SIMULATIONS

Abass Sagna

Chapitre 1

Simulations

1.1 Introduction

1.2 Simulation de variables aléatoires

Simuler un phénomène c'est l'imiter afin de faire paraître une situation réelle attendue lorsqu'on lui soumet à certaines contraintes. Dans le contexte des probabilités, simuler une réalisation d'une variable aléatoire c'est l'imiter afin de générer une valeur parmi l'ensemble des valeurs prises par la variable aléatoire.

La simulation d'une variable aléatoire quelconque X est en général basée sur la simulation d'une variable aléatoire uniforme (ou des variables aléatoires uniformes) qui prend ses valeurs dans $]0,1[$. Comme on l'a vu avec la loi uniforme sur $[0,1]$, la densité de la loi uniforme sur $]0,1[$ s'écrit

$$f(x) = \begin{cases} 1 & \text{si } x \in]0,1[\\ 0 & \text{si } x \notin]0,1[. \end{cases}$$

Nous avons plusieurs méthodes de simulation parmi lesquelles la *méthode d'inversion de la fonction de répartition* que nous énonçons dans la proposition ci-dessous. Ce résultat peut toujours être utilisé lorsque F^{-1} est calculable sans difficulté.

Proposition 1.2.1. *Soit U une v.a. de loi uniforme sur $]0,1[$ et soit X une v.a. de fonction de répartition F et de fonction quantile F^{-1} :*

$$F^{-1}(u) := \inf\{x \in \mathbb{R} : \quad F(x) \geq u\}, \qquad \textit{pour tout } \ u \in]0,1[.$$

Alors X et $F^{-1}(U)$ ont la même loi de probabilité : on note $X \overset{\mathcal{L}}{=} F^{-1}(U)$.

PREUVE. Soit $x \in \mathbb{R}$. Il suffit de démontrer que $F(x) = \mathbb{P}(X \leq x) = \mathbb{P}(F^{-1}(U) \leq x)$. On a

$$\begin{aligned} \mathbb{P}(F^{-1}(U) \leq x) &= \mathbb{P}(U \leq F(x)) && \text{(par croissance de } F) \\ &= F(x) && \text{(par calcul de la f.r. de } U) \end{aligned}$$

Donc les fonctions de répartition de $F^{-1}(U)$ et de X coincident. □

Pour simuler une réalisation de X il suffit donc de pouvoir simuler une réalisation de U et de pouvoir calculer F^{-1}. Cependant il arrive que l'on ne puisse pas calculer explicitement F^{-1} et parfois même la fonction de répartition F ne peut qu'être approcher numériquement. Dans ces situations on fait appel à d'autres méthodes de simulation parmi lesquelles la *méthode du rejet*, la *méthode de composition*, la *méthode de convolution*, celle *utilisant des propriétés spéciales de la variable aléatoire*, ...

1.2.1 Simulation d'une variable aléatoire discrète

Considérons une variable aléatoire X qui prend les valeurs -1, 0 et 1 avec les probabilités suivantes :

$$\mathbb{P}(X=-1)=1/3,\ \mathbb{P}(X=0)=1/6,\ \mathbb{P}(X=1)=1/2.$$

La simulation d'une réalisation de la variable aléatoire X est la mise en ouvre d'un procédé permettant de générer de façon aléatoire (disons pseudo-aléatoire) une des valeurs prises par la v.a. X, c'est-à-dire, -1, 0, ou 1 et en tenant compte des poids associés à chacune d'elles. Comme on a une chance sur trois que X prenne la valeur -1, si au lieu de simuler une réalisation de X, on en simule un très grand nombre, on s'attend intuitivement à ce que la proportion d'apparition de -1 soit proche de $1/3$. Du point de vue théorique, c'est une conséquence de ce qu'on appelle la *loi des grands nombres* qui sera étudiée au chapitre **??**.

La simulation d'un nombre aléatoire, et par conséquent d'une v.a. de loi uniforme sur $]0,1[$, se fait à travers des algorithmes numériques dont le plus utilisé est la méthode de congruence linéaire. Ces algorithmes génèrent des suites de nombres entiers qui peuvent être considérés, en utilisant les tests statistiques, comme indépendants et distribués de façon aléatoire. Ces nombres sont parfois dits pseudo-aléatoires.

Du point de vue pratique, nous supposerons que nous savons simuler une réalisation de la loi uniforme sur $]0,1[$ et, de façon plus générale, une suite $u_1,\dots,u_n$ de n réalisations indépendantes (les u_i n'ont aucun lien les uns les autres) de la loi uniforme sur $]0,1[$. Une telle suite de réalisations est appelée un *échantillon de taille n ou un n-échantillon* de la loi uniforme sur $]0,1[$.

En *langage C*, la génération de nombres pseudo-aléatoires se fait à travers la fonction prédéfinie *rand()* qui est incluse dans la bibliothèque <*stdlib.h*>. La fonction *rand()* génère des nombres pseudo-aléatoires entiers de rang maximal `RAND_MAX` qui est une macro prédéfinie dans <*stdlib.h*>. On simule donc une réalisation d'une loi uniforme sur $]0,1[$ en utilisant l'instruction

$$\texttt{1.0} * \texttt{rand}()/\texttt{RAND\_MAX};\tag{1.2.1}$$

Comme l'algorithme qui génère les nombres pseudo-aléatoires est un algorithme déterministe il dépend du point d'initialisation (`seed`). Avec un point initial donné l'algorithme génère la même séquence de nombres pseudo-aléatoires lorsqu'on fait appel à elle à plusieurs reprises. L'utilisateur peut choisir le point initial en utilisant la fonction *srand()* de la bibliothèque <*stdlib.h*>. Pour éviter l'interaction entre l'utilisateur et l'ordinateur et assurer en même temps que l'algorithme prenne des valeurs initiales différentes il est plus pratique d'utiliser le temps système de l'ordinateur grâce à la fonction *time()* qui se trouve dans la bibliothèque <*time.h*>. Ainsi, pour simuler un échantillon de taille n de la loi uniforme sur $[0,1]$ il suffit de faire appel

à la fonction *rand()* à n reprises comme suit (en incluant bien sûr les bibliothèques qu'il faut !) :

```
int i;
srand((unsigned)time(NULL));
for (i = 0; i < n; i++)
printf("%e \n", 1.0 * rand()/RAND_MAX);
```

Revenons-en à la simulation d'une réalisation de la v.a. X. Nous avons déjà vu que sa fonction de répartition est

$$F(x) = \begin{cases} 0 & \text{si } x < -1 \\ 1/3 & \text{si } x \in [-1, 0[\\ 1/2 & \text{si } x \in [0, 1[\\ 1 & \text{si } x \geq 1. \end{cases}$$

On en déduit que sa fonction quantile est donnée pour $u \in]0, 1[$ par

$$F^{-1}(u) = \begin{cases} -1 & \text{si } u \in]0, 1/3] \\ 0 & \text{si } u \in]1/3, 1/2] \\ 1 & \text{si } u \in]1/2, 1[. \end{cases}$$

Par conséquent pour simuler une réalisation x de X on simulera d'abord une réalisation u de la loi uniforme sur $]0, 1[$ et

1. si $u \in]0, 1/3]$ alors x prendra la valeur -1,
2. si $u \in]1/3, 1/2]$ alors x prendra la valeur 0,
3. si $u \in]1/2, 1[$ alors x prendra la valeur 1.

Simulation 1.2.1. Reprenons l'exemple précédent. Soit X une variable aléatoire à valeurs dans $\{-1, 0, 1\}$ avec

$$\mathbb{P}(X = -1) = 1/3,\ \mathbb{P}(X = 0) = 1/6,\ \mathbb{P}(X = 1) = 1/2.$$

1. Simuler un échantillon de taille $n = 40$ de la loi de X.
2. Faites appel à nouveau à votre fonction qui simule un échantillon de taille 40. Que constatez-vous par rapport à l'échantillon précédent ?
3. On pose $f_n(-1), f_n(0), f_n(1)$, les fréquences d'apparition de $-1, 0, 1$, respectivement, pour un échantillon de taille n. Déterminer les $f_n(i),\ i \in \{-1, 0, 1\}$, pour les échantillons simulés.
4. En faisant varier la taille de l'échantillon n de 1 à 100000 par pas de 1, représenter, sur la même fenêtre graphique, les trois graphiques suivants : pour $i \in \{-1, 0, 1\}$,

   $$n \in \{1, 2, \ldots, 100\} \longmapsto f_n(i)$$

   et les droites d'équations $y = 1/3$, $y = 1/6$, $y = 1/2$.
5. Commenter les graphiques obtenus.
6.

1	1	-1	-1	-1	1	0	1	1	-1
-1	1	-1	-1	-1	-1	1	1	1	1
1	0	1	1	-1	1	-1	1	1	-1
-1	-1	0	-1	1	1	1	-1	1	1

TABLE 1.1 – Echantillon de taille 40 de la loi de X.

1	1	1	-1	1	1	1	1	1	1
0	1	-1	1	-1	-1	1	-1	0	-1
-1	-1	1	0	0	0	-1	0	1	-1
1	1	0	-1	1	0	-1	0	1	1

TABLE 1.2 – Echantillon de taille 40 de la loi de X.

Réponse. 1. On a déjà vu comment simuler une réalisation de la v.a. X. Pour simuler un échantillon de taille $n = 40$ il suffit de faire appel n fois à la fonction qui simule une réalisation de X. Le résultat est donné dans le tableau suivant.

2. En relançant le programme on obtient le jeu de données suivant.

On constate que les deux échantillons sont différents. Remarquons qu'en l'absence de la fonction *srand()* on aurait eu le même échantillon : ce qu'on souhaite justement éviter ! Aussi, il faut noter que même si la suite des nombres générés n'est pas identique lorsqu'on simule différents échantillons de taille n, certains indicateurs doivent rester à peu près égaux lorsque n est grand. C'est le cas par exemple des fréquences d'apparition des valeurs prises par X qui doivent être proches des probabilités associées à chacune de ces valeurs (comme le montre d'ailleurs la Figure 1.1).

3. Pour le premier échantillon on a $f_{40}(-1) = 16/40 = 0.4$, $f_{40}(0) = 3/40 = 0.075$, $f_{40}(1) = 21/40 = 0.525$. Pour le second échantillon on a $f_{40}(-1) = 12/40 = 0.3$, $f_{40}(0) = 9/40 = 0.225$, $f_{40}(1) = 19/40 = 0.475$.

4. Le graphique demandé est affiché à la figure 1.1. On constate que les fréquences d'apparition des valeurs prises par la v.a. $X : -1, 0, 1$, deviennent de plus en plus proches des probabilités qui leur sont associées : c'est-à-dire, $1/3, 1/6, 1/2$, respectivement. Par ailleurs on n'observe pas encore une convergence. C'est-à-dire, une stabilisation des courbes de fréquences (empiriques) près des probabilités (ou fréquences théoriques) associées. Les dernières fréquences obtenues pour l'échantillon de taille $n = 100$ sont de $f_n(-1) = 0.30$, $f_n(0) = 0.16$, $f_n(1) = 0.54$. On peut aussi remarquer que $\mathbb{P}(X = 1) = 1/2$ s'estime mieux que les autres probabilités. Ceci vient du fait que (à peu près) la moitié des observations de notre échantillon est égale à 1 et naturellement, plus on a d'observations, mieux les fréquences s'approchent des probabilités à estimer.

Exercice 1.2.1. On se donne une variable aléatoire X à valeurs dans $\{-2, -1, 0, 1, 2\}$ avec

$$\mathbb{P}(X = -2) = \frac{1}{5},\ \mathbb{P}(X = -1) = \frac{1}{10},\ \mathbb{P}(X = 0) = \frac{2}{5},\ \mathbb{P}(X = 1) = \frac{1}{5},\ \mathbb{P}(X = 2) = \frac{1}{10}.$$

L'objectif de l'exercice est de simuler N réalisations indépendantes de la variable aléatoire X en utilisant le langage C, où le nombre de simulations N sera fixé. Par exemple, si `SimulLoi`

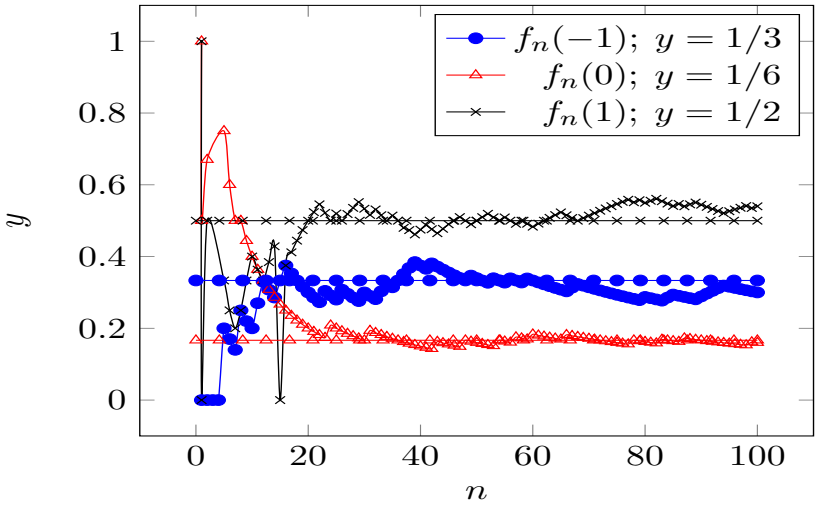

FIGURE 1.1 – En abscisse : taille n de l'échantillon. Ordonnée : Fréquences d'apparition de -1, de 0 et 1 : $f_n(-1)$, $f_n(0)$ et $f_n(1)$.

`Discrete(N, ...)` est la fonction qui simule N réalisations de la variable aléatoire X l'appelle à la fonction

`SimulLoiDiscrete(10, ...)`

génèrera de façon aléatoire 10 valeurs appartenant à l'ensemble $\{-2, -1, 0, 1, 2\}$ des valeurs prises par la variable aléatoire X, par exemple, la suite de valeurs

`-2 0 0 2 2 -2 0 0 2 -1.`

Posons $p_1 = 0.2,\ p_2 = 0.1,\ p_3 = 0.4,\ p_4 = 0.2$ et $p_5 = 0.1$ les probabilités que X prenne les valeurs $a_1 = -2, a_2 = -1, a_3 = 0, a_4 = 1, a_5 = 2$, respectivement.

1. Déterminer la fonction de répartition F de X
2. Calculer la fonction quantile F^{-1} de X.

Vous pouvez utiliser "l'entête" ci-dessous (à compléter éventuellement) pour votre programme C.

```
#define  N       10
#define  p1      0.2
#define  p2      0.1
#define  p3      0.4
#define  p4      0.2
#define  p5      0.1

#define  a1      -2
#define  a2      -1
#define  a3       0
#define  a4       1
```

```
#define  a5       2

#include <stdio.h>
#include <math.h>
#include <stdlib.h>
#include <time.h>
#include <ctype.h>
#include <string.h>
```

1. Ecrivez une fonction `SimulLoiDiscrete` telle que l'appelle à la fonction

   ```
   SimulLoiDiscrete(N, a, a2, a3, a4, a5, p1, p2, p3, p4,
                    p5, ValSim)
   ```

   ne renvoie rien mais remplit le tableau `ValSim` des N valeurs simulées de la variable aléatoire X. Le prototype de la fonction ressemblera a

   ```
   void SimulLoiDiscrete(int n, ...,double * valsim)
   ```

2. Appelez la fonction `SimulLoiDiscrete(N, ...)` pour $N = 10^4$, $N = 10^5$, $N = 10^6$ et pour chaque valeur de N calculer les fréquences d'apparition de chacune des valeurs du tableau `ValSim` et les comparer aux probabilités p_i, $i = 1, 2, 3, 4, 5$. Commenter.

Indications. Le tableau `ValSim` dans la fonction `main` pourra être défini comme suit

```
ValSim = (double *) malloc(N *sizeof(double));
```

1.2.2 Approximation de l'espérance d'une v.a. discrète

Soit X une variable aléatoire réelle. Le calcul de $\mathbb{E}(X)$ peut se faire de façon explicite comme dans le cas des exemples précédents. Dans beaucoup d'applications des mathématiques (à la physique, à la bilogie, à l'économie, à la finance, ...) on est amené à calculer l'espérance mathématique de variables aléatoires. Dans la plupart des cas, le calcul de ces espérances ne peut être mené de façon explicite. On fait alors recours à des méthodes d'estimation dont celle basée sur des simulations. Cette méthode d'estimation est connue sous le nom de la méthode de *Monte Carlo* que nous justifierons théoriquement au chapitre **??**. Pour montrer comment estimer de façon pratique une espérance reconsidérons l'exemple 1.2.1 où X est une variable aléatoire à valeurs dans $\{-1, 0, 1\}$ avec

$$\mathbb{P}(X = -1) = 1/3,\ \mathbb{P}(X = 0) = 1/6,\ \mathbb{P}(X = 1) = 1/2.$$

Ici, $\mathbb{E}(X)$ peut être calculé de façon explicite et vaut $\frac{1}{6} \approx 0.1667$ mais nous allons essayer d'approcher ce résultat à travers les simulations. Rappelons que

$$\mathbb{E}(X) = -1\times\mathbb{P}(X = -1) + 0\times\mathbb{P}(X = 0) + 1\times\mathbb{P}(X = 1).$$

Or nous avons vu à travers les simulations de l'exemple 1.2.1 que lorsque la taille de l'échantillon n est grande, les fréquences d'apparition de $-1, 0, 1$ de l'échantillon, que nous avons noté respectivement $f_n(-1), f_n(0)$ et $f_n(1)$, sont proches des probabilités associées aux valeurs respectives prises par la variable aléatoire X, c'est-à-dire, $f_n(-1) \approx \mathbb{P}(X = -1), f_n(0) \approx$

$\mathbb{P}(X=0), f_n(1) \approx \mathbb{P}(X=1)$. Par conséquent, on peut estimer l'espérance de X par (pour $n = 10000$ par exemple) :

$$\mathbb{E}(X) \approx -1\times f_n(-1) + 0\times f_n(0) + 1\times f_n(1).$$

De manière générale, lorsque X est une variable aléatoire à valeurs dans $E = \{x_1, x_2, \ldots\}$, alors

$$\mathbb{E}(X) = \sum_{i=1}^{+\infty} x_i \mathbb{P}(X = x_i)$$

peut être estimé par $M_n(X)$, défini par

$$M_n(X) = \sum_{i=1}^{+\infty} x_i f_n(x_i) \tag{1.2.2}$$

où les $f_n(x_i)$ sont les fréquences d'apparition des x_i pour un échantillon de taille n assez grand.

Remarque 1.2.1. En notant $X_1(\omega), \ldots, X_n(\omega)$ la suite des valeurs simulées pour l'échantillon de taille n, $n_i = \text{card}\big(\{k \in \{1,\ldots,n\} :\ X_k(\omega) = x_i\}\big)$ et $I = \{i \geq 1,\ n_i \neq 0\}$, l'équation (1.2.2) peut être réécrite comme

$$\begin{aligned} M_n(X) &= \sum_{i=1}^{+\infty} x_i \frac{\text{card}\big(\{k \in \{1,\ldots,n\} :\ X_k(\omega) = x_i\}\big)}{n} \\ &= \frac{1}{n}\sum_{i=1}^{+\infty} n_i x_i = \frac{1}{n}\sum_{i\in I} n_i x_i = \frac{1}{n}\sum_{k=1}^{n} X_k(\omega). \end{aligned}$$

Simulation 1.2.2. Reprenons l'exemple 1.2.1. Soit X une variable aléatoire à valeurs dans $\{-1, 0, 1\}$ avec

$$\mathbb{P}(X=-1) = 1/3,\ \mathbb{P}(X=0) = 1/6,\ \mathbb{P}(X=1) = 1/2.$$

1. A partir de l'échantillon de taille 100 simulé dans l'exemple 1.2.1, représentez sur le même graphique la fonction

   $$n \in \{1, 2, \ldots, 100\} \longmapsto M_n(X)$$

   et la droite d'équation $y = \frac{1}{6}$.
2. Que vaut $M_n(X)$ pour $n = 100$.
3. Déterminer $M_n(X)$ sur run échantillon de taille $n = 10000$. Que constate-t-on ?

Réponse. 1. Le graphique demandé est représenté à la figure 1.2. On constate que $M_n(X)$ ne s'est pas encore stabilisée pour $n = 1, 2, \ldots, 100$.

2. Pour $n = 100$ on a $M_n = 0.24$ (à comparer avec $\mathbb{E}(X) = 1/6 \approx 0.1667$). On a donc une précision de l'ordre de 10^{-1}. Pour augmenter la précision on peut augmenter la taille n de l'échantillon.

3. Pour $n = 10000$ on trouve $M_n(X) = 0.1618$. On constate que $M_{10000}(X)$ est plus proche de $\mathbb{E}(X)$ que $M_{100}(X)$.

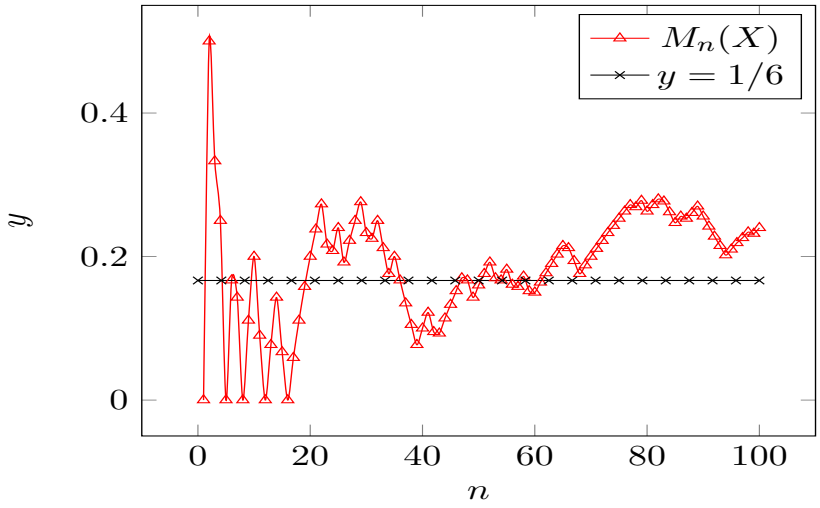

FIGURE 1.2 – En abscisse : taille n de l'échantillon. Ordonnée : $M_n(X)$ et la droite d'équation $y = 1/6$.

Remarque 1.2.2. Remarquons qu'en simulant à plusieurs reprises des échantillon de taille $n = 10000$ dans l'exemple 1.2.2, les valeurs trouvées pour $M_n(X)$ seront identiques lorsqu'elles sont arrondies à 10^{-1} près mais différentes lorsqu'on les arrondit à 10^{-3} près par exemple. Pour avoir l'égalité des $M_n(X)$ à 10^{-3} près, on doit augmenter la taille n de l'échantillon. Par conséquent, l'erreur d'estimation ou la précision de l'estimation, qui est mesurée par l'écart absolu entre $\mathbb{E}(X)$ et la quantité $M_n(X)$ utilisée pour l'estimer, soit $|\mathbb{E}(X) - M_n(X)|$, dépend de n. Plus n est grand, plus $M_n(X)$ s'approche de $\mathbb{E}(X)$ (comme le montre la figure 1.2). Aussi, pour atteindre une précision donnée (par exemple si on veut que l'écart absolu entre $\mathbb{E}(X)$ et $M_n(X)$ soit de l'ordre de 10^{-3}) pour savoir quel ordre de grandeur choisir pour n on s'appuie sur des approximations de la loi asymptotique de $M_n(X)$. Ce résultat est connu sous le nom de *Théorème de la limite centrale* et sera abordé au chapitre **??**. Le théorème de la limite centrale stipule en particulier que l'erreur d'estimation

$$|\mathbb{E}(X) - M_n(X)| = O\Big(\frac{1}{\sqrt{n}}\Big). \tag{1.2.3}$$

Exercice 1.2.2. On reconsidère l'exercice 1.2.1.

1. Calculez explicitement $\mathbb{E}(X)$.
2. Représentez sur la même fenêtre graphique la fonction
$$n \in \{1, 2, \ldots, 100000\} \longmapsto M_n(X)$$
et la droite d'équation $y = \mathbb{E}(X)$. Commenter les graphiques.
3. Représentez sur une nouvelle fenêtre graphique l'erreur d'estimation, c'est-à-dire, la fonction
$$n \in \{1, 2, \ldots, 100000\} \longmapsto |\mathbb{E}(X) - M_n(X)|$$
et la fonction
$$n \in \{1, 2, \ldots, 100000\} \longmapsto \frac{1}{\sqrt{n}}.$$

4. Comparer les deux graphiques de la question précédente.

1.2.3 Approximation de la variance d'une v.a. discrète

Pour voir comment estimer la variance d'une v.a. discrète à valeurs dans $\{x_1, x_2, \ldots\}$ reconsidèrons l'exemple **??** où X est une variable aléatoire qui prend les valeurs -1, 0 et 1 avec les probabilités :

$$\mathbb{P}(X=-1)=1/3,\ \mathbb{P}(X=0)=1/6,\ \mathbb{P}(X=1)=1/2.$$

Nous avons déjà vu que $\mathrm{Var}(X) = \frac{29}{36}$. Voyons maintenant comment l'estimer à travers la simulation. Rappelons que

$$\mathrm{Var}(X) = \mathbb{E}(X^2) - (\mathbb{E}(X))^2.$$

Comme nous savons déjà comment estimer $\mathbb{E}(X)$, il suffit donc de dire comment estimer $\mathbb{E}(X^2)$. Or, nous avons d'après le théorème de transfert

$$\mathbb{E}(X^2) = (-1)^2\times\mathbb{P}(X=-1) + 0\times\mathbb{P}(X=0) + 1\times\mathbb{P}(X=1).$$

On peut donc, pour échantillon de taille n, estimer $\mathbb{E}(X^2)$ par la quantité

$$M_n(X^2) := (-1)^2\times f_n(-1) + 0\times f_n(0) + 1\times f_n(1).$$

Par conséquent, pour un échantillon de taille n, la variance de X peut être estimée par

$$V_n(X) := M_n(X^2) - \big(M_n(X)\big)^2.$$

De façon plus générale, si X est une v.a. à valeurs dans $\{x_1, x_2, \ldots\}$ alors la variance de X peut être estimée par

$$V_n(X) = M_n(X^2) - \big(M_n(X)\big)^2,$$

où

$$M_n(X^2) = \sum_{i=1}^{\infty} x_i^2 f_n(x_i) \qquad \text{et} \quad M_n(X) = \sum_{i=1}^{+\infty} x_i f_n(x_i)$$

avec $f_n(x_i)$ qui désigne la fréquence d'apparition de x_i pour l'échantillon de taille n.

Remarque 1.2.3. En utilisant un raisonnement analogue à la remarque 1.2.1 on peut montrer que

$$M_n(X^2) = \frac{1}{n}\sum_{k=1}^{n} X_k^2(\omega),$$

où $X_1(\omega), \ldots, X_n(\omega)$ est la suite des valeurs simulées de taille n de la variable aléatoire X. D'autre part, $M_n(X) = \frac{1}{n}\sum_{k=1}^{n} X_k$ et on peut montrer après un calcul simple que

$$V_n(X) = M_n(X^2) - \big(M_n(X)\big)^2 = \frac{1}{n}\sum_{k=1}^{n}(X_k - M_n(X))^2.$$

Définition 1.2.1. *Pour une échantillon observé $X_1(\omega), \ldots, X_n(\omega)$ de la loi de X, les quantités $M_n(X)$ et $V_n(X)$ définies précédemment sont appelées respectivement la moyenne empirique et la variance empirique.*

Nous énonçons ci-après le principe général d'estimation de $\mathbb{E}(g(X))$.

Proposition 1.2.2. *Soit g une fonction à valeurs réelles et X une variable aléatoire discrète. Soit $X_1(\omega), \ldots, X_n(\omega)$ un échantillon observé de taille n de la loi de X. Alors la quantité*

$$M_n(g(X)) := \frac{1}{n}\sum_{k=1}^{n} g(X_k)$$

permet d'estimer $\mathbb{E}g(X)$. En particulier, la moyenne empirique $M_n(X)$ et la variance empirique $V_n(X)$ sont estimées respectivement par

$$M_n(X) = \frac{1}{n}\sum_{k=1}^{n} X_k; \quad V_n(X) = M_n(X^2) - \big(M_n(X)\big)^2 = \frac{1}{n}\sum_{k=1}^{n}(X_k - M_n(X))^2. \tag{1.2.4}$$

Simulation 1.2.3. Soit X est une variable aléatoire qui prend les valeurs -1, 0 et 1 avec les probabilités :

$$\mathbb{P}(X=-1) = 1/3,\ \mathbb{P}(X=0) = 1/6,\ \mathbb{P}(X=1) = 1/2.$$

1. Donnez une estimation $V_n(X)$ de la variance de X, pour $n = 10000$, à partir de l'échantillon simulé dans l'exemple **??** et le comparer avec $\mathrm{Var}(X)$ à 10^{-1} près.

Réponse. En utilisant l'échantillon simulé on obtient pour $n = 10000$,

$$M_n(X^2) := (-1)^2 \times f_n(-1) + 0 \times f_n(0) + 1 \times f_n(1) = 0.837.$$

Comme $M_n(X) = 0.1618$, on en déduit (à 10^{-1} près) que $V_n(X) = 0.81 = 29/36 = \mathrm{Var}(X)$.

Une question qu'on peut se poser est de savoir si on peut améliorer l'estimation de l'espérance dans la simulation 1.2.2 pour une taille d'échantillon qui reste invariée et égale à $n = 100$. Nous avons dit que la précision de l'estimation de la moyenne théorique par la moyenne empirique dépend de la taille de l'échantillon et la variance. Comme la taille de l'échantillon est déjà fixée on peut chercher à réduire la variance. Pour cela, on cherchera à réécrire l'espérance de la variable aléatoire X comme l'espérance d'une autre variable aléatoire Y de sorte que $\mathbb{E}(X) = \mathbb{E}(Y)$ mais que $\mathrm{Var}(Y) < \mathrm{Var}(X)$. Cette technique s'appelle la *méthode de réduction de la variance*.

1.2.4 Une méthode de réduction de la variance pour la simulation

Nous illustrons la méthode de réduction de la variance à travers l'exemple utilisé dans la simulation 1.2.2 où on se donne une variable aléatoire X qui prend les valeurs -1, 0 et 1 avec

les probabilités :

$$\mathbb{P}(X=-1)=1/3,\ \mathbb{P}(X=0)=1/6,\ \mathbb{P}(X=1)=1/2$$

et où l'on cherche à estimer $\mathbb{E}(X)=1/6$ par la moyenne empirique. Nous venons de voir que la variance de X vaut $29/36$. Notre objectif est d'améliorer l'estimation de $\mathbb{E}(X)$ par la méthode de réduction de la variance. Nous chercherons une variable alétoire Y telle que

$$\mathbb{E}(X)=\mathbb{E}(Y) \quad \text{et} \quad \mathrm{Var}(Y)<\mathrm{Var}(X).$$

Pour cela rappelons que la variance de X est égale à la distance moyenne entre les valeurs prises par la variable aléatoire X et l'espérance de X. Nous remarquons par ailleurs que 0 est plus proche de $\mathbb{E}(X)=1/6$ que 1 et que 1 est plus proche de $\mathbb{E}(X)$ que -1. Pour garder la même espérance et réduire la variance il suffit donc de définir une variable aléatoire Y qui met plus de poids en 0, la valeur la plus proche de l'espérance, et qui en met moins en -1, la valeur la plus éloignée de l'espérance. Nous n'allons pas nous préoccuper de la recherche du choix optimal ; nous donnons juste deux choix possibles. Soit Y_1 la variable aléatoire telle que $\mathbb{P}(Y_1=0)=1/2$, c'est-à-dire que (à peu près) la moitié des observations vaudra 0 si on simulait Y_1. Déterminons les poids à choisir pour les valeurs -1 et 1. Soit $p_{-1}=\mathbb{P}(Y_1=-1)$, $p_0=\mathbb{P}(Y_1=0)$ et $p_1=\mathbb{P}(Y_1=1)$. Comme on veut que $\mathbb{E}(Y_1)=1/6$ on doit donc avoir $\mathbb{E}(Y_1)=-p_{-1}+p_1=1/6$. D'autre part on a $p_{-1}+p_0+p_1=1$. Ce qui mène au système d'équations

$$\begin{cases} -p_{-1}+p_1 &= \frac{1}{6} \\ \\ p_{-1}+p_1 &= \frac{1}{2}. \end{cases}$$

Par conséquent $p_{-1}=\frac{1}{6}$ et $p_1=\frac{1}{3}$. D'autre part $\mathrm{Var}(Y_1)=\mathbb{E}(Y_1^2)-\big(\mathbb{E}(Y_1)\big)^2=\frac{17}{36}$. On voit donc que $\mathrm{Var}(\mathrm{Y}_1)<\mathrm{Var}(X)$.

Nous pouvons aussi mettre encore plus de poids en 0 en définissant une variable aléatoire Y_2 telle que $\mathbb{P}(Y_2=0)=4/5$. Dans ce cas, en reprenant le même raisonnement que précédemment on obtient : $\mathbb{P}(Y_2=-1)=1/60$ et $\mathbb{P}(Y_2=1)=11/60$. On en déduit que $\mathrm{Var}(Y_2)=\frac{8}{36}<\mathrm{Var}(Y_1)<\mathrm{Var}(X)$. On voit que la variance de Y_2 est presque quatre fois moins que celle de X. Cela a pour effet d'accélérer la convergence de la moyenne empirique $M_n(Y_2)$ vers $\mathbb{E}(Y_2)=\mathbb{E}(X)=1/6$ comme on peut le constater sur la figure 1.3 où on a tracé les moyennes empiriques $M_n(X)$ et $M_n(Y_2)$ associées respectivement à X et Y_2 pour un échantillon de taille $n=100$. On trouve que $M_{100}(Y_2)=0.16$ tandis que $M_{100}(X)=0.26$. Donc l'estimateur $M_n(Y_2)$ est meilleur (on donnera plus de sens à ce terme au chapitre **??**) que $M_n(X)$. Comme conséquence de cette simulation on retiendra que *lorsqu'on a deux estimateurs du même paramètre, on préférera toujours celui qui à la plus petite variance*.

Remarque 1.2.4. Remarquons que dans la pratique on cherche souvent à estimer l'espérance, inconnue, d'une variable aléatoire X. Comme $\mathbb{E}(X)$ est inconnue, on ne peut souvent pas choisir à la main, comme on l'a fait dans l'exemple précédent, la variable aléatoire Y telle que $\mathbb{E}(X)=\mathbb{E}(Y)$ et $\mathrm{Var}(Y)<\mathrm{Var}(X)$. De façon générale, les méthodes de réduction de la variance font appel à des méthodes de changement de probabilité et à des problèmes d'optimisation.

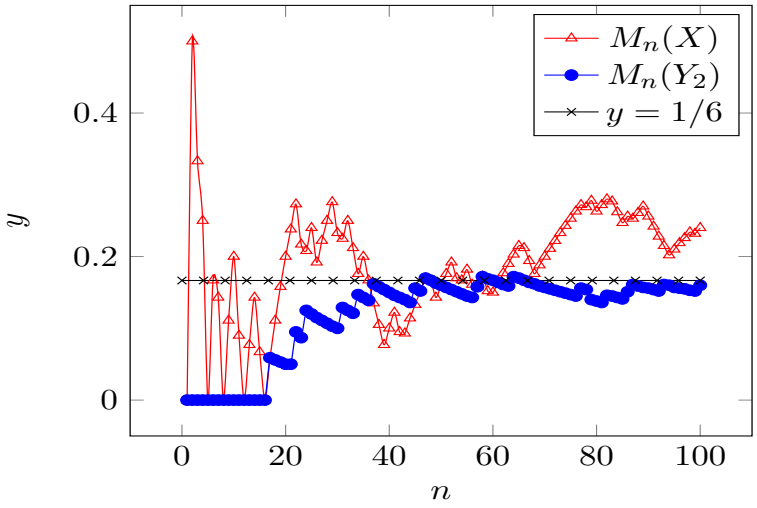

FIGURE 1.3 – En abscisse : taille n de l'échantillon. Ordonnée : $M_n(X)$, $M_n(Y_2)$ et la droite d'équation $y = 1/6$.

1.3 Simulation d'une variable aléatoire continue

La simulation d'une variable aléatoire continue peut être basée sur le même principe que celle des variables aléatoires discrètes, c'est-à-dire, sur la proposition 1.2.1. Pour simuler une variable aléatoire X, il suffit donc de savoir calculer $F^{-1}(U)$ où U est une v.a. uniforme sur $]0,1[$. Considérons l'exemple de la loi dite exponentielle de paramètre $\lambda > 0$ dont la densité s'écrit

$$f(x) = \lambda\, e^{-\lambda x} \mathbf{1}_{\{x>0\}} \begin{cases} \lambda\, e^{-\lambda x} & \text{si } x > 0 \\ 0 & \text{si } x \leq 0. \end{cases}$$

Nous avons déjà vu que la fonction de répartition de X est donnée par

$$F(x) = \begin{cases} 1 - e^{-\lambda x} & \text{si } x > 0 \\ 0 & \text{si } x \leq 0. \end{cases}$$

On en déduit que la fonction quantile est définie pour tout $u \in]0,1[$ par

$$F^{-1}(u) = -\frac{\log(1-u)}{\lambda}.$$

Pour simuler une réalisation x de la loi exponentielle de paramètre λ on simule une réalisation u de la loi uniforme sur $]0,1[$ et on pose $x = -\frac{\log(1-u)}{\lambda}$.

Simulation 1.3.1. Soit X une v.a. de loi exponentielle de paramètre λ.

1. Simuler un échantillon de taille $n = 30$ de X pour $\lambda = 2$.
2. Simuler un échantillon de taille $n = 30$ de X pour $\lambda = 8$.

Réponse. Pour simuler un échantillon $x_1, \ldots, x_n$ de taille n de la loi exponentielle de paramètre λ on simule un échantillon $u_1, \ldots, u_n$ de la loi uniforme sur $]0,1[$ et on pose

$$x_i = -\frac{\log(1-u_i)}{2} \quad \text{pour } \; i = 1, \ldots, n.$$

1.103	0. 279	0.020	0.787	0.195	0.550	1.041	0.840	0.067	1.094
0.047	0.814	0.340	0.304	0.456	0.706	0.169	0.149	0.024	0.555
0.164	0.288	0.878	0.262	0.325	0.287	0.185	0.220	0. 201	0.368

TABLE 1.3 – Echantillon de taille 30 d'une loi exponentielle de paramètre $\lambda = 2$.

1. Le résultat de la simulation est donné dans le tableau 1.3.
2. Le résultat de la simulation est donné dans le tableau 1.4.

0.276	0. 070	0.005	0.197	0.049	0.137	0.260	0.210	0.017	0.274
0.012	0.203	0.085	0.076	0.114	0.177	0.042	0.037	0.006	0.137
0.041	0.072	0.220	0.065	0.081	0.072	0.046	0.055	0.050	0.092

TABLE 1.4 – Echantillon de taille 30 d'une loi exponentielle de paramètre $\lambda = 8$.

1.4 Approximation de $\mathbb{E}g(X)$: méthode de Monte Carlo

Dans beaucoup de problèmes pratiques, on est amené à calculer

$$\mathbb{E}[g(X)], \text{ où } X \text{ est une v.a. à valeurs dans } \mathbb{R}^d, \qquad f : \mathbb{R}^d \to \mathbb{R}. \tag{1.4.1}$$

Lorsque la loi de probabilité de X est connue, il arrive que l'on puisse déterminer une expression analytique à (1.4.1). Par exemple, si $g : \mathbb{R} \mapsto \mathbb{R}$, qui à x associe $g(x) = x^n$, $n \in \mathbb{N}$, et si $X \in \mathcal{N}(0; 1)$, alors

$$\mathbb{E}[g(X)] = \mathbb{E}(X^n) = \frac{(2n)!}{2^n n!}.$$

Si $X = (X_1, X_2)$, $X_i \sim \mathscr{E}(\lambda_i)$, $i = 1, 2$, avec X_1 et X_2 indépendantes, et si $g : \mathbb{R}^2 \mapsto \mathbb{R}$, qui à $x = (x_1, x_2)$ associe $g(x) = x_1 + x_2$, alors

$$\mathbb{E}[g(X)] = \mathbb{E}(X_1 + X_2) = \frac{1}{\lambda_1} + \frac{1}{\lambda_2}.$$

Par ailleurs, lorsque (1.4.1) n'admet pas d'expression analytique, on peut faire recours à des méthodes d'approximation. Une des méthodes d'approximation est la méthode de Monte Carlo (MC). Le principale avantage de la méthode de Monte Carlo par rapport aux méthodes d'approximation connues jusque là est que l'erreur théorique d'approximation induite par la méthode de MC ne dépend pas de la dimension d du vecteur X. D'autre part, cette dernière méthode peut s'appliquer dès lors que la variable aléatoire est simulable (en particulier, on n'a pas forcément besoin de déterminer la loi explicite de X dès lors qu'on sait la simuler).

De façon générale, le méthode de MC peut être utilisée dans les cas suivants.

$\rightsquigarrow$ *La loi de X est connue mais que $\mathbb{E}[g(X)]$ n'admet pas d'expression analytique*. Par exemple, on ne sait pas donner une expression analytique à la quantité $\mathbb{E}[g(X)]$, où $X \sim \mathcal{N}(0; 1)$ et où $g : \mathbb{R} \mapsto \mathbb{R}$, qui à x associe $g(x) = \exp(\sin(x))$, malgré le fait que la loi de probabilité de X est explicite.

$\rightsquigarrow$ *La loi de X n'est pas explicite mais elle peut être simulée*. Considérons le problème du calcul de $\mathbb{E}(Y_m)$, où la variable aléatoire Y_m est définie à travers la formule de récurrence suivante :

$$Y_{k+1} = Y_k + \sigma(Y_k)Z_{k+1}, \quad k = 0, \ldots, n-1,\ Y_0 = 0,$$

où pour tout $k = 1, \ldots, m$, $Z_k \sim \mathscr{E}(1)$ et sont indépendentes. Dans cette situation, la loi de probabilité de Y_m n'est pas connue, pour $m \geq 2$. Par ailleurs, comme on sait simuler Y_m, on peut donner une approximation de $\mathbb{E}(Y_m)$ par la méthode de Monte Carlo.

Maintenant, comment estimer $\mathbb{E}[g(X)]$ par MC lorsqu'on dispose d'un échantillon $X_1(\omega), \ldots, X_n(\omega)$ de taille n (assez grande), de la loi de X, c'est-à-dire, une suite $X_1(\omega), \ldots, X_n(\omega)$ de réalisations indépendantes de la loi de X. Nous allons distinguer les cas de variables aléatoires discrète et continue.

$\triangleright$ *Lorsque X est discrète à valeurs dans $\{x_1, \ldots, x_n, \ldots\}$, $x_i \in \mathbb{R}^d$*. On a dans ce cas

$$\mathbb{E}[g(X)] = \sum_{i=1}^{+\infty} g(x_i)\mathbb{P}(X = x_i).$$

On sait que, comme on l'a déjà vu au Chapitre **??**, lorsque n est assez grande, la fréquence $f_n(x_i)$ d'apparition de x_i devient proche de $\mathbb{P}(X = x_i)$. Ce qui fait que nous pouvons approcher $\mathbb{E}[g(X)]$ par

$$M_n(g(X))(\omega) := \sum_{i=1}^{+\infty} g(x_i)f_n(x_i).$$

En notant $n_i = \mathrm{card}\big(\{k \in \{1, \ldots, n\} :\ X_k(\omega) = x_i\}\big)$ et $I = \{i \geq 1,\ n_i \neq 0\}$, la quantité $M_n(g(X))(\omega)$ peut s'écrire comme

$$\begin{aligned} M_n(g(X))(\omega) &= \sum_{i=1}^{+\infty} g(x_i)\frac{\mathrm{card}\big(\{k \in \{1, \ldots, n\} :\ X_k(\omega) = x_i\}\big)}{n} \\ &= \frac{1}{n}\sum_{i=1}^{+\infty} n_i g(x_i) = \frac{1}{n}\sum_{i=1}^{n} g(x_i) = \frac{1}{n}\sum_{k=1}^{n} g(X_k(\omega)). \end{aligned}$$

Par conséquent, si on dispose d'un échantillon $X_1(\omega), \ldots, X_n(\omega)$ de taille n de la loi de X alors, on peut approcher $\mathbb{E}[g(X)]$ par la moyenne empirique $M_n(g(X))$ associée à $g(X)$. On montre ci-après que cela reste vrai lorsque X est une variable aléatoire continue de densité de probabilité f.

$\triangleright$ *Lorsque X est continue à valeur réelle*. Dans ce cas

$$\mathbb{E}[g(X)] = \int_{-\infty}^{+\infty} g(x)f(x)dx.$$

Maintenant, supposons (quitte à le réordonner) que notre échantillon $(X_1(\omega), \ldots, X_n(\omega)) = (x_1, \ldots, x_n)$ est ordonné avec $x_1 \leq x_2 \leq \ldots \leq x_n$. Posons

$$x_{i-} = \frac{x_i + x_{i-1}}{2}, \quad x_{i+} = \frac{x_i + x_{i+1}}{2}, i = 1, \ldots, n, \text{ avec } x_{0-} = -\infty, x_{n+} = +\infty.$$

On a

$$\begin{aligned}\mathbb{E}[g(X)] &= \int_{-\infty}^{+\infty} g(x)\mathbb{P}_X(dx)\\ &= \sum_{i=1}^{n}\int_{x_{i-}}^{x_{i+}} g(x)\mathbb{P}_X(dx)\\ &\approx \sum_{i=1}^{n} g(x_i)\mathbb{P}(X\in]x_{i-},x_{i+}]).\end{aligned}$$

Comme $]x_{i-},x_{i+}]$ ne contient que le point x_i, on peut donc approcher $\mathbb{P}(X\in]x_{i-},x_{i+}])$ par $f_n(x_i)$ qui est la fréquence d'apparition de x_i et qui vaut n_i/n (où $n_i = \mathrm{card}\big(\{k\in\{1,\dots,n\} : X_k(\omega)=x_i\}\big)$). Par conséquent on peut écrire

$$\mathbb{E}[g(X)] \approx \sum_{i=1}^{n} g(x_i)f_n(x_i) = \frac{1}{n}\sum_{i=1}^{n} g(X_i(\omega)) = M_n(g(X))(\omega).$$

Cette démarche peut s'étendre facilement aux variables aléatoires à valeurs dans $\mathbb{R}^d$. Ce qui fait que de façon générale, si X est une variable aléatoire à valeurs dans $\mathbb{R}^d$ et si g est une fonction de $\mathbb{R}^d$ à valeurs dans $\mathbb{R}$, alors $\mathbb{E}[g(X)]$ peut être approchée par la moyenne empirique

$$M_n(g(X))(\omega) = \frac{1}{n}\sum_{i=1}^{n} g(X_i(\omega)),$$

pour un échantillon $X_1(\omega),\dots,X_n(\omega)$ de taille n de la loi de X. Remarquons que la loi des grands nombres assure la convergence presque sûre de la variable aléatoire $M_n(g(X))$ vers la constante $\mathbb{E}[g(X)]$, lorsque n tend vers $+\infty$.

On peut donc, comme pour le cas d'une v.a. discrète, estimer l'espérance d'une variable aléatoire continue X par la moyenne empirique d'un échantillon de la loi de X en utilisant la méthode de Monte Carlo. De façon plus générale, la proposition 1.2.2 reste vraie pour les variables aléatoires continues.

Proposition 1.4.1. *Soit g une fonction à valeurs réelles et X une variable aléatoire (discrète ou continue). Soit $X_1(\omega),\dots,X_n(\omega)$ un échantillon observé de taille n de la loi de X. Alors la quantité*

$$M_n(g(X)) := \frac{1}{n}\sum_{k=1}^{n} g(X_k)$$

permet d'estimer $\mathbb{E}g(X)$. On dit aussi que $M_n(g(X))$ est un estimateur de $\mathbb{E}(g(X))$. En particulier, la moyenne empirique $M_n(X)$ et la variance empirique $V_n(X)$ définies par

$$M_n(X) = \frac{1}{n}\sum_{k=1}^{n} X_k;\quad V_n(X) = M_n(X^2) - \big(M_n(X)\big)^2 = \frac{1}{n}\sum_{k=1}^{n}(X_k - M_n(X))^2 \quad (1.4.2)$$

sont des estimateurs de $\mathbb{E}(X)$ et de $\mathrm{Var}(X)$.

Simulation 1.4.1. Considérons à nouveau l'exemple 1.3.1 où X est une v.a. continue de densité

$$f(x) = \lambda\, e^{-\lambda x}\mathbf{1}_{\{x>0\}} \begin{cases} \lambda\, e^{-\lambda x} & \text{si } x > 0 \\ 0 & \text{si } x \leq 0. \end{cases}$$

1. Donner une estimation de l'espérance et de la variance de X, pour $\lambda = 2$ et $\lambda = 8$, à travers les échantillons simulés dans l'exemple 1.3.1.
2. Comparez les valeurs précédemment estimées avec les valeurs exactes de l'espérance et de la variance de X pour $\lambda = 2$ et $\lambda = 8$.

Réponse. 1. Si $X_1(\omega), \ldots, X_n(\omega)$ est l'échantillon observé l'espérance peut être estimée par la moyenne empirique

$$M_n(X) = \frac{1}{n}\sum_{i=1}^{n} X_i$$

et la variance peut être estimée par

$$V_n(X) = \frac{1}{n}\sum_{i=1}^{n}(X_i - M_n(X))^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - (M_n(X))^2.$$

On a

— $M_{30}(X) = 0.42$ pour $\lambda = 2$ et $M_{30}(X) = 0.11$ pour $\lambda = 8$.
— $V_{30}(X) = 0.106$ pour $\lambda = 2$ et $V_{30}(X) = 0.007$ pour $\lambda = 8$.

2. Nous avons déjà vu que

$$\mathbb{E}(X) = \frac{1}{\lambda} \qquad \text{et} \qquad \mathrm{Var}(X) = \frac{1}{\lambda^2}.$$

Donc

— $\mathbb{E}(X) = 0.5$ pour $\lambda = 2$ et $\mathbb{E}(X) = 0.125$ pour $\lambda = 8$.
— $\mathrm{Var}(X) = 0.25$ pour $\lambda = 2$ et $\mathrm{Var}(X) = 0.016$ pour $\lambda = 8$.

On constate que l'erreur d'estimation est plus faible pour $\lambda = 8$ que pour $\lambda = 2$. L'explication est en partie donnée par la remarque suivante.

Remarque 1.4.1. On verra ultérieurement que la variance d'une v.a. est un des facteurs qui influe sur l'estimation de l'espérance par la moyenne empirique. En effet, plus la variance est petite plus l'estimation est précise. L'exemple précédent en est une petite illustration.

1.5 Simulation de quelques lois de probabilité usuelles

1.5.1 Quelques lois discrètes et leur simulation

1.5.1.1 Loi de Bernoulli

Pour simuler une réalisation de X on utilise la proposition 1.2.1. Soit X une variable aléatoire de loi de Bernoulli de paramètre $p = \frac{1}{6}$.

1. Simuler un échantillon de taille $n = 40$ de X.
2. Quelles sont les fréquences d'apparition de 0 et de 1. Comparer ces fréquences aux probabilités associées à 0 et 1.

Réponse. 1. La fonction de répartition de X est donnée par

$$F(x) = \begin{cases} 0 & \text{si } x < 0 \\ 1-p & \text{si } x \in [0,1[\\ 1 & \text{si } x \geq 1 \end{cases}$$

de sorte que la fonction de répartition inverse est donnée par

$$F^{-1}(u) = \begin{cases} 0 & \text{si } u \in]0, 1-p[\\ 1 & \text{si } u \in [1-p, 1[. \end{cases}$$

Donc pour simuler une réalisation de X on simulera d'abord une réalisation u de la loi uniforme sur $]0,1[$, puis, on attribuera à X la valeur 0 si $u < 1-p$ et la valeur 1 si $u \geq 1-p$. On répétera ce procédé n fois pour simuler un échantillon de taille n. Le tableau suivant génère un échantillon de taille $n = 40$ de la loi de Bernoulli de paramètre $p = \frac{1}{6}$.

0	0	1	0	1	0	0	0	0	0
0	0	0	1	0	0	0	1	0	0
1	0	0	1	0	0	0	0	0	0
1	1	0	0	0	0	0	0	1	0

TABLE 1.5 – Echantillon de taille 40 de la loi de Bernoulli de paramètre $p = \frac{1}{6}$.

2. On a $f_{40}(0) = \frac{31}{40} = 0.775$ et $f_{40}(1) = \frac{9}{40} = 0.225$. On constate, à 10^{-1} près, que les fréquences d'apparition de 0 et 1 coincident avec les probabilités associées à 0 et 1.

1.5.1.2 Loi Binomiale

Soit X une variable aléatoire de loi binomiale de paramètres $n = 10$ et $p = \frac{1}{2}$.

1. Simuler un échantillon de taille $N = 30$ de X.
2. Quelles sont les fréquences d'apparition de $1, \ldots, 10$. Comparer ces fréquences aux probabilités associées.
3. Répondez à la même question que précédemment pour une taille d'échantillon $N = 100, 200, 300, 400, 500, 1000, 2000, 3000, 4000, 5000, 10000$.

Réponse. 1. Pour simuler une réalisation de la variable aléatoire X nous utilisons la relation (**??**). Il suffit donc de simuler un échantillon de taille 10 de la loi de Bernoulli de paramètre $\frac{1}{2}$ et de compter le nombre de 1 qui apparaît. On répète ce procédé N fois pour simuler un échantillon de taille N de la loi binomiale de paramètres n et $p = \frac{1}{2}$. Nous représentons dans le tableau 1.6 l'échantillon trouvé.

3	5	4	3	5	5	6	3	4	5
6	8	6	7	6	3	4	6	4	6
4	6	6	2	3	8	5	4	3	6

TABLE 1.6 – Echantillon de taille 30 de la loi binomiale de paramètres $n = 10$ et $p = \frac{1}{2}$.

2. On a $f_{30}(0) = f_{30}(1) = f_{30}(9) = f_{30}(10) = 0$; $f_{30}(2) = \frac{1}{30}$, $f_{30}(3) = \frac{6}{30}$, $f_{30}(4) = \frac{6}{30}$, $f_{30}(5) = \frac{5}{30}$, $f_{30}(6) = \frac{9}{30}$, $f_{30}(7) = \frac{1}{30}$, $f_{30}(8) = \frac{1}{30}$.

3. Nous représentons les résultats obtenus dans le tableau 1.7. On constate qu'à partir d'une taille d'échantillon égale à 100 les fréquences d'apparition des valeurs prises par la variable aléatoire X sont égales (à 10^{-1} près) aux probabilités qui leur sont associées. Il faut noter que comme la variance de X vaut $np(1-p)$, elle sera d'autant plus importante (de surcroit l'estimation des probabilités associées aux valeurs prises par X sera moins précise pour une taille d'échantillon N fixée) que le paramètre n est grand.

	0	1	2	3	4	5	6	7	8	9	10
30	0.00	0.00	0.03	0.20	0.20	0.17	0.30	0.03	0.03	0.00	0.00
100	0.00	0.01	0.03	0.17	0.16	0.29	0.23	0.10	0.00	0.01	0.00
200	0.00	0.01	0.05	0.10	0.21	0.26	0.21	0.16	0.02	0.01	0.00
300	0.00	0.02	0.05	0.09	0.24	0.21	0.23	0.11	0.03	0.01	0.00
400	0.00	0.01	0.04	0.13	0.24	0.23	0.20	0.12	0.03	0.01	0.00
500	0.00	0.01	0.04	0.11	0.20	0.25	0.23	0.11	0.04	0.01	0.00
1000	0.00	0.01	0.04	0.11	0.19	0.26	0.21	0.12	0.05	0.01	0.00

TABLE 1.7 – Fréquences d'apparition des valeurs prises par une loi binomiale de paramètres $n = 10$ et $p = 1/2$ pour des échantillons de taille $N \in \{30, 100, 200, 300, 400, 500, 1000\}$.

1.5.1.3 Loi de Poisson

Pour simuler une réalisation x d'une variable aléatoire X de loi de Poisson de paramètre λ on peut utiliser la proposition 1.2.1. Il suffit alors de générer une réalisation u d'une loi uniforme sur $]0, 1[$ et de poser $x = k$ si $F(k-1) < u \leq F(k)$. On peut trouver la valeur de x avec l'algorithme suivant.

- Poser $k = 0$; $p = e^{-\lambda}$; $F = p$;
- Simuler une réalisation u de la loi uniforme sur $]0, 1[$.
- Tant que $u > F$, poser $p = \lambda p/(k+1)$; $F = F + p$; $k = k + 1$;
- Poser $x = k$;

Soit X une variable aléatoire de loi de Poisson de paramètre λ.

1. Simuler un échantillon de taille $n = 30$ de la loi de X pour $\lambda = 0.5$.
2. Simuler un échantillon de taille $n = 30$ de la loi de X pour $\lambda = 2$.
3. En déduire, pour chacune des valeurs de λ, une estimation $M_n(X)$ et $V_n(X)$ de l'espérance et de la variance de X.

Réponse. 1. Nous représentons au tableau 1.8 un échantillon de la loi de Poisson de paramètre $\lambda = 0.5$.

2. L'échantillon de taille 30 d'une loi de Poisson de paramètre $\lambda = 2$ est représenté au tableau 1.9.

1	0	0	0	0	0	0	0	0	0
0	1	0	1	1	1	0	0	0	1
0	2	1	1	1	0	0	0	0	0

TABLE 1.8 – Echantillon de taille 30 de la loi de Poisson de paramètre $\lambda = 0.5$.

3	4	2	5	3	4	2	1	7	0
2	1	4	2	1	1	2	3	3	1
1	1	4	1	1	0	2	1	1	3

TABLE 1.9 – Echantillon de taille 30 de la loi de Poisson de paramètre $\lambda = 2$.

1.5.1.4 Loi géométrique

La simulation une réalisation x d'une variable aléatoire de loi géométrique de paramètre p peut être faite dans l'esprit de la définition. On simulera des réalisations $X_1, X_2, \ldots$ de la loi de Bernoulli de paramètre p et on pose $x = \inf\{k \geq 1, \ \ X_k = 1\}$, c'est-à-dire, le premier instant où Pile apparaît. Nous l'énonçons sous forme de proposition qui se prouve sans difficulté.

Proposition 1.5.1. *Soit $X_1, X_2, \ldots$ des réalisations indépendantes de la loi de Bernoulli de paramètre p. Alors la variable aléatoire*

$$N = \inf\{k \geq 1, \ X_k = 1\}$$

suit la loi géométrique de paramètre de succès p.

Soit donc X une variable aléatoire de loi géométrique de paramètre de succès $p = 1/6$.

1. Simuler un échantillon de taille $n = 30$ de la loi géométrique de paramètres $p = 0.1$ et $p = 0.9$.
2. Donner une estimation de l'espérance et de la variance de X pour les deux échantillons simulés et les comparer avec les valeurs exactes.
3. Représenter sur un tableau les estimations de l'espérance et de la variance pour $n = 30$, 100, 1000, 10000 et pour $p = 0.1$ et $p = 0.9$. Commenter le tableau.
4. Supposons que nous disposons des données simulées mais que la seule information disponible est qu'elles sont générées selon une loi géométrique de paramètre p que nous ne connaissons pas. Comment déduire une estimation du paramètre p pour chacun des échantillons à partir de l'estimation de l'espérance.

Réponse.1. Nous repésentons au tableau 1.10 l'échantillon de taille 30 de la loi géométrique de paramètre $p = 0.1$

et au tableau 1.11 l'échantillon de taille 30 de la loi géométrique de paramètre $p = 0.9$.

2. On a

— $M_{30}(X) = 11.67$ pour $p = 0.1$ et $M_{30}(X) = 1.07$ pour $p = 0.9$

— $V_{30}(X) = 129.09$ pour $p = 0.1$ et $V_{30}(X) = 0.06$ pour $p = 0.9$.

2	4	1	14	3	31	1	23	8	12
12	37	13	1	7	51	4	10	5	18
1	6	21	4	9	17	2	10	12	11

TABLE 1.10 – Echantillon de taille 30 de la loi géométrique de paramètre $p = 0.1$.

1	1	1	1	1	1	2	1	1	1
1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	2	1	1	1

TABLE 1.11 – Echantillon de taille 30 de la loi géométrique de paramètre $p = 0.9$.

D'autre part nous avons

— $\mathbb{E}(X) = 10$ pour $p = 0.1$ et $\mathbb{E}(X) = 1.111$ pour $p = 0.9$

— $\mathrm{Var}(X) = 90$ pour $p = 0.1$ et $\mathrm{Var}(X) = 0.123$ pour $p = 0.9$.

Nous constatons que les estimations sont plus précises pour $p = 0.9$. Cela s'explique par le fait que la variance est beaucoup plus petite lorsque $p = 0.9$ que pour $p = 0.1$.

3. Nous représentons au tableau 1.12 le tableau demandé.

	$n = 30$	$n = 100$	$n = 1000$	$n = 10000$	$n = 10^5$
$p = 0.1$					
$M_n(X)$	11.67	10.34	10.08	10.00	09.97
$V_n(X)$	129.09	95.61	94.55	89.83	89.36
$p = 0.9$					
$M_n(X)$	1.070	1.16	1.109	1.114	1.112
$V_n(X)$	0.062	0.214	0.117	0.130	0.123

TABLE 1.12 – Comparaison des estimations de l'espérance et de la variance d'une v.a. de loi géométrique de paramètre de succès p.

4. Notons $\hat{p}_n$ l'estimation de p à partir d'un échantillon de taille n. Comme

$$\mathbb{E}(X) = \frac{1}{p}$$

et que $\mathbb{E}(X)$ est estimé par $M_n(X)$, on peut déduire une estimation de p, pour n fixé, en résolvant l'équation suivante en $\hat{p}_n$:

$$M_n(X) = \frac{1}{\hat{p}_n}.$$

D'où, pour tout n,

$$\hat{p}_n = \frac{1}{M_n(X)}.$$

Dans le tableau 1.13, nous reconsidérons le tableau 1.12 en rajoutant une colonne pour les $\hat{p}_n$.

	$n=30$	$n=100$	$n=1000$	$n=10000$	$n=10^5$
$p=0.1$					
$\hat{p}_n$	0.086	0.097	0.099	0.100	0.100
$M_n(X)$	11.67	10.34	10.08	10.00	09.97
$V_n(X)$	129.09	95.61	94.55	89.83	89.36
$p=0.9$					
$\hat{p}_n$	0.935	0.862	0.902	0.898	0.899
$M_n(X)$	1.070	1.16	1.109	1.114	1.112
$V_n(X)$	0.062	0.214	0.117	0.130	0.123

TABLE 1.13 – Comparaison des estimations du paramètre de succès, de l'espérance et de la variance d'une v.a. de loi géométrique.

1.5.2 Quelques lois continues et leur simulation

1.5.2.1 Loi uniforme

Il découle de la proposition **??** que pour simuler une réalisation x de la loi uniforme sur $]a,b[$ on simule une réalisation u d'une loi uniforme sur $]0,1[$ et on pose $x = a + (b-a)u$.

1. Simuler un échantillon de taille 30 de la loi uniforme sur $]0,1[$.
2. Simuler un échantillon de taille 30 de la loi uniforme sur $]0,2[$.
3. On présente les deux échantillons à une personne à qui on donne l'information que ceux-ci sont simulés selon une loi uniforme sur $]0,b[$, $b>0$, sans lui révéler la valeur de b pour chaque échantillon.
   (a) Comment peut-elle estimer b pour chaque échantillon.
   (b) Expliquer pourquoi b est mieux estimer pour l'échantillon de la loi uniforme sur $]0,1[$ que pour l'échantillon de la loi uniforme sur $]0,2[$.

Réponse. 1. L'échantillon simulé est présenté à la figure **??**

1.5.2.2 Loi exponentielle

Soit X une variable aléatoire de loi exponentielle de paramètre λ

La simulation d'une réalisation x de la loi exponentielle de paramètre λ est basée sur l'assertion (**??**), comme on l'a déjà vu dans la simulation 1.3.1. Par conséquent, pour simuler un échantillon $x_1,\ldots,x_n$ de taille n de la loi exponentielle de paramètre λ on simule un échantillon $u_1,\ldots,u_n$ de la loi uniforme sur $]0,1[$ et de poser

$$x_i = -\frac{\log(1-u_i)}{2} \quad \text{pour } i=1,\ldots,n.$$

Reconsidérons les deux échantillons obtenus dans la simulation 1.3.1, et qui sont simulés selon une loi exponentielle de paramètre $\lambda=2$ et $\lambda=8$. Nous rappelons les échantillons obtenus aux tableaux 1.14 et 1.15.

1. Si on présente les deux échantillons à une personne à qui on dit que ceux-ci sont simulés selon une loi de exponentielle de paramètre λ sans lui révéler les valeurs de λ, comment peut-elles estimer les valeurs de λ dans chacun des cas.

2. Expliquer pourquoi on peut s'attendre à ce que λ soit mieux estimé pour l'échantillon simulé selon la loi exponentielle de paramètre $\lambda = 8$ que pour celui simulé selon la loi exponentielle de paramètre $\lambda = 2$.

1.103	0. 279	0.020	0.787	0.195	0.550	1.041	0.840	0.067	1.094
0.047	0.814	0.340	0.304	0.456	0.706	0.169	0.149	0.024	0.555
0.164	0.288	0.878	0.262	0.325	0.287	0.185	0.220	0. 201	0.368

TABLE 1.14 – Echantillon de taille 30 d'une loi exponentielle de paramètre $\lambda = 2$.

0.276	0. 070	0.005	0.197	0.049	0.137	0.260	0.210	0.017	0.274
0.012	0.203	0.085	0.076	0.114	0.177	0.042	0.037	0.006	0.137
0.041	0.072	0.220	0.065	0.081	0.072	0.046	0.055	0.050	0.092

TABLE 1.15 – Echantillon de taille 30 d'une loi exponentielle de paramètre $\lambda = 8$.

Réponse. 1. Rappelons que si X est une variable aléatoire de loi exponentielle de paramètre λ alors, $\mathbb{E}(X) = \frac{1}{\lambda}$. Pour estimer λ on utilise le fait que $\mathbb{E}(X)$ peut être estimée par la moyenne empirique $M_n(X) = \frac{1}{n}\sum_{i=1}^{n} X_i$ où $(X_1, X_2, \ldots, X_n)$ est un échantillon de taille n de la loi exponentielle de paramètre λ.

1.5.2.3 Loi Normale

Il existe plusieurs méthodes de simulation d'une variable aléatoire gaussienne parmi lesquelles la méthode de Box-Muller, la méthode de Polar. Nous utiliserons la méthode de Box Muller. Elle est basée sur le résultat suivant qui sera démontré au chapitre **??**.

Proposition 1.5.2. *Soit U_1 et U_2 deux variables aléatoires indépendantes de loi uniforme sur $]0, 1[$. Posons*

$$X_1 = \sqrt{-2\log(U_1)}\cos(2\pi U_2) \qquad \textit{et} \quad X_2 = \sqrt{-2\log(U_2)}\sin(2\pi U_1).$$

Alors X_1 et X_2 sont des variables aléatoires indépendantes de loi gaussienne centrée réduite : $X_1 \sim \mathcal{N}(0; 1)$, $X_2 \sim \mathcal{N}(0; 1)$ et X_1 et X_2 sont indépendantes.

Par conséquent, pour simuler une réalisation x variable aléatoire $X \sim \mathcal{N}(0; 1)$ on simule d'abord une réalisation u_1 de la loi uniforme sur $]0, 1[$ puis une autre réalisation u_2, indépendante de u_1 de la loi uniforme sur $]0, 1[$ et on pose

$$x = \sqrt{-2\log(u_1)}\cos(2\pi u_2) \qquad \text{ou bien} \quad x = \sqrt{-2\log(u_2)}\sin(2\pi u_1).$$

Pour simuler une réalisation d'une variable aléatoire Y de loi gaussienne de moyenne μ et de variance σ^2 : $Y \sim \mathcal{N}(\mu; \sigma^2)$, $\sigma > 0$, on utilise le fait que

$$\text{si } Y \sim \mathcal{N}(\mu, \sigma^2) \text{ alors on peut écrire } Y = \mu + \sigma X, \text{ avec } X \sim \mathcal{N}(0; 1). \tag{1.5.1}$$

Ainsi, pour simuler une réalisation y de $Y \sim \mathcal{N}(\mu, \sigma^2)$ on simule une réalsation x de $X \sim \mathcal{N}(0; 1)$ et on pose $y = \mu + \sigma x$.

1. Simuler un échantillon de taille $n = 100$ de la loi de $Y \sim \mathcal{N}(0; 4)$.
2. On présente les données simulées à une personne à qui on dit que l'échantillon est issu d'une loi gaussienne de moyenne μ et de variance σ^2, sans lui révéler les valeurs de μ et σ^2. Comment peut-elle estimer μ et σ^2.

Réponse. 1. L'échantillon obtenu est représenté au tableau 1.16.

0.52	0.09	-1.84	-1.29	1.61	3.80	-3.80	2.83	-1.87	1.10
0.48	-1.28	3.52	-1.39	-1.01	1.72	0.85	-2.28	3.20	0.55
-1.61	-0.40	-0.35	-0.22	2.65	2.60	-1.99	-1.83	-1.99	-0.33
-3.25	2.03	-1.72	-0.24	-2.07	0.13	0.04	-0.03	-0.13	-0.38
0.35	1.26	2.34	-2.22	2.37	-2.52	2.38	1.63	0.23	-0.01
-2.81	0.20	3.82	-1.14	-1.23	-0.05	-0.21	0. 54	-0.98	-2.69
3.01	-2.85	-2.26	-1.08	-0.22	-0.15	-2.35	-1.39	0.70	-1.56
-0.47	1.49	-0.41	-1.87	-2.47	-1.34	0.38	-1.42	0.35	-3.78
-3.71	-3.06	4.02	0.36	-0.88	-1.09	0.88	0.27	-2.87	2.13
1.83	3.50	0.24	-1.91	0.76	-1.37	-0.90	4.44	0.39	1.40

TABLE 1.16 – Echantillon de taille 100 d'une loi Normale de moyenne 0 et de variance 4.