Feuille n° 3 – Statistiques et Simulations Probabilistes

Exercice 1

Soit (X_n) une suite de variables aléatories i.i.d. de Bernoulli de paramètre p inconnu.

1. Montrer que $\overline{X}_n$ est un estimateur consistant de p. *On rappelle qu'un estimateur consistant est un estimateur qui converge en probabilité vers le paramètre qu'il est destiné à estimer. Dans le cas qui nous occupe on a une suite de variables indépendantes de carré intégrable (il est facile de vérifier que $\mathbb{E}[X_i^2] = p < \infty$ pour chaque $i \in \mathbb{N}$). La loi faible des grands nombres trouve donc à s'appliquer à $\bar{X}_N$ qui converge en probabilité vers $\mathbb{E}[X_1] = p$ et implique donc que $\bar{X}_n$ est donc un estimateur consistant de p.*
2. A partir de $\overline{X}_n$, proposer un estimateur consistant de $p(1-p)$. *Un résultat du cours énonce que si f est une fonction continue sur un intervalle ouvert de $\mathbb{R}$ et T_n est un estimateur consistant d'un paramètre de c alors $f(T_n)$ est un estimateur consistant de $f(c)$. Par conséquent et par le point précédent, pour $c = p$ et $f(x) = x(1-x)$ il est naturel de choisir $\bar{X}_n(1-\bar{X}_n)$ comme estimateur de $p(1-p)$.*

Exercice 2

Soit (X_n) une suite de variables i.i.d. de gaussiennes de paramètres $(0, \theta^2)$ où $\theta > 0$ est inconnu. On considère la variable

$$T_n = \frac{1}{n}\sum_{i=1}^{n} X_i^2 .$$

Est-elle un estimateur consistant de θ^2 ? *On sait que θ^2 est la variance des variables gaussiennes et que celles-ci sont d'espérance nulle. Par conséquent, on sait que $\theta^2 = \mathbb{E}[X_i^2]$ pour chaque $i \in \mathbb{N}$. Par ailleurs, $\mathbb{E}[X_i^4] < \infty$ et donc si on définit pour chaque $n \in \mathbb{N}$, $Y_n = X_n^2$. On peut appliquer la loi faible des grands nombres à*

$$\bar{Y}_n = \frac{1}{n}\sum_{i=1}^{n} Y_i.$$

qui converge vers $\mathbb{E}[Y_1] = \mathbb{E}[X_1^2] = \theta^2$ et il est immédiat que $\bar{Y}_n = T_n$. La réponse est donc oui !

Exercice 3

Soit deux estimateurs S et T d'un paramètre θ tels que $\mathrm{Var}[S] = 40$ et $\mathrm{Var}[T] = 4$.

1. On suppose que $\mathbb{E}[S] = \theta$ et $\mathbb{E}[T] = \theta + 3$. Quel estimateur choisiriez-vous ? Calculons le risque quadratique de chacun des deux estimateurs.

2. (a)

$$\begin{aligned} R_S(\theta) &= b_S^2(\theta) + \mathrm{V}ar(S) \\ &= 0 + \mathrm{V}ar(S) \\ &= 40 \end{aligned} \tag{1}$$

(b)

$$\begin{aligned} R_T(\theta) &= b_T^2(\theta) + \mathrm{V}ar(T) \\ &= 3^2 + \mathrm{V}ar(T) \\ &= 13 \end{aligned} \tag{2}$$

Le risque quadratique de T est plus petit, c'est donc celui qu'on choisit bien que S soit sans biais et que T soit biaisé.

3. On suppose que $\mathbb{E}[S] = \theta$ et $\mathbb{E}[T] = \theta + a$, $a > 0$. Discutez en fonction de a l'estimateur que vous choisiriez. *On a toujours $R_S(\theta) = 40$, en revanche on a à présent $R_T(\theta) = a^2 + 4$ et on choisit l'estimateur T tant que $R_T(\theta) < R_S(\theta)$, c'est à dire $a^2 < 36$.*

Exercice 4

Soit $X_1, \dots, X_n$ un n-échantillon de loi $\mathcal{E}xp(\lambda)$. On souhaite estimer $1/\lambda$.

1. Proposer un estimateur sans biais de $1/\lambda$.
   $\bar{X}_n$ est un estimateur sans biais de $\frac{1}{\lambda}$, en effet on a $\mathbb{E}[X_i] = \frac{1}{\lambda}$ pour chaque i et donc

$$\begin{aligned} \mathbb{E}[\bar{X}_n] &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[X_i] \\ &= \frac{1}{\lambda} \end{aligned}$$

2. Soit $M_n = \min(X_1, \dots, X_n)$. Rappeler la loi de M_n. *Comme on l'a vu dans la feuille de TD2 et TD1, la loi de M_n est $\mathcal{E}xp(n\lambda)$.*
3. A l'aide de l'Exercice 4 de la Feuille n° 2, proposer un autre estimateur sans biais de $1/\lambda$. *Calculons $\mathbb{E}[M_n] = \frac{1}{n\lambda}$, par conséquent $T_n = nM_n$ est un estimateur sans biais de $1/\lambda$.*
4. Lequel de ces deux estimateurs choisiriez-vous ? *Il est nécessaire de calculer le risque quadratique pour pouvoir comparer ces estimateurs. Comme ils sont tous deux sans biais le risque quadratique est donné par la variance.*

$$\begin{aligned} \mathrm{V}ar(\bar{X}_n) &= \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{V}ar(X_i) \\ &= \frac{1}{n\lambda^2} \end{aligned}$$

   La première ligne suit de l'indépendance des variables de l'échantillon qui implique que la variance d'une somme est donnée par la somme des variances. Calculons à présent la variance de T_n

$$\begin{aligned} \mathrm{V}ar(T_n) &= n^2 Var(M_n) \\ &= n^2 \cdot \frac{1}{n^2\lambda^2} \\ &= \frac{1}{\lambda^2} \end{aligned}$$

   On voit donc que dès que $n > \lambda$, le risque quadratique de $\bar{X}_n$ est plus petit que celui de T_n, on choisit donc $\bar{X}_n$.

Exercice 5

On reprend l'Exercice 6 de la Feuille n° 2. Lequel des deux estimateurs proposés choisiriez-vous ?
Comme ils sont sans biais, il suffit de calculer leurs variances pour obtenir leurs risques quadratiques.

$$\begin{aligned} Var(T_1) &= \frac{16}{n^2}Var(N_1) \\ &= \frac{16}{n^2}p_1(1-p_1) \\ &= \frac{1}{n^2}(4-\theta^2) \end{aligned}$$

$$\begin{aligned} Var(T_1) &= \frac{16}{n^2}p_2(1-p_2) \\ &= \frac{1}{n^2}\theta(1-\theta) \end{aligned}$$

On voit que T_1 est meilleur que T_2 tant que $\theta > 4$ et si $\theta < 4$, les rôles s'inversent. Or dans le cas qui nous occupe, on a toujours $\theta < 1$ et donc on choisit toujours T_2.

Exercice 6

Soit $\overline{X}_n$ et $\overline{Y}_m$ les moyennes empiriques de deux échantillons indépendants de tailles respectives n et m de même loi de moyenne μ et de même variance σ^2. On propose un nouvel estimateur

$$T = r\overline{X}_n + (1-r)\overline{Y}_m$$

où r est un réel compris entre 0 et 1.

1. Montrer que T est un estimateur sans biais de μ. *En utilisant les propirétés habituelles de la moyenne empirique, on voit facilement que $\mathbb{E}[T] = r\mu + (1-r)\mu = \mu$*
2. Montrer que la variance de T est minimale pour $r = n/(n+m)$.

$$\begin{aligned} Var(T) &= r^2 Var(\bar{X}_n) + (1-r)^2 Var(\bar{Y}_m) \\ &= r^2\frac{\sigma^2}{n} + (1-r)^2\frac{\sigma^2}{m} \\ &= \sigma^2(\frac{r^2}{n} + \frac{(1-r)^2}{m}). \end{aligned}$$

La première ligne suit de l'indépendance des échantillons et des propriétés habituelles de la variance. Pour arriver à la conclusion, il suffit de minimiser en fonction de r en calculant la dérivée.

Exercice 7

Soit $X_1, \ldots, X_n$ un échantillon de loi de Bernoulli de paramètre p. On considère les estimateurs suivants :

$$T_1 = \frac{1}{n}(X_1 + \ldots + X_n) \quad \text{et} \quad T_2 = \min(X_1, \ldots, X_n).$$

1. T_1 et T_2 sont-ils des estimateurs sans biais de p ? *Déterminons d'abord la loi de T_2, comme cette variable ne prend que les valeurs 0 et 1, il s'agit d'une variable de Bernoulli. De plus $T_2 = 1$ si et seulement si toutes les variables de l'échantillon prennent la valeur 1. Par conséquent $\mathbb{P}[T_2 = 1] = p^n$ et $\mathbb{P}[T_2 = 0] = 1 - p^n$. L'espérance de T_2 est donc $\mathbb{E}[T_2] = p^n$ tandis que l'espérance de T_1 est p (résultat établi à de nombreuses reprises). T_1 est donc sans biais tandis que T_2 est biaisé.*
2. Montrer que

$$\text{MSE}(T_1) = \frac{p(1-p)}{n} \quad \text{et} \quad \text{MSE}(T_2) = p^n - 2p^{n+1} + p^2.$$

T_1 est sans biais et son risque quadratique est donc égal à sa variance que l'on a calculé à de nombreuses reprises :

$$\begin{aligned} Var(T_1) &= \frac{1}{n^2} n Var(X_1) \\ &= \frac{p(1-p)}{n}. \end{aligned}$$

Le biais de T_2 est

$$b_{T_2}(p) = (p^n - p).$$

quant à la variance de $T_2 (\sim Ber(p^n))$, elle est donnée par

$$Var(T_2) = p^n(1-p^n).$$

Le risque quadratique est donc

$$\begin{aligned} R_T(p) &= (p^n - p)^2 + p^n(1-p^n) \\ &= p^n - 2p^{n+1} + p^2. \end{aligned}$$

3. Lequel des deux estimateurs est le plus efficace pour $n = 2$?

Exercice 8

Soit $(X_1, \ldots, X_n)$ un échantillon de loi $\mathcal{E}xp(\lambda)$. On sait que $\overline{X}_n$ est un estimateur sans biais de $1/\lambda$. Considérons des estimateurs plus généraux de la forme

$$T = c(X_1 + \ldots + X_n)\,,$$

où c est un réel.

1. Calculer le risque quadratique moyen de T en fonction de c. *On a* $\mathbb{E}[T] = cn/\lambda$ *et donc :*

$$b_T(\frac{1}{\lambda}) = (cn-1)\frac{1}{\lambda}.$$

Comme d'autre part $Var[T] = c^2 n Var[X_1] = nc^2/\lambda^2$*, on a donc*

$$R_T(\frac{1}{\lambda}) = \frac{(cn-1)^2}{\lambda^2} + \frac{nc^2}{\lambda^2}.$$

2. Pour quelle valeur de c ce risque est-il minimum ? En minimisant cette expression en fonction de c, on trouve que le minimum est en $c = 1/(n+1)$.