13.08.2025

lundi, 13 octobre 2025 07:08

Rappel : Formule de Bayes et Proba totale

Bayes :

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(B \mid A) \cdot \mathbb{P}(A)}{\mathbb{P}(B)}$$

Proba totale : Si $A_1, A_2, \ldots$ disjoints, et $\bigcup_{i \geq 1} A_i = \Omega$

$$\mathbb{P}(B) = \sum_{i \geq 1} \mathbb{P}(B \mid A_i)\, \mathbb{P}(A_i)$$

↳

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(B \mid A)\, \mathbb{P}(A)}{\mathbb{P}(B)}$$

$\mathbb{P}(B) = \mathbb{P}(B \mid A)\, \mathbb{P}(A) + \mathbb{P}(B \mid A^c) \cdot \mathbb{P}(A^c)$

Exemple : Oubli de la fréquence de base (Expl. 2.4.6)

Situation : Maladie dans une population

p = proportion de gens infectés

On a un teste pour vérifier si qqn est infecté ou non.

Teste rate de temps en temps :

- Tst dit "infecté" alors que l'individu est sain : ⇝ proba $\frac{5}{1000}$

- Test dit "sain" alors que l'individu est infecté: $\rightsquigarrow$ proba $\frac{1}{1000}$

Modélisation: $\Omega = \{I_+, I_-\} \times \{T_+, T_-\}$

(I_+: ind. infecté, I_-: ind. sain, T_+: teste pos., T_-: test nég.)

Q1: $\mathbb{P}(\text{test rate}) = ?$

On connait: * $\mathbb{P}(\text{ind infecté}) = \mathbb{P}(\{(I_+, T_-), (I_+, T_+)\}) = \rho$

* $\mathbb{P}(\text{ind sain}) = 1-\rho$

* $\mathbb{P}(\text{test rate} \mid \text{Ind sain}) = \frac{9}{1000}$

* $\mathbb{P}(\text{test rate} \mid \text{Ind inf.}) = \frac{1}{1000}$

$$\mathbb{P}(\text{test rate}) = \underbrace{\mathbb{P}(\text{test rate} \mid \text{Ind sain})}_{=9/1000}\underbrace{\mathbb{P}(\text{ind sain})}_{=1-\rho} + \underbrace{\mathbb{P}(\text{test rate} \mid \text{Ind inf})}_{=1/1000}\underbrace{\mathbb{P}(\text{ind. inf.})}_{=\rho}$$

(proba tot.)

$$= \frac{9}{1000}(1-\rho) + \frac{1}{1000}\rho \underset{\substack{\rho \le 1 \\ 1-\rho \le 1}}{\le} \frac{10}{1000} = \frac{1}{100}$$

↳ test est faible à $99/100$ indépendamment

↪ reste est fiable à 999/1000 indépendamment de la valeur de ρ.

Q2 : Si on test un individu positif, quelle est la proba qu'il soit effectivement infecté ?

$$\mathbb{P}\left(\underbrace{\{(I_+,T_-),(I_+,T_+)\}}_{\text{infecté}} \mid \overbrace{\{(I_-,T_+),(I_+,T_+)\}}^{\text{test positif}}\right) = ?$$

$$\mathbb{P}(\text{infecté} \mid \text{test pos.}) = \frac{\mathbb{P}(\text{test pos} \mid \text{inf.})\,\mathbb{P}(\text{inf}) \quad (=\rho)}{\mathbb{P}(\text{test pos} \mid \text{inf})\,\mathbb{P}(\text{inf}) + \mathbb{P}(\text{test pos} \mid \text{sain})\,\mathbb{P}(\text{sain})}$$

$\mathbb{P}(\text{test pos} \mid \text{inf}) = 1 - \mathbb{P}(\text{test nég} \mid \text{inf}) = 1 - \frac{1}{1000} = \frac{999}{1000}$

$\mathbb{P}(\text{inf}) = \rho$

$\mathbb{P}(\text{test pos} \mid \text{sain}) = \frac{9}{1000}$

$\mathbb{P}(\text{sain}) = 1-\rho$

$$= \frac{\frac{999}{1000}\cdot\rho}{\frac{999}{1000}\cdot\rho + \frac{9}{1000}(1-\rho)} = \frac{999\rho}{990\rho + 9}$$

⚠ cette proba est petite si ρ petit.

Exemple 2.4.7 (Paradoxe de Simpson)

	All		Women		Men	
	Applicants	Admitted	Applicants	Admitted	Applicants	Admitted
Total	12763	41%	4321	35%	8442	44%

Admission à Berkley.

	All		Women		Men	
Dep.	Applicants	Admitted	Applicants	Admitted	Applicants	Admitted
A	933	64%	108	82%	825	62%
B	585	63%	25	68%	560	63%
C	918	35%	593	34%	325	37%
D	792	34%	375	35%	417	33%
E	584	25%	393	24%	191	28%
F	714	6%	341	7%	373	6%
Total	4526	39%	1835	30%	2691	45%

	All		Women		Men	
Dep.	Applicants	Admitted	Applicants	Admitted	Applicants	Admitted
A	933	64%	108	82%	825	62%
C	918	35%	593	34%	325	37%
Total	1851	39%	701	38%	1150	62%

55

$$\mathbb{P}(\text{Admis} \mid H) = \mathbb{P}(\text{Admis} \mid H, \text{dep A})\,\mathbb{P}(\text{dep A} \mid H) + \mathbb{P}(\text{Admis} \mid H, \text{dep C})\,\mathbb{P}(\text{dep C} \mid H)$$

$$= \frac{62}{100} \cdot \frac{825}{1150} + \frac{37}{100} \cdot \frac{325}{1150}$$

$$\mathbb{P}(\text{Admis} \mid F) = \frac{82}{100} \cdot \frac{108}{701} + \frac{34}{100} \cdot \frac{593}{701}$$

Le "paradox" vient des choix des candidats: les hommes postulent plus au département qui accepte plus.

Correlation (section 2.5)

Variance: Soit X un variable aléatoire est

$$\text{Var}(X) := \mathbb{E}\left[(X - \mathbb{E}(X))^2\right] = \mathbb{E}[X^2] - \mathbb{E}[X]^2$$

↳ mesure à quel point X est typiquement

↳ mesure à quel point X est typiquement loin de $E[X]$.

Écart-type : (deviation standard)

$$\sigma_X = \sqrt{Var(X)} \rightsquigarrow \text{"typiquement"}, \ X \in [E(X) - \sigma_X, E(X) + \sigma_X].$$

Remarque : On peut définir alternativement

$$Var(X) = \inf_{\mu \in \mathbb{R}} E[(X-\mu)^2]$$

$$E[X] = \arg \min$$

"Prv" : * $E[(X-\mu)^2] = E[X^2] + \mu^2 - 2\mu E[X]$

* On cherche les minima en μ :

$$2\mu - 2E[X] = 0 \iff \mu = E[X].$$

Covariance : Soient X, Y deux variables aléatoires,

$$Cov(X,Y) := E[XY] - E[X]E[Y].$$

Remarque 1 : Si X et Y sont indépendantes,

$$Cov(X,Y) = E[XY] - E[X]E[Y]$$
$$\overset{\text{indép}}{=} E[X]E[Y] - E[X]E[Y] = 0$$

$\overset{\text{indép}}{=} E[X]E[Y] - E[X]E[Y] = 0.$

Beaucoup de résultats sont "simples" à montrer sous l'hyp. "X et Y sont indép.", l'hyp "$Cov(X,Y) = 0$" est souvent une manière naturelle d'affaiblir l'hyp. d'indép.

Remarque 2: La covariance apparaît naturellement quand on calcule la variance d'une somme:

$$\begin{aligned} Var(X+Y) &= E[(X+Y)^2] - E[X+Y]^2 \\ &= E[X^2] + E[Y^2] + 2E[XY] \\ &\quad - E[X]^2 - E[Y]^2 - 2E[X]E[Y] \\ &= Var(X) + Var(Y) + 2\,Cov(X,Y) \end{aligned}$$

↳ $Cov(X,Y)$ est une correction à "l'additivité" pour la variance.

Propriétés: (lemme 2.5.2 + déf)

1) $Cov(X,X) = Var(X)$
2) $Cov(X,Y) = Cov(Y,X)$
3) $Cov(aX, bY) = a \cdot b \cdot Cov(X,Y) \quad \forall\, a,b \in \mathbb{R}$
4) $Cov(X, Y_1 + Y_2) = Cov(X,Y_1) + Cov(X,Y_2)$

↳ $(X,Y) \mapsto Cov(X,Y)$ est linéaire en chaque

↳ $(X,Y) \mapsto Cov(X,Y)$ est linéaire en chaque coordonnée.

Remarque: Indép $\Rightarrow$ Cov $= 0$

Mais Cov $= 0 \not\Rightarrow$ indép.

Exp!: Soient X, Z deux v.a. indép. telles que

- $\mathbb{P}(X=10) = \mathbb{P}(X=0) = 1/2$
- $\mathbb{P}(Z=-1) = \mathbb{P}(Z=+1) = 1/2$

On définit $Y = X \cdot Z$.

On a
$$\begin{aligned} Cov(X,Y) &= E[XY] - E[X]E[Y] \\ &\overset{\text{déf } Y}{=} E[X^2 Z] - E[X]E[X \cdot Z] \\ &\overset{X,Z \text{ indép}}{=} E[X^2]\underbrace{E[Z]}_{=0} - E[X]^2\underbrace{E[Z]}_{=0} = 0 \end{aligned}$$

Mais X et Y ne sont pas indépendantes:

$$\begin{aligned} \mathbb{P}(X=0, Y=0) &= \mathbb{P}(X=0, XZ=0) \\ &\overset{Z \neq 0 \text{ p.s.}}{=} \mathbb{P}(X=0) = 1/2 \end{aligned}$$

$$\begin{aligned} \mathbb{P}(X=0)\,\mathbb{P}(Y=0) &= \mathbb{P}(X=0)\,\mathbb{P}(XZ=0) \\ &\overset{Z \neq 0 \text{ p.s.}}{=} \mathbb{P}(X=0)^2 = 1/4 \neq 1/2 \end{aligned}$$

On a un cas particulier:

Lemme (2.5.1) [illegible]

Lemme (2.5.1) Si X et Y sont des variables de Bernoulli ($\mathbb{P}(X \in \{0,1\}) = 1$, $\mathbb{P}(Y \in \{0,1\}) = 1$)

Alors X et Y sont indép si et seulement si $\mathrm{Cov}(X,Y) = 0$.

Indication : $\forall c \in \mathbb{R}$, $\mathrm{Cov}(X, c) = 0$.

Coefficient de corrélation : Soient X, Y deux v.a.

$$\rho_{XY} = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\,\mathrm{Var}(Y)}} = \frac{\mathrm{Cov}(X,Y)}{\sigma_X \sigma_Y} \in [-1,1].$$

↳ plus fort

ρ_{XY} est une manière indép. des unités de mesurer la corrélation :

Si deux quantités sont exprimés en X_1, Y_1 m. ou en X_2, Y_2 m. m.

↳ $\mathrm{Cov}(X_1, Y_1) = \mathrm{Cov}\left(\frac{X_2}{1000}, \frac{Y_2}{1000}\right)$

$$= \frac{1}{1000000}\,\mathrm{Cov}(X_2, Y_2)$$

mais $\rho_{X_1 Y_1} = \rho_{X_2 Y_2}$ ▽