**Compléments en Statistique**
**Préparation au Capes**
**Université de Rennes 1**

# Intervalles de fluctuations et intervalles de confiance

## Table des matières

# 1 Introduction

On considère un espace de probabilité $(\Omega, \mathcal{F}, \mathbb{P})$.
Une expérience aléatoire est une expérience réalisée selon des règles bien définies mais dont on ne peut pas prédire le résultat de façon certaine.
On considère dans ce document le cas d'une expérience aléatoire qui n'a que deux résultats possibles : succès (on obtient le résultat espéré) et échec (on n'obtient pas le résultat espéré). On peut par exemple prendre l'exemple du jeu "Pile ou Face", de la roulette au casino, du loto, d'une élection entre deux candidats ...
La recherche d'intervalles de confiance pour des lois plus générales sera rapidement abordée dans la section 5.

La probabilité de succès de l'expérience est notée $p$. On répète l'expérience plusieurs fois de façon indépendante. On définit $X_i$ le résultat de la $i^{\text{ème}}$ réalisation :

$$X_i(\omega) = \begin{cases} 1 & \text{si la } i^{\text{ème}} \text{ réalisation est un succès,} \\ 0 & \text{si la } i^{\text{ème}} \text{ réalisation est un échec.} \end{cases}$$

Par conséquent, pour chaque $i \geqslant 1$, $X_i$ suit la loi de Bernoulli $\mathcal{B}(p)$. On en déduit que pour chaque $i \geqslant 1$, $\mathbb{E}[X_i] = p$ et $Var(X_i) = p(1-p)$.
Le nombre total de succès au bout de $n$ réalisations est

$$S_n = \sum_{i=1}^{n} X_i.$$

Il s'agit d'une variable aléatoire à valeurs dans $\{0, 1, \ldots, n\}$ de loi Binomiale $\mathcal{B}(n, p)$ car on répète de façon indépendante $n$ fois la même expérience de Bernoulli.
Plus le nombre de réalisations $n$ est grand plus $S_n$ peut prendre des grandes valeurs. Regardons maintenant la fréquence de succès sur les $n$ réalisations.

# 2 Moyenne empirique

## 2.1 Définition et propriétés

**Definition 1.** *On considère des variables $X_1, X_2, \ldots, X_n$ indépendantes et de même loi.*
*La* **moyenne empirique** *associée, notée $\overline{X}_n$, est définie par*

$$\overline{X}_n = \frac{\sum_{i=1}^{n} X_i}{n}.$$

Dans le cas qui nous intéresse, chaque $X_i$ est le résultat d'une même expérience ayant deux issus. La moyenne empirique est alors aussi appelée **fréquence de succès**. Il s'agit dans ce cas d'une variable aléatoire à valeurs dans $\{0, \frac{1}{n}, \frac{2}{n}, \ldots, 1\}$.

**Propriété 2.** *Soit $X_1, X_2, \ldots, X_n$ des variables indépendantes et de même loi, d'espérance $m$ et de variance $\sigma^2$ finies. Alors $\overline{X}_n$ est une variable aléatoire d'espérance $m$ et de variance $\sigma^2/n$.*

*Démonstration.* Par linéarité de l'espérance et comme les variables $X_i$ sont d'espérance $m$, on a

$$\mathbb{E}[\overline{X}_n] = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}[X_i] = m.$$

Par indépendance des variables $X_i$ et comme les variables $X_i$ sont de variance $\sigma^2$, on a

$$Var(\overline{X}_n) = \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i) = \frac{\sigma^2}{n}.$$

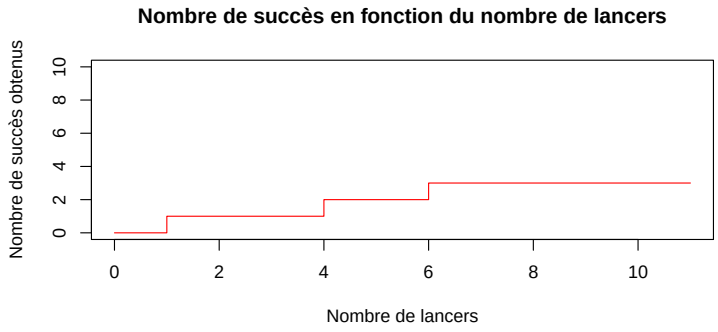
□

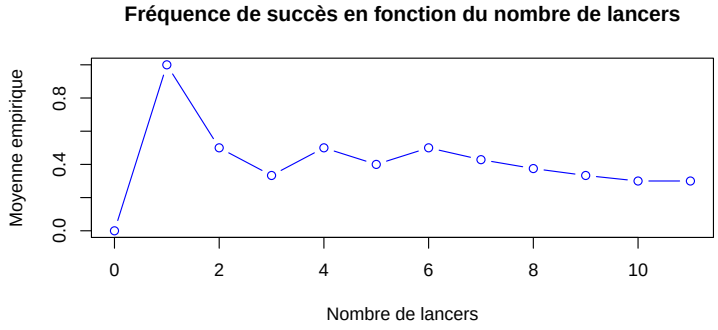
Par conséquent, lorsque les variables $X_i$ suivent la loi de Bernoulli $\mathcal{B}(p)$, $\overline{X}_n$ est une variable aléatoire qui oscille autour de $p$ et dont la variance $\frac{p(1-p)}{n}$ diminue lorsque $n$ grandit, ce qui signifie que pour $n$ grand les oscillations sont d'amplitude de plus en plus faible.

**Exemple.** Considérons l'exemple du jeu "Pile ou Face" avec une pièce bien équilibrée. La probabilité de tomber sur "Pile" est alors $p = 1/2$. Le joueur mise sur "Pile". On répète 10 fois l'expérience et on obtient les résultats suivant :

Pile, Face, Face, Pile, Face, Pile, Face, Face, Face, Face,

Traçons sur des graphiques l'évolution du nombre de "Pile" obtenus et de la fréquence de succès en fonction du nombre de lancers.

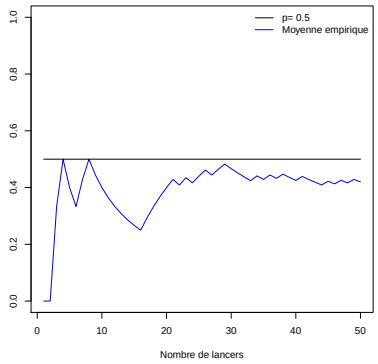


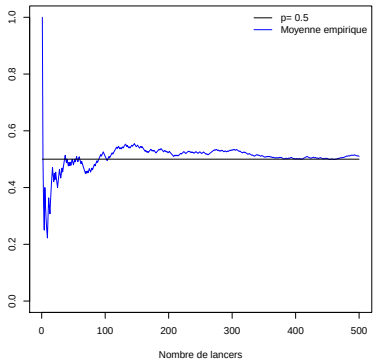
## 2.2 Loi des grands nombres (version faible)

On étudie l'évolution de la moyenne empirique quand on augmente le nombre de réalisations de l'expérience.

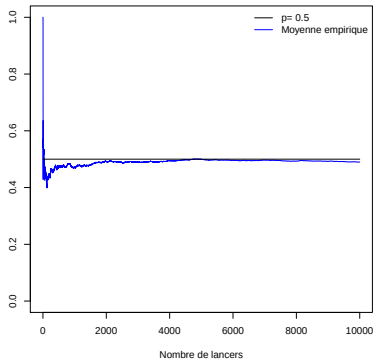
**Exemple.** On reprend l'exemple précédent et on augmente le nombre $n$ de lancers. On obtient les courbes suivantes pour la fréquence du nombre de succès :




pour $n = 50$




pour $n = 500$




pour $n = 10\,000$.

En regardant ces graphiques, on a l'impression que la fréquence de succès converge vers $p = 1/2$ quand $n$ devient très grand.

Cette convergence est formalisée par le théorème de la loi des grands nombres.

**Th eorème 3** (Loi des Grands Nombres)**.**
*On considère $(X_i)_{i\geqslant 1}$ une suite de variables aléatoires indépendantes et de même loi, telle que $\mathbb{E}[|X_1|] < \infty$. On note $m = \mathbb{E}[X_1]$ leur espérance commune.*
*Alors*

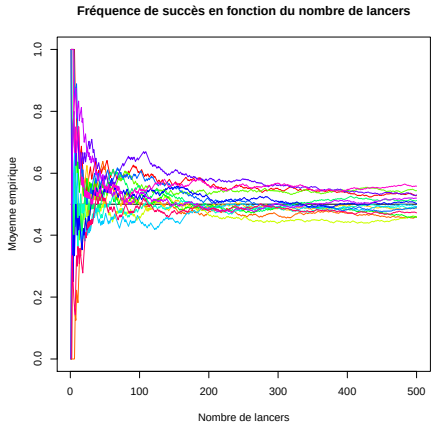
$$\overline{X}_n \text{ converge en probabilité vers } m \text{ lorsque } n \to +\infty.$$

Donc dans la situation qui nous intéresse où les $X_i$ suivent la loi $\mathcal{B}(p)$, la loi de grands nombres nous permet d'affirmer que la fréquence de succès converge vers $p$ lorsque $n$ tend vers l'infini, $p$ étant la probabilité de succès.

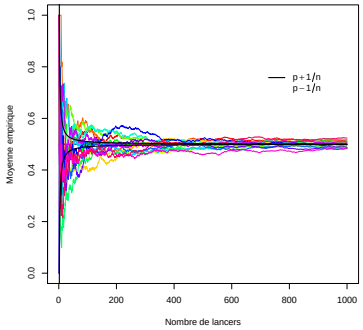
## 2.3 Théorème central limite

On peut alors se demander à quelle vitesse cette convergence a lieu.

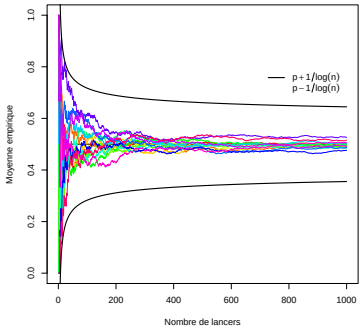
**Exemple.** On reprend notre exemple et on trace plusieurs réalisations de la trajectoire *aléatoire* $n \mapsto \overline{X}_n$ (chaque réalisation ayant une couleur différente sur le graphique). On obtient le résultat suivant



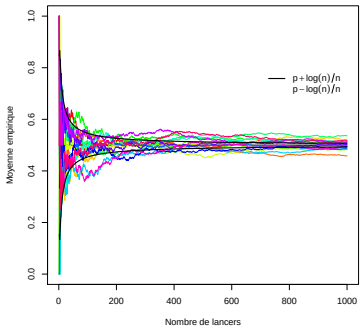

On observe que la convergence est plutôt lente. Essayons différentes fonctionnelles pour évaluer la vitesse de convergence (courbe tracée en noir).



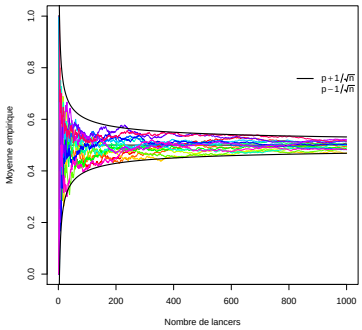

la vitesse semble moins rapide que $n$




la vitesse semble plus rapide que $\ln(n)$



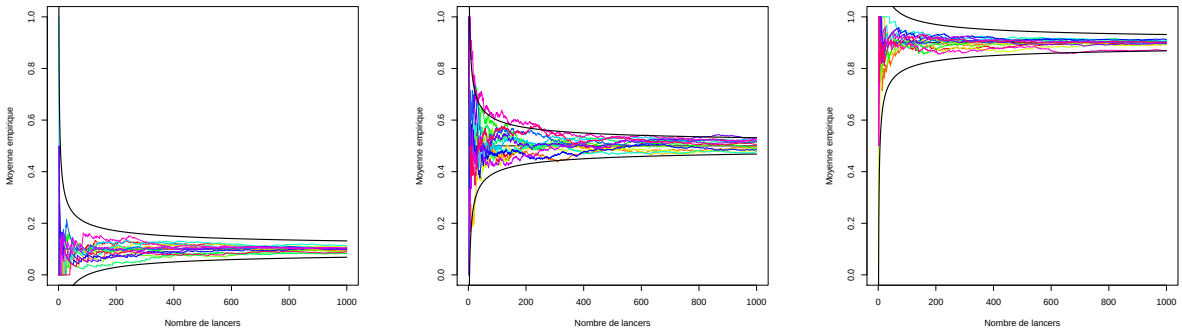

la vitesse semble moins rapide que $\frac{n}{\ln(n)}$




la vitesse semble de l'ordre de $\sqrt{n}$.

Regardons maintenant si la vitesse de convergence dépend peut-être de la valeur de la probabilité $p$ de succès, i.e. regardons si la vitesse $\sqrt{n}$ est toujours satisfaisante lorsqu'on prend différentes valeurs de $p$.

Dans les graphiques ci-dessous, on a tracé plusieurs réalisations de la trajectoire $n \mapsto \overline{X}_n$ et les courbes d'équation $n \mapsto p \pm \frac{1}{\sqrt{n}}$ (en noir), pour différentes valeurs de $p$ :




pour $p = 0.1$, pour $p = 0.5$, pour $p = 0.9$.

La vitesse semble toujours en $\sqrt{n}$ quelque soit la valeur de $p$, même si elle est mieux adaptée lorsque $p = 1/2$.

La vitesse de convergence réelle de la moyenne empirique vers l'espérance est donnée par le théorème central limite.

**Th eorème 4** (Théorème Central Limite)**.**
*On considère $(X_i)_{i\geqslant 1}$ une suite de variables aléatoires indépendantes et de même loi, telle que $\mathbb{E}[X_1^2] < \infty$. On note $m = \mathbb{E}[X_1]$ et $\sigma^2 = Var(X_1)$ leur espérance et leur variance commune.*
*Alors*

$$\sqrt{\frac{n}{\sigma^2}}(\overline{X}_n - m) \text{ converge en loi vers la loi normale centrée réduite } \mathcal{N}(0,1) \text{ lorsque } n \to +\infty.$$

Par conséquent, d'après le théorème central limite, dans la situation qui nous intéresse où les $X_i$ suivent la loi $\mathcal{B}(p)$, la fréquence de succès converge vers $p$ à vitesse $\frac{\sqrt{n}}{\sqrt{p(1-p)}}$.

# 3 Intervalles de fluctuations

On considère toujours la situation d'une expérience aléatoire qui n'a que deux résultats possibles : succès et échec. **La probabilité de succès $p$ est connue**.
On répète l'expérience $n$ fois de façon indépendante et on se demande où se situe la fréquence de succès en fonction du nombre de réalisations $n$.

**Definition 5.** *Soit $X_1, \dots, X_n$ des variables indépendantes de loi de Bernoulli $\mathcal{B}(p)$. Un* **intervalle de fluctuations** *de la fréquence de succès au niveau de confiance $1-\alpha$ est un intervalle déterministe $I_f = [a, b]$, avec $a, b \in \overline{\mathbb{R}}$ tel que*

$$\mathbb{P}(\overline{X}_n \in I_f) = 1 - \alpha.$$

*La quantité $\alpha$ est l'erreur que l'on s'autorise, elle est appelée niveau de risque. Elle est en général petite.*

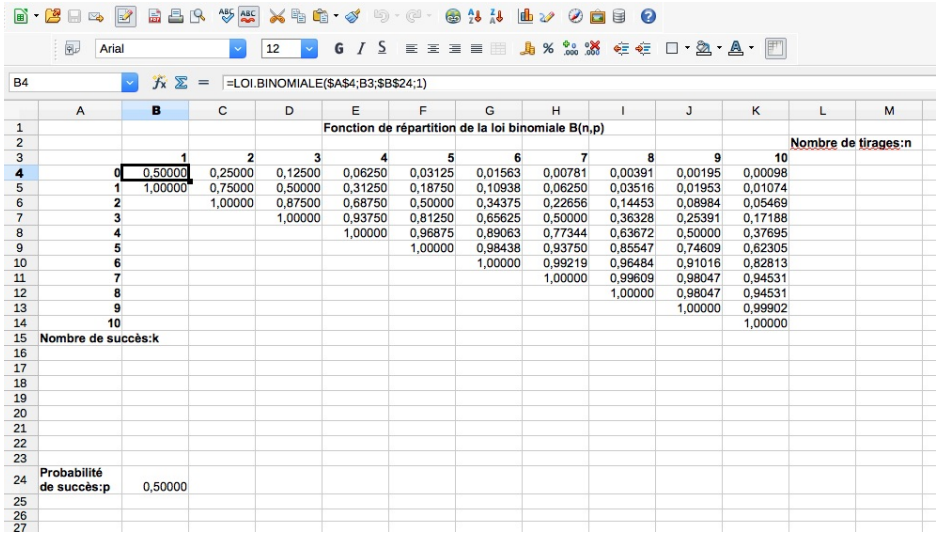
## 3.1 Calcul direct

Dans la situation que l'on considère $S_n = \sum_{i=1}^{n} X_i$ suit la loi Binomiale $\mathcal{B}(n,p)$, qui est une loi bien connue. Par conséquent, pour trouver un intervalle de fluctuations $I_f = [a,b]$ de la fréquence de succès, il faut trouver, à l'aide de la fonction de répartition de la loi binomiale, des réels $a, b$ tels que

$$\begin{aligned}\mathbb{P}(na \leqslant S_n \leqslant nb) &= \mathbb{P}(S_n \leqslant nb) - \mathbb{P}(S_n < na)\\ &= 1-\alpha.\end{aligned}$$

Les valeurs de $a$ et $b$ vont dépendre de $p$, $n$ et de $\alpha$.

À l'aide d'un tableur, on obtient les valeurs de la fonction de répartition de la loi binomiale $\mathcal{B}(n,p)$ pour différentes valeurs de $n$ et de $p$.




| | A | B | C | D | E | F | G | H | I | J | K | L | M |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | | Fonction de répartition de la loi binomiale B(n,p) | | | | | | | | | | | |
| 2 | | | | | | | | | | | | Nombre de tirages:n | |
| 3 | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | | |
| 4 | 0 | 0,50000 | 0,25000 | 0,12500 | 0,06250 | 0,03125 | 0,01563 | 0,00781 | 0,00391 | 0,00195 | 0,00098 | | |
| 5 | 1 | 1,00000 | 0,75000 | 0,50000 | 0,31250 | 0,18750 | 0,10938 | 0,06250 | 0,03516 | 0,01953 | 0,01074 | | |
| 6 | 2 | | 1,00000 | 0,87500 | 0,68750 | 0,50000 | 0,34375 | 0,22656 | 0,14453 | 0,08984 | 0,05469 | | |
| 7 | 3 | | | 1,00000 | 0,93750 | 0,81250 | 0,65625 | 0,50000 | 0,36328 | 0,25391 | 0,17188 | | |
| 8 | 4 | | | | 1,00000 | 0,96875 | 0,89063 | 0,77344 | 0,63672 | 0,50000 | 0,37695 | | |
| 9 | 5 | | | | | 1,00000 | 0,98438 | 0,93750 | 0,85547 | 0,74609 | 0,62305 | | |
| 10 | 6 | | | | | | 1,00000 | 0,99219 | 0,96484 | 0,91016 | 0,82813 | | |
| 11 | 7 | | | | | | | 1,00000 | 0,99609 | 0,98047 | 0,94531 | | |
| 12 | 8 | | | | | | | | 1,00000 | 0,98047 | 0,94531 | | |
| 13 | 9 | | | | | | | | | 1,00000 | 0,99902 | | |
| 14 | 10 | | | | | | | | | | 1,00000 | | |
| 15 | Nombre de succès:k | | | | | | | | | | | | |
| 16 | | | | | | | | | | | | | |
| 17 | | | | | | | | | | | | | |
| 18 | | | | | | | | | | | | | |
| 19 | | | | | | | | | | | | | |
| 20 | | | | | | | | | | | | | |
| 21 | | | | | | | | | | | | | |
| 22 | | | | | | | | | | | | | |
| 23 | | | | | | | | | | | | | |
| 24 | Probabilité de succès:p | 0,50000 | | | | | | | | | | | |
| 25 | | | | | | | | | | | | | |
| 26 | | | | | | | | | | | | | |
| 27 | | | | | | | | | | | | | |

Il n'y a pas unicité des valeurs de $a$ et $b$ satisfaisant les conditions de l'intervalle de fluctuation. Il faut par conséquent faire un choix.

**Exemple.** Une personne achète toutes les semaines un jeu de grattage. La probabilité de succès du jeu est 10%. Cette personne aimerait connaitre au niveau de risque 5% qu'elle va être sa fréquence de succès sur une année.
On a donc $n = 52$ et $p = 0.1$. Comme il n'y a pas unicité de $a$ et $b$, on fait le choix de prendre $a$ tel que $\mathbb{P}(S_{52} < 52 \times a)$ soit de l'ordre de 0.025 et $b$ tel que $\mathbb{P}(S_{52} \leqslant 52 \times b)$ de l'ordre de 0.975. On aura alors $\mathbb{P}(\overline{X}_{52} \in [a,b]) = 0.95$.
En utilisant le tableur, on trouve les valeurs de la fonction de répartition $F(k) = \mathbb{P}(S_{52} \leqslant k)$ de la loi $\mathcal{B}(52, 0.1)$. On obtient pour les premières valeurs

| $k$ | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|---|
| $F(k)$ | 0.00417 | 0.02829 | 0.09663 | 0.22319 | 0.39544 | 0.57918 | 0.73910 | 0.85586 | 0.92884 | 0.96849 |
| $k$ | 10 | 11 | 12 | 13 | 14 | 15 | 16 | | | |
| $F(k)$ | 0.98743 | 0.99546 | 0.99851 | 0.99956 | 0.99988 | 0.99997 | 0.99999 | | | |

On remarque que pour $52 \times a = 2$ et $52 \times b = 10$, on obtient $\mathbb{P}(\overline{X}_{52} \in [a,b]) \simeq 0.96$. Par conséquent, au niveau de risque 4%, la personne aura une fréquence de succès comprise entre 2/52 et 10/52.

Lorsque le nombre de réalisations $n$ est grand, le calcul de la fonction de répartition de la loi binomiale est fastidieux. Au vu des graphiques présentés dans la section 2.3, on peut supposer, pour $n$ assez grand, avec un niveau de risque assez faible (mais dont on ne connaît pas la valeur), que la fréquence de succès est dans l'intervalle

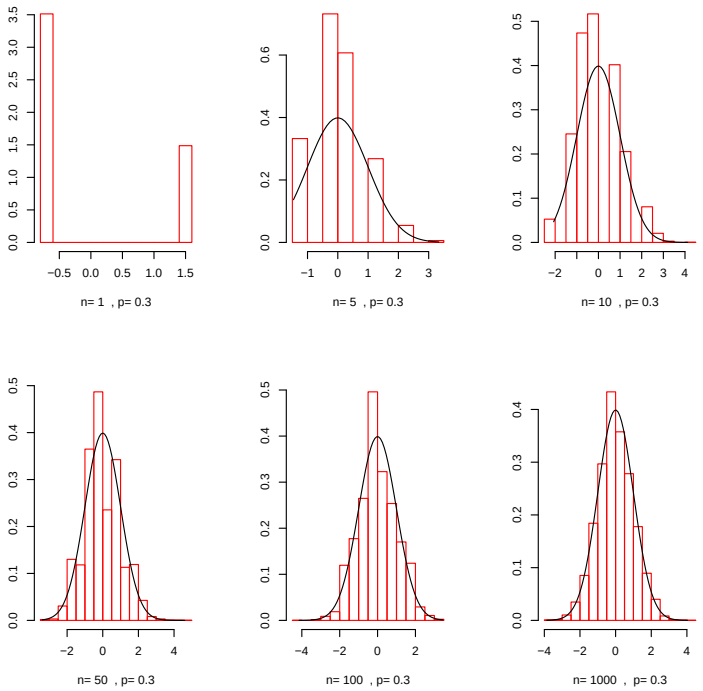
$$I_f = \left[p - \frac{1}{\sqrt{n}}, p + \frac{1}{\sqrt{n}}\right]. \tag{1}$$

Cependant, d'après les derniers graphiques de la section 2.3, ces intervalles ne sont pas forcément optimaux pour toutes les valeurs de $p$. En effet, quand $p$ est proche de 0 ou de 1, on a tendance à encadrer de façon trop grossière la fréquence de succès.

### 3.2 En utilisant le théorème central limite

D'après le théorème central limite, lorsque $n$ est grand, la loi de $\sqrt{\frac{n}{p(1-p)}}\big(\overline{X}_n - p\big)$ est proche de la loi normale $\mathcal{N}(0,1)$.

Sur les graphiques ci-dessous, on a tracé l'évolution de l'histogramme (en rouge) associé à la variable $\sqrt{\frac{n}{p(1-p)}}\big(\overline{X}_n - p\big)$, pour $p = 0.3$, en fonction de $n$. On observe qu'il converge vers la densité normale $\mathcal{N}(0,1)$ (courbe tracée en noir).




Si on trouve un intervalle $I_f = [a,b]$ tel que $\mathbb{P}(Z \in [a,b]) = 1-\alpha$ où $Z \sim \mathcal{N}(0,1)$, alors pour $n$ suffisamment grand

$$\begin{aligned}\mathbb{P}\left(p - \frac{\sqrt{p(1-p)}}{\sqrt{n}}a \leqslant \overline{X}_n \leqslant p + \frac{\sqrt{p(1-p)}}{\sqrt{n}}b\right) &= \mathbb{P}\left(\sqrt{\frac{n}{p(1-p)}}\big(\overline{X}_n - p\big) \in [a,b]\right) \\ &\simeq \mathbb{P}(Z \in [a,b]) = 1-\alpha.\end{aligned}$$

Il n'y a pas unicité de $a$ et $b$ vérifiant $\mathbb{P}(Z \in [a,b]) = 1-\alpha$. On peut par exemple prendre $a = -b$ (ce choix dépend en fait de la situation considérée). Par symétrie de la loi normale $\mathcal{N}(0,1)$, on a

$$\mathbb{P}(|Z| \leqslant t) = 2\mathbb{P}(Z \leqslant t) - 1.$$

Par conséquent, il faut trouver $t_\alpha$ tel que $\mathbb{P}(Z \leqslant t_\alpha) = 1-\alpha/2$. Cette valeur est obtenue en utilisant une table de la loi normale. Par exemple, pour $\alpha = 5\%$, on obtient $t_\alpha = 1.96$.

Conclusion
Soit $t_\alpha$ choisit tel que $\mathbb{P}(|Z| \leqslant t_\alpha) = 1-\alpha$.
Pour $n$ assez grand,

$$I_f = \left[ p - t_\alpha \frac{\sqrt{p(1-p)}}{\sqrt{n}}, p + t_\alpha \frac{\sqrt{p(1-p)}}{\sqrt{n}} \right] \tag{2}$$

est un intervalle de fluctuations de la fréquence de succès au niveau de confiance de l'ordre de $1-\alpha$.

### 3.3 Comparaison des intervalles proposés

On a proposé deux intervalles de fluctuations donnés par les formules (1) et (2). Peut-on comparer ces intervalles ?
On remarque que la fonction $p \mapsto p(1-p)$ est positive et atteint la valeur maximale $1/4$ en $p = 1/2$. On a

$$\forall p \in [0,1], \quad \sqrt{p(1-p)} \leqslant \frac{1}{2}.$$

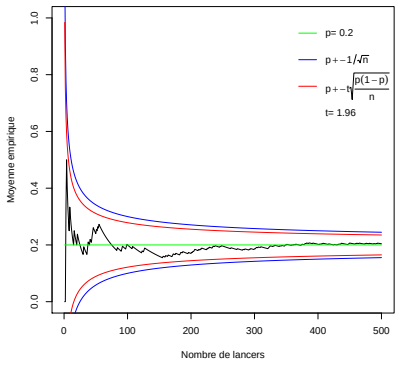
Lorsque $\alpha = 5\%$, on a $t_\alpha = 1.96$ et donc

$$\forall p \in [0,1], \quad 1.96\sqrt{p(1-p)} \leqslant 1.$$

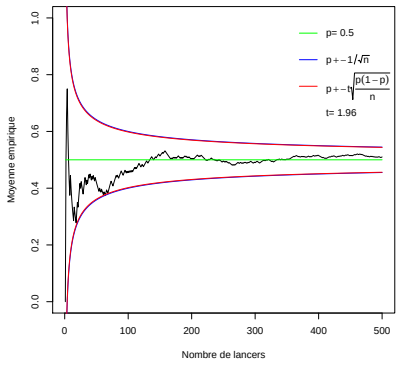
Par conséquent, pour $\alpha = 5\%$, on en déduit

$$\forall p \in [0,1], \quad \left[ p - 1.96 \frac{\sqrt{p(1-p)}}{\sqrt{n}}, p + 1.96 \frac{\sqrt{p(1-p)}}{\sqrt{n}} \right] \subset \left[ p - \frac{1}{\sqrt{n}}, p + \frac{1}{\sqrt{n}} \right]$$

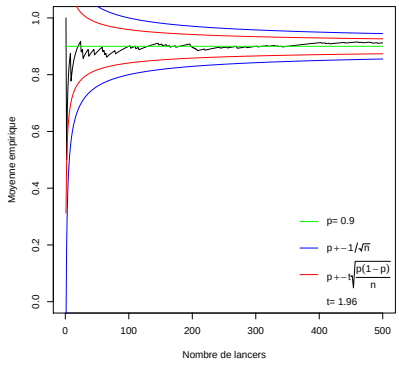
Quand $\alpha = 5\%$, l'intervalle définit par (2) est un meilleur intervalle de fluctuations que celui définit par (1) et quand $\alpha \neq 5\%$, l'intervalle (1) n'a pas de sens.
Sur les graphiques ci-dessous, on compare les intervalles de fluctuations (1) et (2) pour différentes valeurs de $p$ :




pour $p = 0.2$




pour $p = 0.5$




pour $p = 0.9$.

On observe que pour $p = 1/2$, ils sont quasiment confondus, ce qui était prévisible car $1.96\sqrt{p(1-p)}$ est alors très proche de 1.

### 3.4 Erreurs possibles dans la prise de décision

Les intervalles de fluctuations sont un outil intéressant pour la prise de décision. En effet, lorsqu'on ne connaît pas la probabilité de succès d'une épreuve de Bernoulli (probabilité de gagner à un jeu aléatoire, proportion de pièces défectueuses,...), on peut émettre une hypothèse sur sa valeur. On réalise alors plusieurs réalisations de l'épreuve de Bernoulli et au vu de la valeur de la moyenne empirique on pourra rejeter ou pas l'hypothèse de départ.
Si la moyenne empirique n'est pas dans l'intervalle de fluctuations, on aura tendance à rejeter l'hypothèse et si elle est dans l'intervalle, on sera enclin à ne pas rejetter l'hypothèse. C'est ce qu'on appelle une prise de décision. Il faut savoir qu'il y a deux erreurs possibles.

Erreur de première espèce : **rejeter l'hypothèse alors qu'elle est vraie.**

Un intervalle de fluctuation est construit avec un certain niveau de confiance $1-\alpha$ fixé à l'avance. La quantité $\alpha$ correspond à la probabilité de rejeter à tord l'hypothèse. Plus $\alpha$ est petit, moins on rejetera l'hypothèse, ce qui nous amène à l'autre type d'erreur.

Erreur de seconde espèce : **accepter l'hypothèse alors qu'elle n'est pas vraie.**

Autant, la première erreur est contrôlée par $\alpha$, autant cette seconde erreur n'est absolument pas contrôlée. Elle peut arriver avec une forte probabilité.

De manière générale, si la moyenne empirique $\overline{X}_n$ n'est pas dans l'intervalle de fluctuation, on rejette l'hypothèse et si $\overline{X}_n$ est dans l'intervalle de fluctuation, on ne rejette pas l'hypothèse. Ne pas rejeter l'hypothèse ne signifie pas qu'elle est vraie.... mais dans la réalité, il faut prendre une décision et donc souvent on accepte l'hypothèse dans ce cas de figure.

### 3.5 Exercices

**Exercice 1.** Reprennons le cas de la personne qui achète toutes les semaines un jeu de grattage. La probabilité de succès du jeu est 10%. Cette personne aimerait connaitre au niveau de risque 5% qu'elle va être sa fréquence de succès sur une année.
Faire le calcul de trois manières différentes et comparer les intervalles de fluctuations obtenus.

*Corrigé.* Le modèle probabiliste associé à l'expérience est le suivant. On introduit les variables aléatoire $X_i$ qui représentent le résultat du i$^{\text{ème}}$ jeu. On a $X_i = 1$ si le jeu est gagnant et $X_i = 0$ si le jeu est perdant. Les variables $X_i$ sont indépendantes et de loi de Bernoulli $\mathcal{B}(0.1)$. Par ailleurs, il y a 52 semaines dans une année. L'expérience est donc répétée $n = 52$ fois.

- On sait que $S_{52} = \sum_{i=1}^{52} X_i$ suit le loi binomiale $\mathcal{B}(52, 0.1)$. En utilisant la fonction de répartition de la loi binomiale, on trouve l'intervalle $I_1 = [2/52, 10/52] = [0.038, 0.192]$ (cf l'exemple de la section 3.1).
- En utilisant l'intervalle définit par la formule (1), on obtient l'intervalle $I_2 = [-0.038, 0.239]$.
- En utilisant la formule (2) basée sur le théorème central limite, on obtient l'intervalle $I_3 = [0.018, 0.181]$.

Le dernier intervalle obtenu est très proche de celui obtenu à l'aide de la loi binomiale, par contre le second est trop grossier. △

**Exercice 2.** Un joueur qui doit choisir au hasard une carte dans un jeu de 32 cartes obtient certains avantages s'il découvre un roi. On constate qu'il a retourné 11 fois un roi sur 50 essais. Peut-on présumer, au risque de 5%, que ce joueur est un tricheur ?

*Corrigé.* La probabilité de tirer un roi dans un jeu de 32 cartes est $p = 4/32 = 1/8$. On introduit les variables aléatoire $X_i$ qui représentent le résultat du i$^{\text{ème}}$ jeu. On a $X_i = 1$ si le joueur obtient un roi et $X_i = 0$ si le joueur obtient une autre carte qu'un roi.Les variables $X_i$ sont indépendantes et de loi de Bernoulli $\mathcal{B}(1/8)$.

On utilise la formule (2) pour calculer l'intervalle de fluctuations, car la probabilité de succès est proche de 0 et on sait que la formule (1) est trop grossière dans ce cas. Si le joueur ne triche pas, au niveau de confiance de l'ordre 95%, la fréquence de succès doit être dans l'intervalle

$$I_2 = \left[\frac{1}{8} - 1.96\frac{\sqrt{7}}{8\sqrt{50}}, \frac{1}{8} + 1.96\frac{\sqrt{7}}{8\sqrt{50}}\right] = [0.033, 0.217].$$

Le joueur a obtenu une fréquence de succès de $11/50 = 0.22$ qui est hors de l'intervalle de fluctuations. Par conséquent, au niveau de risque 5%, on peut mettre en doute l'honnêteté du joueur. △

# 4 Intervalles de confiance

On considère une expérience aléatoire qui n'a que deux résultats possibles : succès et échec. On suppose maintenant que **la probabilité de succès $p$ est inconnue**, ce qui est une grande différence avec ce que l'on a fait jusqu'à présent.
On peut penser par exemple à une élection entre deux participants ou à la proportion de pièces défectueuses dans un lot de grande taille.
Le but est d'estimer la valeur de $p$. Pour cela, on considère un échantillon $X_1, \dots, X_n$ de variables indépendantes de loi de Bernoulli $\mathcal{B}(p)$.
La valeur de chaque $X_i$ est connue. On a, par exemple, effectué un sondage sur un échantillon de la population pour connaitre les intentions de vote, on a prélevé au hasard un échantillon du lot de pièces usinées pour en comptabiliser le nombre de pièces défectueuses, on a joué plusieurs fois à "Pile ou Face" pour estimer la probabilité de tomber sur Pile, ...

**Definition 6.** *Soit $X_1, \dots, X_n$ des variables indépendantes de loi de Bernoulli $\mathcal{B}(p)$, avec $p \in ]0,1[$ inconnu. Un* **intervalle de confiance** *de la probabilité de succès $p$ au niveau de confiance $1-\alpha$ est un intervalle aléatoire $I_c = [a,b]$, avec $a$ et $b$ qui dépendent de l'échantillon $X_1, \dots, X_n$, tel que*

$$\forall p \in ]0,1[, \quad \mathbb{P}(p \in I_c) = 1 - \alpha.$$

*La quantité $\alpha$ est l'erreur que l'on s'autorise, elle est appelée niveau de risque. Elle est en général petite.*

*Un* **intervalle de confiance *asymptotique*** *pour $p$ au niveau de confiance $1-\alpha$ est une suite d'intervalles aléatoires $I_c^n$ tel que*

$$\forall p \in ]0,1[, \quad \lim_{n\to\infty} \mathbb{P}(p \in I_c^n) = 1 - \alpha.$$

**Attention !** Un intervalle de confiance ne doit pas dépendre de l'inconnue $p$. On doit pouvoir le calculer à partir de la seule connaissance des valeurs de l'échantillon $X_1, \dots, X_n$.

**Remarque 7.** *Un candidat naturel pour estimer la probabilité de succès est la moyenne empirique (aussi appelée fréquence de succès) $\overline{X}_n$, qui comme on l'a vu dans la section 2.2, converge en probabilité vers $p$ lorsque $n \to +\infty$. Comme $\mathbb{E}\left[\overline{X}_n\right] = p$, l'estimateur $\overline{X}_n$ est dit* **estimateur sans biais** *de l'inconnue $p$.*

On souhaite maintenant construire des intervalles de confiance.

## 4.1 Première approche

On ne peut pas travailler directement sur la loi de $\overline{X}_n$, car on ne connait pas le paramètre $p$ de la loi Binomiale $\mathcal{B}(n,p)$. Le Théorème Central Limite permet d'approcher sa loi par une loi normale, quelque soit la valeur de $p$.
Soit $t_\alpha$ tel que $\mathbb{P}(|Z| \leqslant t_\alpha) = 1-\alpha$ où $Z \sim \mathcal{N}(0,1)$. Cette valeur $t_\alpha$ est obtenue en utilisant une table de la loi normale. D'après le théorème central limite, pour $n$ assez grand, on a $\forall p \in ]0,1[$,

$$\begin{aligned}\mathbb{P}\left(p \in \left[\overline{X}_n - \frac{\sqrt{p(1-p)}}{\sqrt{n}}t_\alpha, \overline{X}_n + \frac{\sqrt{p(1-p)}}{\sqrt{n}}t_\alpha\right]\right) &= \mathbb{P}\left(\sqrt{\frac{n}{p(1-p)}}|\overline{X}_n - p| \leqslant t_\alpha\right) \\ &\simeq \mathbb{P}(|Z| \leqslant t_\alpha) = 1-\alpha.\end{aligned}$$

Malheureusement, l'intervalle $\left[\overline{X}_n - \frac{\sqrt{p(1-p)}}{\sqrt{n}}t_\alpha, \overline{X}_n + \frac{\sqrt{p(1-p)}}{\sqrt{n}}t_\alpha\right]$ dépend de l'inconnue $p$ et donc il ne peut pas être un intervalle de confiance.

On peut cependant réutiliser l'argumentation de la section 3.3.
Lorsque $\alpha = 5\%$, on a $t_\alpha = 1.96$ et

$$\forall p \in [0,1], \quad 1.96\sqrt{p(1-p)} \leqslant 1.$$

Par conséquent, pour $\alpha = 5\%$, on a

$$\forall p \in ]0,1[, \quad \left[\overline{X}_n - \frac{\sqrt{p(1-p)}}{\sqrt{n}}1.96, \overline{X}_n + \frac{\sqrt{p(1-p)}}{\sqrt{n}}1.96\right] \subset \left[\overline{X}_n - \frac{1}{\sqrt{n}}, \overline{X}_n + \frac{1}{\sqrt{n}}\right].$$

D'où, pour $n$ assez grand,

$$\forall p \in ]0,1[, \quad \mathbb{P}\left(p \in \left[\overline{X}_n - \frac{1}{\sqrt{n}}, \overline{X}_n + \frac{1}{\sqrt{n}}\right]\right) \geqslant 0.95.$$

L'intervalle $\left[\overline{X}_n - \frac{1}{\sqrt{n}}, \overline{X}_n + \frac{1}{\sqrt{n}}\right]$ est un intervalle de confiance *asymptotique* pour $p$ de niveau de confiance supérieur à 95%, quand $n$ est grand.

## 4.2 Seconde approche

Comme on ne connaît pas $p$, on ne connaît pas non plus la variance des $X_i$ : $Var(X_1) = p(1-p)$. C'est ce qui nous empêche d'utiliser directement le théorème central limite.
Cependant, d'après la loi des grands nombres, la moyenne empirique $\overline{X}_n$ converge vers $p$. On en déduit que,

$$\overline{X}_n\big(1-\overline{X}_n\big) \text{ converge en probabilité vers } Var(X_1) = p(1-p) \text{ lorsque } n \to +\infty.$$

Il est alors naturel d'approcher la valeur de la variance par $\overline{X}_n\big(1-\overline{X}_n\big)$ pour $n$ assez grand et d'utiliser cette approximation pour construire un intervalle de confiance.
Il existe en fait une généralisation du théorème central limite qui permet d'affirmer que

$$\frac{\sqrt{n}}{\sqrt{\overline{X}_n\big(1-\overline{X}_n\big)}}\big(\overline{X}_n - p\big) \text{ converge en loi vers la loi normale centrée réduite } \mathcal{N}(0,1) \text{ lorsque } n \to +\infty.$$

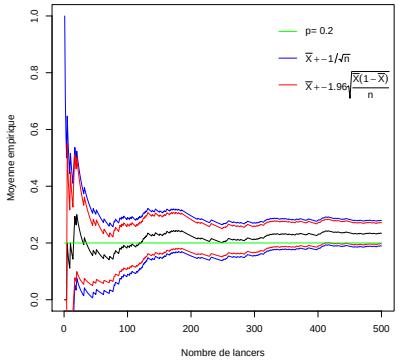
On en déduit que si $t_\alpha$ est choisi tel que $\mathbb{P}(|Z| \leqslant t_\alpha) = 1-\alpha$ où $Z \sim \mathcal{N}(0,1)$, pour $n$ assez grand, on a $\forall p \in [0,1]$,

$$\mathbb{P}\left(p \in \left[\overline{X}_n - \frac{\sqrt{\overline{X}_n(1-\overline{X}_n)}}{\sqrt{n}}t_\alpha, \overline{X}_n + \frac{\sqrt{\overline{X}_n(1-\overline{X}_n)}}{\sqrt{n}}t_\alpha\right]\right) \simeq 1-\alpha.$$

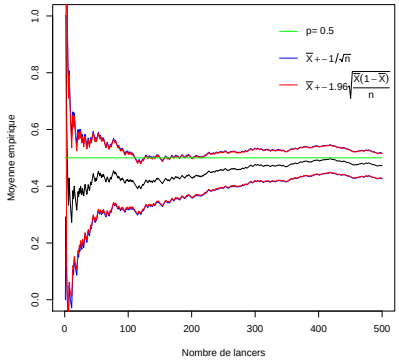
Par conséquent, l'intervalle $\left[\overline{X}_n - \frac{\sqrt{\overline{X}_n(1-\overline{X}_n)}}{\sqrt{n}}t_\alpha, \overline{X}_n + \frac{\sqrt{\overline{X}_n(1-\overline{X}_n)}}{\sqrt{n}}t_\alpha\right]$ est un intervalle de confiance *asymptotique* pour $p$ de niveau de confiance de l'ordre de $1-\alpha$ ($t_\alpha = 1.96$ pour $\alpha = 5\%$).

## 4.3 Comparaison des intervalles de confiance proposés

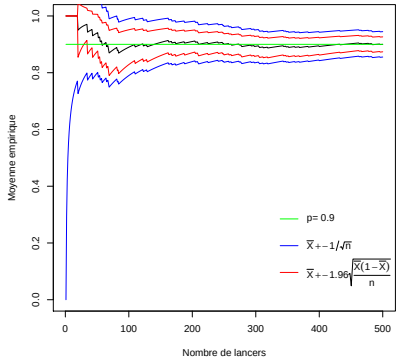
Comme on avait procédé pour les intervalles de fluctuations, on peut comparer les deux intervalles de confiance proposés, le premier n'ayant du sens que lorsque $\alpha = 5\%$.
On pose donc $\alpha = 5\%$. Sur les graphiques ci-dessous on trace l'évolution en fonction de $n$ des deux intervalles de confiance proposés pour différentes valeurs de $p$.




pour $p = 0.2$




pour $p = 0.5$




pour $p = 0.9$.

Ces courbes sont plus chaotiques que celles des intervalles de fluctuations car les intervalles de confiance sont centrés sur $\overline{X}_n$ qui varie en fonction de $n$.
On remarque que la vraie valeur $p$ n'est pas toujours dans l'intervalle de confiance, ce qui est normal car la probabilité d'avoir fait une erreur est de l'ordre de $\alpha = 5\%$.

## 4.4 Exercices

**Exercice 3.** Lors d'une enquête d'opinion, on a dénombré 81 personnes satisfaites d'un produit sur 1681 interrogées. En admettant que les personnes de l'échantillon ont été prises au hasard dans une grande population, donner l'intervalle de confiance de la proportion $p$ de personnes satisfaites dans la population totale, avec une probabilité de confiance de 0.95.

*Corrigé.* Le modèle probabiliste lié à l'expérience est le suivant. On introduit $X_i$ la variable aléatoire représentant l'opinion du i$^{\text{ème}}$ individu. On a $X_i = 1$ si la i$^{\text{ème}}$ personne interrogée est satisfaite et $X_i = 0$ sinon. Les variables $X_i$ sont supposées indépendantes car les personnes sont choisies au hasard et suivent la loi de Bernoulli $\mathcal{B}(p)$, où $p$ est la proportion de personnes satisfaites dans la population entière. Par conséquent, $\left[\overline{X}_n - 1.96\frac{\sqrt{\overline{X}_n(1-\overline{X}_n)}}{\sqrt{n}}, \overline{X}_n + 1.96\frac{\sqrt{\overline{X}_n(1-\overline{X}_n)}}{\sqrt{n}}\right]$ est un intervalle de confiance pour $p$ de niveau de confiance

de l'ordre de 95%, pour $n$ est grand. Ici $n = 1681$ et une réalisation de la fréquence de succès $\overline{X}_n$ est 81/1681. Par conséquent, une réalisation de l'intervalle de confiance au niveau de risque 5% est $[0.038, 0.058]$, ce qui donne une estimation de la valeur de $p$. △

**Exercice 4.** On veut connaître la prévalence d'une maladie chronique dans une population donnée. On extrait au hasard de cette population un échantillon d'effectif 400 et on observe que 16 personnes sont porteurs de la maladie.

1. Déterminer un intervalle de confiance de la prévalence de la maladie dans la population, au risque de 5 %.
2. Quelle doit être la taille minimale de l'échantillon si l'on souhaite une étendue de l'intervalle de confiance inférieure ou égale à 0.02, toujours au risque de 5% ?

*Corrigé.* On ne connaît pas la probabilité $p$ d'être malade. On regarde un échantillon $X_1, \dots, X_{400}$ de loi $\mathcal{B}(p)$ où $X_i = 1$ si le i$^{\text{ème}}$ individu est malade et $X_i = 0$ sinon.

1. L'intervalle $\left[\overline{X}_{400} - 1.96\frac{\sqrt{\overline{X}_{400}(1-\overline{X}_{400})}}{\sqrt{400}}, \overline{X}_{400} + 1.96\frac{\sqrt{\overline{X}_{400}(1-\overline{X}_{400})}}{\sqrt{400}}\right]$ est un intervalle de confiance pour $p$ de niveau de confiance de l'ordre de 95%. Une réalisation de cet intervalle est $[0.021, 0.059]$, ce qui donne une estimation de $p$.
2. Si on change la valeur de $n$, on change a priori la valeur de la moyenne empirique. Cependant, on remarque que quelque soit la valeur de $\overline{X}_n$,

$$\left[\overline{X}_n - 1.96\frac{\sqrt{\overline{X}_n(1-\overline{X}_n)}}{\sqrt{n}}, \overline{X}_n + 1.96\frac{\sqrt{\overline{X}_n(1-\overline{X}_n)}}{\sqrt{n}}\right] \subset \left[\overline{X}_n - \frac{1}{\sqrt{n}}, \overline{X}_n + \frac{1}{\sqrt{n}}\right].$$

Par conséquent, si $\frac{1}{\sqrt{n}} \leqslant 0.01$, alors l'étendue de l'intervalle de confiance sera inférieure ou égale à 0.02. Il faut donc $n \geqslant 10000$.

△

## 5 Généralisation à d'autres lois

La loi des grands nombres et le théorème central limite s'appliquent à n'importe quelle loi ayant un moment d'ordre 2 fini. On peut généraliser la construction d'intervalles de confiance pour la moyenne à des lois plus générales.

**Definition 8.** *On appelle* **échantillon** *(de taille $n$) un $n-$uplet $(X_1, \dots, X_n)$ de variables aléatoires indépendantes et de même loi.*

On considère un échantillon de variables aléatoires $X_1, \dots, X_n$. On note $m = \mathbb{E}[X_1]$ et $\sigma^2 = Var(X_1)$. On ne connaît pas $m$ et on aimerait l'estimer au mieux à partir de l'échantillon.

**Definition 9.** *Soit $X_1, \dots, X_n$ des variables indépendantes de même loi telle que $\mathbb{E}[|X_1|] < \infty$. On note $m = \mathbb{E}[X_1]$ l'espérance commune. Le paramètre $m$ est inconnu.*
*On appelle* **estimateur** *de $m$ toute variable aléatoire $\widehat{m}$ s'écrivant sous la forme $\widehat{m} = f(X_1, \dots, X_n)$.*

Un estimateur de $m$ ne doit évidemment pas dépendre de l'inconnue $m$. Il existe une infinité d'estimateurs. Par exemple, $\widehat{m} = 10$, $\widehat{m} = X_1X_2$, $\widehat{m} = (X_1 + X_n^2)e^{X_3}$ sont des estimateurs de $m$. On va essayer de chercher des estimateurs ayant de "bonnes propriétés", comme par exemple qui convergent vers l'inconnue $m$ lorsque la taille de l'échantillon $n$ tend vers l'infini.
D'après la Propriété 2, on sait que la moyenne empirique $\overline{X}_n = \sum_{i=1}^n X_i/n$ est une variable aléatoire d'espérance $m$ et de variance $\sigma^2/n$. Cette variable oscille par conséquent autour de la valeur inconnue $m$ (elle est dite sans biais) et pour $n$ assez grand les oscillations sont d'amplitude assez faible. Par ailleurs, d'après la loi des grands nombres, $\overline{X}_n$ converge en probabilité vers $m$ lorsque $n \to \infty$ (voir section 2.2).

Trouver un "bon" estimateur de l'inconnu $m$ permet d'estimer *ponctuellement* $m$. Mais on ne sait pas si la vraie valeur de $m$ est proche ou pas de l'estimation ponctuelle choisie. Pour estimer où se trouve la vraie valeur avec une probabilité assez forte, on introduit la notion d'intervalle de confiance.

**Definition 10.** *Soit $X_1, \dots, X_n$ des variables indépendantes de même loi avec $\mathbb{E}[|X_1|] < \infty$, d'espérance commune $m = \mathbb{E}[X_1]$. Un* **intervalle de confiance** *de la moyenne $m$ au niveau de confiance $1-\alpha$ est un intervalle aléatoire $I_c = [a, b]$, avec $a$ et $b$ qui dépendent de l'échantillon $X_1, \dots, X_n$, tel que*

$$\forall m \in \mathbb{R}, \quad \mathbb{P}(m \in I_c) = 1 - \alpha.$$

*Un* **intervalle de confiance *asymptotique*** *de $m$ au niveau de confiance $1-\alpha$ est une suite d'intervalles aléatoires $I_c^n$ tel que*

$$\forall m \in \mathbb{R}, \quad \lim_{n\to\infty} \mathbb{P}(m \in I_c^n) = 1 - \alpha.$$

## 5.1 Intervalles de confiance de la moyenne à variance connue

D'après le théorème central limite, si on choisit $t_\alpha$ tel que $\mathbb{P}(|Z| \leqslant t_\alpha) = 1 - \alpha$ où $Z \sim \mathcal{N}(0, 1)$, on a, pour $n$ grand,

$$\begin{aligned}\mathbb{P}\left(\sqrt{\frac{n}{\sigma^2}}|\overline{X}_n - m| \leqslant t_\alpha\right) &= \mathbb{P}\left(m \in \left[\overline{X}_n - \frac{\sigma}{\sqrt{n}}t_\alpha, \overline{X}_n + \frac{\sigma}{\sqrt{n}}t_\alpha\right]\right) \\ &\simeq \mathbb{P}(|Z| \leqslant t_\alpha) = 1 - \alpha.\end{aligned}$$

Quand la variance $\sigma^2$ est connue, l'intervalle $\left[\overline{X}_n - \frac{\sigma}{\sqrt{n}}t_\alpha, \overline{X}_n + \frac{\sigma}{\sqrt{n}}t_\alpha\right]$ est alors un intervalle de confiance de la moyenne $m$ au niveau de confiance $1-\alpha$ lorsque $n$ est assez grand.

**Exemple.** Regardons le cas particulier où les variables $X_i$ suivent la loi normale $\mathcal{N}(m, 1)$. Dans ce cas, comme les variables sont indépendantes, on connaît explicitement la loi de $\overline{X}_n$ qui est la loi normale $\mathcal{N}(m, 1/n)$.
Par conséquent, $\sqrt{n}\big(\overline{X}_n - m\big)$ suit la loi $\mathcal{N}(0, 1)$. Donc si $t_\alpha$ est choisi tel que $\mathbb{P}(|Z| \leqslant t_\alpha) = 1 - \alpha$ où $Z \sim \mathcal{N}(0, 1)$, on a pour tout $n \geqslant 1$

$$\mathbb{P}\left(m \in \left[\overline{X}_n - \frac{1}{\sqrt{n}}t_\alpha, \overline{X}_n + \frac{1}{\sqrt{n}}t_\alpha\right]\right) = 1 - \alpha.$$

L'intervalle $\left[\overline{X}_n - \frac{1}{\sqrt{n}}t_\alpha, \overline{X}_n + \frac{1}{\sqrt{n}}t_\alpha\right]$ est alors un intervalle de confiance de la moyenne $m$ au niveau de confiance $1-\alpha$, pour tout $n \geqslant 1$ (cet intervalle n'est pas *asymptotique*).

## 5.2 Intervalles de confiance de la moyenne à variance inconnue

Lorsque la variance est inconnue, comme dans le cas de la loi de Bernoulli, on a besoin d'estimer aussi ce paramètre.

### 5.3 Lorsque la variance est de la forme $\sigma^2 = g(m)$

On considère dans cette partie le cas où la variance s'écrit sous la forme $\sigma^2 = g(m)$ où $g$ est une fonction continue. Par exemple, dans le cas où les $X_i$ suivent la loi $\mathcal{B}(p)$, on a $g(x) = x(1-x)$.

D'après la loi des grands nombres, $\overline{X}_n$ est un estimateur de $m$ qui converge en probabilité vers $m$. Comme la fonction $g$ est continue, $g(\overline{X}_n)$ converge en probabilité vers $g(m) = \sigma^2$ quand $m \to \infty$.
Un estimateur naturel de $\sigma^2$ est alors $g(\overline{X}_n)$. D'après la généralisation du théorème central limite, on a

$$\sqrt{\frac{n}{g(\overline{X}_n)}}\big(\overline{X}_n - m\big) \text{ converge en loi vers la loi normale centrée réduite } \mathcal{N}(0,1) \text{ lorsque } n \to +\infty.$$

De ce résultat, en reprenant le même raisonnement que pour la loi de Bernoulli, on en déduit que, si $t_\alpha$ est choisi tel que $\mathbb{P}(|Z| \leqslant t_\alpha) = 1-\alpha$, l'intervalle $\left[\overline{X}_n - \sqrt{\frac{g(\overline{X}_n)}{n}}t_\alpha, \overline{X}_n + \sqrt{\frac{g(\overline{X}_n)}{n}}t_\alpha\right]$ est un intervalle de confiance *asymptotique* pour $m$ de niveau de confiance de l'ordre de $1-\alpha$.

**Exemple.** On considère un échantillon $X_1, \dots, X_n$ de loi de Poisson $\mathcal{P}(\lambda)$ avec $\lambda$ inconnu.
La loi de Poisson satisfait $\mathbb{E}[X_1] = \lambda$ et $Var(X_1) = \lambda$. Par conséquent, $\overline{X}_n$ est à la fois un estimateur sans biais de l'espérance et de la variance. On en déduit que si $t_\alpha$ est choisi tel que $\mathbb{P}(|Z| \leqslant t_\alpha) = 1-\alpha$, $\left[\overline{X}_n - \sqrt{\frac{\overline{X}_n}{n}}t_\alpha, \overline{X}_n + \sqrt{\frac{\overline{X}_n}{n}}t_\alpha\right]$ est un intervalle de confiance asymptotique pour $\lambda$ de niveau de confiance de l'ordre de $1-\alpha$.

### 5.4 Cas général

La variance ne s'écrit pas forcément comme une fonctionnelle de l'espérance. Par ailleurs, même si on peut l'écrire sous la forme $\sigma^2 = g(m)$, en général l'estimateur $g(\overline{X}_n)$ n'est pas centré autour de $\sigma^2$. On introduit maintenant un estimateur de la variance qui a de bonnes propriétés quelque soit la loi des $X_i$.

**Definition 11.** *On considère des variables $X_1, X_2, \dots, X_n$ indépendantes et de même loi.*
*On définit la* **variance empirique** *de l'échantillon par*

$$\widehat{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^n \big(X_i - \overline{X}_n\big)^2.$$

**Propriété 12.** *Soit $X_1, X_2, \dots, X_n$ de variables indépendantes et de même loi, d'espérance $m$ et de variance $\sigma^2$ finies. Alors*

1. $\widehat{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^n X_i^2 - \overline{X}_n^2$.
2. $\widehat{\sigma}_n^2$ *est une variable aléatoire d'espérance* $\frac{n-1}{n}\sigma^2$.
3. $\widehat{\sigma}_n^2$ *converge en probabilité vers $\sigma^2$ quand $n \to \infty$.*

**Remarque 13.** *L'estimateur est dit* biaisé *car son espérance n'est pas égale à $\sigma^2$. Cependant $\frac{n}{n-1}\widehat{\sigma}_n^2$ est un estimateur sans biais de $\sigma^2$ et qui converge en probabilité vers $\sigma^2$ quand $n \to \infty$.*

*Démonstration.* 1. On développe le carré dans la somme et on obtient

$$\widehat{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^n X_i^2 - \frac{2}{n}\sum_{i=1}^n X_i\overline{X}_n + \overline{X}_n^2 = \frac{1}{n}\sum_{i=1}^n X_i^2 - \overline{X}_n^2.$$

2. En utilisant $\mathbb{E}[\overline{X}_n] = m$ et $Var(\overline{X}_n) = \sigma^2/n$ et en développant le carré, on a

$$\begin{aligned}\mathbb{E}\big[\widehat{\sigma}_n^2\big] &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\Big[\big(X_i - m + m - \overline{X}_n\big)^2\Big] \\ &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\Big[(X_i - m)^2\Big] - \frac{2}{n}\sum_{i=1}^{n}\mathbb{E}\big[(X_i - m)\big(\overline{X}_n - m\big)\big] + \mathbb{E}\Big[\big(\overline{X}_n - m\big)^2\Big] \\ &= \sigma^2 - Var(\overline{X}_n) = \frac{n-1}{n}\sigma^2.\end{aligned}$$

3. D'après la loi des grands nombres, $\overline{X}_n$ converge en probabilité vers $m$. Par conséquent, $\overline{X}_n^2$ converge en probabilité vers $m^2$. Par ailleurs, en appliquant la loi des grands nombres à la suite de variables $(X_i^2)_{i\geqslant 1}$ qui sont indépendantes et de même loi, on obtient que $\frac{1}{n}\sum_{i=1}^n X_i^2$ converge en probabilité vers $\mathbb{E}[X_1^2]$. Comme $\sigma^2 = \mathbb{E}[X_1^2] - m^2$, on en déduit que $\widehat{\sigma}_n^2$ converge en probabilité vers $\sigma^2$ quand $n \to \infty$.

□

**Remarque 14.** *Lorsque l'échantillon $X_1, X_2, \ldots, X_n$ suit la loi de Bernoulli $\mathcal{B}(p)$, alors $\widehat{\sigma}_n^2 = \overline{X}_n - \overline{X}_n^2 = \overline{X}_n\big(1 - \overline{X}_n\big)$. En effet, dans ce cas $X_i^2 = X_i$. On retrouve l'estimation de la variance utilisée dans la section 4.2.*

Comme $\widehat{\sigma}_n^2$ converge en probabilité vers $\sigma^2$ quand $n \to \infty$, en utilisant la généralisation du théorème central limite, on obtient que

$$\sqrt{\frac{n}{\widehat{\sigma}_n^2}}\big(\overline{X}_n - m\big) \text{ converge en loi vers la loi normale centrée réduite } \mathcal{N}(0,1) \text{ lorsque } n \to +\infty.$$

Par conséquent, pour $t_\alpha$ est choisi tel que $\mathbb{P}(|Z| \leqslant t_\alpha) = 1 - \alpha$, l'intervalle $\Big[\overline{X}_n - \frac{\widehat{\sigma}_n}{\sqrt{n}}t_\alpha, \overline{X}_n + \frac{\widehat{\sigma}_n}{\sqrt{n}}t_\alpha\Big]$ est un intervalle de confiance *asymptotique* pour $m$ de niveau de confiance de l'ordre de $1 - \alpha$.

## 5.5 Exercices

**Exercice 5.** On considère $X_1, \ldots, X_n$ un échantillon de loi $\mathcal{E}(\lambda)$, avec $\lambda > 0$ inconnu. Trouver un intervalle de confiance à 95% de $\lambda$.

*Corrigé.* Dans le cas de la loi exponentielle, on a $\mathbb{E}[X_1] = 1/\lambda$ et $Var(X_1) = 1/\lambda^2$.
Par conséquent, d'après la loi des grands nombres $\overline{X}_n$ converge en probabilité vers $1/\lambda$ et d'après le théorème central limite généralisé

$$\sqrt{n\overline{X}_n}\left(\overline{X}_n - \frac{1}{\lambda}\right) \text{ converge en loi vers la loi normale centrée réduite } \mathcal{N}(0,1) \text{ lorsque } n \to +\infty.$$

Comme $\alpha = 5\%$, on prend $t_\alpha = 1.96$ et donc pour $n$ assez grand

$$\begin{aligned}\mathbb{P}\left(\sqrt{n\overline{X}_n}\left|\overline{X}_n - \frac{1}{\lambda}\right| \leqslant 1.96\right) &= \mathbb{P}\left(\overline{X}_n - \frac{1.96}{\sqrt{n\overline{X}_n}} \leqslant \frac{1}{\lambda} \leqslant \overline{X}_n + \frac{1.96}{\sqrt{n\overline{X}_n}}\right) \\ &= \mathbb{P}\left(\left(\overline{X}_n + \frac{1.96}{\sqrt{n\overline{X}_n}}\right)^{-1} \leqslant \lambda \leqslant \left(\overline{X}_n - \frac{1.96}{\sqrt{n\overline{X}_n}}\right)^{-1}\right) \simeq 0.95.\end{aligned}$$

Donc $\left[\left(\overline{X}_n + \frac{1.96}{\sqrt{n}\overline{X}_n}\right)^{-1}, \left(\overline{X}_n - \frac{1.96}{\sqrt{n}\overline{X}_n}\right)^{-1}\right]$ est un intervalle de confiance de $\lambda$ de niveau de l'ordre de 95% pour $n$ grand.

Autre méthode : On peut dans le cas particulier de la loi exponentielle, utiliser le Théorème Central Limite classique. En effet, comme $Var(X_1) = 1/\lambda^2$, on a, pour $n$ assez grand,

$$\mathbb{P}\left(\lambda\sqrt{n}\left|\overline{X}_n - \frac{1}{\lambda}\right| \leqslant 1.96\right) = \mathbb{P}\big(\sqrt{n}\big|\lambda\overline{X}_n - 1\big| \leqslant 1.96\big) \simeq \mathbb{P}(|Z| \leqslant 1.96) = 0.95.$$

Cependant,

$$\big\{\sqrt{n}\big|\lambda\overline{X}_n - 1\big|\big\} = \left\{\frac{1}{\overline{X}_n}\left(1 - \frac{1.96}{\sqrt{n}}\right) \leqslant \lambda \leqslant \frac{1}{\overline{X}_n}\left(1 + \frac{1.96}{\sqrt{n}}\right)\right\}.$$

Par conséquent, $I = \left[\frac{1}{\overline{X}_n}\left(1 - \frac{1.96}{\sqrt{n}}\right), \frac{1}{\overline{X}_n}\left(1 + \frac{1.96}{\sqrt{n}}\right)\right]$ est un intervalle de confiance asymptotique au niveau de confiance 95%. △

**Exercice 6.** Vingt adultes francophones ont fait l'objet d'une expérience de mémoire. Le temps pris pour apprendre une liste de 5 verbes allemands a été enregistré pour chaque personne. Ceci a donné les résultats suivants (en minutes) :

| | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 5.1 | 4.8 | 6.3 | 5.0 | 5.5 | 5.0 | 5.2 | 4.9 | 4.5 | 5.8 | 5.3 | 5.2 | 5.6 | 5.5 | 5.2 | 4.9 | 4.7 | 4.7 | 5.8 | 5.5 |

1. Calculer la moyenne et l'écart type de l'échantillon
2. Établir un intervalle de confiance ($\alpha = 5\%$) du temps moyen nécessaire à un francophone pour apprendre la liste des 5 verbes allemands.
3. On dit qu'un francophone ne peut apprendre qu'un verbe par minute. Est-ce que cette affirmation est justifiée par le résultat obtenu dans la question précédente ?

*Corrigé.* On introduit le modèle probabiliste suivant. La variable $X_i$ représente ici le temps mis par la personne $i$ pour apprendre 5 verbes allemand. On ne connaît pas la loi des $X_i$, par conséquent on se permet d'utiliser *abusivement* l'approximation par la loi normale même si $n = 20$ n'est pas si grand.

1. La moyenne de l'échantillon est 5.225 et sa variance 0.199.
2. Ne connaissant pas la loi des $X_i$, on doit utiliser l'intervalle de confiance faisant intervenir la variance empirique. On obtient alors l'intervalle de confiance pour la moyenne $m$ au niveau de confiance 95%

$$I_c = \left[\overline{X}_n - 1.96\frac{\widehat{\sigma}_n}{\sqrt{n}}, \overline{X}_n + 1.96\frac{\widehat{\sigma}_n}{\sqrt{n}}\right]$$

   dont une réalisation est $[5.03, 5.42]$.
3. Selon le résultat du sondage, au niveau de risque 5%, il semble qu'un francophone apprenne moins d'un verbe par minute.

△

**Exercice 7.** Une entreprise reçoit un lot important de pièces fabriquées en série. Dans un échantillon de 200 pièces, 15 sont défectueuses.

1. Donner un intervalle de confiance de la proportion $p$ de pièces défectueuses dans tout le lot.
2. L'entreprise n'accepte la livraison que si la proportion $p$ de pièces défectueuses est de 5%. Que conclure au niveau de risque 1% ?

3. En fait, il est plus réaliste de penser que l'entreprise n'accepte la livraison que si la proportion $p$ de pièces défectueuses est inférieure à 5%. Comment répondre à cette question ?

*Corrigé.* Le modèle probabiliste associé à l'expérience est le suivant. On introduit $X_i$ la variable aléatoire correspondant à l'état de la i$^{\text{ème}}$ pièce et $p$ la proportion de pièces défectueuses dans tout le lot. Comme les pièces sont fabriquées en série, on peut supposer que chaque pièce a la même proportion d'être défectueuse. Les pièces étant choisies au hasard, il est naturel de supposer que les variables $X_i$ sont indépendantes de loi $\mathcal{B}(p)$.

1. Pour $\alpha = 0.01$, on a $t_\alpha = 2.57$. Un intervalle de confiance de $p$ au niveau de confiance 99% est donné par
$$\left[\overline{X}_{200} - 2.57\frac{\sqrt{\overline{X}_{200}\big(1-\overline{X}_{200}\big)}}{\sqrt{200}}, \overline{X}_{200} + 2.57\frac{\sqrt{\overline{X}_{200}\big(1-\overline{X}_{200}\big)}}{\sqrt{200}}\right]$$
dont une réalisation est $[0.027, 0.123]$.
2. Si l'entreprise a raison, la valeur de $p$ est 5%, et donc la moyenne empirique devrait être, au niveau de confiance 99%, dans l'intervalle de fluctuation
$$\left[p - 2.57\frac{\sqrt{p(1-p)}}{\sqrt{200}}, p - 2.57\frac{\sqrt{p(1-p)}}{\sqrt{200}}\right] = [0.01, 0.089]$$
Une réalisation de la moyenne empirique $\overline{X}_{200}$ vaut 0.075 qui est dans cet intervalle de fluctuations, on ne peut donc pas rejeter, au niveau de risque 5%, l'hypothèse que la proportion de pièces défectueuses est 5%. L'entreprise peut prendre la décision d'accepter le lot.
3. Si l'entreprise souhaite que la proportion de pièces défectueuses soient inférieure à un certain niveau, elle va avoir tendance à rejeter le lot si la moyenne empirique est trop grande. Par conséquent, cherchons un intervalle de fluctuations de la forme $[0, a]$. En utilisant la même méthode que dans le cours et notamment le théorème central limite, le réel $a$ va vérifier pour tout $p \in [0, 5\%]$
$$\mathbb{P}\big(\overline{X}_n > a\big) \simeq \mathbb{P}\left(Z > \sqrt{n}\frac{a-p}{\sqrt{p(1-p)}}\right) = \alpha$$
où $Z \sim \mathcal{N}(0,1)$, puisque $\mathbb{P}(\overline{X}_n \in [0,a]) = 1 - \mathbb{P}(\overline{X}_n > a)$. En utilisant la table de la loi normale, on obtient $\sqrt{n}\frac{a-p}{\sqrt{p(1-p)}} = 2.32$, d'où $a = p + 2.32\sqrt{\frac{p(1-p)}{n}}$.
On remarque que pour tout $p \leqslant 5\%$, on a
$$\left[0, p + 2.32\sqrt{\frac{p(1-p)}{n}}\right] \subset \left[0, 0.05 + 2.32\sqrt{\frac{0.05\times 0.95}{n}}\right].$$
Par conséquent, pour tout $p \leqslant 5\%$,
$$\mathbb{P}\left(\overline{X}_n \leqslant 0.05 + 2.32\sqrt{\frac{0.05\times 0.95}{n}}\right) \geqslant \mathbb{P}\left(\overline{X}_n \leqslant p + 2.32\sqrt{\frac{p(1-p)}{n}}\right) \simeq 0.95$$
Si la vraie proportion $p$ de pièces défectueuses est inférieur à 5%, alors avec un niveau de confiance supérieur à 0.95, $\overline{X}_n$ est dans l'intervalle $\left[0, 0.05 + 2.32\sqrt{\frac{0.05.95}{n}}\right]$, qui vaut $[0, 0.085]$ pour $n = 200$. Au niveau de confiance de l'ordre de 95%, on ne rejete pas l'hypothèse que la proportion de pièces défectueuses est inférieure à 5% et l'entreprise accepte le lot.

△

# 6 Codes de simulation

Les codes de simulations des graphes présents dans ce document sont mis à votre disposition à la fois en `R` et `Scilab`. Choisissez entre ces deux logiciels, celui avec le quel vous êtes le plus à l'aise, ils sont tous les deux très efficaces pour ce type de simulation.

## 6.1 Codes en `R`

Les simulations présentes dans ce document ont été effectuées avec le logiciel `R`. Ce logiciel est gratuit et téléchargeable sur la page http://www.r-project.org/ ou sur http://www.rstudio.com/.

N'hésitez pas à utiliser l'aide de `R`, elle est très bien documentée. Vous trouverez, sous l'onglet *Contributed*, sur la page http://cran.r-project.org/ plusieurs manuels (nombreux sont en français), dont notamment celui de Emmanuel Paradis "R pour les débutants".

## 6.2 Codes en `Scilab`

Un autre logiciel gratuit pour réaliser des simulations est `Scilab`. Il est téléchargeable sur la page

http://www.scilab.org/fr.

L'aide de `Scilab` est très bien documentée. Vous trouverez, sous l'onglet *Ressources*, sur la page

http://www.scilab.org/fr/resources/documentation

plusieurs aides et notamment de très bons tutoriels http://www.scilab.org/fr/resources/documentation/tutorials.