Machine Learning
Predictive Analytics

Nistor GROZAVU

nistor@lipn.univ-paris13.fr

Data mining & Machine Learning

«Data mining offers the capability to view data in a new light, discovering associations and patterns not appreciated before. For the humanities domain, it exemplifies the interdisciplinary efforts of digital humanities. »

Jonathan Hagood

Outline

- Introduction : Univariated and bivariated statistics; (2h)
- Projection : PCA (unsupervised) & LDA (supervised) (2h)
- Unuspervised Learning:
  - Clustering and co-clustering : k-means, CAH, weighted k-means, nmf k-means (2h)
  - Neural networks & Self-Organizing Maps (1h)
- Supervised Learning:
  - Decision Trees (1h)
  - Classification & Prediction methods (2h)
- Validation : DB index, Accuracy index, Rand, Shiluette,… (1h)
- Applications (1h)

Data mining & Machine Learning

- Data mining & Machine Learning (applying statistics and pattern recognition to discover knowledge from data)
- a field at the intersection of computer science and statistics
- the efficient discovery of previously unknown, valid, potentially useful, understandable patterns in large datasets
- the analysis of (often large) observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner

Data mining & Machine Learning

- Data mining & Machine Learning (knowledge discovery in databases):
  - Extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) information or patterns from data in large databases
- Alternative names and their “inside stories”:
  - Machine Learning & Data mining: a misnomer?
  - Knowledge discovery(mining) in databases (KDD), knowledge extraction, data/pattern analysis, data archeology, business intelligence, etc.
- Objective: Fit Data to a Model
  - Descriptive
  - Predictive
- Some ‘problems’:
  - Which technique to choose?
    - ARM/Classification/Clustering
    - Answer: Depends on what you want to do with data?
  - Search Strategy – Technique to search the data
    - Interface? Query Language?
    - Efficiency

Why Mine Data?

- Lots of data is being collected and warehoused
  - Web data, e-commerce
  - purchases at department stores
  - Bank/Credit Card transactions

- Computers have become cheaper and more powerful (from 32 bits to 64 bits)
- Competitive Pressure is Strong
  - Provide better, customized services for an *edge* (e.g. in Customer Relationship Management)

Why Mine Data? Scientific Viewpoint

- Data collected and stored at enormous speeds (GB/hour)
  - remote sensors on a satellite
  - telescopes scanning the skies
  - microarrays generating gene expression data
  - scientific simulations generating terabytes of data
- Traditional techniques infeasible for raw data
- Data mining may help scientists
  - in classifying and segmenting data
  - in Hypothesis Formation

- Social need :
  - Meteo prediction;
  - Soil erosion prediction;
  - Inundation, earthquake prediction

Examples: What is (not) Machine Learning?

- **What is not ML?**
  - Look up phone number in phone directory
  - Query a Web search engine for information about "Amazon"
- **What is ML?**
  - Certain names are more prevalent in certain US locations (O'Brien, O'Rurke, O'Reilly… in Boston area)
  - Group together similar documents returned by search engine according to their context (e.g. Amazon rainforest, Amazon.com,)

Data Mining: Classification Schemes

- Decisions in data mining
  - Kinds of databases to be mined
  - Kinds of knowledge to be discovered
  - Kinds of techniques utilized
  - Kinds of applications adapted
- Data mining tasks
  - Descriptive data mining
  - Predictive data mining

Decisions in DM & ML

- **Databases to be mined**
  - Relational, transactional, object-oriented, object-relational, active, spatial, time-series, text, multi-media, heterogeneous, legacy, WWW, etc.
- **Knowledge to be mined**
  - Characterization, discrimination, association, classification, clustering, trend, deviation and outlier analysis, etc.
  - Multiple/integrated functions and mining at multiple levels
- **Techniques utilized**
  - Database-oriented, data warehouse (OLAP), machine learning, statistics, visualization, neural network, etc.
- **Applications adapted**
  - Retail, telecommunication, banking, fraud analysis, DNA mining, stock market analysis, Web mining, Weblog analysis, etc.

DM & ML Tasks

- Prediction Tasks
  - Use some variables to predict unknown or future values of other variables
- Description Tasks
  - Find human-interpretable patterns that describe the data.

Common data mining tasks

- Classification [Predictive]
- Clustering [Descriptive]
- Association Rule Discovery [Descriptive]
- Sequential Pattern Discovery [Descriptive]
- Regression [Predictive]
- Deviation Detection [Predictive]

Example - Association Rule Discovery: Definition

- Given a set of records each of which contain some number of items from a given collection;
  - Produce dependency rules which will predict occurrence of an item based on occurrences of other items.

TID	*Items*
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

Rules Discovered:

{Milk} --> {Coke}

{Diaper, Milk} --> {Beer}

The Sad Truth About Diapers and Beer

- So, don’t be surprised if you find six-packs stacked next to diapers!

Association Rule Discovery: Application 1

- Marketing and Sales Promotion:
  - Let the rule discovered be

    {Bagels, … } --> {Potato Chips}
  - Potato Chips as consequent => Can be used to determine what should be done to boost its sales.
  - Bagels in the antecedent => Can be used to see which products would be affected if the store discontinues selling bagels.
  - Bagels in antecedent *and* Potato chips in consequent => Can be used to see what products should be sold with Bagels to promote sale of Potato chips!

Association Rule Discovery: Application 2

- Supermarket shelf management.
  - Goal: To identify items that are bought together by sufficiently many customers.
  - Approach: Process the point-of-sale data collected with barcode scanners to find dependencies among items.
  - A classic rule --
    - If a customer buys diaper and milk, then he is very likely to buy beer:

$$Diapers \rightarrow Beer,\ support = 20\%,\ confidence = 85\%$$

Free open-source softwares

- Carrot2: Text and search results clustering framework.
- Chemicalize.org: A chemical structure miner and web search engine.
- ELKI: A university research project with advanced cluster analysis and outlier detection methods written in the Java language.
- GATE: a natural language processing and language engineering tool.
- JHepWork: Java cross-platform data analysis framework developed at Argonne National Laboratory.
- KNIME: The Konstanz Information Miner, a user friendly and comprehensive data analytics framework.
- NLTK (Natural Language Toolkit): A suite of libraries and programs for symbolic and statistical natural language processing (NLP) for the Python language.
- Orange: A component-based data mining and machine learning software suite written in the Python language.
- R: A programming language and software environment for statistical computing, data mining, and graphics. It is part of the GNU project.
- RapidMiner: An environment for machine learning and data mining experiments.
- UIMA: The UIMA (Unstructured Information Management Architecture) is a component framework for analyzing unstructured content such as text, audio and video – originally developed by IBM.
- Weka: A suite of machine learning software applications written in the Java programming language.
- ML-Flex: A software package that enables users to integrate with third-party machine-learning packages written in any programming language, execute classification analyses in parallel across multiple computing nodes, and produce HTML reports of classification results.

Commercial softwares

- IBM InfoSphere Warehouse: Intelligent Miner - in-database data mining platform provided by IBM
- Microsoft Analysis Services: data mining software provided by Microsoft
- SAS: Enterprise Miner – data mining software provided by the SAS Institute.
- STATISTICA: Data Miner – data mining software provided by StatSoft.
- Oracle Data Mining: data mining software by Oracle.
- Clarabridge: enterprise class text analytics solution.
- LIONsolver: an integrated software application for data mining, business intelligence, and modeling that implements the Learning and Intelligent OptimizatioN (LION) approach
- MATLAB & Simulink

Steps of a ML & DM process

- Learning the application domain:
  - relevant prior knowledge and goals of application
- Creating a target data set: data selection
- **Data cleaning** and preprocessing: (may take 60% of effort!)
- **Data reduction and transformation**:
  - Find useful features, dimensionality/variable reduction, invariant representation.
- Choosing functions of data mining
  - summarization, classification, regression, association, clustering.
- Choosing the mining algorithm(s)
- **Data mining**: search for patterns of interest
- **Pattern evaluation and knowledge presentation**
  - visualization, transformation, removing redundant patterns, etc.
- Use of discovered knowledge

Data Mining and ML Issues

- Human Interaction
- Overfitting
- Outliers
- Interpretation
- Visualization
- Large Datasets
- High Dimensionality

19

Challenges…

- Different types of the Data (text, images, video…)
- Missing Data
- Irrelevant Data (objects selection)
- Noisy Data (irelevant features)
- Changing Data (data flows)

20

Social Implications

- Privacy preserving
- Profiling peoples
- Unauthorized use

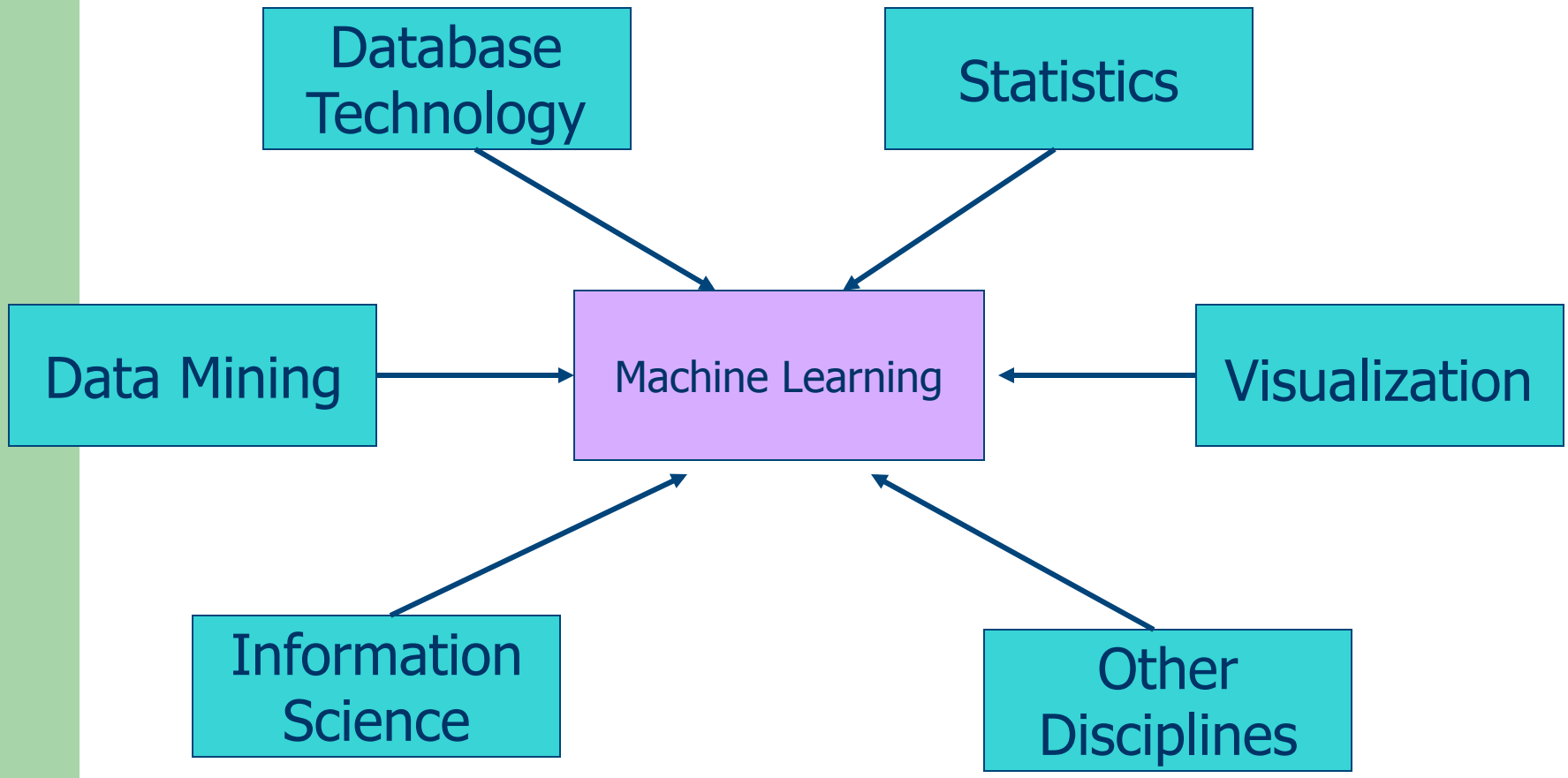
Some solutions as :

Collaborative unsupervised learning;

Transfert Learning, …

21

Data Mining: Confluence of Multiple Disciplines



Query Examples

- Database
  - Find all credit applicants with last name of Smith.
  - Identify customers who have purchased more than $10,000 in the last month.
  - Find all customers who have purchased milk
- Data Mining and Machine Learning
  - Find all credit applicants who are poor credit risks. (classification)
  - Identify customers with similar buying habits. (Clustering)
  - Find all items which are frequently purchased with milk. (association rules)

23

DATA

A data ?

Data is a basic description of a phenomenon

A phenomenon can be described by one or more criteria, these criteria are called variables:

A single criterion: univariate data
Several criteria: Multivariate Data

A phenomenon is thus described by a set of univariate or multivariate data.

More data is generated:

- Bank, telecom, other business transactions ...
- Scientific data: astronomy, biology, etc
- Web, text, and e-commerce

Samples and variables

- Population

  Group or set of individuals that are analyzed.

- Variables

  Set of characteristics of a population.

Classical data matrix

- For *n* objects and *p* variables, we difine the data table :

X which is rectangular matrix containing *n* lines and *p* columns

$$X = (x^1, \ldots, x^p) = \begin{bmatrix} x_1^1 & x_1^2 & & \ldots & & x_1^p \\ x_2^1 & x_2^2 & & & & \\ & & \ddots & & & \\ \vdots & & & x_i^j & & \vdots \\ & & & & \ddots & \\ x_n^1 & & & \ldots & & x_n^p \end{bmatrix}$$

Example of a data table (matrix)

- Objects : x1,x2,…,x10.
- Variables :Y1,Y2,…,Y4.

Matrix of data

	Y1	Y2	Y3	Y4
x1	10	6	45	41
x2	13	8	35	78
x3	15	23	87	64
x4	19	56	96	43
x5	40	47	56	52
x6	45	34	43	42
x7	39	26	12	13
x8	40	12	14	16
x9	11	13	14	15
x10	39	26	12	13

Other definitions :
- Individuals: observations, objets, instances, transactions.
- Variables: attributs, dimensions, description, component.

Data types

- Quantitative variables
- Qualitative variables
- Others : text data

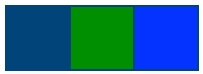
Quantitative variables

- Quantitative variables
  - Continuous (ex: the size, the weight of a person, the time of completion of a task, the volume of an object, the speed of a car)
  - Discrete (ex: counting: the number of persons in a room, the number of items from a list)

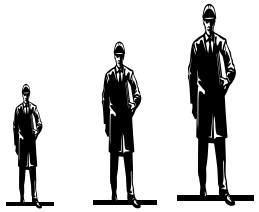
Qualitative variables

- Qualitative variables (categorical)
  Binary data: it can take two states (true or false, 0 or 1, yes or no;).
  Ex: Gender, having a credit or not ...

  Unordered categorical data (nominal):
  Ex: eye color

Colour : 

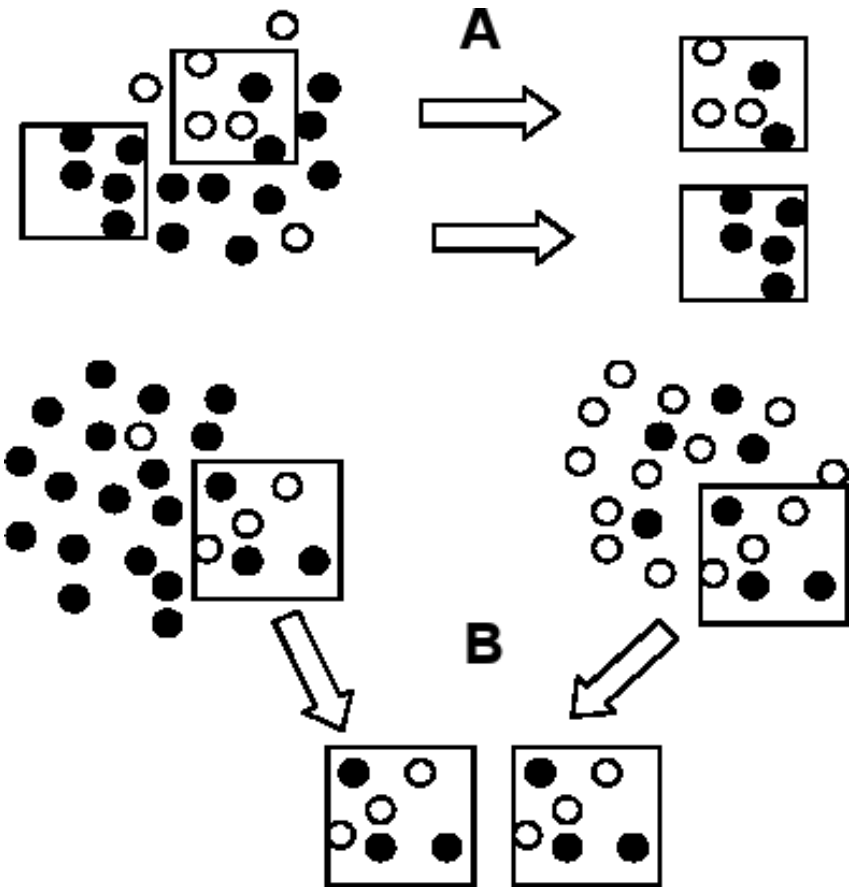
Ordered categorical data: data from a survey (1: very satisfied, 2: satisfied, ..)
Ex: low, medium, high, small, medium, large

Height: 

Goals

- Answer a question or solve a problem
- Make data – intelligible
- Retrieve and select information
- Determine the validity of this information:
  Problem of sampling fluctuations
  Problem of generalization

Objectifs



Why to use the informatics?

- Power (speed) of calculation and processing of large databases
- Visualization (dimension reduction)
- Objectivity!

Quantitative univariate data

Univariate statistics of variables

- First step in the data mining
  Detect anomalies in their distributions (eg extreme or missing values)

  How to discretize continuous variables (if applicable)

  To undesrtand some important information, which can be useful in other analysis (eg age, average income of the population)

  Make a summary of the contents using a graph

Position criteria

A position criterion is a value which represents the best
The corresponding set of data values

Statistics of central tendency

A measure of central tendency is a single value that attempts to describe a set of data by identifying the central position within that set of data. As such, measures of central tendency are sometimes called measures of central location. They are also classed as summary statistics.

It can be used to answer to questions as:

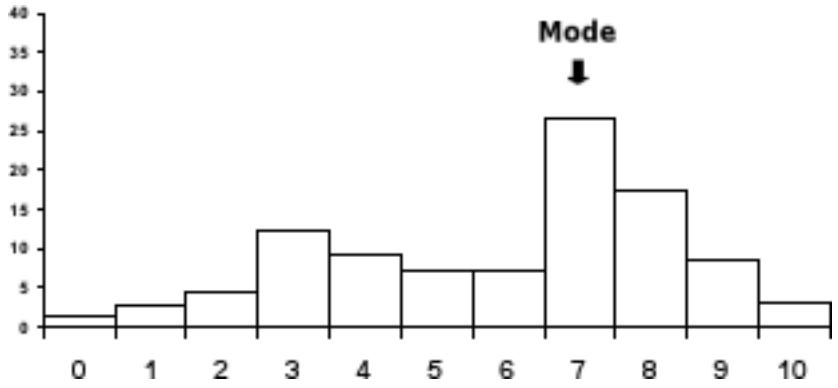
What is the "typical" salary of a footbal player?
How many children has a "typical" French family?
What is the "typical“ note of students for the exam?

The term **central tendency** refers to the "middle" value or perhaps a typical value of the data, and is measured using the **mean**, **median**, or **mode**. Each of these measures is calculated differently, and the one that is best to use depends upon the situation.

Statistics of central tendency

- The mode
- The median
- The mean

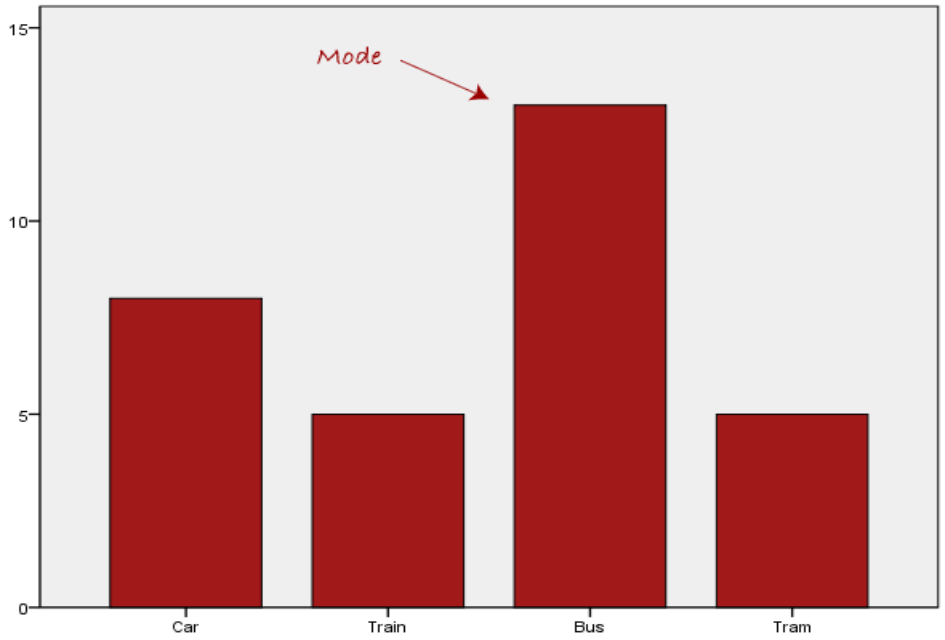
The mode



The mode is the most frequently occurring value in the data set.
(not necessarily unique).

On a histogram it represents the highest bar in a bar chart or histogram.
Therefore, sometimes consider the mode as being the most popular option.

Normally, the mode is used for categorical data where we wish to know which is the most common category :

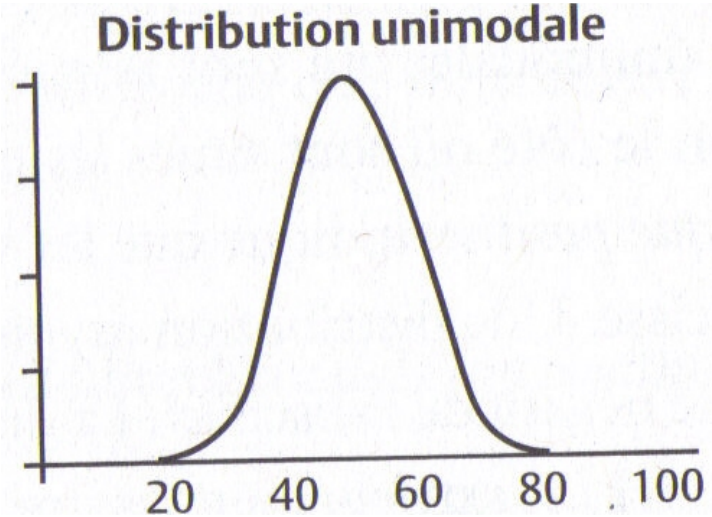


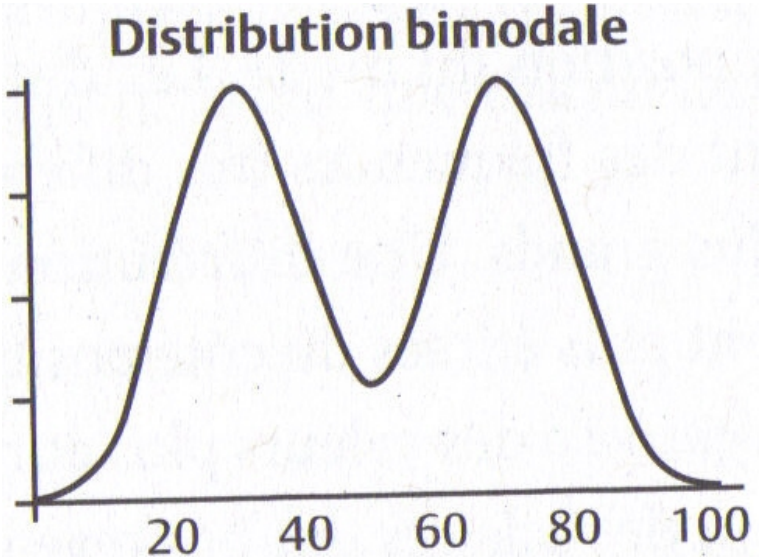
We can see above that the most common form of transport, in this particular data set, is the bus. However, one of the problems with the mode is that it is not unique, so it leaves us with problems when we have two or more values that share the highest frequency.

The mode

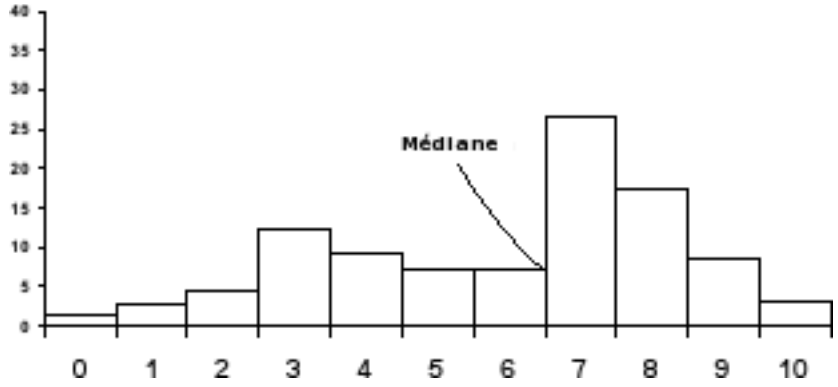
Depending of the distribution of the mode, there are two types of the distributions:

- Unimodal
- Bimodal
- Multimodal





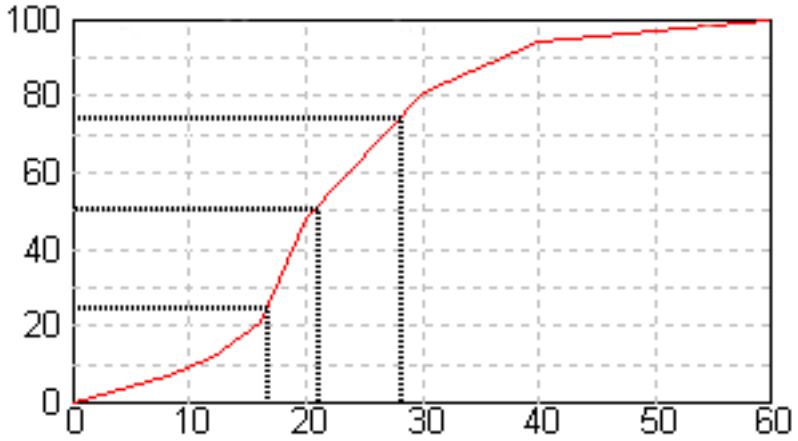
The median



$$\text{median} = \begin{cases} x_{(\frac{n+1}{2})} & \text{si } n \text{ est impaire,} \\ \frac{1}{2}\left(x_{(n/2)} + x_{(1+n/2)}\right) & \text{si } n \text{ est paire.} \end{cases}$$

The median is the measure which allows to define tthe value which cut the distribution in two parts, each of them having the same number of observations.

The médian



The median is the value which cut the population in two populations of the same size.

The median is determined by sorting the data set from lowest to highest values and taking the data point in the middle of the sequence.

- suppose we have the data below:

65	55	89	56	35	14	56	55	87	45	92

- We first need to rearrange that data into order of magnitude (smallest first):

14	35	45	55	55	**56**	56	65	87	89	92

- Our median mark is the middle mark - in this case 56. It is the middle mark because there are 5 scores before it and 5 scores after it.

- This works fine when you have an odd number of scores but what happens when you have an even number of scores? What if you had only 10 scores? Well, you simply have to take the middle two scores and average the result. So, if we look at the example below:

65	55	89	56	35	14	56	55	87	45

- We again rearrange that data into order of magnitude (smallest first):

14	35	45	55	**55**	**56**	56	65	87	89	92

- Only now we have to take the 5th and 6th score in our data set and average them to get a median of 55.5.

The median

So, the median use only the relative position of the observations.

Sample A : 13, 15, 17,19, 23

Sample B : 13, 15, 17,19, 400

Convenients:

- The median is not affected by the outliers
- The median is useful then we have missing data

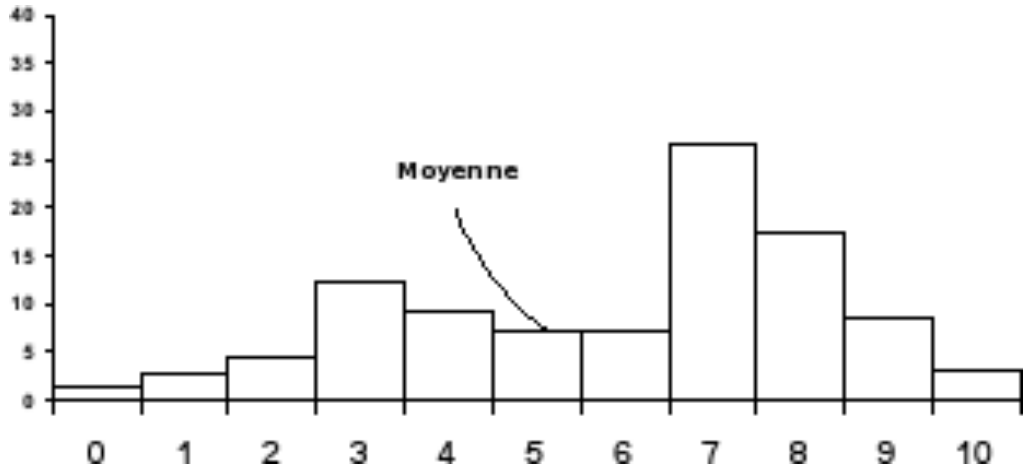
The median

The median often is used when there are a few extreme values that could greatly influence the mean and distort what might be considered typical.

This often is the case with home prices and with income data for a group of people, which often is very skewed. For such data, the median often is reported instead of the mean.

For example, in a group of people, if the salary of one person is 10 times the mean, the mean salary of the group will be higher because of the unusually large salary. In this case, the median may better represent the typical salary level of the group.

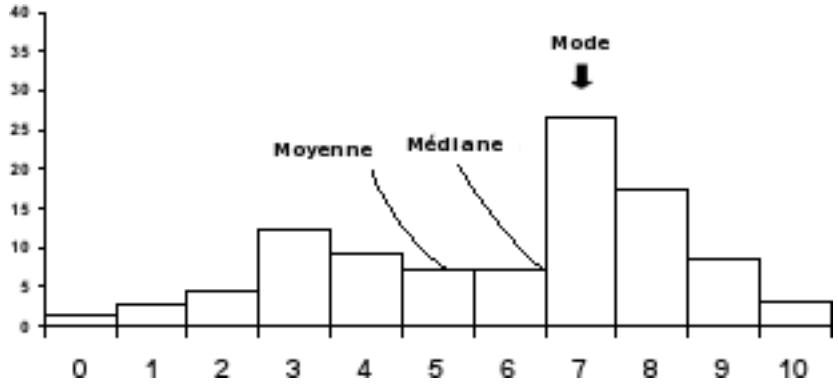
The mean



The mean is the value which could have each of the data sample if they was all identical without changing the total (global) value.
The mean is equal to the sum of all the values in the data set divided by the number of values in the data set.

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

Notes



- All these 3 cretiria gives different information
- The mode use only a set of the distribution values (only the most frequently value is considered)
- The median count only the position of the observations
- The mean is sensitive (depends) on the extram values (outliers)

Dispersion criteria

Definition

A dispersion criteria is a value which represents the homogeneity of the values of a data.

Statistical dispersion (also called statistical variability or variation) is variability or spread in a variable or a probability distribution ;

A measure of statistical dispersion is a nonnegative real number that is zero if all the data are the same and increases as the data become more diverse :

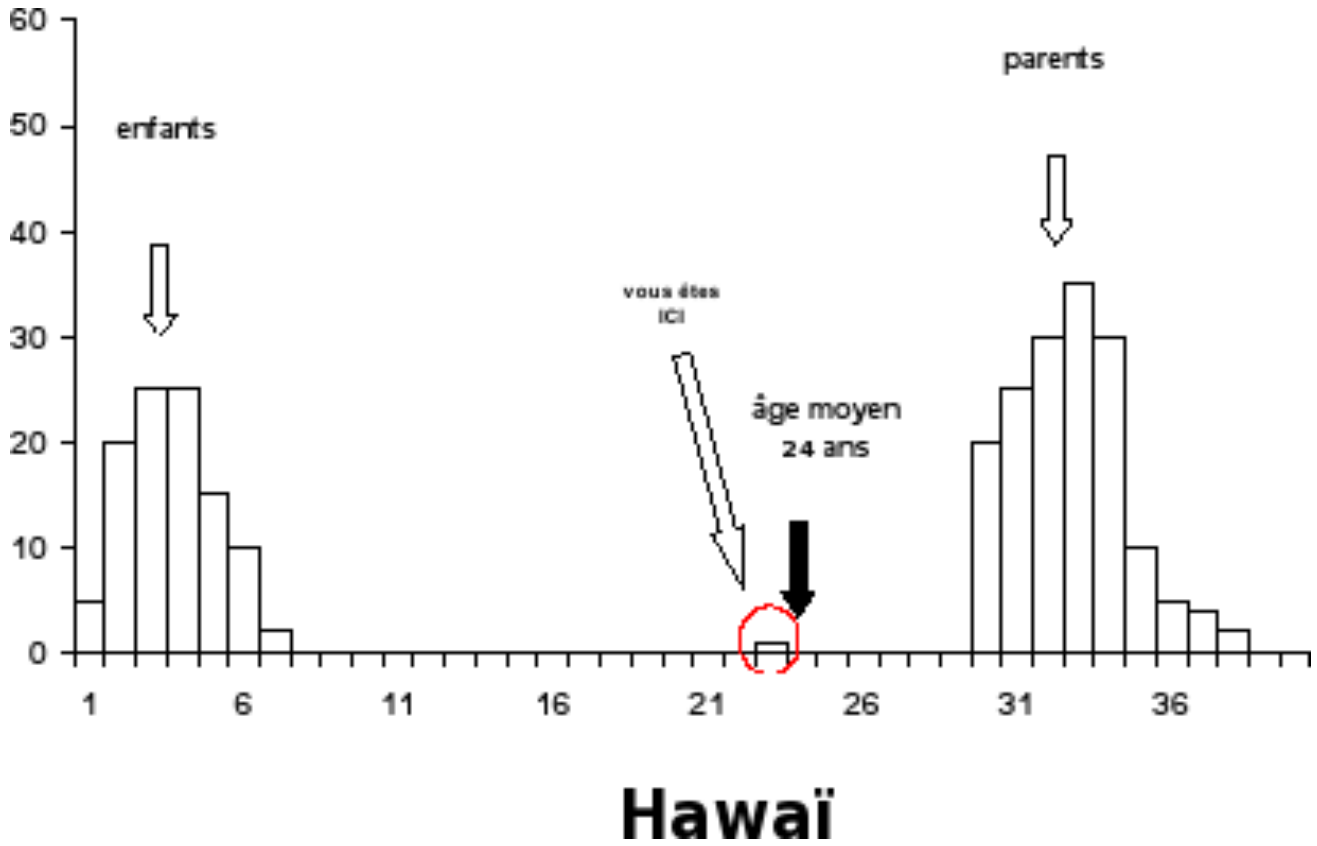
- Standard deviation
- Interquartile range or Interdecile range
- Range
- Mean difference
- Median absolute deviation
- Average absolute deviation (or simply called average deviation)
- Distance standard deviation

Interest

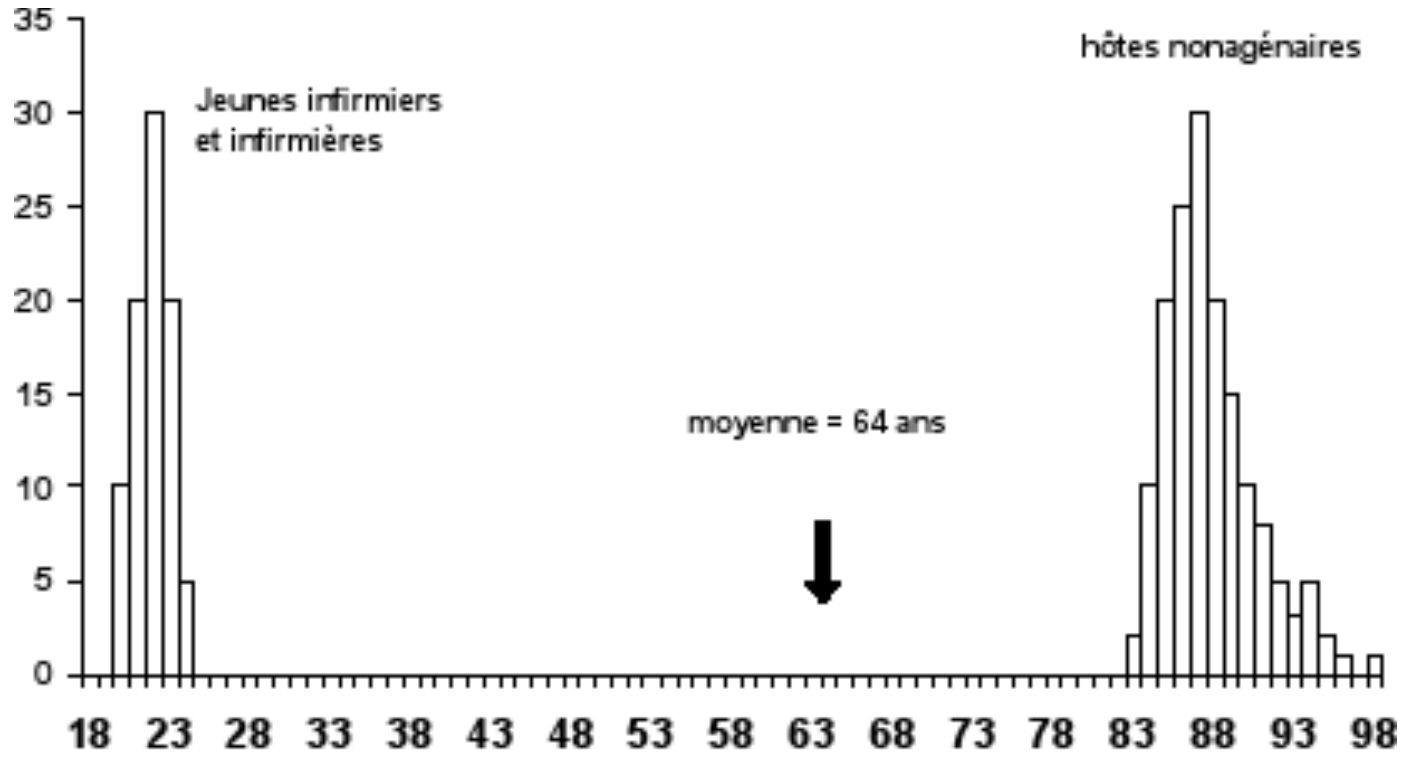
For your holidays you have the choice between:

- A peaceful family pension in Novosibirsk (Siberia): mean age 64 years
- A paradise island a few miles off Hawaii: mean age 24 years

Interest


60
50
40
30
20
10
0
1
6
11
16
21
26
31
36
enfants
parents
vous êtes ICI
âge moyen 24 ans
Hawaï

Interest



Sibérie

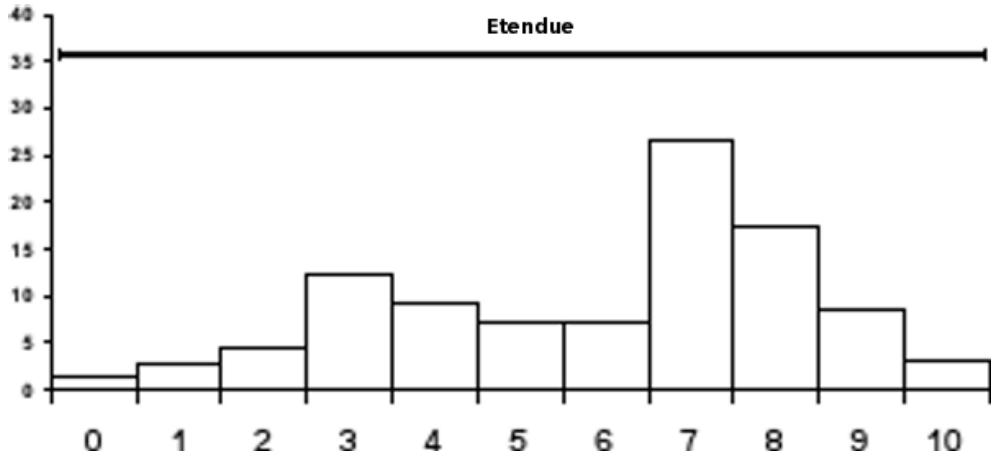
Dispersion measure

- Range
- Standard deviation
- Variance

Range

- The range is the difference between the largest and the smallest observed value

$$Range = X_{\max} - X_{\min}$$



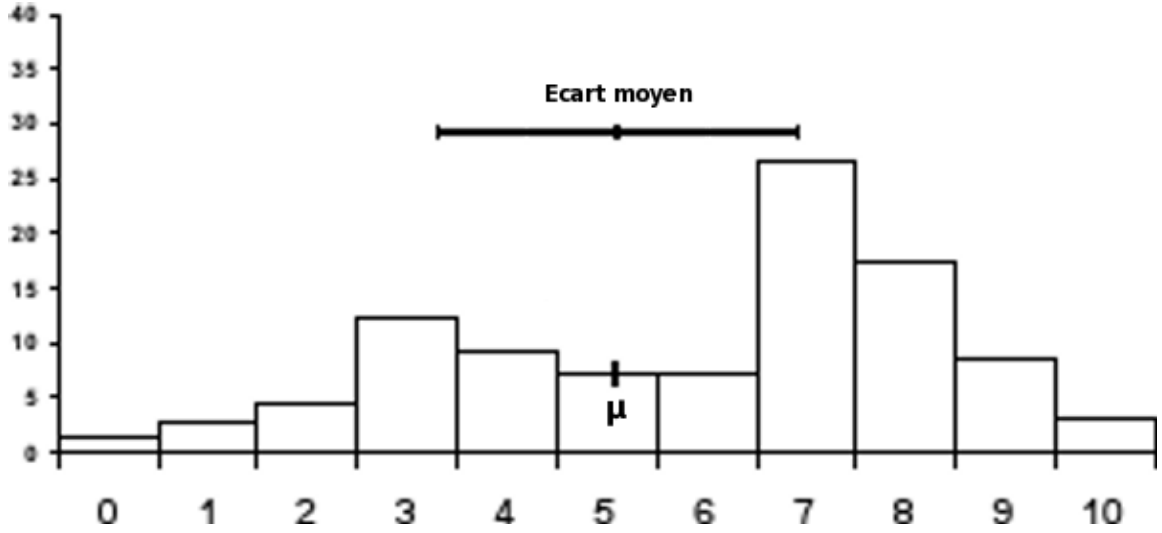
Variance

The variance is the mean of the squared deviation of that variable from its expected value or mean

$$s^2 = \text{var}(x) = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2$$

the variance of a variable has units that are the square of the units of the variable itself

Standard deviation



The standard deviation is the mean of the de labsolute value of the deviation

$$\text{écart moyen} = \frac{1}{n}\sum_{i=1}^{n}|x_i - \mu|$$

Standard deviation

The standard deviation describe the « typical » difference between the observations and the mean (average)

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

Example :
Sample A=(79,80,81)
Sample B=(60,80,100)
mean(A)=mean(B)=80
Standard-Deviation(A)≠ Standard-Deviation (B)
Standard-Deviation(A)=1
Standard-Deviation(B)=20

Remarks

- A series of indicators that gives a partial view of the data: effective, mean, median, variance, standard deviation, minimum, maximum, range, first quartile, 3rd quartile, ...
- These indicators measure the central tendency and the dispersion. We usually use the mean, variance and standard deviation.
- These criteria are sensitive to extreme values
- Mean and standard deviation can be generalized from a sample

Describe a distribution

To describe a distribution, we analyse:

- The mean
- The standard deviation

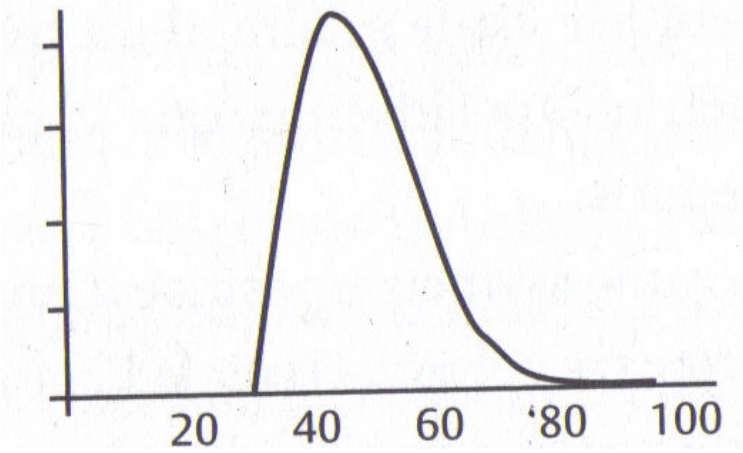
We have also to analyse the distribution for:

- Its degree of **Skewness**
- Its degree of **Kurtosis**

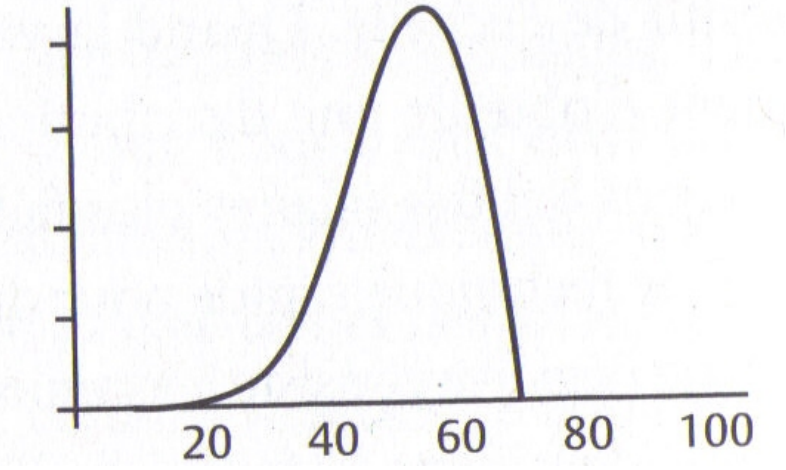
Skewness

$$Sk_x = \frac{\sum_i (x_i - \bar{x})^3}{s_X{}^3} \times \frac{N}{(N-1)(N-2)}$$

Distribution asymétrique positive



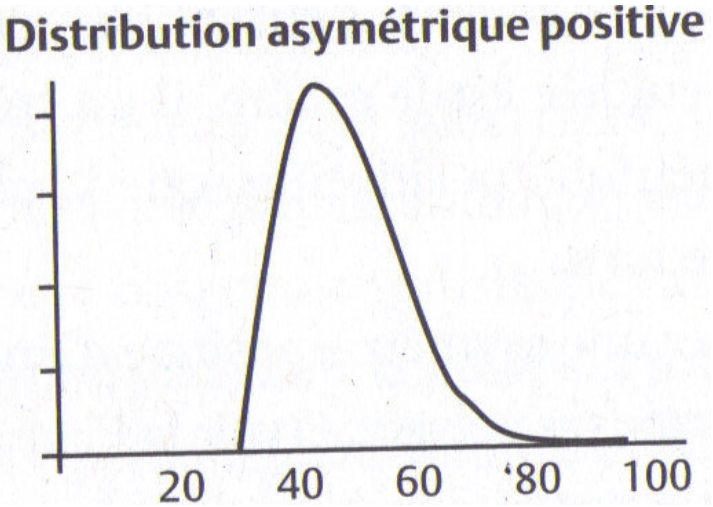
Distribution asymétrique négative

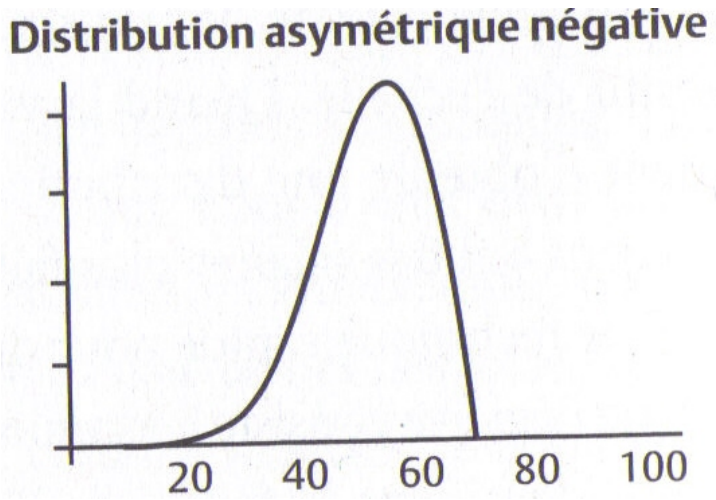


The *skewness*

If :

- $Sk_x = 0$, we have a perfectly symmetrical distribution (values are spread uniformly and also to higher values and lower values of the variable)
- $Sk_x > 0$, - positively skewed (distribution spreads more towards higher values of the variable)
- $Sk_x < 0$, negatively skewed (distribution spreads more towards lower values of the variable)





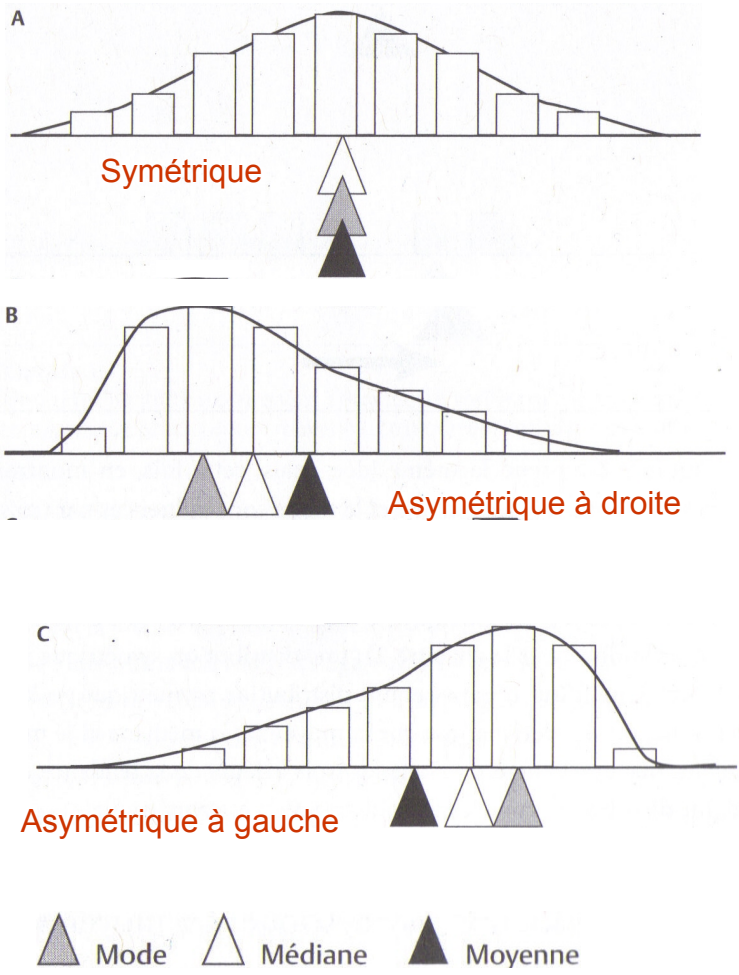
The *skewness*

Another method to determine the skewness of a distribution is:

-Symetric when
mode=median=mean

- Skewed to the right when
mode<median<mean

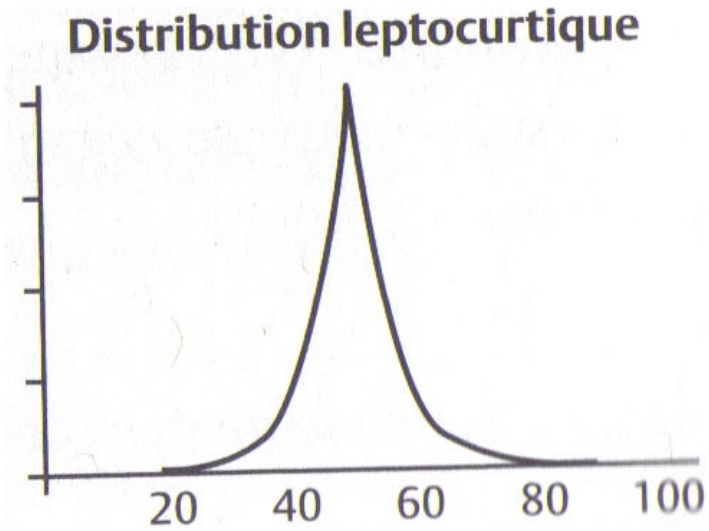
-Skewed to the left when
mode>median>mean

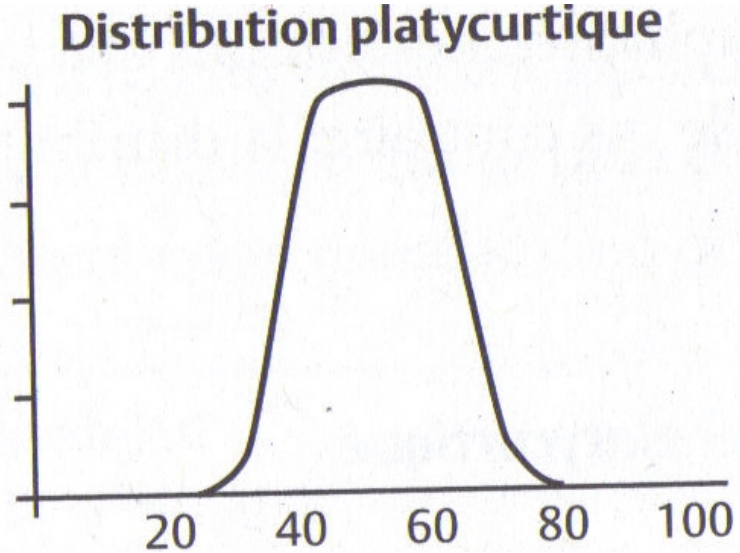


The Kurtosis measure

$$Ku_x = \frac{\sum_i (x_i - \bar{x})^4}{s_x{}^4} \times \frac{N(N+1)}{(N-1)(N-2)(N-3)} - 3\frac{N-1}{(N-2)(N-3)}$$

Distributions with negative or positive excess kurtosis are called **platykurtic distributions** or **leptokurtic distributions** respectively





Quartile

- Distribution function of a random variable :
  - $F(x)=P(X \leq x)$
- Quartiles q_1, q_2, q_3, are defined as:

  - F(q1) = 0,25 is the **first quartile** (designated Q1) = lower quartile = splits lowest 25% of data = 25th percentile

  - $F(q_2) = 0,5$ is the **second quartile** (designated Q2) = median = cuts data set in half = *50th percentile*

  - $F(q_3) = 0,75$ is the **third quartile** (designated Q3) = upper quartile = splits highest 25% of data, or lowest 75% = *75th percentile*

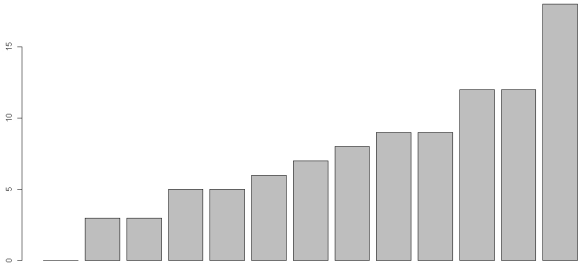
Interquartile range: $IQR=q_3 - q_1$

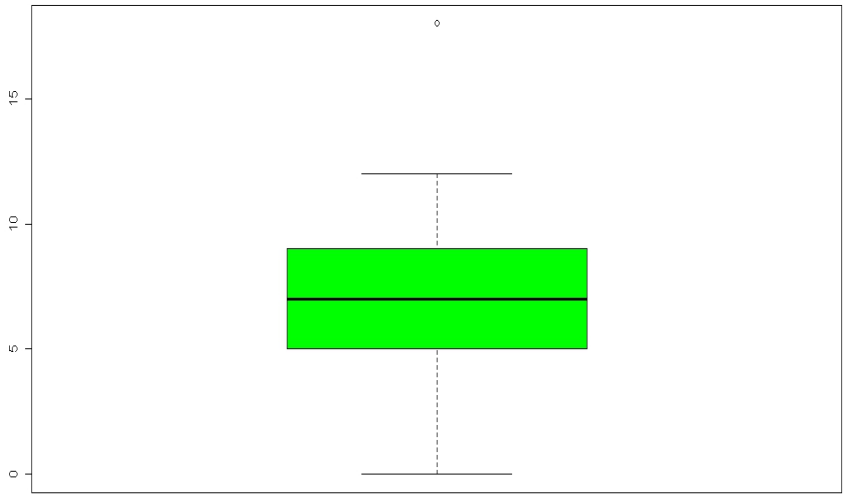
- The difference between the upper and lower quartiles containing 50% of data.

Boxplot

a<-c(0,3,3,5,5,6,7,8,9,9,12,12.18)

barplot(a)

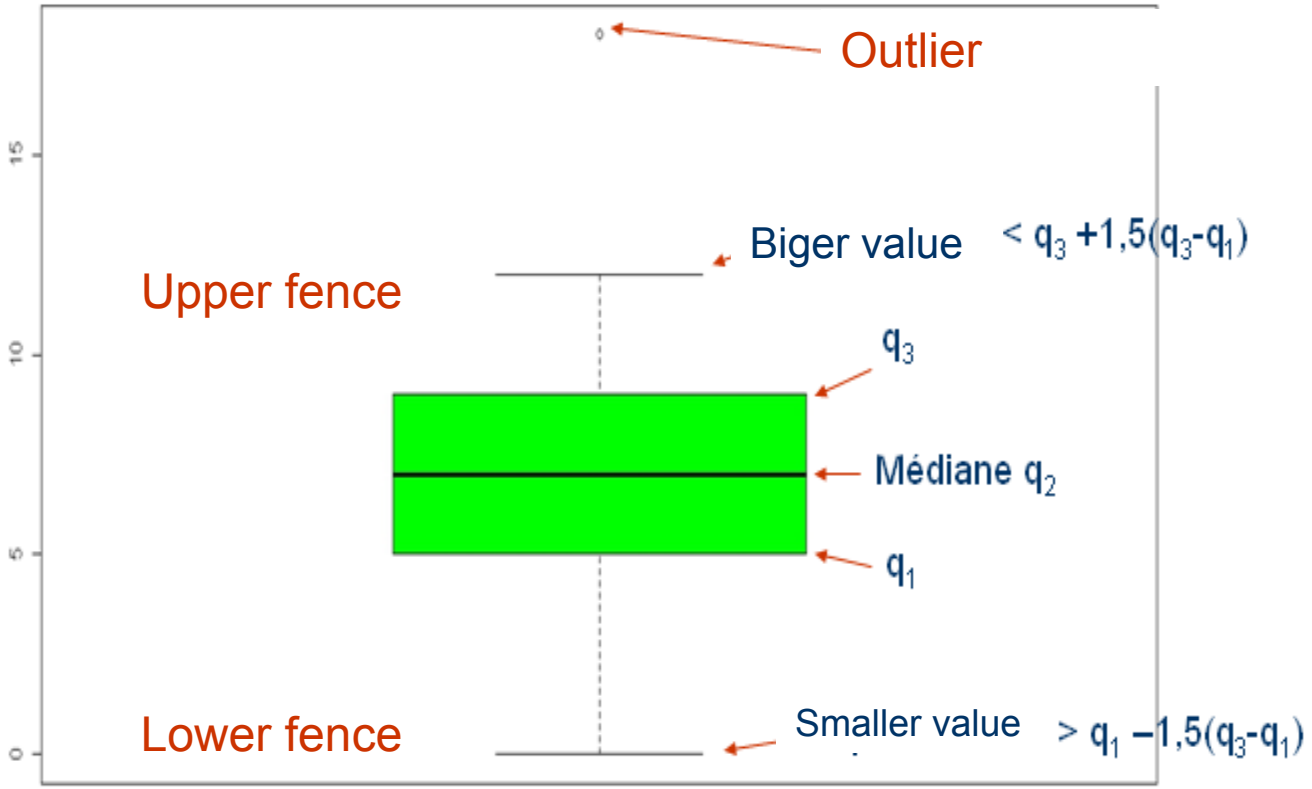




mean=7,46
q_1= 5
median = 7
q_3 = 9
Lower fence= 0
Upper fence= 12

Lower fence > Q1 – 1.5(IQR)
Upper fence < Q3 + 1.5(IQR)
Smallest observation (sample minimum)
Largest observation (sample maximum)

Boxplot



Mean=7,46
q_1=5
Median=7
q_3=9
Lower fence= 0
Upper fence = 12

a<-c(0,3,3,5,5,6,7,8,9,9,12,12,18)

Example : Boxplot

- Create the boxplot of the following data:

52, 18, 26, 40, 8, 50, 63, 42, 21, 7, 44, 14

Example : Boxplot

- Firstly, we order the sample in the ascendend order. Then, find the median.

  7, 8, 14, 18, 21, 26, 40, 42, 44, 50, 52, 63.

- Median = 6,5 value
  = (6st + 7th observations) ÷ 2
  = (26 + 40) ÷ 2
  = **33**

- There arte 6 numbers before the median : 7, 8, 14, 18, 21, 26.

- Q1 = the median of these six variables
  = 3,5 value
  = (3rd + 4th observations) ÷ 2
  = (14 + 18) ÷ 2
  = **16**
- There six numbers after the median : 40, 42, 44, 50, 52, 63.

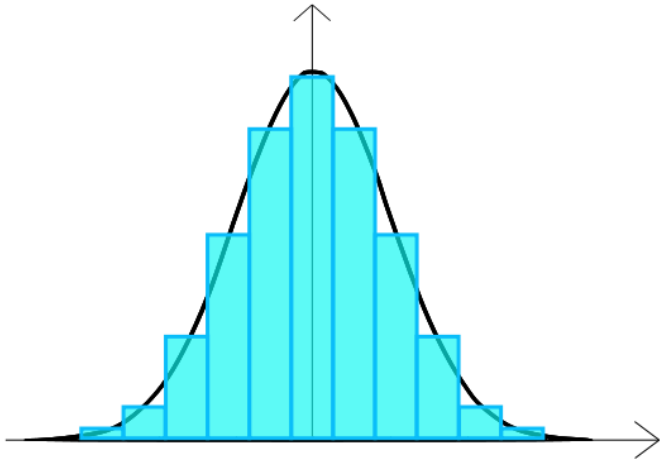
- Q3 = the median of these six elements is
  = (6 + 1) ÷ 2= 3,5 value
  = (3rd + 4th observations) ÷ 2 = (44+50) ÷ 2
  = **47**

Mini = 7
Q1 =16
Median = 33
Q3 = 47
Max = 63

Normal distribution

Normal (or **Gaussian**) **distribution** is a continuous probability distribution that has a bell-shaped probability density function, known as the **Gaussian function**

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$



Recall : The parameter μ - *mean* (location of the peak)
σ is the standard deviation
σ^2 is the variance.

The distribution with $\mu = 0$ and $\sigma^2 = 1$ is called the **standard normal distribution** or the **unit normal distribution**

Normal distribution

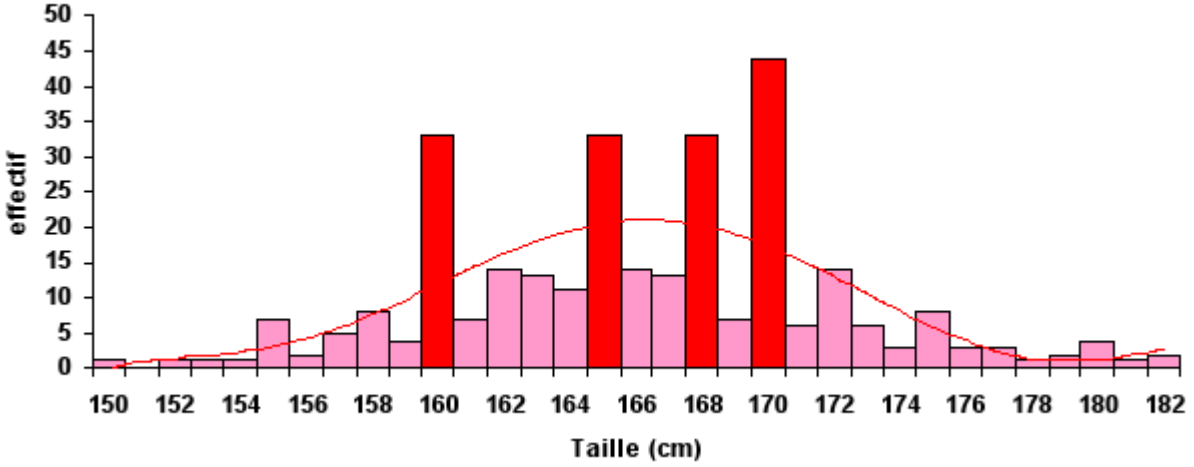
The normal distribution has the following proprieties:

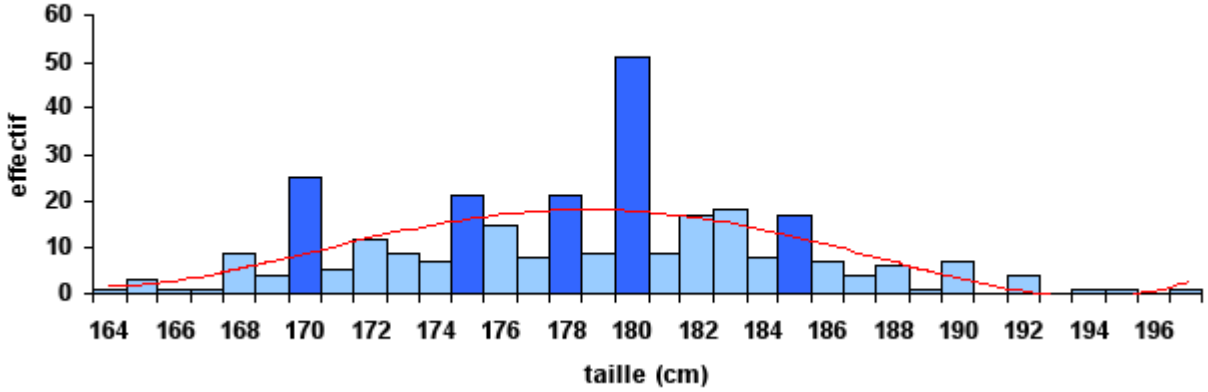
- 68% of the population is situated in the interval: $[\bar{x} - s; \bar{x} + s]$
- 95% of the population is in the range: $[\bar{x} - 2s; \bar{x} + 2s]$
- 99% of the population is in the range: $[\bar{x} - 3s; \bar{x} + 3s]$

If these properties are not satisfied, the distribution is not Gaussian.

Normal distribution

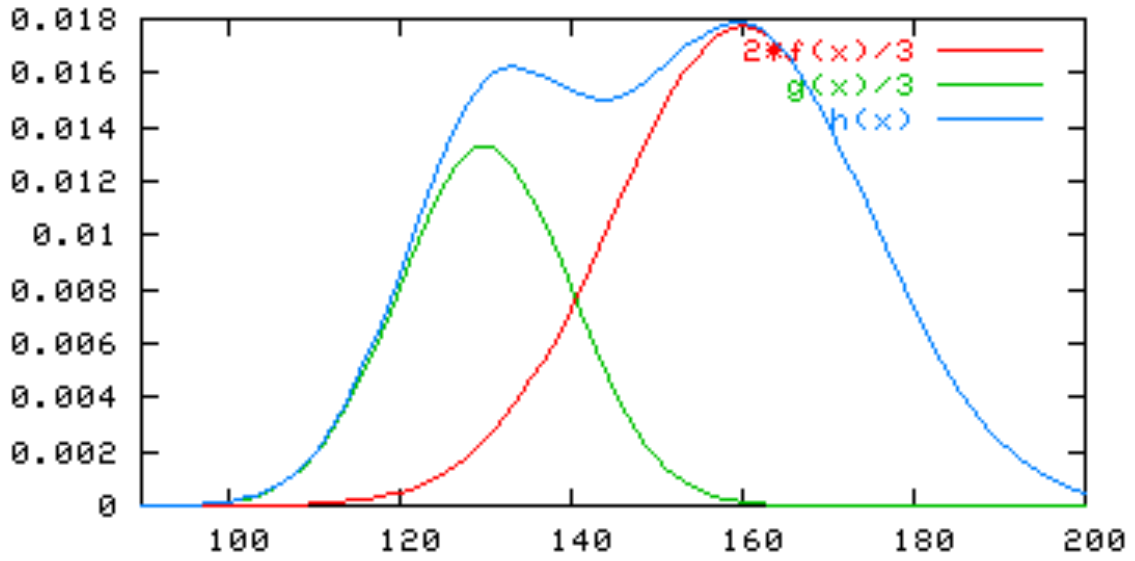
Many natural variables follow the Normal Distribution





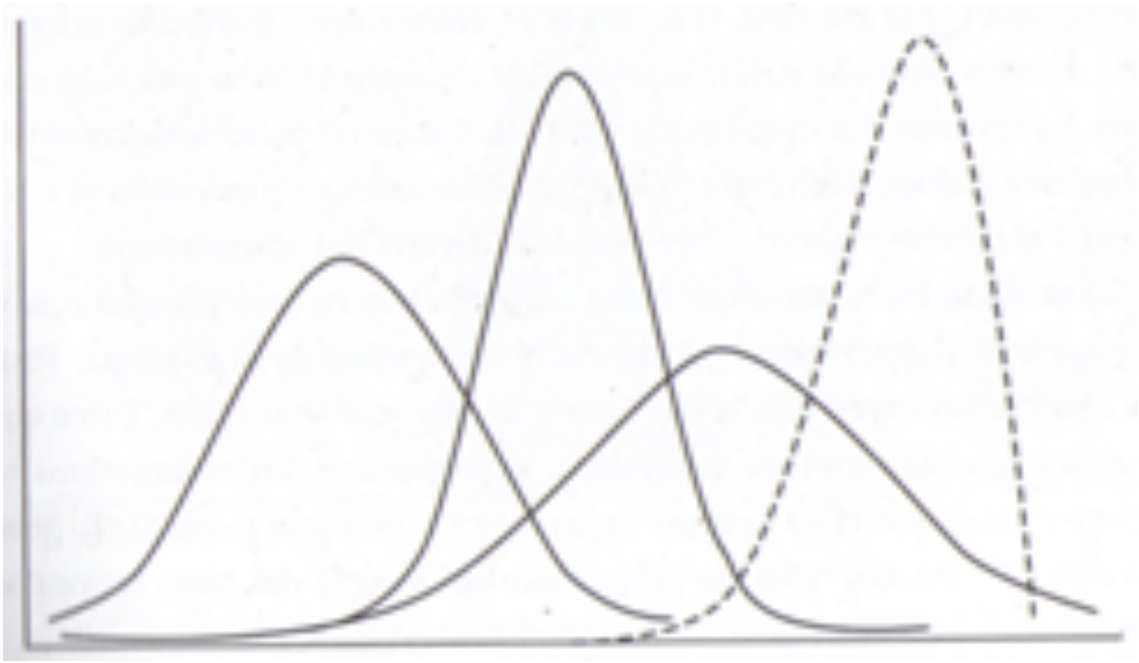
Gaussian distribution

The mixture of two Gaussian populations is not a Gaussian population



The distribution is bimodal

Example



Missing values and outliers

The detection of missing values and outliers is a step to achieve for any type models .

It is important to question why the data is missing, this can help with finding a solution to the problem.

If the values are missing at random there is still information about each variable in each unit but if the values are missing systematically the problem is more severe because the sample cannot be representative of the population.

For example: a research is done about the relation between IQ and income. If participants with an over average IQ do not answer the question ‘What is your salary?’ the results of the research may show that there is no association between IQ and salary, while in fact there is a relationship. Because of these problems, methodologists routinely advise researchers to design research so as to minimize the incidence of missing values (Ader, H.J., Mellenbergh, G.J. 2008).

Missing data

- Ignore the tuple: usually done when class label is missing (assuming the tasks in classification)—not effective when the percentage of missing values per attribute varies considerably.
- Fill in the missing value manually: tedious + infeasible?
- Use a global constant to fill in the missing value: e.g., “unknown”, a new class?!
- Use the attribute mean to fill in the missing value
- Use the attribute mean for all samples belonging to the same class to fill in the missing value: smarter
- Use the most probable value to fill in the missing value: inference-based such as Bayesian formula or decision tree

Outliers

An outlier is an incorrect value corresponding to a bad measure, a miscalculation, a mistake or misrepresentation.

Examples:

- Inconsistent Dates: February 29, a non-leap year, the subscription dates prior to the date of birth of the customer

- Codes "sex" taking more than two different values
- Phone numbers that are not corresponding to phone numbers

Outliers

Several solutions exist to deal with outliers:

Delete the concerned observations, if their number is not too high and sufficiently random distribution

Keep the observation and the variable, tolerating a small margin of error in the model results

Keep the observation and the variable, but replace the outlier by another value that is closest to its true value (mean...)

Keep the observations but don’t use the variables for mining the data.

Data standartization

- Centering

$$y=x-\bar{x}$$

- Centering and reduction (standartization)

$$y=\frac{x-\bar{x}}{s}$$

- Other normalization

Min-Max

$$y=\frac{x-\min(x)}{\max(x)-\min(x)}$$

Logarithmic

$$y=\log(x)$$

Bibliography

- Gregory Piatetsky-Shapiro, KDnuggets
- Ad Feelders, Advanced Data Mining 2011
- Srinivasan Parthasarathy, Introduction to Data Mining
- Min Song, Data Mining
- Fall 2004, CIS, Temple University, CIS527: Data Warehousing, Filtering, and Mining
- Jiawei Han (http://www-sal.cs.uiuc.edu/~hanj/DM_Book.html)
- Vipin Kumar (http://www-users.cs.umn.edu/~kumar/csci5980/index.html)
- Stéphane Tufféry «Data mining et statistique décisionnelle»
- Robert R. Haccoun, Denis Cousineau »Statistiques : Concepts et applications »
- Fenelon, J.P., (1981). « Qu'est-ce que l'analyse des données ? », Lefonen.
- Benzécri, J.-P., (1982). »Histoire et préhistoire de l'analyse des données », Dunod.