Introduction

Mohamed Ndaoud

Presentation

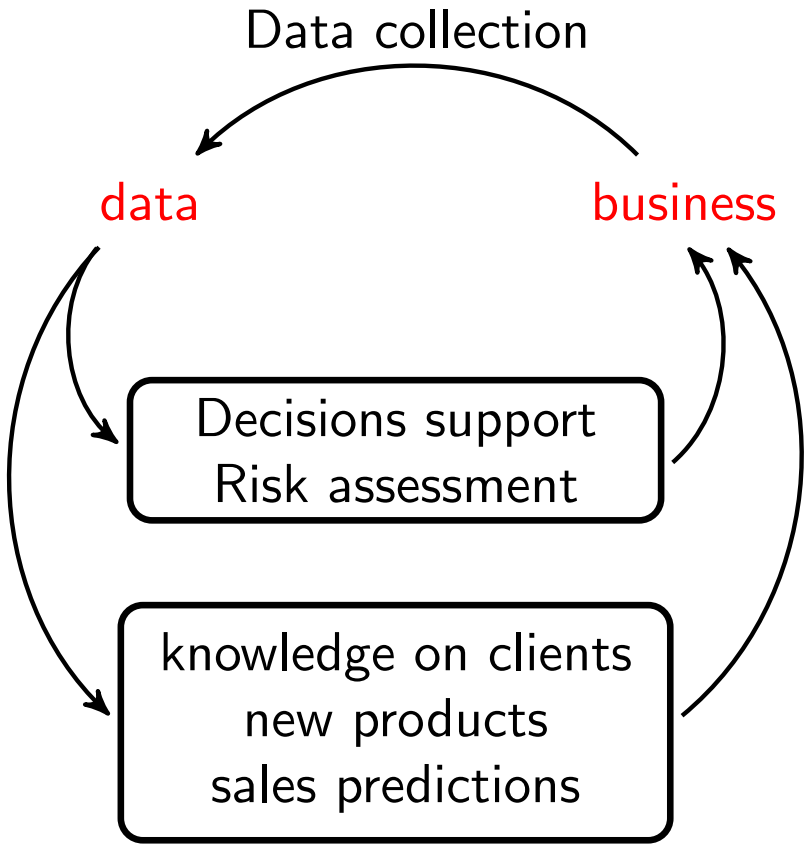
- Associate Professor of Statistics at ESSEC Business School
- Ex-member of the math department at University of Southern California (USC)
- PhD in theoretical statistics (X-ENSAE)
- Research interests:
  - High dimensional statistics
  - Robust statistics
  - Clustering
  - Fairness
- Contact:
  - ndaoud@essec.edu
  - Office : Le Nautile, N324

Course Objectives

- Provide methodological principles of multidimensional data analysis methods
- Learn how to deploy effective decision-making models to a production environment
- Python Programming
- Familiarize with:
  - Data preparation
  - Data visualization
  - Text mining
  - Classification and Prediction
  - Clustering
  - Optimization

Aim of the course: data science for business

- Understand what data can do for business
- Understand how to take decisions using data in an uncertain environment and how to evaluate risk



The virtuous circle between data and business.

Programming language

- free
- easy to install with Anaconda
- user friendly Jupyter Notebooks and Google colab
- Many Notebooks available (Kaggle, Github repos,...)
- Many powerful libraries (pandas, sklearn, XGboost, Keras, etc.)
- Visualization libraries (matplotlib, seaborn)
- large community to help (stackoverflow)

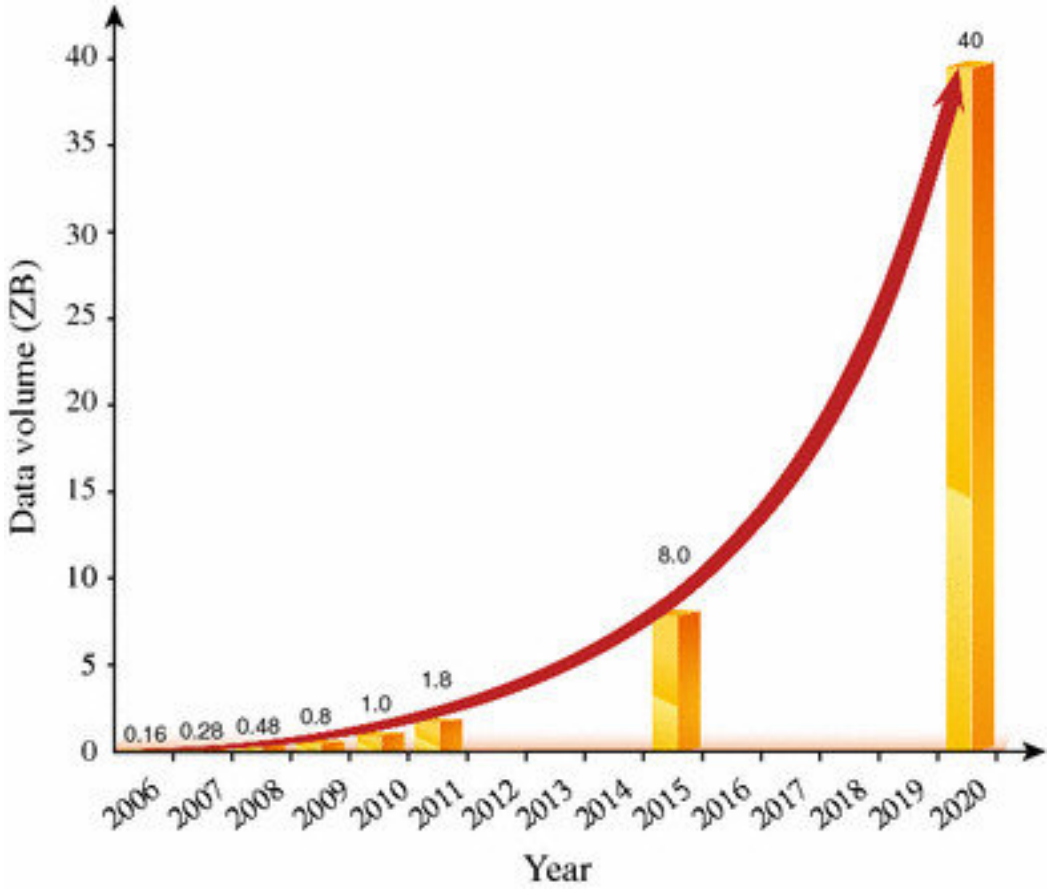
The aim of the course is not to learn python but to know what can be done with data and what python can do with data.

Course organization

- **Course:**
  - Presentation of statistical methods
  - Practical examples using Python
  - Exercises in class
  - Complementary take-home exercises

- **Grading:**
  - Final exam December 8th: 40%
  - Project (groups of 3 students): 40%
  - Participation: 20%

Digital world. Big data: the new gold.



- big data: 1 ZB $= 10^{12}$ GB; 90% of data created during the last 2 years: exponential growth.
- where are the from? What for? challenges raised? opportunities?

Where do data come from? What are they?

⊕ Many human activities went to digital and are now producing data: health, biology, news, finance, marketing, communications, academic world, e-commerce, video streaming, music, advertisement, etc.

⊕ Large databases:

- Genomics : 10^3 patients, 10^6 genes
- Social networks : 3×10^7 FaceBook active users
- French database (cartes Vitale) over the last 10 years: 30To
- 5.6 billion of searches on Google per day (most visited website, 92% of the market share, PageRank algorithm); trends.google.com; Google dataset search.

⊕ The '5V' of big data:

- Volume: only large databases are considered as big data
- Velocity: high speed of accumulation of continuous flow of data
- Variety: data come from heterogeneous sources (texts, pictures, videos, tables, log-files, discussions, etc.)
- Veracity: quality of the data; inconsistencies and uncertainty in data
- Value: Data in itself is of no use or importance but it needs to be converted into something valuable. Because of the last 4V's, data has value (for those who know what to do with them).

Data for Sales predictions

TV-Radio-Newspaper and sales

Budgets of ads campaigns and returned sales

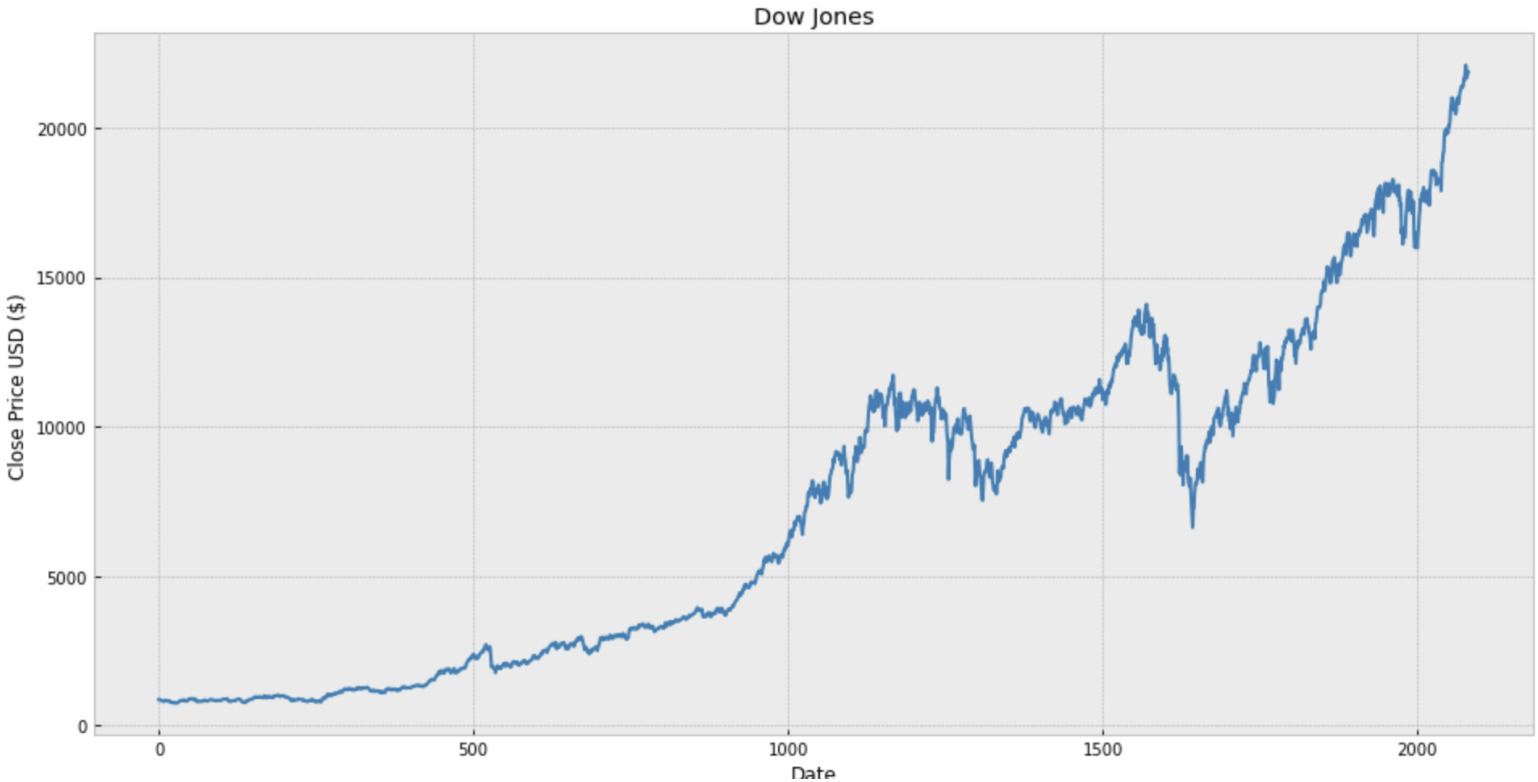
	TV	Radio	Newspaper	Sales
1	230.1	37.8	69.2	22.1
2	44.5	39.3	45.1	10.4
3	17.2	45.9	69.3	9.3
4	151.5	41.3	58.5	18.5
5	180.8	10.8	58.4	12.9

`http://www.insee.fr/`; `https://www.data.gouv.fr/`;
`https://datasetsearch.research.google.com`

TODO: open `linear-regression-Ads-Sales.ipynb`

Data for stock price predictions

Times series from finance



`http://fr.finance.yahoo.com/`;
`http://www.bloomberg.com/enterprise/data/`;
`https://www.investing.com/indices/us-30-historical-data`
`IEX Cloud, a financial data service`

TODO: open `prediction-dow-jones.ipynb`

e-commerce data for recommendation systems

Data of the form:

(user, item) or (user, item, grade)

plus some *metadata*: users data, items data, transaction data

Netflix Prize: recommend movies to users

- 480.189 users (no data on users for privacy), 17.770 movies (title, year of release), 1.408.395 grades + dates of grades
- from October 2006 to September 2009
- 1 million dollar prize for the first team improving by 10% Netflix own algorithm *CineMatch*.
- 40.000 teams over the world
- 2009: class action lawsuit against Netflix for privacy reasons.

e-commerce data for recommendation systems

Data from 12 months logs (Mar. 2016 - Feb. 2017) from 'CI&T's Internal Communication platform (DeskDrop). 73.000 logged users interactions on more than 3.000 public articles shared in the platform.

personId	contentId	eventStrength
-9223121837663643404	-8949113594875411859	1.000000
-9223121837663643404	-8377626164558006982	1.000000
-9223121837663643404	-8208801367848627943	1.000000
-9223121837663643404	-8187220755213888616	1.000000
-9223121837663643404	-7423191370472335463	3.169925

Aim: recommend relevant articles to 'CI&T's employees.

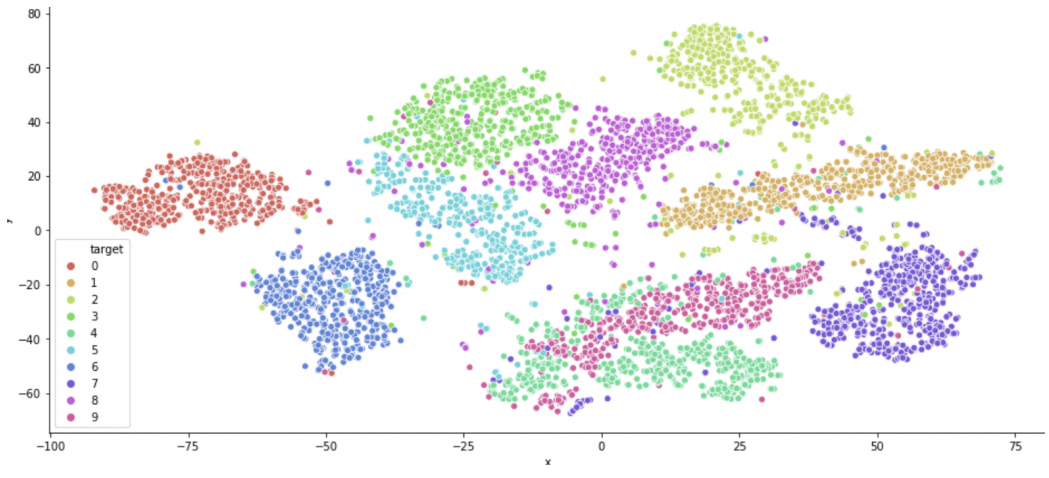
TODO: open `recommender-systems-in-python-101.ipynb`

e-commerce data for customers segmentation

e-commerce data are often big table with rows like

users metadata + historic of purchases

Customer segmentation aka customer analytics is used to help Companies to understand Customer's needs. A good KYC (Know Your Customer) makes the difference for a whole lot of companies. It may be part of a MMM = Marketing Mix modeling strategy.



TODO: open `customer-segmentation-clustering.ipynb`.

e-commerce data for prediction of churn

Type of data:

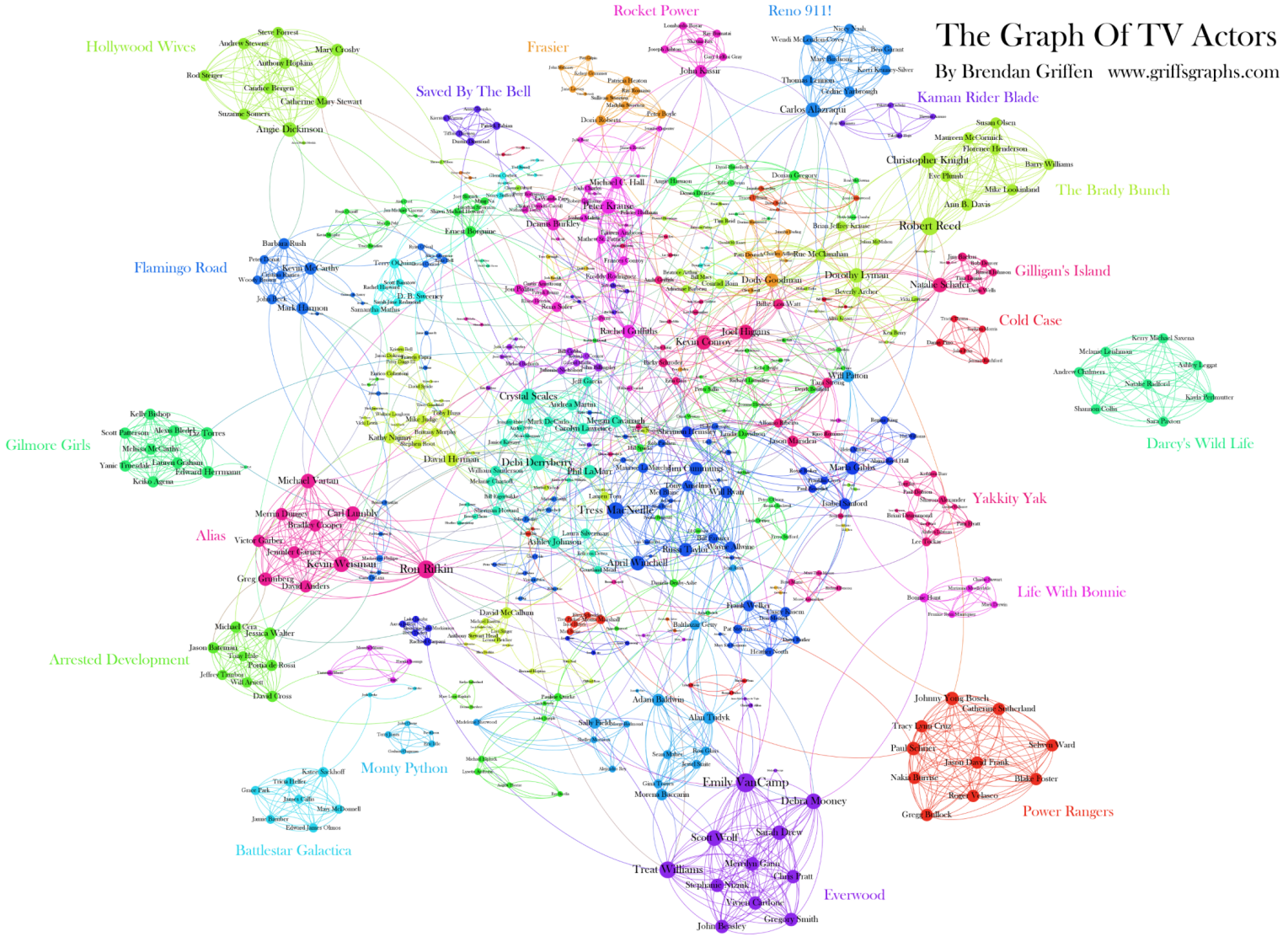
users metadata (gender, partner, type of contract,etc.), churned or not.

Churn rate aka **attrition rate** is a measure of the number of individuals moving out of a collective group over a specific period.
Probability of churn is a key factor that determines the steady-state level of a customer a business will support, in particular, business with respect to a contractual customer base (mobile telephone networks and pay TV operators).

TODO: Go to

Kaggle simple End-to-end project on Telco Customer Churn

Data may also be represented as graphs



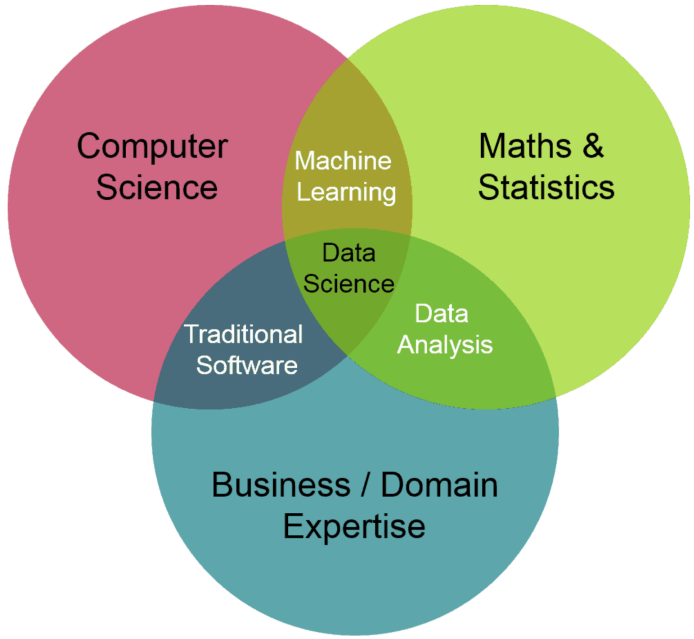
What is a data science for business project?

Aim:

- What are the roles?
- What skills are needed ?

Who is part of a data science for business project?

3 roles: IT experts, data scientists and business experts.



A successful data science for business project is when IT, datascientists and business experts succeed to work together. Usually it is the business expert leading the project and reporting the project.

What is data science for business?

The aim of data science projects is to support business decisions.

Why do we need data to support business? because we need to take decisions in an uncertain environment.

- The fundamental notion behind statistics and machine learning is uncertainty: if there was no uncertainty then there will be no need for statistics and machine learning;
- uncertainty may be looked as a drawback - because it can lead to bad decisions - one may however consider it as an opportunity to make a difference by taking better decisions than others.
- next step is to be followed by the others when taking decisions in uncertain environment: all decisions taken in an uncertain environment are risky.
- the second step is therefore to know how to deal with this risk. Assessing this risk is key to risk control which is key to lead.

Data are used to support decisions and to measure their associated risk in uncertain environments.

How do we deal with uncertainty?

Aim:

- the statistical modelling approach to the decision process and risk control,
- the machine learning approach.

How do we deal with uncertainty in statistics?

Statistical models are used to model uncertainty.

1) Our starting point are the data (like real numbers):

$$\mathtt{x}_1, \ldots, \mathtt{x}_n$$

2) Statistical modeling :
   - Data are realization of random variables:

$$\mathtt{x}_1 = X_1(\omega), \ldots, \mathtt{x}_1 = X_n(\omega)$$

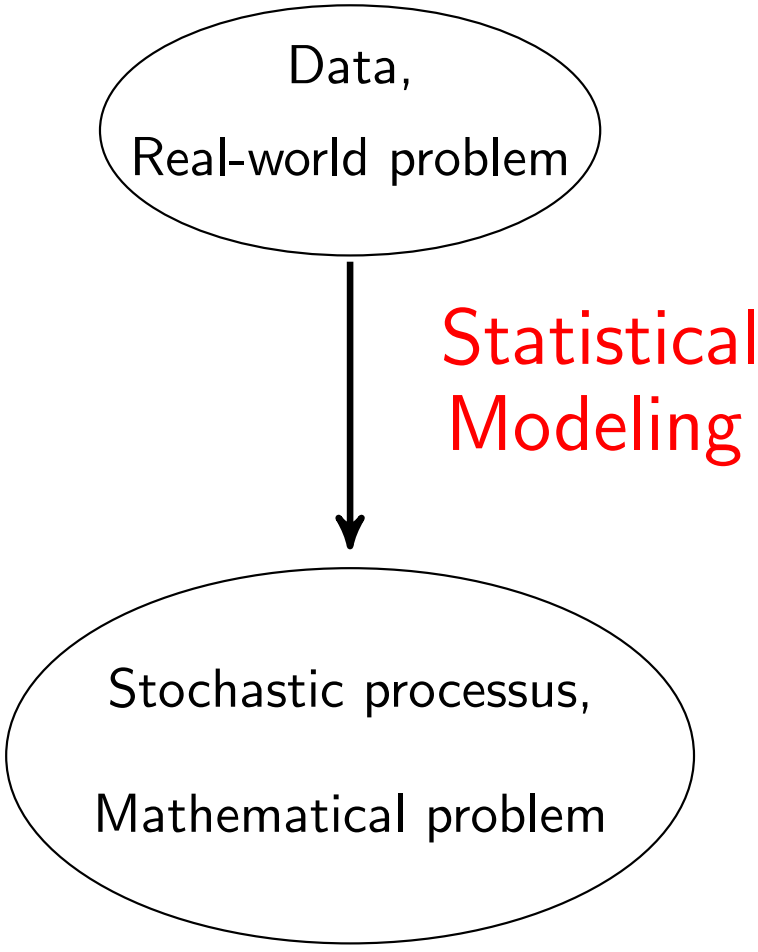
     where $X_1, \ldots, X_n$ are random variables.
   - The probability distribution of $(X_1, \ldots, X_n)$ is unknown. However, we assume a priori that it belongs to a given family of probability distributions

$$\boxed{\left\{\mathbb{P}_\theta^n, \theta \in \Theta\right\}}: \text{ the statistical model}$$

   Idea: We believe that there exists some $\theta \in \Theta$ such that the data are distributed according to $\mathbb{P}_\theta^n$.

3) Standard problems in statistics: Given the observations $(X_1, \ldots, X_n)$, can we estimate θ? test some properties on θ?

What are statistical models used for?



Why are we modelling data?

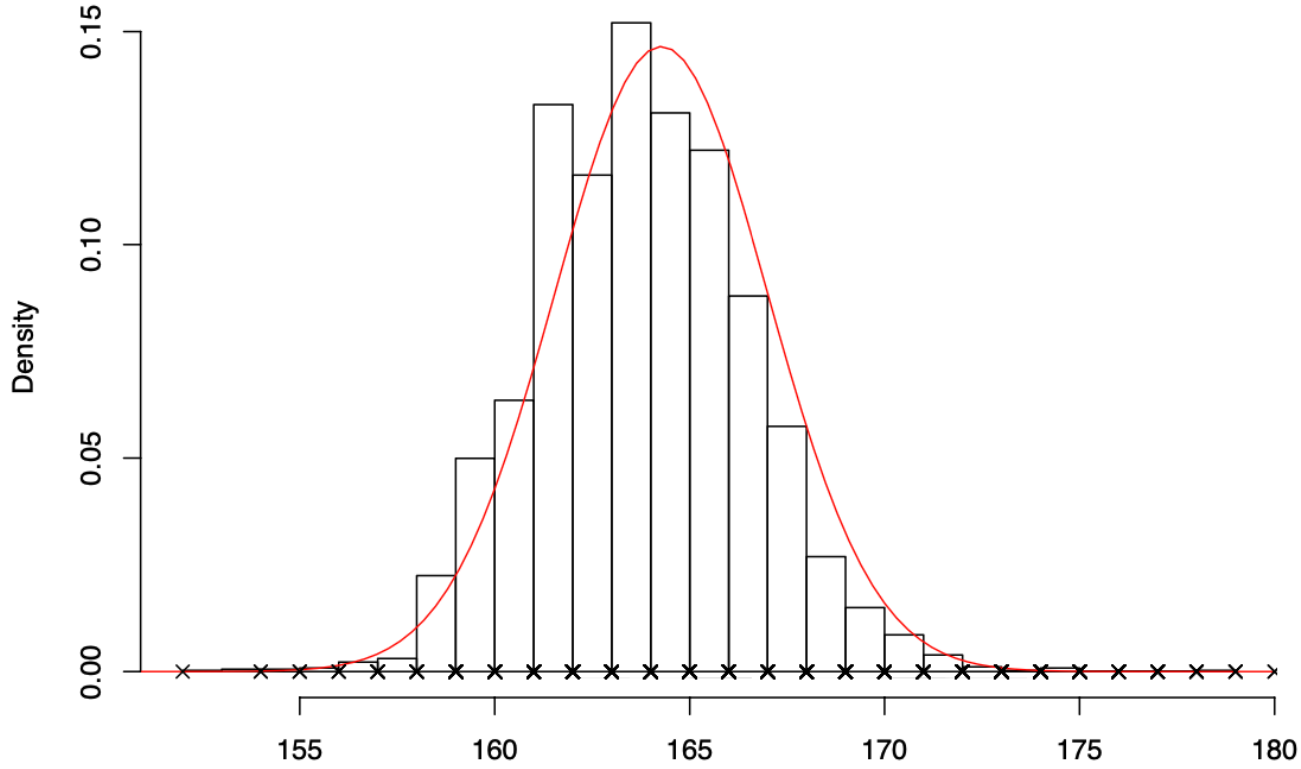
1) Mathematical tools

2) Mathematical results

3) Algorithms

A solution to the mathematical problem may be used to support decisions and evaluate their risks to solve the real-world problem.

Example of statistical modeling: women's height in the US



- We may use a Gaussian distribution $\mathcal{N}(\mu, \sigma)$ where μ is the mean and σ is the square root deviation.
- $\theta = (\mu, \sigma)$ is the parameter of the model
- $\{\mathcal{N}(\mu, \sigma) : \mu > 0, \sigma > 0\}$ is the statistical model.

How do we deal with uncertainty in Machine learning?

In machine learning, all is about generalization.

Generalization: we say that a procedure / machine learning model generalizes well when it performs well on unseen data (i.e. data that were not used for its construction).

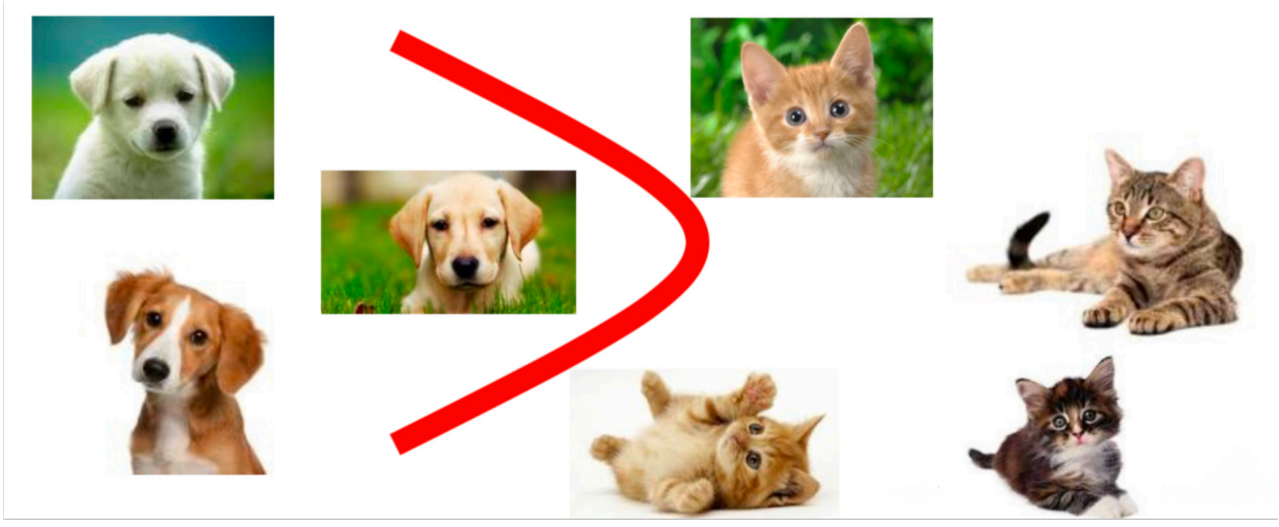
Idea: If the machine learning model generalizes well then we believe it is able to deal with all sources of uncertainty. Moreover, by evaluating its performance on unseen data we can estimate its risk.

Key point in ML methodology: A part of the data is used for assessing the generalization property of the ML procedure as well as estimating its risk i.e. its deviation from reality.
$\Rightarrow$ There is a data splitting step in ML. The most classical ones are:

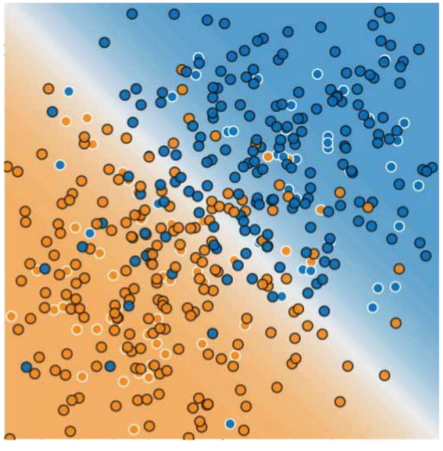
- train/test sample splitting
- cross-validation.

Example: Sample splitting in binary classification

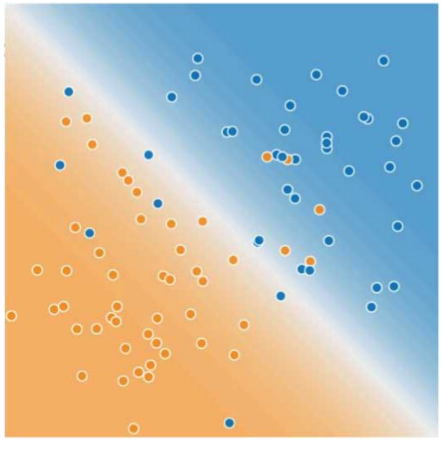
Pb.: construct a procedure that discriminates pictures of dogs and cats



Generalization properties are evaluated on the test sample.



Training Data



Test Data

The six steps of a data science project

1) Define the project + business's expert risk aversion (KPI)

2) Data collection (internal data's compagny + open data)

3) Data cleaning (missing values, outliers)

4) Data visualization (descriptive statistics)

5) Construct procedures (from statistical modelling and ML algorithms) and find the one with the best generalization performance = the one with smallest risk on the test set.

6) Prepare your talk (verbal + visual)

Probability Refresher

M. Ndaoud

Statistical paradigm

1) Starting point : data (ex.: real numbers)

$$\mathtt{x}_1, \ldots, \mathtt{x}_n$$

2) **Statistical modeling** :

- data are realizations

$$X_1(\omega), \ldots, X_n(\omega) \text{ of r.v. } X_1, \ldots, X_n.$$

(in other words, for a certain ω, $X_1(\omega) = \mathtt{x}_1, \ldots, X_n(\omega) = \mathtt{x}_n$)

- The **distribution** $\mathbb{P}^{(X_1,\ldots,X_n)}$ of $(X_1, \ldots, X_n)$ is unknown, but belongs to a given family (a priori)

$$\boxed{\left\{ \mathbb{P}_\theta^n, \theta \in \Theta \right\}} : \text{the model}$$

We believe that there exists $\theta \in \Theta$ such that $\mathbb{P}^{(X_1,\ldots,X_n)} = \mathbb{P}_\theta^n$.

- θ is the parameter and Θ the set of parameters.

Statistical paradigm (Cont'd)

Problem: from the "observation" $X_1, \ldots, X_n$

- **Modeling:** which model to choose ?

Statistical paradigm (Cont'd)

Problem: from the "observation" $X_1, \ldots, X_n$

- **Modeling:** which model to choose ?
- Estimation : construct a function $\phi_n(X_1, \ldots, X_n)$ that approximates the best θ

Statistical paradigm (Cont'd)

Problem: from the "observation" $X_1, \ldots, X_n$

- **Modeling:** which model to choose ?
- Estimation : construct a function $\phi_n(X_1, \ldots, X_n)$ that approximates the best θ
- Test : Establish a decision $\varphi_n(X_1, \ldots, X_n) \in \{\text{set of decisions}\}$ concerning a hypothesis about θ.

Statistical paradigm (Cont'd)

Problem: from the "observation" $X_1, \ldots, X_n$

- **Modeling:** which model to choose ?
- Estimation : construct a function $\phi_n(X_1, \ldots, X_n)$ that approximates the best θ
- Test : Establish a decision $\varphi_n(X_1, \ldots, X_n) \in \{\text{set of decisions}\}$ concerning a hypothesis about θ.
- Prediction : Guess the unobserved value X_{n+1} based on $X_1, \ldots, X_n$

Example of head or tail

- We toss a coin 18 times and observe ($H = 0,\ T = 1$)

$$0, 0, 0, 1, 1, 0, 1, 0, 0, 1, 1, 0, 1, 0, 0, 1, 1, 0$$

- statistical model : we observe $n = 18$ independent random variables X_i, Bernoulli of unknown parameter $\theta \in \Theta = [0, 1]$.

Example of head or tail

- We toss a coin 18 times and observe ($H = 0,\ T = 1$)

$$0, 0, 0, 1, 1, 0, 1, 0, 0, 1, 1, 0, 1, 0, 0, 1, 1, 0$$

- statistical model : we observe $n = 18$ independent random variables X_i, Bernoulli of unknown parameter $\theta \in \Theta = [0, 1]$.
  - Estimation. Estimator $\bar{X}_{18} = \frac{1}{18} \sum_{i=1}^{18} X_i \overset{\text{here}}{=} 8/18 = 0.44$. What precision ?

Example of head or tail

- We toss a coin 18 times and observe ($H = 0,\ T = 1$)

$$0, 0, 0, 1, 1, 0, 1, 0, 0, 1, 1, 0, 1, 0, 0, 1, 1, 0$$

- statistical model : we observe $n = 18$ independent random variables X_i, Bernoulli of unknown parameter $\theta \in \Theta = [0, 1]$.
  - Estimation. Estimator $\bar{X}_{18} = \frac{1}{18} \sum_{i=1}^{18} X_i \overset{\text{here}}{=} 8/18 = 0.44$. What precision ?
  - Test. Decision to make : "is the coin balanced ?". For example: we compare $\bar{X}_{18}$ to 0.5. If $|\bar{X}_{18} - 0.5|$ "small", we accept the hypothesis "the coin is balanced". Otherwise, we reject.
  - Prediction. If we toss the same coin a new time, is the outcome more likely to be head or tail?

Fundamental Theorems

The strong law of large numbers (LLN)

Theorem

Let (X_n) be a sequence of i.i.d. random variables such that $\mathbb{E}\,|X_1| < \infty$. Then

$$\frac{1}{n}\sum_{i=1}^{n} X_i \overset{a.s.}{\to} \mathbb{E}\,X_1$$

Central Limit Theorem (CLT)

Theorem

Let (X_n) be a sequence of i.i.d. random variables such that $\mathbb{E}\, X_1^2 < \infty$. Then

$$\frac{\sqrt{n}}{\sigma}\Big(\frac{1}{n}\sum_{i=1}^{n} X_i - \mathbb{E}\, X_1\Big) \xrightarrow{d} \mathcal{N}(0,1)$$

- CLT : "speed" of convergence in the LLN.
- Interpretation of CLT :

$$\frac{1}{n}\sum_{i=1}^{n} X_i = \mu + \frac{\sigma}{\sqrt{n}}\,\xi^{(n)}, \quad \xi^{(n)} \overset{d}{\approx} \mathcal{N}(0,1).$$

- The type of convergence is a convergence in distribution. (weak convergence).

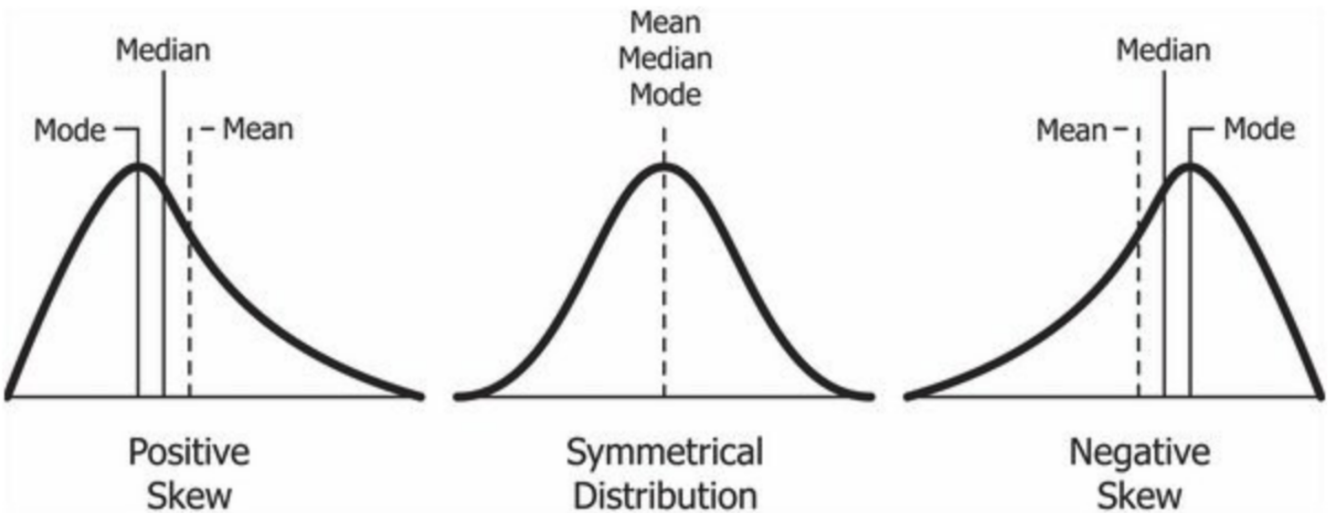
Other parameters of probability distributions

The mode of X is:

- in the discrete case, the values x such that $\mathbb{P}[X = x]$ is the largest one
- in continuous case, the point x where the density function $f(x)$ is the largest one.

The skewness and Kurtosis: of X are

$$\textit{skewness} = \mathbb{E}\left[\left(\frac{X-\mu}{\sigma}\right)^3\right] \text{ and } \textit{kurtosis} = \mathbb{E}\left[\left(\frac{X-\mathbb{E}X}{\sigma}\right)^4\right].$$



(quantities used to test for normality in the Jarque-Bera test).

Graphical Statistics

- Quantiles
- Covariance and correlation

Cumulative distribution function (cdf)

Population cdf :

$$F(x) = \mathbb{P}\left(X \leq x\right), \quad x \in \mathbb{R}$$

Cumulative distribution function (cdf)

Population cdf :

$$F(x) = \mathbb{P}\left(X \leq x\right), \quad x \in \mathbb{R}$$

Empirical cdf :

$$\widehat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} I\left(X_i \leq x\right), \quad x \in \mathbb{R}$$

Some asymptotic properties:

$$\widehat{F}_n(x) \xrightarrow{a.s.} F(x), \quad \left\|\widehat{F}_n - F\right\|_\infty \xrightarrow{a.s.} 0$$

Quantiles

Quantiles

Definition

Let X be a r.v. (of cdf F) and $0 < p < 1$. We call quantile of order p of X (resp. F) :

$$q_p(F) = \inf\{x \in \mathbb{R} : F(x) \geq p\}$$

- When F is continuous and strictly increasing the **quantile of order** p of F is the unique solution to

$$F(q_p) = p \qquad (\text{ i.e. } \quad q_p = F^{-1}(p)\ ).$$

- the **median** $= \mathrm{med}(F) = q_{1/2}(F)$
- the **quartiles** $= \{q_{1/4}(F), \mathrm{med}(F), q_{3/4}(F)\}$

Population and empirical quantiles

The "population" quantile of order p:

$$T(F) = q_p(F) = \inf\{x \in \mathbb{R} : F(x) \geq p\}$$

The "empirical" quantile of order p:

$$T(\widehat{F}_n) = \widehat{q}_{n,p} = \inf\{x \in \mathbb{R} : \widehat{F}_n(x) \geq q\}$$

Definition

Let $X_1, \ldots, X_n$ be a sample of size n of r.v. We call order statistics the n statistics $X_{(1)}, \ldots, X_{(n)}$ such that

$$X_{(1)} \leq \cdots \leq X_{(n)}$$

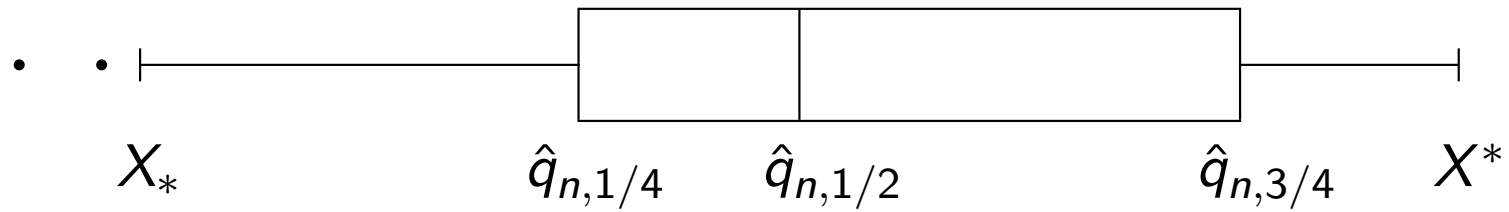
1. For the quantile of order $0 < p < 1$:

$$\boxed{\widehat{q}_{n,p} = X_{(k)} = X_{(\lceil np \rceil)} \text{ when } \frac{k-1}{n} < p \leq \frac{k}{n}}$$

2. In particular, the empirical median satisfies :

$$\boxed{\widehat{q}_{n,1/2} = \mathrm{med}(\widehat{F}_n) = X_{(\lceil n/2 \rceil)}} \text{ where } \lceil t \rceil = \min(n \in \mathbb{N} : n \geq t)$$

The boxplot : synthetic representation of the dispersion of real data



end of the whiskers :

$$X_* = \min\{X_i : |X_i - \hat{q}_{n,1/4}| \leq 1,5\mathcal{I}_n\},$$

$$X^* = \max\{X_i : |X_i - \hat{q}_{n,3/4}| \leq 1,5\mathcal{I}_n\}.$$

Interquartile range:

$$\mathcal{I}_n = \hat{q}_{n,3/4} - \hat{q}_{n,1/4}.$$

Samples beyond the whiskers are considered as *outliers*.

The qq-plot : fit test to some distribution

Given a sample of size n $X_1, \ldots, X_n$ and a cdf F_{ref}, we want to test if the following hypothesis is true :

(H_0) "The X_i are distributed according to F_{ref}"

To "accept or reject visually" this hypothesis, we can draw the qq-plot : it is a scatter plot

$$\boxed{\left(q_{i/n}(F_{ref}), \widehat{q}_{n,i/n}\right)_{i=1}^{n} = \left(q_{i/n}(F_{ref}), X_{(i)}\right)_{i=1}^{n}}$$

1. If the scatter plot is "approximately" aligned with the line $y = x$ then we accept the hypothesis (we also draw the line $y = x$ on the qq-plot)
2. If the scatter plot is "approximately" aligned with a line then the hypothesis is true after centering and scaling (generally, we normalize data first)

Examples of Q-Q plots

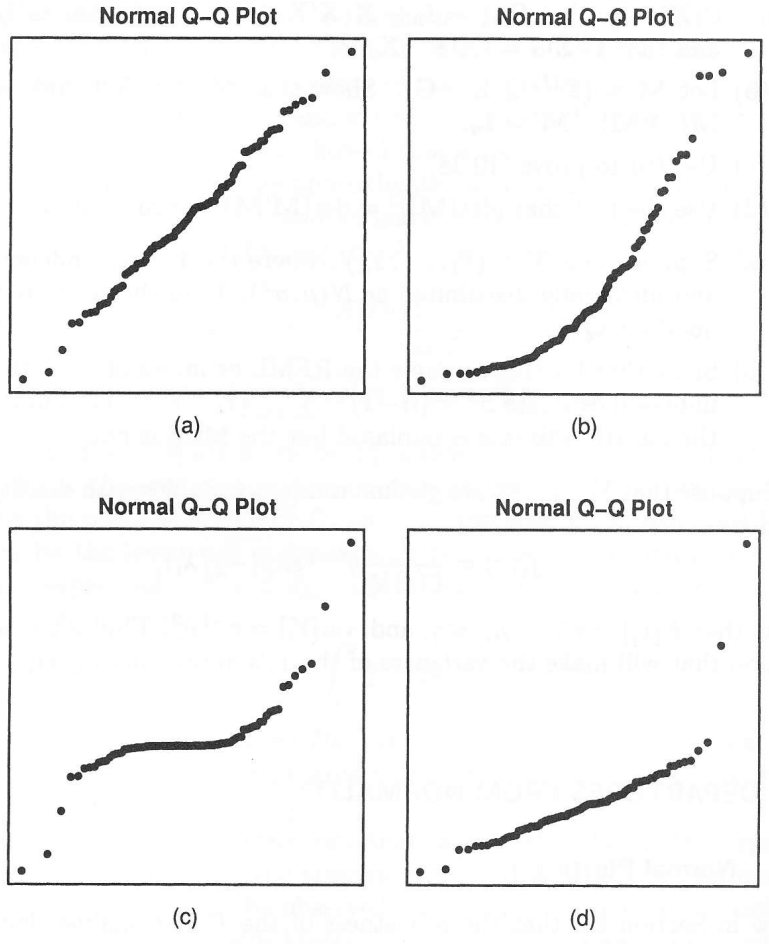


Fig. 10.5 Normal plots of residuals: (a) No indication of non-normality. (b) Skewed errors.(c) Heavy-tailed errors. (d) Outliers.

Covariance et correlation

Dependence between two random variables

- In order to measure the dependence between X and Y, it is relevant to quantify the variation of one variable with respect to the other one.

- If X increases, for example, does Y increase too? If so, what is the level of this dependence?

- In order to address the above questions we introduce the notion of **covariance/correlation**.

The notion of covariance

Definition

Let X and Y be two random variables with finite variances. We define the covariance between X and Y such that

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

- If X and Y are two independent variables then $\mathrm{Cov}(X, Y) = 0$.
- The reverse **is not always true**.
- For applications, it is desirable to have a dimension-free measure of dependence.

The correlation coefficient

Definition

Let X and Y be two r.v. The correlation coefficient between X and Y, that we denote $\mathrm{Cor}(X, Y)$, is given by

$$\mathrm{Cor}(X, Y) = \frac{\mathrm{Cov}(X, Y)}{\sigma_X \sigma_Y} = \mathrm{Cov}(X_*, Y_*)$$

where $X_* = (X - \mathbb{E}(X))/\sigma_X$ and $Y_* = (Y - \mathbb{E}(Y))/\sigma_Y$.

Proposition

For each pair of random variables (X, Y) we have

- $|\mathrm{Cor}(X, Y)| \leq 1$.
- $|\mathrm{Cor}(X, Y)| = 1$ **if and only if** $Y = aX + b$ *for constants a and b in $\mathbb{R}$.*

Estimation of the correlation coefficient

- Suppose that the correlation coefficient between X and Y is not known, but that we observe n i.i.d. copies of $(X_1, Y_1), \ldots, (X_n, Y_n)$. How can we estimate $\mathrm{Cor}(X, Y)$?

- Recall that

$$\boxed{\mathrm{Cor}(X, Y) = \frac{\mathbb{E}(XY) - \mathbb{E}(X)\,\mathbb{E}(Y)}{\sqrt{\mathrm{Var}(X)}\sqrt{\mathrm{Var}(Y)}}}$$

Estimation of the correlation coefficient

- Suppose that the correlation coefficient between X and Y is not known, but that we observe n i.i.d. copies of $(X_1, Y_1), \ldots, (X_n, Y_n)$. How can we estimate $\mathrm{Cor}(X, Y)$?

- Recall that

$$\boxed{\mathrm{Cor}(X, Y) = \frac{\mathbb{E}(XY) - \mathbb{E}(X)\,\mathbb{E}(Y)}{\sqrt{\mathrm{Var}(X)}\sqrt{\mathrm{Var}(Y)}}}$$

- It seems relevant to replace moments by their estimates here.

Estimation of the correlation coefficient

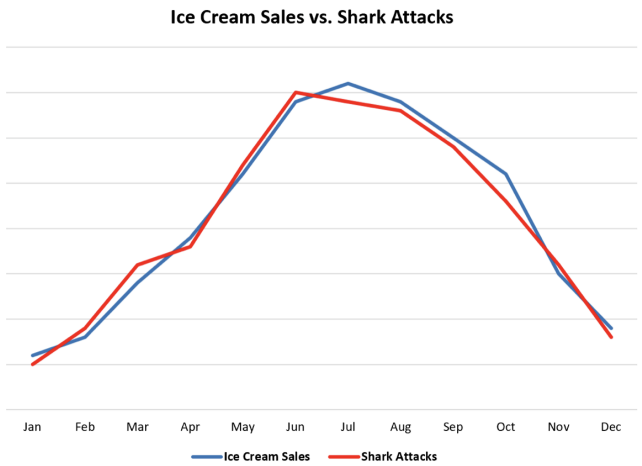
We get then what we call the empirical correlation coefficient given by

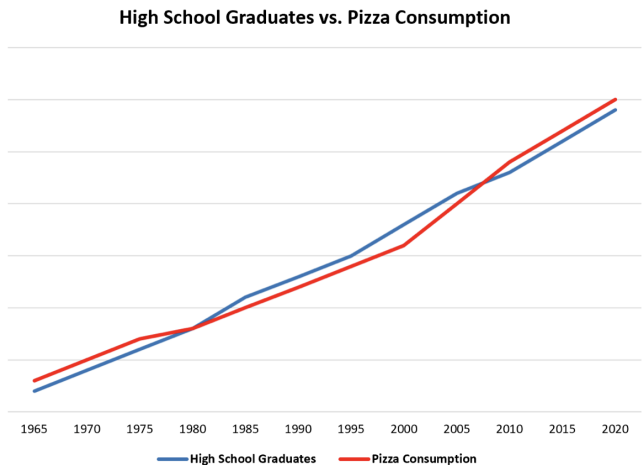
$$\hat{\rho}_n(X,Y) := \frac{n\sum_{i=1}^n X_i Y_i - \left(\sum_{i=1}^n X_i\right)\left(\sum_{i=1}^n Y_i\right)}{\sqrt{n\sum_{i=1}^n X_i^2 - \left(\sum_{i=1}^n X_i\right)^2}\sqrt{n\sum_{i=1}^n Y_i^2 - \left(\sum_{i=1}^n Y_i\right)^2}}$$

Correlation and causality-1

Correlation is not causality

From wikipedia:
'*refers to the inability to legitimately deduce a cause-and-effect relationship between two events or variables solely on the basis of an observed association or correlation between them.*
The idea that "correlation implies causation" is an example of a questionable-cause logical fallacy, in which two events occurring together are taken to have established a cause-and-effect relationship. This fallacy is also known by the Latin phrase cum hoc ergo propter hoc ('with this, therefore because of this').'





Empirical correlation matrix

We observe n i.i.d. copies of a random vector $(X_1, \ldots, X_d)$ and define the empirical correlation matrix $\rho^n \in \mathbb{R}^{d \times d}$ such that

$$\text{For all } i, j \quad \rho^n_{ij} = \hat{\rho}_n(X_i, X_j)$$

- The matrix ρ^n is semi-definite positive.

- We can simply compute the lower triangular part of this matrix.

Visualization of the empirical correlation matrix

In order to graphically visualize the different correlations, we use the tool heatmap in Python.