Vecteurs aléatoires

Probabilités et statistiques

hiver 2022

Retour: variables aléatoires

Résumé de l'épisode précédent

- **Variable aléatoire** X: nombre dont la valeur dépend du hasard
- On peut parler de la **probabilité** qu'elle prenne certaines valeurs

$$0 \leq \mathbb{P}[X \in \mathcal{A}] \leq 1.$$

- Dite **continue** si pour tout x,

$$\mathbb{P}[X = x] = 0,$$

- **discrète** s'il existe une suite de valeurs (x_i) avec

$$\sum_i \underbrace{\mathbb{P}[X = x_i]}_{p_i} = 1.$$

La loi de X

- Peut être décrite grâce à la **fonction de répartition**

$$F_X(x) := \mathbb{P}[X \leq x].$$

- Fonction croissante avec

$$F_X(-\infty) = 0, \qquad F_X(+\infty) = 1.$$

- Sert à évaluer les probabilités par différence:

$$\mathbb{P}[a < X \leq b] = F_X(b) - F_X(a)$$

Mais aussi:

- Sa dérivée est la **« fonction » de densité** $f_X(x)$
- Positive, aire totale sous la courbe 1
- Sert à évaluer les probabilités par intégration:

$$\mathbb{P}[X \in \mathcal{A}] = \int_{x \in \mathcal{A}} f_X(x)\,\mathrm{d}x.$$

En particulier

- Cas discret: F_X est continue par morceaux,

$$f_X(x) = \sum_i p_i \, \delta(x - x_i),$$

  les intégrales se ramènent à des sommes (finies ou non)

- Cas continu: F_X est continue, f_X est une vraie fonction
- Cas mixte: un peu des deux

Mesures de tendance centrale

- L'**espérance**

$$\mu = \mathbb{E}[X] = \int_{-\infty}^{+\infty} x\, f_X(x)\, \mathrm{d}x$$

- mais aussi le **mode**: $f_X(x_m) = \max f_X$

- la **médiane**: $F_X(x_M) = \frac{1}{2}$

Mesure de dispersion

Pour quantifier la dispersion d'une v.a. X, considérons l'espérance de la déviation par rapport à son espérance:

$$\begin{aligned}\mathbb{E}[X - \mathbb{E}[X]\,] &= \mathbb{E}[X] - \mathbb{E}[\mathbb{E}[X]] \\ &= \mathbb{E}[X] - \mathbb{E}[X] \\ &= 0\end{aligned}$$

Oups !

Meilleure idée:

Définition

La **variance** d'une variable aléatoire X est

$$\mathrm{Var}(X) := \mathbb{E}[(X - \mathbb{E}[X])^2] \geq 0.$$

Notation courante: $\mu = \mathbb{E}[X]$, **écart-type** $\sigma := \sqrt{\mathrm{Var}(X)} \geq 0$

Proposition

$$\mathrm{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2$$

Preuve:

$$\mathrm{Var}(X) = \mathbb{E}[(X - \mathbb{E}[X])^2]$$

$$= \mathbb{E}\left[X^2 - 2\,\mathbb{E}[X]X + \mathbb{E}[X]^2\right]$$

$$= \mathbb{E}[X^2] - 2\,\mathbb{E}[X]\,\mathbb{E}[X] + \mathbb{E}[X]^2$$

$$= \mathbb{E}[X^2] - \mathbb{E}[X]^2.$$

Exemples

- $X \sim \mathcal{B}(n,p) \implies \mathrm{Var}(X) = np(1-p)$
- $X \sim \mathcal{G}(p) \implies \mathrm{Var}(X) = \frac{1-p}{p^2}$
- $X \sim \mathcal{N}(\mu,\sigma^2) \implies \mathrm{Var}(X) = \sigma^2$
- \vdots

Écart à l'espérance

L'écart-type est l'unité naturelle pour mesurer la distance à l'espérance.

Théorème (Bienaymé-Tchebychev)

Pour toute variable aléatoire X (d'espérance et variance finies),

$$\mathbb{P}\left[\left|X-\mu\right| \geq n\sigma\right] \leq \frac{1}{n^2}.$$

Bienaymé-Tchebychev

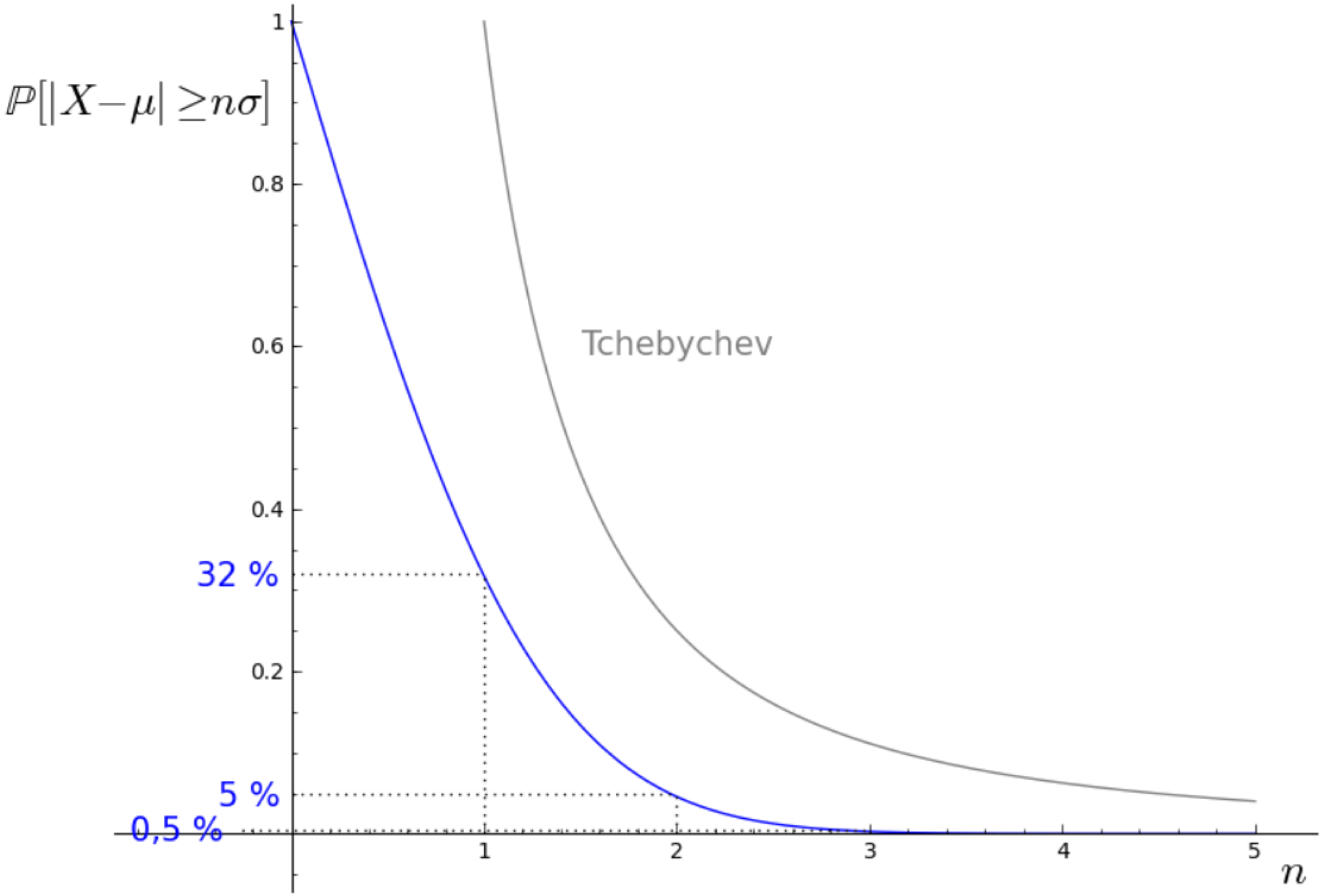
$$\mathbb{P}\Big[|X-\mu| \geq n\sigma\Big] \leq \frac{1}{n^2}.$$

Preuve (cas centré réduit): Si $\mathcal{A}$ désigne l'évènement $|X| \geq n$,

$$1 = \mathbb{E}[X^2] = \int_{x\in\mathcal{A}} x^2 f_X(x)\,\mathrm{d}x + \int_{x\notin\mathcal{A}} x^2 f_X(x)\,\mathrm{d}x$$

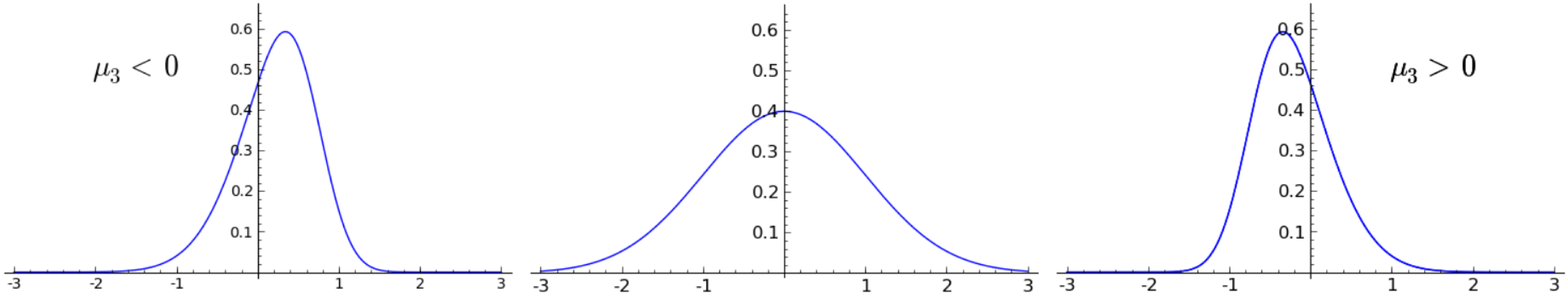
$$\geq \int_{x\in\mathcal{A}} x^2 f_X(x)\,\mathrm{d}x \geq n^2\,\mathbb{P}[\mathcal{A}].$$

La loi normale fait bien mieux



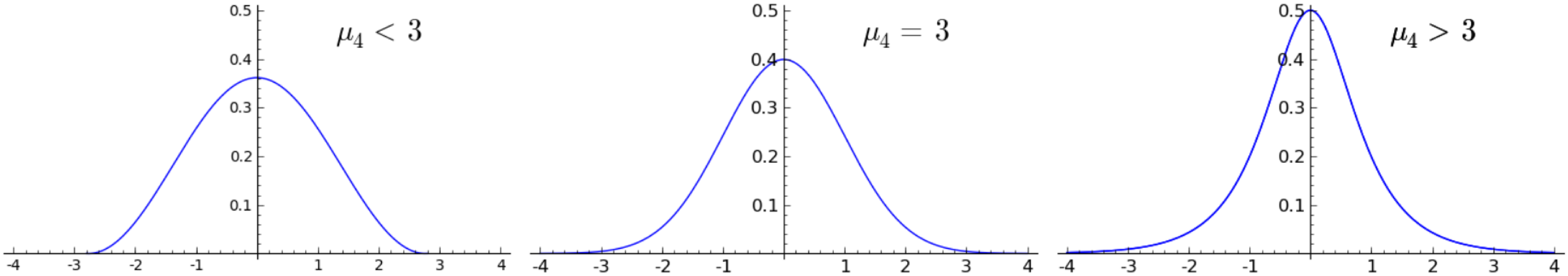
Statistiques d'ordre supérieur

$\mu_3 = \mathbb{E}[X^3]$ **coefficient de dissymétrie**



$$(\mu_1 = 0,\ \mu_2 = 1)$$

$$\mu_4 = \mathbb{E}[X^4]$$ **coefficient d'aplatissement**



$$(\mu_1 = \mu_3 = 0,\ \mu_2 = 1)$$

$$\vdots$$

$$\mu_n = \mathbb{E}[X^n]$$ **n^e moment**

Théorème

La loi d'une variable aléatoire à moments finis est entièrement déterminée par ceux-ci.

Pour s'en convaincre, un outil puissant: la **fonction génératrice des moments** de X

$$g_X(t) = \mathbb{E}[e^{tX}] = \int_{-\infty}^{+\infty} e^{tx} f_X(x)\,\mathrm{d}x.$$

En posant $t = -2\pi i f$ on reconnaît la transformée de Fourier de f_X !

$$g_X(t) = \widehat{f_X}\Big(-\frac{t}{2\pi i}\Big)$$

Preuve: Par transformée inverse, on sait que $g_X(t)$ et $f_X(x)$ contiennent exactement la même information.

$$g_X(t) := \widehat{f_X}\left(-\frac{t}{2\pi i}\right) = \int_{-\infty}^{+\infty} e^{tx} f_X(x)\,\mathrm{d}x$$

Or:

$$g_X(t) = \mathbb{E}[e^{tX}] = \mathbb{E}\left[\sum_{n=0}^{\infty} \frac{(tX)^n}{n!}\right] = \sum_{n=0}^{\infty} \frac{t^n}{n!} \underbrace{\mathbb{E}[X^n]}_{\mu_n}.$$

Conclusion: les μ_n déterminent entièrement g_X (et vice-versa), en fait:

$$\mu_n = g_X^{(n)}(0).$$

« Décrire la loi de X »

- Donner F_X
- Donner f_X (ou les p_i dans le cas discret)
- Donner la suite des moments

$$\mu_n = \mathbb{E}[X^n], \qquad n \in \mathbb{N}$$

- Ou encore, la fonction génératrice

$$g_X(t) = \mathbb{E}[e^{tX}] = 1 + \mu\, t + \mu_2 \frac{t^2}{2} + \mu_3 \frac{t^3}{6} + \cdots$$

Quelques lois courantes

- **Bernoulli** $\mathcal{B}(p)$
- **binomiale** $\mathcal{B}(n, p)$
- **géométrique** $\mathcal{G}(p)$
- **Poisson** $\mathcal{P}(\lambda)$

$\vdots$

- **exponentielle** $\mathcal{E}(\lambda)$
- **uniforme** $\mathcal{U}([a, b])$
- **normale** $\mathcal{N}(\mu, \sigma^2)$

$\vdots$

Vecteurs aléatoires

Plusieurs variables aléatoires

Intéressons-nous maintenant à deux variables aléatoires

vues comme un **couple** ou **vecteur aléatoire**

$$(X, Y) : \Omega \longrightarrow \mathbb{R}^2.$$

(On généralisera ensuite facilement $2 \mapsto n$)

Exemple discret

On lance deux pièces: X_1 le résultat de la première, X_2 de la seconde

(X_1, X_2)	0	1	Σ
0	0,25	0,25	0,5
1	0,25	0,25	0,5
Σ	0,5	0,5	1

Un peu plus intéressant

Couple (X, Y) avec $X = X_1, Y = X_1 + X_2$

(X, Y)	0	1	2	Σ
0	0,25	0,25	0	0,5
1	0	0,25	0,25	0,5
Σ	0,25	0,5	0,25	1

Y donne de l'information sur X, et vice-versa

Terminologie

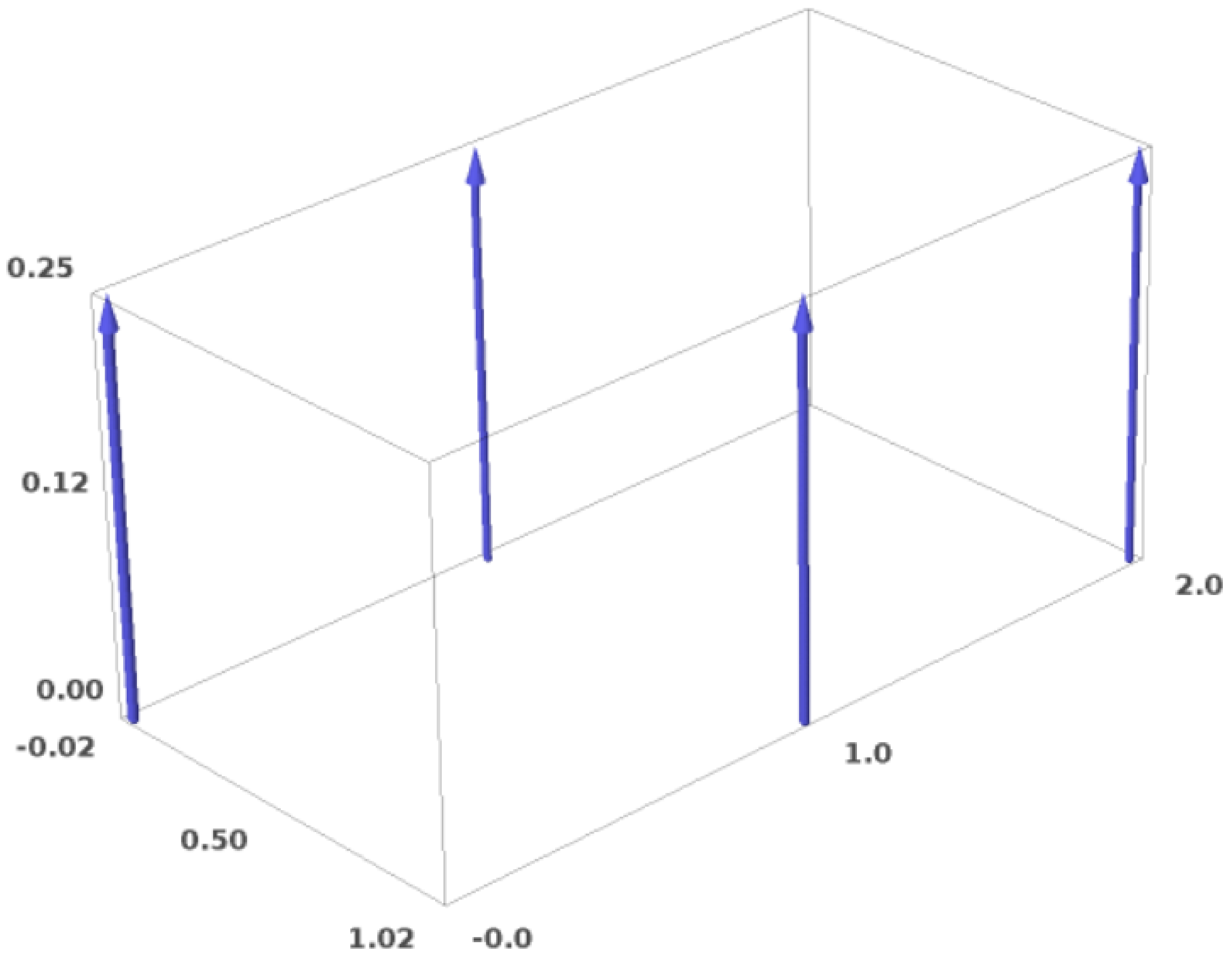
- **Probabilités conditionnelles**, *e.g.*

$$\mathbb{P}[Y = 2 \mid X = 1] = \frac{\mathbb{P}[(X, Y) = (1, 2)]}{\mathbb{P}[X = 1]} = \frac{0{,}25}{0{,}5} = 0{,}5$$

$$\mathbb{P}[X = 0 \mid Y = 0] = \frac{\mathbb{P}[(X, Y) = (0, 0)]}{\mathbb{P}[Y = 0]} = \frac{0{,}25}{0{,}25} = 1$$

- **Probabilités marginales** (somme par ligne ou colonne)

Représentation graphique



Loi conjointe

Définition

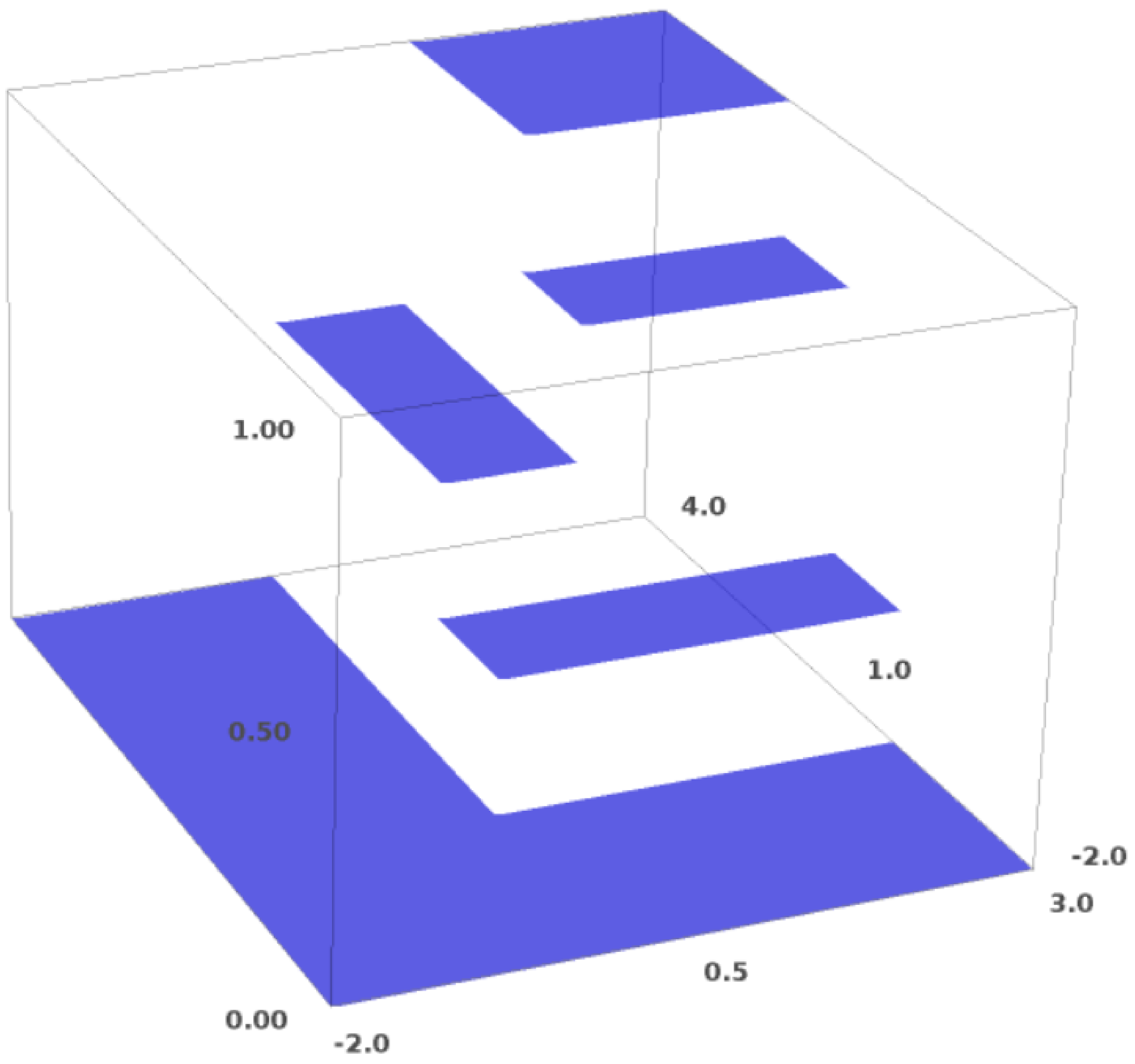
Fonction de répartition

$$F(x, y) = \mathbb{P}[X \leq x \text{ et } Y \leq y]$$

Fonction de densité

$$f(x, y) = \frac{\partial^2 F}{\partial x \partial y}$$

Exemple: fonction de répartition



Utilité

On calcule les probabilités par intégration double

$$\mathbb{P}[(X,Y) \in \mathcal{A}] = \iint_{\mathcal{A}} f(x,y)\,\mathrm{d}x\,\mathrm{d}y$$

Lois marginales:

$$F_X(x) = \mathbb{P}[X \leq x] = \int_{-\infty}^{x} \int_{-\infty}^{+\infty} f(u,y)\,\mathrm{d}y\,\mathrm{d}u$$

$$\implies f_X(x) = \int_{-\infty}^{+\infty} f(x,y)\,\mathrm{d}y$$

et de même pour $f_Y(y)$

Indépendance

Définition

X et Y sont dites **indépendantes** si pour tout $\mathcal{A}, \mathcal{B} \subseteq \mathbb{R}$

$$\mathbb{P}[X \in \mathcal{A} \text{ et } Y \in \mathcal{B}] = \mathbb{P}[X \in \mathcal{A}] \cdot \mathbb{P}[Y \in \mathcal{B}]$$

En d'autres termes:

$$\mathbb{P}[X \in \mathcal{A} \,|\, Y \in \mathcal{B}] = \frac{\mathbb{P}[X \in \mathcal{A} \text{ et } Y \in \mathcal{B}]}{\mathbb{P}[Y \in \mathcal{B}]} = \mathbb{P}[X \in \mathcal{A}]$$

Savoir quelque chose sur l'une n'apporte aucune information sur l'autre

Proposition

X et Y sont indépendantes

$$\iff F(x, y) = F_X(x) \cdot F_Y(y)$$

$$\iff f(x, y) = f_X(x) \cdot f_Y(y)$$

Ce qui simplifie grandement les calculs

Preuve: $(1) \implies (2)$ Par définition de F

$(2) \implies (3)$ en dérivant

$(3) \implies (1)$ en intégrant

Propriétés de l'espérance

Proposition

- Pour tout couple de variables aléatoires,

$$\mathbb{E}[X+Y] = \mathbb{E}[X] + \mathbb{E}[Y].$$

Preuve:

$$\mathbb{E}[X+Y] = \iint_{\mathbb{R}^2} (x+y) f(x,y)\,\mathrm{d}x\,\mathrm{d}y$$

$$= \int_{-\infty}^{+\infty} x \underbrace{\int_{-\infty}^{+\infty} f(x,y)\,\mathrm{d}y}_{f_X(x)}\,\mathrm{d}x + \int_{-\infty}^{+\infty} y \underbrace{\int_{-\infty}^{+\infty} f(x,y)\,\mathrm{d}x}_{f_Y(y)}\,\mathrm{d}y = \mathbb{E}[X] + \mathbb{E}[Y]$$

Propriétés de l'espérance

Proposition

- **Si** X et Y sont indépendantes,

$$\mathbb{E}[X \cdot Y] = \mathbb{E}[X] \cdot \mathbb{E}[Y].$$

Preuve:

$$\mathbb{E}[X \cdot Y] = \iint_{\mathbb{R}^2} x\, y\, f(x,y)\, \mathrm{d}x\, \mathrm{d}y \overset{\text{ind}}{=} \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} x\, y\, f_X(x)\, f_Y(y)\, \mathrm{d}x\, \mathrm{d}y$$

$$= \left(\int_{-\infty}^{+\infty} x\, f_X(x)\, \mathrm{d}x \right) \left(\int_{-\infty}^{+\infty} y\, f_Y(y)\, \mathrm{d}y \right) = \mathbb{E}[X] \cdot \mathbb{E}[Y]$$

Question: la réciproque est-elle vraie ? (indice: non)

L'important cas de la somme

Proposition

Si X et Y sont indépendantes, alors

$$g_{X+Y}(t) = g_X(t) \cdot g_Y(t)$$

Preuve: e^{tX} et e^{tY} sont aussi indépendantes donc

$$g_{X+Y}(t) = \mathbb{E}[e^{t(X+Y)}] = \mathbb{E}[e^{tX} \cdot e^{tY}] = \mathbb{E}[e^{tX}] \cdot \mathbb{E}[e^{tY}] = g_X(t) \cdot g_Y(t).$$

En d'autres termes:

$$\boxed{f_{X+Y} = f_X * f_Y}$$

Statistiques

$$\mathbb{E}[X],\ \mathbb{E}[Y],\ \mathrm{Cov}(X,Y),\ \mathrm{Cor}(X,Y)$$

Variance conjointe

Définition

La **covariance** du couple (X, Y) est

$$\mathrm{Cov}(X, Y) = \sigma_{XY} := \mathbb{E}\left[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])\right]$$

Remarque: $\mathrm{Var}(X) = (\sigma_X)^2 = \sigma_{XX} = \mathrm{Cov}(X, X)$

Indépendance et covariance

Proposition

$$\mathrm{Cov}(X, Y) = \mathbb{E}[X\,Y] - \mathbb{E}[X]\,\mathbb{E}[Y]$$

D'où:

Corollaire

Si X et Y sont indépendantes, **alors** $\mathrm{Cov}(X, Y) = 0$.

Attention: la réciproque n'est pas vraie !

Variance d'une somme

Proposition (Al-Kashi)

$$\mathrm{Var}(X+Y) = \mathrm{Var}(X) + 2\,\mathrm{Cov}(X,Y) + \mathrm{Var}(Y)$$

En particulier, si X et Y sont indépendantes (Pythagore):

$$\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$$

Ou encore:

$$\boxed{\sigma_{X+Y} = \sqrt{\sigma_X^2 + \sigma_Y^2}}$$

Corrélation

On préfère souvent une version normalisée de la covariance:

Définition

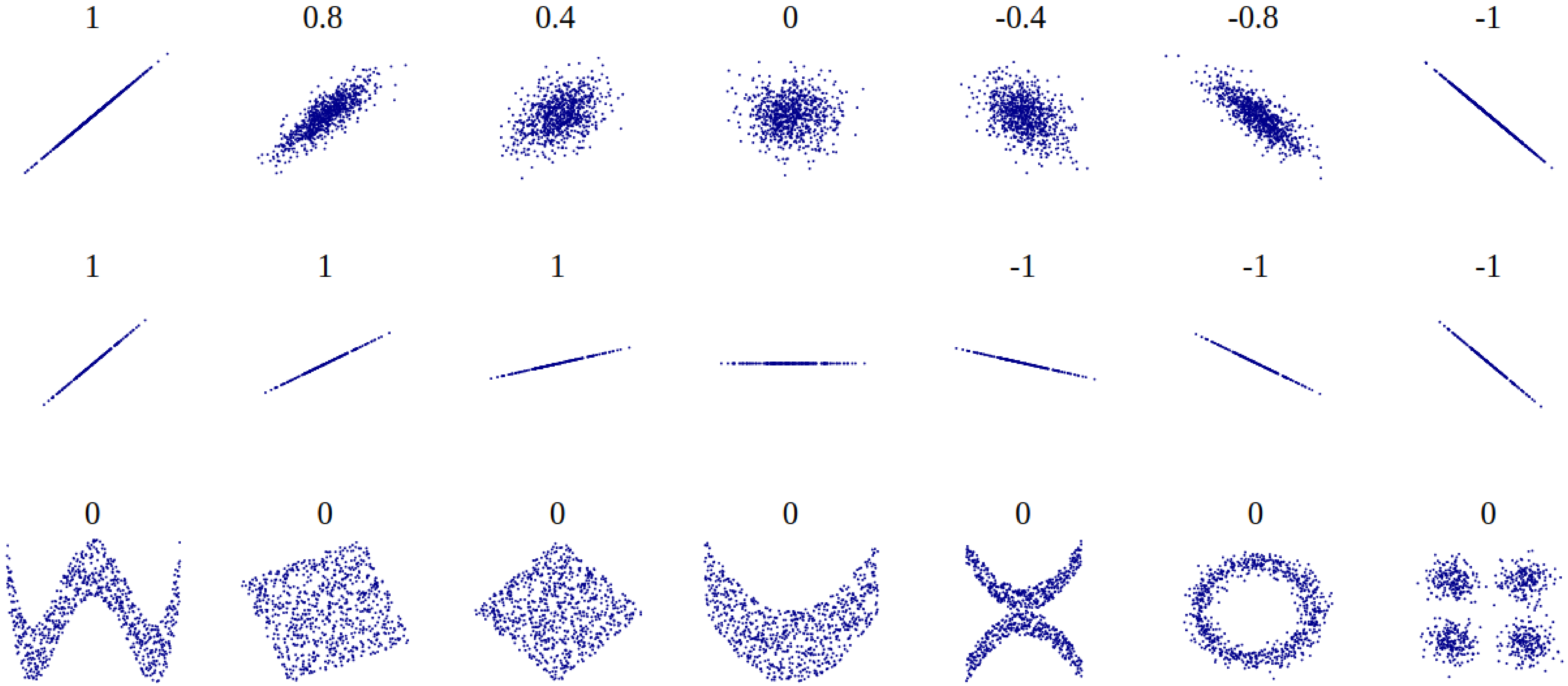
Le **coefficient de corrélation** du couple (X, Y) est

$$-1 \leq \mathrm{Cor}(X, Y) := \frac{\mathrm{Cov}(X, Y)}{\sigma_X\, \sigma_Y} \leq 1$$

Interprétation géométrique:

- $\mathrm{Cov}(X, Y) \simeq$ produit scalaire; $\sigma_X, \sigma_Y \simeq$ normes
- donc $\mathrm{Cor}(X, Y) \simeq \cos\theta$!

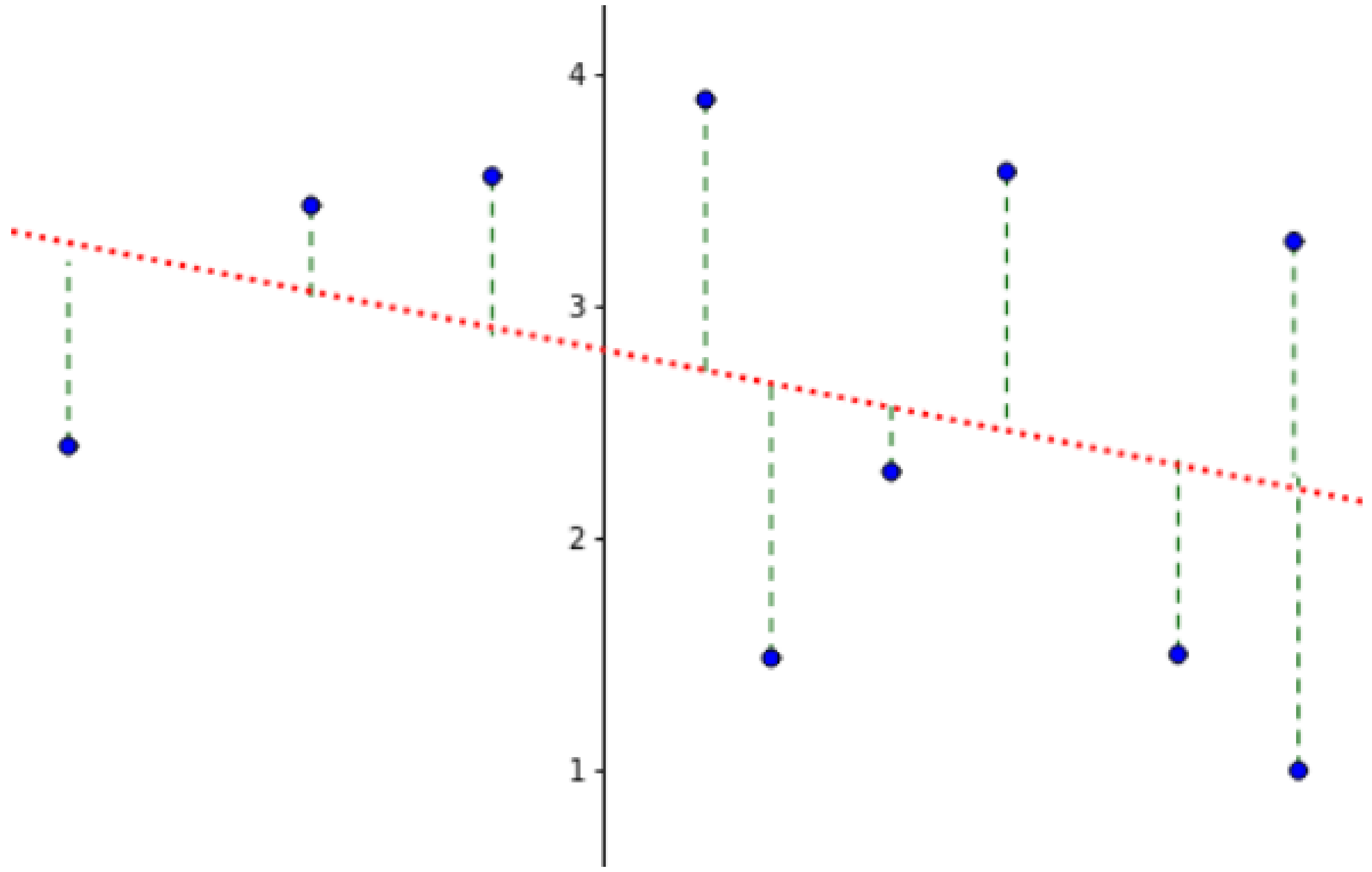
Graphiquement



Lien avec la régression linéaire

Étant données X et Y, on cherche à écrire

$$Y \approx aX + b.$$



On choisit habituellement les coefficients qui minimisent

$$\Delta(a,b) := \mathbb{E}[(aX + b - Y)^2].$$

Coefficients de la droite de régression linéaire $Y \approx aX + b$:

$$\begin{cases} a = \dfrac{\mathbb{E}[XY] - \mathbb{E}[X]\,\mathbb{E}[Y]}{\mathbb{E}[X^2] - \mathbb{E}[X]^2} \\[2em] b = \dfrac{-\mathbb{E}[X]\,\mathbb{E}[XY] + \mathbb{E}[X^2]\,\mathbb{E}[Y]}{\mathbb{E}[X^2] - \mathbb{E}[X]^2} \end{cases}$$

Retenir:

$$\boxed{a = \frac{\mathrm{Cov}(X,Y)}{\mathrm{Var(X)}} = \frac{\sigma_X\,\sigma_Y\,\mathrm{Cor}(X,Y)}{\sigma_X^2} = \frac{\sigma_Y}{\sigma_X}\mathrm{Cor}(X,Y)}$$

Appliquette

Exemple

À quelle(s) condition(s) sur a et b le couple (X, Y) est-il à covariance nulle ? Indépendant ?

(X,Y)	-1	0	1	Σ
-1	a	2b	a	0,5
1	b	2a	b	0,5
Σ	0,25	0,5	0,25	1