Paradossi elementari in statistica

Maria Piera Rogantin

DIMA – Università di Genova

rogantin@dima.unige.it

Secondo voi la statistica serve per ...

gestire le ricerche di Google	V	F
costruire i test psicologici	V	F
produrre un farmaco	V	F
costruire un telefono cellulare	V	F
calcolare il tasso di disoccupazione	V	F
fare le leggi di una regione	V	F
analizzare i censimenti	V	F
fare previsioni del tempo	V	F
vincere alla tombola	V	F
farsi delle opinioni politiche	V	F
commentare una partita di calcio	V	F

Altri usi della statistica ...

Attenti ai numeri!

da un articolo di *David Spiegelhalter* Nove punti per riconoscere statistiche poco raccomandabili comparso sul quotidiano inglese *The Guardian* dopo la Brexit.

`https://www.theguardian.com/science/2016/jul/17/politicians-dodgy-statistics-tricks-guide`

Esaminiamone alcuni

- **Usare un numero vero cambiandone il significato**
  quanto i Paesi membri versano all'UE
  (lordo/netto/quante sovvenzioni ritornano)
  Es. Regno Unito a settimana: 350m £

- **Usare un numero vero cambiandone il significato**
  quanto i Paesi membri versano all'UE
  (lordo/netto/quante sovvenzioni ritornano)
  Es. Regno Unito a settimana: 350m £ lordo
  250m £ netto – 136m £ levate le sovvenzioni dell'UE al RU

- **Usare un numero vero cambiandone il significato**
  quanto i Paesi membri versano all'UE
  (lordo/netto/quante sovvenzioni ritornano)
  Es. Regno Unito a settimana: 350m £ lordo
  250m £ netto – 136m £ levate le sovvenzioni dell'UE al RU
- **Fare in modo che il numero sembri grande ... ma non troppo**
  Nell'es. precedente: 350m £ *a settimana*

- **Usare un numero vero cambiandone il significato**
  quanto i Paesi membri versano all'UE
  (lordo/netto/quante sovvenzioni ritornano)
  Es. Regno Unito a settimana: 350m £ lordo
  250m £ netto – 136m £ levate le sovvenzioni dell'UE al RU
- **Fare in modo che il numero sembri grande ... ma non troppo**
  Nell'es. precedente: 350m £ *a settimana* e non 19 miliardi di £ *all'anno* (non si percepisce un numero così grande) e non 50m £ *al giorno* (troppo poco!)

- **Usare totali o percentuali**

  Es. *"Negli ultimi tre mesi meno del 93% dei pazienti al Pronto soccorso è visitato in meno di 4 ore"*

  Due interpretazioni.

  "È il minimo in 10 anni! (È inacettabile!)"

  oppure

  "Abbiamo raggiunto il massimo numero di pazienti che è visitato in meno di 4 ore! (Grande successo!)"

- **Usare totali o percentuali**

  Es. *"Negli ultimi tre mesi meno del 93% dei pazienti al Pronto soccorso è visitato in meno di 4 ore"*

  Due interpretazioni.

  "È il minimo in 10 anni! (È inacettabile!)"

  oppure

  "Abbiamo raggiunto il massimo numero di pazienti che è visitato in meno di 4 ore! (Grande successo!)"

  (il numero di pazienti negli ultimi 3 mesi è aumentato)

- **Usare totali o percentuali**

  Es. *"Negli ultimi tre mesi meno del 93% dei pazienti al Pronto soccorso è visitato in meno di 4 ore"*

  Due interpretazioni.

  "È il minimo in 10 anni! (È inacettabile!)"

  oppure

  "Abbiamo raggiunto il massimo numero di pazienti che è visitato in meno di 4 ore! (Grande successo!)"

  (il numero di pazienti negli ultimi 3 mesi è aumentato)

  Esempio numerico.

  Accessi negli ultimi 3 mesi: 30 000 (circa 340 al giorno)

  Il 93% è visitato in meno di 4 ore: **27 900**.

  Negli ultimi 10 anni in media sono arrivati al pronto soccorso 25 000 pazienti ogni 3 mesi (circa 280 al giorno) e se ne visitava in meno di 4 ore il 98% (cioè **24 500**)

- **Usare totali o percentuali**

  Es. *"Negli ultimi tre mesi meno del 93% dei pazienti al Pronto soccorso è visitato in meno di 4 ore"*

  Due interpretazioni.

  "È il minimo in 10 anni! (È inacettabile!)"

  oppure

  "Abbiamo raggiunto il massimo numero di pazienti che è visitato in meno di 4 ore! (Grande successo!)"

  (il numero di pazienti negli ultimi 3 mesi è aumentato)

  Esempio numerico.

  Accessi negli ultimi 3 mesi: 30 000 (circa 340 al giorno)

  Il 93% è visitato in meno di 4 ore: **27 900**.

  Negli ultimi 10 anni in media sono arrivati al pronto soccorso 25 000 pazienti ogni 3 mesi (circa 280 al giorno) e se ne visitava in meno di 4 ore il 98% (cioè **24 500**)

- **Se tutto fallisce ... inventarsi i numeri**

Correlazione e causalità – che cosa c'è dietro
Il consumo di caffè...

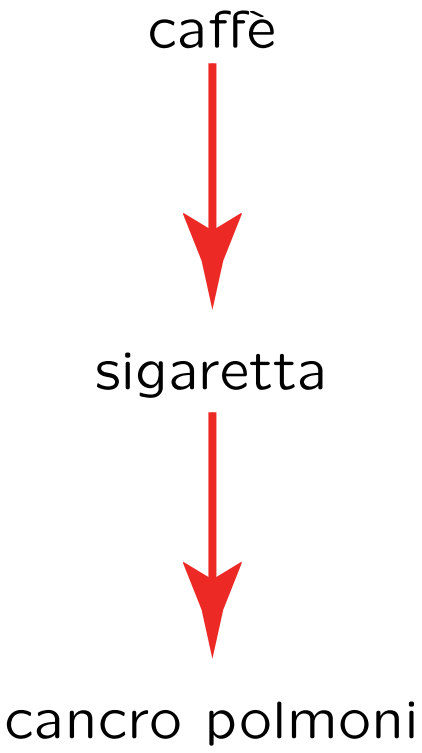
Un'indagine* svolta su 91.767 donne novergesi (seguite per 6-8 anni) evidenzia un significativo

aumento del rischio di cancro polmonare
in presenza di un
elevato consumo di caffè

Che cosa se ne deduce?
visto che non può essere che il cancro ai polmoni induca le donne a bere il caffè
allora
il consumo di caffè favorisce il cancro al polmone

*Eur J Epidemiol. 2016
il legame è negativo per molti altri tipi di cancro

Che cosa c'è dietro?



prima di conoscere l'influenza del fumo sul cancro ai polmoni si era anche pensato che fosse il caffè a provocarlo

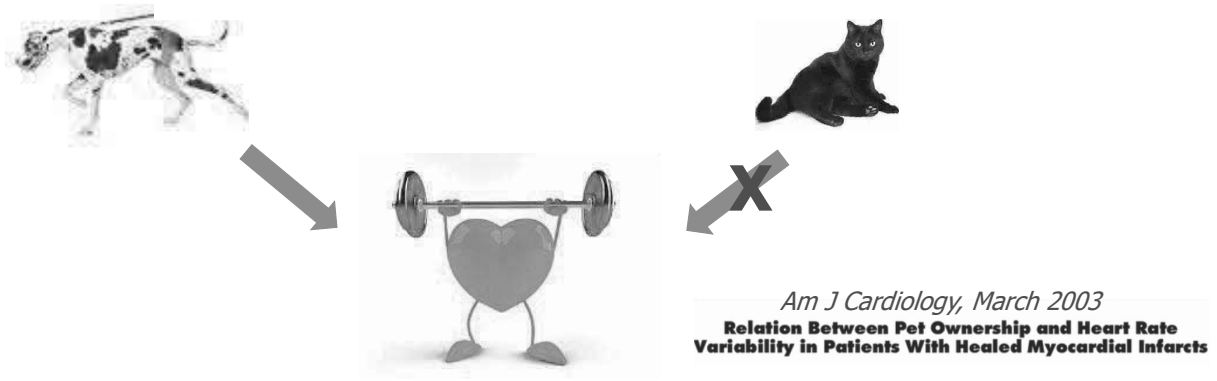
il caffè è un confondente

Un altro esempio. *Dall'intervento della prof. Maria Pia Sormani, Dipartimento di Scienze della salute, Università di Genova, all'evento* **I mestieri dello statistico**, *Genova 22 febbraio 2017*

Cosa fa uno statistico che lavora in campo medico

- Analisi dei dati. Non sempre è come sembra

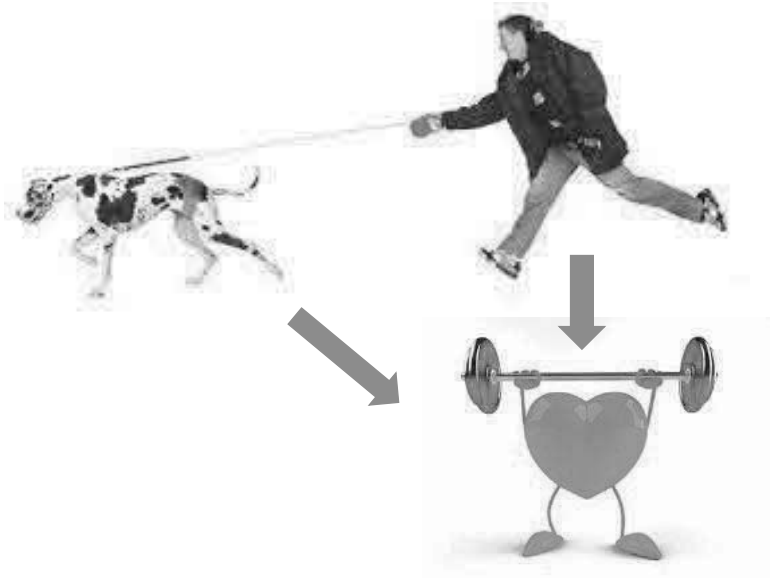
I sopravvissuti a un infarto hanno migliori prestazioni cardiache post-infarto se hanno un cane (ma non un gatto)



Am J Cardiology, March 2003
Relation Between Pet Ownership and Heart Rate Variability in Patients With Healed Myocardial Infarcts

I fattori confondenti

L’esercizio fisico è un fattore confondente : chi possiede un cane fa più eserczio fisico di chi non lo possiede, e l’esercizio fisico è associato a migliori pretazioni cardiache post-infarto



Correzione per fattori confondenti

- L'effetto di un fattore confondente noto si può correggere in fase di analisi statistica
- Avere un cane ha un effetto sulla salute cardiaca post-infarto a parità di esercizio fisico?

Popolazioni e sottopopolazioni – Che cosa c'è dentro

Primo esempio: numero di processi in Florida nel 1976/77 per reati passibili di pena di morte classificati in base alla razza dell'accusato e della pena inflitta

Strumenti: tabelle di contingenza e percentuali riga

		Sentenza		
		pena morte	altro	TOT
Razza accusato	bianca	19	141	160
	nera	17	149	166
	TOT	36	290	326

Popolazioni e sottopopolazioni – Che cosa c'è dentro

Primo esempio: numero di processi in Florida nel 1976/77 per reati passibili di pena di morte classificati in base alla razza dell'accusato e della pena inflitta

Strumenti: tabelle di contingenza e percentuali riga

		Sentenza		
		pena morte	altro	TOT
Razza accusato	bianca	19	141	160
	nera	17	149	166
	TOT	36	290	326

Percentuali riga

		Sentenza		
		pena morte	altro	TOT
Razza accusato	bianca	11.9%	88.1%	100%
	nera	10.2%	89.8%	100%
	TOT	11.0%	89.0%	100%

La sentenza risulta essere indipendente dalla razza dell'accusato

Le cose cambiano se si considera anche la razza della vittima

Suddividiamo la tabella originale

		Sentenza		
		pena morte	altro	TOT
Razza accusato	bianca	19	141	160
	nera	17	149	166
	TOT	36	290	326

in due tabelle a seconda che della razza della vittima

Razza della vittima: nera

		Sentenza	
		pena morte	altro
Razza accusato	bianca	0	9
	nera	6	97

Razza della vittima: bianca

		Sentenza	
		pena morte	altro
Razza accusato	bianca	19	132
	nera	11	52

Calcoliamo le percentuali riga per entrambe le tabelle

Razza della vittima: nera

		Sentenza	
		pena morte	altro
Razza accusato	bianca	0	9
	nera	6	97

Razza della vittima: bianca

		Sentenza	
		pena morte	altro
Razza accusato	bianca	19	132
	nera	11	52

		Sentenza	
		pena morte	altro
Razza accusato	bianca	0%	100%
	nera	6%	94%

		Sentenza	
		pena morte	altro
Razza accusato	bianca	13%	87%
	nera	17%	83%

Se la razza della vittima è:

- nera: a nessun bianco viene inflitta la pena di morte
- bianca: al 13% dei bianchi e al 17% dei neri viene inflitta la pena di morte

Calcoliamo le percentuali riga per entrambe le tabelle

Razza della vittima: nera

Razza accusato	Sentenza: pena morte	Sentenza: altro
bianca	0	9
nera	6	97

Razza della vittima: bianca

Razza accusato	Sentenza: pena morte	Sentenza: altro
bianca	19	132
nera	11	52

Razza accusato	Sentenza: pena morte	Sentenza: altro
bianca	0%	100%
nera	6%	94%

Razza accusato	Sentenza: pena morte	Sentenza: altro
bianca	13%	87%
nera	17%	83%

Se la razza della vittima è:

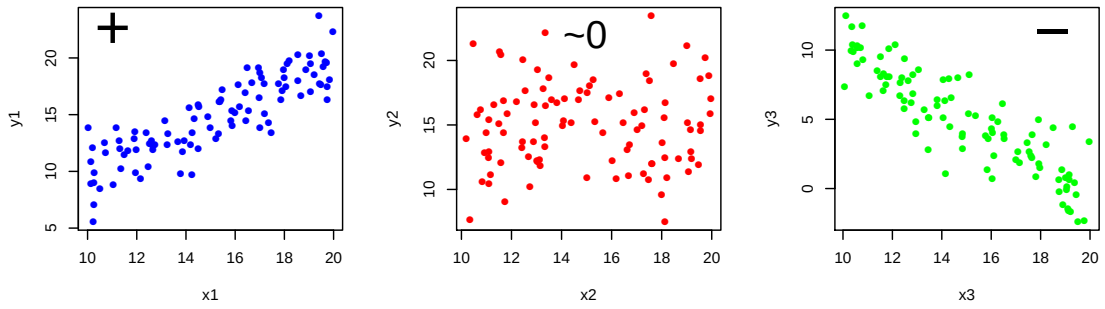
- nera: a nessun bianco viene inflitta la pena di morte
- bianca: al 13% dei bianchi e al 17% dei neri viene inflitta la pena di morte

La sentenza e la razza dell'accusato sono indipendenti, non lo sono se si conosce la razza della vittima

Secondo esempio
Strumenti: covarianza e correlazione

la covarianza e la correlazione sono

- positive se – in media – a alti valori di X corrispondono alti valori di Y (e a bassi valori di X, bassi valori di Y)
- negative se – in media – a alti valori di X corrispondono bassi valori di Y (e a bassi valori di X, alti valori di Y)
- circa 0 altrimenti



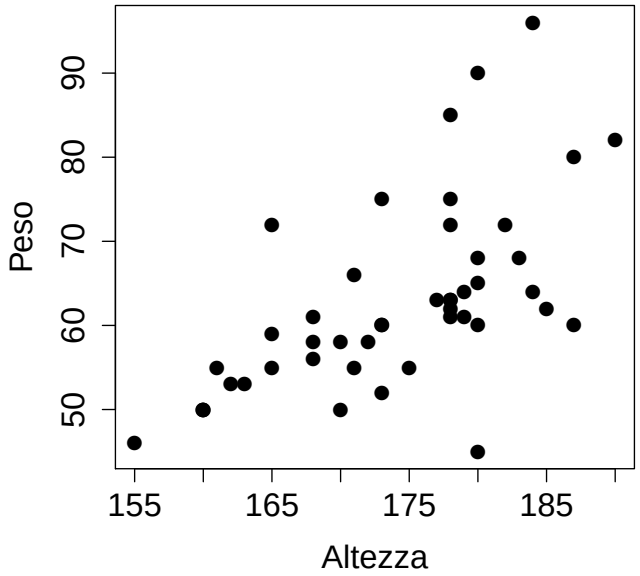
La correlazione è compresa fra -1 e 1

NOTA: correlazione positiva non implica causalità

Covarianza e correlazione
in una popolazione e nei suoi sottogruppi

Esempio:
altezza e peso di ragazzi di primo anno di università

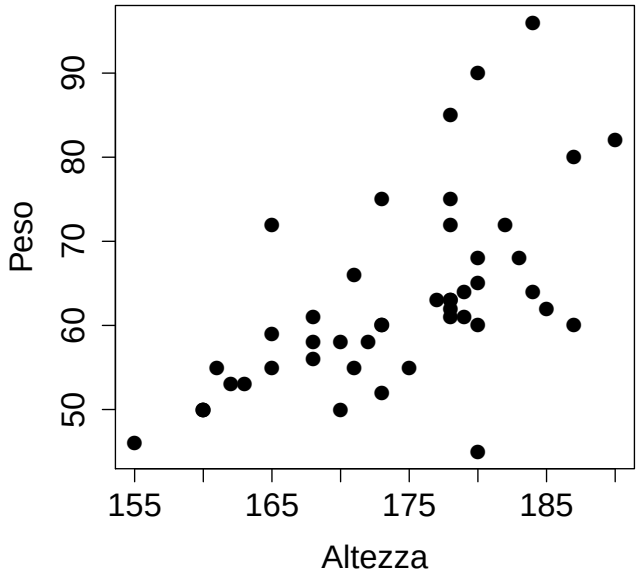
$\rho_{\text{tot}} = 0.61$

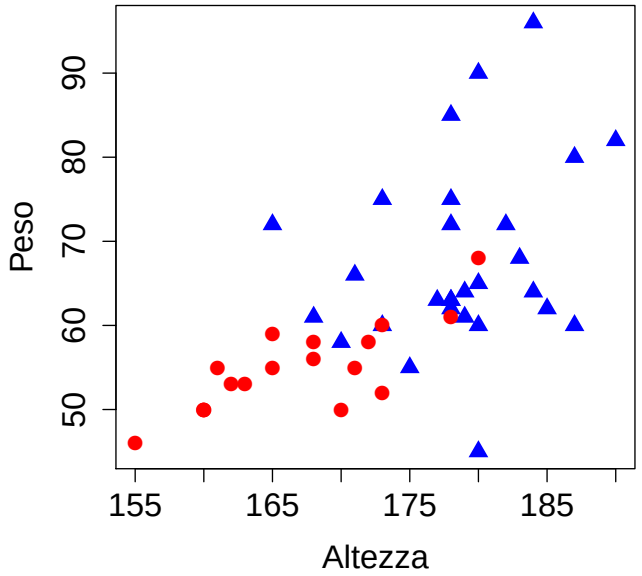


Covarianza e correlazione
in una popolazione e nei suoi sottogruppi

Esempio:
altezza e peso di ragazzi di primo anno di università

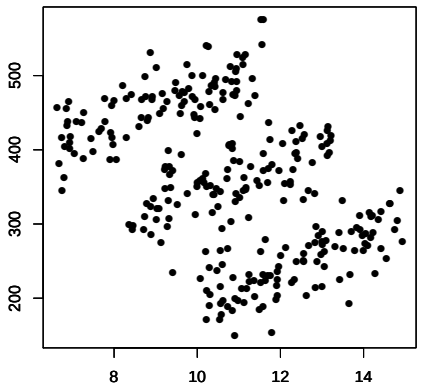
$$\rho_{\text{tot}} = 0.61$$
$$\rho_M = 0.26$$
$$\rho_F = 0.78$$



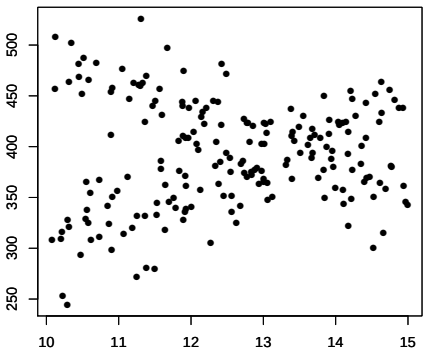


Qualche paradosso

$\rho_{\text{tot}} < 0$



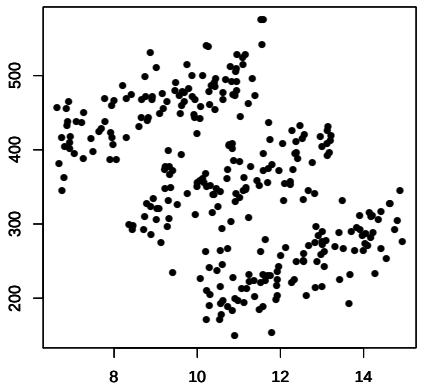
$\rho_{\text{tot}} \simeq 0$

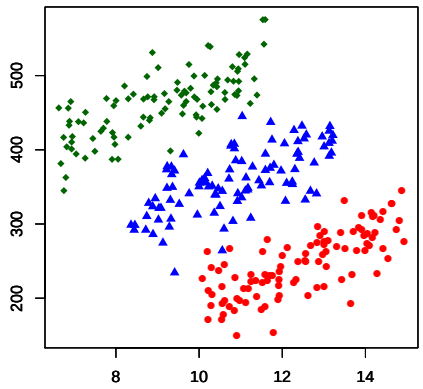


Qualche paradosso

$\rho_{\text{tot}} < 0$

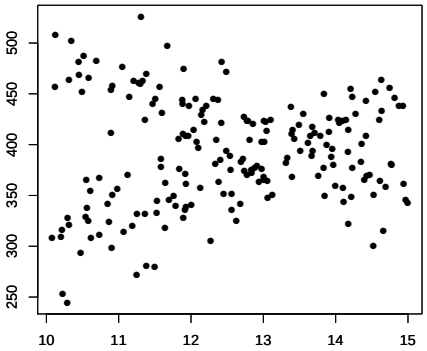
$\rho_1 > 0 \quad \rho_2 > 0 \quad \rho_3 > 0$

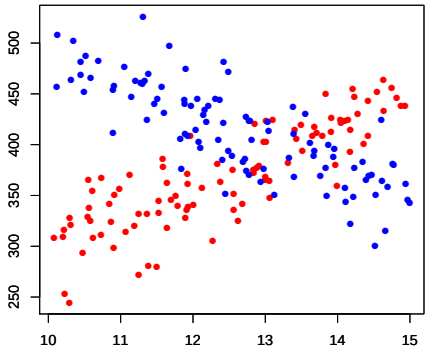




$\rho_{\text{tot}} \simeq 0$

$\rho_1 > 0 \quad \rho_2 < 0$





Che cosa abbiamo capito con questi esempi?

- consumo di caffè – malattie cardiache
- pena di morte e razza accusato/vittima
- correlazione nella popolazione e nei sottogruppi

La soglia dei gruppi

Es.: Che succede se il più furbo dei doriani diventa genoano?

6 doriani - 5 genoani

i genoani sono tutti più furbi dei doriani

disegnati in ordine di furbizia

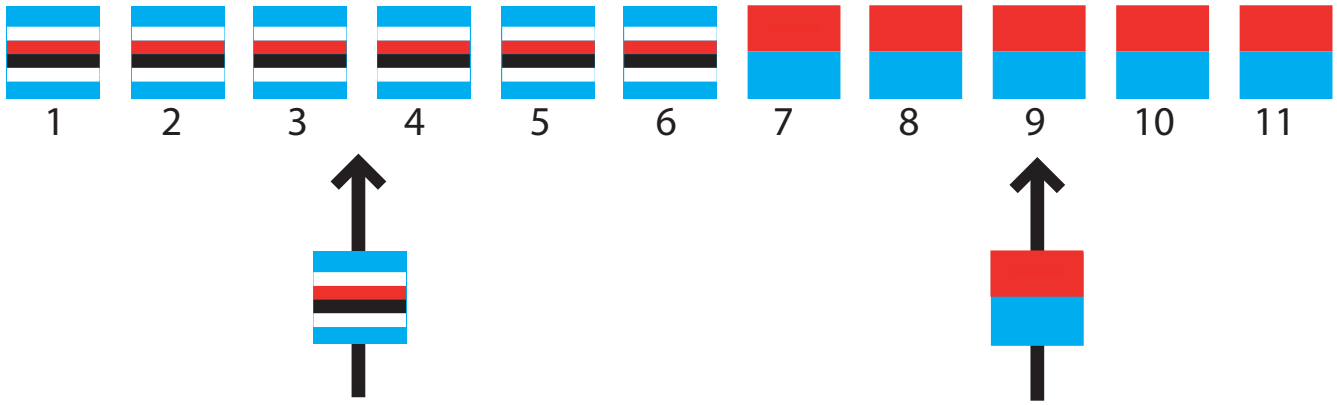
La soglia dei gruppi

Es.: Che succede se il più furbo dei doriani diventa genoano?

6 doriani - 5 genoani

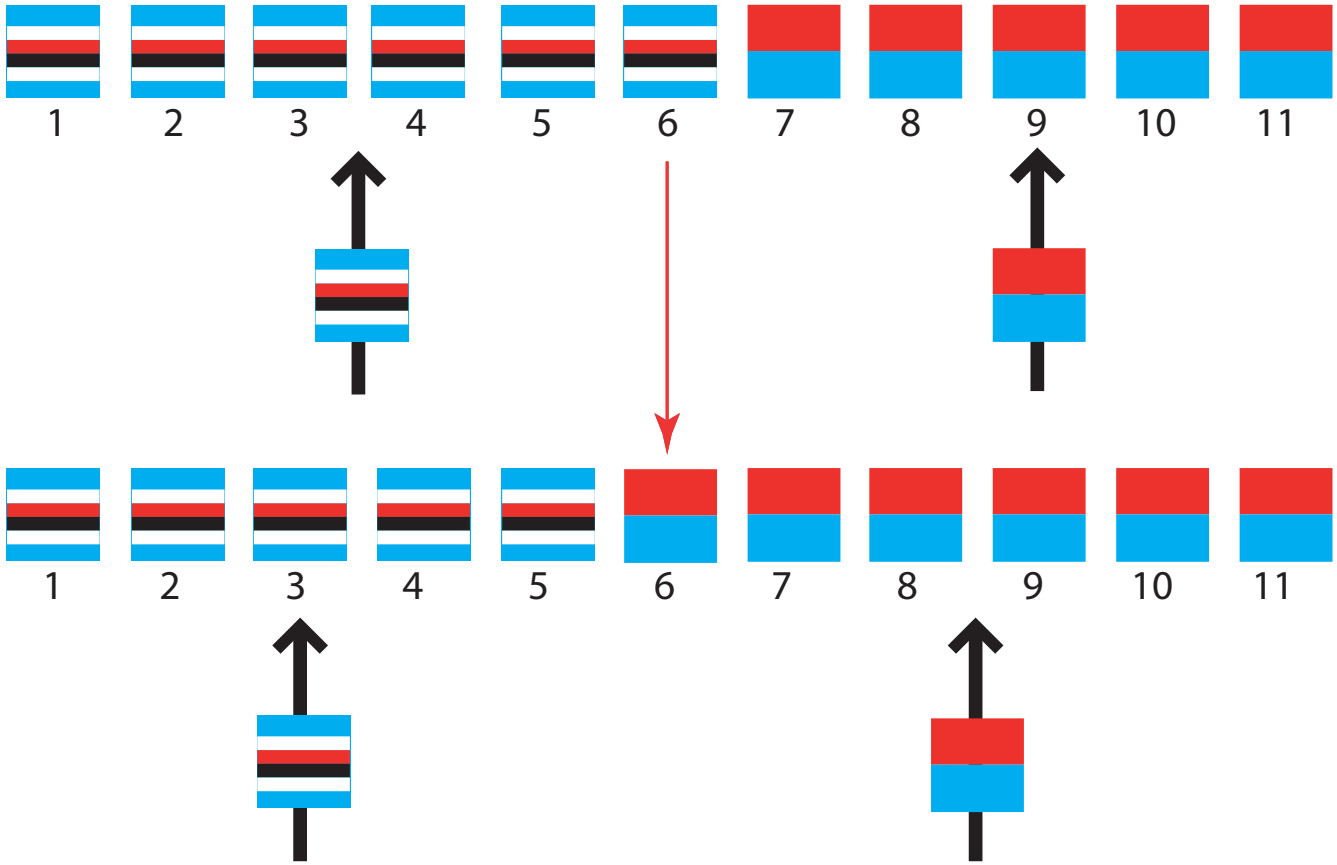
i genoani sono tutti più furbi dei doriani

disegnati in ordine di furbizia



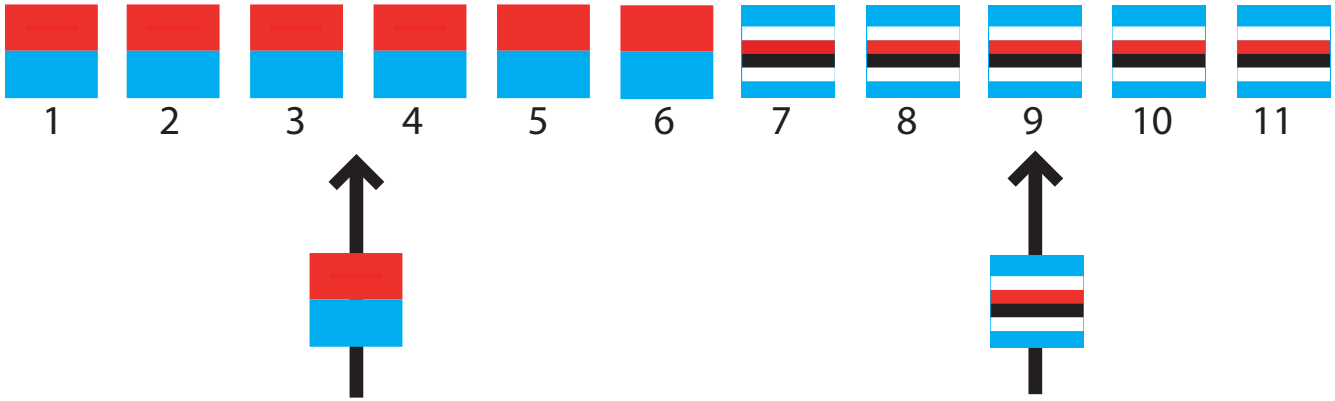
media doriani media genoani

se il più furbo dei doriani diventa genoano ...

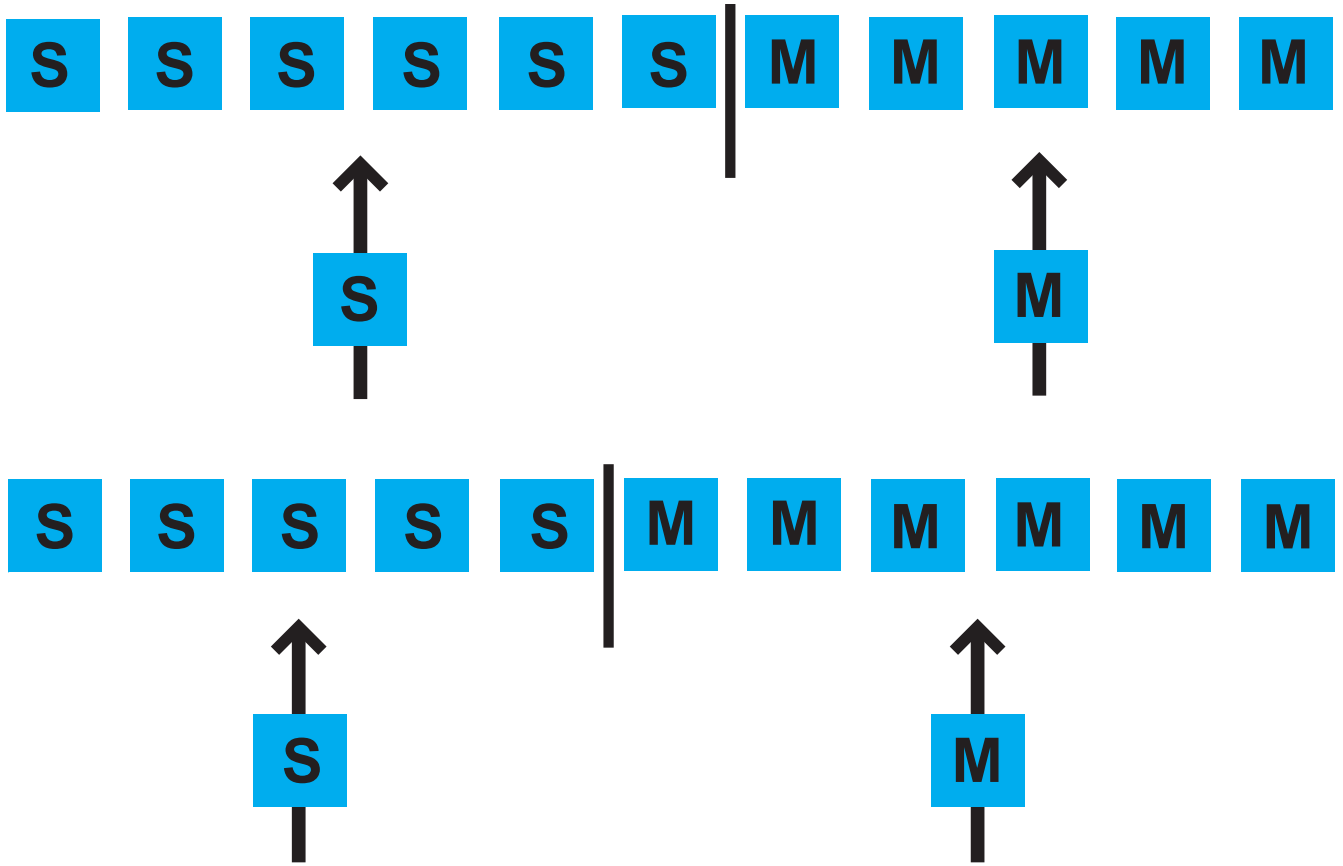


... abbassa la media della furbizia nei due gruppi!!!!

Preferite questo?



È come se avessimo cambiato la soglia della furbizia o della malattia (es. colesterolo nel sangue)



È diminuita la media del colesterolo fra i malati???
Le nuove cure sono migliori???

Il paradosso è attribuito al commediante Will Rogers (1985)

> *When the Okies left Oklahoma and moved to California, they raised the average intelligence level in both states.*

È stato rilevato nell'ambito dello studio della Sclerosi multipla da M.P. Sormani (Università di Genova) e altri nel 2008

E allora aveva ragione Trilussa?

Trilussa (1851-1950) poeta dialettale romano diceva

> *"da li conti che se fanno - seconno le statistiche d'adesso - risurta che te tocca un pollo all'anno: - e, se nun entra nelle spese tue, - t'entra ne la statistica lo stesso - perch'è c'è un antro che ne magna due."*

... non ci si può fidare di niente?

NO, i dati – come i fatti del mondo – devono essere letti e interpretati con attenzione.

Inoltre la conoscenza approfondita delle metodologie statistiche rende consapevoli anche della loro applicabilità alle situazioni concrete.

Un'altra ragione per studiare Statistica

What I like of statistics is that you can do some of the **formality and beauty of mathematics** *but you can also actually* **go out and think and solve problems**.

(Jane Hutton – University of Warwick – Statistica – premiata nell'ottobre 2016 dall'Imperial College of London fra le 12 donne dell'anno distinte nel campo della matematica e dell'informatica)