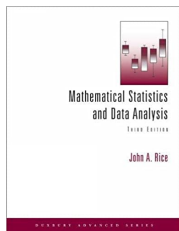


Introduction aux statistiques pour l'Ingénieur

Stéphane Canu

asi.insa-rouen.fr/enseignants/~scanu

scanu@insa-rouen.fr



ITI 3, INSA Rouen Normandie, Février 2023

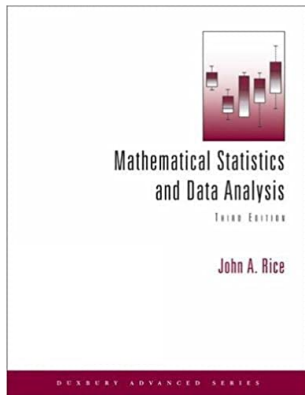
Lecture road map

1 Description d'un couple de variables

- Cas de deux variable quantitatives
 - Covariance et corrélation
 - Espérance conditionnelle et prédiction
 - La régression linéaire
- Tableau de contingence
- Quantitatif vs. Qualitatif

2 Description d'un ensemble de variables

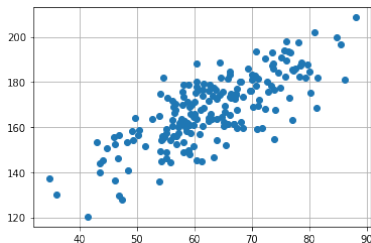
- Cas des variable quantitatives
- Cas des variable qualitatives



Deux variables

$(x_i, y_i), i = 1, \dots, n$

	<i>taille</i>	<i>poids</i>
<i>Alice</i>	x_1	y_1
\vdots	\vdots	\vdots
<i>Bill</i>	x_i	y_i
\vdots	\vdots	\vdots
<i>Sue</i>	x_n	y_n



Mesure de dépendance : Covariance et corrélation

Variance

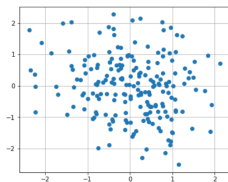
$$V(X) = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$$

Covariance

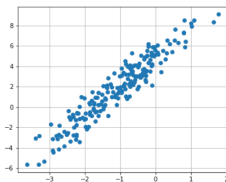
$$\text{cov} = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$$

Corrélation

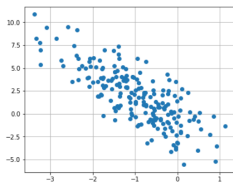
$$\rho = \text{cor}(X, Y) = \frac{\text{cov}(X, Y)}{\sqrt{V(X) V(Y)}}$$



$$\rho = 0$$



$$\rho = 0.9$$



$$\rho = -0.75$$

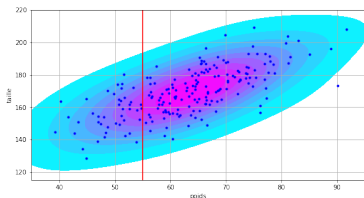
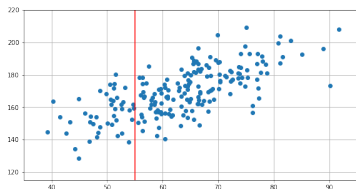
Les OPM associées

$$V(X) = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 \quad \xrightarrow{n \rightarrow \infty} \quad \mathbb{E}_X((X - \mu_x)^2) = \sigma_x^2$$

$$\text{cov} = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) \quad \xrightarrow{n \rightarrow \infty} \quad \mathbb{E}_{X,Y}((X - \mu_x)(Y - \mu_y))$$

$$\rho = \text{cor}(X, Y) = \frac{\text{cov}(X, Y)}{\sqrt{V(X) V(Y)}} \quad \xrightarrow{n \rightarrow \infty} \quad \frac{\mathbb{E}_{X,Y}((X - \mu_x)(Y - \mu_y))}{\sqrt{\sigma_x^2 \sigma_y^2}}$$

Distribution Gaussienne en dimension 2



Soit Z une variable aléatoire Gaussienne bidimensionnelle

$$Z = \begin{pmatrix} X \\ Y \end{pmatrix} \sim \mathcal{N}(\mu, \Sigma)$$

avec la matrice de variance covariance Σ

$$\mathbb{E}(Z) = \mu = \begin{pmatrix} \mu_x = \mathbb{E}(X) \\ \mu_y = \mathbb{E}(Y) \end{pmatrix} \quad \mathbb{E}((z - \mu)(z - \mu)^\top) = \Sigma$$

$$\Sigma = \begin{pmatrix} \sigma_x^2 & cov \\ cov & \sigma_y^2 \end{pmatrix}, \quad \Sigma^{-1} = \frac{1}{\sigma_x^2 \sigma_y^2 - cov^2} \begin{pmatrix} \sigma_y^2 & -cov \\ -cov & \sigma_x^2 \end{pmatrix}$$

Distribution Gaussienne en dimension 2

Soit Z une variable aléatoire Gaussienne bidimensionnelle

$$Z = \begin{pmatrix} X \\ Y \end{pmatrix} \sim \mathcal{N}(\mu, \Sigma)$$

avec la matrice de variance covariance Σ

$$\mu = \begin{pmatrix} \mu_x = \mathbb{E}(X) \\ \mu_y = \mathbb{E}(Y) \end{pmatrix} \quad \Sigma^{-1} = \frac{1}{\sigma_x^2 \sigma_y^2 - \text{cov}^2} \begin{pmatrix} \sigma_y^2 & -\text{cov} \\ -\text{cov} & \sigma_x^2 \end{pmatrix}$$

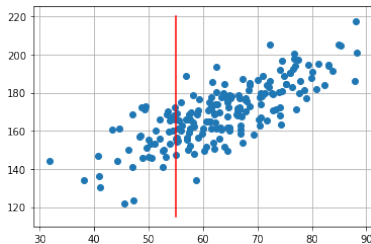
Sa densité s'écrit ($z \in \mathbb{R}^2$)

$$f(z) = \frac{1}{2\pi \det(\Sigma)^{1/2}} \exp \left\{ -\frac{1}{2} (z - \mu)^\top \Sigma^{-1} (z - \mu) \right\}$$

Probabilité et Densité conditionnelle

	Bob	Alice	John	
TV	12	21	12	45
PC	21	12	14	47
VC	22	31	14	67
LV	4	17	5	26
	59	81	45	185

$$\mathbb{P}_{Y|X}(y|x) = \frac{\mathbb{P}_{X,Y}(x,y)}{\mathbb{P}_X(x)}$$



$$f_{Y|X}(y|x) = \frac{f_{X,Y}(x,y)}{f_X(x)}$$

Exemple : densité conditionnelle de gaussiennes

$$f_{Y|X}(y|x) = \frac{f_{X,Y}(z = (x, y))}{f_X(x)} = \frac{\frac{1}{2\pi \det(\Sigma)^{1/2}} \exp\left\{-\frac{1}{2}(z - \mu)^\top \Sigma^{-1}(z - \mu)\right\}}{\frac{1}{(2\pi \sigma_x^2)^{1/2}} \exp\left\{-\frac{1}{2\sigma_x^2}(x - \mu_x)^2\right\}}$$

$$f_{Y|X}(y|x) = \frac{1}{\sigma_y \sqrt{2\pi(1 - \rho^2)}} \exp\left\{-\frac{1}{2\sigma_y^2(1 - \rho^2)} \left(y - \mu_y - \rho \frac{\sigma_y}{\sigma_x}(x - \mu_x)\right)^2\right\}$$

et la loi conditionnelle est aussi normale

$$Y|X \sim \mathcal{N}\left(\underbrace{\mu_y + \rho \frac{\sigma_y}{\sigma_x}(x - \mu_x)}_{\mu_{Y|X}}, \underbrace{\sigma_y^2(1 - \rho^2)}_{\sigma_{Y|X}^2}\right)$$

L'espérance conditionnelle gaussienne

$$\mathbb{E}(Y|X) = \mu_{Y|X} = \mu_y + \rho \frac{\sigma_y}{\sigma_x}(x - \mu_x)$$

Le problème de prédiction

Fixons x et essayons de prédire y .

Soit $h^*(x)$ la meilleure prédiction pour y , (au sens des moindres carrés)

$$\begin{aligned}h^*(x) &= \arg \min_{h(x)} \mathbb{E}_{Y|X} (Y - h(x))^2 \\ &= \arg \min_{h(x)} J(h(x)) \quad \text{avec } J(h(x)) = \int_y (Y - h(x))^2 f_{Y|X}(y|x) dy\end{aligned}$$

or

$$\frac{dJ(h(x))}{dh(x)} = -2 \int_y (Y - h(x)) f_{Y|X}(y|x) dy$$

et

$$\frac{dJ(h(x))}{dh(x)} = 0 \quad \Leftrightarrow \quad h^*(x) = \int_y Y f_{Y|X}(y|x) dy = \mathbb{E}(Y|X).$$

Dans le cas gaussien, la meilleure prédiction est donnée par :

$$\mathbb{E}(Y|X) = \mu_y + \rho \frac{\sigma_y}{\sigma_x} (x - \mu_x)$$

Le problème des moindres carrés

Les données dont nous disposons peuvent aussi être vue comme un système de n équations à $2 + n$ inconnues (a , b et ε). Ce système s'écrit de la manière suivante :

$$\left\{ \begin{array}{l} ax_1 + b + \varepsilon_1 = y_1 \\ \dots \quad \dots \\ ax_i + b + \varepsilon_i = y_i \\ \dots \quad \dots \\ ax_n + b + \varepsilon_n = y_n \end{array} \right.$$

on recherche a et b qui minimisent simultanément tous les $\varepsilon_i, i = 1, n$

$$J(a, b) = \sum_{i=1}^n \underbrace{(y_i - ax_i - b)}_{\varepsilon_i}^2$$

Calcul du gradient

$$\min_{a,b} J(a,b) \quad \text{avec} \quad J(a,b) = \frac{1}{2} \sum_{i=1}^n (ax_i + b - y_i)^2$$

$$\left\{ \begin{array}{l} \frac{\partial J(a,b)}{\partial a} = \sum_{i=1}^n (ax_i + b - y_i) x_i = \sum_{i=1}^n (ax_i^2 + bx_i - y_i x_i) \\ \phantom{\frac{\partial J(a,b)}{\partial a}} = \sum_{i=1}^n ax_i^2 + \sum_{i=1}^n bx_i - \sum_{i=1}^n y_i x_i \\ \phantom{\frac{\partial J(a,b)}{\partial a}} = a \sum_{i=1}^n x_i^2 + b \sum_{i=1}^n x_i - \sum_{i=1}^n y_i x_i \\ \frac{\partial J(a,b)}{\partial b} = \sum_{i=1}^n (ax_i + b - y_i) = \sum_{i=1}^n ax_i + \sum_{i=1}^n b - \sum_{i=1}^n y_i \\ \phantom{\frac{\partial J(a,b)}{\partial b}} = a \sum_{i=1}^n x_i + bn - \sum_{i=1}^n y_i \end{array} \right.$$

Conditions d'optimalité

$$(\hat{a}, \hat{b}) \text{ est solution du problème } \min_{a,b} J(a,b) \Leftrightarrow \begin{cases} \frac{\partial J(\hat{a}, \hat{b})}{\partial a} = 0 \\ \frac{\partial J(\hat{a}, \hat{b})}{\partial b} = 0 \end{cases}$$

$$\begin{cases} \frac{\partial J(\hat{a}, \hat{b})}{\partial a} = 0 \\ \frac{\partial J(\hat{a}, \hat{b})}{\partial b} = 0 \end{cases} \Leftrightarrow \begin{aligned} \hat{a} \sum_{i=1}^n x_i^2 + \hat{b} \sum_{i=1}^n x_i &= \sum_{i=1}^n y_i x_i && \times n \quad (1) \\ \hat{a} \sum_{i=1}^n x_i + \hat{b} n &= \sum_{i=1}^n y_i && \times \sum_{i=1}^n x_i \quad (2) \end{aligned}$$

$$\hat{a} \left(n \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i \right)^2 \right) = n \sum_{i=1}^n y_i x_i - \sum_{i=1}^n y_i \sum_{i=1}^n x_i$$

Deux équations linéaires à deux inconnues

$$\hat{a} = \frac{n \sum_{i=1}^n y_i x_i - \sum_{i=1}^n y_i \sum_{i=1}^n x_i}{n \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i \right)^2} \qquad \hat{b} = \frac{\sum_{i=1}^n y_i - \hat{a} \sum_{i=1}^n x_i}{n}$$

Réécriture de la solution

si l'on note $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$, $\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$ et $V_x = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$ la variance des x_i

$$\begin{aligned}\hat{a} &= \frac{n \sum_{i=1}^n y_i x_i - \sum_{i=1}^n y_i \sum_{i=1}^n x_i}{n \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i\right)^2} \\ &= \frac{\frac{1}{n} \sum_{i=1}^n y_i x_i - \frac{1}{n} \sum_{i=1}^n y_i \frac{1}{n} \sum_{i=1}^n x_i}{\frac{1}{n} \sum_{i=1}^n x_i^2 - \left(\frac{1}{n} \sum_{i=1}^n x_i\right)^2} \\ &= \frac{\frac{1}{n} \sum_{i=1}^n y_i x_i - \bar{y} \bar{x}}{\frac{1}{n} \sum_{i=1}^n x_i^2 - \bar{x}^2} \\ &= \frac{\frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x})}{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2}\end{aligned}$$

$$\begin{aligned}\hat{b} &= \frac{\sum_{i=1}^n y_i - \hat{a} \sum_{i=1}^n x_i}{n} \\ &= \frac{1}{n} \sum_{i=1}^n y_i - \hat{a} \frac{1}{n} \sum_{i=1}^n x_i \\ &= \bar{y} - \hat{a} \bar{x}\end{aligned}$$

$$= \frac{\text{cov}(x, y)}{V_x}$$

Le théorème des moindres carrés

Théorème des moindres carrés

Soient $(x_i, y_i), i = 1, n$ un ensemble de couples d'observations.

La meilleure prédiction de y sachant x , au sens des moindres carrés est donnée par

$$y = h(x) = \mathbb{E}_{Y|X}(y|x) = \mu_y + \rho \frac{\sigma_y}{\sigma_x} (x - \mu_x) = ax + b$$

avec \hat{a} et \hat{b} estimés par :

$$\hat{a} = \frac{\sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\text{cov}(x, y)}{V_x} \quad \text{et} \quad \hat{b} = \bar{y} - \hat{a} \bar{x}$$

\hat{a} et \hat{b} que l'on appelle aussi les estimateurs au sens des moindres carrés.

Causalité \neq Corrélation (selon chatGPT)



Considérons deux variables, X (heures de sommeil) et Y (productivité au travail). Une corrélation positive entre X et Y suggérerait que lorsque X (heures de sommeil) augmente, Y (productivité) augmente également. Cependant, cela n'établit pas la causalité; ce n'est pas parce qu'une personne dort plus que nécessairement que sa productivité s'améliorera.

Considérons maintenant une troisième variable, Z (niveau de stress). Le niveau de stress peut avoir un impact à la fois sur le nombre d'heures de sommeil d'une personne et sur sa productivité au travail. Si une personne a des niveaux de stress élevés, elle peut dormir moins et être également moins productive. Dans ce cas, le niveau de stress (Z) serait une variable causale commune qui affecte à la fois X (heures de sommeil) et Y (productivité).

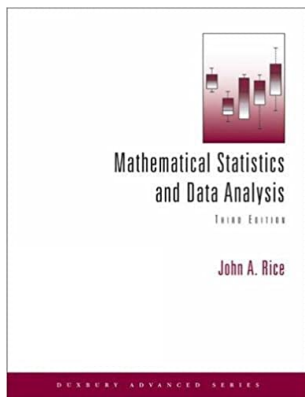
Lecture road map

1 Description d'un couple de variables

- Cas de deux variable quantitatives
 - Covariance et corrélation
 - Espérance conditionnelle et prédiction
 - La régression linéaire
- Tableau de contingence
- Quantitatif vs. Qualitatif

2 Description d'un ensemble de variables

- Cas des variable quantitatives
- Cas des variable qualitatives



Définition

Prenons un exemple : Si l'on considère l'échantillon constitué de l'ensemble des produits vendus par les vendeurs d'une boutique. On peut lui associer un tableau de contingence comme le montre le tableau ??.

(Bob, un pc)

(Percy, un aspirateur)

(Percy, une télé)

(Bob, une télé)

(Percy, une télé)

(Bob, un mac)

(John, un mac)

(Bob, un lave vaisselle)

⇒

	Bob	Percy	John
télévision	1	2	0
ordinateurs	2	0	1
aspirateurs	0	1	0
lave vaisselle	0	1	0

Table: Exemple de tableau de contingence.

Mesures d'indépendance (d'association)

Distance du chi2

$$\chi^2 = n \sum_{i=1}^c \sum_{j=1}^r \frac{n_{ij} - n_{i\bullet}n_{\bullet j}/n}{n_{i\bullet}n_{\bullet j}}$$

Le Φ^2 de Pearson (indépendant de n)

$$\Phi^2 = \frac{\chi^2}{n}$$

Coefficient de Tschuprow (indép. de r et c)

$$T = \sqrt{\frac{\Phi^2}{(r-1)(c-1)}}$$

Coefficient de Cramer

$$C = \sqrt{\frac{\Phi^2}{\min(r, c) - 1}}$$

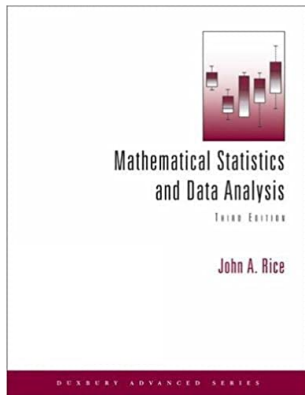
Lecture road map

1 Description d'un couple de variables

- Cas de deux variable quantitatives
 - Covariance et corrélation
 - Espérance conditionnelle et prédiction
 - La régression linéaire
- Tableau de contingence
- Quantitatif vs. Qualitatif

2 Description d'un ensemble de variables

- Cas des variable quantitatives
- Cas des variable qualitatives



Formules de décomposition de la variance

$$(x_i, y_i), i = 1, \dots, n$$

il y a r vendeurs

<i>vendeur</i>	<i>prix</i>
<i>Alice</i>	y_1
\vdots	\vdots
<i>Bill</i>	y_i
\vdots	\vdots
<i>Sue</i>	y_n

Décomposition de la moyenne, \bar{y}_m la moyenne des ventes du vendeur m

$$\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i = \frac{1}{n} \sum_{m=1}^r n_m \bar{y}_m$$

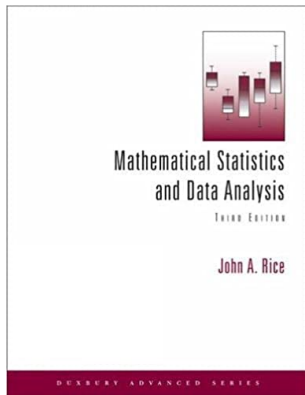
Lecture road map

1 Description d'un couple de variables

- Cas de deux variable quantitatives
 - Covariance et corrélation
 - Espérance conditionnelle et prédiction
 - La régression linéaire
- Tableau de contingence
- Quantitatif vs. Qualitatif

2 Description d'un ensemble de variables

- Cas des variable quantitatives
- Cas des variable qualitatives



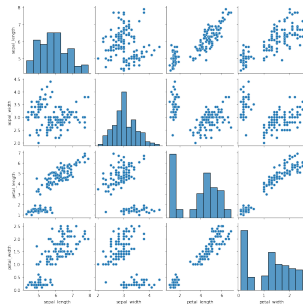


Tableau de Burt

Conclusion