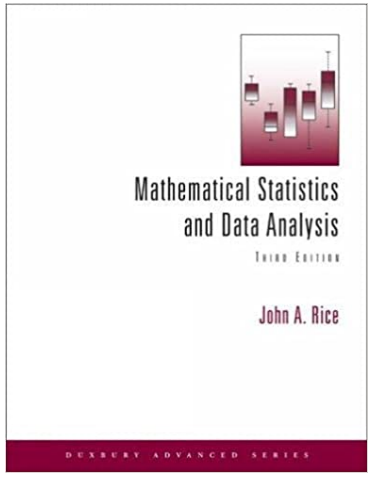
# Introduction aux statistiques pour l'Ingénieur

Stéphane Canu
asi.insa-rouen.fr/enseignants/~scanu
scanu@insa-rouen.fr




ITI 3, INSA Rouen Normandie, Février 2024

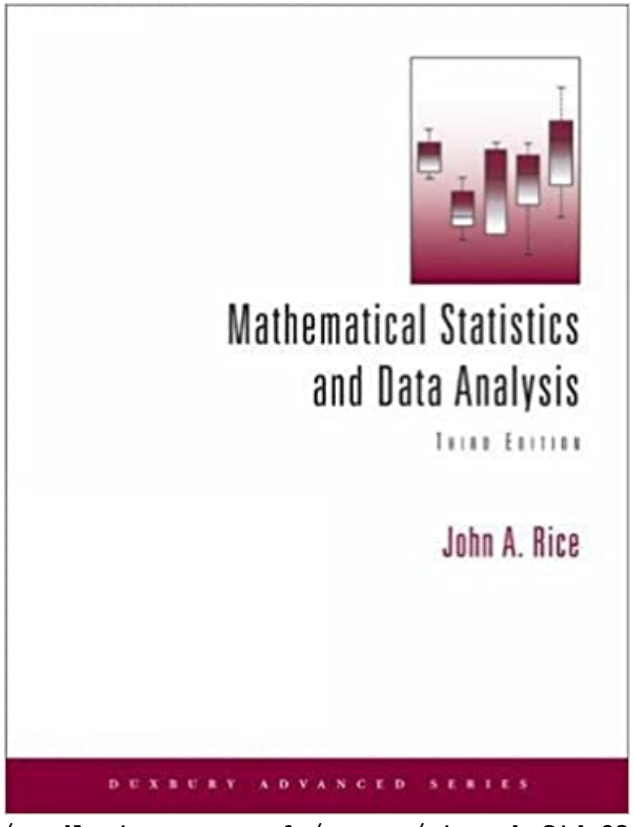
# Lecture road map

## 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

## Description d'un ensemble de variables

- Cas des variable quantitatives
- Cas des variables qualitatives



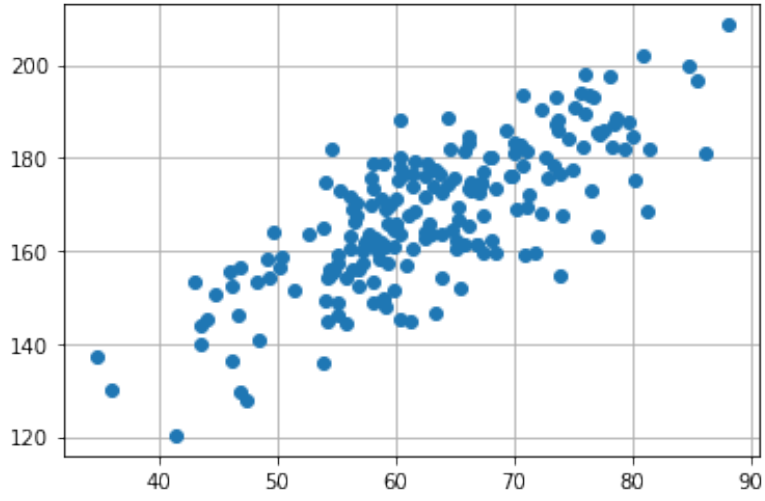

https://moodle.insa-rouen.fr/course/view.php?id=93

# Deux variables quantitatives

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

| | *taille* | *poids* |
|---|---|---|
| *Alice* | $x_1$ | $y_1$ |
| ⋮ | ⋮ | ⋮ |
| *Bill* | $x_i$ | $y_i$ |
| ⋮ | ⋮ | ⋮ |
| *Sue* | $x_n$ | $y_n$ |

Nuage de point (scatterplot)




```
import matplotlib.pyplot as plt

taille = [157.2, 162,5, 181.2, 180.9, 155.8]
poids = [64.0, 53.0, 73.0, 83.4, 53.4]

plt.plot(taille,poids,"ob") # ob = "o" ronds, "b" bleus
```

# Mesure de dépendance : Covariance et corrélation

Variance

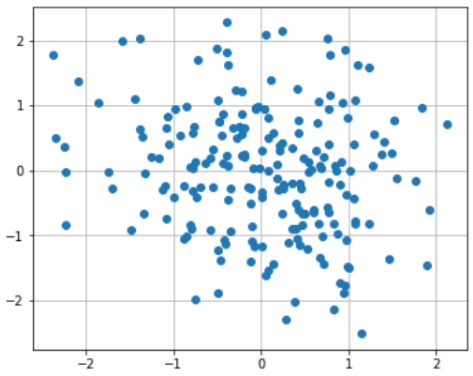
$$V(X) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Covariance

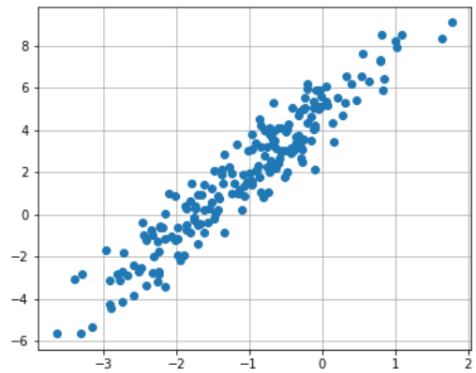
$$cov = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

Corrélation

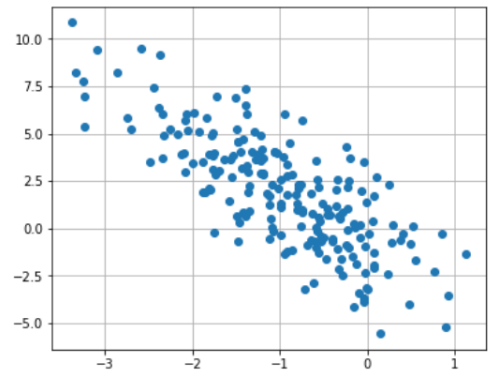
$$\rho = cor(X, Y) = \frac{cov(X, Y)}{\sqrt{V(X)\,V(Y)}}$$



$\rho = 0$



$\rho = 0.9$



$\rho = -0.75$

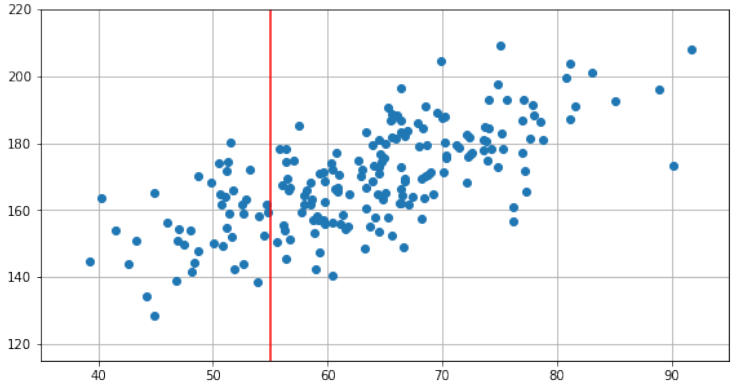
# Les OPM associées

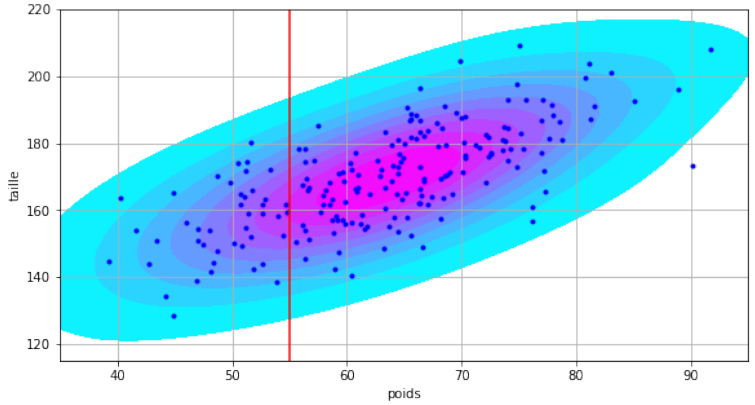
$$V(X) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \quad \underset{n\to\infty}{\longrightarrow} \quad \mathbb{E}_X\big((X-\mu_x)^2\big) = \sigma_x^2$$

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

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

# 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 $\Sigma$

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

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

# 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 $\Sigma$

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

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

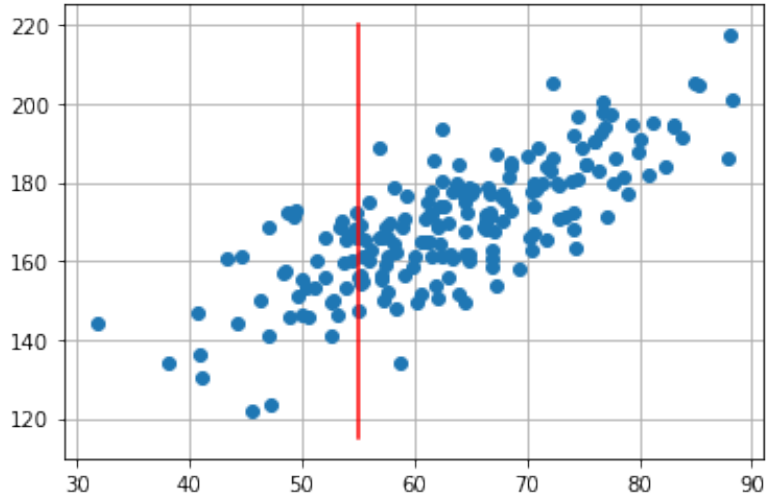
$$f(\mathbf{z}) = \frac{1}{2\pi \det(\Sigma)^{1/2}} \exp\left\{-\tfrac{1}{2}(\mathbf{z}-\mu)^\top \Sigma^{-1}(\mathbf{z}-\mu)\right\}$$

# Probabilité et Densité conditionnelle

Variables qualitatives

| | 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 |

Variables quantitatives




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

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

# Exemple : densité conditionnelle de gaussiennes

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

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

et la loi conditionnelle est aussi normale

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

## L'espérance conditionnelle gaussienne

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

# Le problème de prédiction

Fixons $x$ et essayons de prédire $y$.
Soit $h^\star(x)$ la meilleure prédiction pour $y$, (au sens des moindres carrés)

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

or

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

et

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

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 $\varepsilon$). Ce système s'écrit de la manière suivante :

$$\begin{cases} ax_1 + b + \varepsilon_1 & = y_1 \\ \cdots & \cdots \\ ax_i + b + \varepsilon_i & = y_i \\ \cdots & \cdots \\ ax_n + b + \varepsilon_n & = y_n \end{cases}$$

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

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

# Calcul du gradient

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

$$\begin{cases} \dfrac{\partial J(a,b)}{\partial a} = \displaystyle\sum_{i=1}^{n} (ax_i + b - y_i)\, x_i = \sum_{i=1}^{n} \left(ax_i^2 + bx_i - y_i x_i\right) \\ \qquad = \displaystyle\sum_{i=1}^{n} ax_i^2 + \sum_{i=1}^{n} bx_i - \sum_{i=1}^{n} y_i x_i \\ \qquad = a\displaystyle\sum_{i=1}^{n} x_i^2 + b\sum_{i=1}^{n} x_i - \sum_{i=1}^{n} y_i x_i \\ \dfrac{\partial J(a,b)}{\partial b} = \displaystyle\sum_{i=1}^{n} (ax_i + b - y_i) = \sum_{i=1}^{n} ax_i + \sum_{i=1}^{n} b - \sum_{i=1}^{n} y_i \\ \qquad = a\displaystyle\sum_{i=1}^{n} x_i + bn - \sum_{i=1}^{n} y_i \end{cases}$$

# Conditions d'optimalité

$(\widehat{a}, \widehat{b})$ est solution du problème $\min_{a,b} J(a,b) \quad \Leftrightarrow \quad \begin{cases} \dfrac{\partial J(\widehat{a},\widehat{b})}{\partial a} = 0 \\ \dfrac{\partial J(\widehat{a},\widehat{b})}{\partial b} = 0 \end{cases}$

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

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

## Deux équations linéaires à deux inconnues

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

# Réécriture de la solution

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

$$
\begin{aligned}
\widehat{a} &= \frac{n\sum_{i=1}^{n} y_i x_i - \sum_{i=1}^{n} y_i \sum_{i=1}^{n} x_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} \\
&= \frac{\frac{1}{n}\sum_{i=1}^{n} y_i x_i - \frac{1}{n}\sum_{i=1}^{n} y_i \ \frac{1}{n}\sum_{i=1}^{n} x_i}{\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)^2} \\
&= \frac{\frac{1}{n}\sum_{i=1}^{n} y_i x_i - \overline{y}\,\overline{x}}{\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \overline{x}^2} \\
&= \frac{\frac{1}{n}\sum_{i=1}^{n} (y_i - \overline{y})(x_i - \overline{x})}{\frac{1}{n}\sum_{i=1}^{n} (x_i - \overline{x})^2} \qquad = \frac{\mathrm{cov}(\mathbf{x}, \mathbf{y})}{V_x}
\end{aligned}
$$

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

# Le théorème des moindres carrés

## Théorème des moindres carrées

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

La meilleur prédiction de $y$ sachant $x$, au sens des moindres carrées 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 $\widehat{a}$ et $\widehat{b}$ estimés par :

$$\widehat{a} = \frac{\sum_{i=1}^{n}(y_i - \overline{y})(x_i - \overline{x})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} = \frac{\mathrm{cov}(\mathbf{x}, \mathbf{y})}{V_x} \qquad \text{et} \qquad \widehat{b} = \overline{y} - \widehat{a}\,\overline{x}$$

$\widehat{a}$ et $\widehat{b}$ 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).

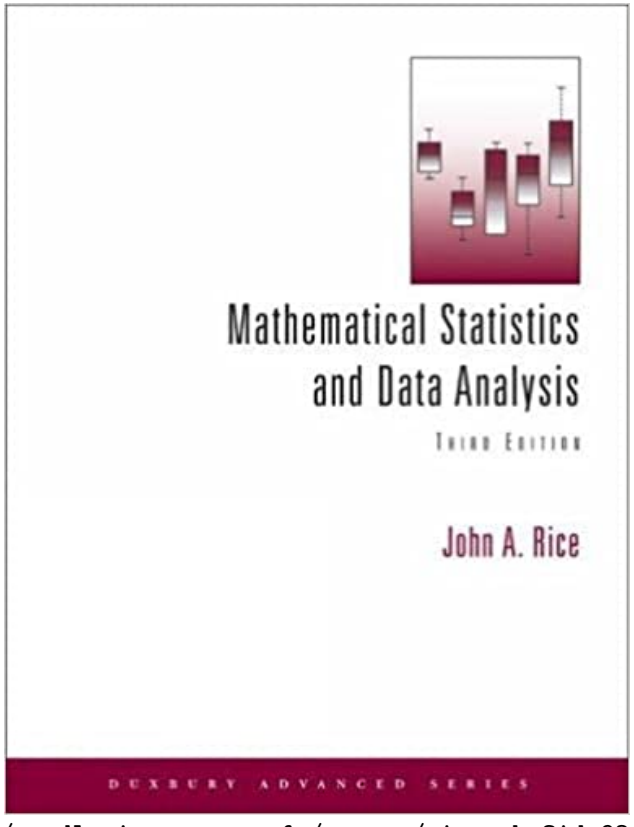
# Lecture road map

## 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

## Description d'un ensemble de variables

- Cas des variable quantitatives
- Cas des variables qualitatives




https://moodle.insa-rouen.fr/course/view.php?id=93

# Codage disjontif complet (one hot vector)

(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)

$\Longrightarrow$

$H =$

| | Bob | Percy | John | télé | ordi | aspi | lave v. |
|---|---|---|---|---|---|---|---|
| v1 | 1 | 0 | 0 | 0 | 1 | 0 | 0 |
| v2 | 0 | 1 | 0 | 0 | 0 | 1 | 0 |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| v8 | 1 | 0 | 0 | 0 | 0 | 0 | 1 |

Table: Exemple de codage disjonctif complet.

# Tableau de contingence

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

(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)

$\Longrightarrow \quad T =$

| | 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.

Matrice de co occurrence $O$ (liée ici au tableau de contingence)

$$O = H^t H$$

Lien entre le codage disjonctif complet et la table de contingence

# Mesures d'indépendance (d'association)

Distance du chi2

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

Le $\Phi^2$ 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

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

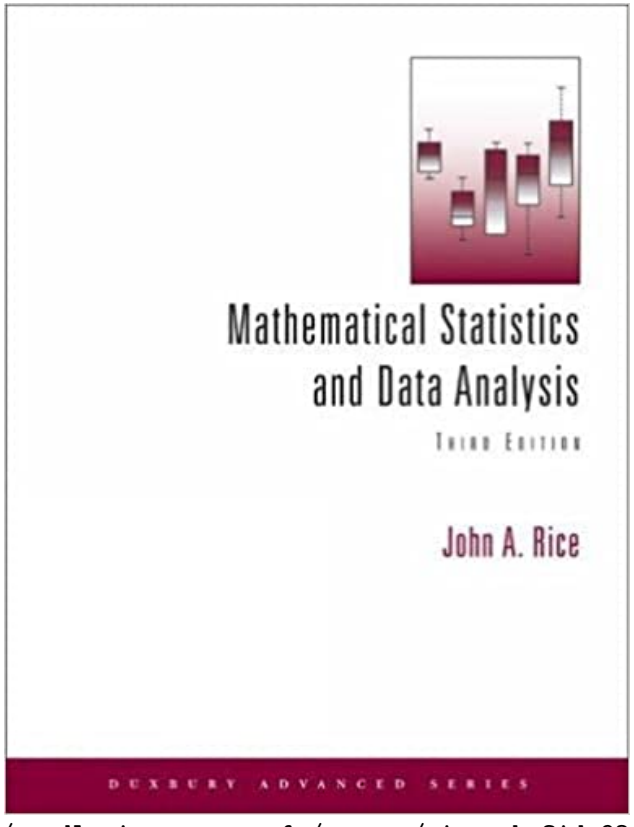
# Lecture road map

## 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

## Description d'un ensemble de variables

- Cas des variable quantitatives
- Cas des variables qualitatives




https://moodle.insa-rouen.fr/course/view.php?id=93

# Formules de décomposition de la moyenne

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

il y a $r$ vendeurs

| *vendeur* | *prix* |
|---|---|
| *Alice* | $y_1$ |
| $\vdots$ | $\vdots$ |
| *Bill* | $y_i$ |
| $\vdots$ | $\vdots$ |
| *Sue* | $y_n$ |

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

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

# Formules de décomposition de la variance

il y a $r$ vendeurs

| *vendeur* | *prix* |
|---|---|
| *Alice* | $y_1$ |
| ⋮ | ⋮ |
| *Bill* | $y_i$ |
| ⋮ | ⋮ |
| *Sue* | $y_n$ |

Décomposition de la variance,

$$\begin{aligned}
\text{SC total} &= \sum_{i=1}^{n}(y_i-\overline{y})^2 = \sum_{m=1}^{r}\sum_{j=1}^{n_m}(y_{mj}-\overline{y})^2 \\
&= \sum_{m=1}^{r}\sum_{j=1}^{n_m}(y_{mj}-\overline{y}_m+\overline{y}_m-\overline{y})^2 \\
&= \sum_{m=1}^{r} n_m(\overline{y}_m-\overline{y})^2 + \sum_{m=1}^{r}\sum_{j=1}^{n_m}(y_{mj}-\overline{y}_m)^2 \\
\text{SC total} &= \text{SC expliqués} + \text{SC résiduels}
\end{aligned}$$

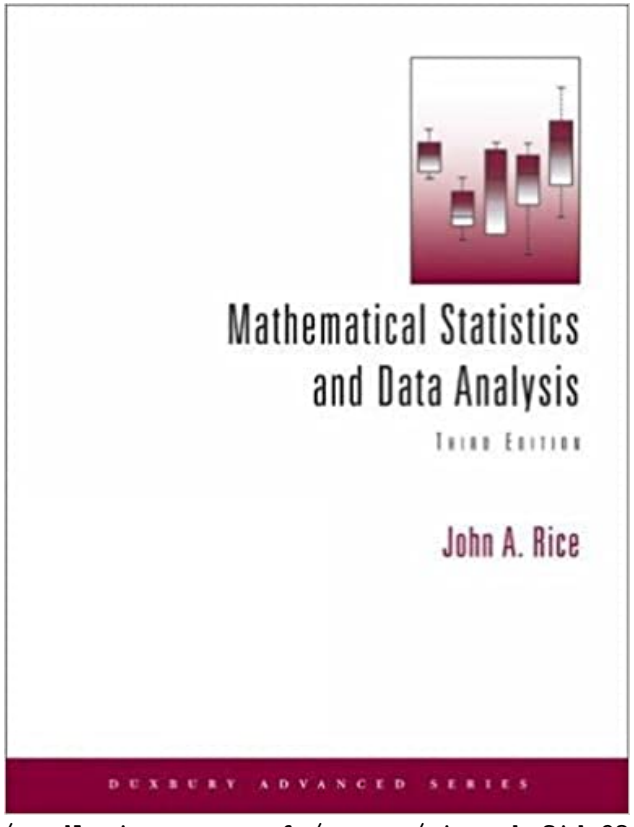
# Lecture road map

## 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

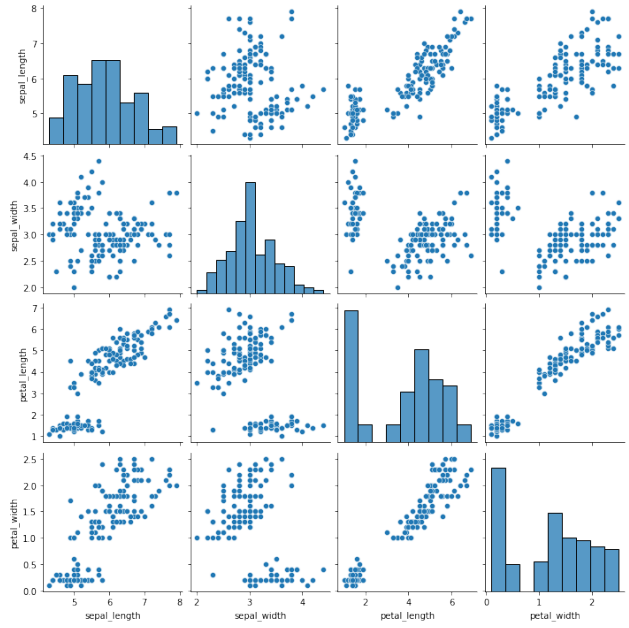
## Description d'un ensemble de variables

- Cas des variable quantitatives
- Cas des variables qualitatives




https://moodle.insa-rouen.fr/course/view.php?id=93

# Variables quantitatives : Multiplot

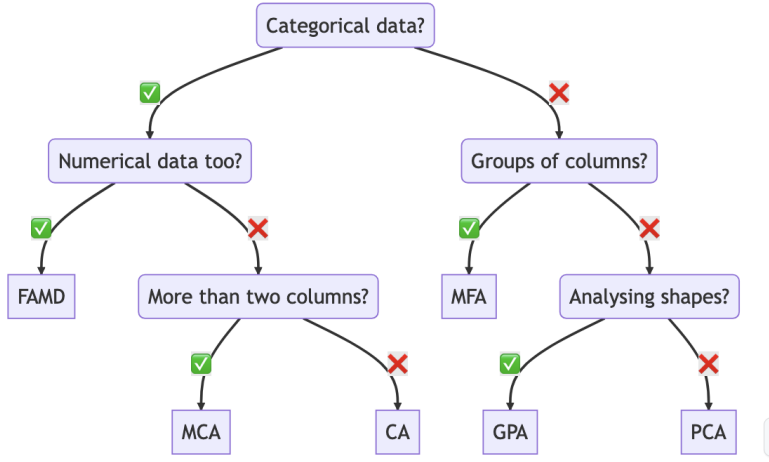



```
from pandas.plotting import scatter_matrix

df = pd.read_csv(data)
scatter_matrix(df)
```

# Variables quantitatives : L'ACP

$$\min_{U,V} \|X - UV^\top\|^2$$




**Principal component analysis (PCA)**

**Correspondence analysis (CA)**

**Multiple correspondence analysis (MCA)**

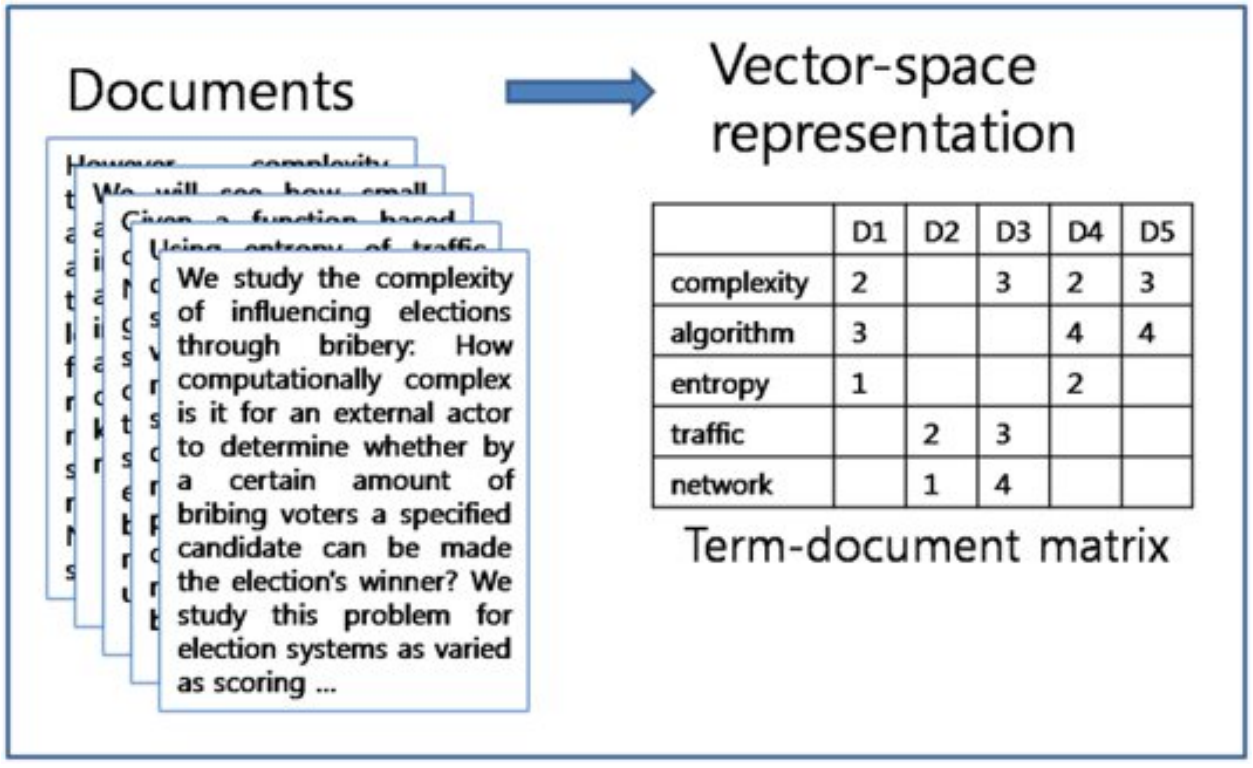
**Multiple factor analysis (MFA)**

**Factor analysis of mixed data (FAMD)**

**Generalized procrustes analysis (GPA)**

`https://github.com/MaxHalford/prince`

# Variables qualitatives : l'exemple des textes




| | D1 | D2 | D3 | D4 | D5 |
|---|---|---|---|---|---|
| complexity | 2 | | 3 | 2 | 3 |
| algorithm | 3 | | | 4 | 4 |
| entropy | 1 | | | 2 | |
| traffic | | 2 | 3 | | |
| network | | 1 | 4 | | |

Term-document matrix

$$\min_{U,V} \|X - UV^\top\|^2$$

# Comment trouver $U$ et $V$?

## Approximation de faible rang de $Y$

$$\min_{U,V} \quad \mathcal{L}(Y, U, V) := \|Y - UV^t\|_F^2 = \sum_{i=1}^{n}\sum_{j=1}^{p} (Y_{ij} - U_{i\bullet}V_{\bullet j})^2$$

$$\text{with} \quad \text{rang}(UV^t) = d$$

Eckart & Young 1936

Solution: SVD $\widehat{Y} = USV^t$

```
from scipy.sparse.linalg import svds

Y = normalise(Y)            # depend de la nature de Y
d = 300                     # taille de la representation
U, S, Vt = svds(Y, k=d)
Vt = np.diag(S)@Vt          # ou U = U*S
```

# Tableau de Burt (variables qualitative)

Table de tables de contingences (plus de deux modalités, deux ou plus variables qualitatives)

| | L2 | L3 | L4 | WN | WV | WA | TC | TR | TD | TS |
|---|---|---|---|---|---|---|---|---|---|---|
| L2 | 1512 | 0 | 0 | 788 | 483 | 241 | 433 | 385 | 399 | 295 |
| L3 | 0 | 375 | 0 | 203 | 23 | 149 | 64 | 82 | 86 | 143 |
| L4 | 0 | 0 | 113 | 62 | 9 | 42 | 3 | 29 | 21 | 60 |
| WN | 788 | 203 | 62 | 1053 | 0 | 0 | 229 | 284 | 273 | 267 |
| WV | 483 | 23 | 9 | 0 | 515 | 0 | 174 | 133 | 125 | 83 |
| WA | 241 | 149 | 42 | 0 | 0 | 432 | 97 | 79 | 108 | 148 |
| TC | 433 | 64 | 3 | 229 | 174 | 97 | 500 | 0 | 0 | 0 |
| TR | 385 | 82 | 29 | 284 | 133 | 79 | 0 | 496 | 0 | 0 |
| TD | 399 | 86 | 21 | 273 | 125 | 108 | 0 | 0 | 506 | 0 |
| TS | 295 | 143 | 60 | 267 | 83 | 148 | 0 | 0 | 0 | 498 |
| | L2 | L3 | L4 | WN | WV | WA | TC | TR | TD | TS |

Exemple avec trois variables de respectivement 3, 4 et 3 modalités.

https://fr.wikipedia.org/wiki/Table_de_Burt

# Conclusion

Attention à la nature des variables

## Comment représenter 2 variables ?

par une matrice... $n \times 2$

- 2 quantitatives -> nuage de point
- 2 qualitatives -> tableau de contingence
- Une de chaque -> $r$ histogrammes

## Ces 2 variables sont elles liées ?

- 2 quantitatives -> corrélation
- 2 qualitatives -> distance du $\chi^2$ (coefficient de Cramer)
- Une de chaque -> décomposition de la variance

Qualitatif -> Quantitatif : valeurs singulières (propre)