

Statistica Descrittiva

Stefania Bartoletti

28 Febbraio 2022

Popolazione

- ▶ In statistica, si è di solito interessati ad ottenere informazioni riguardo un collettivo (un insieme di elementi con qualche caratteristica in comune o soggetti a un comune fenomeno), a cui ci riferiamo col termine **popolazione**.

Popolazione

- ▶ In statistica, si è di solito interessati ad ottenere informazioni riguardo un collettivo (un insieme di elementi con qualche caratteristica in comune o soggetti a un comune fenomeno), a cui ci riferiamo col termine **popolazione**.
- ▶ L'indagine statistica è lo strumento statistico mediante il quale si acquisiscono informazioni su uno o più fenomeni attinenti ad una popolazione. Lo scopo dell'indagine è quello di produrre statistiche, ovvero **descrizioni riassuntive** di carattere quantitativo, riguardanti il collettivo di interesse.

Campione

- ▶ Tuttavia, la popolazione è di solito troppo grande per pensare di esaminare singolarmente i suoi elementi. Per esempio, potremmo essere interessati ad esaminare tutti i residenti di una regione per scopi demografici, oppure tutte le lampadine prodotte in una ditta per studiare gli effetti di un guasto.

Campione

- ▶ Tuttavia, la popolazione è di solito troppo grande per pensare di esaminare singolarmente i suoi elementi. Per esempio, potremmo essere interessati ad esaminare tutti i residenti di una regione per scopi demografici, oppure tutte le lampadine prodotte in una ditta per studiare gli effetti di un guasto.
- ▶ In questi casi, anziché intervistare ogni singolo residente o esaminare ogni singola lampadina, si va ad esaminare un sottoinsieme di elementi, a cui ci riferiamo col termine **campione**.

Campione rappresentativo

- ▶ La selezione del campione è di fondamentale importanza affinché questo sia informativo della popolazione, ovvero rappresentativo.

Campione rappresentativo

- ▶ La selezione del campione è di fondamentale importanza affinché questo sia informativo della popolazione, ovvero **rappresentativo**.
- ▶ Supponiamo di voler studiare la distribuzione di età dei residenti di Ferrara. Per selezionare un campione, intervistiamo le prime 100 persone che entrano nella Biblioteca Ariostea, e scopriamo che l'età media di queste 100 persone è 40.
- ▶ Possiamo concludere che questa sia una buona approssimazione dell'età media della popolazione di Ferrara?

Campione rappresentativo

- ▶ Affinché si possa definire rappresentativo della popolazione, il campione deve essere scelto nella maniera più casuale possibile
- ▶ Questo perchè ogni regola per scegliere il campione può risultare in un bias (distorsione, scostamento) dei valori di un dato
- ▶ Nella pratica, per esempio, non ha senso forzare il campione affinché la metà del campione sia maschio e l'altra metà femmina. Dovremmo sperare di scegliere in campione abbastanza casualmente da rappresentare “per caso” le percentuali vere. Una volta scelto il campione, la **statistica descrittiva** va a fornire una descrizione riassuntive riguardo la caratteristica di interesse.

Statistica Descrittiva

- ▶ I dati di un'indagine statistica sono solitamente organizzati in strutture come liste, tabelle, database con etichette e numeri.

Statistica Descrittiva

- ▶ I dati di un'indagine statistica sono solitamente organizzati in strutture come liste, tabelle, database con etichette e numeri.
- ▶ Per dare un senso ad un database di dati, ci si avvale di statistiche volte a sintetizzarli insieme a strumenti grafici come istogrammi, diagrammi, curve.

Statistica Descrittiva

- ▶ I dati di un'indagine statistica sono solitamente organizzati in strutture come liste, tabelle, database con etichette e numeri.
- ▶ Per dare un senso ad un database di dati, ci si avvale di statistiche volte a sintetizzarli insieme a strumenti grafici come istogrammi, diagrammi, curve.
- ▶ Questi strumenti sono utili per **descrivere** i dati, per esplorarli e comunicarli.

Indice

- ▶ Definizioni di popolazione e variabili
- ▶ Frequenza assoluta e relativa
- ▶ Funzione di distribuzione
- ▶ Indici statistici di sintesi
- ▶ Dati bivariati
- ▶ Coefficiente di correlazione campionaria

Definizioni

- ▶ **Popolazione:** l'insieme di tutti i possibili individui (unità, oggetti) dell'indagine statistica

Definizioni

- ▶ **Popolazione:** l'insieme di tutti i possibili individui (unità, oggetti) dell'indagine statistica
- ▶ **Campione:** sottoinsieme della popolazione preso in considerazione nell'indagine statistica

Definizioni

- ▶ **Popolazione:** l'insieme di tutti i possibili individui (unità, oggetti) dell'indagine statistica
- ▶ **Campione:** sottoinsieme della popolazione preso in considerazione nell'indagine statistica
- ▶ **Variabile:** caratteristica di un individuo, che varia da individuo ad individuo

Definizioni

- ▶ **Popolazione**: l'insieme di tutti i possibili individui (unità, oggetti) dell'indagine statistica
- ▶ **Campione**: sottoinsieme della popolazione preso in considerazione nell'indagine statistica
- ▶ **Variabile**: caratteristica di un individuo, che varia da individuo ad individuo
 - ▶ **Quantitativa**: misurabile

Definizioni

- ▶ **Popolazione**: l'insieme di tutti i possibili individui (unità, oggetti) dell'indagine statistica
- ▶ **Campione**: sottoinsieme della popolazione preso in considerazione nell'indagine statistica
- ▶ **Variabile**: caratteristica di un individuo, che varia da individuo ad individuo
 - ▶ **Quantitativa**: misurabile
 - ▶ continue (es. peso e altezza)

Definizioni

- ▶ **Popolazione**: l'insieme di tutti i possibili individui (unità, oggetti) dell'indagine statistica
- ▶ **Campione**: sottoinsieme della popolazione preso in considerazione nell'indagine statistica
- ▶ **Variabile**: caratteristica di un individuo, che varia da individuo ad individuo
 - ▶ **Quantitativa**: misurabile
 - ▶ continue (es. peso e altezza)
 - ▶ discrete (es. numero dei viaggi)

Definizioni

- ▶ **Popolazione**: l'insieme di tutti i possibili individui (unità, oggetti) dell'indagine statistica
- ▶ **Campione**: sottoinsieme della popolazione preso in considerazione nell'indagine statistica
- ▶ **Variabile**: caratteristica di un individuo, che varia da individuo ad individuo
 - ▶ **Quantitativa**: misurabile
 - ▶ continue (es. peso e altezza)
 - ▶ discrete (es. numero dei viaggi)
 - ▶ **Qualitativa**: non misurabile

Definizioni

- ▶ **Popolazione**: l'insieme di tutti i possibili individui (unità, oggetti) dell'indagine statistica
- ▶ **Campione**: sottoinsieme della popolazione preso in considerazione nell'indagine statistica
- ▶ **Variabile**: caratteristica di un individuo, che varia da individuo ad individuo
 - ▶ **Quantitativa**: misurabile
 - ▶ continue (es. peso e altezza)
 - ▶ discrete (es. numero dei viaggi)
 - ▶ **Qualitativa**: non misurabile
 - ▶ ordinali (es. i giorni della settimana)

Definizioni

- ▶ **Popolazione**: l'insieme di tutti i possibili individui (unità, oggetti) dell'indagine statistica
- ▶ **Campione**: sottoinsieme della popolazione preso in considerazione nell'indagine statistica
- ▶ **Variabile**: caratteristica di un individuo, che varia da individuo ad individuo
 - ▶ **Quantitativa**: misurabile
 - ▶ continue (es. peso e altezza)
 - ▶ discrete (es. numero dei viaggi)
 - ▶ **Qualitativa**: non misurabile
 - ▶ ordinali (es. i giorni della settimana)
 - ▶ nominali (es. colore degli occhi)

Definizioni

- ▶ **Popolazione**: l'insieme di tutti i possibili individui (unità, oggetti) dell'indagine statistica
- ▶ **Campione**: sottoinsieme della popolazione preso in considerazione nell'indagine statistica
- ▶ **Variabile**: caratteristica di un individuo, che varia da individuo ad individuo
 - ▶ **Quantitativa**: misurabile
 - ▶ continue (es. peso e altezza)
 - ▶ discrete (es. numero dei viaggi)
 - ▶ **Qualitativa**: non misurabile
 - ▶ ordinali (es. i giorni della settimana)
 - ▶ nominali (es. colore degli occhi)
- ▶ **Classe**: se le variabili sono continue o possono assumere troppi valori, si dividono in classi

Indagine: quanto leggono i giovani italiani (15-34yrs)

- ▶ **Popolazione**: tutti i giovani italiani dai 15 ai 34 anni
- ▶ **Campione**: sottoinsieme di giovani italiani dai 15 ai 34 anni
- ▶ **Variabile**: numero di libri letti nel 2019 (quantitativa, discreta)

Definizioni

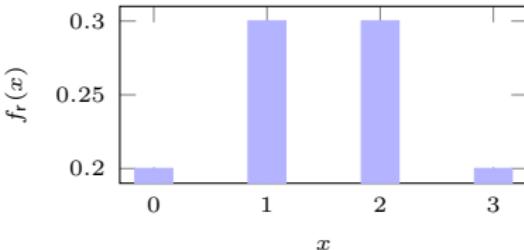
- ▶ n : numero di individui nel campione
- ▶ x_i : valore assunto dalla variabile per l' i -simo individuo
- ▶ **Frequenza assoluta**: la frequenza assoluta di un valore x è il numero di individui per cui la variabile assume tale valore
 $f_a(x) = |\{i : x_i = x\}|$
- ▶ **Frequenza relativa**: la frequenza assoluta rapportata al numero totale di individui del campione

$$f_r(x) = \frac{f_a(x)}{n}$$

Indagine: quanto leggono i giovani italiani (15-34yrs)

Id	Anni	Numero Libri
1	24	3
2	20	0
3	16	2
4	19	1
5	31	2
6	17	3
7	33	1
8	32	1
9	22	0
10	21	2

x	Frequenza assoluta	Frequenza relativa
0	2	0.2
1	3	0.3
2	3	0.3
3	2	0.2

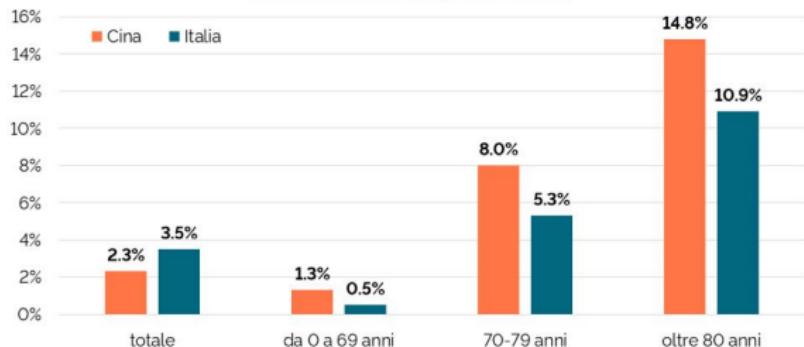


Tasso di letalità

- ▶ **Popolazione:** tutti i contagiati da COVID-19 in varie classi di età
- ▶ **Campione:** sottoinsieme dei contagiati da COVID-19 nelle classi di età considerate (tampone effettuato e analizzato da ISS)
- ▶ **Variabile:** età (quantitativa) e stato di salute (qualitativa)
- ▶ **Classe:**
 - ▶ classi di età (0 – 69), (70 – 79), (> 80)
 - ▶ stato di salute: deceduti (*d*), in vita (*v*)

Tasso di letalità

COVID-19
Tasso di letalità in Italia e in Cina

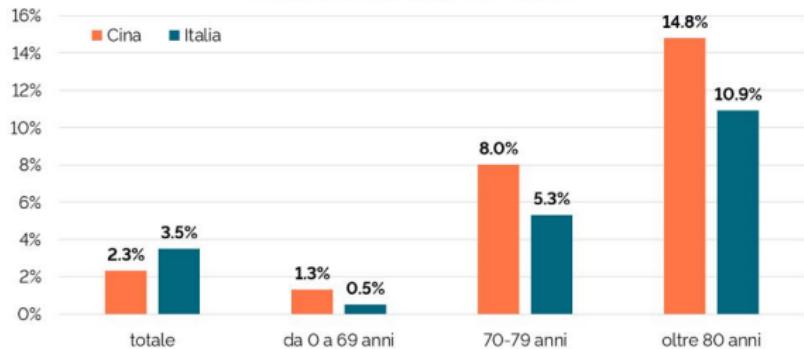


Dati: Istituto superiore di sanità.

(istogramma da inizio pandemia, marzo 2020)

Tasso di letalità

COVID-19
Tasso di letalità in Italia e in Cina



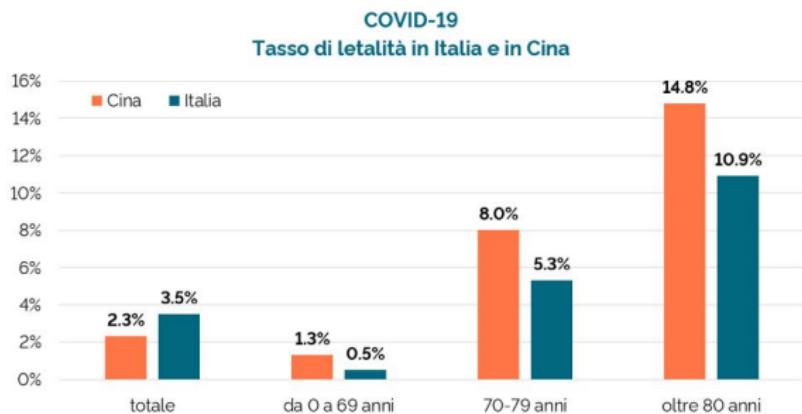
Dati: Istituto superiore di sanità.

$$f_r = \frac{|\{i : x_i^{(\text{tot})} = d\}|}{n}$$

$$f_r^{(k)} = \frac{|\{i : x_i^{(k)} = d\}|}{n^{(k)}}$$

$$k = 1, 2, 3$$

Attenti al campione!

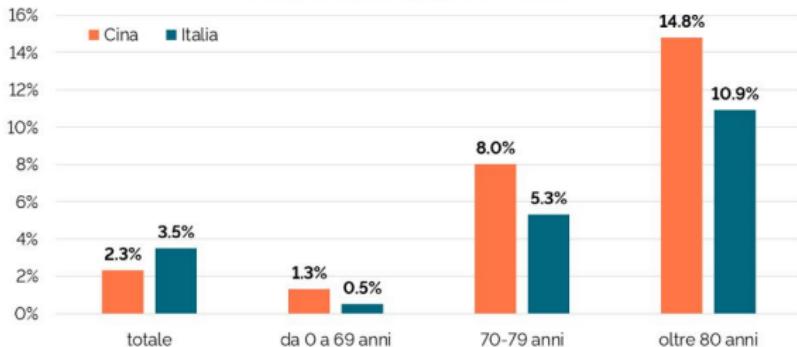


Dati: Istituto superiore di sanità.

$$f_r = \frac{f_a^{(1)} + f_a^{(2)} + f_a^{(3)}}{n} = \frac{f_r^{(1)} n_1 + f_r^{(2)} n_2 + f_r^{(3)} n_3}{n}$$

Attenti al campione!

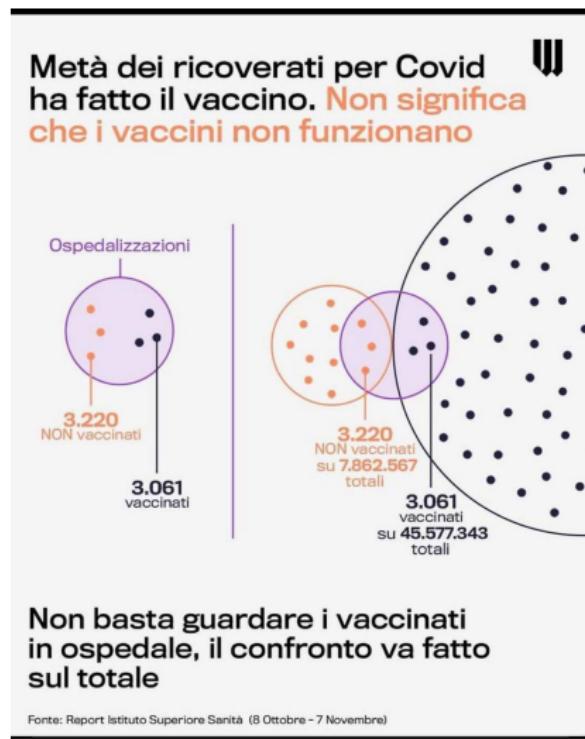
COVID-19
Tasso di letalità in Italia e in Cina



Dati: Istituto superiore di sanità.

- ▶ Su 100 persone italiane, quante sono quelle di età compresa tra 0 e 69 anni, se confrontate con quelle cinesi?

Attenti al campione!



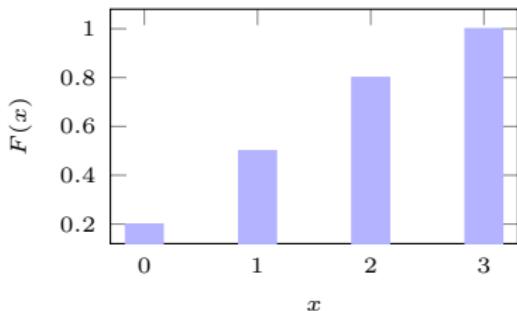
Funzione cumulativa empirica

La funzione cumulativa empirica (funzione di ripartizione empirica) $F(x)$ associa al valore x la frazione di unità nel campione che sono minori o uguali a x

$$F(x) = \frac{|\{i : x_i \leq x\}|}{n} = \sum_{z \leq x} f_r(z)$$

Toy Example:

x	f_a	f_r	$F(x)$
0	2	0.2	0.2
1	3	0.3	0.5
2	3	0.3	0.8
3	2	0.2	1



Funzione cumulativa empirica

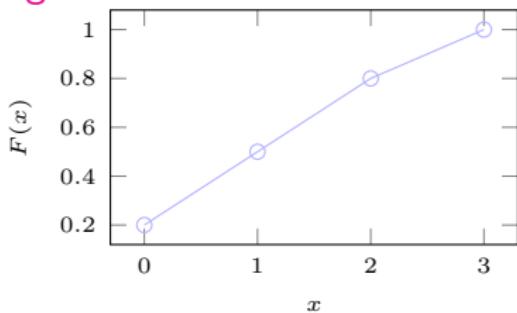
La funzione cumulativa empirica (funzione di ripartizione empirica) $F(x)$ associa al valore x la frazione di unità nel campione che sono minori o uguali a x

$$F(x) = \frac{|\{i : x_i \leq x\}|}{n} = \sum_{z \leq x} f_r(z)$$

Toy Example:

x	f_a	f_r	$F(x)$
0	2	0.2	0.2
1	3	0.3	0.5
2	3	0.3	0.8
3	2	0.2	1

Ogiva:



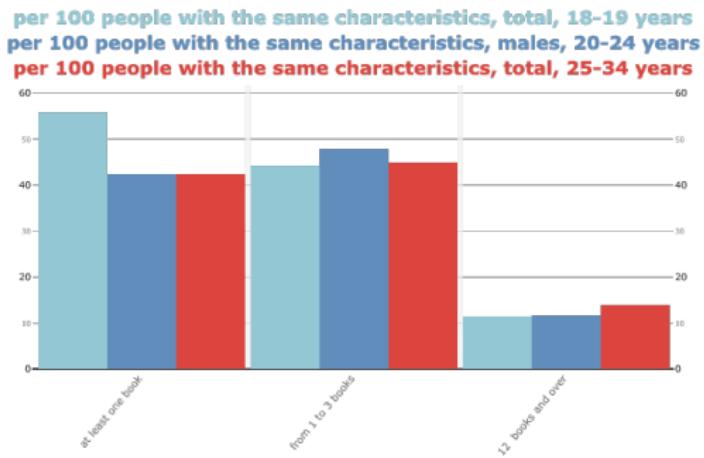
Indagine: quanto leggono i giovani italiani (15-34yrs)

Dataset: Aspetti della vita quotidiana

	Misura	per 100 persone con le stesse caratteristiche			
	Sesso	totale			
	Selezione periodo	2019			
	Tipo dato	persone di 6 anni e più per lettura di libri negli	almeno un libro	da 1 a 3 libri	12 e più libri
Classe di età					
15-17 anni		54,1	44,9		12,2
18-19 anni		55,9	44,3		11,4
20-24 anni		50,5	43,7		12,3
25-34 anni		42,5	44,9		14,1

Dati estratti il 22 feb 2020 12:31 UTC (GMT) da I.Stat

Indagine: quanto leggono i giovani italiani (15-34yrs)



Source: ISTAT, Aspetti della vita quotidiana

Indici statistici di sintesi

Un insieme di dati $\{x_1, x_2, \dots, x_n\}$ può essere descritto da indici statistici di sintesi:

- ▶ Misure di tendenza centrale
- ▶ Misure di dispersione
- ▶ Misure di forma

Misure di tendenza centrale

Descrivono il centro di un insieme di dati $\{x_1, x_2, \dots, x_n\}$

1. Media campionaria

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

La media campionaria è un indice **lineare**: prese due costanti a e b e un nuovo insieme di dati costituito da $\{y_1, y_2, \dots, y_n\}$ con $y_i = ax_i + b$, la media campionaria del nuovo insieme è

$$\bar{y} = a\bar{x} + b$$

Misure di tendenza centrale

Descrivono il centro di un insieme di dati $\{x_1, x_2, \dots, x_n\}$

1. Media campionaria

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

La media campionaria è un indice **lineare**: prese due costanti a e b e un nuovo insieme di dati costituito da $\{y_1, y_2, \dots, y_n\}$ con $y_i = ax_i + b$, la media campionaria del nuovo insieme è

$$\bar{y} = a\bar{x} + b$$

Misure di tendenza centrale

Descrivono il centro di un insieme di dati $\{x_1, x_2, \dots, x_n\}$

1. Media campionaria

Siano $v \in \mathcal{X}$ i valori assunti dalla variabile x con frequenza relativa $f_r(x)$

$$\bar{x} = \sum_{v \in \mathcal{X}} v \cdot f_r(v)$$

Misure di tendenza centrale

Descrivono il centro di un insieme di dati $\{x_1, x_2, \dots, x_n\}$

1. Media campionaria

Siano $v \in \mathcal{X}$ i valori assunti dalla variabile x con frequenza relativa $f_r(x)$

$$\bar{x} = \sum_{v \in \mathcal{X}} v \cdot f_r(v)$$

Toy Example:

x	Fr. Ass.	$f_r(x)$	$F(x)$
0	2	0.2	0.2
1	3	0.3	0.5
2	3	0.3	0.8
3	2	0.2	1

Misure di tendenza centrale

Descrivono il centro di un insieme di dati $\{x_1, x_2, \dots, x_n\}$

1. Media campionaria

Siano $v \in \mathcal{X}$ i valori assunti dalla variabile x con frequenza relativa $f_r(x)$

$$\bar{x} = \sum_{v \in \mathcal{X}} v \cdot f_r(v)$$

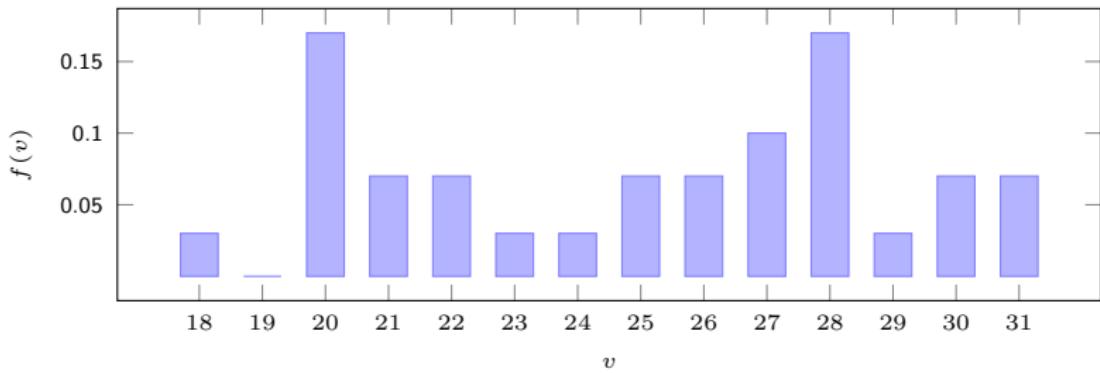
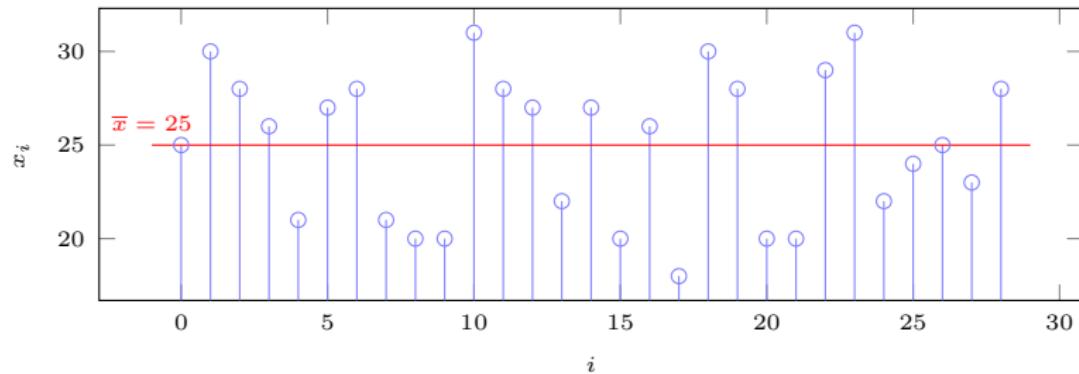
Toy Example:

x	Fr. Ass.	$f_r(x)$	$F(x)$
0	2	0.2	0.2
1	3	0.3	0.5
2	3	0.3	0.8
3	2	0.2	1

$$\bar{x} = \sum_{v \in \mathcal{X}} v \cdot f(v) = 0 \cdot 0.2 + 1 \cdot 0.3 + 2 \cdot 0.3 + 3 \cdot 0.2 = 1.5$$

Voti degli studenti che hanno superato un esame

$$n = 29, \mathcal{X} = \{18, 19, \dots, 30, 31\}$$



Misure di tendenza centrale

Descrivono il centro di un insieme di dati $\{x_1, x_2, \dots, x_n\}$

2. Mediana campionaria Si ordini l'insieme di dati in ordine crescente, i.e.

$$x_{o(1)} \leq x_{o(2)} \leq \dots \leq x_{o(n-1)} \leq x_{o(n)}$$

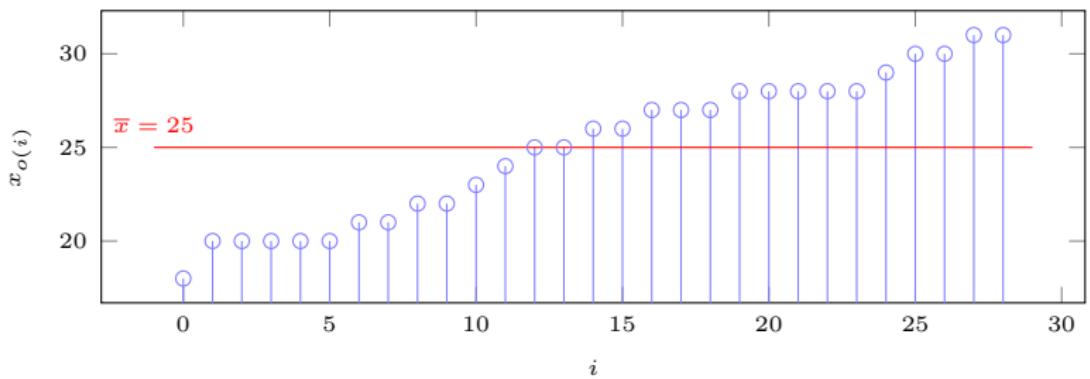
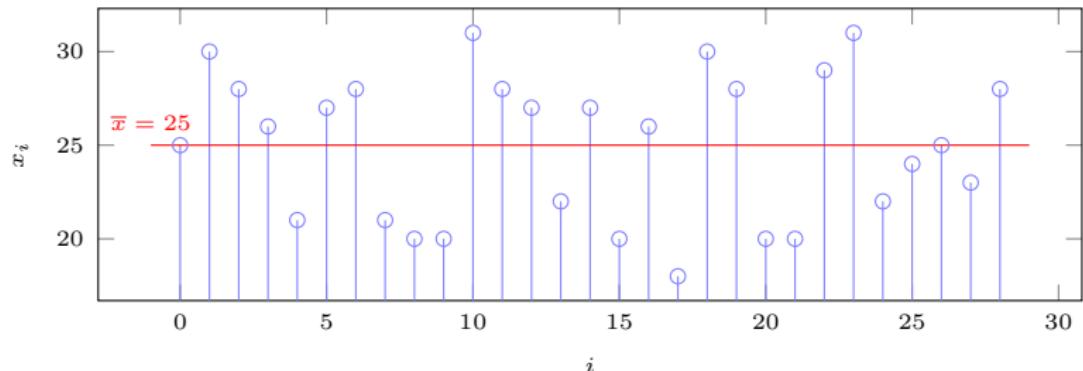
Toy Example:

$$\mathbf{x} = \{3, 0, 2, 1, 2, 3, 1, 1, 0, 2\}$$

$$\{0, 0, 1, 1, 1, 2, 2, 2, 3, 3\}$$

$$o(1) = 2, o(2) = 9, o(3) = 4, o(4) = 7, o(5) = 8 \dots$$

Voti degli studenti che hanno superato un esame



Misure di tendenza centrale

Descrivono il centro di un insieme di dati $\{x_1, x_2, \dots, x_n\}$

2. **Mediana campionaria** Si ordini l'insieme di dati in ordine crescente, i.e.

$$x_{o(1)} \leq x_{o(2)} \leq \dots \leq x_{o(n-1)} \leq x_{o(n)}$$

La mediana è il valore centrale della successione di dati ordinata

$$x_{\text{med}} = \begin{cases} x_{o((n+1)/2)} & n \text{ dispari} \\ \frac{x_{o(n/2)} + x_{o(n/2+1)}}{2} & n \text{ pari} \end{cases}$$

Misure di tendenza centrale

Descrivono il centro di un insieme di dati $\{x_1, x_2, \dots, x_n\}$

2. **Mediana campionaria** Si ordini l'insieme di dati in ordine crescente, i.e.

$$x_{o(1)} \leq x_{o(2)} \leq \dots \leq x_{o(n-1)} \leq x_{o(n)}$$

La mediana è il valore centrale della successione di dati ordinata

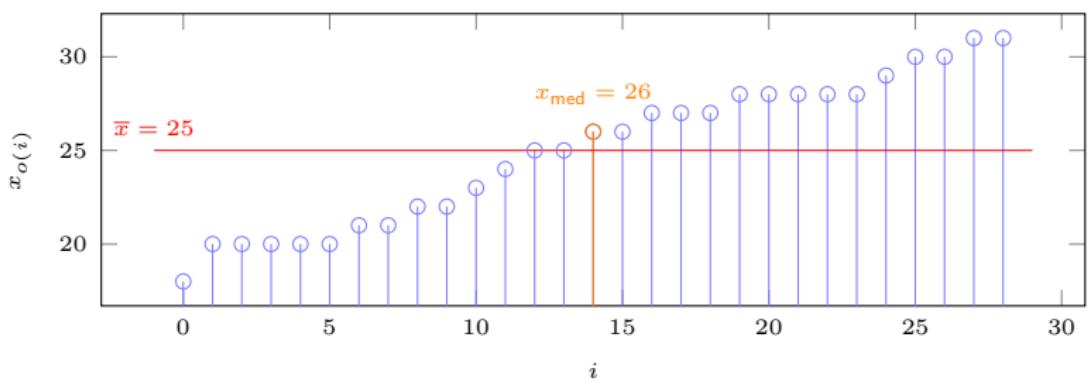
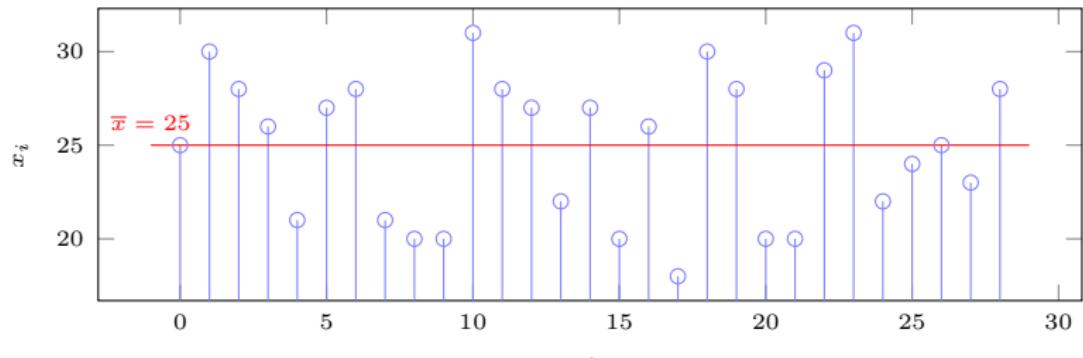
$$x_{\text{med}} = \begin{cases} x_{o((n+1)/2)} & n \text{ dispari} \\ \frac{x_{o(n/2)} + x_{o(n/2+1)}}{2} & n \text{ pari} \end{cases}$$

Toy Example:

$$\{0, 0, 1, 1, 1, 2, 2, 2, 3, 3\}, n = 10$$

$$x_{\text{med}} = 1.5$$

Voti degli studenti che hanno superato un esame



Misure di tendenza centrale

Descrivono il centro di un insieme di dati $\{x_1, x_2, \dots, x_n\}$

3. Moda campionaria

La *moda campionaria* di un insieme di dati, se esiste, è l'unico valore che ha frequenza massima. Se non vi è un solo valore con frequenza massima, ciascuno di essi è detto *valore modale*.

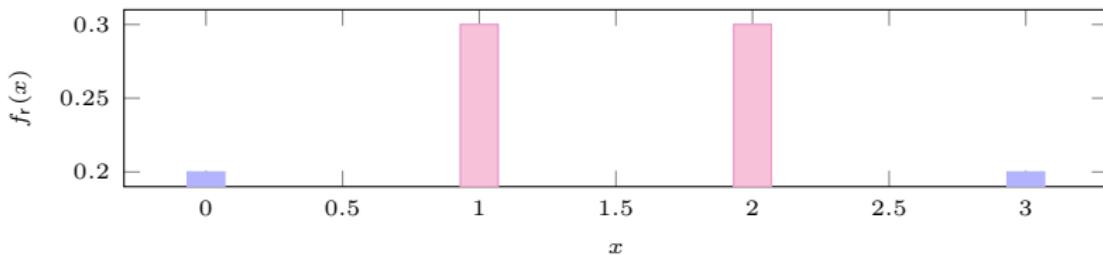
Misure di tendenza centrale

Descrivono il centro di un insieme di dati $\{x_1, x_2, \dots, x_n\}$

3. Moda campionaria

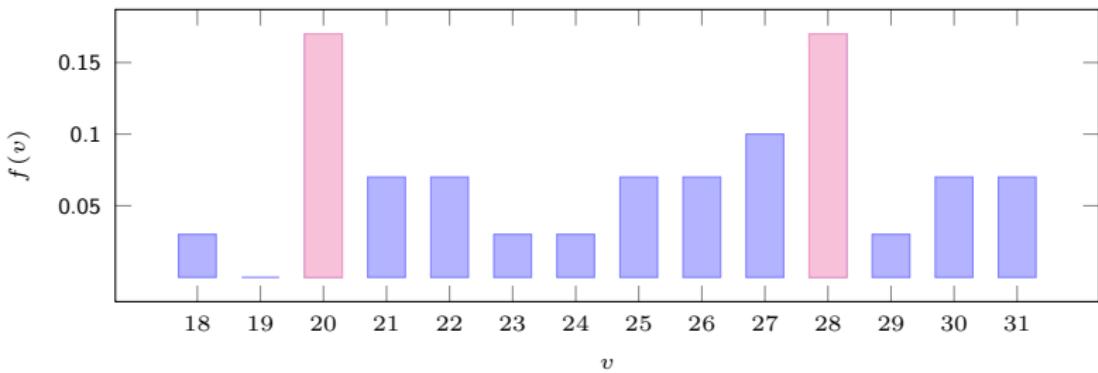
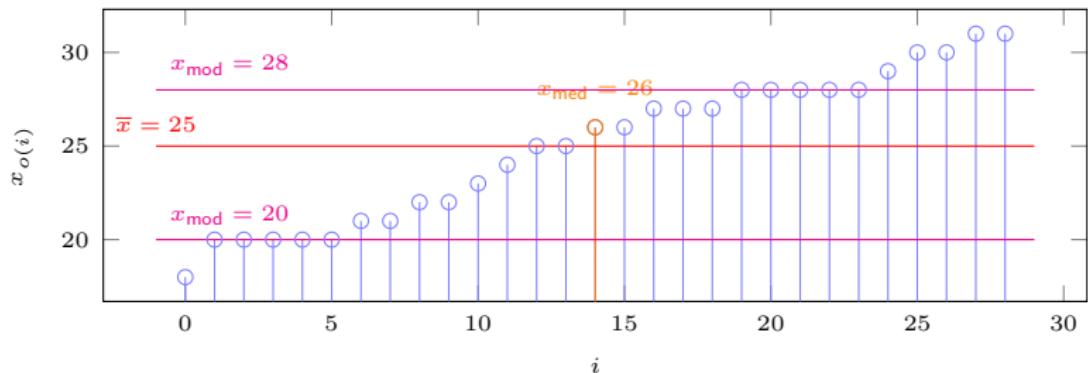
La *moda campionaria* di un insieme di dati, se esiste, è l'unico valore che ha frequenza massima. Se non vi è un solo valore con frequenza massima, ciascuno di essi è detto *valore modale*.

Toy Example:



Voti degli studenti che hanno superato un esame

$$n = 29, \mathcal{X} = \{18, 19, \dots, 30, 31\}$$



Indici statistici di sintesi

Un insieme di dati $\{x_1, x_2, \dots, x_n\}$ può essere descritto da indici statistici di sintesi:

- ▶ **Misure di tendenza centrale**

- ▶ media campionaria
- ▶ mediana campionaria
- ▶ moda campionaria

Indici statistici di sintesi

Un insieme di dati $\{x_1, x_2, \dots, x_n\}$ può essere descritto da indici statistici di sintesi:

- ▶ Misure di tendenza centrale
 - ▶ media campionaria
 - ▶ mediana campionaria
 - ▶ moda campionaria
- ▶ Misure di dispersione

Misure di dispersione

Descrivono quanto i valori in $\{x_1, x_2, \dots, x_n\}$ sono concentrati rispetto alla media.

Misure di dispersione

Descrivono quanto i valori in $\{x_1, x_2, \dots, x_n\}$ sono concentrati rispetto alla media.

1. Varianza campionaria

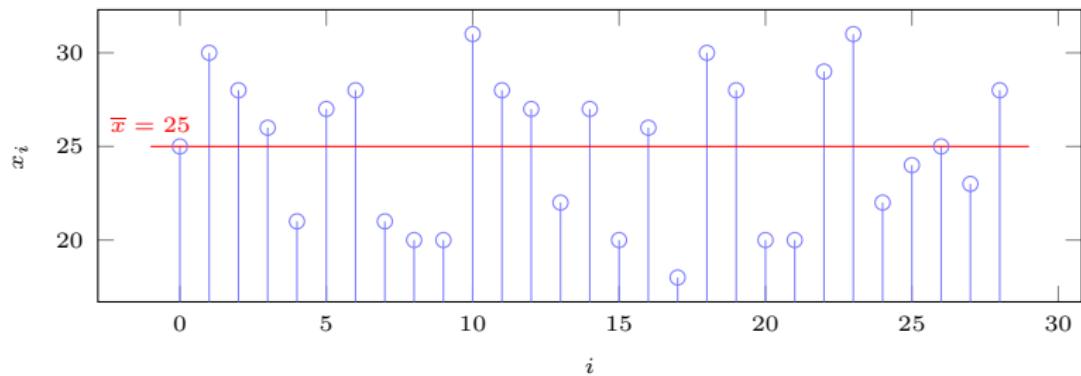
$$s_x^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$$

Misure di dispersione

Descrivono quanto i valori in $\{x_1, x_2, \dots, x_n\}$ sono concentrati rispetto alla media.

1. Varianza campionaria

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



Misure di dispersione

Descrivono quanto i valori in $\{x_1, x_2, \dots, x_n\}$ sono concentrati rispetto alla media.

1. Varianza campionaria

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

La varianza campionaria **non** è un indice lineare: prese due costanti a e b e un nuovo insieme di dati costituito da $\{y_1, y_2, \dots, y_n\}$ con $y_i = ax_i + b$, la varianza campionaria del nuovo insieme è

$$s_y^2 = a^2 s_x^2$$

Misure di dispersione

Descrivono quanto siano concentrati rispetto alla media i valori in $\{x_1, x_2, \dots, x_n\}$

1. Varianza campionaria

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

Toy Example:

$$\mathbf{x} = \{3, 0, 2, 1, 2, 3, 1, 1, 0, 2\}$$

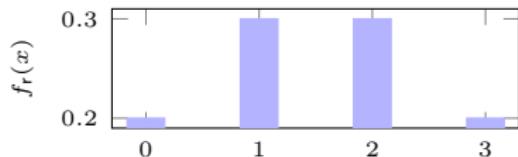
$$\bar{x} = 1.5$$

$$s_x^2 = 1.17$$

$$\mathbf{y} = \{3, 0, 2, 2, 2, 2, 1, 1, 0, 2\}$$

$$\bar{y} = 1.5$$

$$s_y^2 = 0.94$$



Misure di dispersione

Descrivono quanto siano concentrati rispetto alla media i valori in $\{x_1, x_2, \dots, x_n\}$

1. Varianza campionaria

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

Misure di dispersione

Descrivono quanto siano concentrati rispetto alla media i valori in $\{x_1, x_2, \dots, x_n\}$

1. Varianza campionaria

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

1b. Deviazione standard campionaria

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

La deviazione standard ha la stessa unità di misura della variabile.

Disuguaglianza di Chebyshev (versione campionaria)

- ▶ Sia $\{x_1, x_2, \dots, x_n\}$ un insieme di dati con media \bar{x} e deviazione standard campionaria $s_x > 0$
- ▶ Si consideri $\mathcal{S}_k = \{i : |x_i - \bar{x}| < ks_x\}$, con cardinalità $|\mathcal{S}_k|$
- ▶ Allora

$$\frac{|\mathcal{S}_k|}{n} \geq 1 - \frac{n-1}{nk^2} > 1 - \frac{1}{k^2} \quad (1)$$

Disuguaglianza di Chebyshev (versione campionaria)

- ▶ Sia $\{x_1, x_2, \dots, x_n\}$ un insieme di dati con media \bar{x} e deviazione standard campionaria $s_x > 0$
- ▶ Si consideri $\mathcal{S}_k = \{i : |x_i - \bar{x}| < ks_x\}$, con cardinalità $|\mathcal{S}_k|$
- ▶ Allora

$$\frac{|\mathcal{S}_k|}{n} \geq 1 - \frac{n-1}{nk^2} > 1 - \frac{1}{k^2} \quad (1)$$

Toy Example:

$$\mathbf{x} = \{3, 0, 2, 1, 2, 3, 1, 1, 0, 2\}$$

$$\bar{x} = 1.5$$

$$s_x = 1.08$$

$$k = 1, ks_x = 1.08$$

$$k = 2, ks_x = 2.16$$

$$\frac{|\mathcal{S}_1|}{10} \geq 1 - \frac{9}{10} > 1 - \frac{1}{1}$$

$$\frac{|\mathcal{S}_2|}{10} \geq 1 - \frac{9}{40} > 1 - \frac{1}{4} = 0.75$$

Disuguaglianza di Chebyshev (versione campionaria)

- ▶ Fornisce un limite **inferiore** alla frazione di dati di un campione che si allontanano dalla loro media campionaria più di un multiplo della deviazione standard
- ▶ Vale per tutti gli insiemi di dati
- ▶ In alcune situazioni in questo limite può essere notevolmente migliorato

Percentili

In un insieme di dati numerici, per ogni valore k esiste un dato che è contemporaneamente maggiore o uguale di almeno il k percento dei dati, e minore o uguale di almeno il $100 - k$ percento dei dati.

Percentili

In un insieme di dati numerici, per ogni valore k esiste un dato che è contemporaneamente maggiore o uguale di almeno il k percento dei dati, e minore o uguale di almeno il $100 - k$ percento dei dati.

Tale dato si definisce percentile k -esimo. Se non è unico, allora sono esattamente due, e in questo caso il percentile k -esimo è definito come la loro media aritmetica.

Calcolo del percentile k -esimo di un campione di numerosità n :

- ▶ $p = k/100$
- ▶ $\lceil np \rceil$ dati sono minori o uguali al percentile k -esimo
- ▶ $\lceil n(1 - p) \rceil$ dati sono maggiori o uguali al percentile k -esimo

Toy example:

$$0 \leq 0 \leq 1 \leq 1 \leq 1 \leq 2 \leq 2 \leq 2 \leq 3 \leq 3$$

$$k = 30, p = 0.3$$

Calcolo del percentile k -esimo di un campione di numerosità n :

- ▶ $p = k/100$
- ▶ $\lceil np \rceil$ dati sono minori o uguali al percentile k -esimo
- ▶ $\lceil n(1 - p) \rceil$ dati sono maggiori o uguali al percentile k -esimo

Toy example:

$$0 \leq 0 \leq 1 \leq 1 \leq 1 \leq 2 \leq 2 \leq 2 \leq 3 \leq 3$$

$$k = 30, p = 0.3$$

$$\lceil np \rceil = \lceil 10 \cdot 0.3 \rceil = \lceil 3 \rceil = 3$$

Calcolo del percentile k -esimo di un campione di numerosità n :

- ▶ $p = k/100$
- ▶ $\lceil np \rceil$ dati sono minori o uguali al percentile k -esimo
- ▶ $\lceil n(1 - p) \rceil$ dati sono maggiori o uguali al percentile k -esimo

Toy example:

$$0 \leq 0 \leq \boxed{1} \leq 1 \leq 1 \leq 2 \leq 2 \leq 2 \leq 3 \leq 3$$

$$k = 30, p = 0.3$$

$$\lceil np \rceil = \lceil 10 \cdot 0.3 \rceil = \lceil 3 \rceil = 3$$

Calcolo del percentile k -esimo di un campione di numerosità n :

- ▶ $p = k/100$
- ▶ $\lceil np \rceil$ dati sono minori o uguali al percentile k -esimo
- ▶ $\lceil n(1 - p) \rceil$ dati sono maggiori o uguali al percentile k -esimo

Toy example:

$$0 \leq 0 \leq \boxed{1} \leq 1 \leq 1 \leq 2 \leq 2 \leq 2 \leq 3 \leq 3$$

$$k = 30, p = 0.3$$

$$\lceil np \rceil = \lceil 10 \cdot 0.3 \rceil = \lceil 3 \rceil = 3$$

$$\lceil n(1 - p) \rceil = \lceil 10 \cdot 0.7 \rceil = \lceil 7 \rceil = 7$$

Calcolo del percentile k -esimo di un campione di numerosità n :

- ▶ $p = k/100$
- ▶ $\lceil np \rceil$ dati sono minori o uguali al percentile k -esimo
- ▶ $\lceil n(1 - p) \rceil$ dati sono maggiori o uguali al percentile k -esimo

Toy example:

$$0 \leq 0 \leq 1 \leq \boxed{1} \leq 1 \leq 2 \leq 2 \leq 2 \leq 3 \leq 3$$

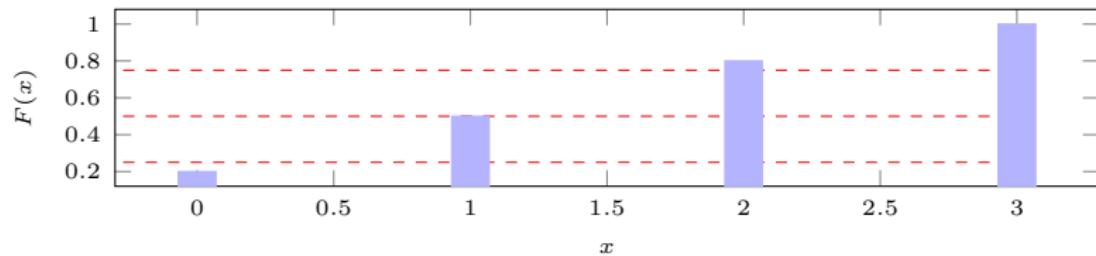
$$k = 30, p = 0.3$$

$$\lceil np \rceil = \lceil 9 \cdot 0.3 \rceil = \lceil 2.7 \rceil = 3$$

$$\lceil n(1 - p) \rceil = \lceil 9 \cdot 0.7 \rceil = \lceil 6.3 \rceil = 7$$

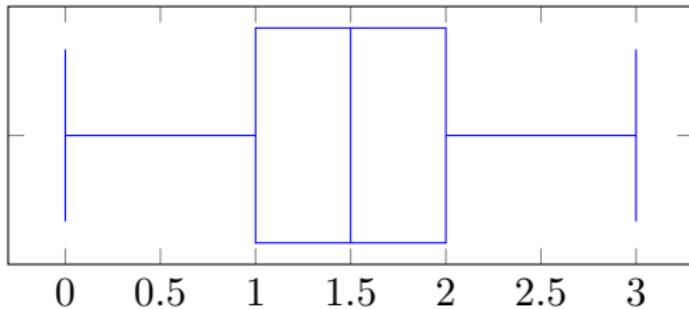
Quartili

- ▶ Il 25-esimo percentile si dice primo quartile;
- ▶ il 50-esimo si dice mediana campionaria o secondo quartile;
- ▶ il 75-esimo si dice terzo quartile.



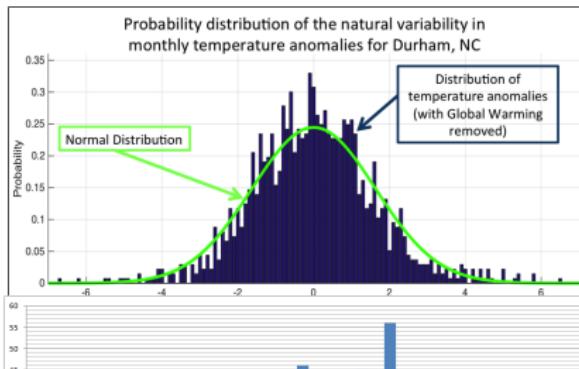
Box plot

Uno strumento utile a visualizzare alcune delle statistiche rappresentative dei dati è il box plot. Si ottiene sovrapponendo ad una linea orizzontale che va dal minore al maggiore dei dati, un rettangolo (il box) che va dal primo al terzo quartile, con una linea verticale che lo divide al livello del secondo quartile.



Misure di forma

In presenza di molti dati, che provengono dai contesti piú disparati, si puó notare una distribuzione con forma caratteristica: un solo massimo, in corrispondenza della mediana; la frequenza decresce da entrambi i lati simmetricamente, secondo una curva a campana.



Distribuzione normale

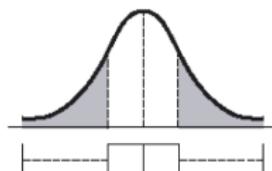
Se un campione numerico è approssimativamente **normale**, ha media campionaria \bar{x} e deviazione standard campionaria s_x , allora possiamo migliorare la diseguaglianza di Chebychev.

La **regola empirica** indica che:

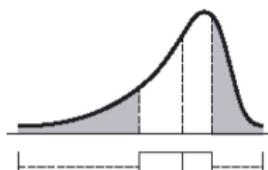
- ▶ Circa il **68%** dei dati cade nell'intervallo $\bar{x} \pm s$
- ▶ Circa il **95%** dei dati cade nell'intervallo $\bar{x} \pm 2s$
- ▶ Circa il **99.7%** dei dati cade nell'intervallo $\bar{x} \pm 3s$

Asimmetria: skewness

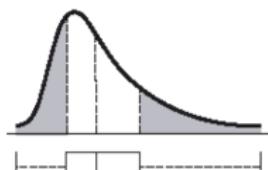
Se un insieme di dati presenta un istogramma che è sensibilmente asimmetrico rispetto alla mediana, si parla di campione **skewed** (ovvero sbilanciato).



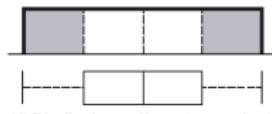
(a) Distribuzione simmetrica a campana



(b) Distribuzione obliqua a sinistra



(c) Distribuzione obliqua a destra



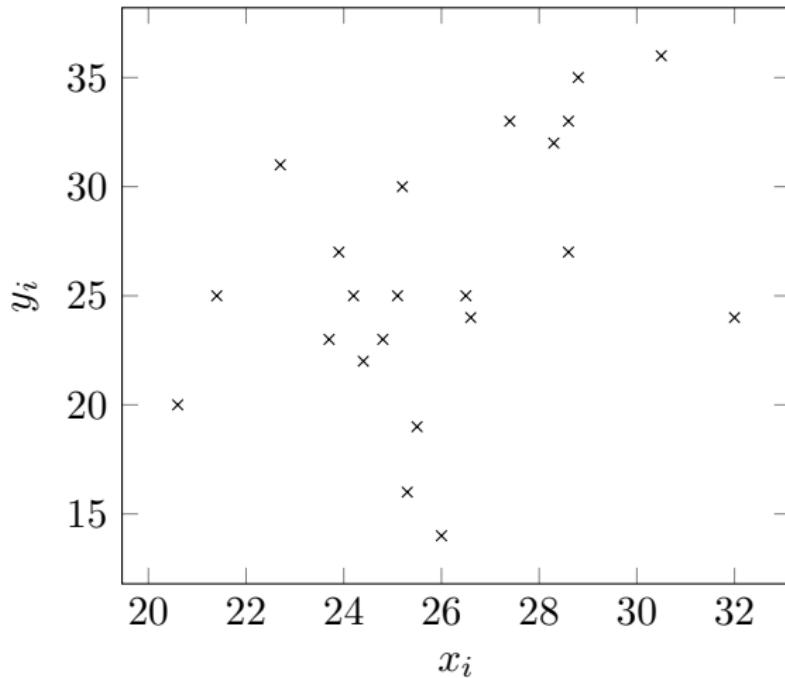
(d) Distribuzione uniforme (rettangolare)

Dati bivariati

Se siamo interessati alla descrizione di coppie di variabili che hanno tra loro qualche associazione, ogni coppia è da considerarsi un'osservazione

- ▶ (x_i, y_i) : i -esima osservazione
- ▶ Il campione è detto **bivariato**.

Diagramma di dispersione



Il diagramma di dispersione può fornire una prima risposta intuitiva ma grossolana sull' associazione tra le due variabili.

Covarianza campionaria

Sia dato un campione bivariato (x_i, y_i) per $i = 1, 2, \dots, n$, dove le singole variabili hanno medie campionarie \bar{x} e \bar{y} .

Covarianza campionaria

Sia dato un campione bivariato (x_i, y_i) per $i = 1, 2, \dots, n$, dove le singole variabili hanno medie campionarie \bar{x} e \bar{y} .

- ▶ La covarianza campionaria è

$$c = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{(n - 1)}$$

- ▶ La covarianza assume valore positivo (negativo) se, mediamente, le due variabili subiscono oscillazioni concordi (discordi).

Covarianza campionaria

Sia dato un campione bivariato (x_i, y_i) per $i = 1, 2, \dots, n$, dove le singole variabili hanno medie campionarie \bar{x} e \bar{y} .

- ▶ La covarianza campionaria è

$$c = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{(n - 1)}$$

- ▶ La covarianza assume valore positivo (negativo) se, mediamente, le due variabili subiscono oscillazioni concordi (discordi).
 - ▶ Es. quando x supera il valor medio anche y supera il valor medio (quando x supera il valor medio y non lo supera, e viceversa) La covarianza assume valore zero se le oscillazioni sono indipendenti.

Covarianza campionaria

Sia dato un campione bivariato (x_i, y_i) per $i = 1, 2, \dots, n$, dove le singole variabili hanno medie campionarie \bar{x} e \bar{y} .

- ▶ La covarianza campionaria è

$$c = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{(n - 1)}$$

- ▶ La covarianza assume valore positivo (negativo) se, mediamente, le due variabili subiscono oscillazioni concordi (discordi).
 - ▶ Es. quando x supera il valor medio anche y supera il valor medio (quando x supera il valor medio y non lo supera, e viceversa) La covarianza assume valore zero se le oscillazioni sono indipendenti.
- ▶ La sola covarianza può assumere qualsiasi valore e quindi non indica la forza della relazione lineare tra i termini.

Coefficiente di correlazione

Sia dato un campione bivariato (x_i, y_i) per $i = 1, 2, \dots, n$, dove le singole variabili hanno medie campionarie \bar{x} e \bar{y} e deviazioni standard campionarie s_x e s_y .

Coefficiente di correlazione

Sia dato un campione bivariato (x_i, y_i) per $i = 1, 2, \dots, n$, dove le singole variabili hanno medie campionarie \bar{x} e \bar{y} e deviazioni standard campionarie s_x e s_y .

Il coefficiente di correlazione campionaria è

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{(n - 1)s_x s_y} = \frac{c}{s_x s_y}$$

Coefficiente di correlazione

Sia dato un campione bivariato (x_i, y_i) per $i = 1, 2, \dots, n$, dove le singole variabili hanno medie campionarie \bar{x} e \bar{y} e deviazioni standard campionarie s_x e s_y .

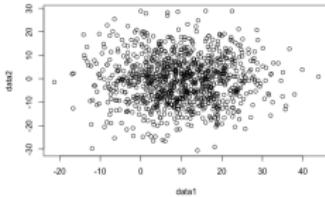
Il coefficiente di correlazione campionaria è

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{(n - 1)s_x s_y} = \frac{c}{s_x s_y}$$

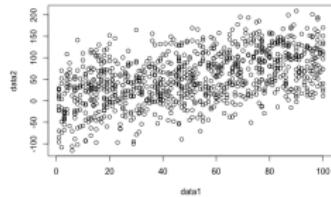
Proprietà

- ▶ $-1 \leq r \leq 1$
- ▶ quando $r > 0$ i dati sono correlati **positivamente**, quando $r < 0$ i dati sono correlati **negativamente**
- ▶ se $y_i = ax_i + b$, allora

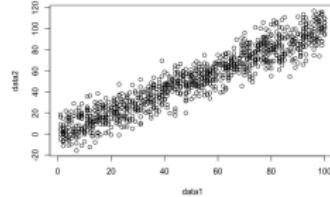
$$r = \begin{cases} 1 & \text{se } a > 0 \\ -1 & \text{se } a < 0 \end{cases}$$



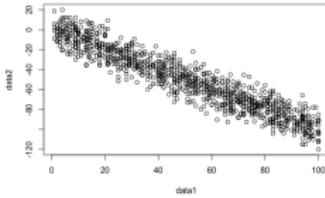
(a) $r = 0.01$



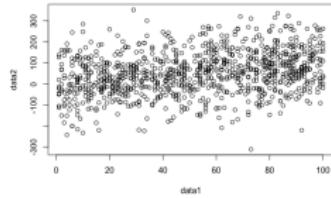
(c) $r = 0.5$



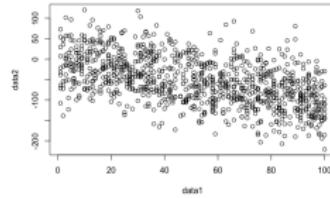
(e) $r = 0.94$



(b) $r = -0.94$



(d) $r = 0.27$



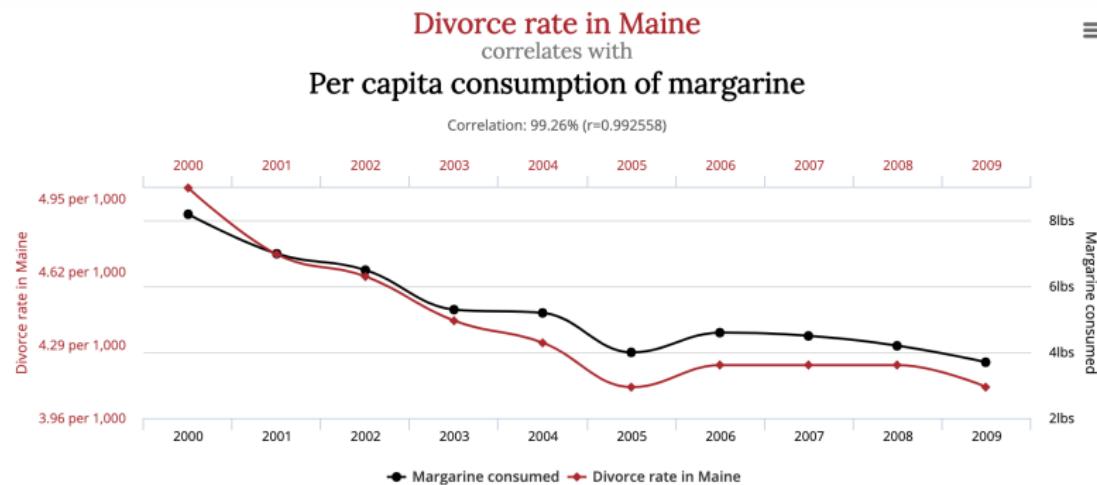
(f) $r = -0.5$

Correlation is not causation

Sebbene la correlazione indichi una associazione tra gli andamenti di due variabili, tale associazione **non implica una relazione di causa-effetto.**

Correlation is not causation

Sebbene la correlazione indichi una associazione tra gli andamenti di due variabili, tale associazione **non implica una relazione di causa-effetto.**

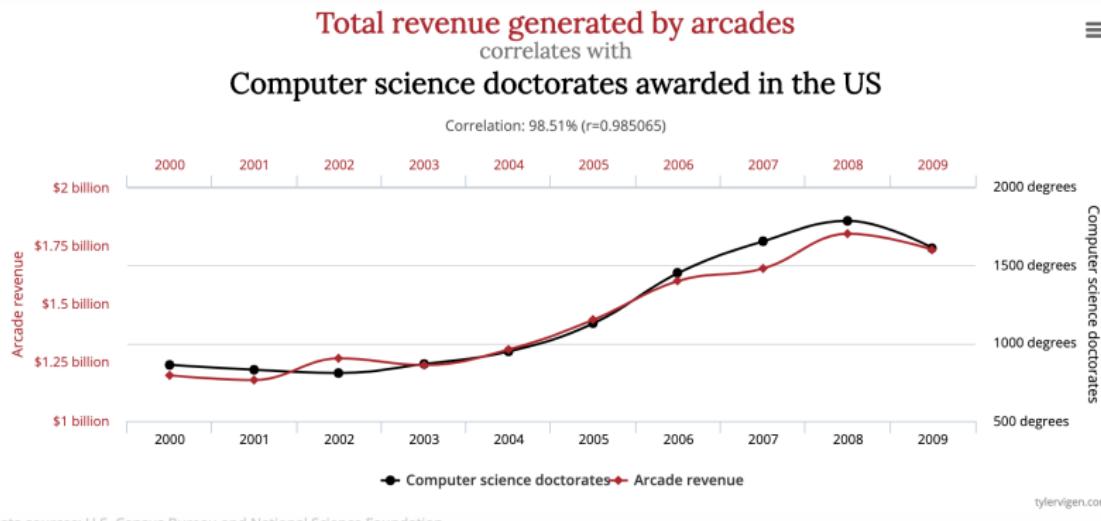


Data sources: National Vital Statistics Reports and U.S. Department of Agriculture

tylervigen.com

<http://www.tylervigen.com/spurious-correlations>

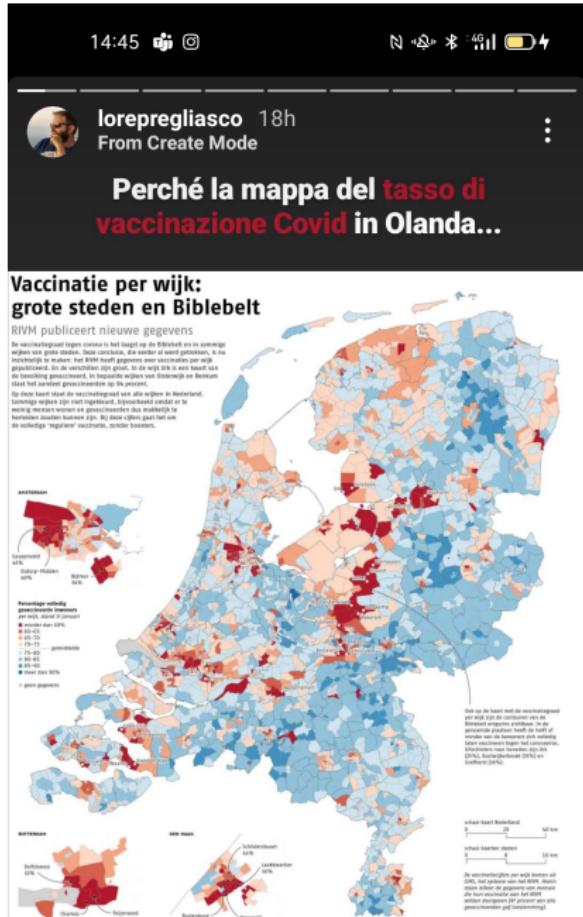
Correlation is not causation



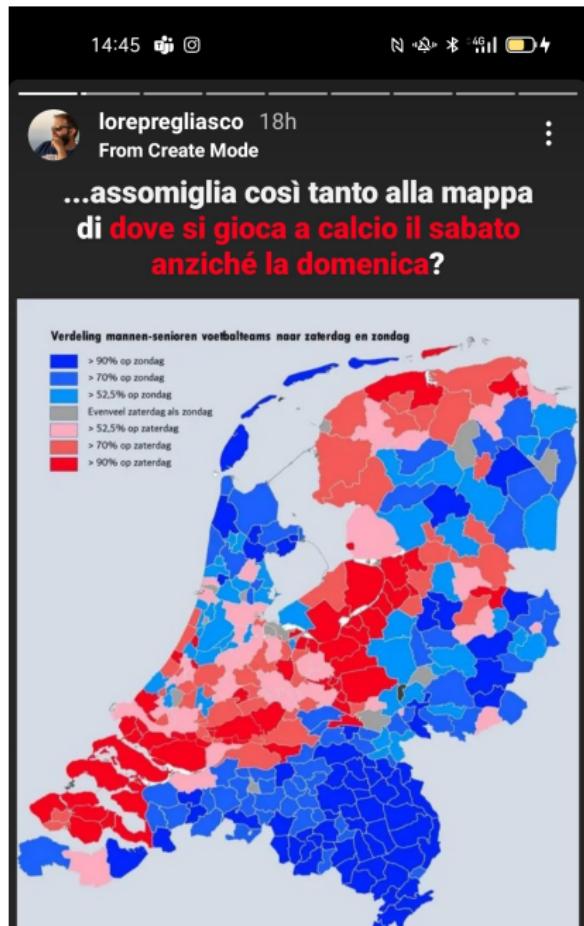
Data sources: U.S. Census Bureau and National Science Foundation

<http://www.tylervigen.com/spurious-correlations>

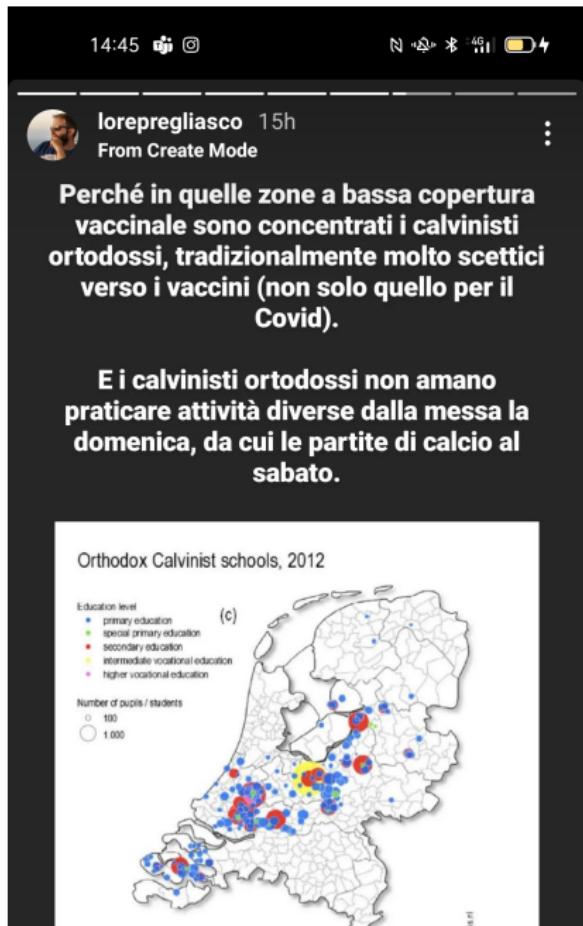
Correlation is not causation



Correlation is not causation



Correlation is not causation



Correlation is not causation

