- **ELEMENTI DI STATISTICA**

DATA IS ALSO VERY COMPLEX

- Multiple types of data: tables, time series, images, graphs, etc
- Spatial and temporal aspects
- Interconnected data of different types:
  - From the mobile phone we can collect, location of the user, friendship information, check-ins to venues, opinions through twitter, images though cameras, queries to search engines

EXAMPLE: GENOMIC SEQUENCES

- http://www.1000genomes.org/page.php
- Full sequence of 1000 individuals
- $3 \cdot 10^9$ nucleotides per person → $3 \cdot 10^{12}$ nucleotides
- Lots more data in fact: medical history of the persons, gene expression data

EXAMPLE: ENVIRONMENTAL DATA

- Climate data (just an example)

http://www.ncdc.gov/oa/climate/ghcn-monthly/index.php

- "a database of temperature, precipitation and pressure records managed by the National Climatic Data Center, Arizona State University and the Carbon Dioxide Information Analysis Center"

- "6000 temperature stations, 7500 precipitation stations, 2000 pressure stations"
  - Spatiotemporal data

DATA DEFINITION

- Collection of data objects and their attributes
- An attribute is a property or characteristic of an object
  - Examples: eye color of a person, temperature, etc.
  - Attribute is also known as variable, field, characteristic, or feature
- A collection of attributes describe an object
  - Object is also known as record, point, case, sample, entity, observations, or instance

Attributes

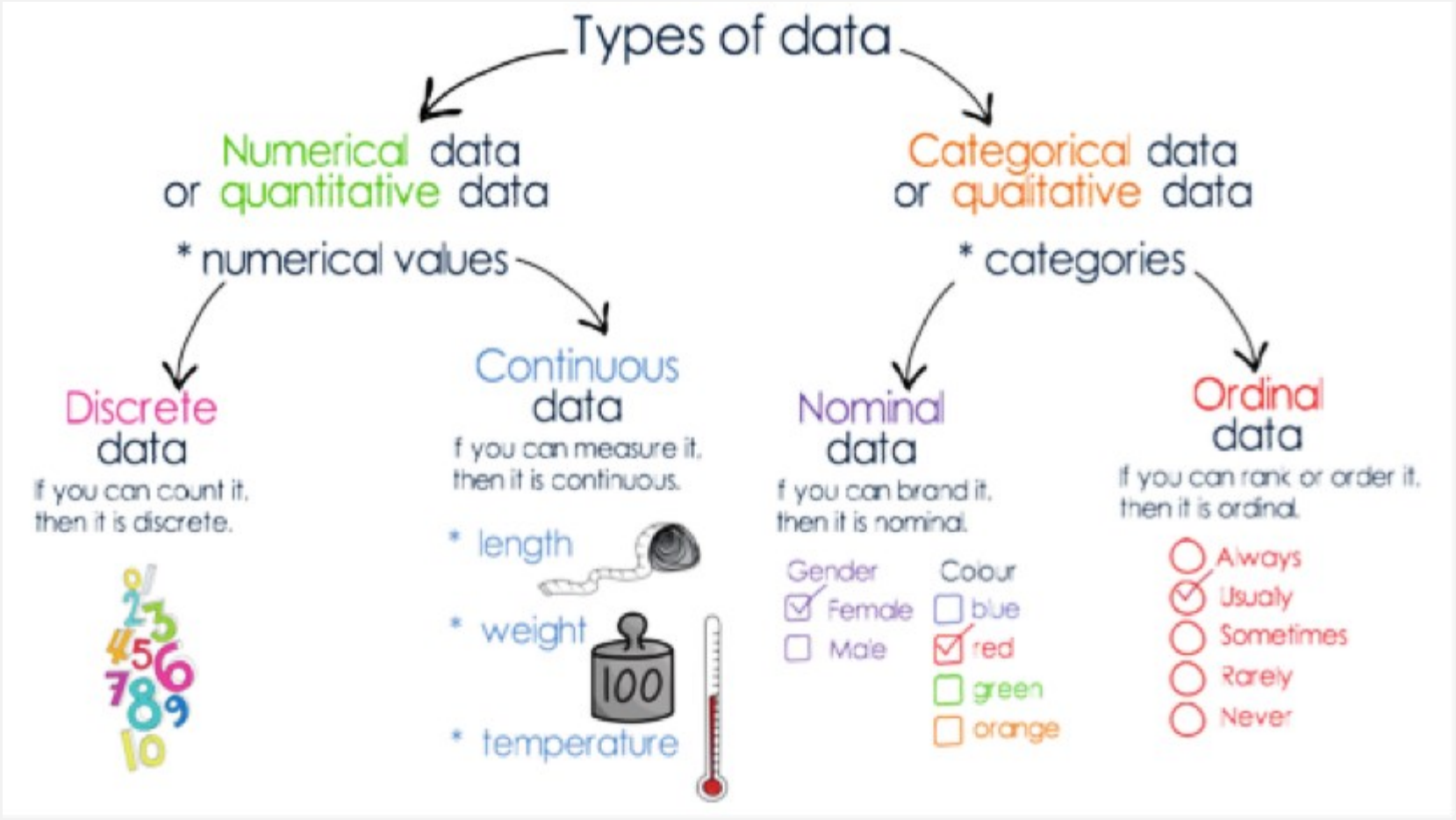
Objects

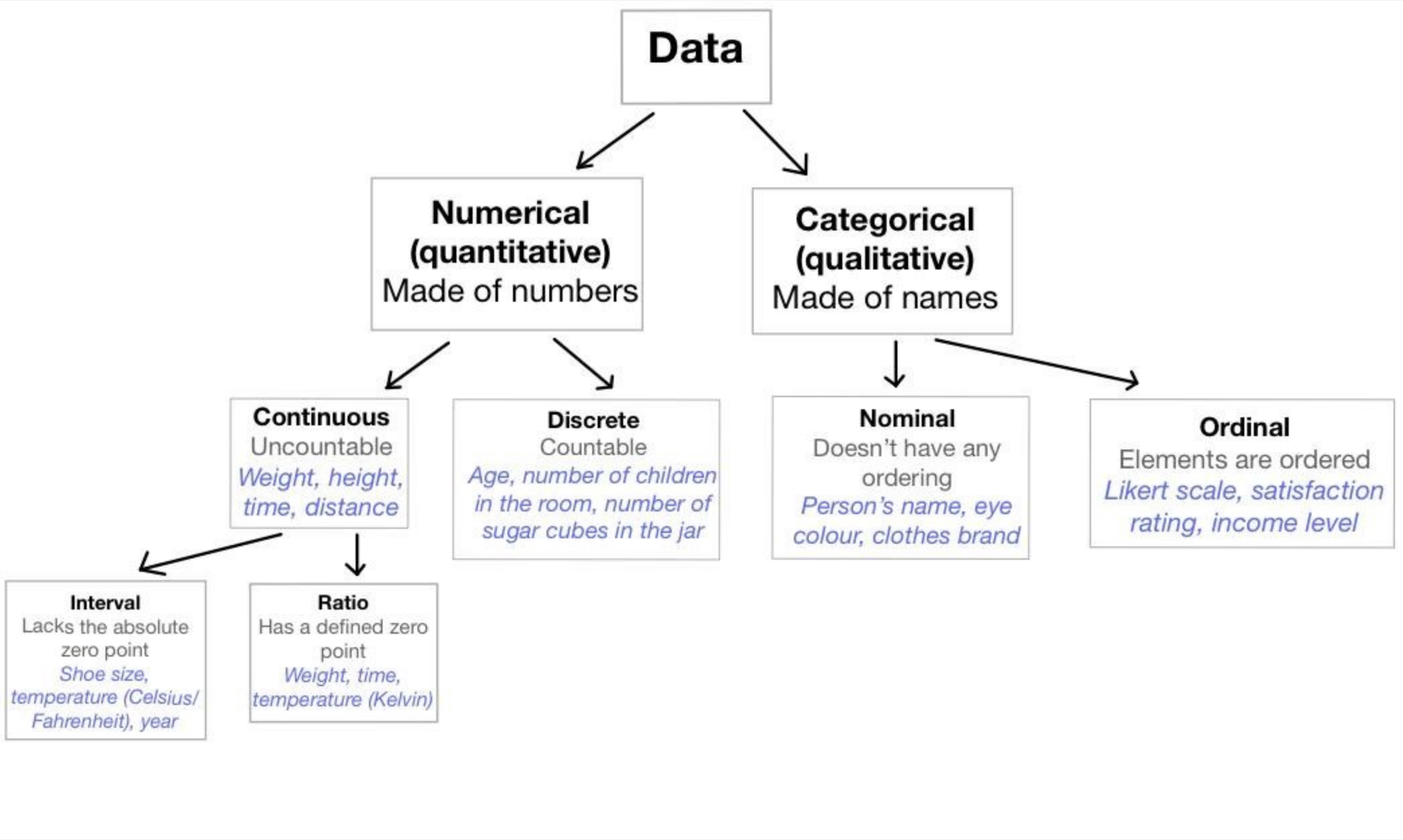
Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Size: Number of objects
Dimensionality: Number of attributes
Sparsity: Number of populated object-attribute pairs

ATTRIBUTES/VARIABLES/FEATURES

- There are different types of attributes
  - Categorical
    - Examples: eye color, zip codes, words, rankings (e.g, good, fair, bad), height in {tall, medium, short}
    - Nominal (no order or comparison) vs Ordinal (order but not comparable)
  - Numeric
    - Examples: temperature, time, length, value, count.
    - Discrete (counts) vs Continuous (temperature)
    - Special case: Binary attributes (yes/no, exists/not exists)


Types of data
Numerical data
or quantitative data
* numerical values
Discrete
data
If you can count it,
then it is discrete.
Continuous
data
If you can measure it,
then it is continuous.
* length
* weight
* temperature
100
Categorical data
or qualitative data
* categories
Nominal
data
If you can brand it,
then it is nominal.
Gender
Female
Male
Colour
blue
red
green
orange
Ordinal
data
If you can rank or order it,
then it is ordinal.
Always
Usualy
Sometimes
Rarely
Never


Data
Numerical (quantitative)
Made of numbers
Categorical (qualitative)
Made of names
Continuous
Uncountable
Weight, height, time, distance
Discrete
Countable
Age, number of children in the room, number of sugar cubes in the jar
Nominal
Doesn't have any ordering
Person's name, eye colour, clothes brand
Ordinal
Elements are ordered
Likert scale, satisfaction rating, income level
Interval
Lacks the absolute zero point
Shoe size, temperature (Celsius/ Fahrenheit), year
Ratio
Has a defined zero point
Weight, time, temperature (Kelvin)

NUMERIC RECORD DATA

- If data objects have the same fixed set of numeric attributes, then the data objects can be thought of as points in a multi-dimensional space, where each dimension represents a distinct attribute

- Such data set can be represented by an n-by-d data matrix, where there are n rows, one for each object, and d columns, one for each attribute

Projection of x Load	Projection of y load	Distance	Load	Thickness
10.23	5.27	15.22	2.7	1.2
12.65	6.25	16.22	2.2	1.1

CATEGORICAL DATA

- Data that consists of a collection of records, each of which consists of a fixed set of categorical attributes

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	High	No
2	No	Married	Medium	No
3	No	Single	Low	No
4	Yes	Married	High	No
5	No	Divorced	Medium	Yes
6	No	Married	Low	No
7	Yes	Divorced	High	No
8	No	Single	Medium	Yes
9	No	Married	Medium	No
10	No	Single	Medium	Yes

ORDERED DATA

- Genomic sequence data

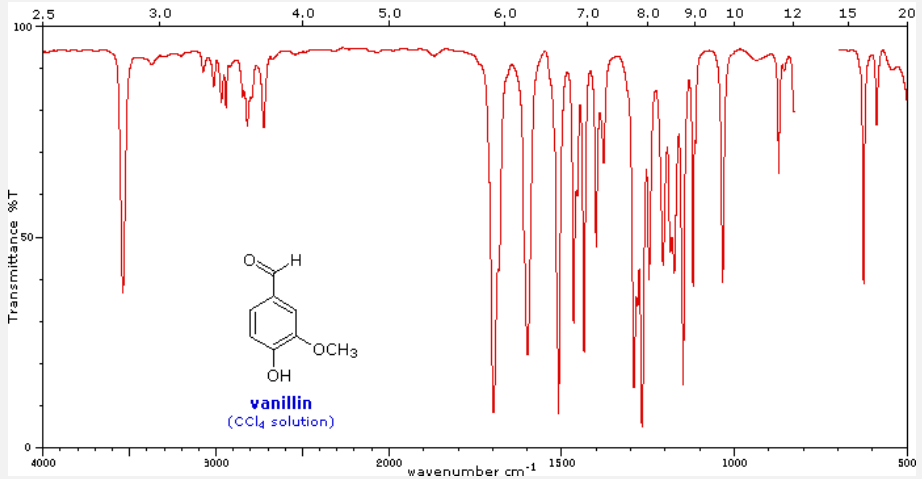
```
GGTTCCGCCTTCAGCCCCGCGCC
CGCAGGGCCCGCCCCGCGCCGTC
GAGAAGGGCCCGCCTGGCGGGCG
GGGGGAGGCGGGGCCGCCCGAGC
CCAACCGAGTCCGACCAGGTGCC
CCCTCTGCTCGGCCTAGACCTGA
GCTCATTAGGCGGCAGCGGACAG
GCCAAGTAGAACACGCGAAGCGC
TGGGCTGCCTGCTGCGACCAGGG
```

- Data is a long ordered string

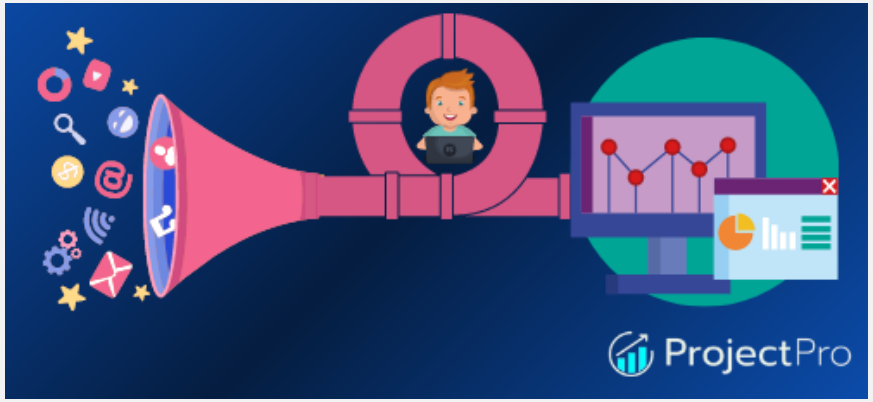
ORDERED DATA

- Time series
  - Sequence of ordered (over “time”) numeric values.

- Spectra
  - Sequence of ordered (over «wavenumber») numeric values of transmittance



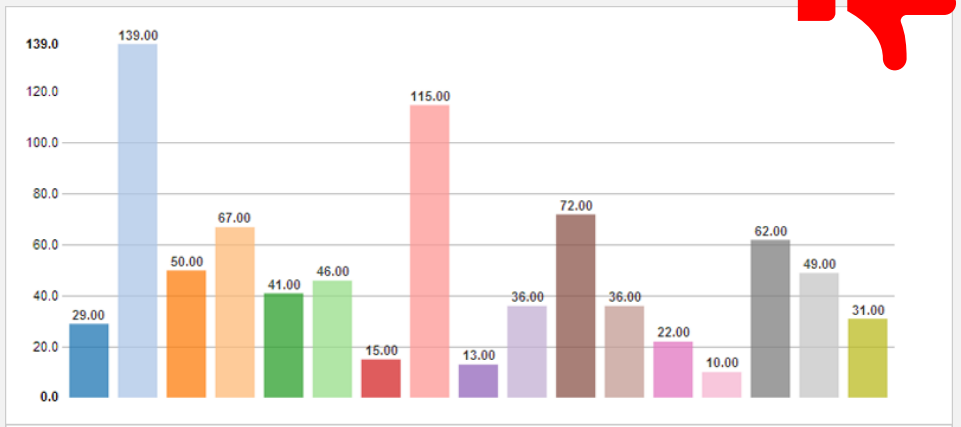
DATA ANALYSIS PIPELINE

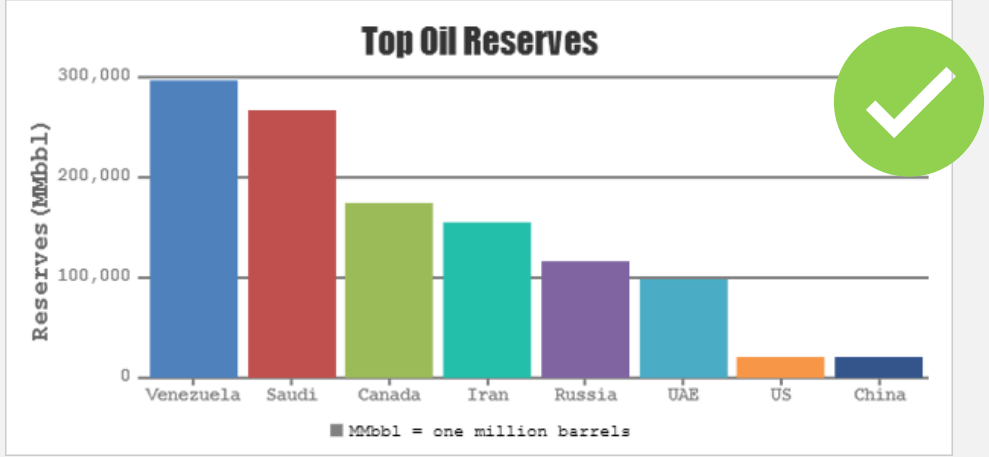


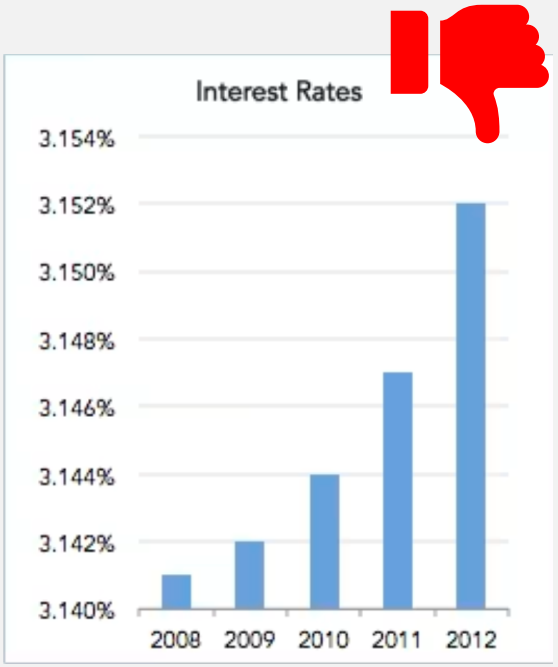
- Mining is not the only step in the analysis process

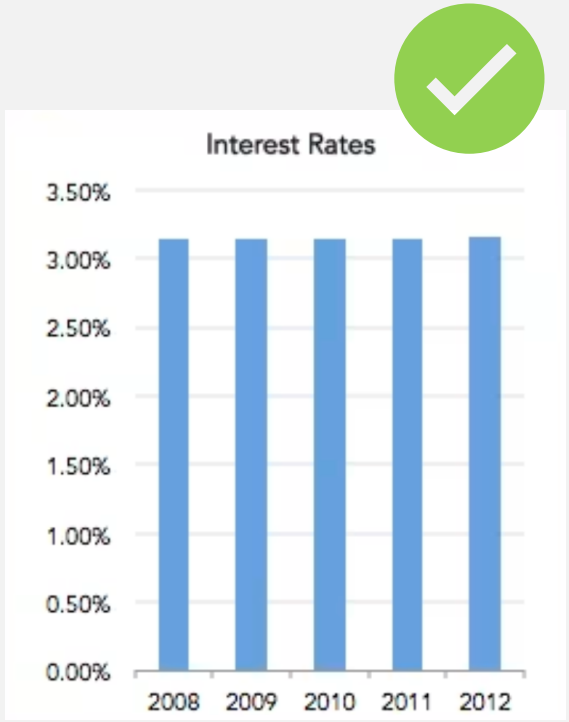
- Preprocessing: real data is noisy, incomplete and inconsistent. Data cleaning is required to make sense of the data
  - Techniques: Sampling, Dimensionality Reduction, Feature selection.
  - A dirty work, but it is often the most important step for the analysis.
- Post-Processing: Make the data actionable and useful to the user
  - Statistical analysis of importance
  - Visualization.
- Pre- and Post-processing are often data mining tasks as well

GOOD AND BAD VISUALIZATION

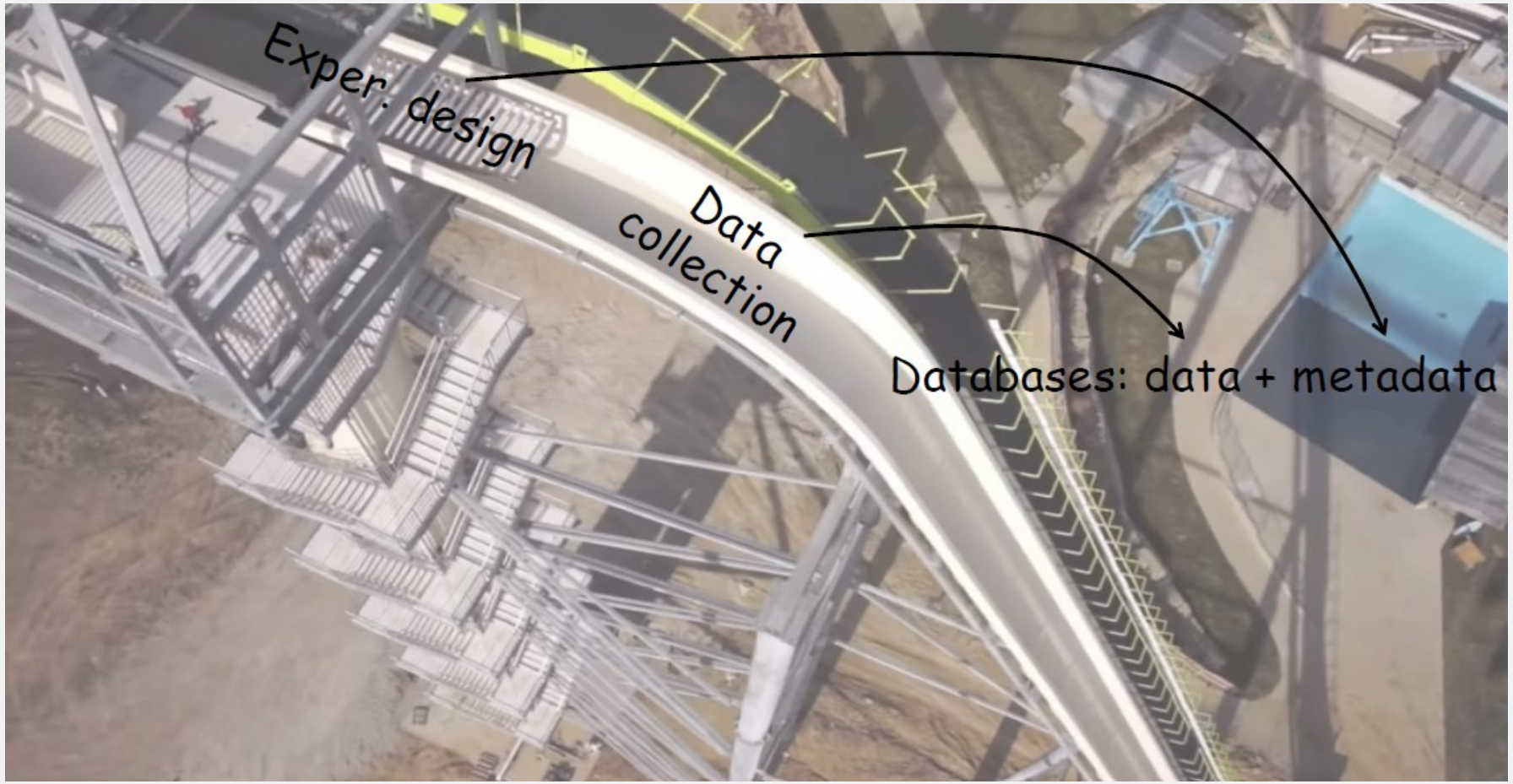








DATA ANALYSIS 'SLIDE'/'PIPELINE'



SAMPLING

- Sampling is the main technique employed for data selection.
  - It is often used for both the preliminary investigation of the data and the final data analysis.
- Statisticians sample because obtaining the entire set of data of interest is too expensive or time consuming.
- Sampling is used in data mining because processing the entire set of data of interest is too expensive or time consuming.

SAMPLING AND SAMPLE SIZE

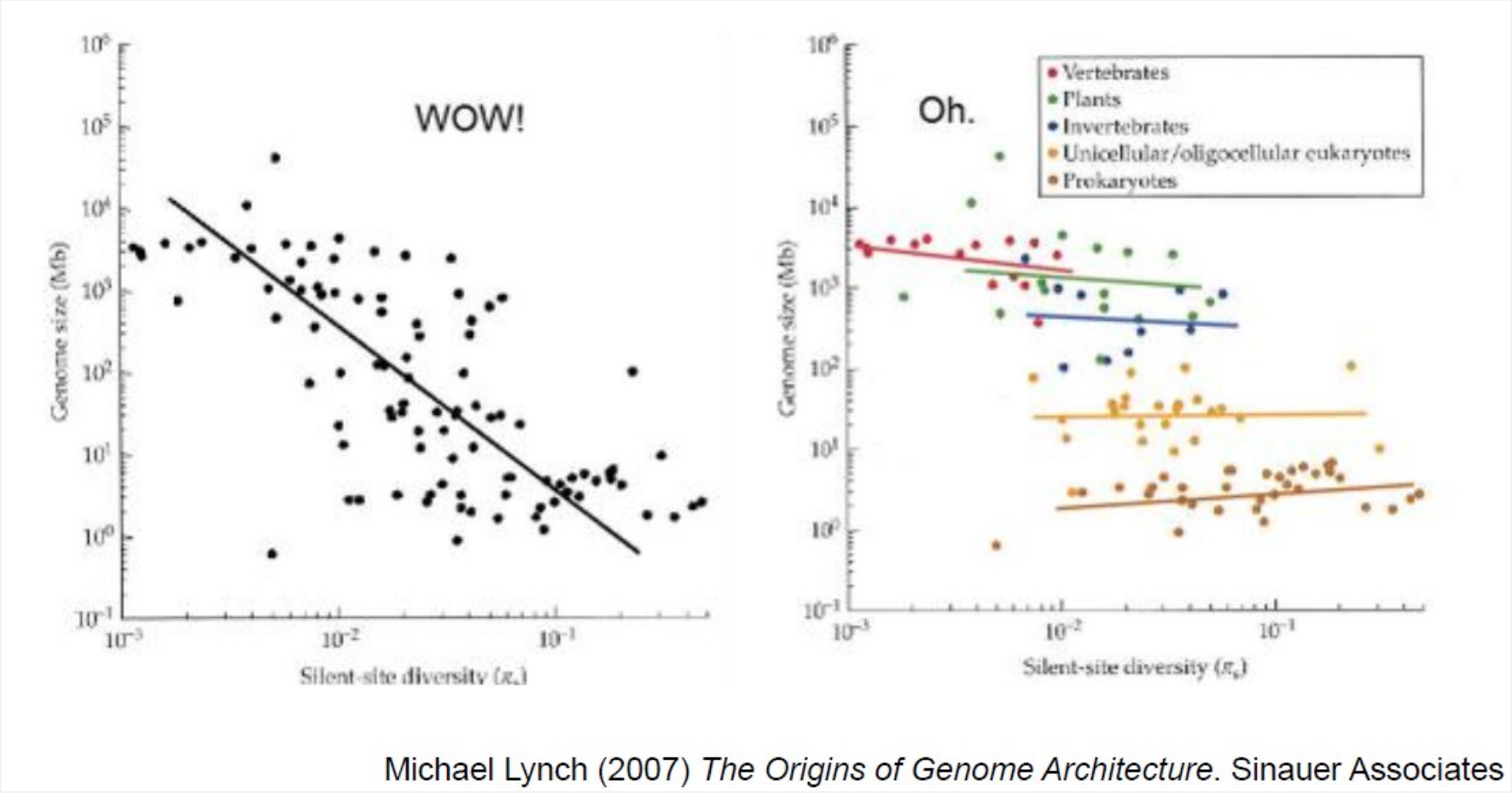
- The key principle for effective sampling is the following:
  - using a sample will work almost as well as using the entire data sets, if the sample is representative
  - A sample is representative if it has approximately the same property (of interest) as the original set of data

8000 points

2000 Points

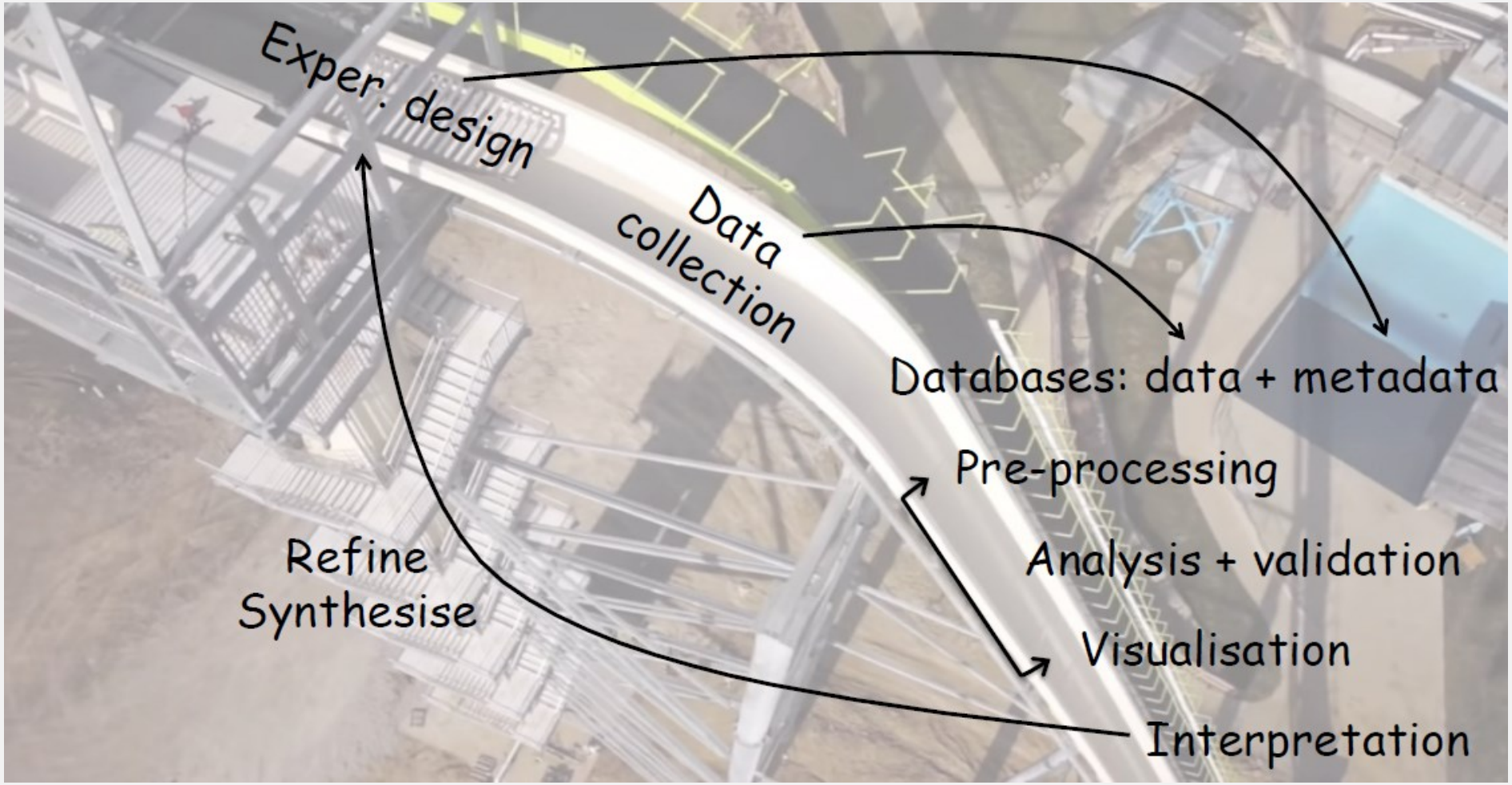
500 Points

CAPTURING METADATA IS IMPORTANT! ARE YOU AWARE OF ALL THE FACTORS?



Michael Lynch (2007) *The Origins of Genome Architecture*. Sinauer Associates

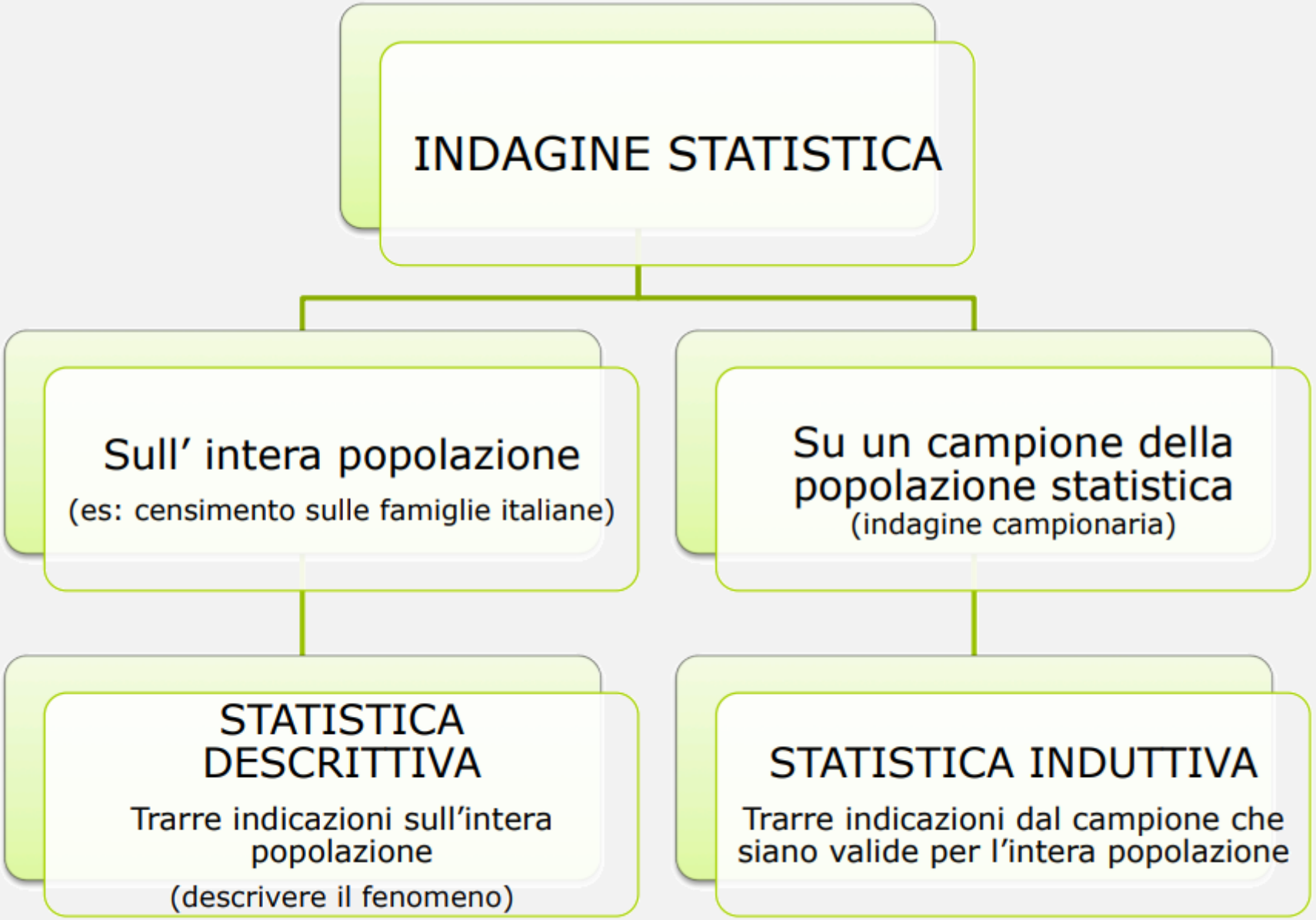
DATA ANALYSIS 'SLIDE'/'PIPELINE'



- Statistica: raccolta di metodi e strumenti matematici atti ad organizzare una o più serie di dati che descrivono una categoria di fatti

  - È la scienza che studia i fenomeni collettivi o di massa.
  - Esempi: numero di componenti delle famiglie di una data area geografica, l'età dei cittadini di un certo paese, la lunghezza delle foglie di un tipo di pianta, la durata delle lampadine di una certa marca,…

- La statistica insegna a individuare i modi in cui un fenomeno si manifesta, a descriverlo sinteticamente, e a trarne da esso conclusioni più generali di fenomeni più ampi.


INDAGINE STATISTICA
Sull' intera popolazione
(es: censimento sulle famiglie italiane)
Su un campione della popolazione statistica
(indagine campionaria)
STATISTICA DESCRITTIVA
Trarre indicazioni sull'intera popolazione
(descrivere il fenomeno)
STATISTICA INDUTTIVA
Trarre indicazioni dal campione che siano valide per l'intera popolazione

Popolazione, unità, campione statistico

- **Popolazione statistica**: insieme degli elementi a cui si riferisce l'indagine statistica:
  - opinione degli americani riguardo una nuova elezione presidenziale: tutti i cittadini USA
  - geni sovra-espressi nelle persone che soffrono di obesità: tutte le persone obese
  - …

- **Unità statistica**: ogni elemento della popolazione statistica, la minima unità della quale si raccolgono i dati:
  - Un cittadino, una persona obesa….

- **Campione statistico (sample)**: un qualsiasi insieme di unità statistiche prese da tutta la popolazione. Un campione è dunque un sottoinsieme di misurazioni selezionate dalla popolazione
  - 50 persone con problemi di obesità (estratte a caso).

Variabile casuale

- Il *fenomeno collettivo* si presenta secondo modalità diverse nelle varie unità statistiche, perciò lo chiameremo variabile casuale.

- Il valore assunto dalla variabile casuale in una data unità statistica lo chiameremo **osservazione**.

  - Esempio:
  - **variabile casuale**: livello di espressione del gene AAA
  - **osservazione**: il gene AAA della persona X ha un livello di espressione pari a 12.3, il gene AAA della persona Y ha un livello di espressione di 10.2, il gene AAA della persona Z….

Variabile quantitativa o qualitativa

- **Variabile quantitativa**: quando assume valori numerici:
  - **Continua**: assume valori continui in un intervallo (peso e statura di una persona, livelli di intensità dei campioni su microarray, livello di espressione genica, etc.)
  - **Discreta**: assume valori discreti come numero di campioni, numero di geni sovra-espresso, numero di pazienti, etc.

- **Variabile qualitativa**: quando assume valori non numerici
  - **Ordinale**: i dati sono in un ordine (buono-medio-cattivo, freddo-tiepido-caldo…)
  - **Categorica**: uomo/donna, fenotipo, gruppi di pazienti malati/sani, etc.

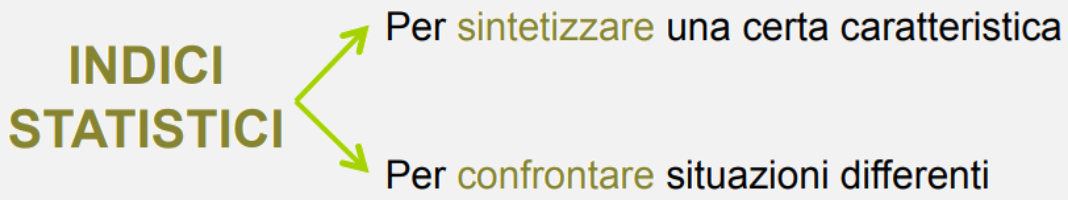
La data table

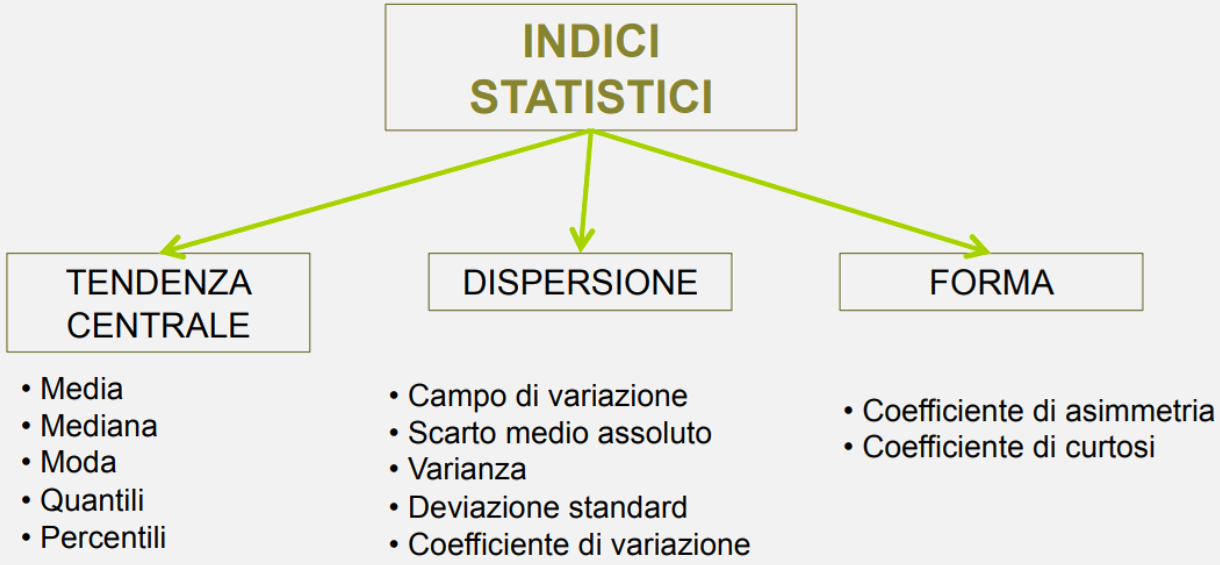
- I dati codificati di una rilevazione statistica effettuata su n **unità statistiche** con riferimento a p **variabili**, vengono raccolti in una tabella che viene chiamata "**matrice dei dati**"

N.	Sesso	Titolo di studio	Età	Peso	N. Ricoveri
1	M	Licenza media inferiore	36	65	3
2	F	Laurea	45	70	1
…	…	…	…	…	…
N	F	Diploma	60	55	6

Analisi dei dati

- Quando i dati sono molti, l'analisi diretta della matrice non consente di cogliere in via immediata gli aspetti salienti del fenomeno. Occorre perciò ottenere una sintesi attraverso un'**elaborazione statistica dei dati**





Frequenze assolute e percentuali

- Quando il campione di cui vogliamo descrivere le variabili statistiche è molto grande, anziché considerare tutti i valori, si possono scrivere solo valori distinti e riportare, per ogni valore, quante volte compare.

Sia $Y = (y_1, y_2, \dots, y_N)$ una variabile statistica discreta. Definiamo **modalità** i valori distinti tra $y_1, \dots, y_N$ e **frequenza assoluta** di una modalità il numero di volte che viene osservata nell'espressione della variabile statistica.

$$Y = (60, 80, 92, 100, 83, 84, 96, 74, 63, 80, 100, 90, 75, 74, 92)$$

in termini di modalità e frequenze assolute, attraverso la seguente tabella:

Modalità	60	80	92	100	83	84	96	74	63	90	75
Frequenza assoluta	1	2	2	2	1	1	1	2	1	1	1

Ovviamente la somma di tutte le frequenze assolute dà la numerosità N del campione:

$$N = 1 + 2 + 2 + 2 + 1 + 1 + 1 + 2 + 1 + 1 + 1 = 15$$

Frequenze assolute e percentuali

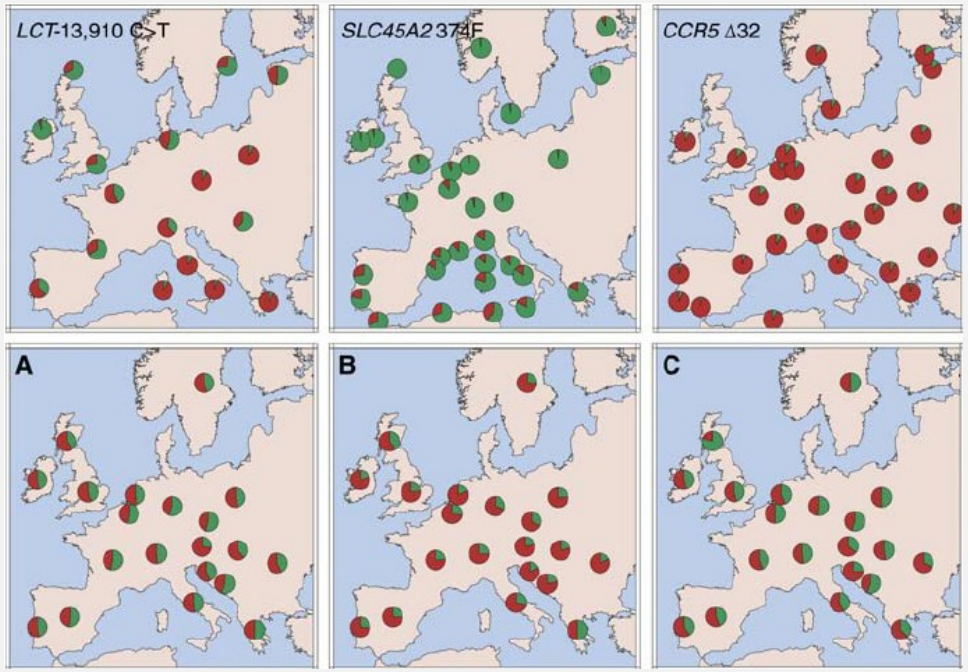
Siano Y una variabile statistica e f la frequenza assoluta della modalità z. Definiamo **frequenza relativa** della modalità z il rapporto f/N, ove N è il numero di elementi del campione. La **frequenza percentuale** è data dalla frequenza relativa moltiplicata per 100.

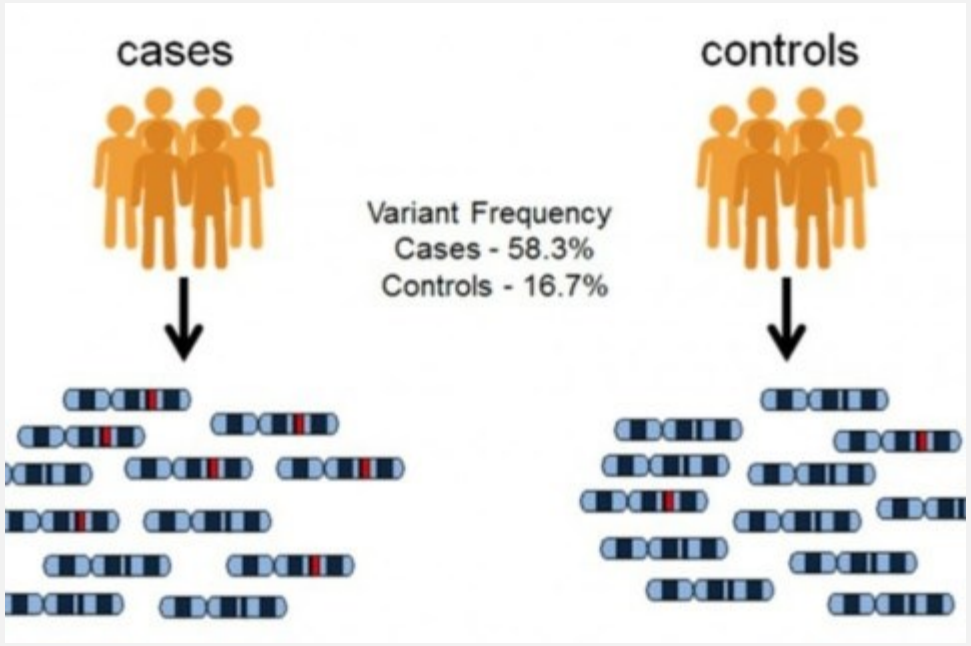
Supponiamo di avere la variabile statistica corrispondente al voto dell'esame di Matematica e Statistica per un campione di 300 studenti. Anziché rappresentare la variabile statistica attraverso i suoi 300 valori (cioè i voti conseguiti dagli studenti), come visto in precedenza, utilizziamo una separazione in classi, dove abbiamo stabilito, arbitrariamente, di considerare intervalli di 2 punti. Esprimiamo dunque i dati nella seguente tabella:

Classi	$[18, 20)$	$[20, 22)$	$[22, 24)$	$[24, 26)$	$[26, 28)$	$[28, 30]$
F. A.	80	55	65	40	35	25
F. R.	$0.2\overline{6}$	$0.18\overline{3}$	$0.21\overline{6}$	$0.1\overline{3}$	$0.11\overline{6}$	$0.08\overline{3}$

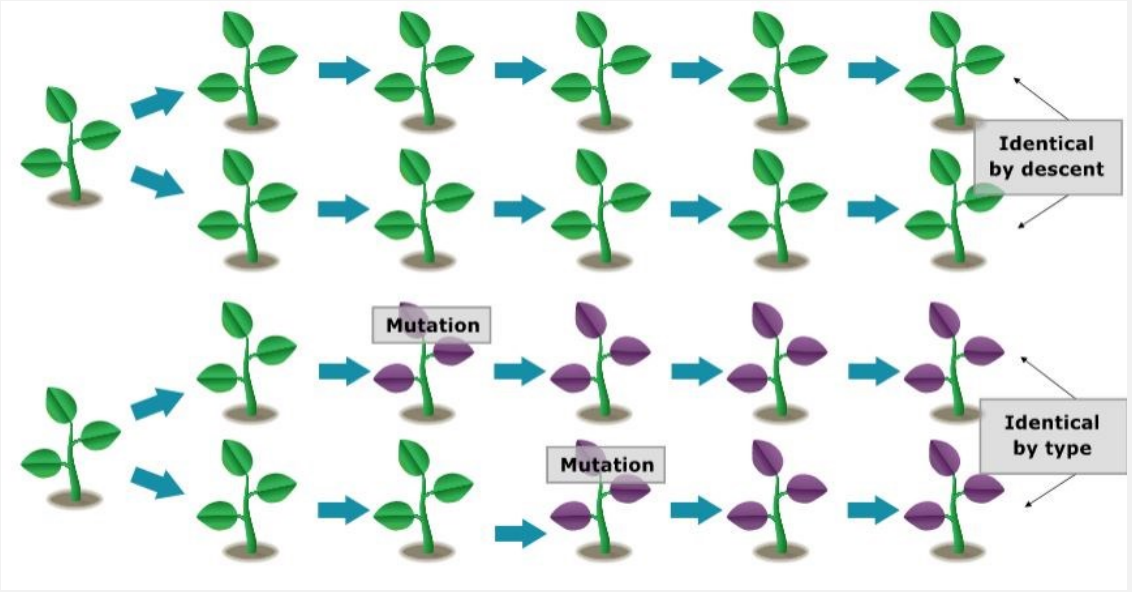
Dalla tabella possiamo leggere subito, per esempio, che 40 studenti hanno ottenuto un voto compreso tra 24 e 26, e rappresentano il 13% circa del campione considerato.

Possiamo anche vedere che 100 studenti, cioè il 33% circa del campione, ha conseguito un voto uguale o superiore a 24/30. Tale dato si può ottenere rapidamente facendo la somma delle frequenze assolute, rispettivamente, relative per le classi individuate dagli intervalli: $[24, 26)$, $[26, 28)$, $[28, 30]$.





Maps showing the frequency distribution of individual genetic variants



Media

- **Media di una popolazione**: somma di tutti i valori delle variabili della popolazione diviso il numero di unità della popolazione (N)

$$\mu = \frac{\sum_{i=1}^{N} X_i}{N}$$

Dove N è il numero di elementi della popolazione, X_i è la i-esima osservazione della variabile X_i

- **Media di un campione**: somma di tutti i valori delle variabili di un sottoinsieme della popolazione diviso il numero di unità di tale campione (n)

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

Dove n è il numero di elementi del campione della popolazione, X_i è la i-esima osservazione della variabile X_i

Dato il seguente set di misurazioni di livello di espressione dei geni:

55.20	18.06	28.16	44.14	61.61	4.88	180.29	399.11	97.47	56.89	271.95	365.29	807.80

Media della popolazione:

$$\mu = \frac{\sum_{i=1}^{13} 55.20 + 18.06 + 28.16 + 44.14 + 61.61 + \ldots + \ldots + 807.80}{13} = \frac{2390{,}85}{13} = 183.9115$$

Media del campione (55.20; 18.06; 28.16; 44.14):

$$\overline{X} = \frac{55.20 + 18.06 + 28.16 + 44.14}{4} = \frac{145.56}{4} = 36.39$$

La media di qualsiasi campione $\bar{X}$ può essere molto diversa da quella dell'intera popolazione μ . Più è numeroso il campione, più la media del campione sarà vicina a quella della popolazione

Media

- **Media ponderata di una popolazione**: si assegna ad ogni variabile un peso; si sommano tutti i valori delle variabili, moltiplicate per il peso, e si divide il numero ottenuto per la somma dei pesi

$$\mu = \frac{\sum_{i=1}^{N} p_i X_i}{\sum_{i=1}^{N} p_i}$$

- Il valore atteso di una variabile X, indicato con $E[X]$ è definito come la media di X calcolata su un grande numero di esperimenti

Esempio

Esame	*Voto*	*Crediti (cfu)*
Economia politica	21	7
Ragioneria	25	10
Diritto commerciale	26	6
Matematica	24	5
...	...	...

$$media.p = \frac{(voto \times cfu) + (voto \times cfu) + (...)}{(cfu + cfu + ...)}$$

$$media.p = \frac{(21 \times 7) + (25 \times 10) + (26 \times 6) + (24 \times 5)}{(7 + 10 + 6 + 5)} = \boxed{24.04}$$

Moda

- La **moda** è il valore più frequente di una distribuzione, o meglio, la <u>modalità più ricorrente della variabile</u> (cioè quelle a cui corrisponde la frequenza più elevata).

962	1005	1003	768	980	965	1030	1005	975	989	955	783	1005

La moda di questo campione è 1005, in quanto compare 3 volte.

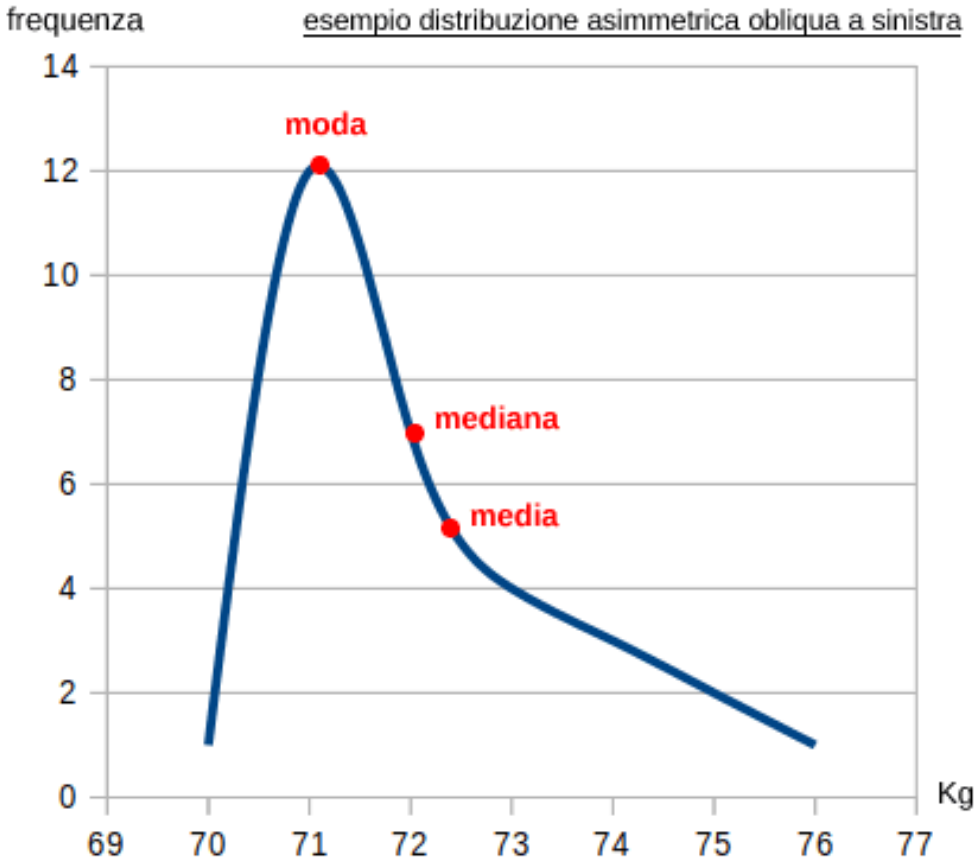
Caratteristiche della moda:

- viene utilizzata solamente a scopi descrittivi, perché è meno stabile e meno oggettiva delle altre misure di tendenza centrale
- per individuare la moda di una distribuzione si possono usare metodi grafici, come istogrammi
- può differire nella stessa serie di dati, quando si formano classi di distribuzione (intervalli) con ampiezza differente
- per individuare la moda entro una classe di frequenza, non conoscendo come i dati sono distribuiti, si ricorre all'ipotesi della ripartizione uniforme.

kg	frequenza
70	1
71	12
72	7
73	4
74	3
75	2
76	1

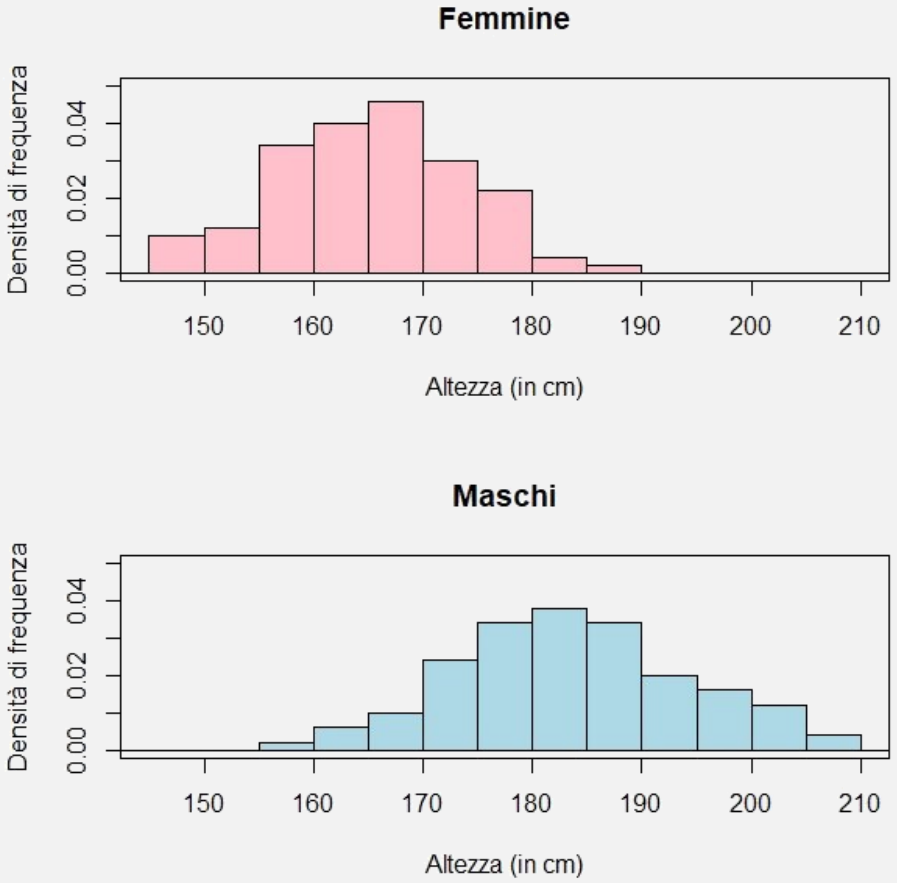
moda	71
mediana	72
media	72,2

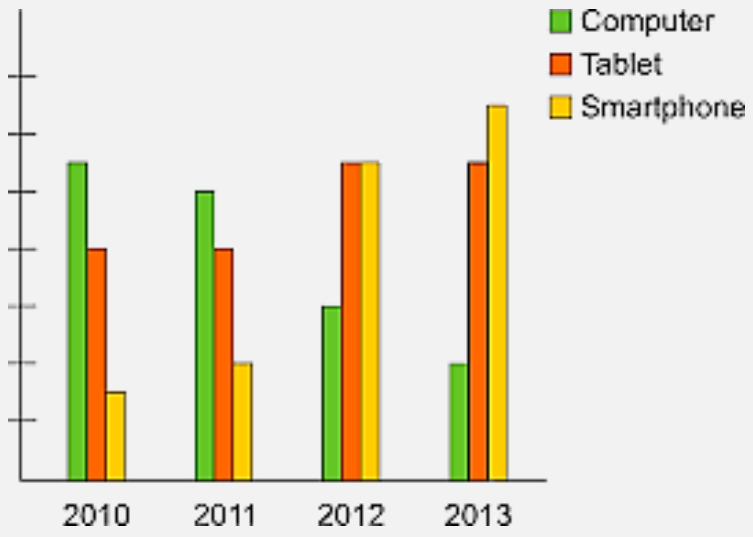
WWW.ANDREAMININI.ORG



Istogrammi

- Un istogramma descrive la frequenza relativa dei dati compresi in un intervallo (a, b) ed è utilizzato per visualizzare la distribuzione dati.





Distribuzioni unimodali/bimodali

- Una distribuzione può presentare più mode:
  - Distribuzioni **unimodali**: distribuzioni di frequenza che hanno una sola moda, ossia un solo un punto di massimo (che rappresenta sia il massimo relativo che il massimo assoluto)
  - Distribuzioni **bimodali o k-modali**: distribuzioni di frequenza che presentano due o più mode, ossia che hanno due (o k) massimi relativi
    - Esempio: misurando le altezze di un gruppo di giovani in cui la parte maggiore sia formata da femmine e la minore da maschi si ottiene una distribuzione bimodale, con una moda principale ed una secondaria.

Distribuzione zeromodale

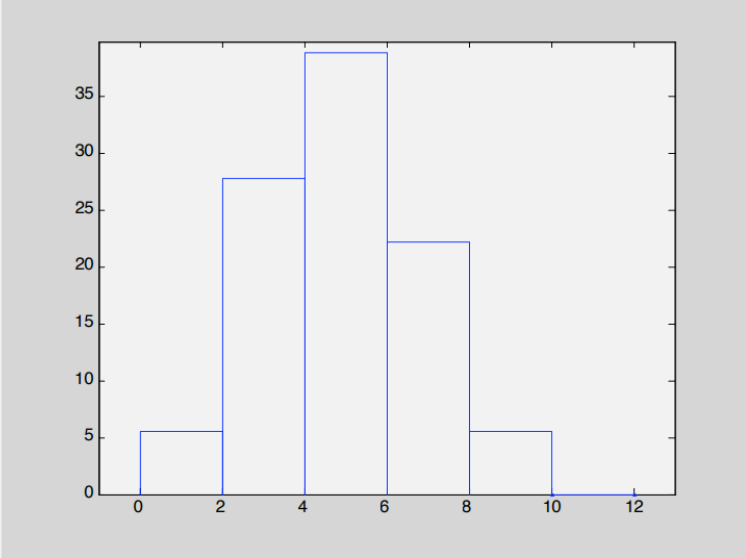
- Nessun valore ha una frequenza più elevata degli altri:

$$A = \{1, 1, 1, 2, 2, 2, 3, 3, 3, 4, 4, 4, 5, 5, 5, 6, 6, 6\}$$

Distribuzione unimodale

C’è un solo valore con una frequenza più elevata degli altri

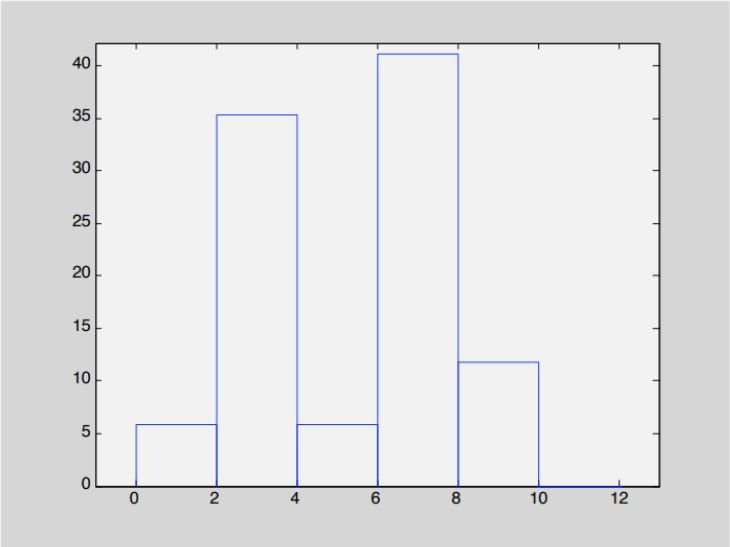
A = {1, 2, 2, 3, 3, 3, 4, 4, 4, 4, 5, 5, 5, 6, 6, 7, 7, 8}

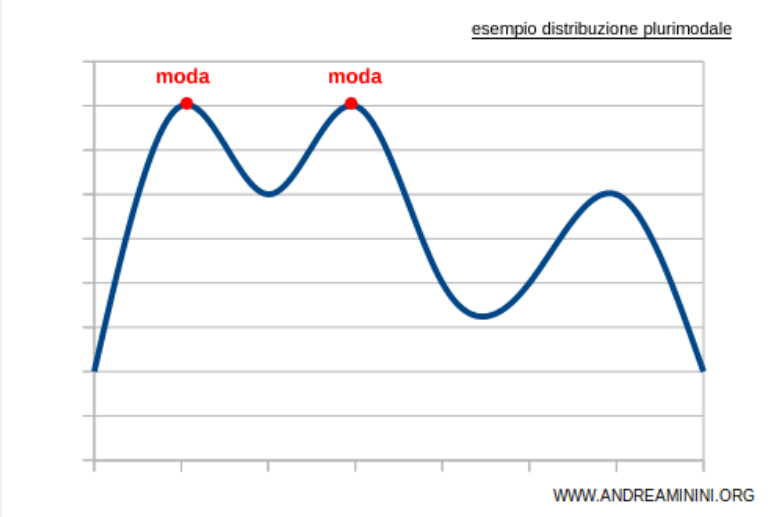


Distribuzione bimodale

Ci sono due valori con una frequenza più elevata degli altri.

A = {1, 2, 2, 3, 3, 3, 3, 5, 6, 6, 6, 6, 6, 7, 7, 8, 8}



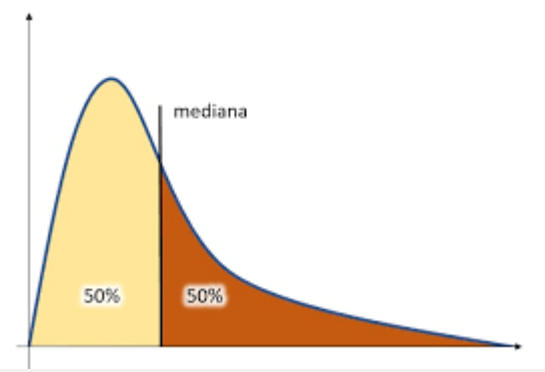


Mediana

- La **mediana** è il valore che occupa la posizione centrale in un insieme ordinato di dati.

- È una misura robusta, in quanto poco influenzata dalla presenza di dati anomali.

- Caratteristiche:
  - si ricorre al suo uso quando si vuole attenuare l'effetto di valori estremi;
  - in una distribuzione o serie di dati, ogni valore estratto a caso ha la stessa probabilità di essere inferiore o superiore alla mediana.



Calcolo della Mediana

Per calcolare la mediana di un gruppo di dati, bisogna:

1. disporre i valori in ordine crescente oppure decrescente e contare il numero totale n di dati;
2. se il numero (n) di dati è dispari, la mediana corrisponde al valore numerico del dato centrale, quello che occupa la posizione $(n+1)/2$;
3. se il numero (n) di dati è pari, la mediana è stimata utilizzando i due valori centrali che occupano le posizioni $n/2$ e $n/2+1$:
   a. con poche osservazioni, come mediana viene assunta la media aritmetica di queste due osservazioni intermedie;
   b. con molte osservazioni raggruppate in classi, si ricorre talvolta alle proporzioni.

Esempio.
Consideriamo il seguente campione:

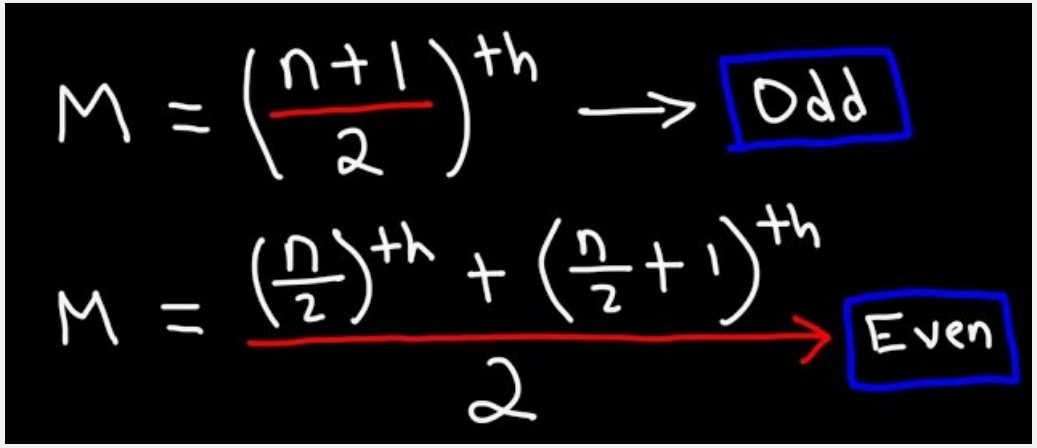
96	78	90	62	73	89	92	84	76	86

1. Ordiniamo i campioni in ordine crescente:

62	73	76	78	**84**	**86**	89	90	92	95

2. Dal momento che il numero di campioni è pari ($n = 10$), la mediana è calcolata come la media dei due elementi centrali:

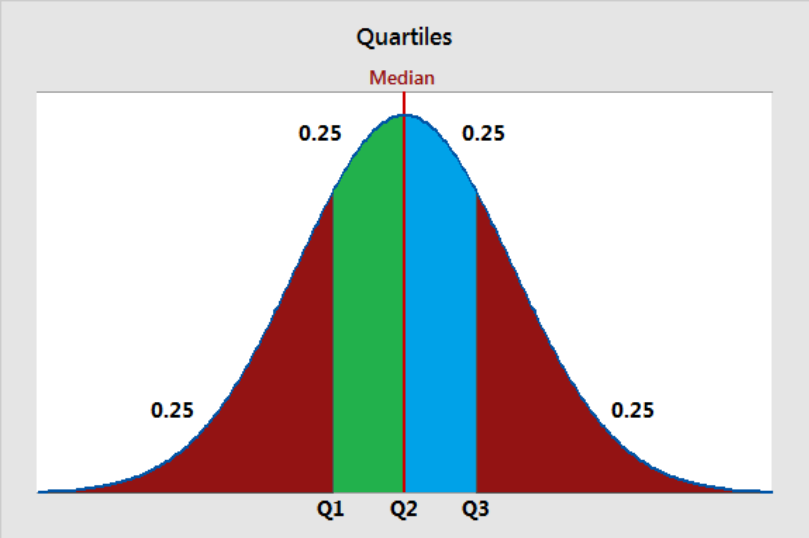
$$mediana = \frac{84 + 86}{2} = 85$$



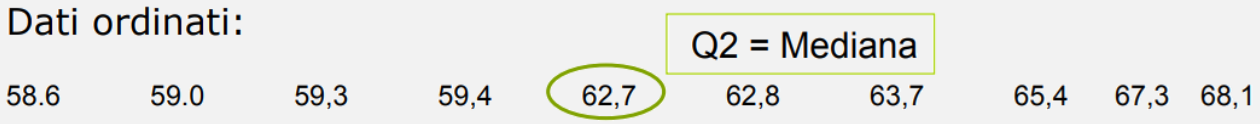
Quantili

- I **quantili** sono una famiglia di misure, a cui appartiene anche la mediana, che si distinguono a seconda del numero di parti uguali in cui suddividono una distribuzione.

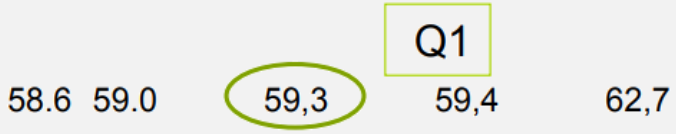
- I **quartili** ripartiscono la distribuzione in 4 parti di pari frequenza, dove ogni parte contiene la stessa frazione di osservazioni:
  - Il **primo quartile** è definito come il numero $q1$ per il quale il 25% dei dati statistici è minore o uguale a $q1$.
  - Il **secondo quartile** è definito come il numero $q2$ per il quale il 50% dei dati statistici è minore o uguale a $q2$. Il secondo quartile corrisponde alla mediana.
  - Il **terzo quartile** è definito come un numero $q3$ per il quale il 75% dei dati statistici è minore o uguale a $q3$.



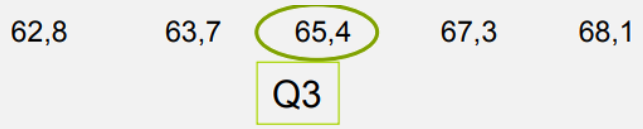
Esempio.
Consideriamo uno studio che esamina i tempi d'attesa al ristorante in un campione di 10 clienti:



La mediana è pari a 62,75
Si considera la metà inferiore dei dati, ovvero tutti i valori inferiori alla mediana e su questo sottoinsieme di dati si calcola la mediana, il valore trovato è Q1

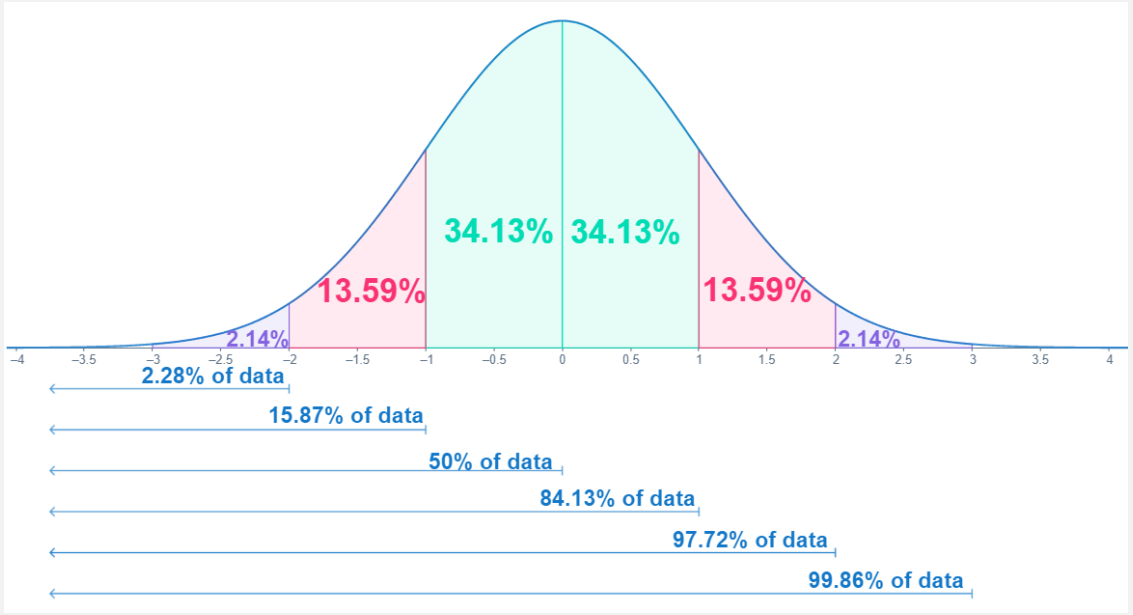


Si considera la metà superiore dei dati, ovvero tutti i valori superiori della mediana e su questo sottoinsieme di dati si calcola la mediana il valore trovato è Q3



Decili e percentili

- In modo analogo, si definiscono:
  - **Decili**: 9 punti che dividono la distribuzione ordinata in 10 parti uguali
  - **Percentili**: 99 punti che dividono la distribuzione ordinata in 100 parti uguali



Campo di variazione (o «range»)

- Il **campo di variazione** di una distribuzione è la differenza tra il dato più grande e quello più piccolo della distribuzione:

$$C = x_{max} - x_{min}$$

- Questo indice è abbastanza grossolano non dicendo nulla sulla variabilità dei dati intermedi.

  - Esempio: il campo di variazione della seguente distribuzione:

$$25 - 26 - 28 - 29 - 30 - 32 \rightarrow C = 32 - 25 = 7$$

Scarto

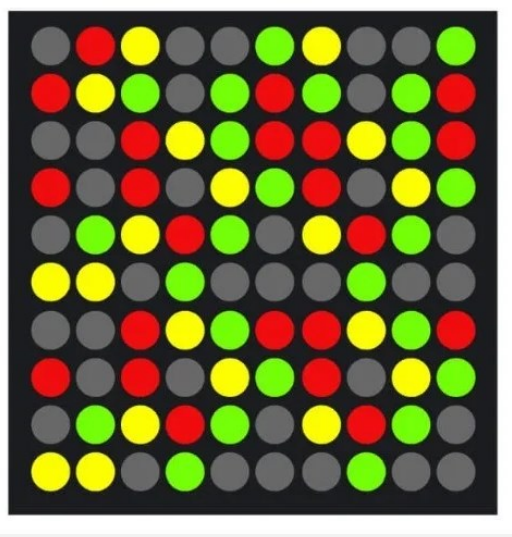
- Lo **scarto** misura quanto ciascun dato x_i si discosta dal valor medio, ovvero $s = x_i - \bar{X}$

  - Esempio: Consideriamo le seguenti intensità rilevate dagli spot dei microarray:

$$435.02, 678.14, 235.35, 956.12, \dots, 1127.82, 456.43$$

  - La media di questi valori è: 515.13; i loro scarti sono:

$$435.02 - 515.13 = -80.11$$
$$678.14 - 515.13 = 163.01$$
$$235.35 - 515.13 = -279.78$$
$$956.12 - 515.13 = 440.99$$



Scarto assoluto

Usando s possono essere ricavati diversi altri indici di variabilità

- Si chiama **scarto medio assoluto** e si indica con s_m la media aritmetica dei valori assoluti degli scarti

$$s_m = \frac{\sum_{i=1}^{n} |x_i - \bar{X}|}{n}$$

Varianza

- **Varianza della popolazione**: misura che caratterizza molto bene la variabilità di una popolazione. Indicatore della dispersione di una variabile o distribuzione statistica che ottengo calcolando la media dei quadrati degli scarti della media aritmetica (μ).

$$\sigma^2 = \frac{\sum_{i=1}^{N}(X_i - \mu)^2}{N}$$

Dove N è il numero di osservazioni dell'intera popolazione; μ è la media della popolazione; x_i è l'i-esimo dato statistico osservato

- **Varianza di un campione**:

$$s^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}$$

Dove n è il numero di osservazioni del campione; $\bar{X}$ è la media del campione; x_i è l'i-esimo dato statistico osservato

Quando n è grande, le differenze fra le due formule sono minime; quando n è piccolo, le differenze sono sensibili.

Esempio.

Consideriamo il seguente campione di osservazioni:

$$\{2,3,6,9,15\}$$

Calcolo della media:

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{2+3+6+9+15}{5} = 7$$

Calcolo della varianza campionaria:

$$s^2 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{n-1} = \frac{(2-7)^2 + (3-7)^2 + (6-7)^2 + (9-7)^2 + (15-7)^2}{4} = 27.5$$

Devianza

Nel calcolo di alcune statistiche, si ricorre alla devianza, data dal numeratore della varianza. È una misura della dispersione. Si ottiene calcolando la somma dei quadrati delle deviazioni dei dati di una distribuzione rispetto alla media.

$$Dev = \sum_{i=1}^{N} (X_i - \bar{X})^2$$

La varianza è la media della devianza, ovvero indica quanto i valori della distribuzione varia o rispetto alla media

Deviazione standard o scarto quadratico medio

- La varianza ha lo svantaggio di essere una grandezza quadratica e quindi non direttamente confrontabile con la media o con gli altri valori della distribuzione.

- Per trovare una misura espressa nella stessa unità di misura della variabile di partenza è sufficiente estrarre la radice quadrata della varianza.

- La **deviazione standard** è una misura di distanza dalla media e quindi ha sempre un valore positivo.

- È una misura della dispersione della variabile casuale intorno alla media.

Deviazione standard o scarto quadratico medio

- **Deviazione standard della popolazione**:

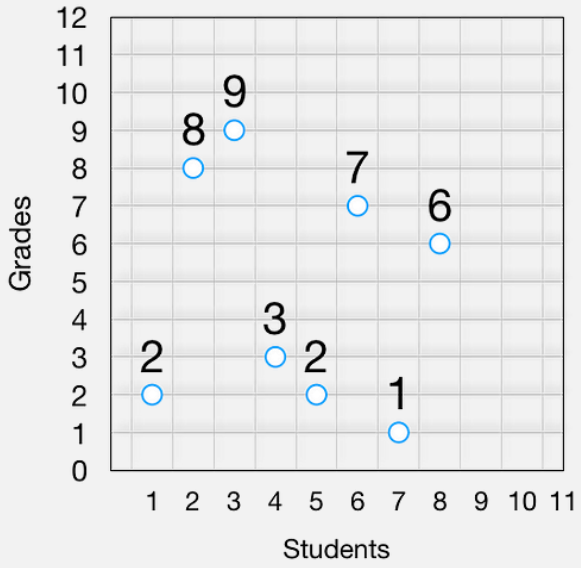
$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(X_i - \mu)^2}{N}}$$

Dove N è il numero di osservazioni dell'intera popolazione; μ è la media della popolazione; X_i è l'i-esimo dato statistico osservato.

- **Deviazione standard di un campione**:

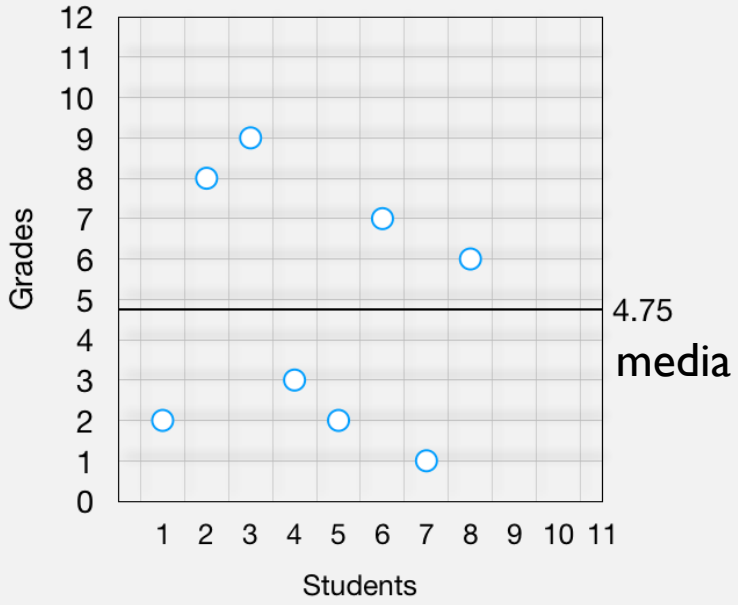
$$s = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}}$$

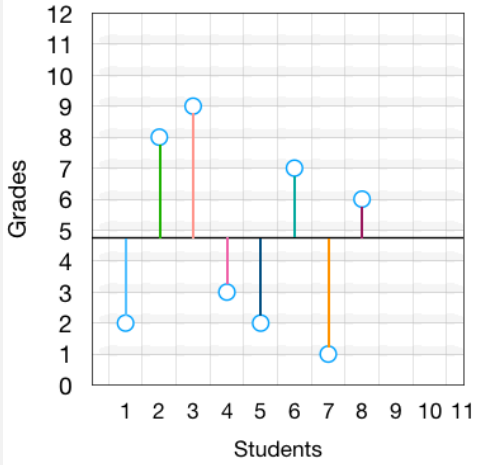
Dove n è il numero di osservazioni del campione; $\bar{X}$ è la media del campione; x_i è l'i-esimo dato statistico osservato



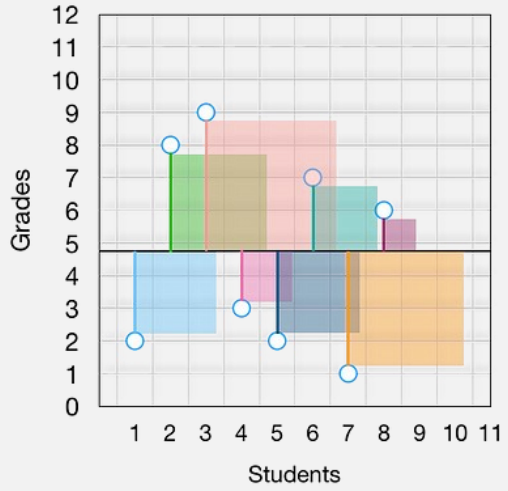
Student	Grade
1	2
2	8
3	9
4	3
5	2
6	7
7	1
8	6

$$\bar{x} = \frac{\sum_{n=1}^{N} x_n}{N} = \frac{2+8+9+3+2+7+1+6}{8} = \frac{38}{8} = 4.75$$





$$x - \bar{x} =$$
$$(2-4.75) + (8-4.75)$$
$$+ (9-4.75) + (3-4.75)$$
$$+ (2-4.75) + (7-4.75)$$
$$+ (1-4.75) + (6-4.75)$$



$$\sum (x_n - \bar{x})^2 =$$

$$7.5625 + 10.5625 + 18.0625 + 3.0625 + 7.5625 + 5.0625 + 14.0625 + 1.5625 = 67.5$$

Varianza

$$\frac{\sum (x_n - \bar{x})^2}{N} = \frac{67.5}{8} = 8.45 \text{ points}^2$$

Deviazione standard

$$\sqrt{\frac{\sum (x_n - \bar{x})^2}{N}}$$

Mean ( ) = 8.45

$\sqrt{}$ 8.45 2.91 = 2.91

Esempio.
Consideriamo i voti di due studenti:
- Anna: $30, 30, 28, 27, 26$
- Stefano: $21, 30, 30, 30, 30$

Entrambi hanno la stessa media dei voti (28.2)

Calcoliamo la deviazione standard:
- $\sigma_{Anna} = 1.78$
- $\sigma_{Stefano} = 4.02$

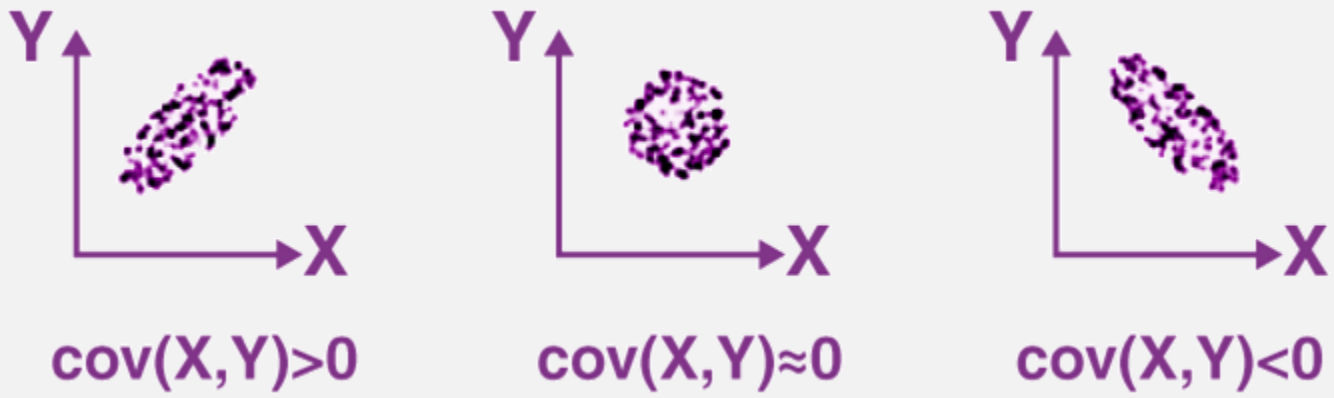
Cosa significa? I voti di Anna sono più concentrati (vicini) rispetto a quelli di Stefano

Covarianza

- Indice che consente di verificare se fra due variabili statistiche esiste un legame lineare.

- Considerando due serie $\{x_i\}$ e $\{y_i\}, i = 1,2, \dots, n$, pone a confronto le coppie di scarti $(x_i - \bar{x})$ e $(y_i - \bar{y})$:

$$Cov(X, Y) = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$



La Covarianza può essere:

- **POSITIVA**: quando X e Y variano tendenzialmente nella stessa direzione, cioè al crescere della X tende a crescere anche Y e al diminuire della X tende a diminuire anche Y
- **NEGATIVA**: quando le due variabili variano tendenzialmente in direzione opposta, cioè quando al crescere di una variabile l'altra variabile tende a diminuire (e viceversa)
- **NULLA**: quando non vi è alcuna tendenza delle 2 variabili a variare nella stessa direzione o in direzione opposta. Quando $Cov(X, Y) = 0$ si dice anche che X ed Y sono non correlate o linearmente indipendenti.

Correlazione

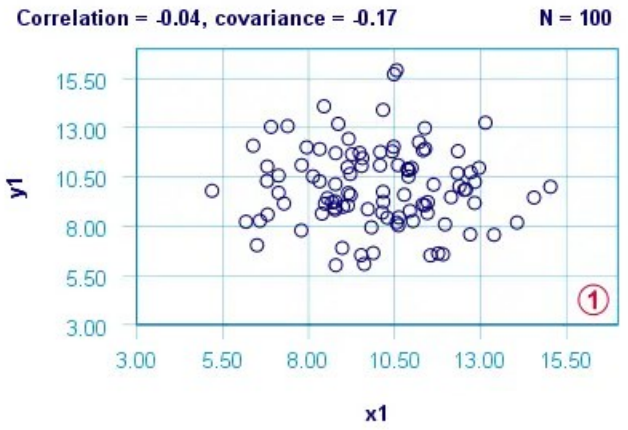
- È una modalità più rigorosa che consente di studiare il **grado di intensità** del legame lineare tra coppie di variabili

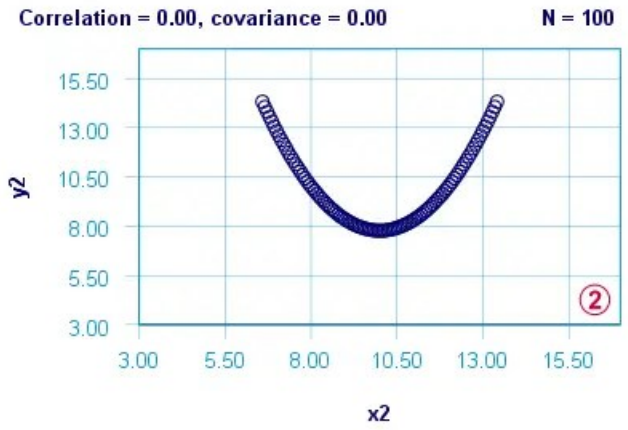
$$r_{xy} = \frac{Cov(X,Y)}{\sqrt{(VarX)(VarY)}}$$

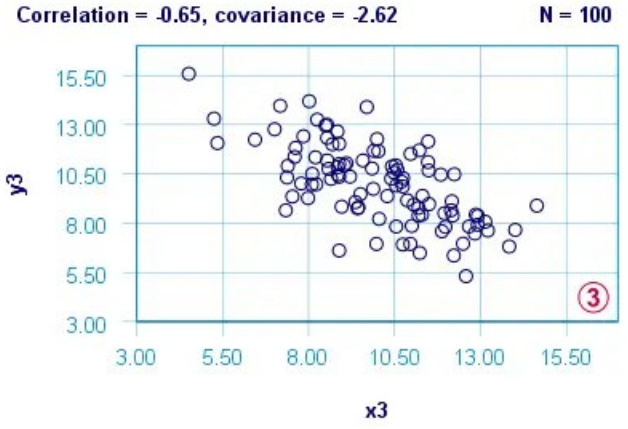
- Coefficiente di Pearson

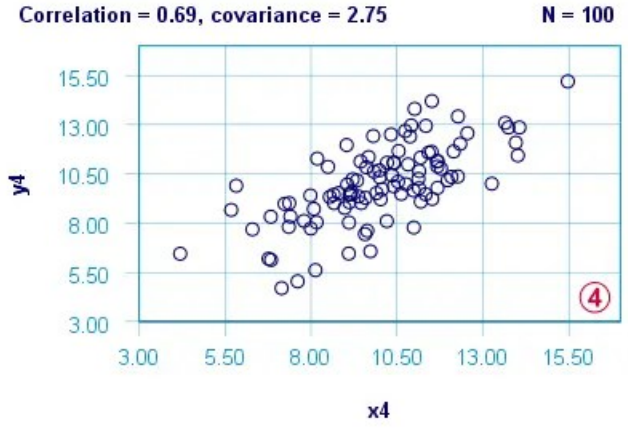
Il coefficiente di correlazione ci permette di:
- riassumere la forza della relazione **lineare** fra le variabili
- verificare l'apparente associazione fra le variabili

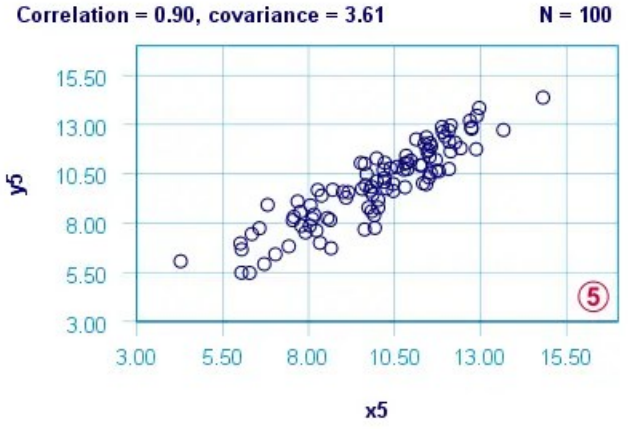
Il coefficiente di correlazione:
- varia da -1 a 1 (se uguale a 1 o a -1: perfettamente correlate)
- è positivo quando i valori delle variabili crescono insieme
- è negativo quando i valori di una variabile crescono al decrescere dei valori dell'altra
- non è influenzato dalle unità di misura

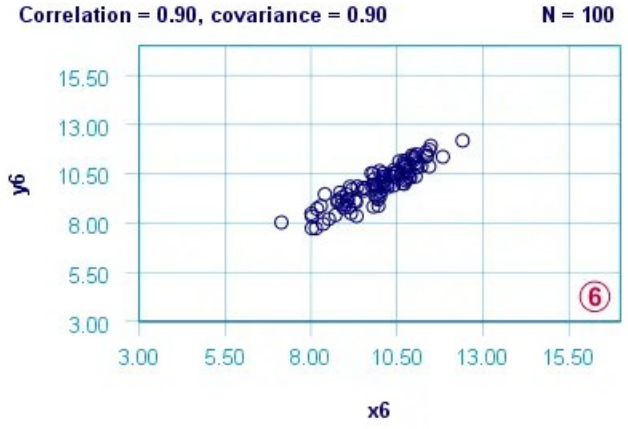

Correlation = -0.04, covariance = -0.17
N = 100
y1
x1
15.50
13.00
10.50
8.00
5.50
3.00
3.00
5.50
8.00
10.50
13.00
15.50
①

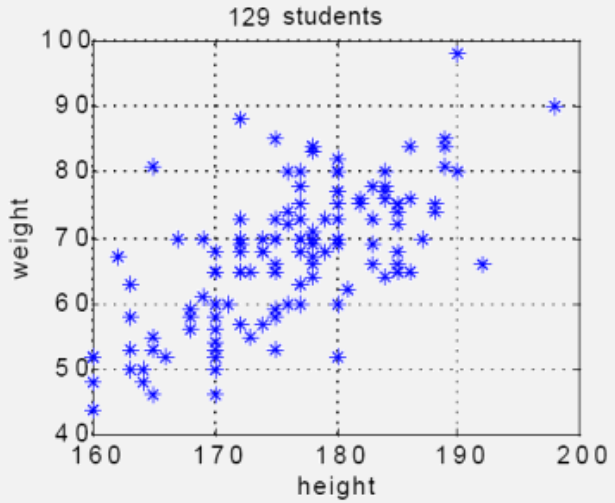

Correlation = 0.00, covariance = 0.00
N = 100
y2
x2
15.50
13.00
10.50
8.00
5.50
3.00
3.00
5.50
8.00
10.50
13.00
15.50
②

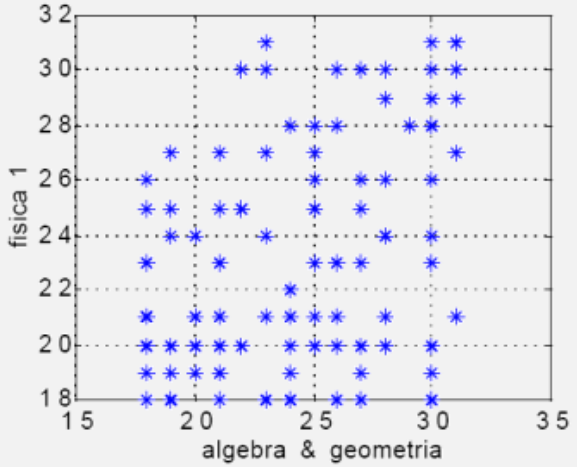

Correlation = -0.65, covariance = -2.62
N = 100
y3
x3
15.50
13.00
10.50
8.00
5.50
3.00
3.00
5.50
8.00
10.50
13.00
15.50
③

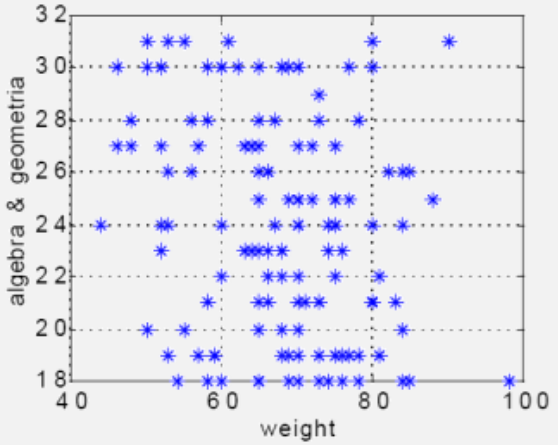

Correlation = 0.69, covariance = 2.75
N = 100
y4
x4
15.50
13.00
10.50
8.00
5.50
3.00
3.00
5.50
8.00
10.50
13.00
15.50
④

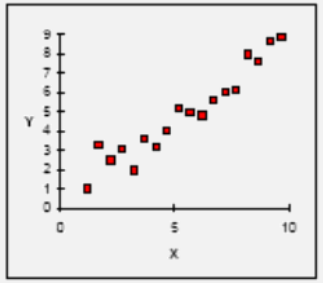

Correlation = 0.90, covariance = 3.61
N = 100
y5
x5
15.50
13.00
10.50
8.00
5.50
3.00
3.00
5.50
8.00
10.50
13.00
15.50
⑤


Correlation = 0.90, covariance = 0.90
N = 100
y6
x6
15.50
13.00
10.50
8.00
5.50
3.00
3.00
5.50
8.00
10.50
13.00
15.50
⑥



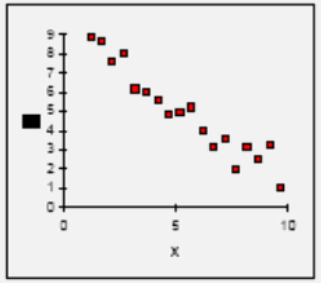


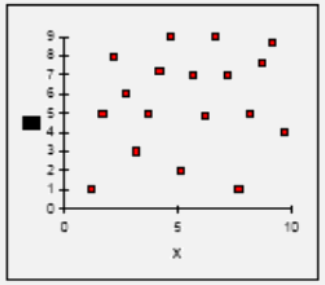




r=0,96

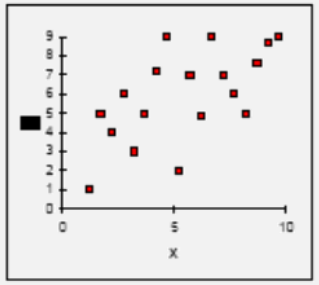
r=-0,96



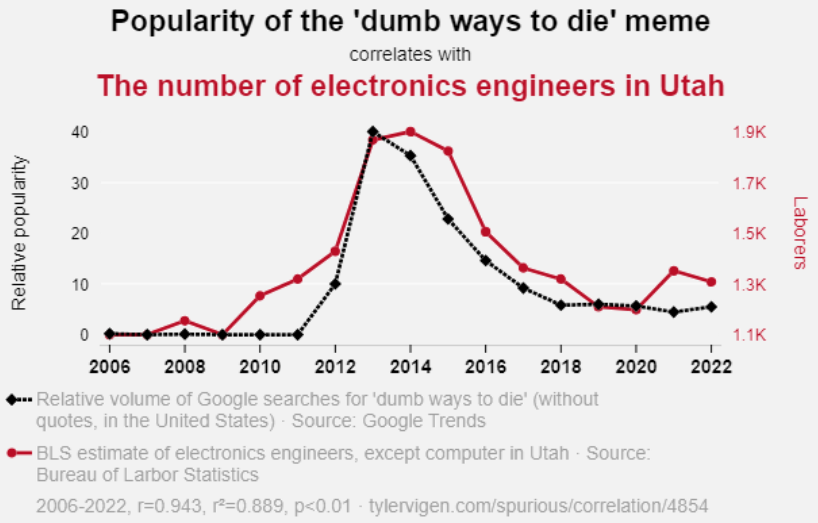


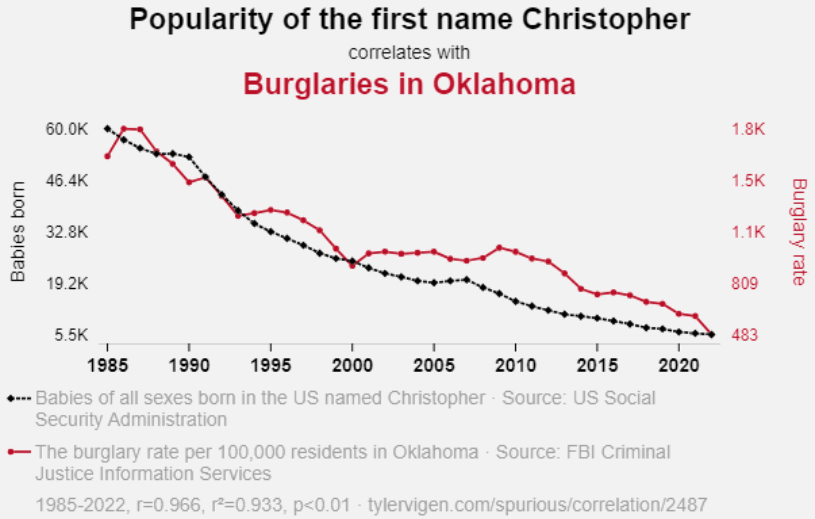
r=0,12

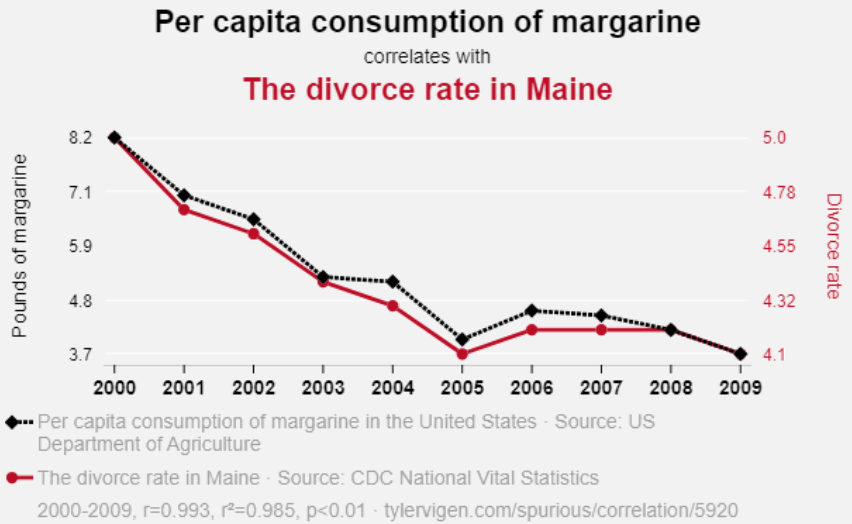
r=0,62



CORRELATION IS NOT CAUSATION





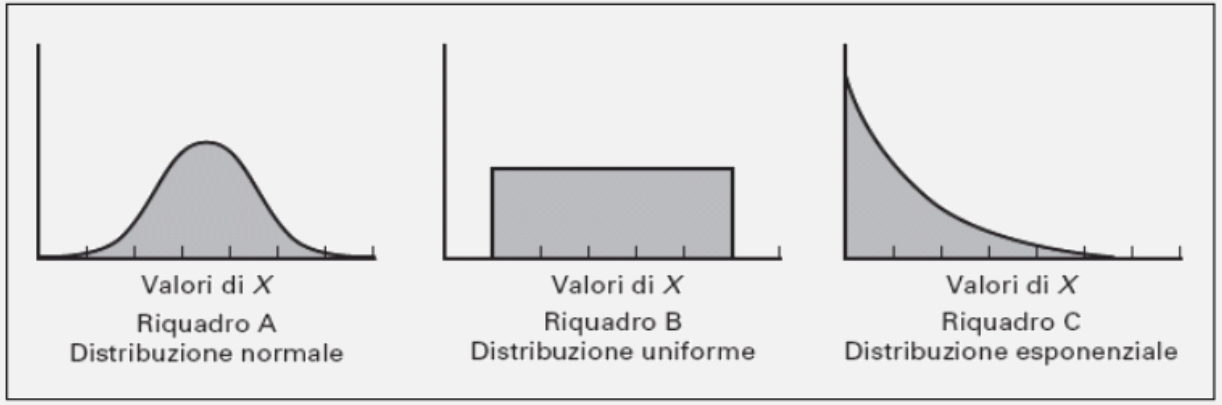


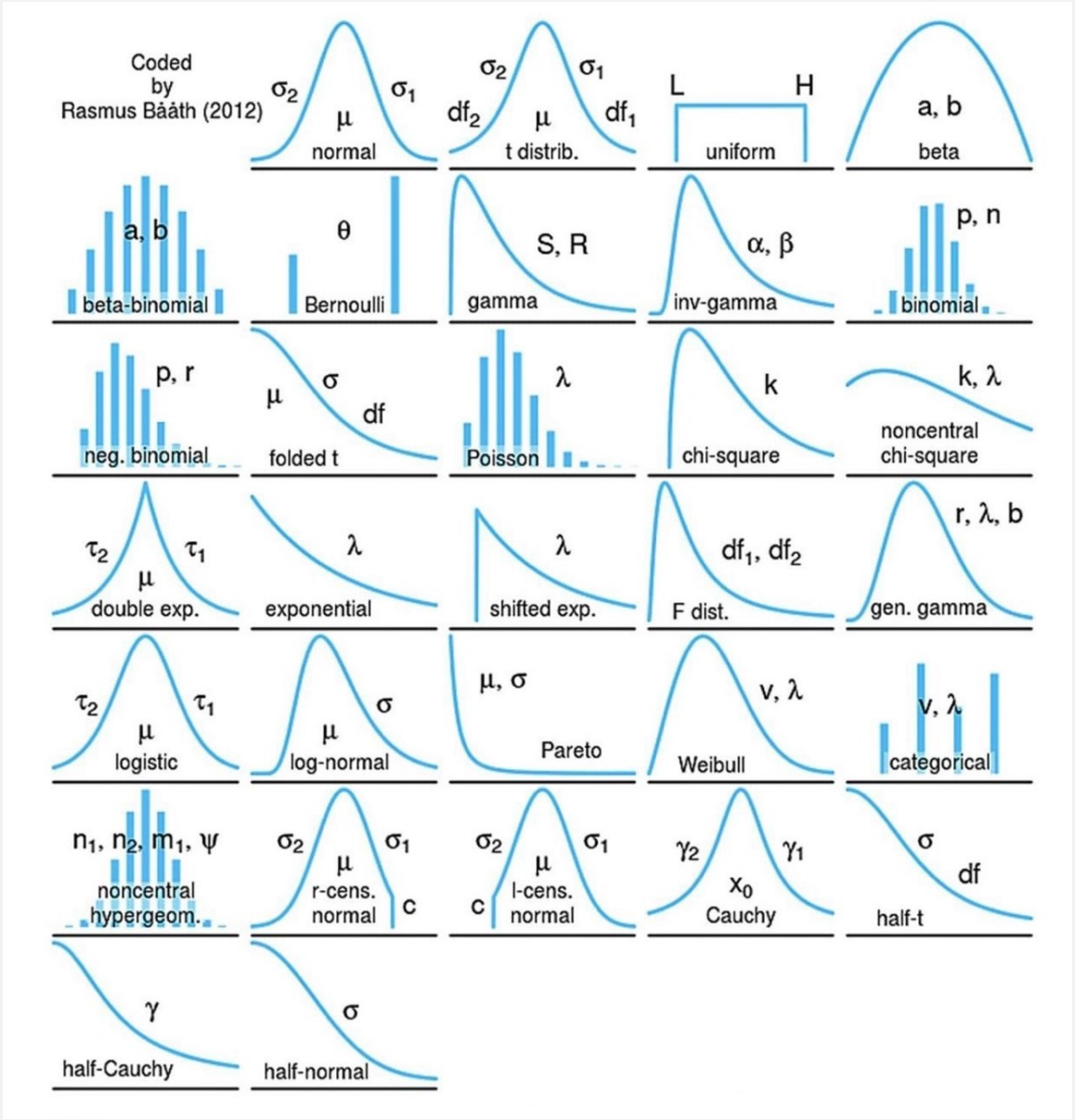
Distribuzioni di probabilità continue

- Una funzione di densità di probabilità continua è un modello che definisce analiticamente come si distribuiscono i valori assunti da una variabile aleatoria continua

- Quando si dispone di un'espressione matematica adatta alla rappresentazione di un fenomeno continuo, siamo in grado di calcolare la probabilità che la variabile aleatoria assuma valori compresi in intervalli

- La figura rappresenta graficamente tre funzioni di densità di probabilità: normale, uniforme ed esponenziale

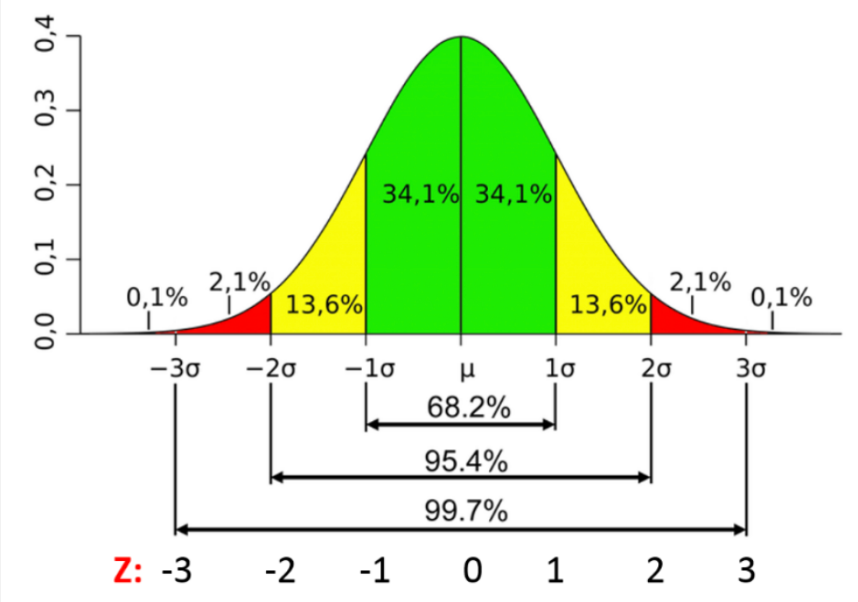


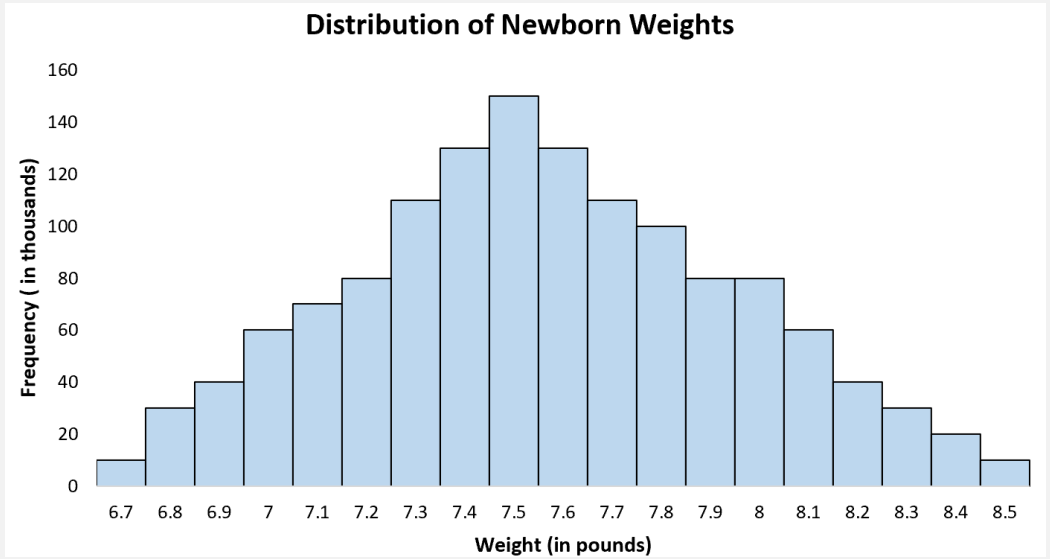

Coded
by
Rasmus Bååth (2012)
σ2
σ1
μ
normal
σ2
σ1
df2
μ
df1
t distrib.
L
H
uniform
a, b
beta
a, b
beta-binomial
θ
Bernoulli
S, R
gamma
α, β
inv-gamma
p, n
binomial
p, r
neg. binomial
μ
σ
df
folded t
λ
Poisson
k
chi-square
k, λ
noncentral
chi-square
τ2
τ1
μ
double exp.
λ
exponential
λ
shifted exp.
df1, df2
F dist.
r, λ, b
gen. gamma
τ2
τ1
μ
logistic
μ
σ
log-normal
μ, σ
Pareto
ν, λ
Weibull
ν, λ
categorical
n1, n2, m1, ψ
noncentral
hypergeom.
σ2
σ1
μ
r-cens.
normal
c
σ2
σ1
μ
c
l-cens.
normal
γ2
γ1
x0
Cauchy
σ
df
half-t
γ
half-Cauchy
σ
half-normal

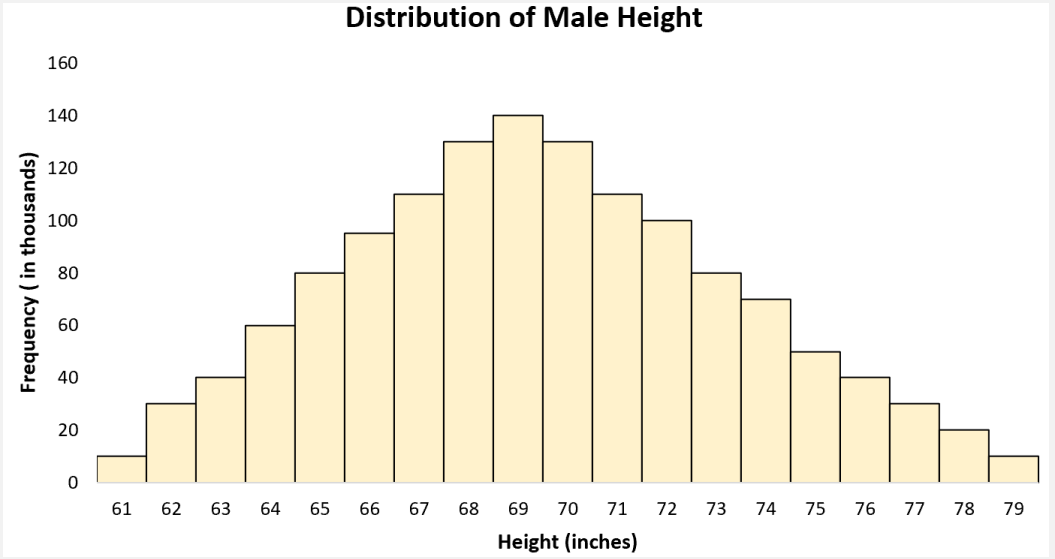
Distribuzione normale

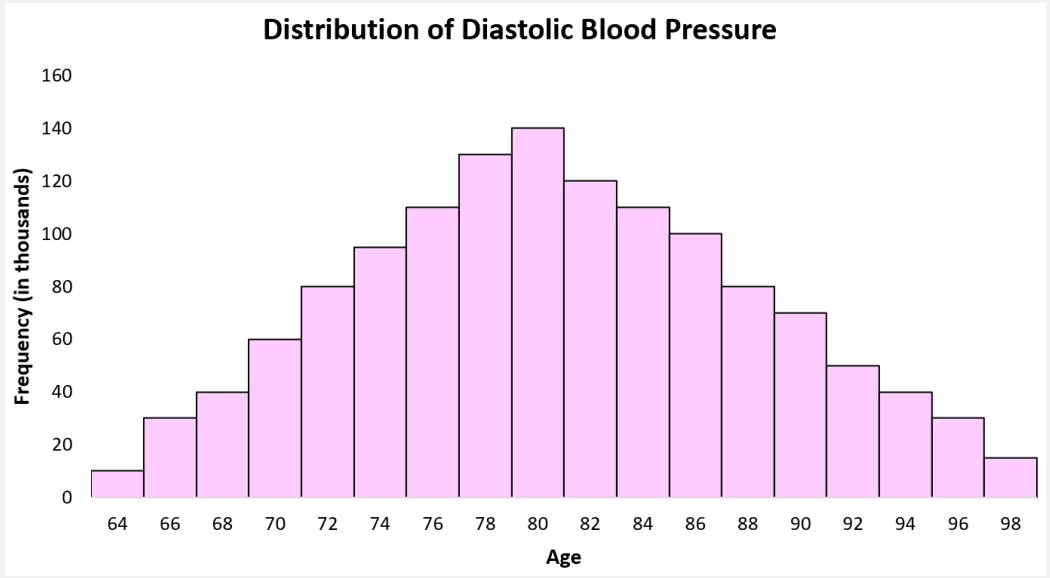
- La distribuzione normale (o distribuzione Gaussiana) è la distribuzione continua più utilizzata in statistica.

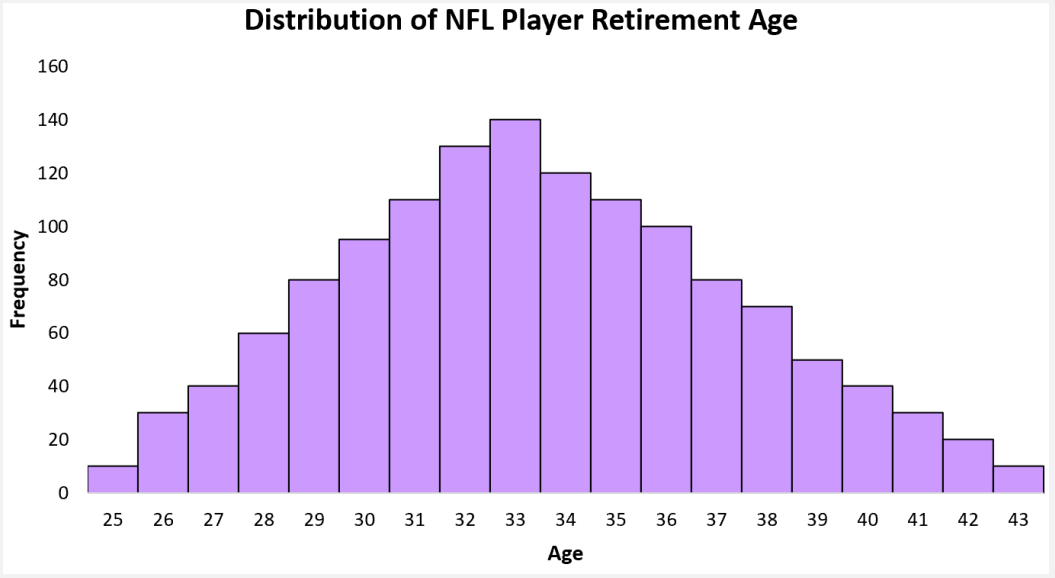
- La distribuzione normale è importante in statistica per tre motivi fondamentali:
  1. Diversi fenomeni continui sembrano seguire, almeno approssimativamente, una distribuzione normale.
  2. La distribuzione normale può essere utilizzata per approssimare numerose distribuzioni di probabilità discrete.




Distribution of Newborn Weights
Frequency (in thousands)
0
20
40
60
80
100
120
140
160
6.7
6.8
6.9
7
7.1
7.2
7.3
7.4
7.5
7.6
7.7
7.8
7.9
8
8.1
8.2
8.3
8.4
8.5
Weight (in pounds)


Distribution of Male Height
Frequency (in thousands)
0
20
40
60
80
100
120
140
160
61
62
63
64
65
66
67
68
69
70
71
72
73
74
75
76
77
78
79
Height (inches)


Distribution of Diastolic Blood Pressure
Frequency (in thousands)
0
20
40
60
80
100
120
140
160
64
66
68
70
72
74
76
78
80
82
84
86
88
90
92
94
96
98
Age

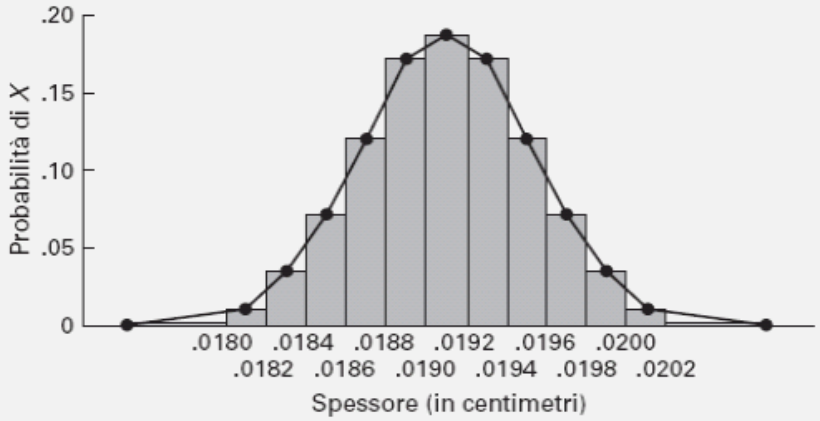

Distribution of NFL Player Retirement Age
Frequency
0
20
40
60
80
100
120
140
160
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
Age

Distribuzione normale

La distribuzione normale ha alcune importanti caratteristiche:

- La distribuzione normale ha una forma campanulare e simmetrica
- Le sue misure di posizione centrale (valore atteso, mediana) coincidono
- Il suo range interquartile è pari a 1.33 volte lo scarto quadratico medio, cioè copre un intervallo compreso tra $\mu - 2/3\,\sigma$ e $\mu + 2/3\,\sigma$
- La variabile aleatoria con distribuzione normale assume valori compresi tra $-\infty$ e $+\infty$

Consideriamo lo spessore misurato in centimetri di 10000 rondelle di ottone prodotte da una grande società metallurgica. Il fenomeno aleatorio continuo di interesse, lo spessore delle rondelle, si distribuisce approssimativamente come una normale



Distribuzione normale

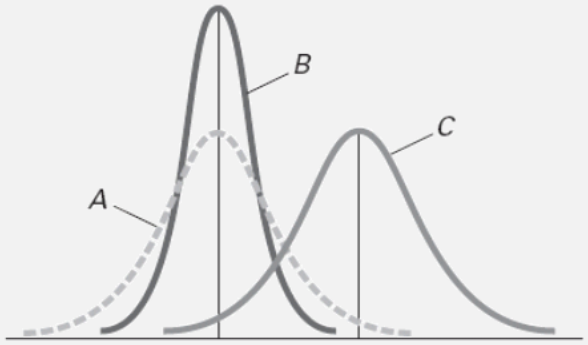
Utilizzeremo il simbolo $f(X)$ per denotare l'espressione matematica di una funzione di densità di probabilità. Nel caso della distribuzione normale la funzione di densità di probabilità normale è data dalla seguente espressione:

Funzione di densità di probabilità normale

$$f(X) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\left(\frac{1}{2}\right)\left[\frac{X-\mu}{\sigma}\right]^2}$$

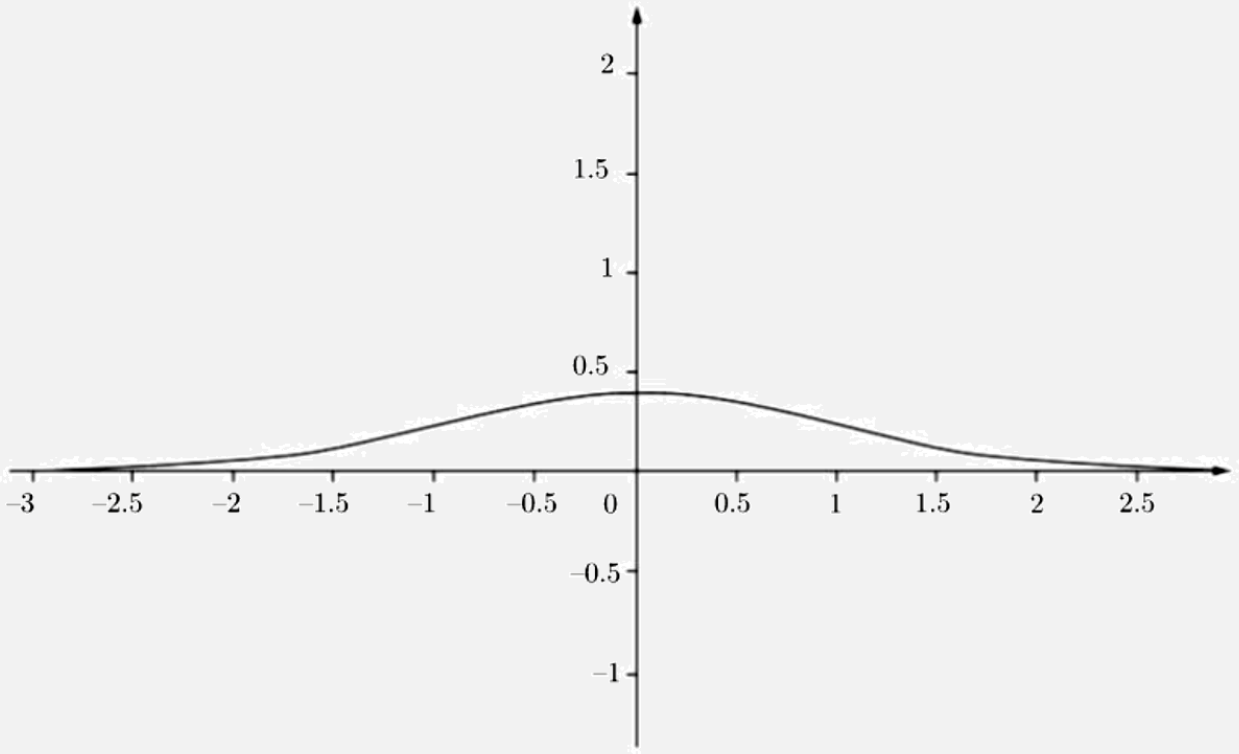
Dove μ è il valore atteso della popolazione; σ è lo scarto quadratico medio della popolazione; X rappresenta i valori assunti dalla variabile aleatoria, $-\infty < X < +\infty$

Notiamo che, essendo e e π delle costanti matematiche, le probabilità di una distribuzione normale dipendono soltanto dai valori assunti dai due parametri μ e σ. Specificando particolari combinazioni di μ e σ, otteniamo differenti distribuzioni di probabilità normali.



Vediamo il grafico per $\mu = 0$ e $\sigma = 1$; la curva gaussiana in questo caso prende il nome di *curva normale standardizzata*:

$$y = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

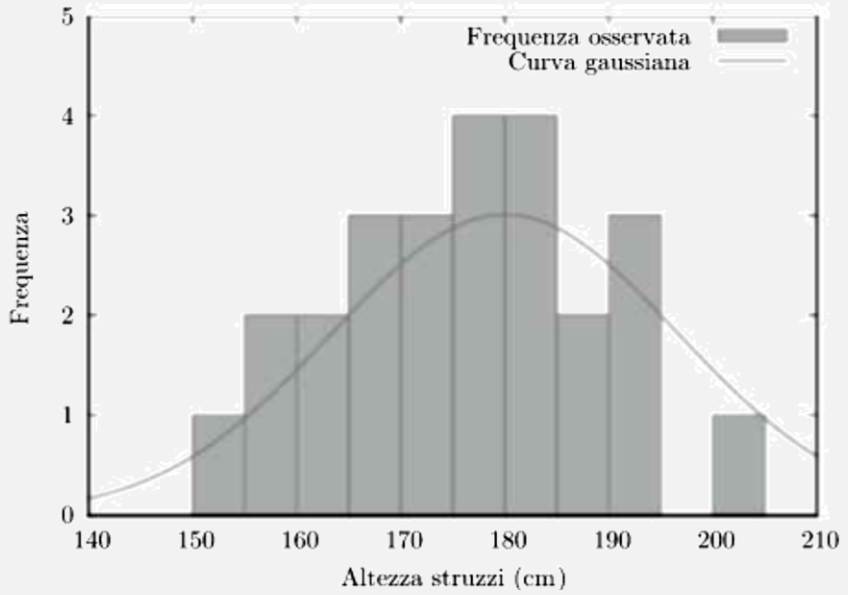


Supponiamo di avere una tabella che riporta i dati dell'altezza in cm (arrotondata) degli struzzi in un allevamento in termini di modalità e frequenze assolute:

Altezza	205	200	195	190	185	180	175	170	165	160	155
F. A.	1	0	3	2	4	4	3	3	2	1	1

Calcoliamo sia varianza sia valor medio ottenendo:

$$\sigma^2 = 16.58, \qquad \mu = 180$$



Esercizio.

Supponiamo di misurare le altezze in centimetri al garrese di dieci cani partecipanti a un concorso canino, ottenendo la variabile statistica:

$$Y = (40, 42, 38, 41, 40, 45, 46, 42, 42, 41)$$

Si determini la mediana, la media aritmetica, la varianza.

Svolgimento

Per determinare la mediana dobbiamo disporre i dati in ordine crescente:

$$38, 40, 40, 41, 41, 42, 42, 42, 45, 46$$

Media: somma dei dati diviso 10 → 41.7 cm

Mediana: se il numero (n) di dati è pari, la mediana è stimata utilizzando i due valori centrali che occupano le posizioni $n/2$ e $n/2 + 1$→ media del quinto e del sesto valore → 41.5 cm

Varianza: $\sigma^2 = \frac{\sum_{i=1}^{N}(X_i-\mu)^2}{N}$ → 5.01

Esercizio.

Uno studente di biologia ha conseguito i seguenti voti:

- Biologia 1 (6 CFU): 27
- Matematica (8 CFU): 24
- Genetica (6 CFU): 30
- Chimica (6 CFU): 25

Calcolare la media pesata secondo i CFU. Che cosa accade alla media pesata se lo studente consegue il voto 26 nell'esame di Istologia (4 CFU)?

Svolgimento

Calcoliamo la media pesata:

$$(27 \times 6 + 24 \times 8 + 30 \times 6 + 25 \times 6)/26 = 26.31$$

Anche senza calcoli possiamo affermare che la media pesata certamente diminuirà, anche se di una quantità molto piccola, dato che l'ultimo voto conseguito è vicino a tale media. Infatti:

$$(27 \times 6 + 24 \times 8 + 30 \times 6 + 25 \times 6 + 26 \times 4)/30 = 26.27$$

Media pesata: $\mu = \frac{\sum_{i=1}^{N} p_i X_i}{\sum_{i=1}^{N} p_i}$ → (27x6 + 24x8 + 30x6 + 25x6)/26=26.31

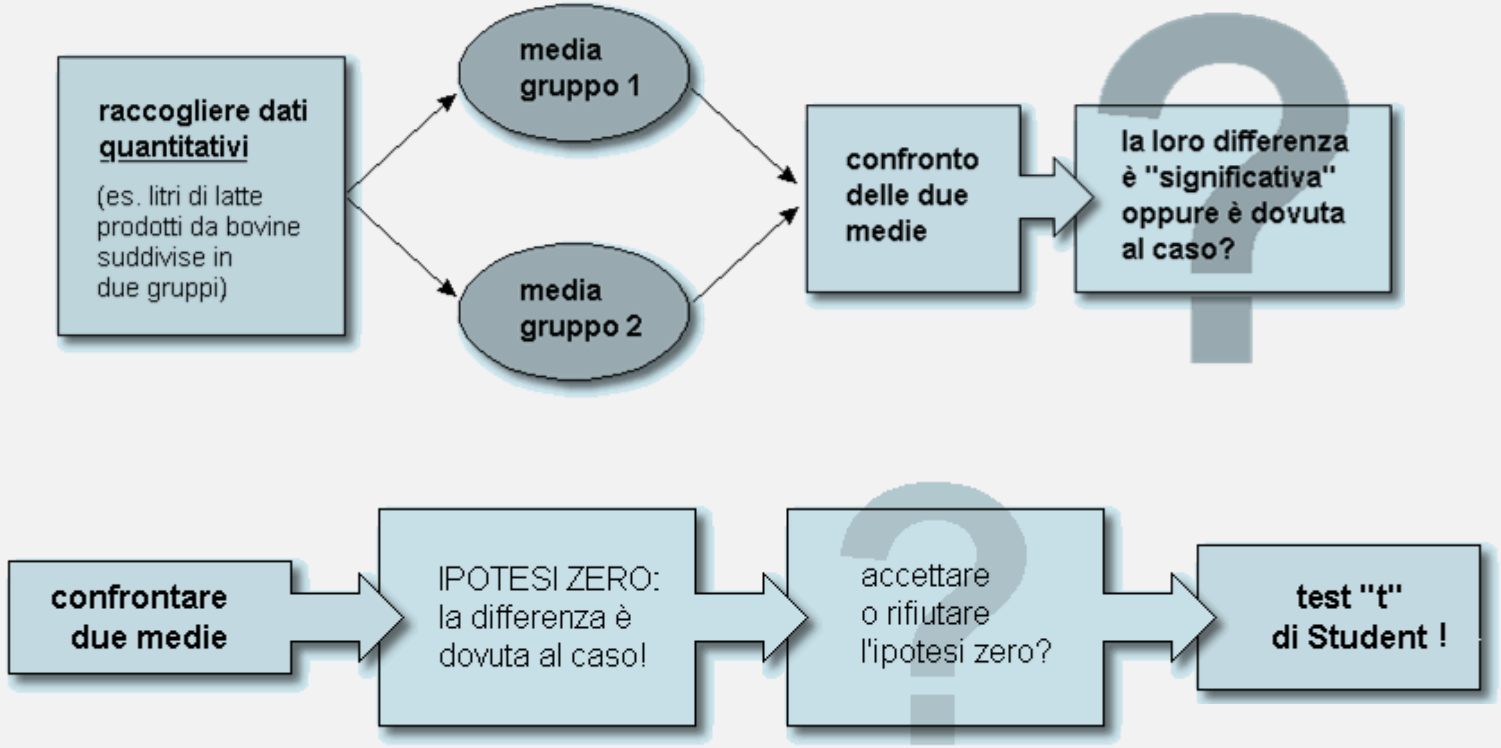
Dal momento che l'ultimo conseguito è vicino alla media pesata degli altri esami, anche senza calcoli possiamo intuire che la media pesata finale diminuirà di una quantità molto piccola:

(27x6 + 24x8 + 30x6 + 25x6 + 26x4)/26=26.27

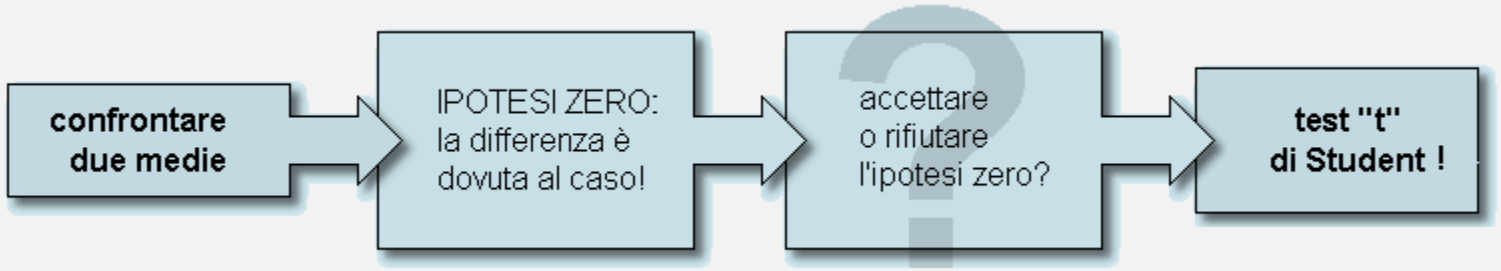
Confrontare due medie: il test *t* di Student

la differenza fra le medie dei due campioni è significativa?

Puoi affermare che la differenza osservata non è dovuta al caso ma che, invece, esiste veramente una diversità tra le medie delle due popolazioni?



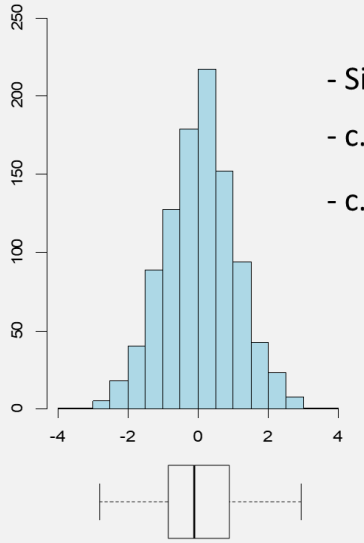
Confrontare due medie: il test *t* di Student



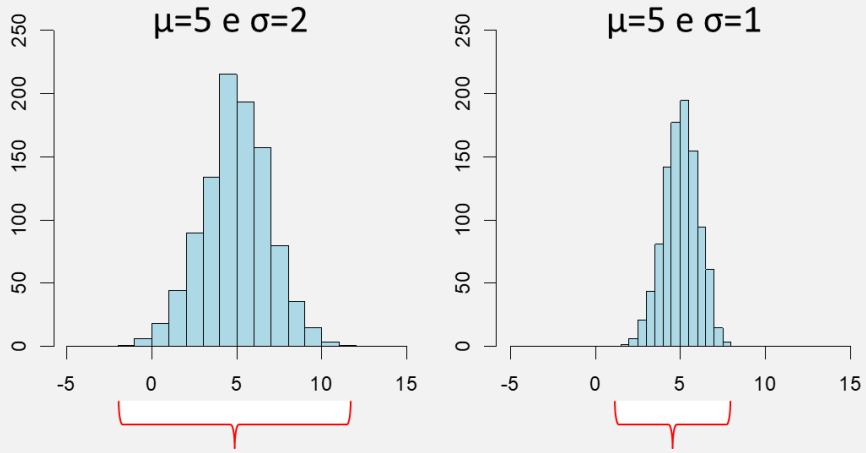
Condizioni:

1. Indipendenza delle osservazioni
2. Normalità delle popolazioni a confronto
3. Omogeneità della varianza

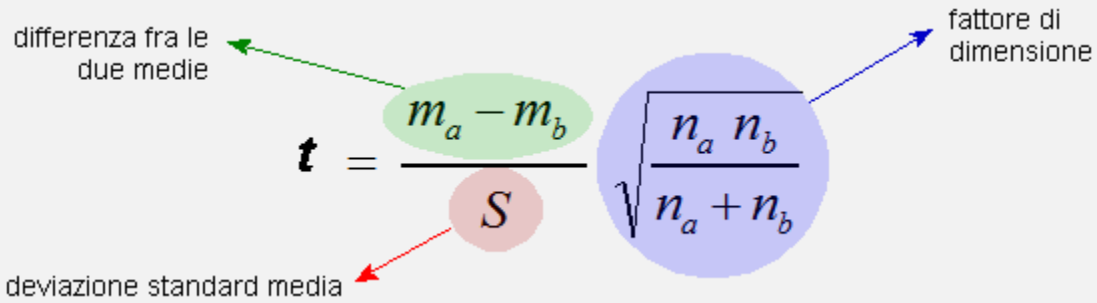
Analisi dell'istogramma



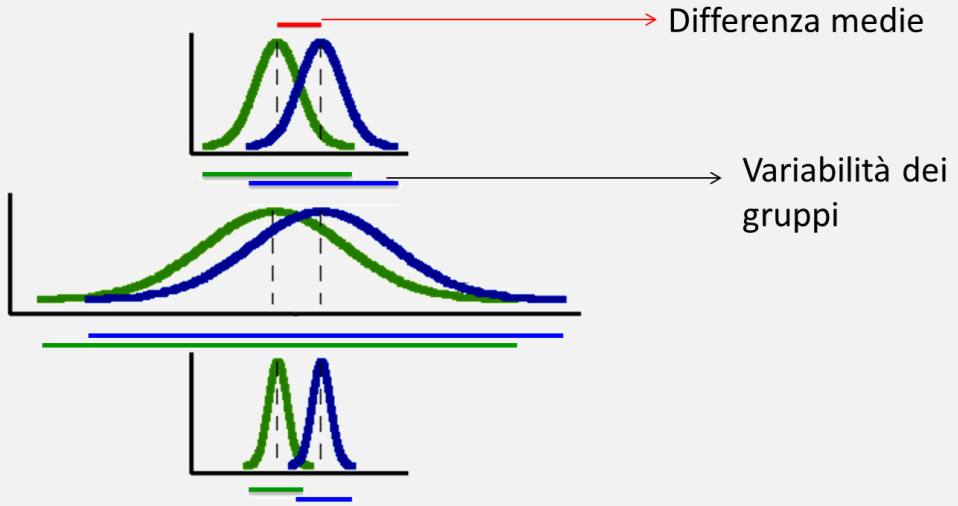
- Simmetria (media ≈mediana)
- c. 2/3 dei dati in un intervallo μ±σ
- c. 95% dei dati in un intervallo μ±2σ

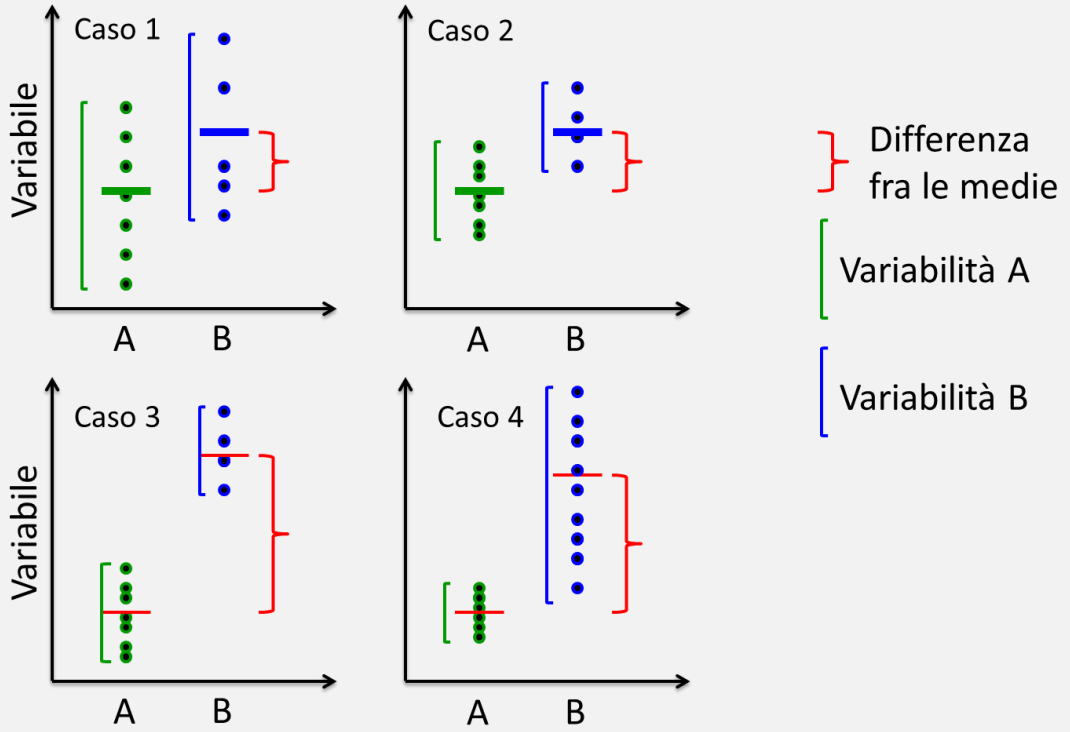


Confrontare due medie: il test *t* di Student

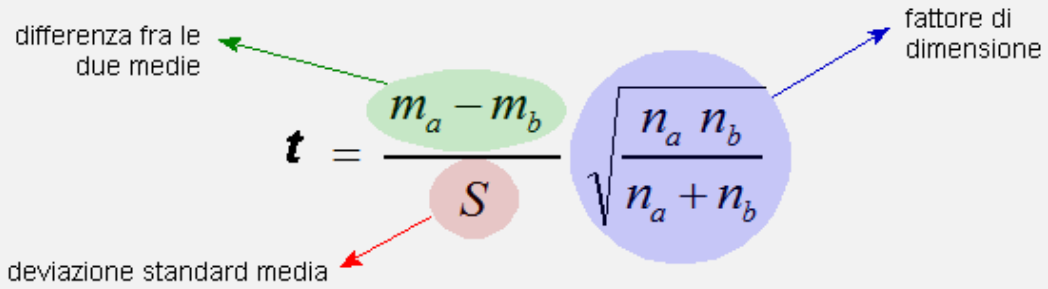


$$t_{calcolato} = \frac{\text{Misura legata alla differenza fra le medie}}{\text{Misura di variabilità dentro i gruppi}}$$

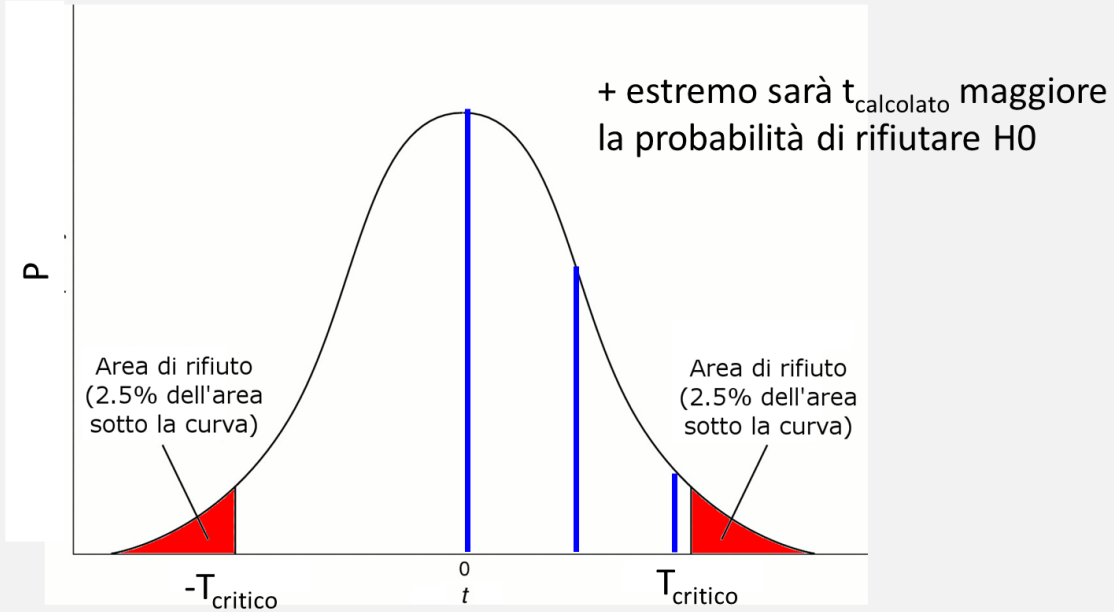




Confrontare due medie: il test *t* di Student



$$t_{calcolato} = \frac{\text{Differenza fra le medie}}{\text{Errore standard della differenza}}$$



+ estremo sarà $t_{calcolato}$ maggiore la probabilità di rifiutare H0