Université de Lille

Méthodes de réduction de la variance

Enjeux dans le contexte des événements rares

Reda Mdair

2024

Table des matières

1 Introduction

Les simulations de Monte Carlo constituent une approche incontournable pour l'estimation de quantités probabilistes complexes, en particulier lorsque les expressions analytiques sont inaccessibles. Leur simplicité conceptuelle et leur convergence garanties en font des méthodes puissantes et polyvalentes. Cependant, leur efficacité pratique est souvent limitée par un inconvénient majeur : la variance des estimateurs obtenus peut être élevée, rendant les résultats imprécis à moins d'augmenter significativement la taille des échantillons.

Ce problème est d'autant plus crucial dans le contexte des événements rares, où la probabilité recherchée est extrêmement faible. Dans ces situations, une approche naïve de Monte Carlo peut s'avérer inefficace, nécessitant des milliards de simulations pour obtenir une estimation exploitable. C'est ici que les méthodes de réduction de la variance interviennent.

L'objectif de ce projet est d'explorer ces méthodes et de montrer comment elles permettent d'améliorer drastiquement la précision des estimations tout en réduisant le coût computationnel. Nous examinerons notamment trois techniques majeures :
- l'échantillonnage préférentiel (Importance Sampling), qui consiste à modifier la distribution des simulations pour mieux capturer les régions critiques de l'espace des événements rares.
- l'utilisation de variables de contrôle, qui exploite la corrélation entre variables aléatoires pour réduire la variance des estimateurs.
- la stratification, qui segmente l'espace d'échantillonnage en strates homogènes afin d'obtenir des estimations plus précises par regroupement des simulations.

À travers une approche théorique rigoureuse, nous illustrerons comment ces méthodes transforment l'efficacité des simulations de Monte Carlo, en particulier pour les événements rares. L'application finale démontrera leur pertinence dans un cas réel, ancré dans le domaine du sport, en évaluant la probabilité d'un score exceptionnel dans un match de football.

2 Méthodes de réduction de la variance

Les méthodes de Monte Carlo sont des approches très couramment utilisées dans l'estimation d'événements rares. La probabilité associée à ces derniers peut s'exprimer comme une intégrale de la forme :

$$I = E_f[h(X)] = \int h(x)f(x)dx \quad (*)$$

où f est la densité de probabilité de la variable aléatoire X et h est une fonction de X. Cette intégrale peut être estimée grâce à la méthode de Monte Carlo classique dont on rappelle le principe :

Définition (Monte Carlo classique)
Soit $(X_1, \ldots, X_n)$ un échantillon de n variables aléatoires i.i.d simulées selon une loi f. L'estimateur de Monte Carlo classique (MCC) de $E_f[h(X)]$ est alors donné par :

$$\delta_{MCC} = \frac{1}{n}\sum_{i=1}^{n} h(X_i) \qquad (2.0)$$

Cette méthode est un résultat puissant en théorie des probabilités et repose sur la loi forte des grands nombres. Son avantage est que l'estimateur de MCC est sans biais et fortement consistant.

Cependant, une limitation des méthodes de Monte Carlo repose sur leur précision qui dépend de la variance de l'estimateur. Cette dernière peut notamment être élevée dans le contexte des événements rares.
Afin de résoudre ce défi, diverses méthodes de réduction de la variance ont été élaborées pour améliorer l'efficacité des simulations de Monte Carlo classique dans l'estimation des phénomènes extrêmes.

2.1 Échantillonnage préférentiel

Le principe de l'échantillonnage préférentiel, ou *« importance sampling »* en anglais, consiste à simuler non pas selon la densité cible f de X mais selon une densité instrumentale $\tilde{f}$. Cette dernière doit être choisie judicieusement afin d'améliorer les performances des estimateurs basés sur Monte Carlo.
Une application concrète sera consacrée à cette méthode dans la section 5.

Rappelons que l'on cherche à approcher $E_f(h(X))$ avec X une variable aléatoire de densité f. Pour toute densité $\tilde{f}$, nous pouvons alors réécrire (2.0) sous la forme :

$$I = E_f[h(X)] = \int h(x)f(x)dx = \int \left(\frac{h(x)f(x)}{\tilde{f}(x)}\right)\tilde{f}(x)dx$$

$$= E_{\tilde{f}}\left[\frac{h(Y)f(Y)}{\tilde{f}(Y)}\right]$$

où Y est une variable aléatoire de loi $\tilde{f}$.

Sachant que l'approximation de (MCC) est donnée par $\delta_{MCC} = \frac{1}{n}\sum_{i=1}^{n} h(X_i)$ avec $(X_1, \ldots, X_n)$ un n-échantillon simulé selon la loi f , l'estimateur par échantillonnage préférentiel (EP) de $E_f(h(X))$ est lui construit comme :

$$\boxed{\delta_{EP} = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{h(Y_i)f(Y_i)}{\tilde{f}(Y_i)}\right]} \tag{2.1}$$

où $(Y_1, \ldots, Y_n)$ est un n-échantillon simulé selon la loi $\tilde{f}$.

Pour que cet estimateur soit bien défini, il faut s'assurer que le support de hf soit inclus dans celui de $\tilde{f}$ afin de ne pas diviser par 0. Autrement dit, la condition nécessaire est : $h(y)f(y) > 0 \Rightarrow \tilde{f}(y) > 0$.

Biais de l'estimateur : L'estimateur par (EP) est sans biais, sous réserve que $supp(hf) \subseteq supp(\tilde{f})$.

Convergence de l'estimateur : La loi forte des grands nombres appliquée à la suite de variables aléatoires i.i.d. $(\frac{h(Y_i)f(Y_i)}{f(\tilde{Y}_i)})_{i\geq 1}$ (et d'espérance finie sous $\tilde{f}$) donne :

$$\delta_{EP} = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{h(Y_i)f(Y_i)}{\tilde{f}(Y_i)}\right] \xrightarrow[n\to\infty]{p.s.} E_{\tilde{f}}\left[\frac{h(Y)f(Y)}{\tilde{f}(Y)}\right] = I.$$

Comparons désormais la variance de l'estimateur par échantillonnage préférentiel avec celle de (MCC). En notant X et Y des variables aléatoires de densités respectives f et $\tilde{f}$, on obtient les résultats :

- $Var(\delta_{MCC}) = Var\left(\frac{1}{n}\sum_{i=1}^{n} h(X_i)\right)$
  $= \frac{1}{n}Var(h(X))$
  $= \frac{1}{n}(E_f[h(X)^2] - I^2)$

- $Var(\delta_{EP}) = Var\left(\frac{1}{n}\sum_{i=1}^{n}\frac{h(Y_i)f(Y_i)}{\tilde{f}(Y_i)}\right)$
  $= \frac{1}{n}Var\left(\frac{h(Y)f(Y)}{\tilde{f}(Y)}\right)$
  $= \frac{1}{n}\left(E_{\tilde{f}}\left[\frac{h(Y)^2 f(Y)^2}{\tilde{f}(Y)^2}\right] - I^2\right)$
  $= \frac{1}{n}\left(E_f\left[\frac{h(X)^2 f(X)}{\tilde{f}(X)}\right] - I^2\right)$
  $= \frac{1}{n}\left(E_f[h(X)^2 w(X)] - I^2\right)$

où l'on a utilisé la notation $w(X)$ pour désigner le poids d'importance $\frac{f(X)}{\tilde{f}(X)}$. Ces deux variances diffèrent d'un facteur près : le poids d'importance. Il suffit donc de choisir une loi instrumentale $\tilde{f}$ adaptée afin que la pondération w minimise au plus le terme $E_f[h(X)^2 w(X)]$.

Cette densité minimisant la variance possède bien une solution explicite, donnée par le lemme suivant :

Lemme (Loi instrumentale optimale)
Soit Y une variable aléatoire de densité $\tilde{f}$ vérifiant $E_{\tilde{f}}[h(Y)^2 w(Y)^2] < \infty$. Alors la densité minimisant la variance est donnée par :

$$\tilde{f}(y) = \frac{|h(y)|f(y)}{\int |h(y)|f(y)dy}$$

Preuve :

$$\begin{aligned} Var\left(h(Y)\frac{f(Y)}{\tilde{f}(Y)}\right) &= E_{\tilde{f}}\left[h(Y)^2\frac{f(Y)^2}{\tilde{f}(Y)^2}\right] - E_{\tilde{f}}\left[h(Y)\frac{f(Y)}{\tilde{f}(Y)}\right]^2 \\ &= E_{\tilde{f}}\left[h(Y)^2 w(Y)^2\right] - \left(\int h(y)f(y)dy\right)^2 \end{aligned}$$

Le second terme ne dépendant pas de $\tilde{f}$, minimiser la variance de l'estimateur par (EP) revient donc à minimiser la quantité $E[w(Y)^2 h(Y)^2]$. L'inégalité de Jensen induit par la suite :

$$E[w(Y)^2 h(Y)^2] = E[(w(Y)|h(Y)|)^2] \geq E[w(Y)|h(Y)|]^2 = E[|h(X)|]^2$$

La dernière inégalité se transforme en égalité uniquement si la variable aléatoire $w(Y)|h(Y)|$ est constante p.s. Sachant que Y est de loi $\tilde{f}$, cela se traduit par l'existence d'une constante c telle que, pour tout y vérifiant $\tilde{f}(y) > 0$, on a :

$$w(y)|h(y)| = c \Leftrightarrow \tilde{f}(y) = \frac{|h(y)|f(y)}{c} \Leftrightarrow \tilde{f}(y) = \frac{|h(y)|f(y)}{\int |h(y)|f(y)dy}$$

où l'on a utilisé le fait que $\tilde{f}$ est une densité pour en déduire la constante c à la dernière équivalence. □

Cependant, cette densité minimisant la variance est hors de portée, puisqu'elle dépend précisément de la quantité I que l'on cherche à estimer. Néanmoins, ce résultat nous fournit une indication intéressante : une grande réduction de la variance peut être obtenu en choisissant $\tilde{f}$ telle que le ratio $\frac{|h|f}{\tilde{f}}$ soit presque constant et de variance finie.

Mesure de la qualité des échantillons générés :
Hormis la possibilité de comparer l'estimateur par (EP) à celui de (MCC) via la variance, il peut également être intéressant de s'assurer que le choix de la densité instrumentale $\tilde{f}$ est bon. On définit alors le critère *"effective sample size"* donné par le coefficient :

$$ESS = \frac{(\sum_{i=1}^n w_i)^2}{\sum_{i=1}^n w_i^2}$$

Cette quantité vaut n si tous les poids d'importance valent 1, c'est-à-dire si $\tilde{f} = f$ (soit le critère de MCC). Elle représente donc le nombre d'observations apportant de l'information. Une valeur proche approchant n indique une faible variance des poids, et induit par conséquent une estimation fiable.

Limites de l'échantillonnage préférentiel :
Le choix de la loi instrumentale reste néanmoins soumis à une contrainte. En effet, la variance de l'estimateur δ_{EP} peut ne pas être finie. Dans le cas où $Var_f(h(X)) < \infty$, il suffit alors que $\tilde{f}$ ait des queues de distribution plus lourdes que celles de f (ou autrement dit que le ratio $\frac{f}{\tilde{f}}$ soit borné) afin de contourner ce problème. Dans le cas où le facteur n'est pas borné, les poids d'importance seront très variables, ce qui entraîne alors des changements brusques d'une itération à l'autre.

2.2 Variables de contrôle

La méthode des variables de contrôle est une technique très utilisée dans le contexte de la réduction de variance. Sa particularité réside dans son exploitation de la corrélation entre les variables aléatoires. Nous verrons notamment que, dans un cas particulier, elle diminue forcément la variance de l'estimateur classique.

Dans cette section, on définit $\theta = E[X]$ la quantité d'intérêt à estimer.
Soit Y une variable aléatoire de moyenne connue $y = E[Y]$. La méthode des variables de contrôle consiste alors à générer $(X_1, Y_1), ..., (X_n, Y_n)$ i.i.d, de même loi que (X,Y), et de combiner les moyennes empiriques $\bar{X_n}$ et $\bar{Y_n}$ afin d'améliorer l'estimateur classique : $\bar{X_n} = \frac{1}{n}\sum_{i=1}^{n} X_i$.

On définit δ_{VC} l'estimateur associé à la méthode des variables de contrôle comme :

$$\boxed{\delta_{VC} = \frac{1}{n}\sum_{i=1}^{n}(X_i - c(Y_i - y)) = \bar{X_n} - c(\bar{Y_n} - y)} \tag{2.2}$$

où la constante c est à déterminer librement. Par défaut, le choix naïf est $c = 1$.

Biais de l'estimateur :

$$\begin{aligned} E[\delta_{VC}] &= E[\bar{X_n}] - cE[\bar{Y_n}] + cE[y] \\ &= E[X] \end{aligned}$$

L'estimateur par (VC) est bien sans biais.

Convergence de l'estimateur :
La loi forte des grands nombres appliquée pour les suites de variables aléatoires i.i.d. $(X_i)_{1\leq i\leq n}$ et $(Y_i)_{1\leq i\leq n}$ donne :

$$\begin{cases} \frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow[n\to\infty]{p.s.} E[X] \\ \frac{1}{n}\sum_{i=1}^{n} Y_i - y \xrightarrow[n\to\infty]{p.s.} 0 \end{cases} \quad \Rightarrow \delta_{VC} \xrightarrow[n\to\infty]{p.s.} E[X]$$

Variance de l'estimateur :

$$\begin{aligned} Var(\delta_{VC}) &= Var(\bar{X_n} - c(\bar{Y_n} - E[Y])) \\ &= Var(\bar{X_n}) + c^2 Var(\bar{Y_n}) - 2cCov(\bar{X_n}, \bar{Y_n}) \\ &= Var(\bar{X_n}) + \frac{1}{n}\left(c^2 Var(Y) - 2cCov(X,Y)\right) \end{aligned}$$

Cette expression admet bien un choix optimal de c pour minimiser la variance, donné par la proposition qui suit :

Proposition (Constante c optimale)
L'estimateur par variables de contrôle δ_{VC} a une variance inférieure à l'estimateur classique $\bar{X_n}$ si et seulement si :

$$c^2 Var(Y) - 2cCov(X,Y) < 0$$

L'estimateur de variance minimale δ^*_{VC} est alors obtenu pour :

$$c^* = \underset{c\in\mathbb{R}}{\operatorname{argmin}}(Var(\delta_{VC}(c)) = \frac{Cov(X,Y)}{Var(Y)}$$

Preuve : Notons $V = Var(\delta_{VC}(c))$. Cette fonction est polynomiale, et vérifie :
$\frac{\partial V}{\partial c} = \frac{1}{n}(2cVar(Y) - 2Cov(X,Y))$ avec $\frac{\partial V}{\partial c} = 0 \Leftrightarrow c = \frac{Cov(X,Y)}{Var(Y)}$
$\frac{\partial^2 V}{\partial c^2} = \frac{2}{n}Var(Y) > 0$
Donc V est strictement convexe et c^* est bien le minimiseur de la variance. □

Le choix optimal de c^* nous conduit ainsi à la variance :

$$Var(\delta^*_{VC}) = Var(\bar{X_n})(1 - \rho(X,Y)^2)$$

où $\rho(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}}$ désigne le coefficient de corrélation entre X et Y.
Ce dernier calcul nous montre notamment que :

$$Var(\delta^*_{VC}) \leq Var(\bar{X_n})$$

L'estimateur de variance minimale δ^*_{VC} a donc une variance inférieure à celle de $\bar{X_n}$ dès lors que la corrélation entre X et Y n'est pas nulle. Cet écart est d'autant plus accentué que la corrélation est élevée.

Comment déterminer c^* sans connaître $\mathbf{Cov(X,Y)}$ et $\mathbf{Var(Y)}$?
Dans la plupart des cas, ces termes sont inconnus. On utilise alors l'estimateur basé sur l'échantillon :

$$\hat{c}^*_n = \frac{\sum_{i=1}^n (Y_i - y)(X_i - \bar{X_n})}{\sum_{i=1}^n (Y_i - y)^2}$$

Lien avec la régression linéaire :

Il existe également une relation intéressante entre les variables de contrôle et l'analyse de régression. En effet, le calcul de l'estimateur optimal δ^*_{VC} revient à effectuer une régression de X sur la variable de contrôle Y.
Considérons le modèle suivant :

$$X = aY + b + \epsilon$$

où a est le coefficient de régression, b est l'intercept, et $\epsilon \sim \mathcal{N}(0, \sigma^2)$ capture l'erreur aléatoire.
Les estimateurs des moindres carrés de a et b sont donnés par :

$$\hat{a} = \frac{\widehat{Cov(X,Y)}}{\widehat{Var(Y)}} = \frac{\sum_{i=1}^{n}(Y_i - y)(X_i - \bar{X_n})}{\sum_{i=1}^{n}(Y_i - y)^2} \qquad \hat{b} = \hat{X_n} - \hat{a}\hat{Y_n}$$

Le lien avec la méthode de (VC) réside donc dans la relation : $c^* = \hat{a}$. En calculant par la suite $\hat{b} + \hat{a}y$, on trouve :

$$\begin{aligned} \hat{b} + \hat{a}y &= \hat{X_n} - \hat{a}\hat{Y_n} + \hat{a}y \\ &= \hat{X_n} - \hat{a}(\hat{Y_n} - y) \\ &= \hat{X_n} - c^*(\hat{Y_n} - y) \\ &= \delta^*_{VC} \end{aligned}$$

Le modèle linéaire évalué en $y = E[Y]$ nous permet donc d'obtenir l'estimation par variable de contrôle optimale.

Exemple numérique (estimation par VC)

On considère une variable aléatoire $U \sim \mathcal{U}[0,1]$ et on cherche à estimer la quantité :

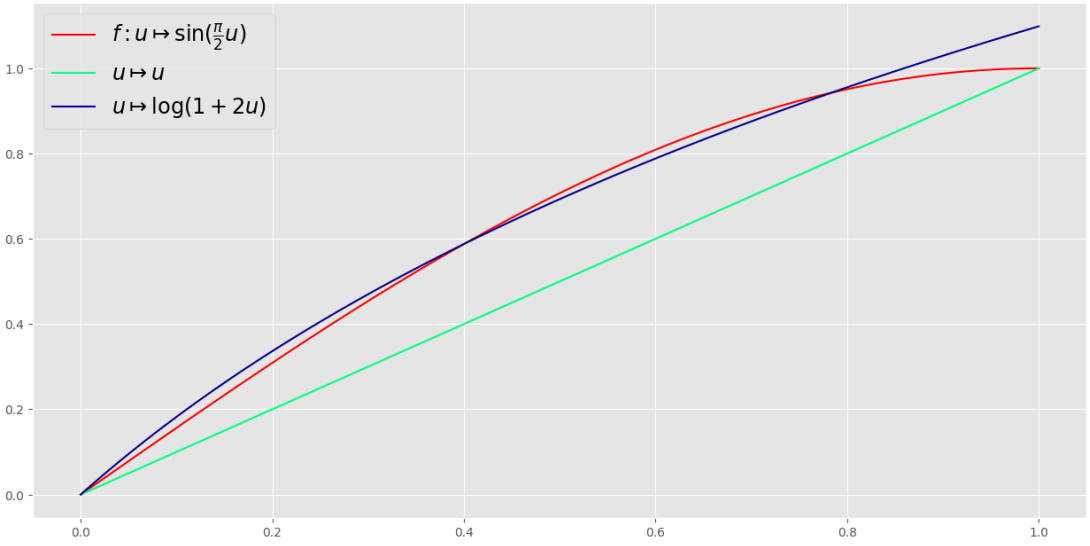
$$I = E\left[sin\left(\frac{\pi}{2}U\right)\right]$$

La valeur exacte vaut 0.6366198 (obtenue par la commande *integrate* sur R).
Pour $u \in [0,1]$, notons $f : u \mapsto sin(\frac{\pi}{2}u)$ la fonction associée à la quantité d'intérêt. On propose l'approximation linéaire $u \mapsto u$ et la deuxième approximation moins évidente $u \mapsto ln(1+2u)$. Les trois fonctions sont bien définies sur [0,1] et prennent les mêmes valeurs sur les bords de l'intervalle.
On peut ainsi proposer 2 variables de contrôle, dont on comparera les performances :

$$Y_1 = U \qquad Y_2 = ln(1 + 2U)$$

Leurs espérances sont bien connues, et définies par :

$$E[Y_1] = \frac{1}{2} = y_1 \qquad E[Y_2] = 0.6479184 = y_2$$



On obtient donc deux représentations de I :

$$I = E\left[sin\left(\frac{\pi}{2}U\right) - (U - y_1)\right] = E\left[sin\left(\frac{\pi}{2}U\right) - (ln(1+2U) - y_2)\right]$$

Nous pouvons ainsi définir deux estimateurs associés à la méthode de (VC) :

$$\delta^1_{VC} = \frac{1}{n}\sum_{i=1}^{n}\left[sin\left(\frac{\pi}{2}U_i\right) - \left(U_i - \frac{1}{2}\right)\right] \quad \delta^2_{VC} = \frac{1}{n}\sum_{i=1}^{n}\left[sin\left(\frac{\pi}{2}U_i\right) - (\ln(1+2U_i) - 0.65)\right]$$

où $(U_1, ..., U_n)$ sont i.i.d de même loi que U.
Remarque : On a choisi $c = 1$ comme constante par défaut.

Comparons désormais la variance des estimateurs δ^1_{VC} et δ^2_{VC} avec celle de l'estimateur classique :

$$\delta_{MCC} = \frac{1}{n}\sum_{i=1}^{n} sin\left(\frac{\pi}{2}U_i\right)$$

Pour $n = 10.000$, les résultats obtenus sous R sont donnés par le tableau suivant :

	δ_{MCC}	δ^1_{VC}	δ^2_{VC}
Moyenne	0.6356	0.6354	0.6361
Variance	9.57×10^{-2}	4.14×10^{-3}	6.88×10^{-4}

L'estimateur δ^2_{VC} est celui qui approche le mieux la valeur exacte de I, avec notamment la variance la plus faible. Ceci était logiquement attendu en visualisant les fonctions associées aux variables de contrôle. En effet, la fonction $u \mapsto ln(1+2u)$ approche très bien la fonction d'intérêt $f : u \mapsto sin(\frac{\pi}{2}u)$.
Ceci induit une forte corrélation entre les variables aléatoires $sin(\frac{\pi}{2}U)$ et $ln(1+2U)$, et par conséquent une variance bien plus faible via l'estimateur δ^2_{VC}.
Concernant l'estimateur δ^1_{VC}, il reste néanmoins plus précis que l'estimateur de (MCC), avec une variance bien inférieure. Ceci malgré le fait que la fonction linéaire $u \mapsto u$ approche moyennement la fonction f.

2.3 Stratification

On considère $\mathcal{X}$ l'ensemble des valeurs prises par une variable aléatoire X, et $(\mathcal{X}_1, ..., \mathcal{X}_N)$ une partition en N groupes (appelés **strates**) de $\mathcal{X}$:

$$\mathbb{P}[X \in \bigcup_{i=1}^{N} \mathcal{X}_i] = 1 \qquad avec \quad \mathcal{X}_i \bigcap \mathcal{X}_j = \emptyset, \ i \neq j$$

La méthode de stratification consiste à former des groupes homogènes, puis combiner l'information contenue dans chacune des strates en y effectuant des échantillonnages indépendants.
En notant $p_i = \mathbb{P}(X \in \mathcal{X}_i)$, on peut formuler notre quantité d'intérêt comme suit :

$$I = E[h(X)] = \sum_{i=1}^{N} E[h(X)|X \in \mathcal{X}_i]\mathbb{P}(X \in \mathcal{X}_i) = \sum_{i=1}^{N} p_i E[h(X)|X \in \mathcal{X}_i]$$

Cette formulation nécessite de savoir simuler selon les lois conditionnelles. Dans ce cas précis, on peut approcher chacune des moyennes conditionnelles et obtenir l'estimateur par stratification :

$$\boxed{\delta_{ST} = \sum_{i=1}^{N} p_i \left[\frac{1}{n_i} \sum_{j=1}^{n_i} h\left(X_j^{(i)}\right) \right]} \tag{2.3}$$

où, pour tout $i \in \{1, \ldots, N\}$, $(X_1^{(i)}, X_2^{(i)}, ..., X_{n_i}^{(i)})$ sont des tirages i.i.d. de loi $\mathcal{L}(X|X \in \mathcal{X}_i)$, $(n_i)_{1 \leq i \leq N}$ (appelés **allocations**) sont les choix du nombre de tirages pour cette loi, et $n_1 + ... + n_N = n$ (nombre total de simulations).

Biais de l'estimateur : Pour tout $i \in \{1, \ldots, N\}$, on a :

$$\begin{aligned} E\left[\frac{1}{n_i} \sum_{j=1}^{n_i} h(X_j^{(i)}) \right] &= \frac{1}{n_i} n_i E[h(X)|X \in \mathcal{X}_i]\mathbf{1}_{\{n_i>0\}} \\ &= E[h(X)|X \in \mathcal{X}_i]\mathbf{1}_{\{n_i>0\}} \end{aligned}$$

On en déduit ainsi :

$$\begin{aligned} E[\delta_{ST}] &= \sum_{i=1}^{N} p_i E[h(X)|X \in \mathcal{X}_i]\mathbf{1}_{\{n_i>0\}} \\ &= E[h(X)]\mathbf{1}_{\{n_i>0\}} \end{aligned}$$

L'estimateur par stratification est sans biais, sous réserve que $n_i > 0$ pour tout $i \in \{1, \ldots, N\}$, c'est-à-dire qu'au moins un tirage doit être effectué dans chacune des strates.

Convergence de l'estimateur : Pour $i \in \{1, \dots, N\}$ tel que $n_i > 0$, la loi forte des grands nombres appliquée à la suite de variables aléatoires i.i.d. $(X_j^{(i)})_{1 \leq j \leq n_i}$ donne :

$$\frac{1}{n_i} \sum_{j=1}^{n_i} h\left(X_j^{(i)}\right) \xrightarrow[n_i \to \infty]{p.s.} E[h(X)|X \in \mathcal{X}_i]$$

$$\Rightarrow \delta_{ST} \xrightarrow[n \to \infty]{p.s.} \sum_{i=1}^{N} p_i E[h(X)|X \in \mathcal{X}_i] = I$$

L'estimateur par stratification est convergent vers I, sous réserve de la même hypothèse.

Variance de l'estimateur : Les hypothèses d'indépendance et de distributions identiques donnent :

$$\begin{aligned} Var(\delta_{ST}) &= \sum_{i=1}^{N} \frac{{p_i}^2}{{n_i}^2} Var\left(\sum_{j=1}^{n_i} h\left(X_j^{(i)}\right)\right) \\ &= \sum_{i=1}^{N} \frac{{p_i}^2}{{n_i}^2} n_i Var(h(X)|X \in \mathcal{X}_i) \\ &= \sum_{i=1}^{N} \frac{{p_i}^2}{n_i} \sigma_i^2 \end{aligned}$$

où $\sigma_i^2 = Var(h(X)|X \in \mathcal{X}_i) = E[h^2(X)|X \in \mathcal{X}_i] - E[h(X)|X \in \mathcal{X}_i]^2$
Cette quantité peut être approchée par l'estimateur naturel :

$${\tilde{\sigma}_i}^2 = \frac{1}{n_i} \sum_{j=1}^{n_i} h^2\left(X_j^{(i)}\right) - \left(\frac{1}{n_i} \sum_{j=1}^{n_i} h\left(X_j^{(i)}\right)\right)^2$$

On obtient ainsi la variance approchée :

$$Var(\delta_{ST}) = \sum_{i=1}^{N} \frac{{p_i}^2}{n_i} {\tilde{\sigma}_i}^2$$

La variance dépend donc notamment de deux critères : le choix des strates $\mathcal{X}_i$ et le choix des allocations $(n_1, \dots, n_N)$.
Minimiser la variance dans le cas de la stratification revient donc à choisir judicieusement les allocations. Dans ce contexte, un premier résultat nous permet d'assurer une réduction de la variance par rapport à l'estimateur de MCC : l'allocation proportionnelle.

Proposition (Allocation $(n_i)_{1\leq i\leq N}$ proportionnelle)
Pour chaque strate $\mathcal{X}_i, i = 1, \ldots, N$, considérons l'allocation proportionnelle à la probabilité d'appartenance à la strate $n_i = np_i$.
Dans ce cas, on obtient un estimateur stratifié δ_{ST}^p de variance inférieure à l'estimateur classique :

$$Var(\delta_{ST}^p) = \frac{1}{n}\sum_{i=1}^{N} p_i\sigma_i^2 \leq Var(\delta_{MCC})$$

Un second résultat nous permet d'obtenir une variance minimale, en optimisant le choix des allocations, comme l'énonce la proposition suivante :

Proposition (Allocation $(n_i)_{1\leq i\leq N}$ optimale)
L'estimateur de variance minimale δ_{ST}^* est construit à partir du choix d'allocation

$$(n_1^*, \ldots, n_N^*) = \left(n\frac{p_1\sigma_1}{\sum_{i=1}^N p_i\sigma_i}, \ldots, n\frac{p_N\sigma_N}{\sum_{i=1}^N p_i\sigma_i}\right)$$

et admet comme variance :

$$Var(\delta_{ST}^*) = \frac{1}{n}\left(\sum_{i=1}^{N} p_i\sigma_i\right)^2 \leq Var(\delta_{ST}^p)$$

Ce dernier dépend des termes σ_i qui sont inconnus, mais qui peuvent être approchés par les $\tilde{\sigma}_i$ vus précédemment. Ainsi, pour tirer profit de ce dernier théorème, nous pouvons dans un premier temps effectuer une simulation d'estimation, puis dans un second temps poser l'allocation quasi optimale suivante :

$$(n_1^*, \ldots, n_N^*) = \left(n\frac{p_1\tilde{\sigma_1}}{\sum_{i=1}^N p_i\tilde{\sigma_i}}, \ldots, n\frac{p_N\tilde{\sigma_N}}{\sum_{i=1}^N p_i\tilde{\sigma_i}}\right)$$

Exemple numérique (estimation par ST)
On considère une variable aléatoire $U \sim \mathcal{U}[0, 1]$ et on cherche à estimer la quantité :

$$I = E[U] = \int_0^1 u\, du$$

On choisit de définir 5 strates, correspondant à des intervalles de même longueur :

$$\mathcal{X}_1 = [0, 0.2];\ \ \mathcal{X}_2 = [0.2, 0.4];\ \ \mathcal{X}_3 = [0.4, 0.6];\ \ \mathcal{X}_4 = [0.6, 0.8];\ \ \mathcal{X}_5 = [0.8, 1]$$

Les probabilités p_i associées sont toutes égales :

$$p_i = \mathbb{P}(U \in \mathcal{X}_i) = \frac{1}{5} \qquad i = 1, \ldots, 5$$

On effectue $n = 10.000$ simulations avec une allocation proportionnelle :

$$n_i = np_i = \frac{10000}{5} = 2000 \qquad i = 1, \ldots, 5$$

Les estimateurs de MCC et de stratification obtenus sont donc :

$$\delta_{MCC} = \frac{1}{n}\left(\sum_{i=1}^{5}\sum_{j=1}^{2000} U_j^{(i)}\right)$$

$$\delta_{ST} = \sum_{i=1}^{5} p_i \left(\frac{1}{n_i}\sum_{j=1}^{n_i} U_j^{(i)}\right) = \sum_{i=1}^{5}\frac{1}{5}\left(\frac{1}{2000}\sum_{j=1}^{2000} U_j^{(i)}\right)$$

Les variances de ces deux estimateurs sont formulées comme :

$$Var(\delta_{MCC}) = \frac{1}{n}\left(E[U^2] - I^2\right) = \frac{1}{n}\left(1 - (1/2)^2\right)$$

$$Var(\delta_{ST}) = \frac{1}{n}\sum_{i=1}^{5} p_i {\tilde{\sigma}_i}^2 = \frac{1}{n}\sum_{i=1}^{5}\frac{1}{5}\left(\frac{1}{2000}\sum_{j=1}^{2000}\left(U_j^{(i)}\right)^2 - \left(\frac{1}{2000}\sum_{j=1}^{2000} U_j^{(i)}\right)^2\right)$$

L'application numérique de ces formules sous R nous donne :

Cas $n = 10000$

	δ_{MCC}	δ_{ST}^p
Moyenne	0.500168	0.5003644
Variance	8.333333-06	3.300696e-07

Cas $n = 100000$

	δ_{MCC}	δ_{ST}^p
Moyenne	0.4996247	0.4999157
Variance	8.333333e-07	3.339183e-08

L'estimateur (ST) fournit une meilleure approximation que celui de (MCC) en augmentant le nombre de simulations. En termes de variance, l'avantage est nettement en faveur de la stratification.
Utilisons désormais l'allocation optimale pour le choix des n_i. La variance de δ_{ST}^* est alors donnée par :

$$Var(\delta_{ST}^*) = \frac{1}{n}\left(\sum_{i=1}^{5}\frac{1}{5}\sqrt{\frac{1}{2000}\sum_{j=1}^{2000}\left(U_j^{(i)}\right)^2 - \left(\frac{1}{2000}\sum_{j=1}^{2000} U_j^{(i)}\right)^2}\right)^2$$

L'application numérique fournit alors :

Cas $n = 10.000$

	δ_{ST}^*
Variance	3.300587e-07

Cas $n = 100.000$

	δ_{ST}^*
Variance	3.339155e-08

L'allocation optimale fournit logiquement une variance encore plus faible que l'allocation proportionnelle, mais l'écart est très faible, voire négligeable.

3 Application à un domaine concret

Dans cet exemple, notre objectif est de mettre en pratique l'échantillonnage préférentiel pour estimer un phénomène extrême dans le contexte du sport. Le code associé se trouve en annexes à la section 4. Le jeu de données figure dans la même section et peut être téléchargé via le lien suivant : **Consulter les données CSV**

L'objet de notre étude concerne le LOSC (équipe de football de Lille) et on s'intéresse à la question : **Quelle est la probabilité que le LOSC gagne un match par au moins 8 buts d'écarts en championnat ?**
À noter que cet exploit n'est plus arrivé pour le club depuis 2005. Afin de résoudre ce problème, nous recueillons un échantillon $(X_1, \ldots, X_n)$ de $n = 746$ observations basé sur les 20 dernières saisons du LOSC en championnat. Le tableau des données, construit à partir des informations disponibles sur le site officiel du championnat `https://ligue1.fr`, se trouve en annexes dans la section 4.
Les observations X_i de l'échantillon sont calculées par la différence entre le nombre de buts marqués et le nombre de buts concédés par match.

$$X_i \begin{cases} > 0 & \text{en cas de victoire} \\ = 0 & \text{en cas de de match nul} \\ < 0 & \text{en cas de défaite} \end{cases}$$

Les principales données observées sont les suivantes :

$$\mu = 0.42 \qquad \sigma = 1.63 \qquad \min(X_i) = -6 \qquad max(X_i) = 8$$

Notre évènement peut s'écrire sous la forme :

$$I = \mathbb{P}(X \geq 8) = E_f[\mathbf{1}_{X \geq 8}]$$

avec X de densité f (inconnue) donnée par la loi des observations. En construisant l'histogramme des observations, nous obtenons la figure :

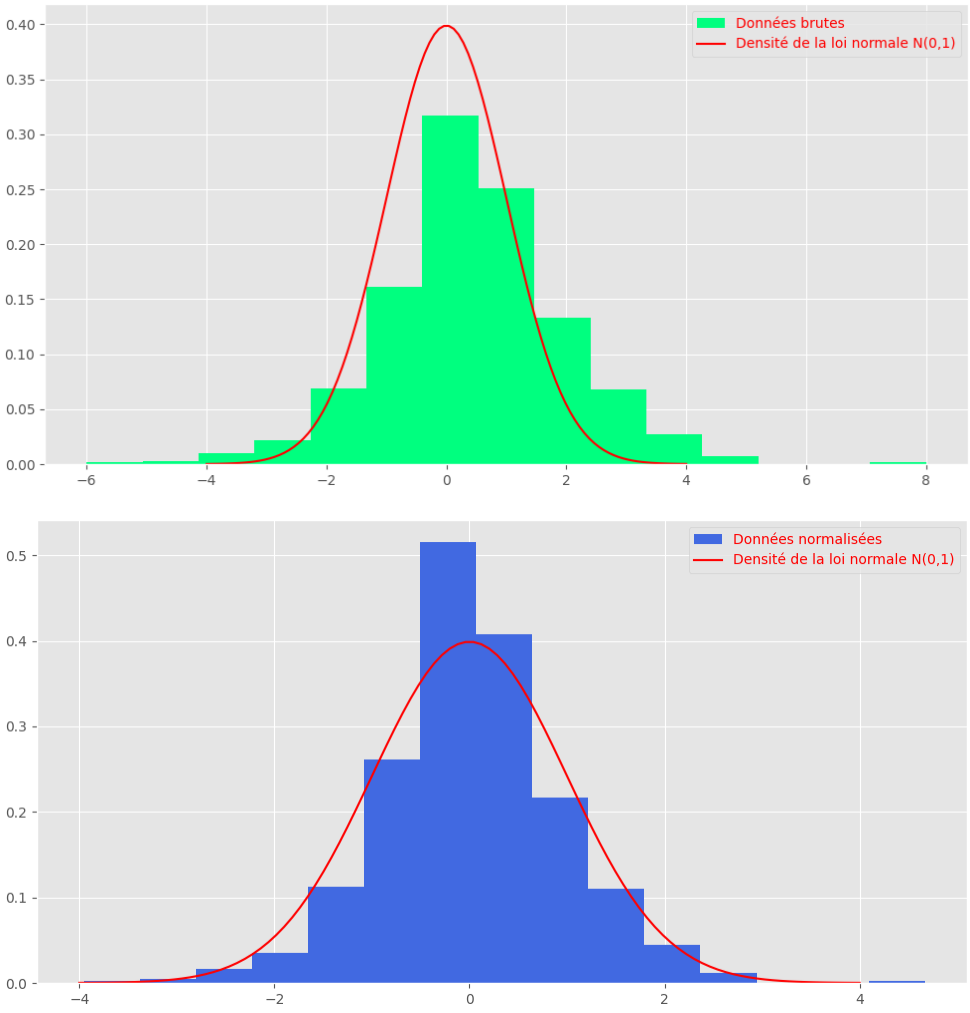
On observe une distribution légèrement décentrée de la loi normale $\mathcal{N}(0, 1)$. La moyenne et l'écart-type de notre échantillon sont donnés par : $\mu = 0.42$ et $\sigma = 1.63$. Ainsi, il est utile de normaliser les données afin d'obtenir :

$$N = \frac{X - \mu}{\sigma} = \frac{X - 0.42}{1.63}$$

Les données normalisées $(N_i)_{1 \leq i \leq n}$ sont alors représentées comme suit :

Cette distribution est très proche d'une loi normale $\mathcal{N}(0, 1)$. Nous pouvons approcher la loi de N par cette dernière. Pour notre estimation, on va ainsi utiliser l'expression de N dont la loi est connue :

$$I = \mathbb{P}(X \geq 8) = \mathbb{P}\left(N \geq \frac{8 - \mu}{\sigma}\right) = E_g[\mathbf{1}_{N \geq 4.66}]$$



avec g la densité de N, qui sera la loi cible de notre étude (en l'occurrence celle de la loi normale centrée réduite).
Concernant la loi instrumentale, nous proposons une loi de Student à 2 degrés de liberté, donnée par la densité :

$$\tilde{g}(x) = \frac{1}{\sqrt{2\pi}}\Gamma\left(\frac{3}{2}\right)\left(1+\frac{x^2}{2}\right)^{-\frac{3}{2}}$$

Cette densité a bien son support qui contient celui de g, approche bien g et accorde plus de poids sur les queues de distribution (dans notre cas, c'est celle de droite qui nous intéresse). C'est donc un bon candidat pour notre méthode.
Pour $(Y_i)_{1\leq i\leq n}$ une suite de n variables aléatoires i.i.d. de loi $\tilde{g}$, on obtient l'estimateur par échantillonnage préférentiel :

$$\delta_{EP} = \frac{1}{n}\sum_{i=1}^{n}\frac{g(Y_i)}{\tilde{g}(Y_i)}h(Y_i) = \frac{1}{n}\sum_{i=1}^{n}\frac{e^{-\frac{1}{2}Y_i^2}}{\Gamma(\frac{3}{2})}\left(1+\frac{Y_i^2}{2}\right)^{\frac{3}{2}}\mathbf{1}_{\{Y_i\geq 4.66\}} = 1.655109\times 10^{-6}$$

Notons que l'estimateur de Monte Carlo classique donne : $\delta_{MCC} = 1.340483\times 10^{-3}$ (une seule valeur de l'échantillon est supérieure ou égale à 8).

La valeur exacte de l'événement peut être approchée à l'aide de la commande *"pnorm"* sur R et donne : 1.55528×10^{-6}. Comme attendu, l'estimateur (EP) s'y rapproche largement plus que l'approche classique.
Comparons désormais la variance des deux estimateurs :

- $Var(\delta_{MCC}) = \frac{1}{n} Var\left(\mathbf{1}_{N \geq 4.66}\right) = 1.338686 \times 10^{-3}$
- $Var(\delta_{EP}) = \frac{1}{n} Var\left(\mathbf{1}_{Y \geq 4.66} \frac{g(Y)}{\tilde{g}(Y)}\right) = 9.695668 \times 10^{-10}$

L'estimateur par échantillonnage préférentiel réduit la variance d'un facteur $\approx 10^6$ par rapport à la méthode de Monte Carlo classique, démontrant ainsi l'intérêt de l'approche pour des événements rares.
Il est également intéressant de calculer la taille effective de l'échantillon afin d'évaluer la qualité du choix de $\tilde{g}$. L'application de sa formule donne :

$$ESS = \frac{(\sum_{i=1}^{n} w_i)^2}{\sum_{i=1}^{n} w_i^2} = \frac{\left(\sum_{i=1}^{n} \frac{e^{-\frac{1}{2}Y_i^2}}{\Gamma\left(\frac{3}{2}\right)} \left(1 + \frac{Y_i^2}{2}\right)^{\frac{3}{2}}\right)^2}{\sum_{i=1}^{n} \frac{e^{-Y_i^2}}{\Gamma(\frac{3}{2})^2} \left(1 + \frac{Y_i^2}{2}\right)^3} = 661.3292$$

Cette valeur se rapproche bien de notre taille d'échantillon $n = 746$. Par conséquent, cela confirme le choix judicieux de la loi de Student $t(2)$. De plus, le résultat est en adéquation avec la faible variance trouvée pour l'estimateur par (EP).

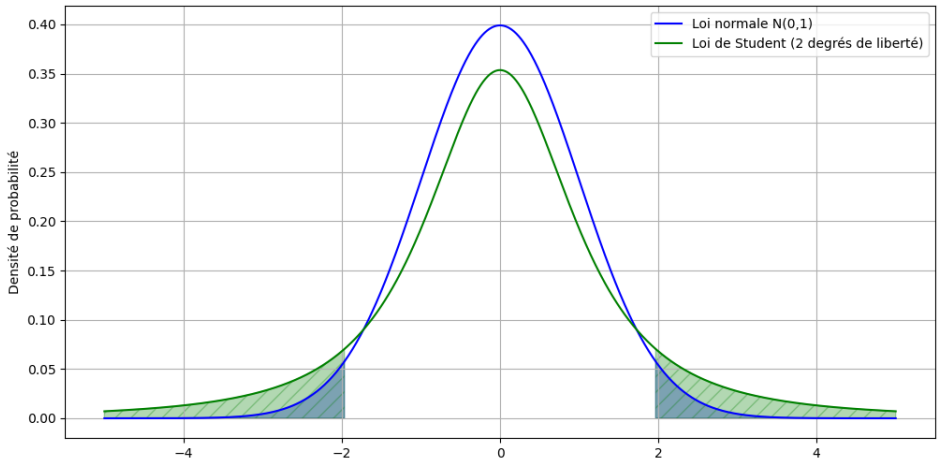


FIGURE 1 – Comparaison des queues de distribution entre la loi cible et la loi instrumentale

4 Annexes

Code R (Exemple numérique : estimation par VC)

```r
set.seed(1)

f <- function(u) {
  return(sin(0.5*pi*u))
}

monte_carlo <- function(sample, n) {
  moyenne <- mean(sample)
  variance <- ((n-1)/n) * var(sample)
  return(c(moyenne = moyenne, variance = variance))
}

n <- 10^4

sample_U <- runif(n)

sample_MCC <- f(sample_U)
sample_VC1 <- f(sample_U) - (sample_U - 1/2)
sample_VC2 <- f(sample_U) - (log(1 + 2*sample_U) - 0.6479184)

tableau <- data.frame(
  MCC = monte_carlo(sample_MCC, n),
  VC1 = monte_carlo(sample_VC1, n),
  VC2 = monte_carlo(sample_VC2, n)
)

rownames(tableau) <- c("Moyenne", "Variance")
tableau
```

Code R (Exemple numérique : estimation par ST)

```
set.seed(1)

n <- 10^4
nb_strates <- 5
valeurs_strates <- seq(0, 1, length.out = nb_strates + 1)

ech_unif <- runif(n)

f <- function(i) {
  return( mean(ech_unif[ech_unif >= valeurs_strates[i]
           & ech_unif < valeurs_strates[i + 1]]) )
}
g <- function(i) {
  return( mean(ech_unif[ech_unif >= valeurs_strates[i]
         & ech_unif < valeurs_strates[i + 1]]^2) )
}
mean_strate <- sapply(1:nb_strates, function(i) f(i))
mean_sq_strate <- sapply(1:nb_strates, function(i) g(i))

estimateur_MCC <- sum(ech_unif) / n
estimateur_ST <- sum((1/5) * mean_strate)

variance_MCC <- (integrate(function(u) u^2 *
        dunif(u, 0, 1), -Inf, Inf)$value - (1/2)^2)/ n
variance_ST <- sum((1/5) * (mean_sq_strate - mean_strate^2)) / n
variance_ST_minimale <- sum((1/5) * sqrt(mean_sq_strate -
   mean_strate^2))^2/ n

cat("Estimateur par MCC :", estimateur_MCC)
cat("Estimateur par ST :", estimateur_ST)

cat("Variance par MCC:", variance_MCC)
cat("Variance par ST :", variance_ST)
cat("Variance minimale par ST :", variance_ST_minimale)
```

Code R (Application du LOSC)

```r
set.seed(4)

data <- read.csv2("nomrepertoire/Donnees_LOSC.csv", header=F,
       stringsAsFactors = FALSE, na.strings = c(""))
X <- as.vector(as.matrix(data))
X <- X[!is.na(X)]

n <- length(X)
mu <- mean(X)
sigma <- sd(X)
N <- (X-mu) / sigma

valeur_exacte <- 1-(pnorm(4.663381, 0, 1))

MCC <- mean(ifelse(X >= 8, 1, 0))

Y <- rt(n, 2)
w <- dnorm(Y) / dt(Y,2)
EP <- mean(w * ifelse(Y >= 4.663381, 1, 0))

ESS <- function(n) {
  sum_1 <- 0
  sum_2 <- 0
  for (i in 1:n) {
    sum_1 <- sum_1 + w[i]
    sum_2 <- sum_2 + w[i]^2
  }
  return(sum_1^2 / sum_2)
}

cat("Estimation par MCC:", MCC)
cat("Estimation par EP:", EP)
cat("Taille effective ESS:", ESS(n))

var_MCC <- ((n-1)/n) * var(ifelse(N >= 4.663381, 1, 0))
var_EP <- ((n-1)/n) * var(ifelse(Y >= 4.663381, 1, 0) * w)
cat( "Variance par MCC:", var_MCC )
cat( "Variance par EP:", var_EP )
```

	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022	2023	2024
J1	2	1	1	0	0	-1	0	0	1	1	0	-1	-1	3	2	1	0	0	3	0
J2	-3	0	4	1	-2	-1	0	-1	0	-1	1	0	1	-3	0	-1	1	-4	0	2
J3	0	-1	-1	0	-1	0	0	1	0	1	2	0	0	-2	3	3	1	0	-6	-3
J4	-1	0	3	0	1	-3	0	1	-1	0	0	1	-3	0	-1	-2	0	1	2	1
J5	1	0	0	0	0	1	3	2	0	-2	2	0	-1	0	1	1	2	-1	-1	0
J6	2	2	-2	-1	1	0	1	0	0	2	0	0	-1	-1	1	0	3	-1	2	-1
J7	2	4	-3	0	2	0	0	0	-2	3	-1	-1	-2	-4	-1	2	4	1	-1	2
J8	4	-1	1	0	1	1	2	0	2	0	1	-1	1	-3	3	0	0	1	1	0
J9	1	2	0	0	0	0	-2	2	0	3	-3	2	-1	0	2	0	0	2	-1	1
J10	1	-1	1	3	0	-1	-2	2	1	1	-1	0	1	-1	1	-1	-1	-1	1	2
J11	-1	0	0	-1	0	4	0	2	2	1	-2	-1	-1	-1	1	3	4	0	3	0
J12	1	2	3	-3	2	2	2	0	1	2	0	0	-1	3	-1	-1	0	-1	1	0
J13	2	0	1	1	-1	2	3	0	-2	0	-2	0	-1	2	0	0	1	0	-1	2
J14	0	2	0	0	3	-2	1	0	-1	1	1	0	0	-3	-2	-2	1	0	0	2
J15	0	-1	0	-2	0	4	0	2	0	1	-1	-2	2	1	0	1	2	0	1	0
J16	0	1	2	0	0	1	3	1	0	1	0	1	1	1	1	1	0	1	2	0
J17	0	-1	0	-1	0	4	1	2	2	-1	0	1	1	-2	0	1	1	1	0	-1
J18	1	0	-2	3	1	3	3	0	0	1	3	3	-2	-3	1	1	-1	0	0	3
J19	0	2	0	1	2	4	0	0	3	0	-1	0	0	0	-1	-4	1	1	4	0
J20	-1	2	0	0	0	2	2	-2	0	-1	1	0	0	1	2	-1	1	2	-1	4
J21	2	-1	1	0	1	-1	1	3	-2	-2	-1	-1	0	-1	1	-2	1	0	0	-2
J22	1	3	1	2	0	1	0	1	-1	0	-1	-2	1	-1	1	1	1	-2	2	3
J23	0	0	0	0	1	0	2	-1	0	-1	0	0	-1	1	4	1	3	-4	2	-2
J24	0	2	-1	0	-2	2	-1	1	2	2	1	1	-1	-3	2	2	2	1	-1	1
J25	0	-2	-1	2	1	1	0	0	2	0	0	0	-1	0	0	-1	0	0	1	0
J26	0	0	-1	0	-2	-1	1	0	2	0	-1	0	1	0	0	3	3	1	0	0
J27	0	2	4	-1	2	0	1	-1	1	1	1	1	-1	-1	1	1	0	4	0	1
J28	-2	1	0	0	1	1	1	4	2	2	1	-3	0	-1	1	1	2	0	2	2
J29	0	4	-2	5	1	-2	2	3	-1	0	3	2	1	0	-1		0	1	2	-1
J30	0	0	-2	1	2	3	-1	1	1	0	-2	1	0	-1	1		-1	0	-1	1
J31	8	0	-3	-2	0	-1	0	-2	5	1	2	1	1	-1	0		1	0	1	1
J32	-2	0	2	0	-1	0	0	3	0	1	1	3	-1	-1	4		2	-1	0	-1
J33	1	0	-2	1	-1	4	5	2	1	1	2	3	-2	0	0		0	-1	3	1
J34	1	0	-1	2	-2	1	1	1	0	0	-5	2	3	-4	5		1	1	-1	0
J35	1	1	-1	1	0	2	1	1	-2	1	2	0	3	2	0		2	-3	0	
J36	2	-2	1	0	0	2	1	3	3	0	-1	1	-2	1	1		3	-1	1	
J37	-2	4	-1	1	1	1	0	-1	0	-2	-4	0	-4	1	5		0	2	1	
J38	2	0	0	0	1	-1	1	3	0	3	3	1	3	-5	-2		1	0	0	

TABLE 1 – Données de l'échantillon d'observation (Application LOSC)
Lien téléchargement

5 Bibliographie

S. Rasmussen & P. W. Glynn, *Stochastic Simulation : Algorithms and Analysis*, Springer, 2009.
- Chapitre 5 : "Variance-Reduction Methods"
- Chapitre 6 : "Rare-Event Simulation"

Christian P. Robert & George Casella, *Monte Carlo Statistical Methods*, Springer, 2004.

Rubinstein & Kroese, *Simulation and the Monte Carlo Method*, Wiley, 2016.

Pour la Science, n°385, édition française de *Scientific American*, novembre 2009.
- Chapitre : "Hasard et incertitude, les défis qu'ils posent".