Data Mining - Lecture 1

Introduction to data analysis, Univariate statistics of variables, Random variables

Dr. Issam Falih

issam.falih@uca.fr

Course Components

Organization

- 10.5h Lectures ($3 \times 3.5h$)
- 14h Laboratory ($4 \times 3.5h$): R programming

Evaluation

- Exam: 70%
- Laboratory + Project: 30%

Course Outline

- Univariate Statistics
- Bivariate and Multivariate Statistics
- Projection: Principal Component Analysis
- Linear Regression, Logistic Regression
- Unsupervised Learning: K-Means, HCA
- Time series predictions
- Data Visualization

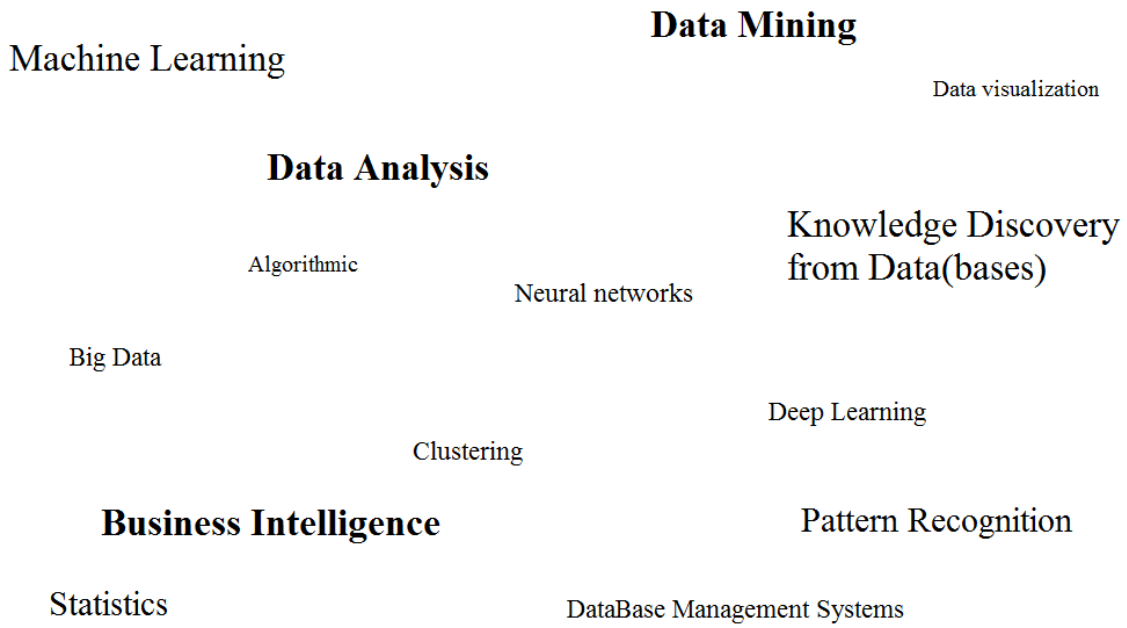
Outline

Outline

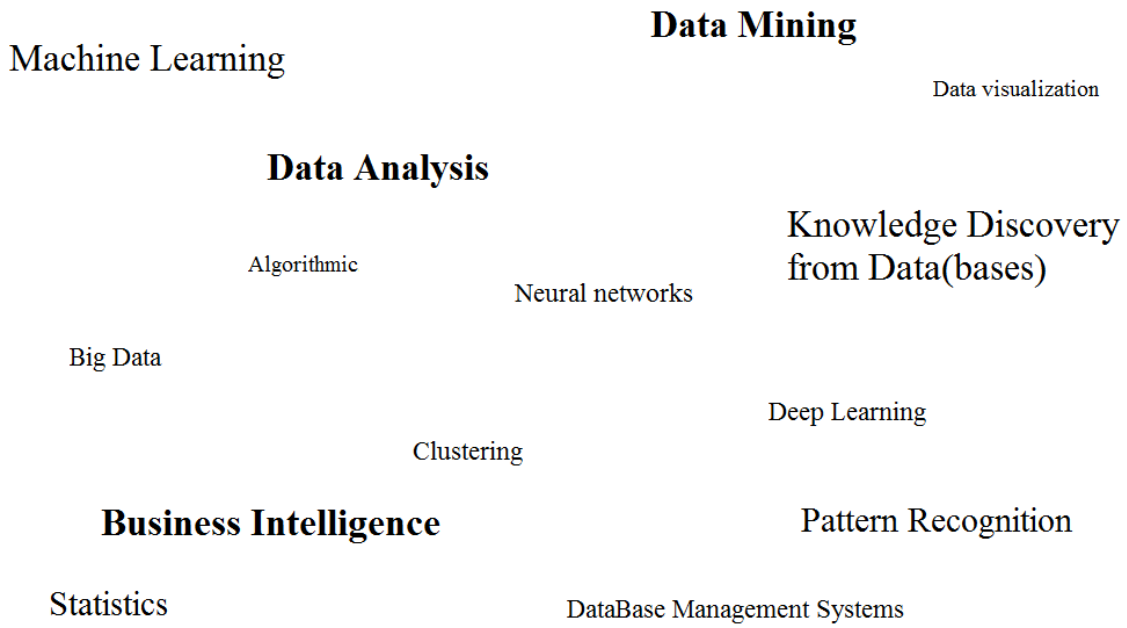
What is data analysis ?

- Data analysis is a field at the intersection between statistics and computer science.
- It aims at efficiently discovering previously unknown, valid, potentially useful and understandable knowledge from data.
- It can help with the analysis of (often large) data sets: finding internal relationships or structures, extracting clusters, summarize or visualize in an understandable way.

Data analysis and related terms (1/5)



Data analysis and related terms (1/5)



Open questions

- How do we define these terms exactly ?
- Do these fields overlap ?
- What are the differences between them ?

Data analysis and related terms (2/5)

Data Mining, Data Analysis and Business Intelligence

- **Data Mining** : The process of extracting knowledge from data
- **Data Analysis** : The process of extracting knowledge from data and analyzing it
- **Business Intelligence** : Data analysis followed by decision making for corporate purposes

Data analysis and related terms (2/5)

Data Mining, Data Analysis and Business Intelligence

- **Data Mining** : The process of extracting knowledge from data
- **Data Analysis** : The process of extracting knowledge from data and analyzing it
- **Business Intelligence** : Data analysis followed by decision making for corporate purposes

Knowledge Discovery from Data(bases) : KDD

- **Knowledge Discovery from Data** : The overall process of discovering knowledge from data and to make sense of it.
- KDD and Data analysis are somewhat synonymous.

Data analysis and related terms (3/5)

Machine Learning and Statistics

- **Machine Learning** : A Science field concerned with the development of algorithms for computers to process, analyze and learn from data.
- **Statistics** : A subfield of mathematics containing various tools for trend analysis, probabilistic predictions and solid theoretical bases to build various kind of models.

Data analysis and related terms (3/5)

Machine Learning and Statistics

- **Machine Learning** : A Science field concerned with the development of algorithms for computers to process, analyze and learn from data.
- **Statistics** : A subfield of mathematics containing various tools for trend analysis, probabilistic predictions and solid theoretical bases to build various kind of models.

- Remark: Machine Learning and statistics are Science fields, while Knowledge Discovery from Data, Data Mining, Data Analysis and Business intelligence are processes.

Data analysis and related terms (4/5)

Machine Learning and statistics as tools for Data Analysis

Data Mining and **Data Analysis** use various tools from **statistics** and algorithms from **Machine Learning** to extract and analyze information from data.

Data analysis and related terms (4/5)

Machine Learning and statistics as tools for Data Analysis

Data Mining and **Data Analysis** use various tools from **statistics** and algorithms from **Machine Learning** to extract and analyze information from data.

Pattern Recognition

- **Pattern Recognition** is an ambiguous term that may be applied to either Machine Learning or Data Mining applied to images and video data.
- **Pattern Recognition** is also sometimes used to refer to Data Analysis applied to time data.

Data analysis and related terms (5/5)

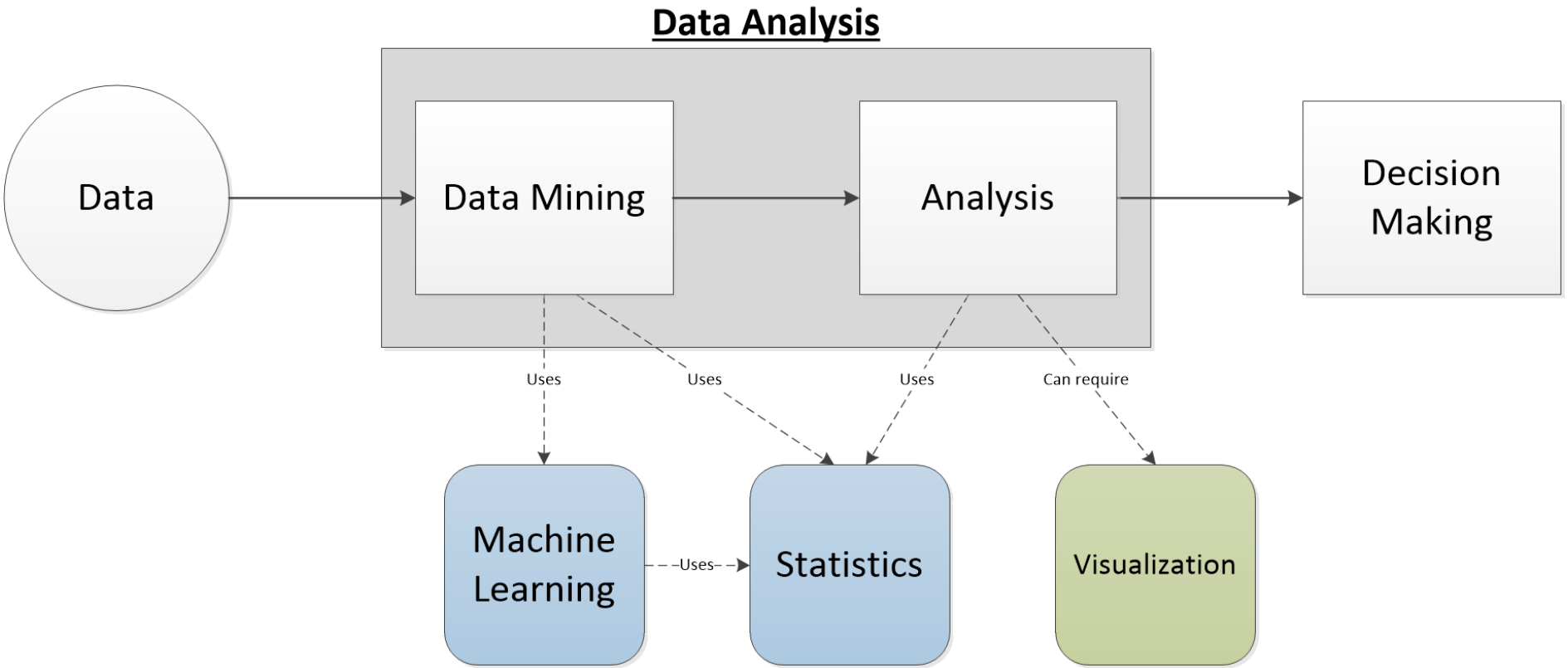
Data Visualization

- **Data Visualization** is a sub-field of algorithmic and programming which helps visualizing data and results to better understand them : plots, figures, schemes, etc.

Data analysis and related terms (5/5)

Data Visualization

- **Data Visualization** is a sub-field of algorithmic and programming which helps visualizing data and results to better understand them : plots, figures, schemes, etc.



Data analysis in your cursus

- A concrete application of statistics and probability
- A strong basis to understand Machine Learning concepts

Why analyzing data ? (1/3)

Lots of data are collected and stored in warehouses

- Customer data: e-commerce, purchases at department stores, banking data, health records, etc.
- Image data: remote sensing satellite data, medical images
- Web data: web traffic, social networks, web navigation
- Science data: genetic data, weather data, traffic data, electricity consumption data.
- ...

All these data are too many and too big to be analyzed manually using traditional techniques. Yet, there is a social need (real or artificial) to process these data.

Why analyzing data ? (2/3)

Social needs that require data analysis

- Weather forecast and prediction of natural disasters
- Network management (electrical, web, traffic, etc.)
- Recommender systems: e-commerce, custom recommendations, etc.
- Medical applications: genome analysis, automated diagnosis, medical image analysis
- Financial: Credit risk analysis, stock market analysis, fraud detection, etc.
- ...

Why analyzing data ? (3/3)

Data analysis helps people from any fields: scientists, bankers, salesmen, engineers, medical doctors, etc.

- Describing, classifying and segmenting data
- Formulating hypotheses
- building models
- Making predictions

Computers have become cheaper and more powerful

- Can run more and more complex simulations
- Data collected and stored at enormous speeds

Databases and knowledge

- **Databases to be mined**:
  - Relational, transactional, object-oriented, object-relational, active, spatial, time-series, text, multi-media, heterogeneous, legacy, web, image, categorical, numerical, etc.
- **Knowledge to be mined**:
  - Characterization, discrimination, association rules, classification, clusters, trends, deviation and outlier detection, etc.
  - Statistical models
- **Techniques**:
  - Data warehouse (OLAP), Machine Learning, statistics, pattern recognition, neural networks, visualization, etc.

Data Analysis tasks

Prediction Tasks

Using known variables and patterns to predict unknown and future data.

- Regression
- Time series predictions
- Classification

Description Tasks

Finding patterns and structures in a data set that can be understood by a human.

- Clustering
- Association rules mining
- Sequential pattern discovery

Example: Rule mining

Given a set of records, rule mining consists in producing dependency rules which will predict the occurrence of an item based on the occurrence of the others.

ID	Items
1	bread, coke, milk
2	beer, bread
3	beer, coke, diaper, milk
4	beer, bread, coke, diaper, milk
5	coke, bread, milk

Rule mined:

$\{\text{milk}\} \rightarrow \{\text{coke}\}$

$\{\text{diaper, milk}\} \rightarrow \{\text{beer}\}$

The sad truth about diapers and beer

Figure: Don't be surprised if you find beers stacked next to the diapers !

Steps of a KDD process

- Learning the application domain
  - Relevant prior knowledge, goals and applications
- Creating a target data set: data selection
- **Data cleaning** and pre-processing (may be 60% of the process)
- **Data reduction and transformation**
  - Find useful features, dimensionality/variable reduction, invariant representation
- Choosing a Data Mining task
  - Summarization, classification, regression, clustering, etc.
- Choosing a **Machine Learning** algorithm
- **Data Mining**: Search for interesting patterns
- **Results evaluation and knowledge presentation**
  - Visualization, transformation, removing redundant patterns, etc.
- Use/re-use mined knowledge

Goals of Data Mining and KDD

- Answer a question or solve a problem.
- Make data intelligible
- Retrieve and select patterns, information and structures
- Determine the validity of the extracted information: problem sampling fluctuation, problem generalization.

Data Mining and KDD : issues and challenges

- Human interaction: Data visualization, model interpretation.
- Data pre-processing: Irrelevant data (object selection), outliers, noisy data (irrelevant or noisy features), missing data.
- Large data sets: too many objects, too many features, scalability.
- Dynamic data sets: on-line data sets (real time).
- Model and algorithm selection
- Result evaluation
- Overfitting
- Hybrid data: text, images, videos, etc.

Social implications

- Privacy issues
- Profiling people
- Unauthorized use
- Systematic processing without human control
- ...

Examples of data analysis software

Free and open-source software

- Software: R, Weka, KNIME, RapidMiner, JHepWork, Scilab, etc.
- Toolkits and packages: NLTK, Carot2, ML-flex, UIMA, etc.

Commercial software

SAS, Matlab & Simulink, IBM InfoSphere Warehouse, Microsoft Analysis Services, STATISTICA, Oracle Data Mining, LIONsolver, etc.

Examples

What is **not** data analysis

- Looking up phone numbers in a data base
- Querying a web search engine for "Amazon"
- Aggregating columns from database tables using SQL commands

What is data analysis

- Searching for the prevalence of certain names in certain US locations (e.g O'Brien, O'Rurke, O'Reilly, ... in the Boston area).
- Processing and sorting the results returned by a search engine based on their similarity (e.g. the amazon rain forest and amazon.com).

Outline

A data ?

A **data** *(datum) is a basic description of a thing, an individual, a fact, an instruction or a phenomenon.*

- A data is made of one or several descriptive criteria called **variables** or **features**:
  - A single criterion: Univariate data
  - Several criteria: Multivariate data (bivariate data for 2 criteria)

A data ?

A **data** *(datum) is a basic description of a thing, an individual, a fact, an instruction or a phenomenon.*

- A data is made of one or several descriptive criteria called **variables** or **features**:
  - A single criterion: Univariate data
  - Several criteria: Multivariate data (bivariate data for 2 criteria)

A **data set** *is a set containing several data, usually identified by an id*.
The id is not considered a variable.

- Other names for a data: individual, object, data object, observation, point.
- Other names for a data set: population
- Other names for a variable: feature, descriptor, criterion

A data matrix

A data set can usually be represented in the form of a matrix with N lines (for N objects) and D columns (for D variables):

$$X = \{x_1, ...x_N\} = \begin{pmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,D} \\ x_{2,1} & x_{2,2} & & \\ & & \ddots & \\ \vdots & & x_{i,j} & \vdots \\ & & & \ddots \\ x_{N,1} & & \cdots & x_{N,D} \end{pmatrix}$$

A data array

A data set can also be represented as an array:

	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

- Objects: x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9 and x_{10}.
- Variables: $Y1$, $Y2$, $Y3$ and $Y4$.

Types of variables

Quantitative variables

- **Continuous variables**: Size, weight, time, volume, speed, etc.
- **Discrete variables**: Counting the number of item in a room, number of items, etc.

Qualitative variables

Categorical variables: Binary data, colors, gender, having a credit or not, labels, etc.

- **Ordered**: survey result (not satisfied, satisfied, very satisfied), nominal sizes (small, medium, tall, very tall), etc.
- **Unordered**: eye color

Others

Text, images, videos, etc.

Dealing with missing data

Important question : Why are there missing values ?

- Methodology issues ? Incomplete data ? Impossible fields ? Unanswered questions in a survey ? Data encryption issues ?

Dealing with missing data

Important question : Why are there missing values ?

- Methodology issues ? Incomplete data ? Impossible fields ? Unanswered questions in a survey ? Data encryption issues ?

Dealing with missing values

- Ignoring the tuples with missing values
- Fill in the missing values when feasible:
  - Fill with the mean value for the missing attribute
  - Deduce the missing values from similar data (smarter)
  - Used the most probable value for the missing attribute (e.g. bayesian inference)

Outliers

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

Types of outliers

- Inconsistent values: February 29 on a non-leap year, subscription dates prior to the birht date of a customer, Invalid postcode, More than 2 (sometimes 3) values on a "sex" field, etc.
- Values that are way out of range compared with the rest of the data set.

Dealing with outliers

There are several ways to deal with outliers:

- Ignoring them or deleting them, if their numbers are not too high and their distribution sufficiently random.
- Keeping them and tolerating a small margin of error.
- Treating outlier variables as missing variables and replacing them.
- Keeping the observation partially, by ignoring the aberrant variables (can prove complicated)

Outline

Introduction

Univariate statistics of variables are applicable to **data sets with a single variable**, or to **individual variables** in a data set with several features.

- Extracting and understanding important information: mean value, range, variance, distribution, etc.
- Discretizing continuous variables when applicable.
- Summarizing a content using graphs.
- Detecting anomalies: missing variables, extreme values, etc.

Measures of central tendency

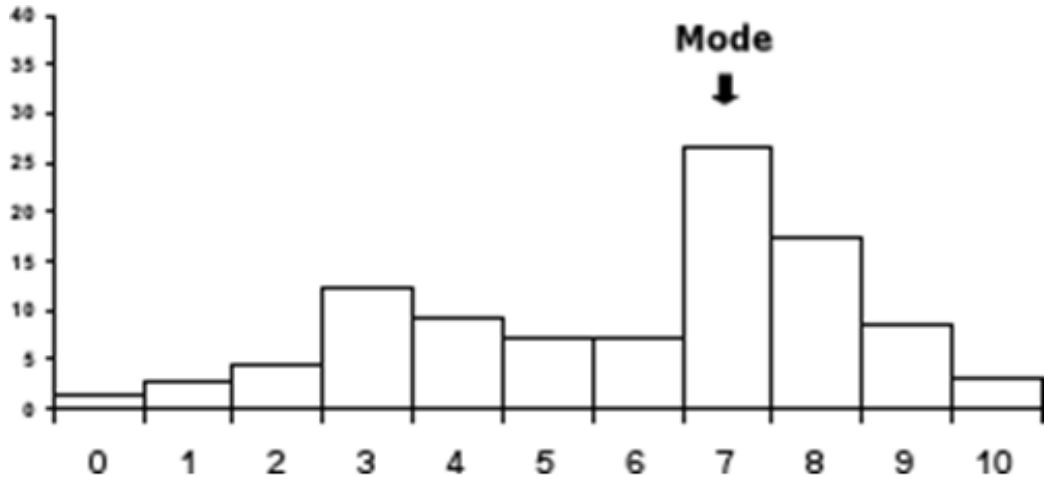
Definition

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

They can be used to answer to questions such as: What is the "median" salary of a football player? How many children has a "typical" French family? What is the "average" grade for this exam ?

- The term central tendency refers to the "middle" value or perhaps a typical value of the data, and is measured using the **mean**, the **median**, or the **mode**.
- Each of these measure is calculated differently, and their relevance depends upon the situation.

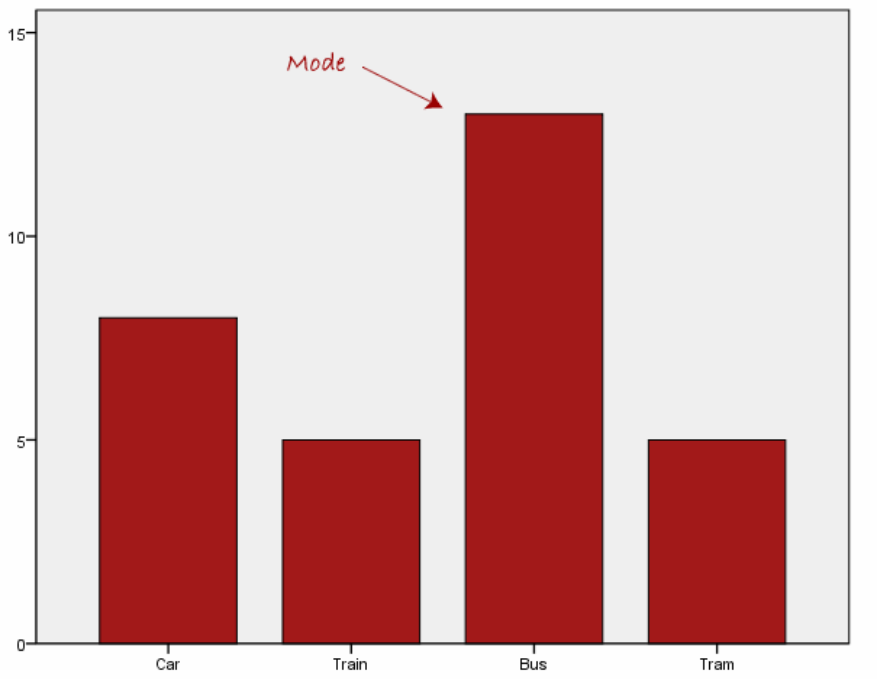
The Mode



The mode(s) *is the most frequently occurring value(s) in the data set.*

- The mode is most easily identified in an ordered frequency histogram.
- On a histogram it represents the highest bar.
- The mode is not necessarily unique.
- In surveys, the mode is the most popular option.

The Mode

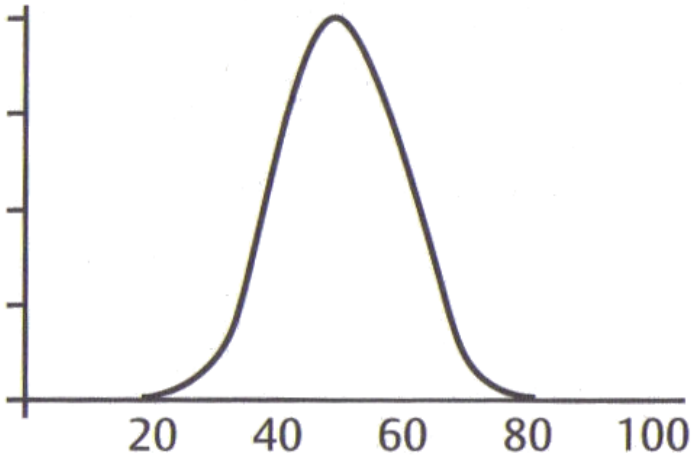


- The mode can be used with categorical unordered data.
- In the histogram above, one can see that the bus is the most common form of transport.
- However, it is unclear whether there are one or two modes ...

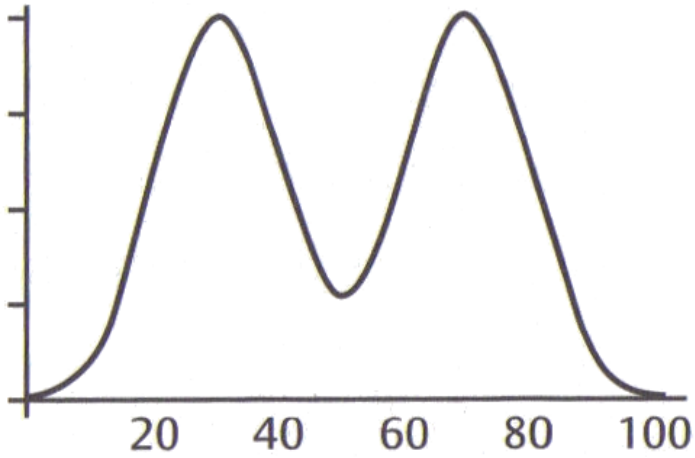
The Mode

Modes can also be found in data distributions. A distribution can be:

- Unimodal
- Bimodal
- Multimodal



(a) Unimodal distribution



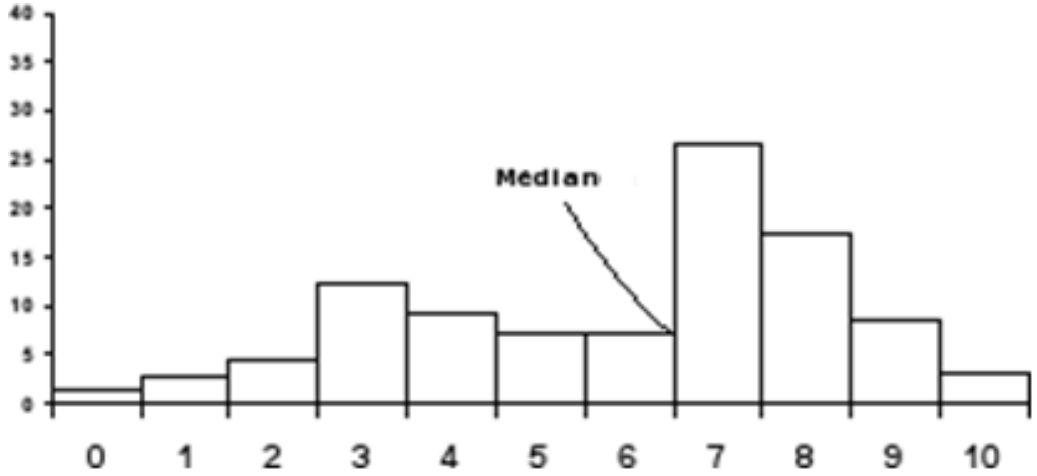
(b) Bimodal distribution

Median

The median *is a measure that allows to define the value which cuts the distribution in two equal parts.*

- In an ordered data set, the median is computed as follows:

$$median = \begin{cases} x_{\frac{N+1}{2}} & \text{if } N \text{ is odd} \\ \frac{1}{2}\left(x_{\frac{N}{2}} + x_{\frac{N+1}{2}}\right) & \text{if } N \text{ is even} \end{cases}$$



Median

The data set needs to be sorted from lowest to highest values in order to compute the median.

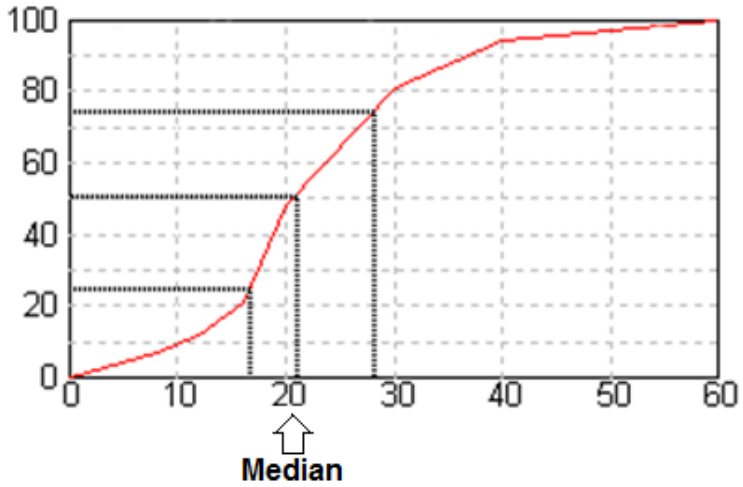


Figure: The median is the value which cuts an ordered set in two subsets of the same size

Median

The data set needs to be sorted from lowest to highest values in order to compute the median.

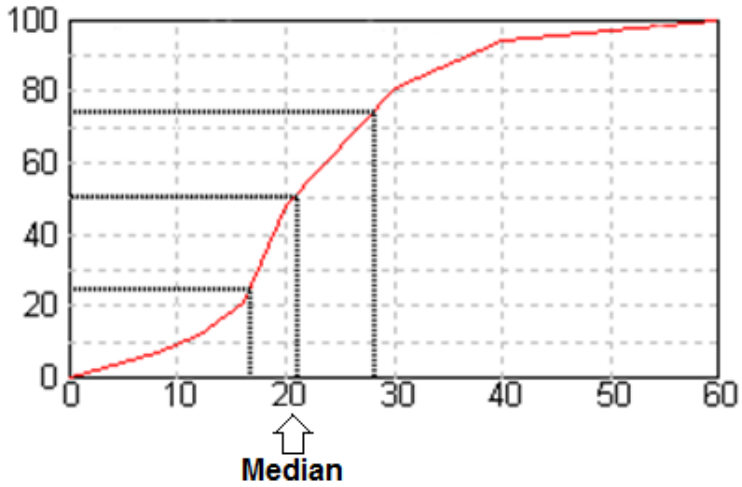


Figure: The median is the value which cuts an ordered set in two subsets of the same size

The median can be used to compute the **quartiles**:

- The median is the 2nd quartile.
- Computing the median of the subsequent subsets gives the 1st and 3rd quartiles.

Median: Example 1

- Let's consider the following data:

65	55	89	56	35	14	56	55	87	45	92

- We first need to rearrange the data:

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.

Median: Example 2

- What happens when you have an even number of scores ? What if you had only 10 scores ? Let's take a look at the example bellow:

65	55	89	56	35	14	56	55	87	45

- Again, we first need to sort the data:

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

- We have to take the 5th and 6th score in our data set and consider there mean value. The new median is **55.5** !

Median: Advantages

The median only uses the relative position of the observation in the ordered data set.

- Sample A: 13, 15, 17, 19, 23
- Sample B: 13, 15, 17, 19, 400

In both cases, the median is **17**.

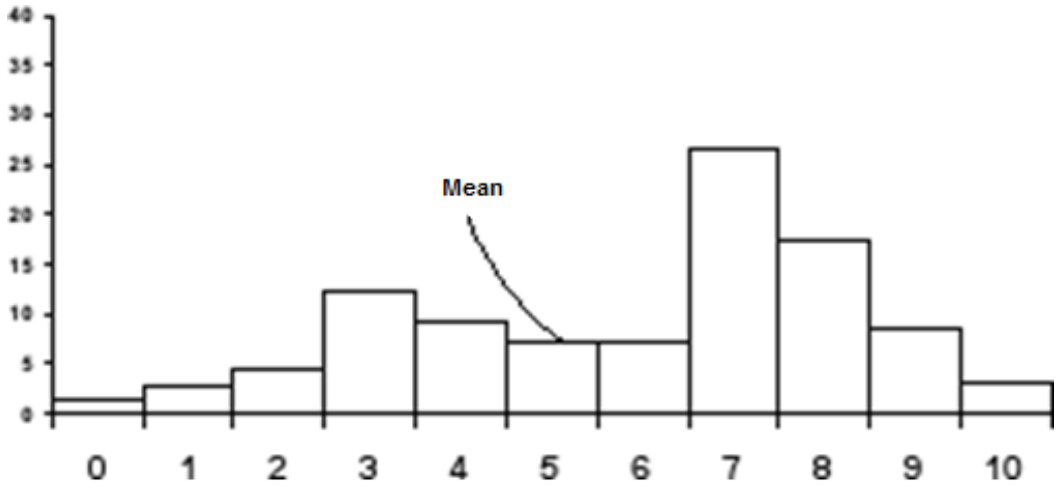
Advantages

- The median is not affected by outliers.
- The median is useful in cases where there are missing data.

Median: Advantages

- The median is often used when there are 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 are often very **skewed**. For such data, the median often is reported instead of the mean.
- For example, in a group of people where the salary of one person is 10 times the mean, the mean salary of the group may be unusually dragged up. In this case, the median may better represent the typical salary level of the group.

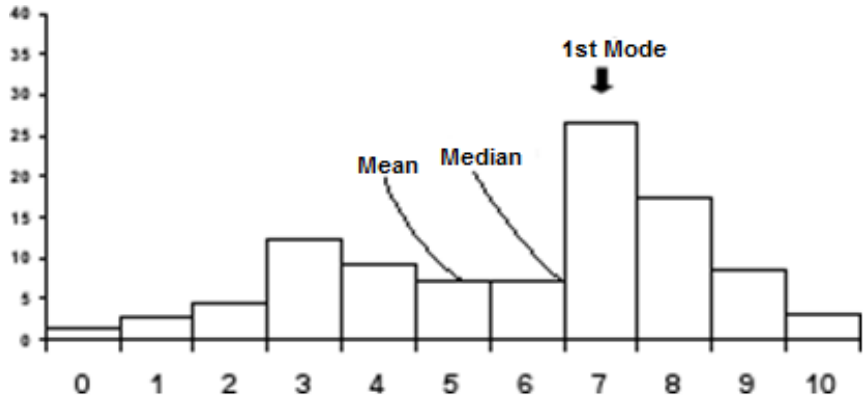
Mean



- The mean value is traditionally denoted $\bar{x}$, μ, or $\mathbb{E}[X]$ for the **expected value** of a random variable X.
- 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:

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

Measures of central tendency: Example



The 3 criteria give different information and have different issues:

- The mode only use the most frequent values of the distribution.
- The median only uses the position of the observations.
- The mean is sensitive to extreme values (outliers).

Measures of central tendency: Summary

Type of variable	Best measure of central tendency
Nominal	Mode
Ordinal (not skewed)	Mean
Ordinal (skewed)	Median

Real examples

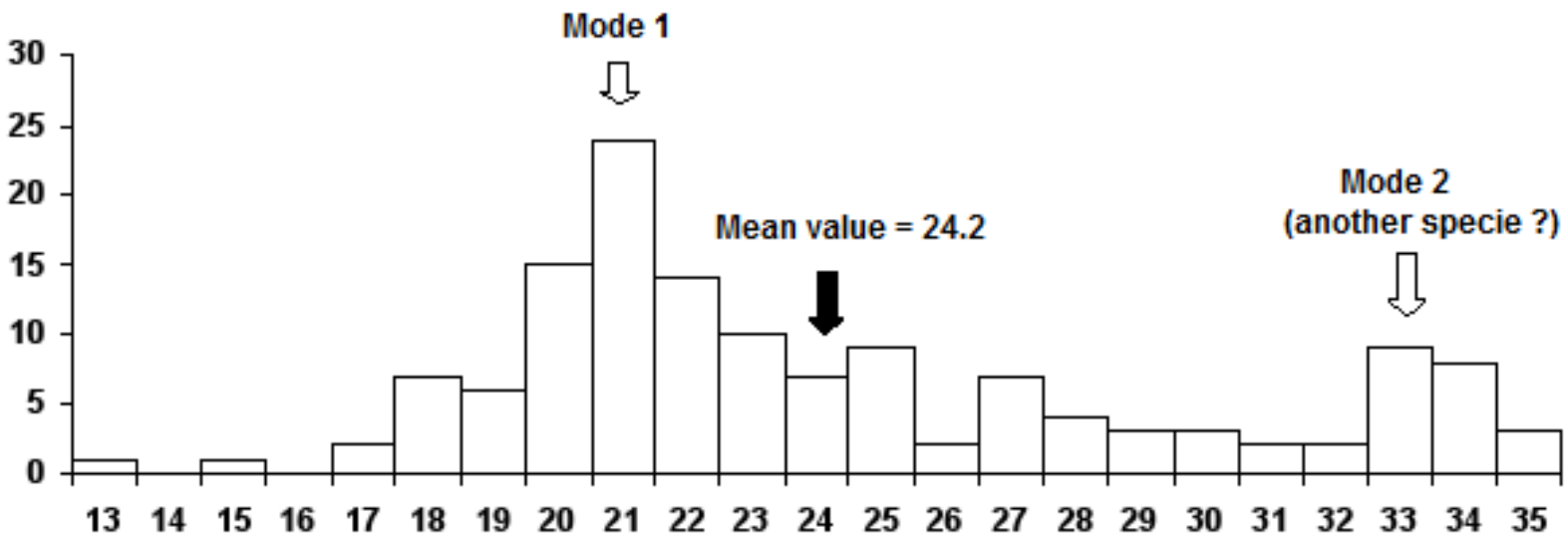


Figure: Number of petals on daisy flowers

Real examples

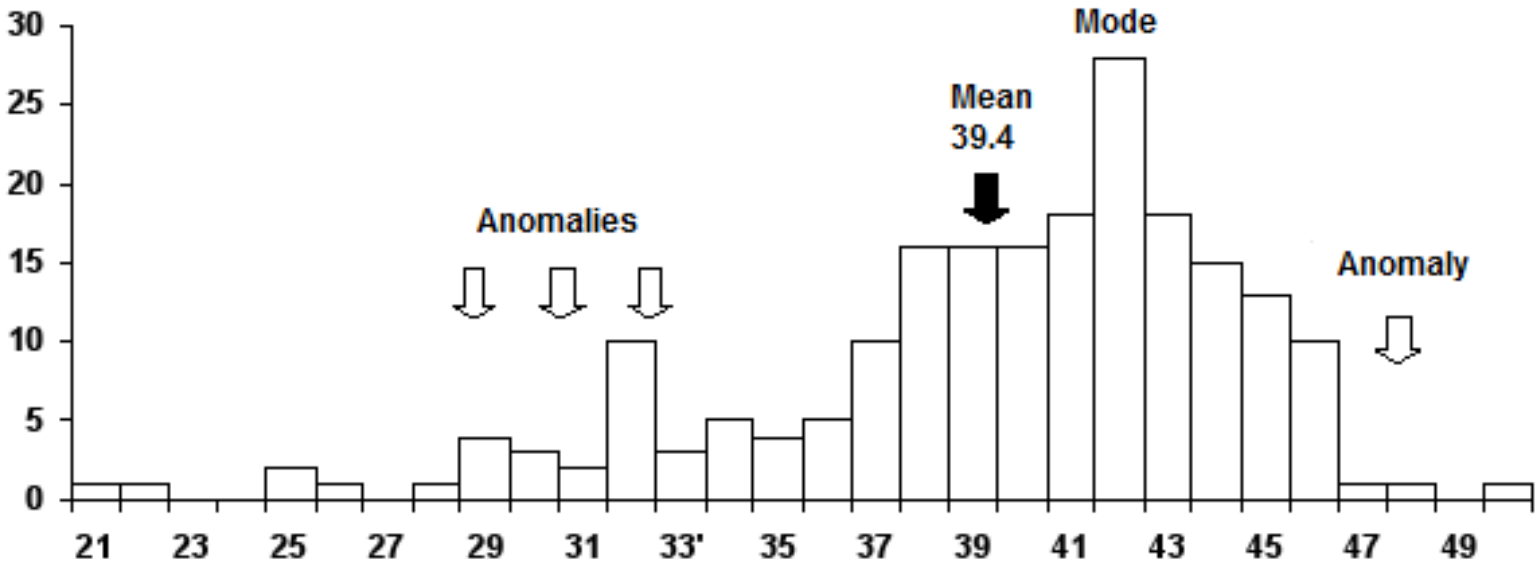


Figure: Maple seeds size

Definition

A **dispersion criterion** *is a value which represents the homogeneity in the values of a variable.*

- **Statistical dispersion** (also called **statistical variability** or variation) is the variability in a variable or a probability distribution.
- A **measure of statistical dispersion** is a NON-NEGATIVE real number equal to zero if all the data are the same and that increases as the data become more diverse:
  - Standard deviation
  - Interquartile range, interdecile range
  - Median absolute deviation
  - Mean absolute deviation (also simply called mean deviation)

Practical interest

For your holidays, you have the choice between:

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

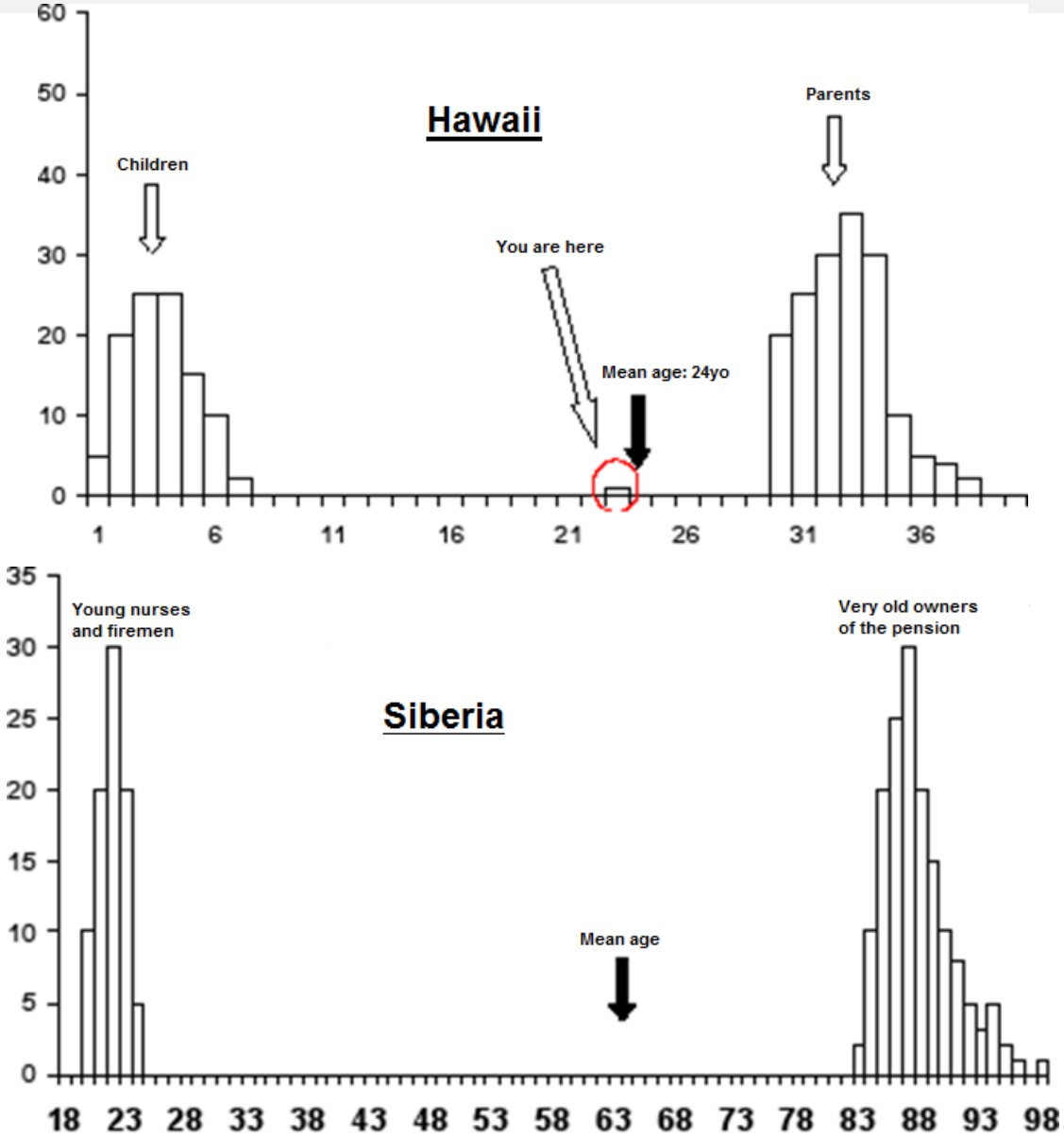
Practical interest

For your holidays, you have the choice between:

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

Where would you go ?

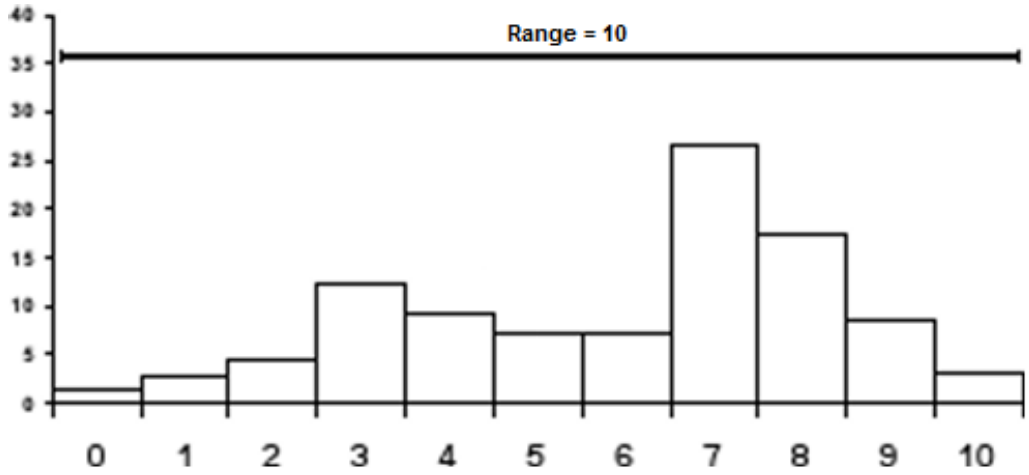
Practical interest



Range

The **range** *is the difference between the largest and the smallest observed value.*

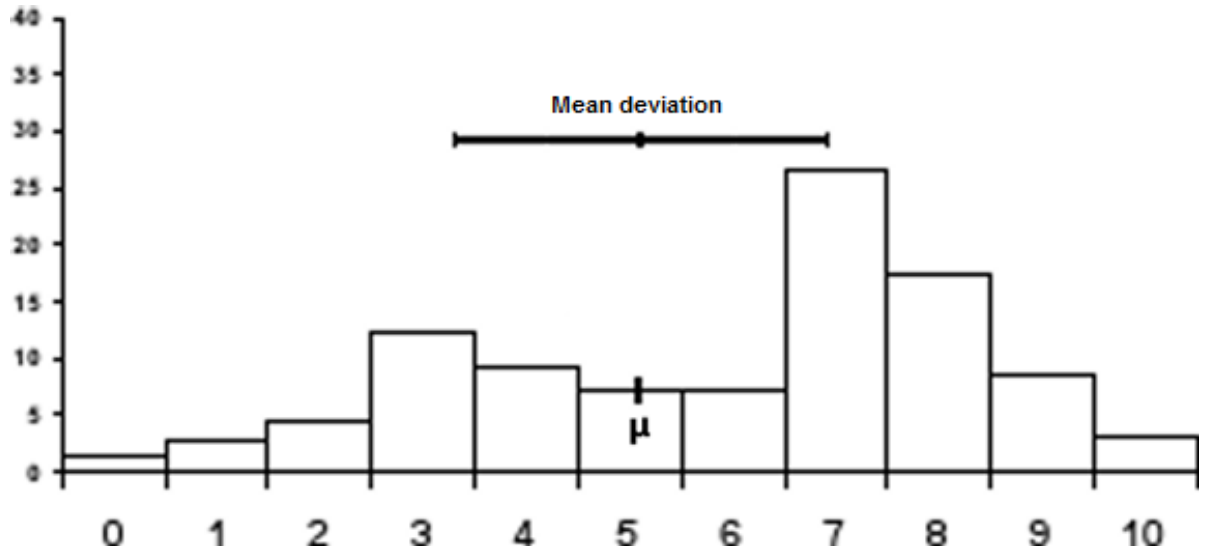
$$Range = X_{max} - X_{min}$$



Mean Deviation

The **mean deviation** *is the average deviation from the mean value.*

$$\sigma_{\bar{x}} = \frac{1}{N} \sum_{i=1}^{N} |x_i - \bar{x}|$$



Variance

The **variance** *is the mean of the squared deviation of a variable from its expected value or mean.*

$$\sigma_X^2 = var(X) = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2$$

- The variance is denoted σ^2, $Var(X)$, or $V(X)$.

Variance

The **variance** *is the mean of the squared deviation of a variable from its expected value or mean.*

$$\sigma_X^2 = var(X) = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2$$

- The variance is denoted σ^2, $Var(X)$, or $V(X)$.

Remark

If the data are ponderated, we have:

$$\sigma^2 = \sum_{i=1}^{N}(p_i x_i - \bar{x})^2, \qquad \bar{x} = \frac{\sum_{i=1}^{N} p_i x_i}{\sum_{i=1}^{N} p_i}$$

Variance: König-Huygens Theorem

The variance can also be computed using the König-Huygens Theorem: The variance of a random variable X is the expected value of the squared deviation from the mean of X, $\mu = \mathbb{E}[X]$.

$$Var(x) = \mathbb{E}\left[(X - \mathbb{E}[X])^2\right] \tag{1}$$
$$Var(x) = \mathbb{E}\left[(X^2 - 2X\mathbb{E}[X] + \mathbb{E}[X]^2\right] \tag{2}$$
$$Var(x) = \mathbb{E}[X^2] - 2\mathbb{E}[X]\mathbb{E}[X] + \mathbb{E}[X]^2 \tag{3}$$
$$Var(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2 \tag{4}$$

Standard deviation

The **standard deviation** *describes the "typical" difference between the observations and the mean value.*

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

- The standard deviation is denoted s or σ.

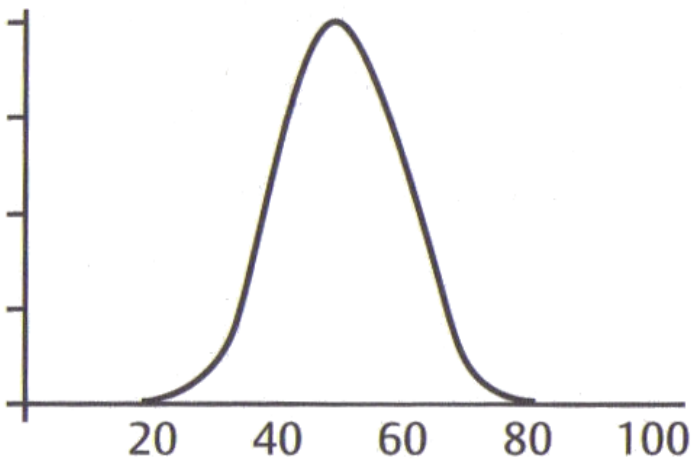
Example:

Sample	Mean	Standard deviation
79,80,81	80	$\sqrt{\frac{2}{3}}$
60,80,100	80	$20\sqrt{\frac{2}{3}}$

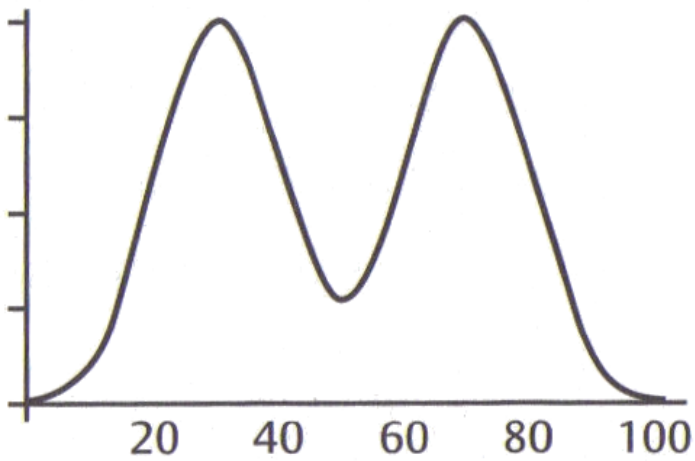
Data distribution

A **distribution** *describes the likelihood (probability) for a variable to take a given value.*

- A distribution can be seen as the continuous version of a frequency histogram.



(a) Unimodal distribution



(b) Bimodal distribution

Describing distribution

To describe a unimodal distribution, we analyse:

- Its mean
- Its standard deviation

To further analyze a distribution, one may also want to assess its shape:

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

Describing distribution

To describe a unimodal distribution, we analyse:

- Its mean
- Its standard deviation

To further analyze a distribution, one may also want to assess its shape:

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

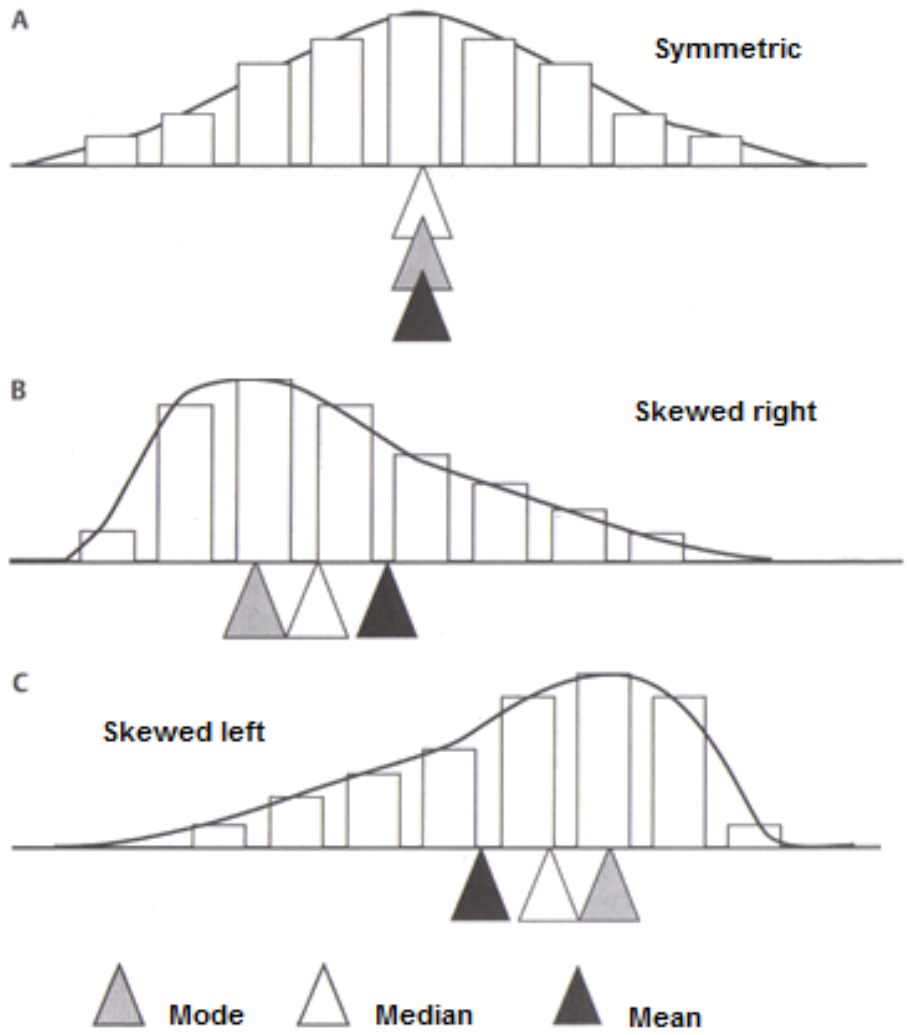
Adaptation to multimodal distributions

The same process can be done for multimodal distributions by analyzing the standard deviations, degrees of Skewness and degree of Kurtosis around each mode (instead of the mean).

Skewness

The skewness of a distribution can be determined as follows:

- Symmetric when:
  $mode = median = mean$
- Skewed right when:
  $mode < median < mean$
- Skewed left when:
  $mode > median > mean$



Fisher-Pearson coefficient of Skewness

Fisher-Pearson coefficient of Skewness:

$$Sk_X = \frac{\sum_{i=1}^{N}(x_i - \bar{x})^3}{N \times s^3(X)}$$

Adjusted Fisher-Pearson coefficient of Skewness:

$$ASk_X = \frac{\sqrt{N(N-1)}}{N-1} \frac{\sum_{i=1}^{N}(x_i - \bar{x})^3}{N \times s^3(X)}$$

- If $Sk_X = 0$, we have a perfectly symmetrical distribution (the values are spread uniformly around the mean).
- If $Sk_X > 0$, positively skewed/skewed right (the distribution spreads more towards higher values).
- If $Sk_X < 0$, negatively skewed/skewed left (the distribution spreads more towards lower values).

Kurtosis measure

Kurtosis Measure:

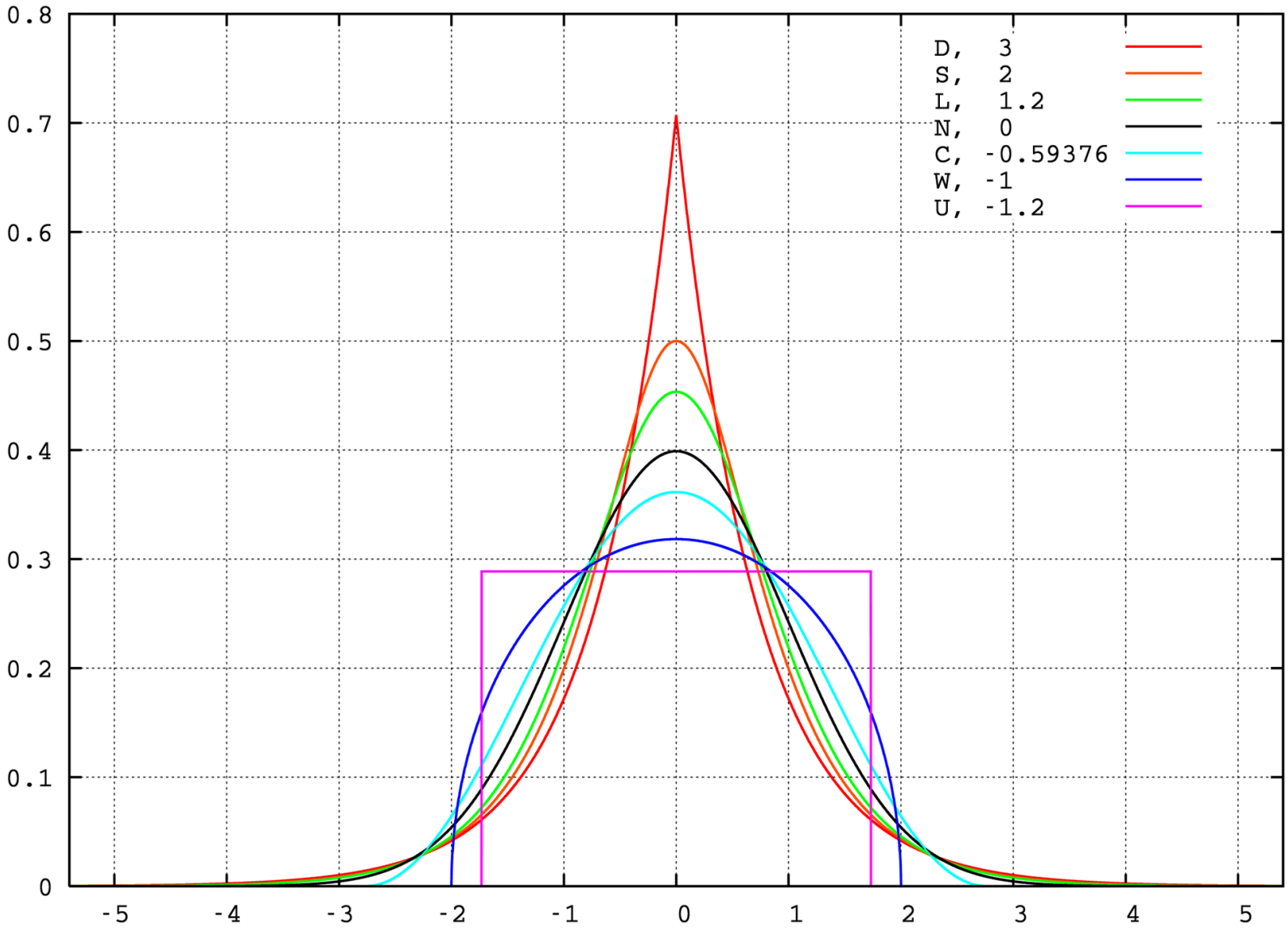
$$\tilde{Ku}_X = \frac{\sum_{i=1}^{N}(x_i - \bar{x})^4}{N \times s^4(X)}$$

In practice, we use the normalized Kurtosis measure (for a standard normal distribution):

$$Ku_X = \frac{\sum_{i=1}^{N}(x_i - \bar{x})^4}{N \times s^4(X)} - 3$$

- When $Ku_X < 0$, the distribution is **Leptokurtic**.
- When $Ku_X > 0$, the distribution is **Platykurtic**.

Kurtosis measure



Outline

Random variables

A **random variable** *is a* **variable** *whose value is subject to variations due to chance or randomness. It can take on a set of possible* **different values** *(continuous or not), and is required to be* **measurable** *so that it is possible to determine* **the probability** *that the variable takes any given value.*

Random variables

A **random variable** *is a* **variable** *whose value is subject to variations due to chance or randomness. It can take on a set of possible* **different values** *(continuous or not), and is required to be* **measurable** *so that it is possible to determine* **the probability** *that the variable takes any given value.*

Random variable: Properties

A random variable is a **measurable function**: $\mathrm{X} : \Omega \to E$, with:

- Ω the probability space containing all possibles values for the variable.
- E a measurable space: e.g. $\mathbb{R}$.
- A random variable **does not return a probability**. The probability of a measurable set of outcomes (i.e. an event) is given by the probability measure P with which Ω is equipped.
- X returns a numerical quantity of outcomes in Ω: e.g. the number of heads in a random collection of coin flips.

Random variables

- Any measure of a phenomenon for which is impossible to know the outcome in advance is a random variable:
  - Dice rolls, coins launches, weather prediction, etc.
- Any data from a random sampling is a random variable.

Distribution functions

In probability, a **distribution function** *of a random variable (feature)* X *is the function* F_{X} *that for any value* $x \in \Omega$ *associates the* **probability** *that* X *has a value lower or equal to* x*.*

Distribution function: Properties

- $F_{\mathrm{X}} : \Omega \to [0, 1]$
- $F_{\mathrm{X}}(x) = P(\mathrm{X} \leq x)$
- F_{X} is increasing
- $\lim_{x \to -\infty} F_{\mathrm{X}}(x) = 0$
- $\lim_{x \to +\infty} F_{\mathrm{X}}(x) = 1$

The distribution function is sometimes called **cumulative distribution function**. It is the continuous version of a cumulative frequency function.

Distribution function

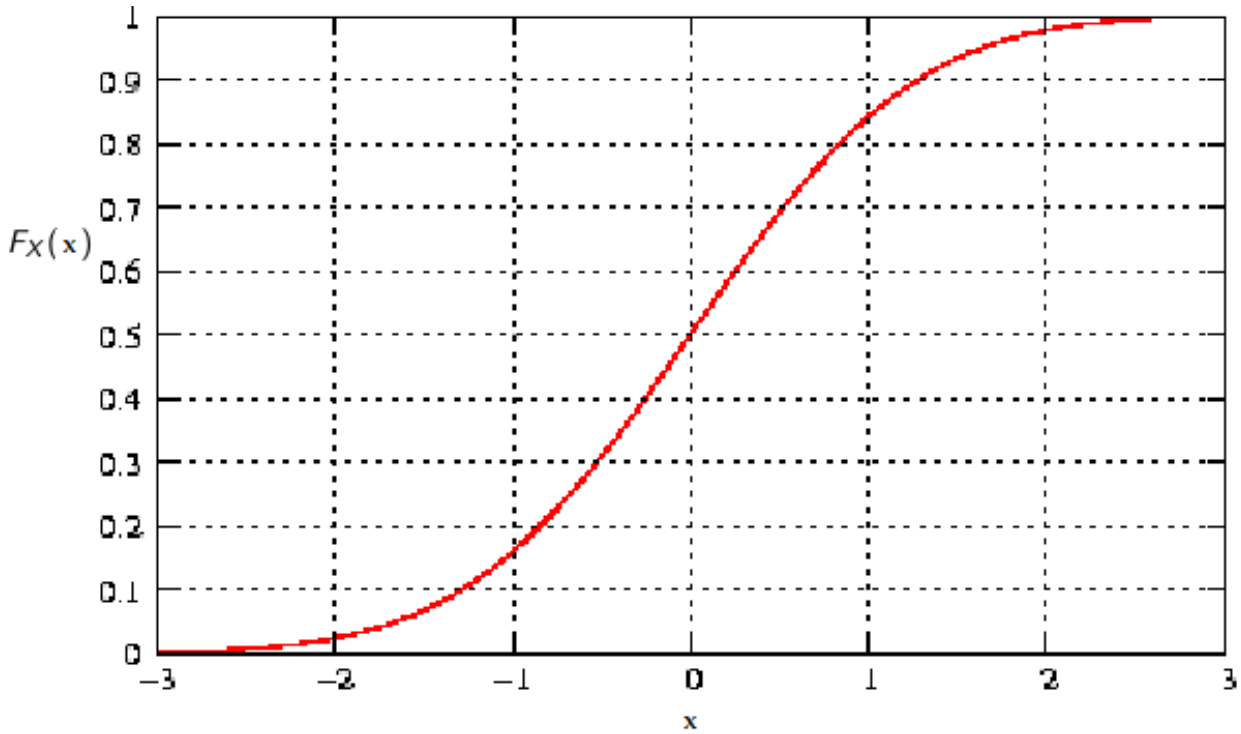


Figure: The distribution function is similar to the cumulative frequency curve, except that it is a probability.

Distribution function and quartiles

The **quartiles** q_1, q_2 and q_3 are defined as follows:

- $F(q_1) = 0.25$ is the first quartile and contains 25% of the data. It is also called the lower quartile.
- $F(q_2) = 0.5$ is the second quartile. It splits the data in two equal parts. It is the median.
- $F(q_3) = 0.75$ is the third quartile. It seperate the last 25% of the data. It is also called the upper quartile.

Interquartile range

The interquartile range or IQR contains the 50% of data closest to the median. $IQR = q_3 - q_1$

Probability density

In probability, the **probability density** *of a random continuous real variable X is a function that describes the* **relative likelihood** *for this random variable to take a given value.*

- This function is denoted $f_{\mathrm{X}}(x)$, $p_{\Omega}(x)$ or just $p(x)$.
- In practice, we like f_{X} to express a probability rather than just a likelihood, so that $p(x) = P(\mathrm{X} = x)$.

Probability density function

- $p(x) : \Omega \to [0, 1]$

$$p(x) = \frac{\mathrm{d}F_{\mathrm{X}}(x)}{\mathrm{d}x}$$

Probability density

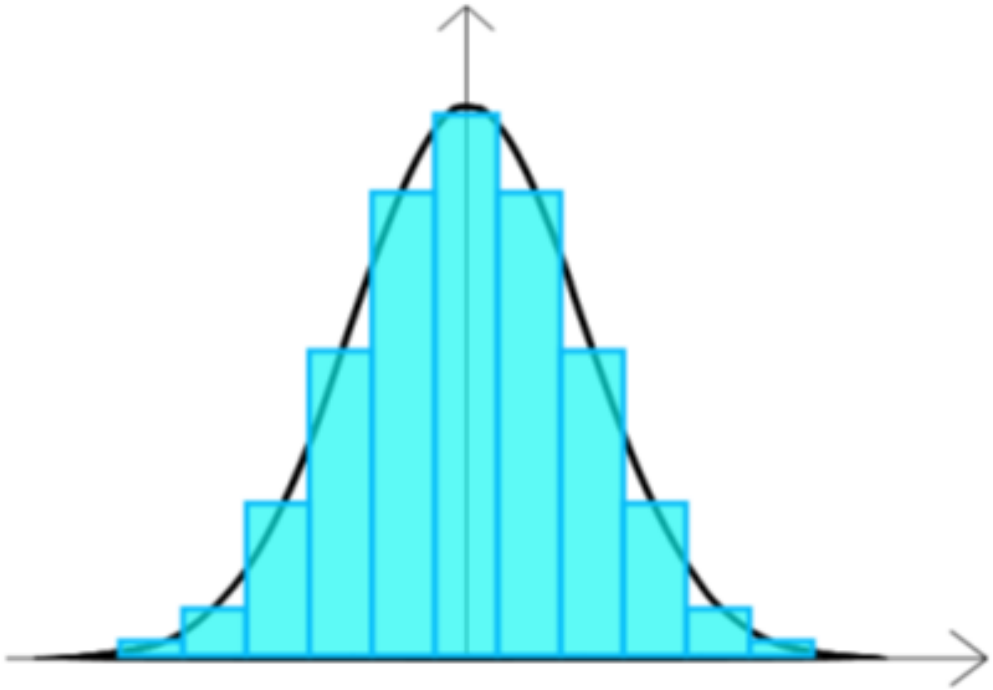
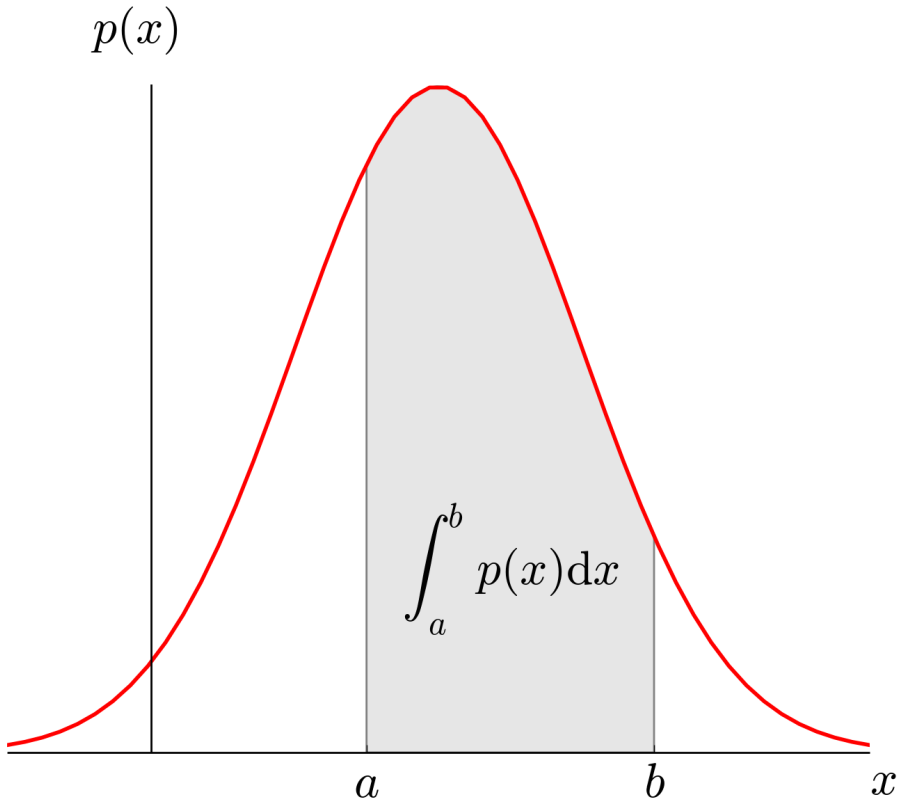


Figure: A density function can be seen as a continuous version of a frequency diagram

Probability density



The probability $P(a \leq \mathrm{X} \leq b)$ is the area under the curve on the interval [i:j]:

$$P(a \leq \mathrm{X} \leq b) = \int_a^b f_{\mathrm{X}}(x)\mathrm{d}x$$

Density functions: properties

Density functions have the following properties:

- $\forall x \in \mathbb{R} \quad p(x) \geq 0$
- $p(x)$ is integrable on $\mathbb{R}$
- $p(x)$ verifies:

$$\int_{\mathbb{R}} p(x)\mathrm{d}x = 1$$

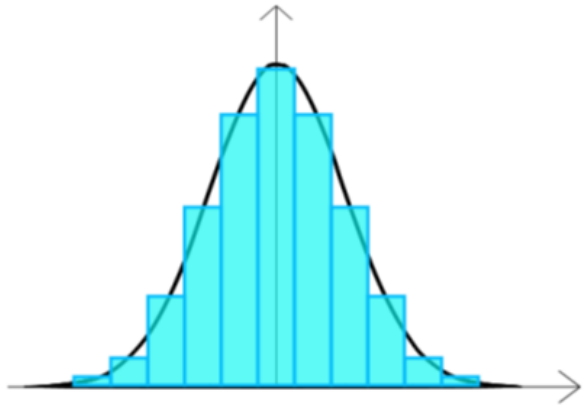
- $p(x)$ verifies:

$$F_X(a) = \int_{-\infty}^{a} p(x)\mathrm{d}x$$

Example: Gaussian distributions

In probability, a random variable X *follows a* **normal distribution** *with a mean* μ *and a standard deviation* σ *if it admits a probability density such that*:

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



The distribution is also called a **Gaussian distribution**.

Standard normal distribution

We call **standard normal distribution** *or* **unit normal distribution** *the Gaussian distribution whose average is zero, and whose standard deviation is the unity ($\mu = 0$ and $\sigma = 1$).*

$$p(x) = \frac{1}{\sqrt{2\pi}} \exp^{-\frac{x^2}{2}}$$

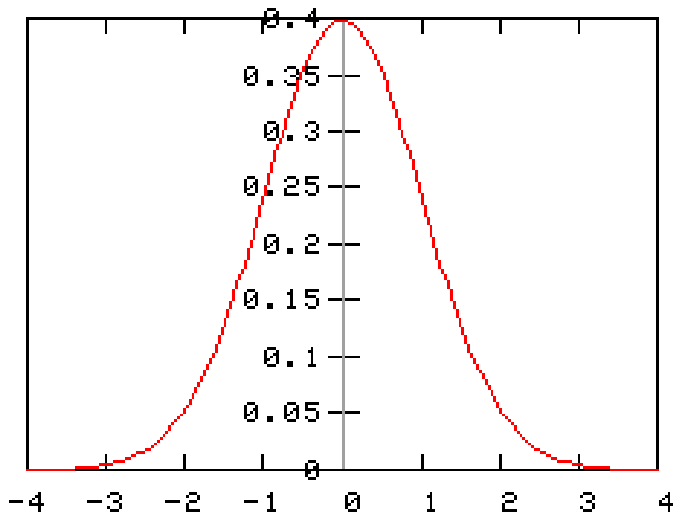


Figure: Density function of a standard normal distribution

Standard normal distribution: distribution function

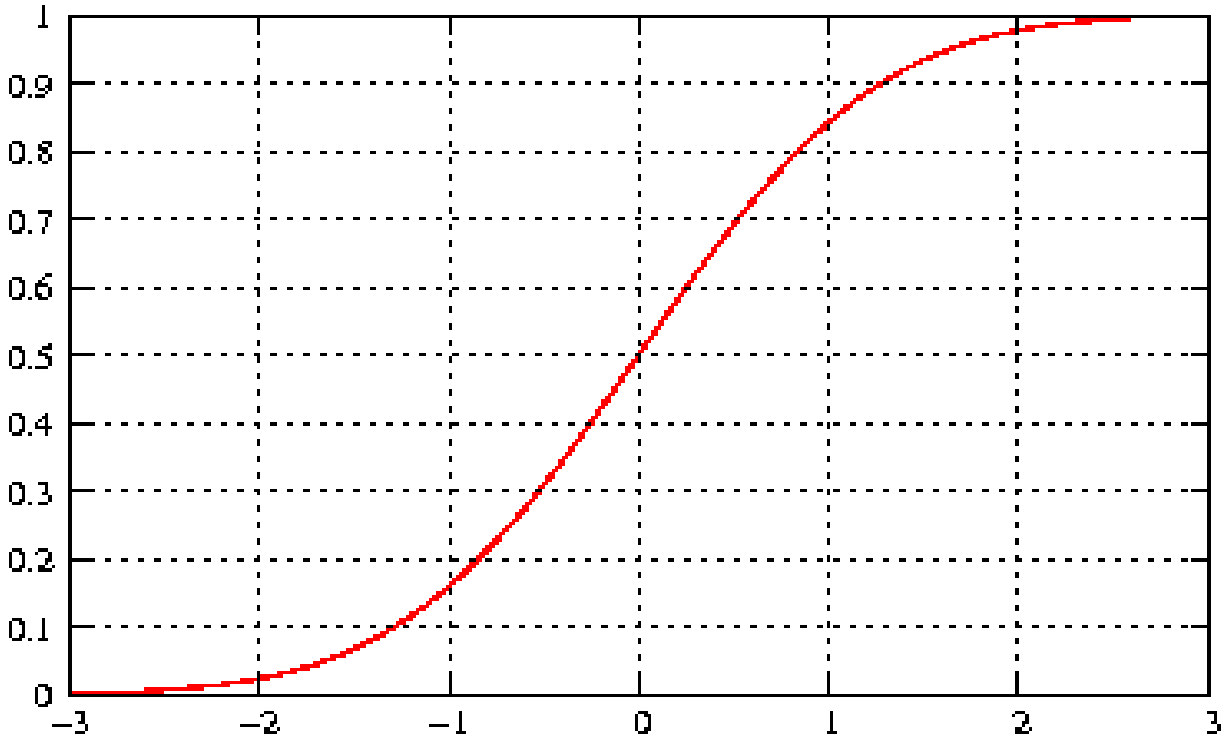
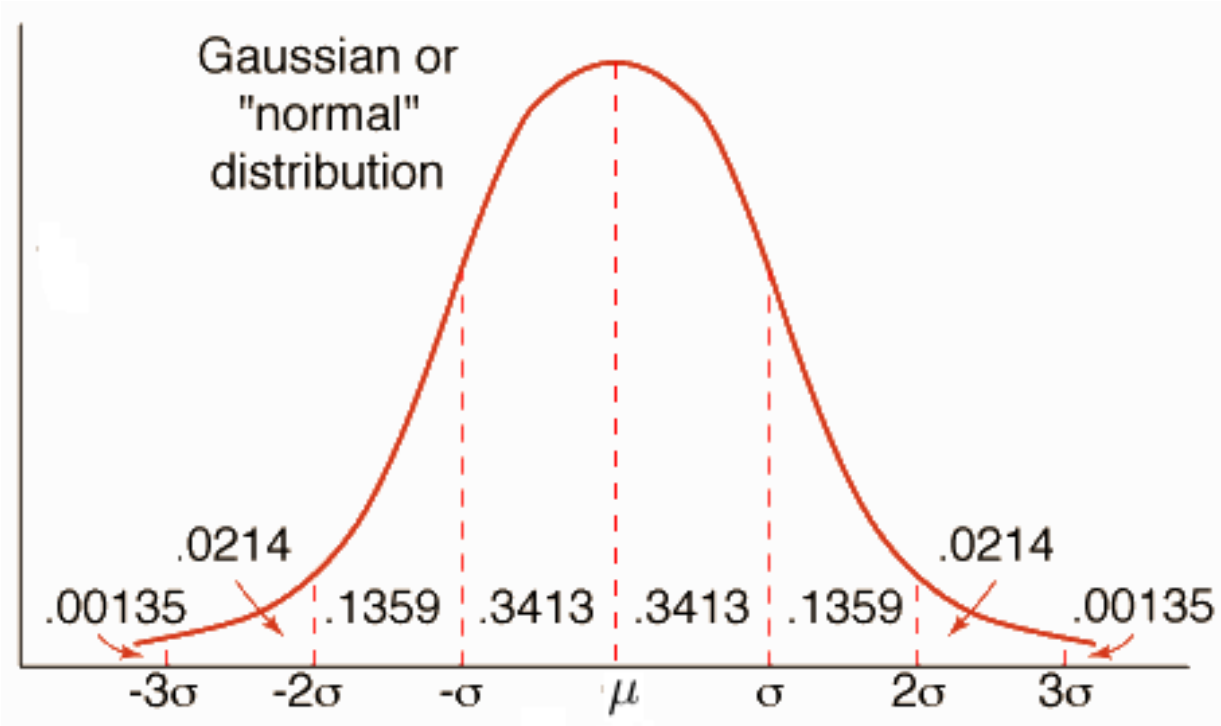


Figure: Distribution function of a standard normal distribution

Gaussian distributions: properties

The Gaussian distribution has the following properties:

- 90% of the data are in the interval $[\mu - 1.64\sigma; \mu + 1.64\sigma]$.
- 95% of the data are in the interval $[\mu - 1.96\sigma; \mu + 1.96\sigma]$.
- 99% of the data are in the interval $[\mu - 2.58\sigma; \mu + 2.58\sigma]$.



Gaussian distributions: properties

- We note $\mathcal{N}(\mu, \sigma^2)$ a normal distribution with the mean μ and variance σ^2.
- For any random variable X that follows a gaussian distribution $\mathcal{N}(\mu, \sigma^2)$, the random variable $\tilde{\mathrm{X}} = \frac{X-\mu}{\sigma}$ follows the standard normal distribution.
- Let $\mathrm{X} \sim \mathcal{N}(\mu, \sigma^2)$ and $(\alpha, \beta) \in \mathbb{R}^*$, then the random variable $\alpha\mathrm{X} + \beta$ follows a gaussian distribution $\mathcal{N}(\alpha\mu + \beta, \alpha^2\sigma^2)$

Gaussian distributions: properties

Multimodal distributions can be made of several Gaussians. It is called a **Gaussian mixture**.

- The mixture of two Gaussian populations is not a Gaussian !

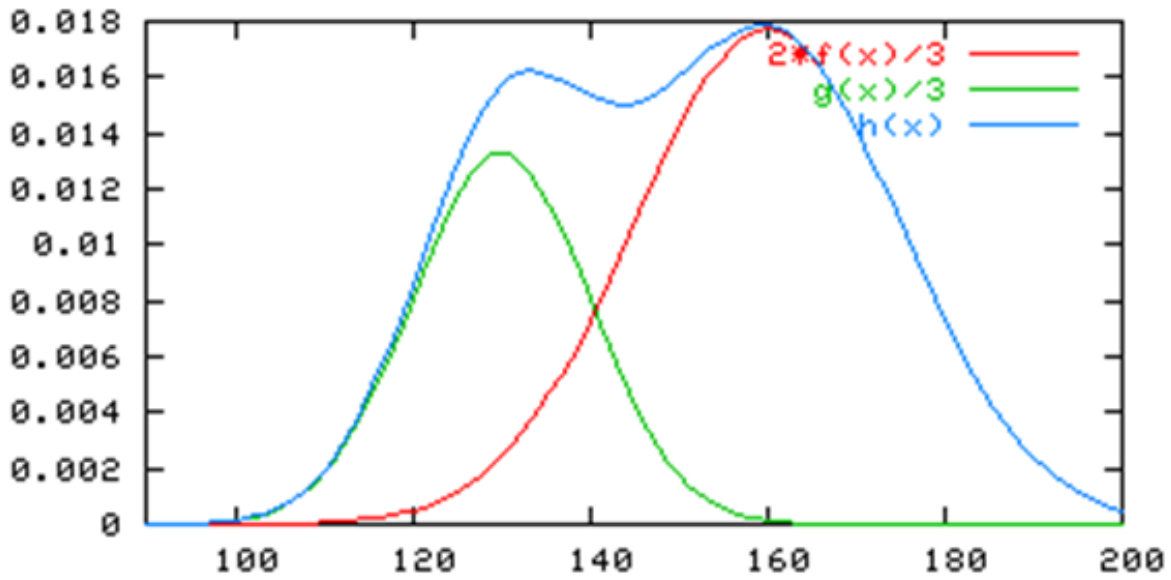


Figure: A bimodal distribution (in blue) with 2 gaussians

Gaussian distributions: properties

Central-limit theorem

Let us consider n random variables $\mathrm{X}_1, \mathrm{X}_2, \cdots, \mathrm{X}_n$ so that:

1. These variables are independent two by two.
2. They all have the same mean μ and the same variance σ^2
3. They have the same probability densities.

Then, if we note $\mathrm{Y} = \sum_{i=1}^{n} \mathrm{X}_i$, we have:

- $\mathrm{Y} \sim \mathcal{N}(n\mu, n\sigma^2)$ when n tends towards infinity.

Gaussian distributions: properties

Central-limit theorem

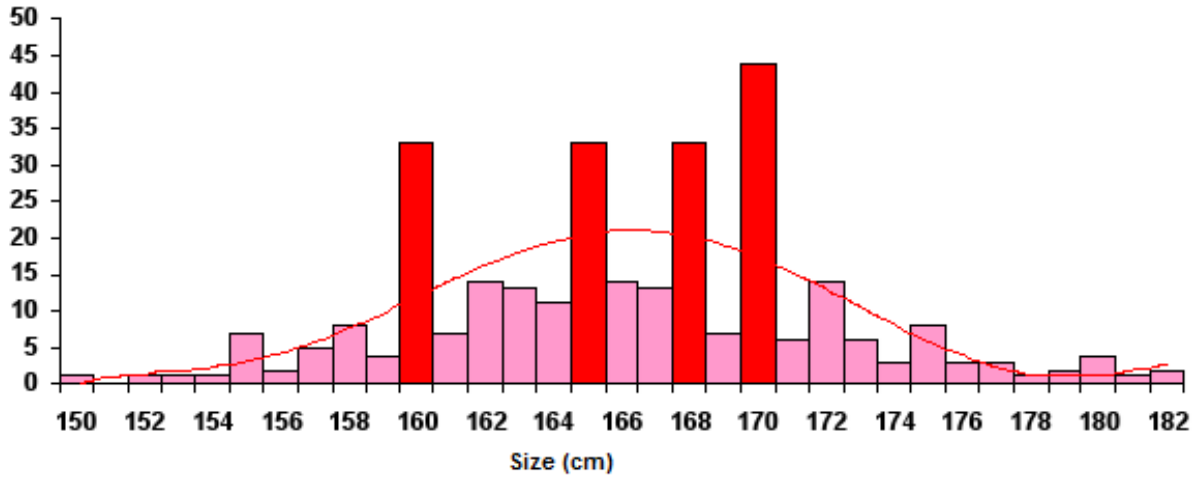
In practice, it has been proved that the theorem remains valid even when conditions 1 and 2 are not met.

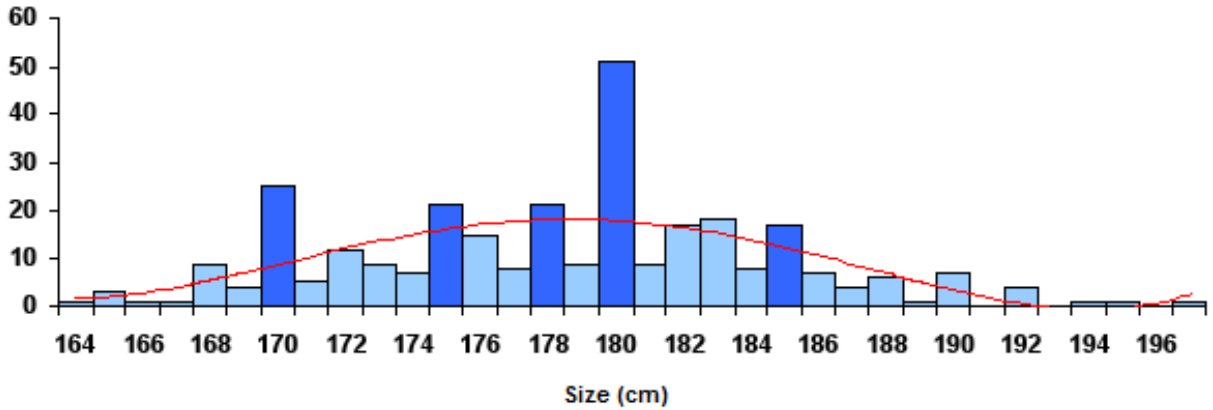
- The independence condition is rarely met in nature.
- The second condition is not vital so long that each local variance is "small" compared with the global variance.

As a result, many natural variables can be approximated using normal distributions.

Gaussian distributions: examples

Many natural phenomenons follow Gaussian distributions:





Gaussian distributions: Shapiro-Wilk Test

The Shapiro–Wilk Test

The Shapiro–Wilk test is a frequentist statistic test of normality : It tests the null hypothesis that a sample $x_1, \cdots, x_n$ came from a normally distributed population (following a gaussian distribution).

- The null hypothesis of this test is that the sample follows a normal distribution.
- If the **p-value** is less than the chosen α (typically 5%), then the null hypothesis is rejected and the data tested are most likely not from a normally distributed population.

Outline

Estimations in statistics

In statistics, **estimation** refers to the process by which one makes inferences about a population, based on information obtained from a **sample**.

Types of estimations

- Point estimate: It consists in estimating a population parameter such as a mean, a median, or a variance.
- Interval estimate: An interval estimate is defined by two numbers, between which a population parameter is said to lie.

- Estimations are often used in **survey methodology**.
- Estimators are often denoted with a hat ($\hat{\mu}$, $\hat{x}$, $\hat{\sigma_X}$)

Position criteria

A **position criterion** *calculated from* **a sample** *of a data set is the* **best estimator** *of the standard position of the population*

- Let $\bar{X}$ be a sample of size n of a population X, in the case of the mean value, the best estimate is to use the original formula:

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

  is the best estimator of the average for the full set X.
- The median and the modes can also be approximated from a sample using their original formulas.

Dispersion measures

Estimation dispersion measures

- Dispersion measures of a sample computed with the original formula tend to be underestimated.
- It is necessary to *correct* them by dividing by $n-1$ instead of n: significant correction for small n, negligible for a bigger n.

- Let $\bar{X}$ be a sample of size n of a population X,

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

is the best estimator of the variance for the full set X.

What are confidence intervals ?

- We have just seen than when analyzing samples and random variables, we compute **estimate** values of different measures like the mean or the variance.
- It raises several questions:
  - Can we know how far/close they are from the truth ?
  - How do we know if we can trust such estimations ? Is there a confidence measure ?

What are confidence intervals ?

- We have just seen than when analyzing samples and random variables, we compute **estimate** values of different measures like the mean or the variance.
- It raises several questions:
  - Can we know how far/close they are from the truth ?
  - How do we know if we can trust such estimations ? Is there a confidence measure ?

Confidence intervals

Confidence intervals are statistical tools that can be used to determine a **range of value** in which the real value of an estimated measure should fall **given a desired level of confidence** (0 to 99%)

Standard error of the mean value

It is possible to evaluate the standard error on the estimated mean value:

$$SE_{\bar{X}} = \frac{s}{\sqrt{n}}$$

- We usually want n to be *"big enough"* ($n > 30$).

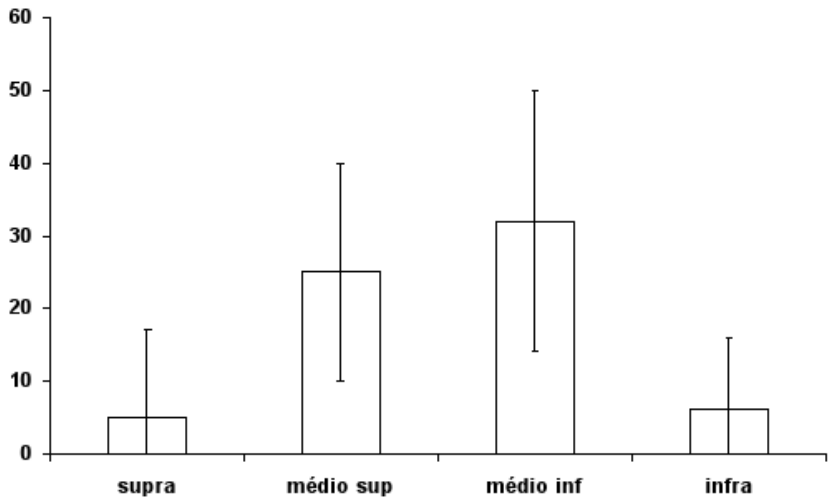


Figure: Example of a graph featuring standard errors bars ($+/-$). The larger the bar, the less reliable the results

Confidence intervals

The **confidence** *interval is the range of value in which the real value of the approximated property can be found with a probability of our choice (usually 95%).*

Confidence Interval of the mean: IC_{95}

In a **normal distribution**, 95% of the values lie within a range of ± 1.96 standard deviations around its mean. Therefore:

$$\hat{\mu} = \mu \pm 1.96 \frac{s}{\sqrt{n}}$$

$$\implies \mu = \hat{\mu} \pm 1.96 \frac{s}{\sqrt{n}}$$

Confidence intervals

Confidence interval of the difference of means: IC_{95}

- We consider two random variables X_A and X_B that follow any distribution of mean μ_a and μ_b and of variances σ_A^2 and σ_B^2 (estimated s_A^2 and S_B^2.
- If X_A and X_B are independent, the difference D between their two estimated means follows a normal distribution:

$$D = (m_A - m_B) \rightarrow \mathcal{N}\left(\mu_A - \mu_B, \frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}\right)$$

Confidence interval of D

$$\Delta = \left[D - 1.96\sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}}; D + 1.96\sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}}\right]$$

Confidence interval of anything

While the confidence interval for the approximated mean value of a normal distribution can be manually computed. For other distributions or criteria, it requires the use of time and computationally expensive methods.

Bootstrap method

1. Take your sample from n data (objects/observations).
2. Take a random sub-sample from your data (with replacement).
3. Compute the value of the parameter you are interested in for this sub-sample.
4. *Repeat at least* **a thousand time**.
5. Delete the 2.5% of highest values, and the 2.5% of lower values.
6. The remaining values are situated in the 95% confidence interval (IC95).

Confidence intervals: The shampoo example

- A famous brand sells a miracle shampoo that prevents hair loss in 75% of men.
- The small writings say that it was tested on a sample of 12 people.
- Should I buy it ?

Computing the confidence intervals that it works

Confidence intervals: The shampoo example

- A famous brand sells a miracle shampoo that prevents hair loss in 75% of men.
- The small writings say that it was tested on a sample of 12 people.
- Should I buy it ?

Computing the confidence intervals that it works

$\hat{\mu} = 0.75$ $n = 12$

Confidence intervals: The shampoo example

- A famous brand sells a miracle shampoo that prevents hair loss in 75% of men.
- The small writings say that it was tested on a sample of 12 people.
- Should I buy it ?

Computing the confidence intervals that it works

$$\hat{\mu} = 0.75 \qquad n = 12 \qquad \hat{s} = \frac{9 \times 0.25^2 + 3 \times 0.75^2}{12-1} = 0.205$$

Confidence intervals: The shampoo example

- A famous brand sells a miracle shampoo that prevents hair loss in 75% of men.
- The small writings say that it was tested on a sample of 12 people.
- Should I buy it ?

Computing the confidence intervals that it works

$$\hat{\mu} = 0.75 \qquad n = 12 \qquad \hat{s} = \frac{9 \times 0.25^2 + 3 \times 0.75^2}{12-1} = 0.205$$

$$SE = \frac{0.205}{\sqrt{12}} = 0.059$$

Confidence intervals: The shampoo example

- A famous brand sells a miracle shampoo that prevents hair loss in 75% of men.
- The small writings say that it was tested on a sample of 12 people.
- Should I buy it ?

Computing the confidence intervals that it works

$$\hat{\mu} = 0.75 \qquad n = 12 \qquad \hat{s} = \frac{9 \times 0.25^2 + 3 \times 0.75^2}{12-1} = 0.205$$

$$SE = \frac{0.205}{\sqrt{12}} = 0.059 \qquad IC95_{\mu} = [0.63; 0.86]$$

Confidence intervals: The shampoo example

- A famous brand sells a miracle shampoo that prevents hair loss in 75% of men.
- The small writings say that it was tested on a sample of 12 people.
- Should I buy it ?

Computing the confidence intervals that it works

$$\hat{\mu} = 0.75 \qquad n = 12 \qquad \hat{s} = \frac{9 \times 0.25^2 + 3 \times 0.75^2}{12-1} = 0.205$$

$$SE = \frac{0.205}{\sqrt{12}} = 0.059 \qquad IC95_{\mu} = [0.63; 0.86]$$

- What if it was tested on a sample of **12 monkeys**, knowing that shampoos tested on monkeys produce the same results on humans in only 60% of cases ?

Outline

Bibliography

- Ad Feelders, Advanced Data Mining 2011
- Srinivasan Parthasarathy, Introduction to Data Mining
- Min Song, Data Mining

Questions ?

Questions ?

For next week

If you plan on using your own computer:

- Install R (`http://cran.r-project.org`)
- Install RStudio

If you are not familiar with R programming:

- Take a look at this tutorial: `http://www-bcf.usc.edu/~gareth/ISL/ISLR%20Seventh%20Printing.pdf` p42-52.