**Université de Picardie Jules Verne** **2011-2012**
**UFR des Sciences - LAMFA CNRS UMR 6140**

**PAF Amiens - Formation Enseignement des Mathématiques - 20 janvier 2012**

# Mathématiques : statistiques et simulation

## Introduction : échantillonage et proportion

*(Extrait du document ressource pour la classe de seconde)*
*Dans le sens commun des sondages, un échantillon est un sous-ensemble obtenu par prélèvement aléatoire dans une population.*
*En statistique, un échantillon de taille n est est la liste des n résultats obtenus par n répétitions indépendantes de la même expérience. Par exemple :*
*- 100 lancers d'une pièce, en observant les apparitions de Pile ;*
*- 100 lancers d'un dé à 6 faces, en observant les apparitions du 4 ;*
*- 100 tirages succcessifs avec remise d'une boule dans une urne contenant 2 boules blanches et 1 boule verte, en observant les apparitions d'une boule blanche.*
*Ces trois exemples relèvent du modèle de Bernoulli qui affecte la probabilité p au nombre 1, et la probabilité* $1-p$ *au nombre* 0.

Dans les deux premiers exemples, $p$ peut être vue "directement" comme une probabilité. Dans le dernier cas, $p = \frac{2}{3}$ pourrait être d'abord vue comme la proportion de boules blanches dans l'urne, mais c'est aussi la probabilité d'obtenir une boule blanche lors d'un tirage.
Dans les trois cas, ce calcul de probabilité (sur un ensemble fini) relève du programme de seconde.

Le résultat d'une expérience aléatoire à 2 issues peut être représenté par la

**Loi de Bernoulli** $\mathcal{B}(p)$
Soit $(\Omega, \mathcal{A}, P)$ un espace probabilisé.
Une variable aléatoire $X$ suit la loi de Bernoulli de paramètre $p \in ]0,1[$, que l'on note $\mathcal{B}(p)$, si et seulement si $X$ est à valeurs dans $\{0;1\}$, et $P(X=1) = p$ et $P(X=0) = 1-p$.
On a alors $E(X) = p$ et $Var(X) = p(1-p)$.

**Exemple** d'une urne contenant une proportion $p = \frac{2}{3}$ de boules blanches
On tire une boule au hasard dans l'urne : le nombre de "boule blanche" obtenu en un tirage est une variable aléatoire $X$ de loi de Bernoulli $\mathcal{B}(p)$ : $P(X=1) = p = \frac{2}{3}$ et $P(X=0) = 1-p = \frac{1}{3}$. On a $E(X) = \frac{2}{3}$ et $Var(X) = \frac{2}{3} \times \frac{1}{3} = \frac{2}{9}$.
Si on effectue $n = 100$ tirages avec remise d'une boule, on observe la réalisation de $X_1$ , $X_2$ , ... , $X_n$ , variables aléatoires indépendantes de même loi que $X$. On dit que l'on a un échantillon aléatoire simple de taille $n = 100$ de loi de Bernoulli de paramètre $p = \frac{2}{3}$.
Le nombre de "boules blanches" obtenues en $n = 100$ tirages est la variable aléatoire $\sum_{i=1}^{n} X_i$.
La fréquence de "boules blanches" obtenue est la variable aléatoire : $F_n = \frac{\sum_{i=1}^{n} X_i}{n}$.

C'est ce point de vue qui est suivi pour le travail de simulation de ces situations. Remarquons qu'à ce stade, il n'est pas nécessaire de connaître les lois de probabilité de ces variables aléatoires. Il faut seulement savoir simuler des 1 et des 0, avec probabilité $p$ et $1-p$.

Ayant procédé par répétitions indépendantes d'expériences, $nF_n = \sum_{i=1}^{n} X_i$ suit la loi Binomiale $\mathcal{B}(n,p) = \mathcal{B}\left(100; \frac{2}{3}\right)$.

**Loi Binomiale** $\mathcal{B}(n,p)$ (vue en première)
Soit $(\Omega, \mathcal{A}, P)$ un espace probabilisé.
Une variable aléatoire $X$ suit la loi de Binomiale des paramètres $n \in \mathbb{N}^*$ et $p \in ]0,1[$, que l'on note $\mathcal{B}(n,p)$, si et seulement si $X$ est à valeurs dans $\{0,1,\ldots,n\}$, et pour tout $P(X=k) = \binom{n}{k} p^k (1-p)^{n-k}$.
On a alors $E(X) = np$ et $Var(X) = np(1-p)$.

Revenant à $F_n$, on a alors $nE(F_n) = E(nF_n) = np$ et $n^2 Var(F_n) = Var(nF_n) = np(1-p)$, d'où $E(F_n) = p = \frac{2}{3}$ et $Var(F_n) = \frac{p(1-p)}{n} = \frac{2}{9n}$.
On constate donc que lorsqu'on augmente la taille $n$ de l'échantillon, l'espérance de $F_n$ reste constante, alors que la variance diminue.

Ce que l'on peut observer sur les simulations (en seconde), et qui est confirmé par la notion de variable aléatoire (en première et terminale), c'est que pour $n$ répétitions indépendantes d'expériences de Bernoulli :
- différents échantillons de taille $n$ peuvent donner différentes fréquences $f_n$ d'apparition du nombre 1 ;
- ces différentes fréquences $f_n$ fluctuent autour de la valeur $p$, en restant "presque toutes" dans un intervalle centré en $p$.

## 1. Intervalle de fluctuation : en seconde, en première, en terminale

### 1.1. En seconde : simulations et prise de décision

Conjecture à partir des simulations :
$F_n$ appartient à l'intervalle de fluctuation $\left[ p - \frac{1}{\sqrt{n}} \ ; \ p + \frac{1}{\sqrt{n}} \right]$ avec une probabilité d'au moins $0,95$.
Conditions d'application : $n \geq 25$, $p$ compris entre 0, 2 et 0, 8, seuil 95%.
Sur les simulations, on observe que $f_n$ est dans l'intervalle $\left[ p - \frac{1}{\sqrt{n}} \ ; \ p - \frac{1}{\sqrt{n}} \right]$ pour environ 95% des échantillons.

Utilisation de l'intervalle de fluctuation pour la prise de décision :
- si $p$ est connue, et que $f_n$ n'appartient pas à $\left[ p - \frac{1}{\sqrt{n}} \ ; \ p + \frac{1}{\sqrt{n}} \right]$, on considère que l'échantillon n'est pas représentatif ; l'observation de $f_n$ n'est pas compatible avec la valeur de $p$, au sens où une observation en dehors de l'intervalle de fluctuation ne s'obtient que pour environ 5% des échantillons.
- si l'on fait une hypothèse sur sa valeur de $p$, disons $p = p_0$, et que $f_n$ n'appartient pas à $\left[ p_0 - \frac{1}{\sqrt{n}} \ ; \ p_0 + \frac{1}{\sqrt{n}} \right]$, on considère que l'observation de $f_n$ n'est pas compatible avec la valeur $p_0$ supposée de $p$, que l'on rejettera avec un risque d'erreur de 5%.

Exemple d'application de l'intervalle de fluctuation
Dans une certaine espèce de rongeur, on a compté 206 mâles sur 400 naissances. Cela est-il conforme à l'hypothèse d'équiprobabilité male/femelle à chaque naissance ?
On peut considérer la situation suivante.
Population : les rongeurs d'une certaine espèce.
Variable : le sexe, à deux modalités (mâle et femelle), représenté par une variable aléatoire de loi de Bernoulli $\mathcal{B}(p)$, où $p$ est la proportion de mâles dans la population ; on a ainsi $P(X=1) = p$ et $P(X=0) = 1-p$.

Echantillon $(X_1, X_2, \ldots, X_n)$ de taille $n = 400$ de $X$.
Observation de l'échantillon : $(x_1, x_2, \ldots, x_n) = (1, 1, 0, 1, \ldots, 0)$.
Estimateur de la proportion $p$ : $F_n = \dfrac{\sum_{i=1}^{n} X_i}{n}$, proportion (ou fréquence) de mâles dans l'échantillon, où $\sum_{i=1}^{n} X_i$ représente le nombre de mâles de l'échantillon.
Estimation ponctuelle de la proportion $p$ : $f_n = \dfrac{\sum_{i=1}^{n} x_i}{n} = \dfrac{206}{400} = 0.515$, fréquence (ou proportion) de mâles dans l'observation de l'échantillon.
Supposons l'équiprobabilité male/femelle à chaque naissance, autrement dit que $p = 0,5$.
Pour un échantillon de $n = 400$ naissances, l'intervalle de fluctuation de $F_n$ est
$\left[ p - \frac{1}{\sqrt{n}} \ ; \ p + \frac{1}{\sqrt{n}} \right] = \left[ 0.5 - \frac{1}{\sqrt{400}} \ ; 0.5 + \frac{1}{\sqrt{400}} \right] = \left[ 0.45 \ ; \ 0.55 \right]$.
Notre observation $f_n = 0.515$ appartient à l'intervalle de fluctuation : elle est donc conforme à l'hypothèse $p = 0,5$, qui n'est donc pas rejetée au rique 5%.

### 1.2. En première : avec la loi Binomiale

*On s'appuie ici sur le document d'accompagnement qui précise le contenu « Utilisation de la loi binomiale pour une prise de décision à partir d'une fréquence » et la capacité correspondante, « Exploiter l'intervalle de fluctuation à un seuil donné, déterminé à l'aide de la loi binomiale, pour rejeter ou non une hypothèse sur une proportion », des programmes du lycée.*

Considérons une variable aléatoire $X$ de loi Binomiale $\mathcal{B}(n, p)$. Cette variable aélatoire est à valeurs entières dans l'intervalle $[0, n]$.
On cherche à partager l'intervalle $[0, n]$, où $X$ prend ses valeurs, en trois intervalles $[0, a-1]$, $[a, b]$ et $[b+1, n]$ de sorte que $X$ prenne ses valeurs dans chacun des intervalles extrêmes avec une probabilité proche de 0,025, sans dépasser cette valeur.
En tabulant les probabilités cumulées $P(X \leq k)$, pour $k$ allant de 0 à $n$, il suffit de déterminer le plus petit entier $a$ tel que $P(X \leq a) > 0,025$ et le plus petit entier $b$ tel que $P(X \leq b) \geq 0,975$, c'est-à-dire $P(X > b) \leq 0,025$. Autrement dit, $a$ est le plus grand entier tel que $P(X < a) \leq 0.25$. On observe aussi que $a < b$.
On a ainsi $P((X < a) \cup (X > b)) = P(X < a) + P(X > b) \leq 0.05$
et donc $P(a \leq X \leq b) = P\left(\overline{(X < a) \cup (X > b)}\right) \geq 0.95$, en étant "assez proche" de 0.95.
Comme $F_n = \frac{X}{n}$, on a ainsi $P\left(\frac{a}{n} \leq F_n \leq \frac{b}{n}\right) \geq 0.95$, en étant "assez proche" de 0.95.
La règle de décision est la suivante : si la fréquence observée $f_n$ appartient à l'**intervalle de fluctuation à 95 %** $\left[ \frac{a}{n}, \frac{b}{n} \right]$, on considère que l'hypothèse selon laquelle la proportion est $p$ dans la population n'est pas remise en question et on l'accepte ; sinon, on rejette l'hypothèse selon laquelle cette proportion vaut $p$.
Pour $n \geq 30$, $np \geq 5$ et $n(1-p) \geq 5$, on observe que l'intervalle de fluctuation $\left[ \frac{a}{n}, \frac{b}{n} \right]$ est sensiblement le même que l'intervalle $\left[ p - \frac{1}{\sqrt{n}} \ , \ p + \frac{1}{\sqrt{n}} \right]$ proposé dans le programme de seconde.

Utilisation pour la prise de décision : analogue à ce qui est fait en seconde

### 1.3. En terminale : avec la loi Normale

On s'appuie sur le fait que si $n$ est suffisamment grand, $U = \dfrac{F_n - p}{\sqrt{\frac{p(1-p)}{n}}}$ suit approximativement la loi normale $\mathcal{N}(0; 1)$.
On détermine le réel $u_\alpha$ tel que $P(-u_\alpha \leq U \leq u_\alpha) = 1 - \alpha$. Pour $\alpha = 5\%$, on a $u_{0.05} = 1.96$.

On en déduit l'intervalle de fluctuation asymptotique $IF_p = \left[ p - \sqrt{\frac{p(1-p)}{n}}\, u_\alpha \,;\, p + \sqrt{\frac{p(1-p)}{n}}\, u_\alpha \right]$ au niveau $1-\alpha$.

Conditions d'application : $n \geq 30$, $np \geq 5$ et $n(1-p) \geq 5$, seuil $1-\alpha$.

Utilisation pour la prise de décision : analogue à ce qui est fait en seconde et première.

Lien avec l'intervalle de fluctuation au seuil 95% donné en seconde :

- pour tout $p \in [0,1]$, on a $0 \leq p(1-p) \leq \frac{1}{4}$, $0 \leq \sqrt{p(1-p)} \leq \frac{1}{2}$, $\sqrt{\frac{p(1-p)}{n}}\, u_\alpha \leq \frac{1.96}{2}\frac{1}{\sqrt{n}} \leq \frac{1}{\sqrt{n}}$ et donc on a l'inclusion d'événements $IF_p \subset IF'_p = \left[ p - \frac{1}{\sqrt{n}} \,;\, p + \frac{1}{\sqrt{n}} \right]$ et donc $P(F_n \in IF'_p) \geq P(F_n \in IF_p) = 1-\alpha = 0.95$ : $IF'_p = \left[ p - \frac{1}{\sqrt{n}} \,;\, p + \frac{1}{\sqrt{n}} \right]$ est un intervalle de fluctuation de $F_n$ de niveau (supérieur ou égal à) $1-\alpha = 0.95$.

- pour tout $p \in [0.2, 0.8]$, on a $0.16 \leq p(1-p) \leq 0.25$, $0.4 \leq \sqrt{p(1-p)} \leq 0.5$ et donc $0.784\frac{1}{\sqrt{n}} = 0.4 \times 1.96\frac{1}{\sqrt{n}} \leq \sqrt{\frac{p(1-p)}{n}}\, u_\alpha \leq 0.5 \times 1.96\frac{1}{\sqrt{n}} = 0.98\frac{1}{\sqrt{n}}$.

## 2. Intervalle de confiance : en seconde et en terminale

### 2.1. En seconde

L'appartenance de $F_n$ à l'intervalle de fluctuation $\left[ p - \frac{1}{\sqrt{n}} \,;\, p + \frac{1}{\sqrt{n}} \right]$ équivaut à l'appartence de $p$ à $\left[ F_n - \frac{1}{\sqrt{n}} \,;\, F_n + \frac{1}{\sqrt{n}} \right]$.

Ainsi, l'intervalle de confiance $\left[ F_n - \frac{1}{\sqrt{n}} \,;\, F_n + \frac{1}{\sqrt{n}} \right]$ contient $p$ avec une probabilité d'au moins 0,95.

Sur les simulations, on observe que $p$ est dans l'intervalle $\left[ f_n - \frac{1}{\sqrt{n}} \,;\, f_n - \frac{1}{\sqrt{n}} \right]$ pour environ 95% des échantillons.

Utilisation de l'intervalle de confiance pour l'estimation de $p$ inconnue : à partir de l'observation $f_n$, on obtient une fourchette contenant $p$ au niveau de confiance 0,95.

Exemple d'intervalle de confiance

Reprenons l'exemple précédent, pour lequel on a observé $f_n = 0.515$ sur un échantillon de taille $n = 400$.

Intervalle de confiance de la proportion $p$ :

$$\left[ f_n - \frac{1}{\sqrt{n}} \,;\, f_n - \frac{1}{\sqrt{n}} \right] = \left[ 0.515 - \frac{1}{\sqrt{400}} \,;\, 0.515 + \frac{1}{\sqrt{400}} \right] = \left[ 0.465 \,;\, 0.565 \right].$$

### 2.2. En terminale

Le résultat vu en seconde est validé par le calcul de l'intervalle de fluctuation $IF'_p$ vu ci-dessus.

L'intervalle de confiance $\left[ f_n - \sqrt{\frac{f_n(1-f_n)}{n-1}}\, u_\alpha \;,\; f_n + \sqrt{\frac{f_n(1-f_n)}{n-1}}\, u_\alpha \right]$ vu dans le supérieur n'est pas au programme.

# 3. Nouvelles notions en terminale

## 3.1. Variable centrée réduite

Une variable aléatoire $X$ est dite centrée réduite si son espérance est nulle et son écart-type vaut 1, ce qui s'écrit $E(X) = 0$ et $\sigma(X) = 1$

Soit maintenant une variable aléatoire $X$ telle que $E(X) = m$ et $\sigma(X) = \sigma \neq 0$ ; on a donc $\sigma > 0$.

La variable aléatoire $X - m$ a une espérance nulle.

La variable aléatoire $\frac{X}{\sigma}$ a un écart-type égal à 1.

La variable aléatoire $Z = \frac{X - m}{\sigma}$ a une espérance nulle et un écart-type égal à 1. On dit que $Z$ est la variable aléatoire centrée réduite associée à $X$.

Valeurs de $X$ et valeurs de $Z$.

Si $X$ prend ses valeurs entre $a$ et $b$, $X - m$ prend ses valeurs entre $a - m$ et $b - m$ et $Z = \frac{X - m}{\sigma}$ prend ses valeurs entre $\frac{a - m}{\sigma}$ et $\frac{b - m}{\sigma}$.

On a alors, pour toute valeur $k$ de $X$, $P(X = k) = P\left(Z = \frac{k - m}{\sigma}\right)$.

Exemple avec $X_n$ de loi Binomiale $\mathcal{B}(n; p)$, avec $p \in ]0, 1[$

On a $E(X_n) = np$ et $\sigma(X_n) = \sqrt{np(1-p)}$.

Comme $X_n$ prend ses valeurs entre 0 et $n$, $X_n - np$ prend ses valeurs entre $-np$ et $n - np$ et $Z_n = \frac{X_n - np}{\sqrt{np(1-p)}}$ prend ses valeurs entre $\frac{-np}{\sqrt{np(1-p)}} = -\sqrt{\frac{np}{1-p}}$ et $\frac{n - np}{\sqrt{np(1-p)}} = \sqrt{\frac{n(1-p)}{p}}$.

On a alors, pour tout entier $k$ compris entre 0 et $n$, $P(X_n = k) = P\left(Z_n = \frac{k - m}{\sigma}\right) = \binom{n}{k} p^k (1-p)^{n-k}$.

On a $E(Z_n) = 0$ et $\sigma(Z_n) = 1$.

Visualisation graphique sur la feuille Excel.

Intérêt de centrer et réduire : l'espérance et l'écart-type de $Z_n$ ne dépendent plus de ceux de $X_n$.

## 3.2. Loi normale centrée réduite $\mathcal{N}(0; 1)$

Considérons une variable aléatoire $X_n$ de loi Binomiale $\mathcal{B}(n; p)$, avec $n = 100$ et $p = 0,5$, et la variable aléatoire $Z_n = \frac{X_n - 50}{5}$ centrée réduite associée à $X_n$.

Construisons le diagramme en bâtons de la loi de probabilité de $Z_n$.

Construisons un histogramme associé : à chaque valeur $k$ on fait correspondre un rectangle dont l'aire est égale à $P(X_n = k) = P\left(Z_n = \frac{k - 50}{5}\right)$ et de base de longueur $\frac{1}{5}$ centrée sur $\frac{k - 50}{5}$.

Les sommets des bâtons, comme les bords supérieurs des rectangles, font apparaître une courbe en cloche, l'aire située sous cette courbe étant voisine de l'aire de la réunion des rectangles.

De Moivre a découvert que cette courbe représente la fonction définie par $f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$.

On obtiendrait par exemple :

$$P(45 \leq X_n \leq 60) = \sum_{k=45}^{60} P(X_n = k) = \sum_{k=45}^{60} \binom{100}{k} \left(\frac{1}{2}\right)^{100} \simeq 0.8467.$$

$$= P\left(\frac{45 - 50}{5} \leq \frac{X_n - 50}{5} \leq \frac{60 - 50}{5}\right) = P(-1 \leq Z_n \leq 2) = \sum_{k=45}^{60} P\left(Z_n = \frac{k - 50}{5}\right)$$

$$\simeq \int_{-1}^{2} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx = 0.8186.$$

Conjecture.

Lorsque $n$ devient grand, à $p$ fixé, la largeur des rectangles $\frac{1}{\sqrt{np(1-p)}}$ est de plus en plus petite.

L'aire correspondant à $P(a \leq Z_n \leq b)$ semble se rapprocher de l'aire située sous la courbe en cloche, c'est-à-dire de $\int_a^b \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx$.

**Théorème de Moivre-Laplace**

On suppose que pour tout entier $n$, la variable aléatoire $X_n$ suit la loi Binomiale $\mathcal{B}(n,p)$.

On pose $Z_n = \frac{X_n - np}{\sqrt{np(1-p)}}$, variable centrée réduite associée à $X_n$.

Alors, pour tous réels $a$ et $b$ tels que $a < b$, on a : $\lim_{n\to+\infty} P(a \leq Z_n \leq b) = \int_a^b \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx$.

**Définition**. **Loi normale centrée réduite** $\mathcal{N}(0;1)$.

Posons $f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$ pour tout réel $x$

Une variable aléatoire $X$ suit la loi normale $\mathcal{N}(0;1)$ si, pour tous réels $a$ et $b$ tels que $a < b$, on a : $P(a \leq X \leq b) = \int_a^b f(x)dx = \int_a^b \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx$.

La fonction $f$ est appelée fonction de densité (ou densité de probabilité) de la loi $\mathcal{N}(0;1)$.

On a $E(X) = 0$, $Var(X) = 1$ et $\sigma(X) = 1$.

Conséquence sur la loi de $F_n$.

Rappelons que $F_n = \frac{X_n}{n}$. On a alors $Z_n = \frac{X_n - np}{\sqrt{np(1-p)}} = \frac{F_n - p}{\sqrt{\frac{p(1-p)}{n}}}$.

Le théorème de Moivre-Laplace indique alors que, pour $n$ assez grand, $Z_n = \frac{F_n - p}{\sqrt{\frac{p(1-p)}{n}}}$ suit approximativement la loi normale $\mathcal{N}(0;1)$.

**Théorème**

Si $X$ est une variable aléatoire suivant la loi normale $\mathcal{N}(0;1)$, alors, pour tout $\alpha \in ]0,1[$, il existe un unique réel positif $u_\alpha$ tel que $P(-u_\alpha \leq U \leq u_\alpha) = 1 - \alpha$.

### 3.3. Lois normales $\mathcal{N}(\mu;\sigma)$

**Définition**.

Une variable aléatoire $X$ suit la loi normale $\mathcal{N}(\mu;\sigma)$ si la variable aléatoire $Z = \frac{X-\mu}{\sigma}$ suit la loi normale $\mathcal{N}(0;1)$.

C'est une loi à densité, en ce sens qu'il existe une fonction $g$ définie sur $\mathbb{R}$ telle que, pour tous réels $a$ et $b$ tels que $a < b$, on a : $P(a \leq X \leq b) = \int_a^b g(x)dx$.

On a $E(X) = \mu$, $Var(X) = \sigma^2$ et $\sigma(X) = \sigma$.

### 3.4. Loi uniforme sur un intervalle $[a,b]$

**Définition**.
Soient deux réels $a$ et $b$ tels que $a < b$.
Une variable aléatoire $X$ suit la loi uniforme sur $[a,b]$ si pour tous réels $c$ et $d$ tels que $a \leq c < d \leq b$, on a $P(c \leq X \leq d) = \int_c^d f(x)dx$, avec $f(x) = \frac{1}{b-a}$ pour tout $x$ dans $[a,b]$.
Ce qui donne $P(c \leq X \leq d) = \frac{d-c}{b-a}$.

**Proposition**
Si une variable aléatoire $X$ suit la loi uniforme sur $[0,1]$, alors la variable aléatoire $Y = (b-a)X + a$ suit la loi uniforme sur $[a,b]$.

Ce résultat montre que l'on peut simuler la loi uniforme sur $[a,b]$ à partir la loi uniforme sur $[0,1]$.
On a même un résultat plus général.

**Proposition**
Soit $X$ une variable aléatoire et $F_X$ sa fonction de répartition.
Si $F_X$ est continue et strictement croissante, alors
- $Y = F_X(X)$ est une variable aléatoire de loi uniforme sur $[0,1]$.
- si $U$ suit la loi uniforme sur $[0,1]$, alors $Z = F_X^{-1}(U)$ suit la même loi que $X$.

On peut utiliser ce dernier résultat pour simuler la loi exponentielle.

### 3.5. Loi exponentielle

**Définition**.
Soit un réel $\lambda > 0$.
Une variable aléatoire $X$ suit la loi exponentielle de paramère $\lambda$ si pour tous réels $c$ et $d$ tels que $0 \leq c < d$, on a : $P(c \leq X \leq d) = \int_c^d f(x)dx$, avec $f(x) = \lambda e^{-\lambda x}$ pour tout $x \geq 0$.
Ce qui donne $P(c \leq X \leq d) = e^{-\lambda c} - e^{-\lambda d}$, et en particulier $P(X \leq d) = 1 - e^{-\lambda d}$.

## 4. Méthode de Monte Carlo et applications

Voir [D-M].
La méthode de Monte Carlo peut être définie comme toute technique numérique de résolution de problème au moyen d'un modèle stochastique dans lequel on utilise des nombres aléatoires.
Développée vers 1949, elle est attribuée à John von Neumann et Stanislav Ulam (mathématicien américain. La référence au casino rappelle que la roulette permet de générer des nombres aléatoires.
Elle peut être utilisée pour :
- l'estimation d'une surface
- l'estimation d'une intégrale
- les problèmes de files d'attente
- la gestion des stocks,
- le rendement d'un investissement.

### 4.1. Estimation d'une aire

Exemple introductif.
Supposons que nous voulions estimer l'aire d'un carré de coté 0,5.
Bien sûr, on sait que cette aire est 0.25. Comment trouver une estimation de ce résultat.
Ce carré $C = [0, 0.5] \times [0, 0.5]$ est évidemment inclus dans le carré $\Omega = [0, 1] \times [0, 1]$.
Si l'on place $n$ points aléatoirement dans le carré $\Omega$, certains seront dans $C$, d'autres non.

Les coordonnées d'un point aléatoire correspondent à deux variables aléatoires indépendantes $X$ et $Y$ de même loi Uniforme sur $[0, 1]$.
On a alors $p = P((X, Y) \in C) = P((0 \leq X \leq 0.5) \cap (0 \leq Y \leq 0.5)) = \cdots = 0.5 \times 0.5 =$ aire de $C$.

Ainsi, lors de l'expérience de Bernoulli "placer un point dans $\Omega$", l'événement $A$ "le point est dans $C$" a pour probabilité $p =$ aire de $C$.
Un échantillon de $n$ points, obtenu par $n$ répétition indépendantes de l'expérience, nous fournira alors une estimation $f_n$ de $p$.

Simulation
Nous devons simuler deux échantillons de taille $n$ de la loi Uniforme sur $[0, 1]$. Le premier donnera les abscisses, le second les ordonnées des $n$ points aléatoires.
Pour chaque point, on regarde s'il est dans $C$ ou pas. On compte le nombre $n_C$ de points dans $C$ puis on obtient $f_n = \frac{n_C}{n}$.
Il s'agit de la méthode dite "du rejet"

Cas général
1) Si $R$ est une région de $\Omega = [0, 1] \times [0, 1]$, alors on a $P((X, Y) \in R) =$ aire de $R$. On procède alors comme dans l'exemple précédent.
2) Si $R$ est une région de $\Omega = [a, b] \times [c, d]$, alors on a $P((X, Y) \in R) = \frac{\text{aire de } R}{(b-a)(d-c)}$.
On procède façon analogue, en simulant deux échantillons de taille $n$ de la loi Uniforme sur $[a, b]$ et $[c, d]$.

### 4.2. Estimation d'une intégrale

Si $h$ est une fonction positive, alors $\int_a^b h(x)dx$ est l'aire située sous la courbe représentative de $f$. On peut donc appliquer l'estimation précédente.

Méthode de l'espérance
On peut estimer l'intégrale $\int_a^b h(x)dx$ dès lors que $h = gf$ avec $f$ densité de probabilité.
Considérons une variable aléatoire $X$ de densité $f$ et posons $\widetilde{g}(x) = \begin{array}{c} g(x) \text{ si } x \in [a, b] \\ 0 \text{ sinon} \end{array}$.
On a alors $E(\widetilde{g}(X)) = \int_{-\infty}^{+\infty} \widetilde{g}(x)f(x)dx = \int_a^b g(x)f(x)dx = \int_a^b h(x)dx$.
Le problème est donc d'estimer $E(\widetilde{g}(X))$.
A partir d'un échantillon $(X_1, \ldots, X_n)$ de taille $n$ de $X$, on a l'estimateur $\frac{1}{n}\sum_{i=1}^{n} \widetilde{g}(X_i)$.

Exemple : estimation de $\int_0^{+\infty} xe^{-x}dx$.
Cette intégrale n'est autre que $E(X)$, avec $X$ de loi exponentielle de paramètre 1.
Pour la simulation, on peut utiliser le fait que si $U$ suit la loi uniforme sur $[0, 1]$, alors $X = -\ln(1 - U)$ suit la loi exponentielle de paramètre 1.

# 5. Applications diverses

## 5.1. Marche aléatoire sur un graphe à trois sommets

Voir exercice 8 de la deuxième liste d'exercices.

Une puce se déplace indéfiniment entre trois points $A$, $B$ et $C$.

Au départ (étape 0), elle est en $A$. A chaque étape, elle quitte sa position et gagne indifféremment l'un des deux autres points.

On suppose construit un espace probabilisé $(\Omega, \mathcal{A}, P)$ modélisant cette suite infinie de déplacements.

Pour tout entier naturel $n$, on considère l'événement $A_n$ (respectivement $B_n$ et $C_n$) : "la puce est en $A$ (respectivement $B$ et $C$)" à l'issue de la $n$-ème étape, et la probabilité $\alpha_n$ (respectivement $\beta_n$ et $\gamma_n$) de l'événement $A_n$ (respectivement $B_n$ et $C_n$).

On pose $\alpha_0 = 1$, $\beta_0 = 0$ et $\gamma_0 = 0$.

L'objectif est de calculer $\alpha_n$, $\beta_n$ et $\gamma_n$ pour tout entier $n \geq 0$ ; autrement dit la probabilité que la puce se retrouve en $A$, $B$ et $C$ au bout de $n$ déplacements.

A l'aide de la formule des probabilités totales (utilisant des probabilités conditionnelles), on peut démontrer que pour tout entier $n \geq 0$, $\begin{cases} \alpha_{n+1} = \frac{1}{2}\beta_n + \frac{1}{2}\gamma_n \\ \beta_{n+1} = \frac{1}{2}\alpha_n + \frac{1}{2}\gamma_n \\ \gamma_{n+1} = \frac{1}{2}\alpha_n + \frac{1}{2}\beta_n \end{cases}$.

Ce qui peut s'écrire matriciellement $X_{n+1} = \begin{pmatrix} \alpha_{n+1} \\ \beta_{n+1} \\ \gamma_{n+1} \end{pmatrix} = M \begin{pmatrix} \alpha_n \\ \beta_n \\ \gamma_n \end{pmatrix} = MX_n$, avec

$M = \begin{pmatrix} 0 & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & 0 & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} & 0 \end{pmatrix}$.

Les termes de la matrice $M$ sont les probabilités de transition d'un point du graphe à un autre. Par exemple, $P(A_{n+1}/A_n) = 0$ et $P(A_{n+1}/B_n) = \frac{1}{2}$.

On démontre alors, par récurrence, que pour tout entier $n \geq 0$, $X_n = M^n X_0$.

On est donc ramener au calcul de $M^n$, ce qui utilise en général la diagonalisation/trigonalisation/... de $M$.

Sur le cas étudié qui est relativement simple, on ppeut repartir du système et trouver des relations de récurrence vérifiées séparément par chacune des trois suites.

Ici, on a, pour tout entier $n \geq 0$, $\beta_n = \gamma_n$ et $\alpha_{n+1} = \frac{1}{2}(1 - \alpha_n)$.

L'étude de la suite $(\alpha_n)_{n\geq 0}$, qui est arithmético-géométrique, conduit à l'expression $\alpha_n = \frac{2}{3}\left(-\frac{1}{2}\right)^n + \frac{1}{3}$.

On en déduit alors que $\beta_n = \gamma_n = \alpha_{n+1} = \frac{2}{3}\left(-\frac{1}{2}\right)^{n+1} + \frac{1}{3} = -\frac{1}{3}\left(-\frac{1}{2}\right)^n + \frac{1}{3}$.

Une activité de simulation pour estimer $p = \alpha_N$.

On se fixe une valeur de $N$, par exemple $N = 10$.

Utiliser le tableur ou la calculatrice pour :

- pour simuler les $N = 10$ déplacements de la puce ;
- répéter $n = 100$ fois les $N = 10$ déplacements ;
- obtenir la fréquence de réalisation de l'événement $A_N$ ;
- comparer avec la probabilité de l'événement $A_N$.

## 5.2. Modèle de diffusion d'Ehrenfest

Extrait de [M-B-J] sur http://edutice.archives-ouvertes.fr/docs/00/05/45/65/PDF/co24th2.pdf

Le modèle

C'est un modèle de diffusion d'un gaz à travers une paroi proposé par les physiciens Ehrenfest (Mr et Mme) au début du siècle dernier.

Une boite séparée en 2 compartiments A et B contient au total $N$ particules. A chaque top d'une horloge, une particule et une seule, choisie au hasard parmi les $N$, change de compartiment.

A l'instant initial toutes les particules sont en A.

Dans ce modèle il est important de remarquer plusieurs points :

- la probabilité pour une particule donnée de passer de A à B, ou de B à A est la même, égale à $1/N$ ;
- cette probabilité ne dépend pas du temps ;
- le comportement d'une particule est indépendant de celui des autres particules.

L'objectif

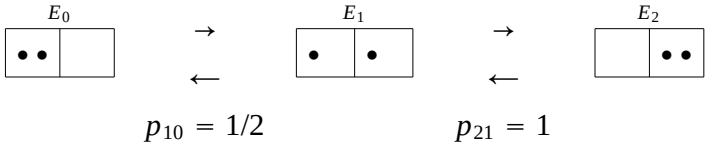
Etudier l'évolution de la répartition des particules au bout d'un grand nombre de déplacements de particules.

Pour les physiciens Ehrenfest l'un des objectifs était de lever le «paradoxe » de l'irréversibilité.

Ils voulaient donc montrer comment, à partir de particules aux évolutions réversibles, on pouvait obtenir, en combinant ces évolutions, une situation macroscopique irréversible.

Un exemple : $N = 2$ particules.

Il y a $N + 1 = 3$ états possibles : $E_0$ $E_1$ $E_2$




D'où la matrice de transition $\begin{pmatrix} 0 & 1 & 0 \\ 1/2 & 0 & 1/2 \\ 0 & 1 & 0 \end{pmatrix}$ avec $p_{i,j}$ = probabilité de passer de l'état $i$ et à l'état $j$.

On peut procéder comme dans l'exemple précédent pour obtenir les probabilités d'être dans chacun des trois états après $n$ déplacements.

La simulation

La simulation peut se réaliser avec le tableur Excel et le langage de programmation associé VBA (visual basic).

Le graphique (nuage de points) et le choix aléatoire de la particule qui va se déplacer (par la fonction ENT(ALEA( )*$N$+1) peuvent se faire directement sous Excel sans avoir recours à la programmation.

Par contre, pour que la même opération se répète un grand nombre de fois, il est plus pratique de recourir à la programmation (utilisation d'une boucle : For i = 1 to n …….Next i ).

L'analyse des résultats

L'irréversibilité

Il semble impossible, en regardant la simulation pour $N = 100$, de revenir à l'état initial (toutes les particules dans le compartiment A).

Le système s'équilibre apparemment autour de la position 50/50.

Cet état d'équilibre paraît encore plus stable pour un nombre de particules plus grand ($N = 1000$).

L'évolution de l'urne d'Ehrenfest est irréversible puisque le système s'équilibre dans un état différent de l'état initial.

On peut remarquer toutefois que, lorsque le nombre de particules est très faible ($N = 2, 3$ ……) le comportement de l'urne est totalement réversible.

Il faut donc l'accumulation d'un grand nombre de particules réversibles pour créer de l'irréversibilité.

La simulation «tempsretour.xls » permet de constater le comportement réversible de l'urne pour $N$ petit.

Le temps de retour

Un système tel que l'urne d'Ehrenfest possède $N+1$ états, chaque état correspondant au nombre $k$ de particules dans le volume A, $k = 0, \ldots, N$, qui peuvent être choisies de $\binom{N}{k}$ façons.

Ce qui donne $\sum_{k=0}^{N} \binom{N}{k} = 2^N$ situations possibles pour l'urne, chaque situation ne conduisant pas forcément à un état différent.

Pour un nombre $N$ de particules très grand (ce qui correspond à la réalité), on conçoit aisément, vu le grand nombre de situations possibles, que le retour à l'état initial sera extrêmement rare puisqu'une seule situation conduit à cet état.

Quoiqu'il en soit, on démontre dans l'étude des chaînes de Markov que le retour à l'état initial est quasi-certain pour un tel système.

Nous avons étudié (et simulé informatiquement) le temps de retour pour l'urne d'Ehrenfest. L'espérance du temps de retour à l'état initial pour l'urne d'Ehrenfest est $2^N$.

Pour le nombre de particules traitées dans une situation macroscopique le temps de retour est donc quasiment infini à notre échelle (et même par rapport à l'âge de l'univers).

L'irréversibilité apparente de la physique statistique est donc en grande partie due à la différence entre l'échelle de temps de l'observateur et celle du temps de retour.

Conclusion sur l'irréversibilité

L'irréversibilité de l'urne d'Ehrenfest n'est donc qu'une illusion.

D'une part elle dépend du nombre de particules mises en jeu, et d'autre part elle disparaît si le temps d'observation est illimité.

Cependant, pour les situation macroscopiques usuelles (grands nombres de particules, temps d'observation limité), l'urne présente un comportement irréversible.

La plupart des activités proposées permettent d'évaluer le temps de retour pour se convaincre de l'irréversibilité du phénomène.

### 5.3. Etude du principe du calcul de la pertinence d'une page web (page rank google)

Voir [RIG] sur http://www.discmath.ulg.ac.be/mam/pratique.html : la matrice cachée de Google

## Références

Références - Ouvrages

[CHV] Gérard Chauvat et al, Mathématiques BTS/DUT, Probabilités et statistique

[C-T] Hubert Carnec et Marc Thomas, Itinéraires en Statistiques & Probabilités

[D-M] Yadolah Dodge et Giuseppe Melfi, Premiers pas en simulation

Références - En ligne

[M-B-J] Alain Marie-Jeanne, Frédéric Beau, Michel Janvier - Simulation de l'urne d'Ehrenfest - http://edutice.archives-ouvertes.fr/docs/00/05/45/65/PDF/co24th2.pdf

[RIG] Michel Rigo - La matrice cachée de Google - http://www.discmath.ulg.ac.be/mam/pratique.html

[SN] St@tNet - Les techniques de la statistique - http://www.agro-montpellier.fr/cnam-lr/statnet/index.htm