S_n = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2$ est donc un estimateur sans biais de σ^2. □

Solution de l'exercice 3.

1. On est dans le cadre du calcul de la fonction de répartition d'un maximum :

$$\begin{aligned}
\mathbf{P}([M_n \leqslant i]) &= \mathbf{P}([\max\{X_1,\dots,X_n\} \leqslant i]) \\
&= \mathbf{P}\left(\bigcap_{k=1}^{n} [X_k \leqslant i]\right) \\
&= \prod_{k=1}^{n} \mathbf{P}([X_k \leqslant i]), \text{ en utilisant l'indépendance} \\
&= \prod_{k=1}^{n} \mathbf{P}([X_1 \leqslant i]), \text{ car } X_1,\dots,X_n \text{ sont de mêmes lois} \\
&= \mathbf{P}([X_1 \leqslant i])^n \\
&= \left(\sum_{\ell=1}^{i} \mathbf{P}([X_1 = \ell])\right)^n, \text{ car } X_1 \text{ prend les valeurs } 1,\dots,N \\
&= \left(\frac{i}{n}\right)^n, \text{ car } X_1 \hookrightarrow \mathscr{U}([\![1,N]\!])
\end{aligned}$$

2. En utilisant la formule précédente,

$$\mathbf{P}([M_n = N]) = \mathbf{P}([M_n \leqslant N]) - \mathbf{P}([M_n \leqslant N-1]) = 1 - \left(\frac{N-1}{N}\right)^n.$$

Comme $-1 < \frac{N-1}{N} < 1$, alors $\lim\limits_{n\to+\infty} \mathbf{P}([M_n = N]) = 1$.

3. On va effectuer un changement dans l'ordre des sommations :

$$\begin{aligned}
\sum_{i=1}^{N} \mathbf{P}(Y \geqslant i) &= \sum_{i=1}^{N}\sum_{j=i}^{N} \mathbf{P}(Y = j) \\
&= \sum_{1\leqslant i\leqslant j\leqslant N} \mathbf{P}(Y = j) \\
&= \sum_{j=1}^{N}\sum_{i=1}^{j} \mathbf{P}(Y = j) \\
&= \sum_{j=1}^{N} j\mathbf{P}(Y = j) = \mathbf{E}[Y].
\end{aligned}$$

4. En utilisant les questions précédentes, comme M_n est à valeurs dans $[\![1,N]\!]$,

$$\begin{aligned}
\mathbf{E}[M_n] &= \sum_{i=1}^{N} \mathbf{P}([M_n \geqslant i]) \\
&= \sum_{i=1}^{N} (1 - \mathbf{P}([M_n \leqslant i-1])) \\
&= \sum_{i=1}^{N} \left(1 - \left(\frac{i-1}{N}\right)^n\right) \\
&= N - \sum_{i=1}^{N} \left(\frac{i-1}{N}\right)^n \\
&= N - \sum_{i=0}^{N-1} \left(\frac{i}{N}\right)^n.
\end{aligned}$$

5. D'après la question précédente, pour $j \in [\![0, N-1]\!]$,

$$\begin{aligned}
\frac{0}{N} \leqslant \frac{j}{N} &\leqslant \frac{N-1}{N} \\
0 \leqslant \left(\frac{j}{N}\right)^n &\leqslant \left(\frac{N-1}{N}\right)^n \\
-\left(\frac{N-1}{N}\right)^n \leqslant -\left(\frac{j}{N}\right)^n &\leqslant 0 \\
-\sum_{j=0}^{N-1}\left(1-\frac{1}{N}\right)^n \leqslant -\sum_{j=0}^{N-1}\left(\frac{j}{N}\right)^n &\leqslant 0 \\
N - N\left(1-\frac{1}{N}\right)^n \leqslant \mathbf{E}[M_n] &\leqslant N.
\end{aligned}$$

Comme $\lim\limits_{n\to+\infty} \left(1-\frac{1}{N}\right)^n = 0$, alors d'après le théorème d'encadrement,

$$\lim_{n\to+\infty} \mathbf{E}[M_n] = N.$$

Remarque. Pour obtenir l'encadrement, il n'y avait pas besoin de la formule sur l'espérance car

* d'une part, $M_n \leqslant N$, donc $\mathbf{E}[M_n] \leqslant \mathbf{E}[N] = N$,
* d'autre part, $\mathbf{E}[M_n] = \sum_{k=1}^{N} k\mathbf{P}(M_n = k) \geqslant N\mathbf{P}(M_n = N)$,

ce qui donne l'encadrement annoncé. □

II - Comparaison d'estimateurs

Solution de l'exercice 4.

1. Comme $\overline{X}_n$ et T_n sont des fonctions de $X_1, \ldots, X_n$, alors ce sont des estimateurs.
En utilisant la linéarité de l'espérance,

$$\mathbf{E}\left[\overline{X}_n\right] = \frac{1}{n}\sum_{i=1}^{n}\mathbf{E}[X_i] = \frac{1}{n}\sum_{i=1}^{n}\lambda = \lambda.$$

Ainsi, $\overline{X}_n$ est un estimateur sans biais de λ.

En utilisant les propriétés de la somme,

$$\begin{aligned}
T_n &= \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2 \\
&= \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i^2 - 2X_i\overline{X}_n + \overline{X}_n^2\right) \\
&= \frac{1}{n-1}\left[\sum_{i=1}^{n}X_i^2 - 2\overline{X}_n\sum_{i=1}^{n}X_i + n\overline{X}_n^2\right] \\
&= \frac{1}{n-1}\left[\sum_{i=1}^{n}X_i^2 - n\overline{X}_n^2\right].
\end{aligned}$$

Ainsi,

$$\begin{aligned}
\mathbf{E}[T_n] &= \frac{1}{n-1}\left[\sum_{i=1}^{n}\mathbf{E}\left[X_i^2\right] - n\mathbf{E}\left[\overline{X}_n^2\right]\right] \\
&= \frac{n}{n-1}\left[\mathbf{E}\left[X_1^2\right] - \mathbf{E}\left[\overline{X}_n^2\right]\right] \\
&= \frac{n}{n-1}\left[\mathbf{V}(X_1) + \mathbf{E}[X_1]^2 - \mathbf{V}\left(\overline{X}_n\right) - \mathbf{E}\left[\overline{X}_n\right]^2\right].
\end{aligned}$$

Or, comme les $X_1, \ldots, X_n$ sont indépendantes,

$$\begin{aligned}
\mathbf{V}\left(\overline{X}_n\right) &= \mathbf{V}\left(\frac{1}{n}\sum_{i=1}^{n}X_i\right) = \frac{1}{n^2}\mathbf{V}\left(\sum_{=1}^{n}X_i\right) \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\mathbf{V}(X_i) = \frac{\lambda}{n}.
\end{aligned}$$

Finalement,

$$\mathbf{E}[T_n] = \frac{n}{n-1}\left[\lambda + \lambda^2 - \frac{\lambda}{n} - \lambda^2\right] = \frac{n}{n-1}\frac{n-1}{n}\lambda = \lambda.$$

Ainsi, T_n est un estimateur sans biais de λ.

2. Comme $\overline{X}_n$ est un estimateur sans biais, son risque quadratique est égal à sa variance. Ainsi, d'après les calculs réalisés à la question précédente,

$$R_\lambda\left(\overline{X}_n\right) = \mathbf{V}\left(\overline{X}_n\right) = \frac{\lambda}{n}.$$

3. Comme $\frac{2\lambda^2}{n-1} > 0$, alors $\overline{X}_n$ (qui est la moyenne empirique) est un estimateur préférable de λ que T_n (qui est l'estimateur sans biais de la variance) car son risque quadratique est plus faible. □

Solution de l'exercice 5.

1. Commme $\overline{X}_n$ est une fonction de $X_1, \ldots, X_n$, alors $\overline{X}_n$ est un estimateur.
De plus, d'après la linéarité de l'espérance,

$$\mathbf{E}\left[\overline{X}_n\right] = \frac{1}{n}\sum_{i=1}^{n}\mathbf{E}[X_i] = \theta.$$

Ainsi, $\overline{X}_n$ est un estimateur sans biais de θ.

Comme $X_1, \ldots, X_n$ sont indépendantes, d'après les propriétés de la variance,

$$\mathbf{V}\left(\overline{X}_n\right) = \frac{1}{n^2}\mathbf{V}\left(\sum_{i=1}^{n}X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n}\mathbf{V}(X_i) = \frac{1}{n^2}\sum_{i=1}^{n}1 = \frac{1}{n}.$$

2. Comme Y_n est une fonction de $X_1, \ldots, X_n$, alors Y_n est un estimateur. De plus, d'après la linéarité de l'espérance,

$$\mathbf{E}\left[Y_n\right] = \sum_{i=1}^{n} \alpha_i \mathbf{E}\left[X_i\right] = \theta \sum_{i=1}^{n} \alpha_i.$$

Comme $\theta \neq 0$, alors Y_n est un estimateur sans biais si et seulement si $\sum_{i=1}^{n} \alpha_i = 1$.

3. En utilisant la bilinéarité de la covariance,

$$\begin{aligned}\mathrm{Cov}\left(\overline{X}_n, Y_n\right) &= \mathrm{Cov}\left(\frac{1}{n}\sum_{i=1}^{n} X_i, \sum_{j=1}^{n} \alpha_j X_j\right)\\ &= \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_j \mathrm{Cov}\left(X_i, X_j\right).\end{aligned}$$

De plus, en utilisant l'indépendance,

$$\mathrm{Cov}\left(X_i, X_j\right) = \begin{cases} 0 & \text{si } j \neq i \\ \mathbf{V}\left(X_i\right) & \text{si } i = j \end{cases}.$$

Alors,

$$\begin{aligned}\mathrm{Cov}\left(\overline{X}_n, Y_n\right) &= \frac{1}{n}\sum_{i=1}^{n}\left(\alpha_i \underbrace{\mathbf{V}\left(X_i\right)}_{=1} + \sum_{j\neq i} 0\right)\\ &= \frac{1}{n}.\end{aligned}$$

Ainsi, d'après la positivité de la variance,

$$\begin{aligned}0 \leqslant \mathbf{V}\left(\overline{X}_n - Y_n\right) &= \mathbf{V}\left(\overline{X}_n\right) + \mathbf{V}\left(Y_n\right) - 2\mathrm{Cov}\left(\overline{X}_n, Y_n\right)\\ &= \frac{1}{n} + \mathbf{V}\left(Y_n\right) - 2\frac{1}{n} = \mathbf{V}\left(Y_n\right) - \mathbf{V}\left(\overline{X}_n\right).\end{aligned}$$

Ainsi, $\mathbf{V}\left(\overline{X}_n\right) \leqslant \mathbf{V}\left(Y_n\right)$ et $\overline{X}_n$ a un risque quadratique inférieur à celui de Y_n.

L'égalité $\mathbf{V}\left(\overline{X}_n\right) = \mathbf{V}\left(Y_n\right)$ a lieu si et seulement si $\mathbf{V}\left(\overline{X}_n - Y_n\right) = 0$, i.e. $\overline{X}_n - Y_n = c$ presque sûrement. Comme $\mathbf{E}\left[\overline{X}_n\right] = \mathbf{E}\left[Y_n\right] = \theta$, alors $c = 0$ et $\overline{X}_n = Y_n$ presque sûrement.

4. Ainsi, parmi les estimateurs sans biais qui sont des combinaisons linéaires de $X_1, \ldots, X_n$, l'estimateur de θ qui a le plus faible risque quadratique est l'estimateur de la moyenne empirique. $\square$