Probabilités et statistiques

II – Indépendance et corrélation

G. Chênevert

20 novembre 2023

JUNIA ISEN

Au menu aujourd'hui

Variables aléatoires (suite et fin)

Vecteurs aléatoires

Statistiques conjointes

Résumé de l'épisode précédent

- **Variable aléatoire** X : nombre qui 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_n) avec

$$\sum_n \underbrace{\mathbb{P}[X = x_n]}_{p_n} = 1.$$

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

$$\int_{-\infty}^{+\infty} f_X(x)\,\mathrm{d}x = 1.$$

- 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

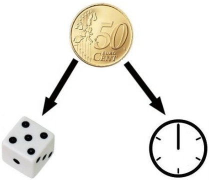
- Cas discret : F_X est continue par morceaux,

$$f_X(x) = \sum_n p_n \, \delta(x - x_n),$$

  et 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

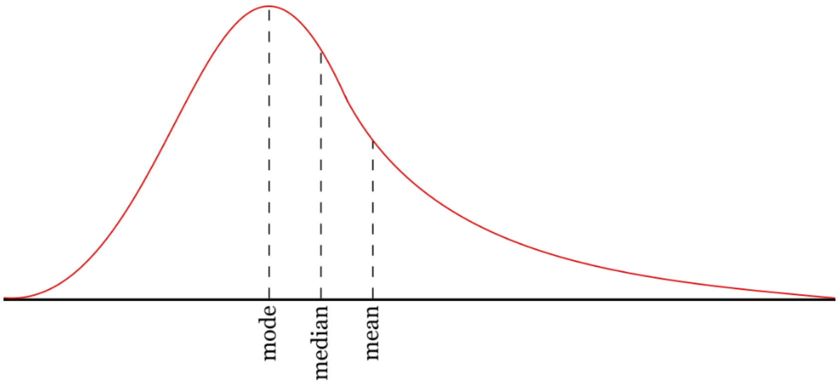
- L'**espérance**

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

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

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

Trois notions disctinctes



Mesure de dispersion

Pour quantifier la dispersion d'une v.a. X,

considérons *l'espérance de la déviation par rapport à son espérance* :

$$\mathbb{E}[X - \mathbb{E}[X]] = \mathbb{E}[X] - \mathbb{E}[\mathbb{E}[X]] = \mathbb{E}[X] - \mathbb{E}[X] = 0$$

Oups ! En fait c'est une caractérisation de l'espérance : le nombre μ pour lequel

$$\mathbb{E}[X - \mu] = 0.$$

Meilleure idée

Définition

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

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

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

Proposition

$$\mathrm{Var}(X) = \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2 = \mu_2 - \mu^2$$

Exemples

- $X \sim \mathcal{B}(n, p) \implies \mathrm{Var}(X) = np(1-p)$
- $X \sim \mathcal{G}(p) \implies \mathrm{Var}(X) = \frac{1-p}{p^2}$
- $X \sim \mathcal{U}([a, b]) \implies \mathrm{Var}(X) = \frac{(b-a)^2}{12}$
- $X \sim \mathcal{N}(\mu, \sigma) \implies \mathrm{Var}(X) = \sigma^2$
- $\vdots$

Écart à l'espérance

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

Théorème (Bienaymé-Tchebychev)

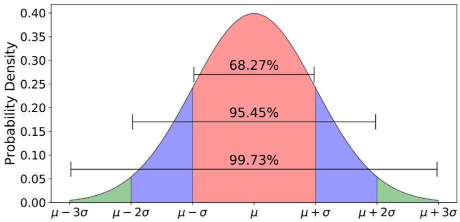
Pour toute variable aléatoire X (d'espérance et variance finies),

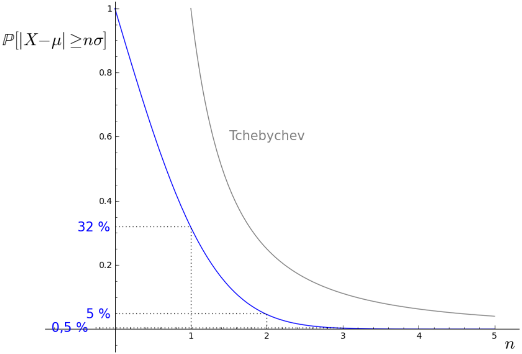
$$\mathbb{P}\big[\big|X-\mu\big| \geq n\sigma\big] \;\leq\; \frac{1}{n^2}.$$

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

$$\begin{aligned} 1 = \mathbb{E}\left[X^2\right] &= \int_{x\in\mathcal{A}} x^2\, f_X(x)\,\mathrm{d}x + \int_{x\notin\mathcal{A}} x^2\, f_X(x)\,\mathrm{d}x \\ &\geq \int_{x\in\mathcal{A}} x^2 f_X(x)\,\mathrm{d}x \geq n^2\,\mathbb{P}[\mathcal{A}]. \end{aligned}$$

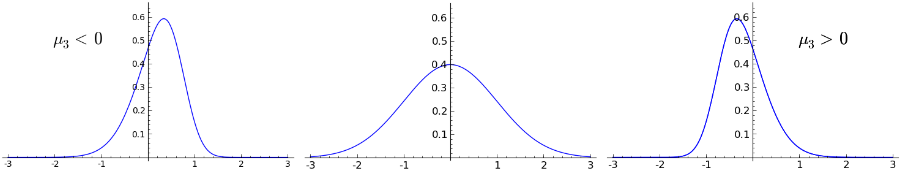
La loi normale fait bien mieux





Statistiques d'ordre supérieur

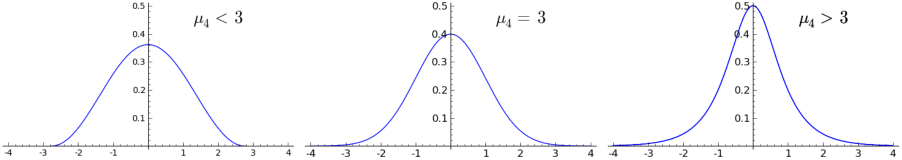
$\mu_3 = \mathbb{E}\left[X^3\right]$ coefficient de **dissymétrie** (*skewness*)



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

Statistiques d'ordre supérieur

$\mu_4 = \mathbb{E}\left[X^4\right]$ coefficient d'**aplatissement** (*kurtosis*)



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

Rappel : ≪ Décrire la loi de X ≫

- Donner F_X
- Donner f_X (ou les p_n dans le cas discret)
- Donner la suite des moments $\mu_n = \mathbb{E}[X^n]$, $n \in \mathbb{N}$
- Ou encore, la fonction génératrice

$$g_X(t) = \mathbb{E}\left[e^{tX}\right] = 1 + \mu\, t + \mu_2 \frac{t^2}{2} + \mu_3 \frac{t^3}{6} + \cdots = \sum_{n=0}^{\infty} \mu_n \frac{t^n}{n!}$$

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

Au menu aujourd'hui

Variables aléatoires (suite et fin)

Vecteurs aléatoires

Statistiques conjointes

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

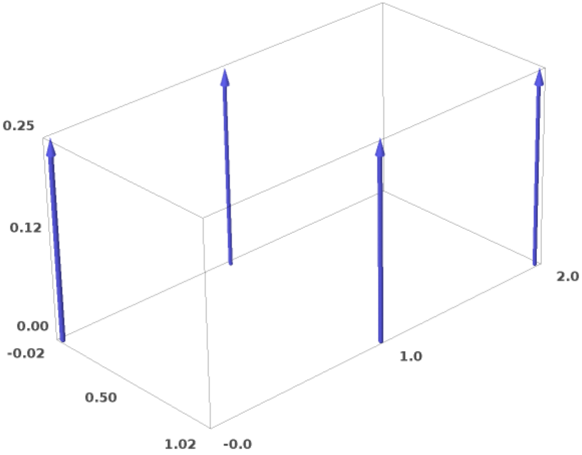
- **Probabilités conditionnelles**, *e.g.*

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

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

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

Représentation graphique



Loi conjointe

Définition

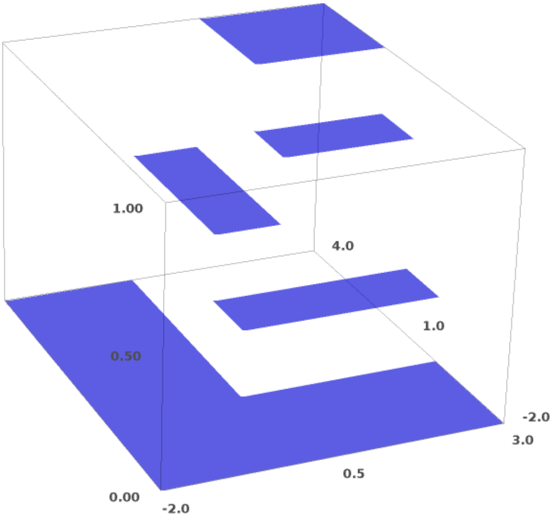
Fonction de répartition

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

Densité de probabilité

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

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 :

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

$$\implies f_X(x) = \int_{-\infty}^{+\infty} f(x, y)\,\mathrm{d}y \qquad \text{et de même pour } f_Y(y)$$

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 :

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

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

Proposition

$$X \text{ et } Y \text{ sont indépendantes } \overset{\text{déf}}{\iff} \mathbb{P}[X \in \mathcal{A} \text{ et } Y \in \mathcal{B}] = \mathbb{P}[X \in \mathcal{A}] \cdot \mathbb{P}[Y \in \mathcal{B}]$$

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

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

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

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

En d'autres termes :

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

Au menu aujourd'hui

Variables aléatoires (suite et fin)

Vecteurs aléatoires

Statistiques conjointes

Variance conjointe

Définition

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

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

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

Indépendance et covariance

Proposition

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

D'où :

Corollaire

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

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

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

Ou encore :

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

Corrélation

On préfère souvent une version normalisée de la covariance :

Définition

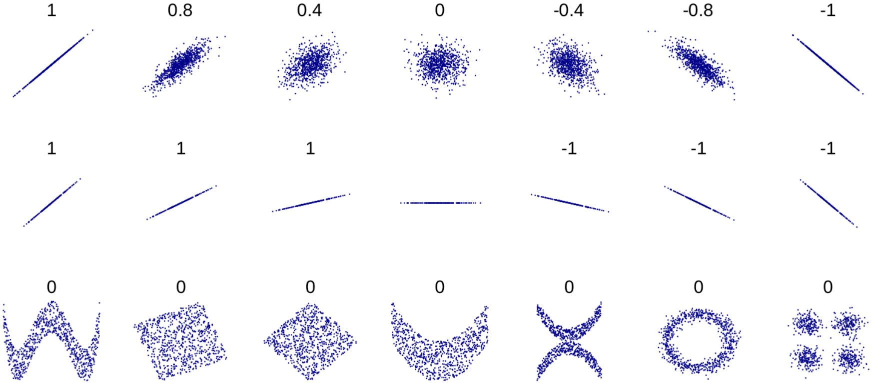
Le **coefficient de corrélation** (linéaire) du couple (X, Y) est

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

Interprétation géométrique :

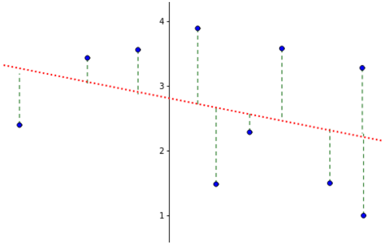
- $\mathrm{Cov}(X, Y) \simeq$ produit scalaire ; $\sigma_X, \sigma_Y \simeq$ normes
- donc $\mathrm{Cor}(X, Y) \simeq \cos\theta$!

Graphiquement



Régression linéaire

Étant données X et Y, on cherche à écrire



On choisit habituellement les coefficients qui minimisent

$$\Delta(a,b) := \mathbb{E}\big[(aX+b-Y)^2\big]$$ **droite des moindres carrés**

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

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

Retenir :

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