# Concentration, loi des grands nombres (non évaluable à l'épreuve écrite)

**Dans ce troisième chapitre, l'inégalité de Bienaymé-Tchebychev explicite le rôle de la variance comme indicateur de dispersion. Tous ces outils se conjuguent pour établir l'inégalité de concentration pour la moyenne d'un échantillon d'une variable aléatoire, ce qui justifie l'apparition du facteur $1/\sqrt{n}$ en théorie de l'estimation aperçue expérimentalement en classe de seconde, et ce qui va permettre d'aboutir à la démonstration de la loi des grands nombres, fondement de l'étude probabiliste.**

## I. Quelles sont les formules de majoration à connaître ?

**Propriété :** Inégalité de Bienaymé-Tchebychev

Soient $X$ une variable aléatoire discrète, et $\delta$ un réel strictement positif.

$p(|X - E(X)| \geq \delta) \leq \frac{V(X)}{\delta^2}$ ou $p(|X - E(X)| < \delta) \geq 1 - \frac{V(X)}{\delta^2}$.

**Exemple 1 :** Si, pour une certaine variable aléatoire $X$, $E(X) = 10$ et $V(X) = 0{,}01$, alors sans aucune indication particulière sur la loi de $X$, on sait que $X$ prendra des valeurs entre 9,7 et 10,3 avec une probabilité supérieure à 0,88. Évidemment, si on connaît la loi de $X$, on peut calculer directement la valeur de cette probabilité.

**Remarque :** Ayant un caractère universel, l'inégalité de Bienaymé-Tchebychev possède cependant un défaut : les majorations obtenues ne sont pas très précises.

**Exemple 2 :** Soit $X : B(10 ; 0{,}5)$. On a $E(X) = 10 \times 0{,}5 = 5$ et $V(X) = 10 \times 0{,}5 \times (1 - 0{,}5) = 2{,}5$.

Soit $\delta = 4$. On a alors : $p(|X - 5| \geq 4) \leq \frac{2{,}5}{4^2} \Leftrightarrow p(|X - 5| \geq 4) \leq 0{,}15625$.

Or $|X - 5| \geq 4 \Leftrightarrow X - 5 \geq 4$ ou $-(X - 5) \geq 4 \Leftrightarrow X \geq 9$ ou $X \leq 1$.

$p(|X - 5| \geq 4) = p(\{X \geq 9\} \cup \{X \leq 1\}) = p(X \geq 9) + p(X \leq 1)$

$p(|X - 5| \geq 4) = p(X = 0) + p(X = 1) + p(X = 9) + p(X = 10) \approx 0{,}021$.

Ainsi, l'inégalité $p(|X - 5|$

$\geqslant$

$4)$

$\leqslant$

0,15625 est vraie mais peu « précise ».

**Propriété : Inégalité de concentration**

Soient $n$ un entier naturel non nul, $(X_1, X_2, ..., X_n)$ un échantillon d'une loi de probabilité d'espérance $\mu$ et de variance $V$, $\delta$ un réel strictement positif. On pose $M_n = \frac{X_1 + X_2 + ... + X_n}{n}$. On a : $p(|M_n - \mu| \geq \delta) \leq \frac{V}{n\delta^2}$.

**Exemple :** Combien de fois faut-il lancer une pièce équilibrée pour que la fréquence d'apparition de la face pile soit comprise strictement entre 0,45 et 0,55 avec une probabilité d'au moins 0,99 ?

Soit $X : B(1 ; 0{,}5)$. On a : $E(X) = p = 0{,}5$ et $V(X) = p(1 - p) = 0{,}25$. Soient $n$ un entier non nul et $(X_1, X_2, ..., X_n)$ un échantillon de taille $n$ de la variable aléatoire $X$. On pose $M_n = \frac{X_1 + X_2 + ... + X_n}{n}$.

On cherche l'entier $n$ tel que $p(0{,}45 < M_n < 0{,}55)$

$\geqslant$

0,99, c'est-à-dire $p(|M_n - 0{,}5|$

$\geqslant$

$0{,}05)$

$\leqslant$

0,01.

D'après l'inégalité de concentration, on a : $p(|M_n - 0{,}5| \geq 0{,}05) \leq \frac{V(X)}{n \times 0{,}05^2}$. D'où

$p(|M_n - 0{,}5| \geq 0{,}05) \leq \frac{0{,}25}{n \times 0{,}0025} \Leftrightarrow p(|M_n - 0{,}5| \geq 0{,}05) \leq \frac{100}{n}$.

Il suffit donc d'avoir $\frac{100}{n} \leq 0{,}01$, c'est-à-dire $n$

$\geqslant$

10 000.

Il suffit donc de lancer la pièce 10 000 fois pour que la fréquence d'apparition de la face pile soit comprise strictement entre 0,45 et 0,55 avec une probabilité d'au moins 0,99.

## II. Que faut-il retenir sur la loi faible des grands nombres ?

**Propriété : Loi faible des grands nombres**

Soient $n$ un entier naturel non nul, $(X_1, X_2, ..., X_n)$ un échantillon d'une loi de probabilité d'espérance $\mu$, et $\delta$ un réel strictement positif. On pose $M_n = \frac{X_1 + X_2 + ... + X_n}{n}$. On a : $\lim_{n \to +\infty} p(|M_n - \mu| \geq \delta) = 0$.

**Remarques :** On peut reformuler la propriété en : « Lorsque $n$ est grand, sauf exception, la fréquence observée est proche de la probabilité. »

La limite de la propriété est équivalente à : $mj$\lim_{n\rightarrow +\infty }p(\mu -\delta .

**Démonstration :** On sait que : $E(M_n) = E(X) = \mu$ et $V(M_n) = \frac{V(X)}{n}$.

En utilisant l'inégalité de Bienaymé-Tchebychev, on a : pour tout réel strictement positif δ, $p(|M_n - \mu| \geq \delta) \leq \frac{\frac{V(X)}{n}}{\delta^2}$, soit $p(|M_n - \mu| \geq \delta) \leq \frac{V(X)}{n\delta^2}$. Or $\lim_{n\to+\infty} \frac{V(X)}{n\delta^2} = 0$.
Donc, grâce au théorème des gendarmes, on obtient : $\lim_{n\to+\infty} p(|M_n - \mu| \geq \delta) = 0$.

**Exemple : La méthode de Monte-Carlo**

On souhaite déterminer, à l'aide d'une simulation, une valeur approchée du nombre réel π.
On va pour cela imaginer une cible carrée ABCD de côté 1. On construit le quart de cercle de centre A et de rayon AB. On note $E$ l'expérience aléatoire : « On lance une fléchette sur la cible ». On estime que la fléchette atteint toujours la cible et que la probabilité que la fléchette atteigne une zone est proportionnelle à l'aire de cette zone.

Si on pose A l'événement : « La fléchette touche le quart de disque », alors on peut calculer que $p(\text{A}) = \frac{\frac{\pi - 1^2}{4}}{1^2} = \frac{\pi}{4}$.
Pour déterminer une valeur approchée de π, on va simuler un très grand nombre de lancers de fléchettes, et on va calculer la proportion de fléchettes atteignant le quart de disque.
La loi faible des grands nombres nous garantit que plus la taille de l'échantillon est grande, plus la proportion des fléchettes atteignant le quart de disque va se rapprocher de $\frac{\pi}{4}$.
Voici un script permettant donc de déterminer une valeur approchée de π :

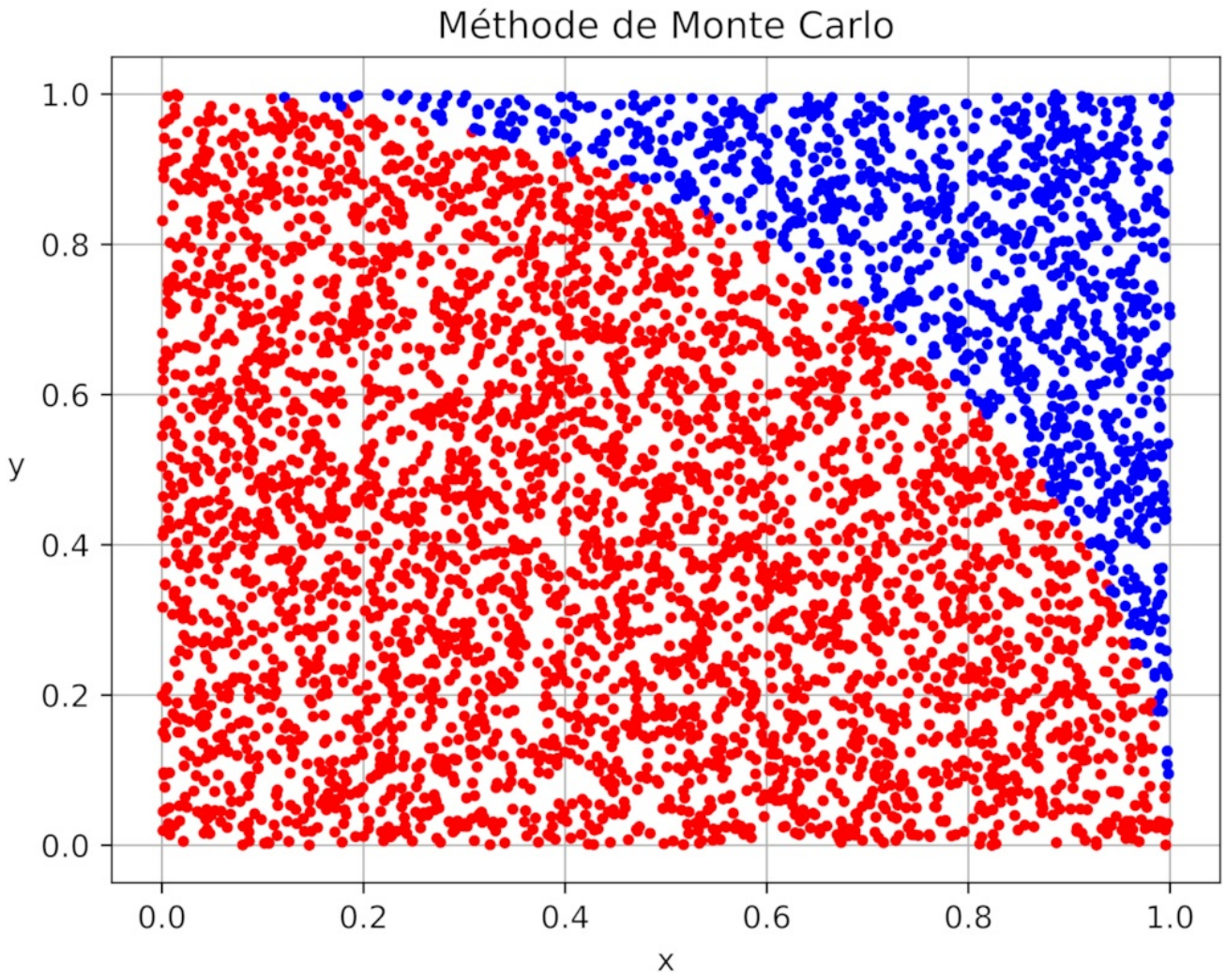
```
import matplotlib.pyplot as plt
from math import *
from random import *

n = int(input("nombre de points = "))
c = 0
for i in range(n):
    x = random()
    y = random()
    if x**2+y**2 < 1 :
        plt.plot(x,y,'r.-')
        c = c+1
    else:
        plt.plot(x,y,'b.-')

print(4*c/n)

plt.xlabel('x')
plt.ylabel('y')
plt.title('Méthode de Monte Carlo')
plt.grid(True)
plt.savefig('test.png')
plt.show
```

En entrant $n$ = 10 000, on a obtenu une valeur approchée de 3,1276, et cette cible :




## III. Comment peut-on simuler une variable aléatoire ?

### Démarrage

Pour simuler une variable aléatoire, il faut donner la loi de probabilité de cette variable aléatoire, c'est-à-dire renseigner les valeurs qu'elle peut prendre et les probabilités associées. On utilisera des listes pour ces deux données.

**Exemple :** Simulation d'un lancer de dé. $\Omega = \{1 ; 2 ; 3 ; 4 ; 5 ; 6\}$, avec pour chaque valeur une probabilité de $\frac{1}{6}$.

• Le script simul permet, grâce à la fonction random, de simuler le choix aléatoire d'une valeur que peut prendre la variable aléatoire en tenant compte du poids des probabilités affectées.

• Le script un_echantillon permet, en affectant à la variable taille un entier non nul, de créer un échantillon. De même, le script `N_echantillon` permet de créer `N` échantillons.

```python
from random import random
from mat import sqrt

def simul(valeurs, probabilites):
    nb = random
    pcumul = 0
    for i in range(len(valeurs)):
        if pcumul <= nb < pcumul+probabilites[i]:
            return valeurs[i]
        pcumul = pcumul+probabilites[i]

def un_echantillon(valeurs, probabilites, taille):
    ech = []
    for j in range(taille):
        ech.append(simul(valeurs, probabilites))
    return ech

def N_echantillon(valeurs, probabilites, taille, N):
    Grand_ech = []
    for k in range(N):
        Grand_ech.append(un_echantillon(valeurs, probabilites, taille))
    return Grand_ech
```

**Paramètres**

On crée trois scripts pour calculer l'espérance, la variance et l'écart type d'une variable aléatoire discrète.
La fonction `N_moyennes` permet de créer une liste qui, après avoir créé `N` échantillons, contient la moyenne de chaque échantillon.
La fonction ecarttype_des_N_moyennes permet de calculer l'écart type de la liste précédente.

```python
def esperance(valeurs, probabilites):
    e = 0
    a = len(valeurs) # la longueur de la liste
    for i in range(a):
        e = e + probabilites[i]*valeurs[i] # somme des pi*xi
    return e/sum(probabilites)

def variance(valeurs, probabilites):
    var = 0
    e = esperance(valeurs, probabilites)
    n = len(valeurs)
    for i in range(n):
        var = var+probabilites[i]*(valeurs[i]-e)**2
    return var

def ecarttype(valeurs, probabilites):
    sigma = sqrt(variance(valeurs, probabilites))
    return sigma

def N_moyennes(valeurs, probabilites, taille, N):
    Grand_ech = N_echantillon(valeurs, probabilites, taille, N)
    liste_de_moyennes = []
    for i in range(len(Grand_ech)):
        liste_de_moyennes.append(sum(Grand_ech[i])/len(Grand_ech[i]))
    return liste_de_moyennes

def ecarttype_des_N_moyennes(valeurs, probabilites, taille, N):
    liste = N_moyennes(valeurs, probabilites, taille, N)
    un = [1 for i in range(N)]
    s = ecarttype(liste, un)
    return s
```

**Écart et proportion**

La fonction ecart1 permet de mesurer, après une simulation de N échantillons, l'écart entre l'écart type de la variable aléatoire $X$ et l'écart type des moyennes de N échantillons.
Plus la variable « taille » est grande, plus cet écart est proche de 0.
La fonction proportion permet de déterminer la proportion des échantillons pour lesquels l'écart entre l'espérance de $X$ et la moyenne de l'échantillon est inférieure ou égal à $\frac{k\sigma(X)}{\sqrt{n}}$ pour $k$ un entier compris entre 1 et 3.

```python
def ecart1(valeurs, probabilites, taille, N):
    return abs(ecarttype_des_N_moyennes(valeurs, probabilites, taille, N)-ecarttype(valeurs, probabilites)/sqrt(taille))

def proportion(valeurs, probabilites, taille, N, k):
    prop = 0
    Grand_ech = N_echantillon(valeurs, probabilites, taille, N)
    for i in range(N):
        if abs(esperance(valeurs, probabilites)-(sum(Grand_ech[i])/taille)) <= k*ecarttype(valeurs, probabilites)/sqrt(taille):
            prop = prop+1
    return prop/N
```

## Pour aller plus loin : Estimation

**Fréquence**

En statistique, la fréquence d'une valeur est le quotient entre son effectif et la taille de la population. On exprime cette fréquence sous la forme d'un pourcentage ou d'un nombre décimal.

### Intervalle de confiance et lien avec la fluctuation

Lorsque l'on veut connaître une information sur l'ensemble d'une population, il est compliqué d'interroger l'ensemble des personnes concernées. On constitue un échantillon représentatif, puis on étend les résultats obtenus.
L'expérience montre que, lorsqu'on choisit un autre échantillon représentatif, on obtient des résultats assez proches. Aussi, pour avoir une meilleure approximation du résultat, on va donner un « intervalle de confiance » qui permet de limiter les effets de la fluctuation en fonction des échantillons.

### « Au seuil de 95 % de la fréquence »

La phrase « au seuil de 95 % de la fréquence » signifie « avec une marge d'erreur de 5 % ».

### Intervalle de fluctuation au seuil de 95 % de la fréquence

Soit $X$ une variable aléatoire qui suit la loi binomiale de paramètres $n$ et $p$, $X : B(n, p)$, avec $0 < p < 1$, $n \geqslant 30$, $np > 5$ et $n(1 - p) > 5$.
On appelle intervalle asymptotique au seuil de 95 % de la fréquence l'intervalle :
$\left[p - 1,96\sqrt{\frac{p(1-p)}{n}}; p + 1,96\sqrt{\frac{p(1-p)}{n}}\right]$

### Intervalle de confiance

Il s'agit de savoir comment estimer la proportion $p$ d'individus dans une population ayant une propriété identique. Pour cela, on utilise la fréquence $f$ observée sur un échantillon de la population.
On appelle intervalle de confiance de la proportion $p$ avec un niveau de confiance de 95 %, l'intervalle $\left[f - \frac{1}{\sqrt{n}}; f + \frac{1}{\sqrt{n}}\right]$, où $n$ est la taille de l'échantillon interrogé.

## Histoire des mathématiques

C'est grâce aux fréquences que Jacques Bernoulli établit l'un des résultats majeurs de son ouvrage l'*Ars Conjectandi* : son « théorème d'or », l'actuelle « loi des grands nombres ». Ce théorème, qui relie fréquences et probabilités, valide le principe de l'échantillonnage et est le premier exemple de « théorème limite » en théorie des probabilités.
Les mathématiciens français et russe, Bienaymé et Tchebychev, ont démontré l'inégalité qui porte leur nom, en parlant de fréquences d'échantillons plutôt que de variables aléatoires. Ils fournissent ainsi la possibilité d'une démonstration plus simple de la loi des grands nombres.
Au début du XIX$^{e}$ siècle, la modélisation des erreurs de mesure va devenir centrale pour faire de la statistique une science à part entière. Lagrange et Laplace ont développé une approche probabiliste de la théorie des erreurs. Gauss imagine une méthode des moindres carrés (après Legendre), qu'il applique avec succès à la prédiction de la position d'un astéroïde. Il y propose de comprendre l'écart type comme une « erreur moyenne à craindre ».
Finalement, l'introduction de méthodes statistiques en sociologie est l'œuvre du mathématicien et astronome belge Quételet dans les années 1830. Il réfléchit alors à la distribution de données autour de la moyenne. Ce procédé sera approfondi par l'anglais Galton, qui est l'inventeur de nombreuses méthodes statistiques couramment employées, comme l'étalonnage et la corrélation, qui lui ont notamment servi dans ses recherches sur la théorie de l'évolution.