

 **Attention : utiliser un lecteur de pdf adapté**

 Ce polycopié contient plusieurs animations, il est donc conseillé d'utiliser un lecteur de pdf capable de lire les animations (comme Adobe Reader, Foxit PDF Reader, Okular ou autres).

Table des matières

1	Loi faible des grands nombres	2
2	Convergence en loi et théorème central limite	2
2.1	Convergence en loi	2
2.2	Théorème central limite	3
3	Introduction aux tests : test de conformité à la moyenne	4
3.1	Test de conformité à la moyenne	4
3.2	Test de conformité à une proportion	5

1 Loi faible des grands nombres



Définition de la moyenne empirique et de la variance empirique d'un n -uplet de variables aléatoires

Soit $X_1, \dots, X_n$ des variables aléatoires. On appelle **moyenne empirique** $M_n = \frac{1}{n} \sum_{k=1}^n X_k$, **écart-type empirique**

$$S_n = \sqrt{\frac{1}{n} \sum_{k=1}^n (X_k - M_n)^2} \text{ et } \text{variance empirique } S_n^2 = \frac{1}{n} \sum_{k=1}^n (X_k - M_n)^2.$$

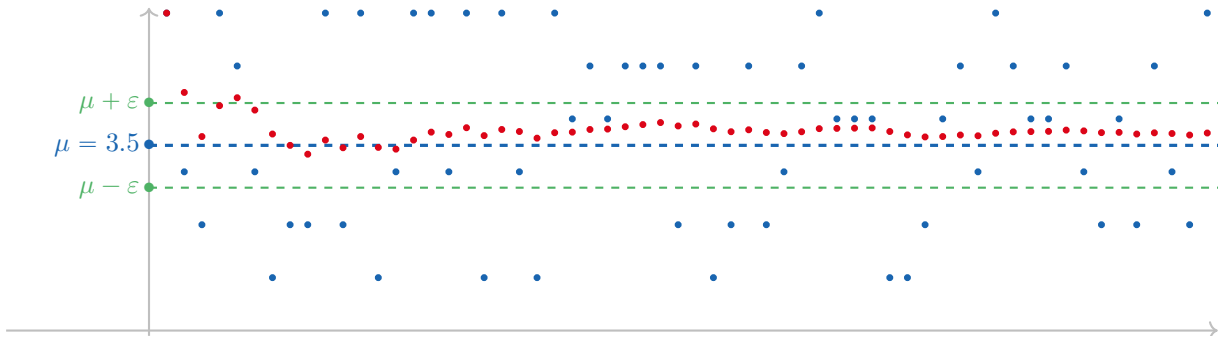
Remarque 1. Si $X_1, \dots, X_n$, des variables aléatoires indépendantes, admettent toutes la même espérance μ et toutes la même variance σ^2 , alors M_n admet aussi une espérance et une variance et $\mathbb{E}(M_n) = \mu$ et $\mathbb{V}(M_n) = \sigma^2/n$. Ainsi, si on a un échantillon de X , la moyenne empirique vaut «en moyenne» μ . On dit que M_n est un **estimateur sans biais** de μ .



Théorème n° 1 : loi faible des grands nombres

Soit $(X_n)_{n \in \mathbb{N}^*}$ une suite de VA indépendantes admettant une même espérance μ et une même variance. Alors, pour tout $\varepsilon > 0$, $\mathbb{P}(|M_n - \mu| > \varepsilon) \xrightarrow{n \rightarrow \infty} 0$.

Remarque 2. On dit que la suite de variables aléatoires (M_n) tend vers μ **en probabilité**.



Exemple de simulation : X_n (en **bleu**) suit une loi uniforme sur $\llbracket 1; 6 \rrbracket$ (lancé d'un dé équilibré), et M_n (en **rouge**). Pour n assez grand, $|M_n - \mu| > \varepsilon$ est possible mais peu probable et donc on ne finit par ne pas l'observer sur la simulation.

2 Convergence en loi et théorème central limite

2.1 Convergence en loi



Définition de la convergence en loi

On dit qu'une suite de VA $(X_n)_{n \in \mathbb{N}}$ **converge en loi** vers une VA X si $F_{X_n}(x) \xrightarrow{n \rightarrow \infty} F_X(x)$ pour tout $x \in \mathbb{R}$ où F_X est continue.

Remarque 3.

- Si (X_n) converge en loi vers X une VA à densité, $\mathbb{P}(X_n \in I) \xrightarrow{n \rightarrow \infty} \mathbb{P}(X \in I)$ pour tout I intervalle.
- Si X est une variable à densité, alors F_X est continue sur $\mathbb{R}$, ainsi la limite est vraie pour tout $x \in \mathbb{R}$.

Exemple 1. On suppose que pour tout $n \in \mathbb{N}^*$, $X_n \sim \mathcal{U}([0; 1/n])$, étudier la convergence en loi de $(X_n)_{n \in \mathbb{N}}$.

**Proposition n° 1 : caractérisation de la convergence en loi pour des variables aléatoires entières**

Si X et X_n , pour $n \in \mathbb{N}$, sont des VA à valeurs dans $\mathbb{N}$, alors $(X_n)_n$ converge vers X en loi ssi pour tout $k \in \mathbb{N}$, $\mathbb{P}(X_n = k) \xrightarrow{n \rightarrow \infty} \mathbb{P}(X = k)$.

**Proposition n° 2 : convergence en loi d'une suite de binomiales vers une loi de Poisson**

Si, pour tout $n \in \mathbb{N}^*$, $X_n \sim \mathcal{B}(n, p_n)$ et $np_n \xrightarrow{n \rightarrow \infty} \lambda > 0$. Alors (X_n) converge en loi vers X où $X \sim \mathcal{P}(\lambda)$.

Convergence en loi d'une suite $(X_n)_n$ ou pour $n \geq 8$, $X_n \sim \mathcal{B}(n, 7/n)$ vers une loi de Poisson de paramètre 7. En **rouge** la loi de X_n en **bleu** celle de Poisson de paramètre 7.

Remarque 4. Dans la pratique pour $n \geq 30$ et $p \leq 0.1$, si $X \sim \mathcal{B}(n, p)$ et $Y \sim \mathcal{P}(np)$, on considère que l'on peut approximer $\mathbb{P}(X = k)$ par $\mathbb{P}(Y = k)$ (ces conditions ne sont pas à connaître). Cela justifie pourquoi les lois de Poisson sont utilisées pour les évènements rares.

2.2 Théorème central limite

**Théorème n° 2 central limite (première forme)***(admis)*

Soit $(X_n)_{n \in \mathbb{N}^*}$ une suite de variables aléatoires indépendantes de même loi, admettant une espérance μ et une variance $\sigma^2 > 0$, alors $(M_n^*)_{n \in \mathbb{N}^*}$ converge en loi vers Z où $Z \in \mathcal{N}(0, 1)$:

$$\forall (\alpha, \beta) \in \mathbb{R}^2 \text{ tel que } \alpha \leq \beta \quad \mathbb{P} \left(\alpha \leq \frac{M_n - \mu}{\sigma/\sqrt{n}} \leq \beta \right) \xrightarrow{n \rightarrow \infty} \frac{1}{\sqrt{2\pi}} \int_{\alpha}^{\beta} e^{-\frac{t^2}{2}} dt = \Phi(\beta) - \Phi(\alpha)$$

Prenons, pour tout $n \in \mathbb{N}^*$, $X_n \sim \mathcal{U}([1; 6])$. Affichage de l'histogramme des valeurs de M_n^* pour (l'histogramme a été réalisé avec un échantillon de taille 10^5) différentes valeurs de n .



Théorème n° 3 de Moivre-Laplace

Si, pour tout $n \in \mathbb{N}^*$, $Y_n \sim \mathcal{B}(n, p)$ où $p \in]0; 1[$, alors la suite (Y_n^*) converge en loi vers Z où $Z \sim \mathcal{N}(0, 1)$:

$$\forall x \in \mathbb{R} \quad \mathbb{P} \left(\frac{Y_n - np}{\sqrt{np(1-p)}} \leq x \right) \xrightarrow{n \rightarrow \infty} \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-\frac{t^2}{2}} dt$$

$Y_n \sim \mathcal{B}(n, p)$ avec $p = 0.3$, représentation en **rouge** de la fonction de répartition de $\frac{Y_n - np}{\sqrt{np(1-p)}}$ et en **bleu pointillé** la fonction de répartition de la loi normale centrée réduite.

Remarque 5. Le théorème central limite, affirme que M_n oscille autour de μ avec des fluctuations qui dépendent de σ . Or, on ne connaît pas toujours σ , on peut alors le remplacer dans le théorème central limite par l'écart-type empirique :



Théorème n° 4 central limite (seconde forme)

(admis)

Soit $(X_n)_{n \in \mathbb{N}^*}$ une suite de variables aléatoires indépendantes de même loi, admettant une espérance μ et une variance $\sigma^2 > 0$, alors $\left(\frac{M_n - \mu}{S_n / \sqrt{n}} \right)_{n \in \mathbb{N}^*}$ converge en loi vers Z où $Z \sim \mathcal{N}(0, 1)$:

$$\forall (\alpha, \beta) \in \mathbb{R}^2 \text{ tel que } \alpha < \beta \quad \mathbb{P} \left(\alpha \leq \frac{M_n - \mu}{S_n / \sqrt{n}} \leq \beta \right) \xrightarrow{n \rightarrow \infty} \frac{1}{\sqrt{2\pi}} \int_{\alpha}^{\beta} e^{-\frac{t^2}{2}} dt = \Phi(\beta) - \Phi(\alpha)$$



Proposition n° 3 : espérance de la variance empirique

(non exigible)

Si $X_1, \dots, X_n$, des VA indépendantes, admettent une même espérance et une même variance σ^2 . Alors, S_n^2 admet une espérance : $\mathbb{E}(S_n^2) = \frac{n-1}{n} \sigma^2$.

Remarques 6. • Comme $\mathbb{E}(S_n^2) \neq \sigma^2$, on dit que S_n^2 est un **estimateur biaisé** de σ^2 .

• Ainsi, $\mathbb{E}(S_n^2) \xrightarrow{n \rightarrow \infty} \sigma^2$, on dit que S_n^2 est un **estimateur asymptotiquement sans biais** de σ^2 .

• Observons que $\frac{n}{n-1} \mathbb{E}(S_n^2) = \sigma^2$, par linéarité de l'espérance, on obtient que $\mathbb{E} \left(\frac{n}{n-1} S_n^2 \right) = \sigma^2$. Posons $S'_n =$

$\sqrt{\frac{n}{n-1}} S_n = \sqrt{\frac{1}{n-1} \sum_{k=1}^n (X_k - M_n)^2}$ (écart-type corrigé) de sorte que $\mathbb{E}(S_n'^2) = \sigma^2$, ainsi, $S_n'^2$ est un **estimateur sans biais** de σ^2 .

• Dans le théorème central limite seconde forme, on obtient le même résultat en remplaçant S_n par S'_n .

3 Introduction aux tests : test de conformité à la moyenne

3.1 Test de conformité à la moyenne

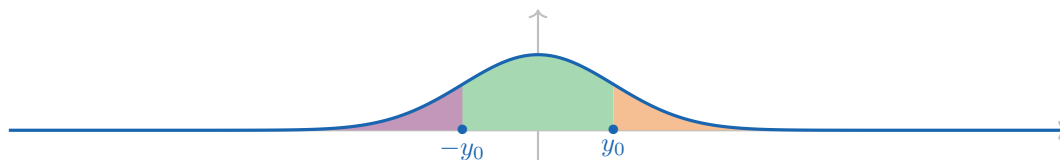
Pour un TIPE, des élèves étudient un pot biodégradable censé favoriser la croissance d'une plante. La longueur moyenne de la plante sans pot est connue et vaut $m = 11.1$. Les élèves mesurent la longueur moyenne des plantes qui ont poussé dans leur pot et veulent savoir si la moyenne obtenue est significativement différente de m pour conclure à un effet de leur

pot. On note X la variable aléatoire qui mesure la longueur d'une plante ayant poussé dans leur pot. Les élèves ont donc des valeurs d'échantillon de taille n de $X : X_1, \dots, X_n$ avec $\mathbb{E}(X) = \mu$ (inconnue) et une certaine variance $\mathbb{V}(X) = \sigma^2$ (inconnue). On pose l'hypothèse nulle $\mathcal{H}_0 : \mu = m$. On cherche un intervalle auquel $\mu = \mathbb{E}(X)$ appartient avec une grande probabilité. Si m (que l'on connaît) n'est pas cet intervalle, on rejettera l'hypothèse $\mathcal{H}_0$. Si n est assez grand, d'après le théorème central limite, les lois de $\frac{M_n - \mu}{\sigma/\sqrt{n}}, \frac{M_n - \mu}{S_n/\sqrt{n}}, \frac{M_n - \mu}{S'_n/\sqrt{n}}$ sont approximées par une loi normale centrée réduite. Soit $y_0 > 0$:

$$\begin{aligned}\mathbb{P}\left(-y_0 \leq \frac{M_n - \mu}{\sigma/\sqrt{n}} \leq y_0\right) &= \mathbb{P}\left(M_n - y_0 \frac{\sigma}{\sqrt{n}} \leq \mu \leq M_n + y_0 \frac{\sigma}{\sqrt{n}}\right) \\ \mathbb{P}\left(-y_0 \leq \frac{M_n - \mu}{S_n/\sqrt{n}} \leq y_0\right) &= \mathbb{P}\left(M_n - y_0 \frac{S_n}{\sqrt{n}} \leq \mu \leq M_n + y_0 \frac{S_n}{\sqrt{n}}\right) \\ \mathbb{P}\left(-y_0 \leq \frac{M_n - \mu}{S'_n/\sqrt{n}} \leq y_0\right) &= \mathbb{P}\left(M_n - y_0 \frac{S'_n}{\sqrt{n}} \leq \mu \leq M_n + y_0 \frac{S'_n}{\sqrt{n}}\right)\end{aligned}$$

Ainsi, μ est dans trois intervalles avec une certaine probabilité. Seulement, le premier intervalle a un inconvénient : il dépend de σ dont on ne connaît pas forcément la valeur. Concentrons-nous donc sur le deuxième intervalle (on pourrait faire de même avec le troisième). Posons $Y = \frac{M_n - \mu}{S_n/\sqrt{n}}$. Soit $\varepsilon \in]0; 1[$ et cherchons $y_0 > 0$ tel que $\mathbb{P}(-y_0 \leq Y \leq y_0) \geq 1 - \varepsilon$. Or, par parité de la densité de la loi centrée réduite et par croissance stricte de Φ :

$$\mathbb{P}(-y_0 \leq Y \leq y_0) \simeq \Phi(y_0) - \Phi(-y_0) = 2\Phi(y_0) - 1 \geq 1 - \varepsilon \iff \Phi(y_0) \geq 1 - \frac{\varepsilon}{2} \iff y_0 \geq \Phi^{-1}\left(1 - \frac{\varepsilon}{2}\right)$$



Si $Z \sim \mathcal{N}(0, 1)$ et $y_0 = \Phi^{-1}(1 - \varepsilon/2)$, alors $\mathbb{P}(Z \leq -y_0) = \mathbb{P}(Z \geq y_0) = \frac{\varepsilon}{2}$ et $\mathbb{P}(-y_0 \leq Z \leq y_0) = 1 - \varepsilon$.

Ainsi plus ε diminue, plus $1 - \varepsilon/2$ augmente plus y_0 augmente : plus le niveau de confiance augmente plus la précision baisse. Nos élèves ont un échantillon de $n = 35$ plantes avec une taille moyenne (espérance empirique) de 12cm et un écart-type (écart-type empirique) de 3cm. On donne $\Phi(1.64) \simeq 0.95$ et $\Phi(1.95) \simeq 0.975$:

- Pour $\varepsilon = 0.1$, $\Phi^{-1}(1 - \varepsilon/2) = \Phi^{-1}(0.95) \simeq 1.64$. On prend donc $y_0 = 1.64$, ainsi, $12 - \frac{1.64S_n}{\sqrt{35}} \simeq 11.2$ et $12 + \frac{1.64S_n}{\sqrt{35}} \simeq 12.8$. Or, $11.1 \notin [11.2; 12.8]$ donc notre échantillon n'est pas conforme, on rejette $\mathcal{H}_0$. Cela ne garantit pas que $\mathcal{H}_0$ est fausse, mais cela veut dire que si $\mathcal{H}_0$ était vraie, alors 9 fois sur 10, on n'aurait pas rejeté et seule une fois sur 10, on l'aurait rejeté à tort.
- Pour $\varepsilon = 0.05$, $\Phi^{-1}(1 - \varepsilon/2) = \Phi^{-1}(0.975) \simeq 1.95$. on obtient l'intervalle $[11.0; 13.0]$, donc on ne peut plus rejeter l'hypothèse nulle. Cela ne veut pas dire qu'elle est vraie, seulement qu'on ne réussit pas à affirmer qu'elle est fausse avec assez de conviction.

Remarque 7. Si on dispose d'un intervalle I (dépendant de nos observations) tel que $\mathbb{P}(\mathbb{E}(X) \in I) \geq 1 - \varepsilon$, on dit que I est un **intervalle de confiance au niveau** de confiance ε . Faire un test d'adéquation, consiste à trouver un tel intervalle I , puis tester si la valeur censée valoir $\mathbb{E}(X)$ (ou pour tout autre quantité que l'on souhaiterait estimer) appartient ou non à cet intervalle. Quand ce n'est pas le cas, on rejette cette valeur (avec une probabilité de ε de se tromper si cette valeur était en fait la bonne).

3.2 Test de conformité à une proportion

Dans une forêt, se trouvent deux types d'écureuils : les roux et les gris. La proportion d'écureuils roux est connue et vaut $p_0 = 87\%$. À côté de cette forêt, il y a un zoo avec des écureuils (roux et gris), on a perdu les archives et on ignore si les écureuils venaient de cette forêt ou non (on suppose que la proportion d'écureuils roux est constante dans le zoo et dans la forêt). On prélève un échantillon de taille n d'écureuils avec $(X_1, X_2, \dots, X_n)$ une famille de n variables aléatoires $X_i \sim \mathcal{B}(p)$ indépendantes (X_i vaut 1 si le i -ième écureuil prélevé du zoo est roux) et $\mathbb{E}(X) = p$ et $\mathbb{V}(X) = p(1 - p)$ où p désigne la proportion d'écureuils roux dans le zoo. Il y a ainsi deux cas : soit $p \neq p_0$, dans ce cas on conclut qu'ils ne viennent pas de cette forêt, soit $p = p_0$ et dans ce cas, on ne peut répondre à la question (ils pourraient venir de cette forêt ou d'une autre qui a des proportions similaires). On pose donc l'hypothèse nulle $\mathcal{H}_0 : p = p_0$. On cherche un intervalle de confiance de $p = \mathbb{E}(X)$. Si p_0 (que l'on connaît) n'est pas cet intervalle, on rejettera l'hypothèse $\mathcal{H}_0$. Si n est assez

grand, d'après le théorème central limite, la loi de $M_n^* = \frac{M_n - p}{\sigma/\sqrt{n}}$ est approximée par une loi normale centrée réduite. Soit $y_0 > 0$:

$$\mathbb{P}(-y_0 \leq M_n^* \leq y_0) = \mathbb{P}\left(-y_0 \leq \frac{M_n - p}{\frac{\sigma}{\sqrt{n}}} \leq y_0\right) = \mathbb{P}\left(M_n - y_0 \frac{\sqrt{p(1-p)}}{\sqrt{n}} \leq p \leq M_n + y_0 \frac{\sqrt{p(1-p)}}{\sqrt{n}}\right)$$

Ainsi, p est dans un certain intervalle avec une certaine probabilité. Malheureusement, cet intervalle dépend lui-même de p et p est non connu. On utilise ainsi la majoration $p(1-p) \leq 1/4$, on obtient un intervalle plus grand :

$$\mathbb{P}(-y_0 \leq M_n^* \leq y_0) \leq \mathbb{P}\left(M_n - \frac{y_0}{2\sqrt{n}} \leq p \leq M_n + \frac{y_0}{2\sqrt{n}}\right)$$

Soit $\varepsilon \in]0; 1[$, on cherche $y_0 > 0$ tel que $\mathbb{P}(-y_0 \leq M_n^* \leq y_0) \geq 1 - \varepsilon$. Or, par symétrie et par croissance stricte de Φ :

$$\mathbb{P}(-y_0 \leq M_n^* \leq y_0) \simeq \Phi(y_0) - \Phi(-y_0) = 2\Phi(y_0) - 1 \geq 1 - \varepsilon \iff \Phi(y_0) \geq 1 - \frac{\varepsilon}{2} \iff y_0 \geq \Phi^{-1}\left(1 - \frac{\varepsilon}{2}\right)$$

Sur un échantillon de 65 écureuils, 43 sont roux, ainsi, ici $n = 65$ et $M_n = \frac{43}{65}$. On donne $\Phi(1.64) \simeq 0.95$ et $\Phi(1.95) \simeq 0.975$:

- Pour $\varepsilon = 0.1$, $\Phi^{-1}(1 - \varepsilon/2) = \Phi^{-1}(0.95) \simeq 1.64$. On prend donc $y_0 = 1.64$, ainsi, $\frac{43}{65} - \frac{1.64}{2\sqrt{65}} \simeq 0.55$ et $\frac{43}{65} + \frac{1.64}{2\sqrt{65}} \simeq 0.76$. Or, $0.87 \notin [0.55; 0.76]$, on rejette donc $\mathcal{H}_0$.
- Pour $\varepsilon = 0.05$, $\Phi^{-1}(1 - \varepsilon/2) = \Phi^{-1}(0.975) \simeq 1.95$. on obtient l'intervalle $[0.6; 0.89]$ et on ne peut donc plus rejeter l'hypothèse nulle.