

STT 1000 - STATISTIQUES

ARTHUR CHARPENTIER



Fonction de répartition et quantiles

Fonction de répartition (cdf)

Let X denote a random variable, its **cumulative distribution function (c.d.f.)** is

$$F(x) = \mathbb{P}(X \leq x), \text{ for all } x \in \mathbb{R}.$$

More formally, $F(x) = \mathbb{P}(\{\omega \in \Omega | X(\omega) \leq x\})$.

Observe that

- F is an increasing function on \mathbb{R} with values in $[0, 1]$,
- $\lim_{x \rightarrow -\infty} F(x) = 0$ and $\lim_{x \rightarrow +\infty} F(x) = 1$.

Quantile

Let X denote a random variable, its **quantile function** is

$$Q(p) = F^{-1}(p) = \inf\{x \in \mathbb{R} : F(x) > p\}, \text{ for all } p \in [0, 1]$$

Densité

The **density** or **probability function** of X is

$$f(x) = \begin{cases} \frac{dF(x)}{dx} = F'(x) & \text{in the (absolutely) continuous case} \\ \mathbb{P}(X = x) & \text{in the discrete case, } x \in \mathbb{N} \\ dF(x), & \text{in a more general context} \end{cases}$$

Example: if $X \sim \mathcal{U}_{[0,1]}$, $f(x) = \mathbf{1}_{[0,1]}(x)$, $x \in \mathbb{R}$.

Example: if $X \sim \mathcal{U}_{[a,b]}$, $f(x) = \frac{1}{b-a} \mathbf{1}_{[a,b]}(x)$, $x \in \mathbb{R}$.

Moments (Espérance & Moyenne)

Espérance

$$\mathbb{E}[Y] = \int_{\mathbb{R}} yf(y) dy, \text{ ou } \sum_{y \in \mathbb{N}} yf(y) \text{ si } X \sim f$$

Example: if $X \sim \mathcal{U}_{[0,1]}$, $\mathbb{E}(X) = \frac{1}{2}$

La version empirique, pour un échantillon $\{y_1, y_2, \dots, y_n\}$ est la moyenne

Moyenne (empirique)

$$\bar{y} = \sum_{i=1}^n y_i \frac{1}{n} = \frac{1}{n} \sum_{i=1}^n y_i$$

Moments (Variance)

Variance

$$\text{Var}[Y] = \mathbb{E}[(Y - \mathbb{E}[Y])^2] = \int_{\mathbb{R}} (y - \mathbb{E}[Y])^2 f(y) dy,$$

Example: if $X \sim \mathcal{U}_{[0,1]}$, $\text{Var}(X) = \frac{1}{12}$

Note: $\text{Var}[Y] = \mathbb{E}[Y^2] - \mathbb{E}[Y]^2$.

La version empirique, pour un échantillon $\{y_1, y_2, \dots, y_n\}$ est aussi appelée variance

Variance

$$s^2 = \sum_{i=1}^n (y_i - \bar{y})^2 \frac{1}{n} = \frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^2$$

Vecteur aléatoire

Let $\mathbf{Z} = (X, Y)$ be a random vector. The cumulative distribution function of \mathbf{Z} is

$$F(\mathbf{z}) = F(x, y) = \mathbb{P}(X \leq x, Y \leq y), \text{ for all } \mathbf{z} = (x, y) \in \mathbb{R} \times \mathbb{R}.$$

Let $\mathbf{Z} = (X, Y)$ be a random vector. The density of \mathbf{Z} is

$$f(\mathbf{z}) = f(x, y) = \begin{cases} \frac{\partial^2 F(x, y)}{\partial x \partial y} & \text{in the continuous case, } \mathbf{z} = (x, y) \in \mathbb{R}^2 \\ \mathbb{P}(X = x, Y = y) & \text{in the discrete case, } \mathbf{z} = (x, y) \in \mathbb{N}^2 \end{cases}$$

Indépendance

X et Y sont indépendants ($X \perp\!\!\!\perp Y$) si et seulement si

$$\begin{cases} F(x, y) &= F_X(x) \cdot F_Y(y), \quad \forall x, y \in \mathbb{R} \\ f(x, y) &= f_X(x) \cdot f_Y(y), \quad \forall x, y \in \mathbb{R} \end{cases}$$

Vecteur aléatoire

$$\text{Var}(\mathbf{Z}) = \mathbb{E}[(\mathbf{Z} - \mathbb{E}[\mathbf{Z}])(\mathbf{Z} - \mathbb{E}[\mathbf{Z}])^\top] = \begin{pmatrix} \text{Var}[X] & \text{Cov}[X, Y] \\ \text{Cov}[X, Y] & \text{Var}[Y] \end{pmatrix}$$

Si $X \perp\!\!\!\perp Y$, $\text{Cov}[X, Y] = 0$... mais la réciproque est fausse.

Indépendants et Identiquement Distribués (i.i.d.)

X_1, X_2, \dots, X_n sont i.i.d. si

$$\begin{cases} X_1, \dots, X_n \text{ sont indépendants : } f(x_1, \dots, x_n) = \prod_{i=1}^n f_i(x_i) \\ X_1, \dots, X_n \text{ ont la même loi : } f_i = f \ \forall i \end{cases}$$

Lois Binomiales

Loi de Bernoulli $\mathcal{B}(p)$

$Y \sim \mathcal{B}(p)$:

$$\mathbb{P}[Y = y] = p^y(1-p)^{1-y} \begin{cases} p \text{ si } y = 1 \\ 1 - p \text{ si } y = 0 \end{cases}, \text{ où } y \in \{0, 1\}$$

cf loi de Bernoulli, où $p = \mathbb{P}[Y = 1] = \mathbb{E}[Y] \in [0, 1]$.

Loi binomiale $\mathcal{B}(n, p)$

$Y \sim \mathcal{B}(n, p)$:

$$\mathbb{P}[Y = y] = \binom{n}{y} p^y(1-p)^{n-y} \text{ où } y \in \{0, 1, 2, \dots, n\}$$

cf loi binomiale, où $\mathbb{E}[Y] = np$.

Lois Binomiales

$Y \sim \mathcal{B}(n, p)$:

$$\mathbb{P}[Y = y] = \binom{n}{y} p^y (1 - p)^{n-y}$$

où $y \in \{0, 1, 2, \dots, n\}$ cf loi binomiale,

où $\mathbb{E}[Y] = np$ et $\text{Var}[Y] = np(1 - p)$.

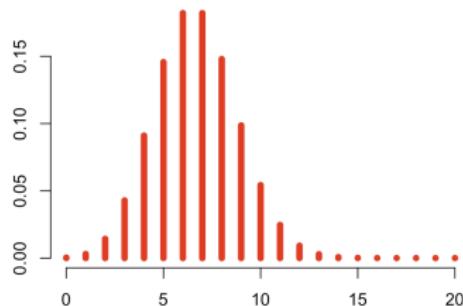
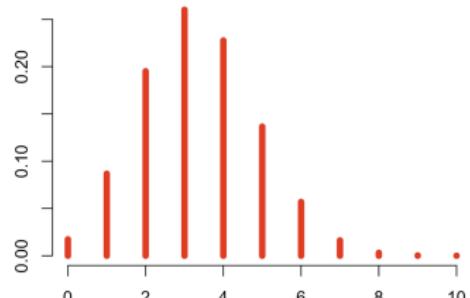
Note: when $pn \sim \lambda$ and n large enough,

$$\mathcal{B}(n, p) \approx \mathcal{P}(\lambda), \text{ where } \lambda = \frac{p}{n}$$

and when $n \rightarrow \infty$,

$$\mathcal{B}(n, p) \approx \mathcal{N}(np, np(1 - p))$$

(see Galton's quincunx)



Lois Binomiales & Multinomiales

$\mathbf{Y} = (Y_1, \dots, Y_d) \sim \mathcal{M}(\mathbf{p})$ où $\mathbf{p} = (p_1, \dots, p_d)$ si

$$Y_1 + \dots + Y_d = 1 \text{ et } Y_j \sim \mathcal{B}(p_j), \quad \forall j \in \{1, \dots, d\}$$

i.e. $\mathbf{Y} = (\mathbf{1}_{C_1}, \mathbf{1}_{C_2}, \dots, \mathbf{1}_{C_d})$

$\mathbf{Y} = (Y_1, \dots, Y_d) \sim \mathcal{M}(n, \mathbf{p})$ où $\mathbf{p} = (p_1, \dots, p_d)$ si

$$Y_1 + \dots + Y_d = n \text{ et } Y_j \sim \mathcal{B}(n, p_j), \quad \forall j \in \{1, \dots, d\}$$

cf loi multinomiale. Pour

$$(y_1, \dots, y_d) \in \mathcal{S}_{d,n} = \{(y_1, \dots, y_d) \in \mathbb{N}^d : (y_1 + \dots + y_d = n)\}$$

$$\mathbb{P}[(Y_1, \dots, Y_d) = (y_1, \dots, y_d)] = \frac{n!}{y_1! \cdots y_d!} p_1^{y_1} \cdots p_d^{y_d}$$

Example: $\mathbf{Y} = (Y_0, Y_1) \sim \mathcal{M}(n, \mathbf{p})$ où $\mathbf{p} = (p_0, p_1)$.

Lois de Poisson

Loi de Poisson $\mathcal{P}(\lambda)$

$Y \sim \mathcal{P}(\lambda)$:

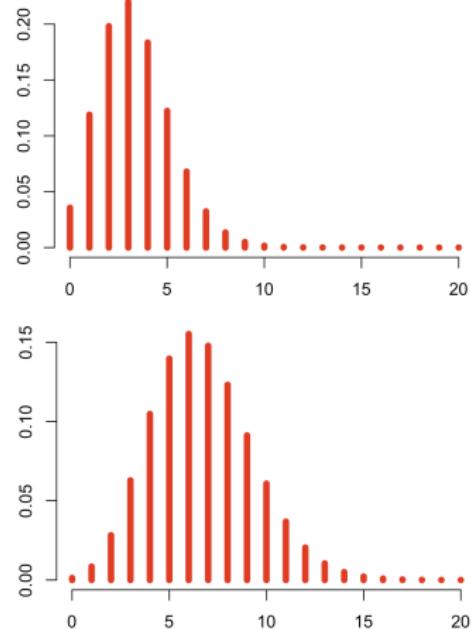
$$\mathbb{P}[Y = y] = e^{-\lambda} \frac{\lambda^y}{y!}$$

où $y \in \{0, 1, 2, \dots\}$

cf [loi de Poisson](#),

où $\mathbb{E}[Y] = \lambda$ et $\text{Var}[Y] = \lambda$.

Note: si $Y \sim \mathcal{P}(\lambda)$, $\mathbb{P}[Y = 0] = e^{-\lambda}$,
aussi $\lambda = -\log \mathbb{P}[Y = 0]$.



Distribution Géométrique

Loi géométrique $\mathcal{G}(p)$

$Y \sim \mathcal{G}(p)$, pour $y \in \{0, 1, 2, \dots\}$

$$\mathbb{P}[Y = y] = p(1 - p)^y$$

cf loi géométrique,

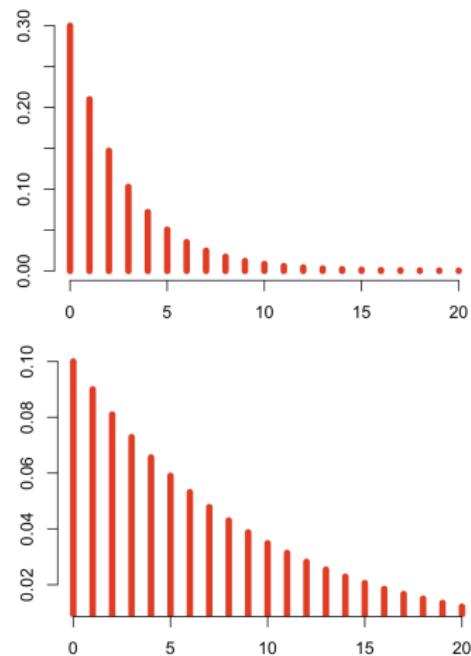
$$\mathbb{E}[Y] = \frac{1}{p} - 1 \text{ and } \text{Var}[Y] = \frac{1 - p}{p^2}$$

Note: possible alternative representation

$$\mathbb{P}[Y = y] = p(1 - p)^{y-1}$$

où $y \in \{1, 2, \dots\}$, et dans ce cas

$$\mathbb{E}[Y] = \frac{1}{p} \text{ and } \text{Var}[Y] = \frac{1 - p}{p^2}$$

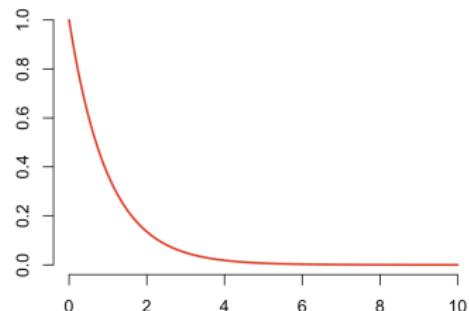


Distribution Exponentielle

Loi exponentielle $\mathcal{E}(\lambda)$

$Y \sim \mathcal{E}(\lambda)$: pour $y \in \mathbb{R}_+$,

$$f(y) = \lambda e^{-\lambda y} \text{ ou } \mathbb{P}[Y > y] = e^{-\lambda y}$$

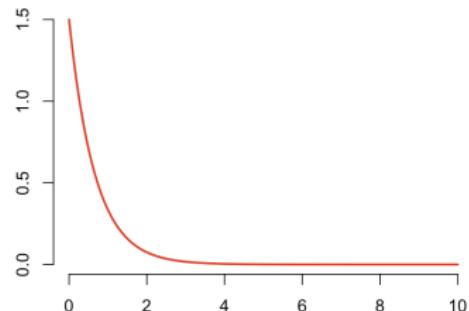


cf [loi exponentielle](#),

$$\mathbb{E}[Y] = \frac{1}{\lambda} \text{ and } \text{Var}[Y] = \frac{1}{\lambda^2}$$

Absence de mémoire: pour $h > 0$,

$$\mathbb{P}[Y > y + h | Y > y] = \frac{e^{-\lambda(y+h)}}{e^{-\lambda y}} \mathbb{P}[Y > h]$$

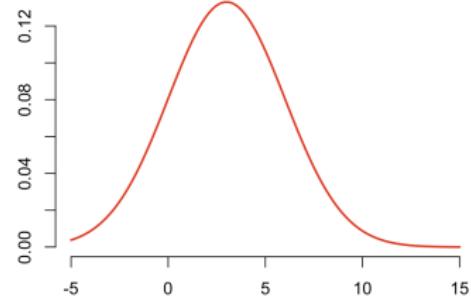


Distribution Gaussienne

Loi normale / Gaussienne $\mathcal{N}(\mu, \sigma^2)$

$Y \sim \mathcal{N}(\mu, \sigma^2)$: pour $y \in \mathbb{R}$,

$$f(y) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y - \mu)^2}{2\sigma^2}\right)$$



$$\mathbb{E}[Y] = \mu \text{ and } \text{Var}[Y] = \sigma^2$$

Si $Z \sim \mathcal{N}(0, 1)$, $Y = \mu + \sigma Z \sim \mathcal{N}(\mu, \sigma^2)$, et $\frac{Y - \mu}{\sigma} \sim \mathcal{N}(0, 1)$.

Note: $\mathbb{P}[Z \leq z] = \Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^z \exp\left(-\frac{x^2}{2}\right) dx$

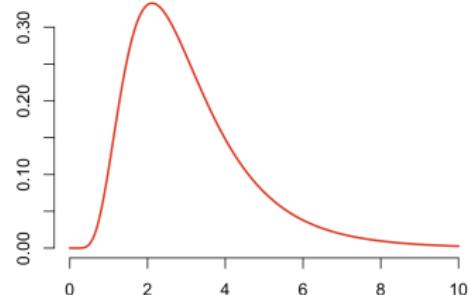
```
1 > pnorm(1.96)
2 [1] 0.9750021
3 > qnorm(.95)
4 [1] 1.644854
```

Distribution Log-Normale

Loi lognormale $LN(\mu, \sigma^2)$

$Y \sim LN(\mu, \sigma^2)$: pour $y \in \mathbb{R}_+$,

$$f(y) = \frac{1}{y\sqrt{2\pi\sigma^2}} \exp\left(-\frac{[\log(y) - \mu]^2}{2\sigma^2}\right)$$

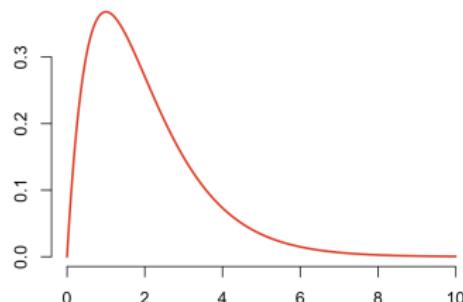


cf loi log-normale,

$$\mathbb{E}[Y] = \exp\left(\mu + \frac{\sigma^2}{2}\right)$$

$$\text{Var}[Y] = (e^{\sigma^2} - 1) \exp\left(2\mu + \sigma^2\right)$$

Si $Y \sim LN(\mu, \sigma^2)$, $\log Y \sim \mathcal{N}(\mu, \sigma^2)$



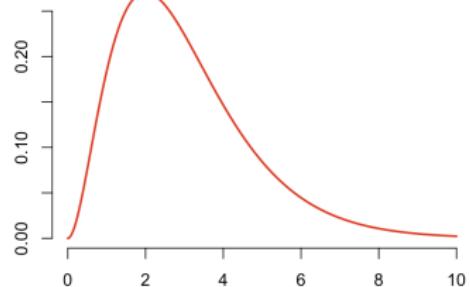
Note: $\mathbb{E}(\log Y) \neq \log \mathbb{E}(Y)$

Distribution Gamma

Loi Gamma $\mathcal{G}(\alpha, \beta)$

$Y \sim \mathcal{G}(\alpha, \beta)$: pour $y \in \mathbb{R}_+$,

$$f(y) = \frac{\beta^\alpha y^{\alpha-1} e^{-\beta y}}{\Gamma(\alpha)}$$



cf loi gamma,

$$\mathbb{E}[Y] = \frac{\alpha}{\beta} \text{ et } \text{Var}[Y] = \frac{\alpha}{\beta^2}$$

α est appelé *shape* et β *rate*

Note: si $X \sim \mathcal{G}(\nu/2, 1/2)$ alors $X \sim \chi^2(\nu)$, cf loi du chi-deux.

Convergence & Probabilité

Convergence en probabilité d'une suite (X_n)

$X_n \xrightarrow{\mathbb{P}} Y$ lorsque $n \rightarrow \infty$ si pour tout $\varepsilon > 0$,

$$\lim_{n \rightarrow \infty} \mathbb{P}[|X_n - Y| < \varepsilon] = 0$$

Convergence presque sûre d'une suite (X_n)

$X_n \xrightarrow{p.s.} Y$ lorsque $n \rightarrow \infty$ si $\mathbb{P} \left[\lim_{n \rightarrow \infty} X_n = Y \right] = 1$, p.s.

Convergence

Convergence en moyenne d'une suite (X_n)

$X_n \xrightarrow{L^p} Y$ lorsque $n \rightarrow \infty$ si $\lim_{n \rightarrow \infty} \mathbb{E}[|X_n - Y|^p] = 0$, $p \in \mathbb{N}_*$

Si $p = 2$, on parle de convergence en moyenne quadratique.

Convergence en loi d'une suite (X_n)

$X_n \xrightarrow{\mathcal{L}} Y$ lorsque $n \rightarrow \infty$ si pour tout $x \in \mathbb{R}$, $F_n(x) \rightarrow G(x)$
lorsque $n \rightarrow \infty$ où $F_n(x) = \mathbb{P}[X_n \leq x]$ et $G(x) = \mathbb{P}[Y \leq x]$

Convergence

The sequence (X_n) converges in probability towards X , denoted $X_n \xrightarrow{\mathbb{P}} X$, if

$$\forall \varepsilon > 0, \lim_{n \rightarrow \infty} \mathbb{P}(|X_n - X| > \varepsilon) = 0.$$

Let $f : \mathbb{R} \rightarrow \mathbb{R}$ be a continuous function, if $X_n \xrightarrow{\mathbb{P}} X$ then $f(X_n) \xrightarrow{\mathbb{P}} f(X)$. More generally, let $f : \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}$ be a continuous function, if $X_n \xrightarrow{\mathbb{P}} X$ and $Y_n \xrightarrow{\mathbb{P}} y \in \mathbb{R}$ then $f(X_n, Y_n) \xrightarrow{\mathbb{P}} f(X, y)$ (Slutsky Theorem).

A sufficient condition to have $X_n \xrightarrow{\mathbb{P}} a$ is that

$$\lim_{n \rightarrow \infty} \mathbb{E}[X_n] = a \text{ and } \lim_{n \rightarrow \infty} \text{Var}(X_n) = 0$$

Suppose X_i 's are i.i.d. with finite expected value $\mu = \mathbb{E}(X_i)$, then $\bar{X}_n \xrightarrow{\mathbb{P}} \mu$ as $n \rightarrow +\infty$, see Law of Large Numbers

Convergence

Loi (faible) des Grands Nombres

Soit (Y_n) une suite de variables aléatoires réelles indépendantes et de même loi admettant une espérance μ .

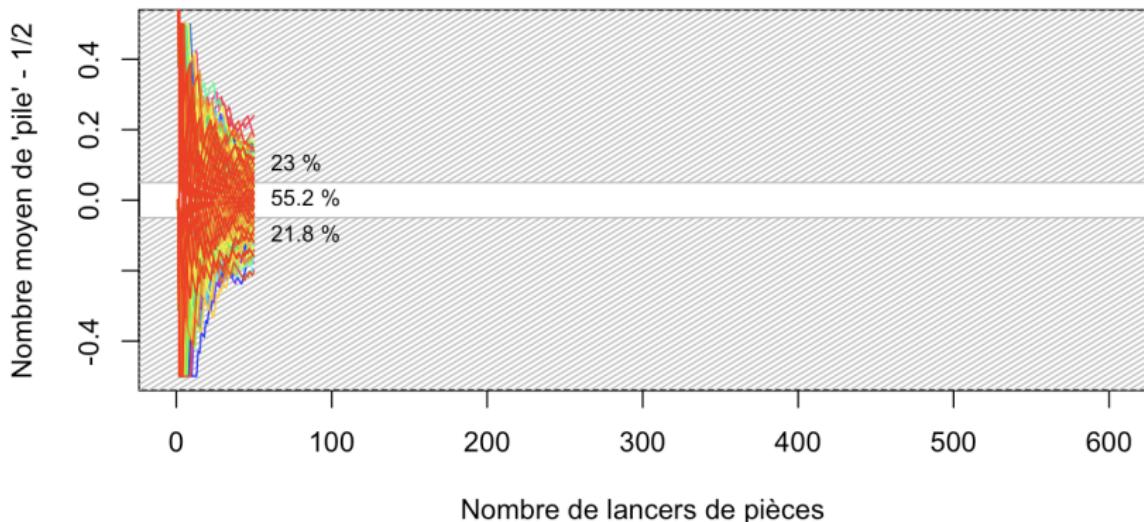
La moyenne empirique $\bar{X}_n = \frac{1}{n} \sum_{k=1}^n Y_k$ converge en probabilité vers l'espérance : pour tout $\varepsilon > 0$,

$$\lim_{n \rightarrow +\infty} \mathbb{P} \left(|\bar{X}_n - \mu| > \varepsilon \right) = 0$$

Example Lancer de pièces (pile ou face)

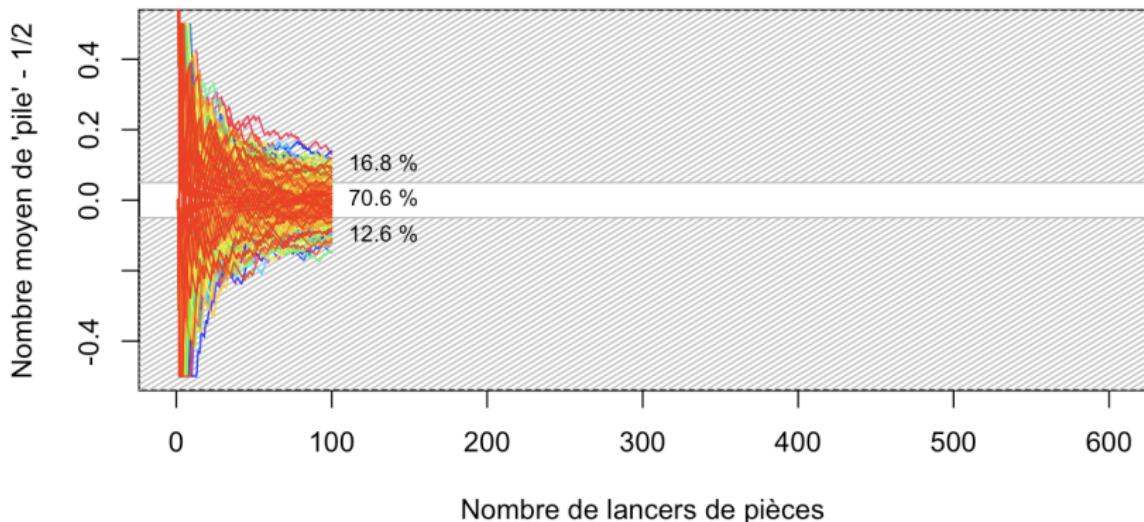
Convergence, loi des grands nombres

$$\mathbb{P} \left(\left| \bar{X}_{50} - \mu \right| > 5\% \right) \sim 45\% \text{ } (n = 50 \text{ lancers})$$



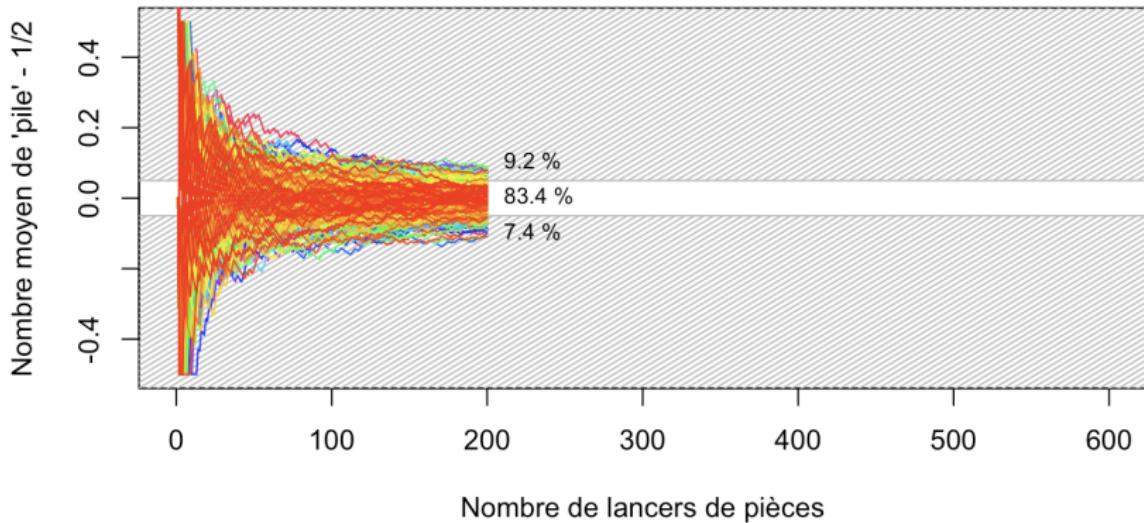
Convergence, loi des grands nombres

$$\mathbb{P} \left(\left| \bar{X}_{100} - \mu \right| > 5\% \right) \sim 30\% \text{ } (n = 100 \text{ lancers})$$



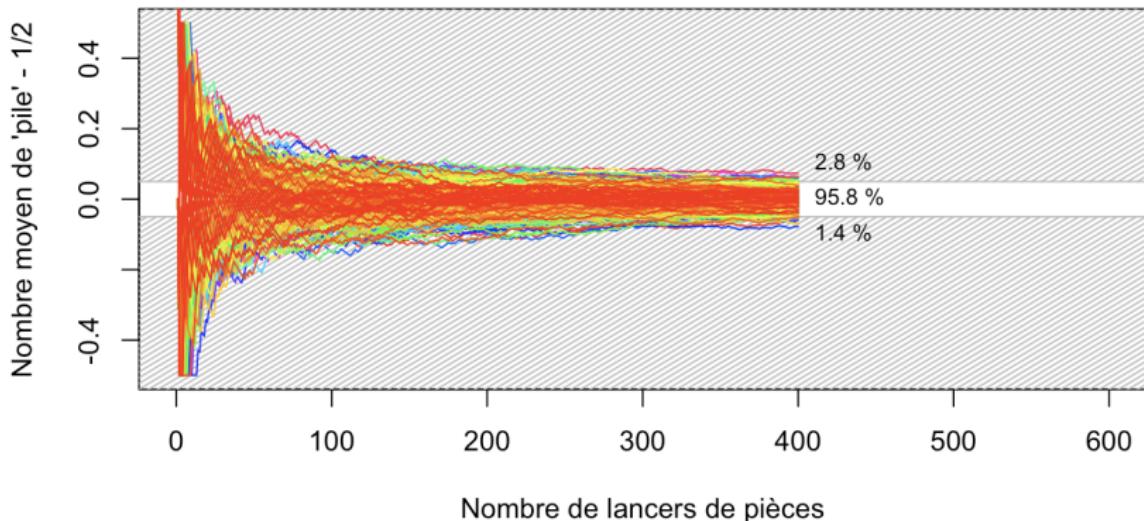
Convergence, loi des grands nombres

$$\mathbb{P} \left(\left| \bar{X}_{200} - \mu \right| > 5\% \right) \sim 17\% \text{ } (n = 200 \text{ lancers})$$



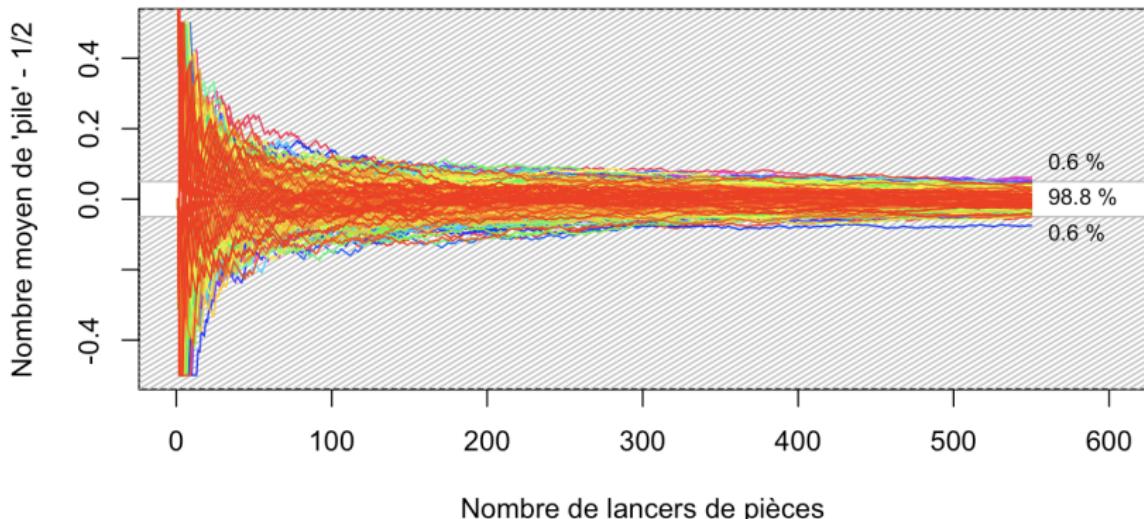
Convergence, loi des grands nombres

$$\mathbb{P} \left(\left| \bar{X}_{400} - \mu \right| > 5\% \right) \sim 6\% \text{ } (n = 400 \text{ lancers})$$



Convergence, loi des grands nombres

$$\mathbb{P} \left(\left| \bar{X}_{500} - \mu \right| > 5\% \right) \sim 1\% \text{ } (n = 500 \text{ lancers})$$



Convergence

Sequence (X_n) converges in distribution towards X , denoted $X_n \xrightarrow{\mathcal{L}} X$, if for any continuous function h

$$\lim_{n \rightarrow \infty} \mathbb{E}[h(X_n)] = \mathbb{E}(h(X)).$$

Convergence in distribution is the same as convergence of distribution function $X_n \xrightarrow{\mathcal{L}} X$ if for any $t \in \mathbb{R}$ where F_X is continuous

$$\lim_{n \rightarrow \infty} F_{X_n}(t) = F_X(t).$$

Let $X_1, X_2 \dots$ denote i.i.d. random variables with mean μ and variance σ^2 , then :

$$\frac{\bar{X}_n - \mathbb{E}(\bar{X}_n)}{\sqrt{\text{Var}(\bar{X}_n)}} = \sqrt{n} \left(\frac{\bar{X}_n - \mu}{\sigma} \right) \xrightarrow{\mathcal{L}} X \text{ where } X \sim \mathcal{N}(0, 1)$$

see Central Limit Theorem

Convergence

Théorème Central Limite

Soit (Y_n) une suite de variables aléatoires réelles indépendantes et de même loi admettant une espérance μ et une variance σ^2 .

La moyenne empirique $\bar{X}_n = \frac{1}{n} \sum_{k=1}^n Y_k$ centrée et réduite converge vers une loi normale :

$$\lim_{n \rightarrow \infty} \mathbb{P} \left(\sqrt{n} \cdot \frac{\bar{X}_n - \mu}{\sigma} \leq z \right) = \Phi(z)$$

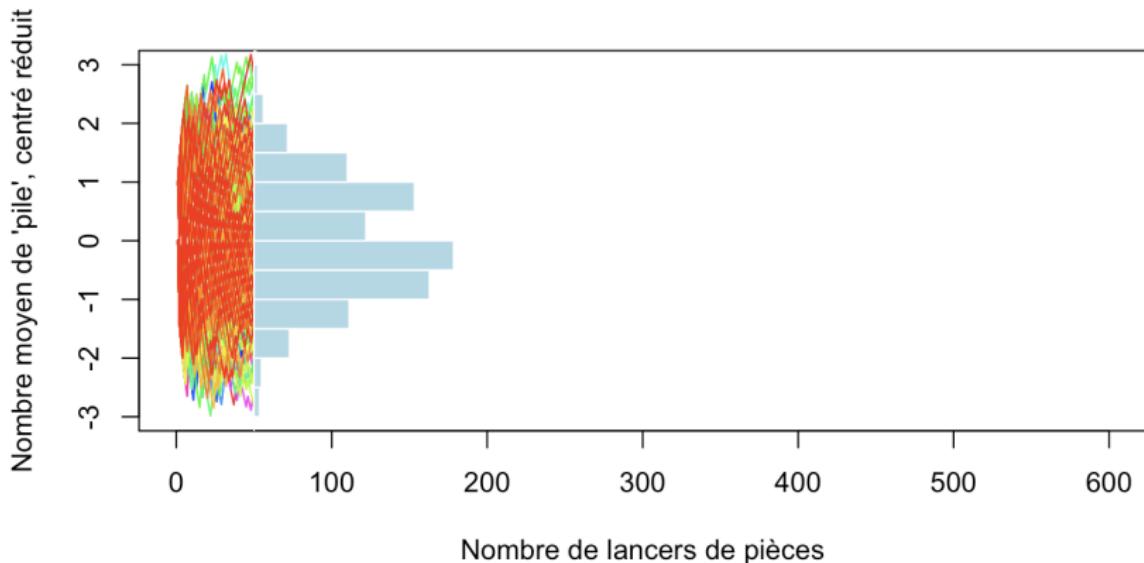
où

$$\Phi(z) = \int_{-\infty}^z \frac{1}{\sigma \sqrt{2\pi}} \exp \left[-\frac{1}{2} \left(\frac{x - \mu}{\sigma} \right)^2 \right]$$

Example Lancer de pièces (pile ou face)

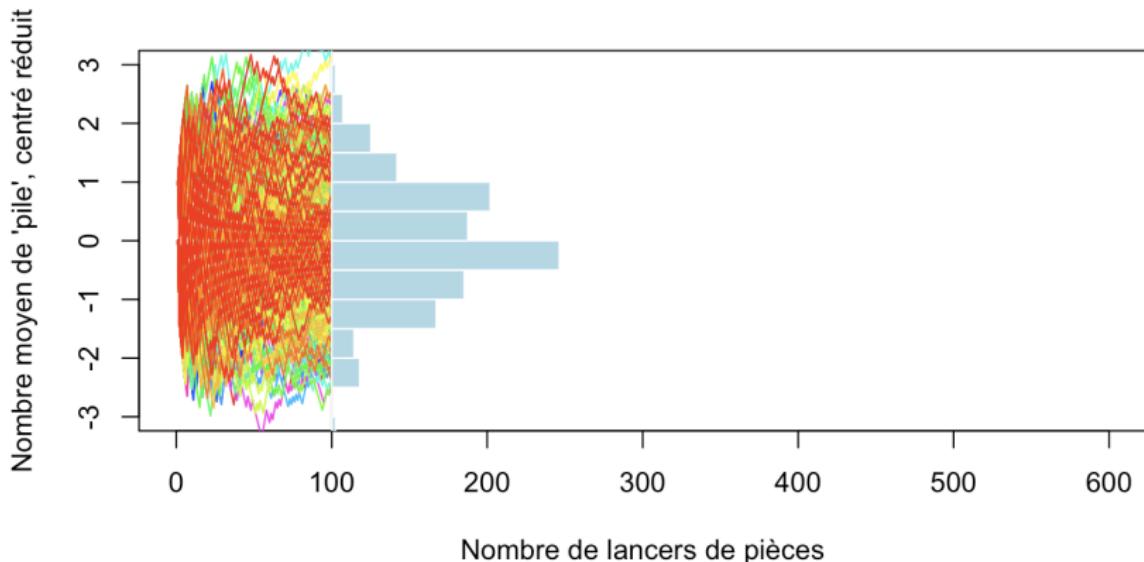
Convergence, théorème central limite

Histogramme de $\sqrt{50} \cdot \frac{\bar{X}_{50} - \mu}{\sigma}$ ($n = 50$ lancers)



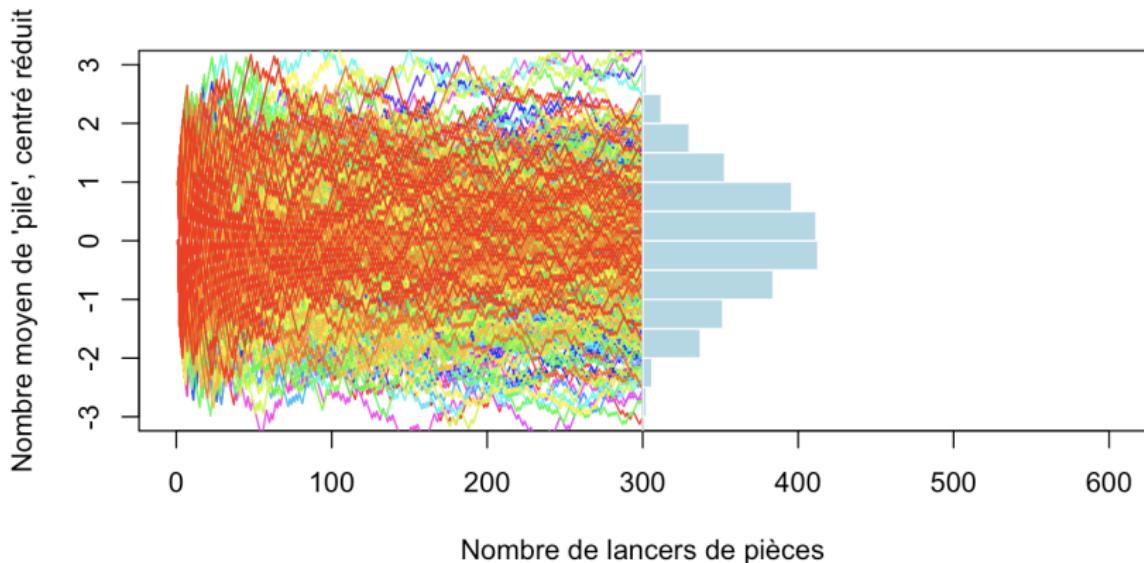
Convergence, théorème central limite

Histogramme de $\sqrt{100} \cdot \frac{\bar{X}_{100} - \mu}{\sigma}$ ($n = 100$ lancers)



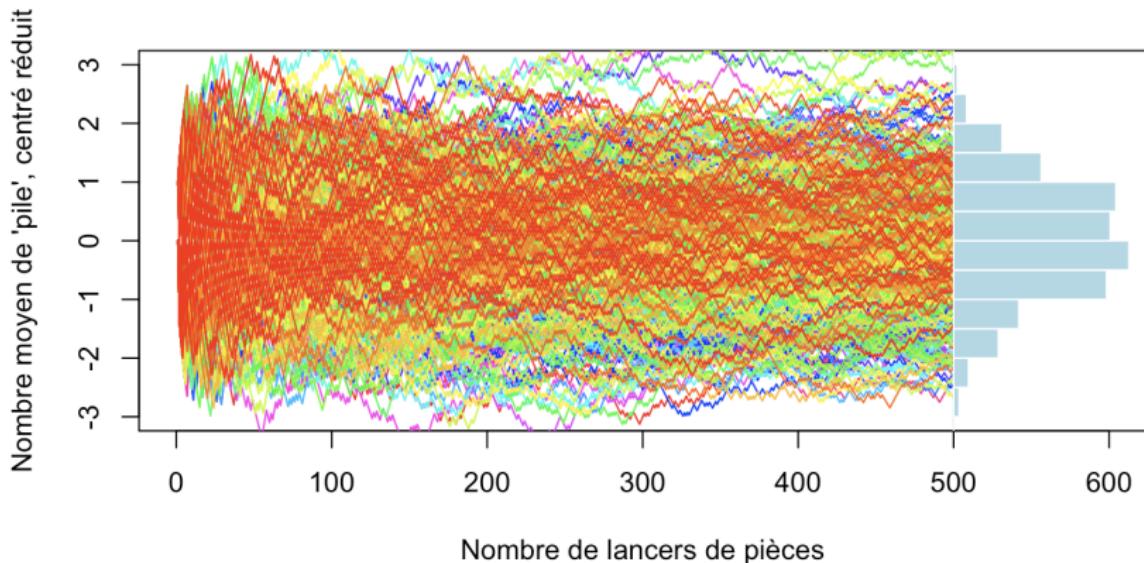
Convergence, théorème central limite

Histogramme de $\sqrt{300} \cdot \frac{\bar{X}_{300} - \mu}{\sigma}$ ($n = 300$ lancers)



Convergence, théorème central limite

Histogramme de $\sqrt{500} \cdot \frac{\bar{X}_{500} - \mu}{\sigma}$ ($n = 500$ lancers)



Transformation

Transformation

Let X be an absolutely continuous random variable with density $f(x)$. If function ϕ is a differentiable one-to-one mapping, then variable $Y = \phi(X)$ has a density g satisfying

$$g(y) = \frac{f(\phi^{-1}(y))}{\phi'(\phi^{-1}(y))}.$$

Let X be an absolutely continuous random variable with cdf F , i.e. $F(x) = \mathbb{P}(X \leq x)$. Then $Y = F(X)$ has a uniform distribution on $[0, 1]$.

Let Y be a uniform distribution on $[0, 1]$ and F denote a cdf. Then $X = F^{-1}(Y)$ is a random variable with cdf F .

Delta Method

Delta-Method

In the **univariate case**, if there is a sequence of random variables X_n satisfying $\sqrt{n}[X_n - \theta] \xrightarrow{\mathcal{L}} \mathcal{N}(0, \sigma^2)$, where θ and σ^2 are two constants, then

$$\sqrt{n}[g(X_n) - g(\theta)] \xrightarrow{\mathcal{L}} \mathcal{N}(0, \sigma^2 \cdot [g'(\theta)]^2)$$

for any function g satisfying the property that $g'(\theta)$ exists and is non-zero valued.

Faisons un développement de Taylor,

$$g(X_n) \approx g(\theta) + g'(\theta)(X_n - \theta),$$

soit, en réarrangeant et en multipliant par \sqrt{n}

$$\begin{aligned} \sqrt{n}[g(X_n) - g(\theta)] &\approx g'(\theta) \underbrace{\sqrt{n}[X_n - \theta]}_{\xrightarrow{\mathcal{L}} \mathcal{N}(0, \sigma^2)} . \end{aligned}$$

Delta Method

On conclut en rappelant que si $Z \sim \mathcal{N}(0, \sigma^2)$, $tZ \sim \mathcal{N}(0, t^2\sigma^2)$.

Delta-Method

In the **multivariate case**, if there is a sequence of random vectors \mathbf{X}_n satisfying $\sqrt{n}[\mathbf{X}_n - \boldsymbol{\theta}] \xrightarrow{\mathcal{L}} \mathcal{N}(\mathbf{0}, \Sigma)$, where $\boldsymbol{\theta}$ is a vector in \mathbb{R}^d and Σ is a symmetric positive $d \times d$ matrix, then

$$\sqrt{n}[g(\mathbf{X}_n) - g(\boldsymbol{\theta})] \xrightarrow{\mathcal{L}} \mathcal{N}(\mathbf{0}, \nabla g(\boldsymbol{\theta})^\top \Sigma \nabla g(\boldsymbol{\theta}))$$

for any function g satisfying the property that $\nabla g(\boldsymbol{\theta})$ exists and is non-zero valued.