

Principi di Inferenza

Statistica descrittiva: Indici di posizione

Laura D'Angelo
Università Milano Bicocca

Gli **indici di posizione** sono singoli valori di sintesi che descrivono “dove” si concentrano i dati.

Vedremo:

- ▶ media,
- ▶ mediana,
- ▶ quantili.

Ci occuperemo solo di indici di posizione per **variabili quantitative** (discrete o continue).

1

Media aritmetica

Consideriamo $x_1, \dots, x_n$ i dati osservati di una variabile X .

Definizione

Dati $x_1, \dots, x_n$ **quantitativi**, la **media aritmetica** è

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

Nota: definita e significativa per variabili quantitative (discrete o continue).

- **Rappresentatività:** se $x_i = a$ per tutti gli $i = 1, \dots, n$ (ovvero i dati sono costanti), allora $\bar{x} = a$.
- **Internalità:** indicando $x_{(1)} = \min\{x_1, \dots, x_n\}$, $x_{(n)} = \max\{x_1, \dots, x_n\}$,

$$x_{(1)} \leq \bar{x} \leq x_{(n)}$$

- **Formula ricorsiva:** data una nuova osservazione x_{n+1} , la media $\bar{x}_{new}$ del campione $(x_1, \dots, x_n, x_{n+1})$ si può calcolare a partire da $\bar{x}$ attraverso la media pesata:

$$\bar{x}_{new} = \frac{n\bar{x} + x_{n+1}}{n+1}$$

Proprietà

Per ogni costante $k \in \mathbb{R}$,

$$S(k) = \sum_{i=1}^n (x_i - k)^2$$

è minimizzata in $k = \bar{x}$.

Idea: abbiamo dei valori $(x_1, \dots, x_n)$ e vogliamo trovare un valore k che sia **il più vicino possibile a tutti i dati**.

- ▶ $(x_i - k)$ = scarto tra x_i e k , ovvero, quanto x_i dista da k . Lo calcoliamo per ciascun $i = 1, \dots, n$;
- ▶ $(x_i - k)^2$ = per evitare che gli scarti positivi e negativi si cancellino, li eleviamo al quadrato;
- ▶ $S(k) = \sum_{i=1}^n (x_i - k)^2$: infine, li sommiamo tutti, perchè vogliamo che k sia *complessivamente* il più vicino a *tutti i dati*;
- ▶ La media $\bar{x}$ è precisamente il valore che minimizza $S(k)$.

La media è **molto sensibile** ai valori anomali, ovvero gli outlier.

Esempio (numero di figli, 14 famiglie):

- ▶ Dati: $(0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 2, 3) \Rightarrow \bar{X} \approx 0.5$
- ▶ Cambiando un solo valore 3 in 10:
Dati: $(0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 2, 10) \Rightarrow \bar{X} \approx 1.0$ (**raddoppia**)

Consideriamo $x_1, \dots, x_n$ i dati osservati di una variabile X .

Consideriamo ora la **trasformazione lineare** dei dati:

$$y_i = a + b x_i$$

con $a, b \in \mathbb{R}$.

Proprietà

La media $\bar{y}$ dei dati trasformati può essere ottenuta a partire dalla media $\bar{x}$ calcolata sui dati originali come

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

Per trasformazioni *non lineari* $y_i = f(x_i)$, in generale $\overline{f(x)} \neq f(\bar{x})$ (la media dei valori trasformati è diversa dalla trasformazione della media).

2

Mediana

Consideriamo $x_1, \dots, x_n$ i dati osservati di una variabile X .

Inoltre, indichiamo con $x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$ i dati ordinati, così che $x_{(i)}$ indica l' i -esimo valore osservato in ordine (crescente) di grandezza.

Idea: la **mediana** Me è un valore tale che circa il 50% delle osservazioni è $\leq Me$ e circa il 50% è $\geq Me$.

Definizione

Consideriamo i dati ordinati: $x_{(1)} \leq \dots \leq x_{(n)}$:

$$Me = \begin{cases} x_{(\frac{n+1}{2})} & \text{se } n \text{ è dispari} \\ \frac{1}{2} \left(x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)} \right) & \text{se } n \text{ è pari.} \end{cases}$$

Esempio 1 (n dispari).

- ▶ Dati: 1, 7, 2, 9, 4 $\Rightarrow n = 5$
- ▶ Dati ordinati: 1, 2, 4, 7, 9
- ▶ La mediana è l'elemento in posizione $\frac{n+1}{2} = 3$ (l'elemento centrale)
 $\Rightarrow Me = x_{(3)} = 4.$

Esempio 2 (n pari).

- ▶ Dati: 1, 2, 1, 5 $\Rightarrow n = 4$
- ▶ Dati ordinati: 1, 1, 2, 5
- ▶ qualunque $Me \in (1, 2)$ divide al 50/50; convenzionalmente spesso si usa la media tra i due valori $x_{(\frac{n}{2})}$ e $x_{(\frac{n}{2}+1)}$, ovvero in questo caso $x_{(2)} = 1$ e $x_{(3)} = 2$
 $\Rightarrow Me = \frac{x_{(2)} + x_{(3)}}{2} = \frac{1+2}{2} = 1.5.$

- ▶ **Rappresentatività:** se $x_i = c$ per tutti gli $i = 1, \dots, n$, allora $Me = c$.
- ▶ **Internalità:** $x_{(1)} \leq Me \leq x_{(n)}$.
- ▶ **Minimizzazione scarti assoluti:** $Me \in \arg \min_k \sum_{i=1}^n |x_i - k|$.
 - ▶ Se n è dispari, Me è *unico*.
 - ▶ Se n è pari, *ogni* $k \in [x_{(n/2)}, x_{(n/2+1)}]$ minimizza la somma.
- ▶ **Robustezza ai valori anomali:** la mediana è robusta agli outlier: varia poco se cambiano pochi valori estremi.

Esempio (numero di figli, 14 famiglie): confronto con la media

- ▶ Dati: $(0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 2, 3) \Rightarrow \bar{X} \approx 0.5, Me_X = 0$
- ▶ Cambiando un solo valore 3 in 10:
Dati: $(0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 2, 10) \Rightarrow \bar{X} \approx 1.0, Me_X = 0$
la mediana resta 0 anche cambiando un valore da 3 a 10

Consideriamo $x_1, \dots, x_n$ i dati osservati di una variabile X , e chiamiamo Me_X la mediana. Consideriamo la trasformazione dei dati $y_i = f(x_i)$. Siamo interessati alla mediana Me_Y calcolata sui dati trasformati $y_1, \dots, y_n$.

- ▶ Se f è **monotona crescente**, allora $Me_Y = f(Me_X)$.
- ▶ Se f è **monotona decrescente**, allora $Me_Y = f(Me_X)$ (l'ordine si inverte ma la posizione mediana corrisponde).
- ▶ Se f **non è monotona**, in generale la relazione non vale.

Esempi:

- ▶ monotona: $f(x) = a + bx$ monotona se $b \neq 0$;
 $f(x) = \log x$ per $x \in (0, \infty)$, monotona crescente;
- ▶ non monotona: $f(x) = x^2$ non monotona su $\mathbb{R}$.

La mediana ideale soddisfa $F(Me) = 0.5$, dove F è la funzione di ripartizione.

Tuttavia, la funzione di ripartizione empirica $\hat{F}$, che calcoliamo sui dati, è una **funzione a salti**, quindi:

- ▶ **può non assumere mai il valore 0.5**: se $\hat{F}$ “attraversa 0.5” con un salto, la mediana è il valore Me in corrispondenza del salto.
- ▶ se $\hat{F}$ è **uguale a 0.5** in un intervallo $[a, b]$, qualunque $Me \in [a, b]$ è una mediana.

Esempio: morti da calcio di cavallo

Dati (Prussia, 1875–1894):

1, 1, 2, 1, 1, 3, 0, 4, 0, 1, 0, 3, 2, 1, 0, 2, 1, 1, 0, 0

Dati ordinati:

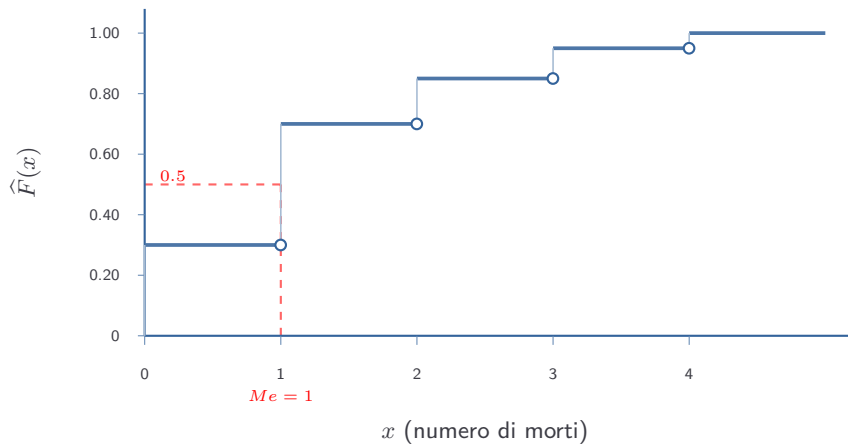
0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 1, 2, 2, 2, 3, 3, 4

$$n = 20 \quad \Rightarrow \quad Me = \frac{x_{(10)} + x_{(11)}}{2} = 1, \quad \text{ma } \hat{F}(1) = 0.7$$

Morti	Frequenza assoluta	Frequenza relativa	Frequenza relativa cumulata
0	6	0.30	0.30
1	8	0.40	0.70
2	3	0.15	0.85
3	2	0.10	0.95
4	1	0.05	1

Esempio: morti da calcio di cavallo

Esempio sui dati Prussia (frequenze relative cumulate)



$$\hat{F}(x) = \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\{x_i \leq x\}} .$$

Esempio: chiamate perse per turno

Dati:

0, 1, 3, 0, 4, 1, 3, 0, 1, 3, 0, 3, 4, 1, 0, 6, 3, 1, 3, 4

Dati ordinati:

0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 3, 3, 3, 3, 3, 3, 4, 4, 4, 6

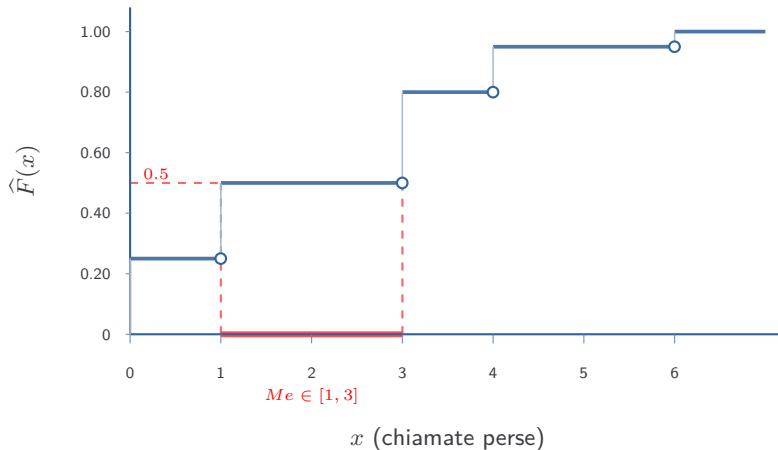
$$n = 20 \implies Me = \frac{x_{(10)} + x_{(11)}}{2} = \frac{1 + 3}{2} = 2, \quad \text{ma } \hat{F}(x) = 0.5 \quad \forall x \in [1, 3)$$

Chiamate perse	Frequenza assoluta	Frequenza relativa	Frequenza relativa cumulata
0	5	0.25	0.25
1	5	0.25	0.50
3	6	0.30	0.80
4	3	0.15	0.95
6	1	0.05	1

Il valore 2 non compare mai nei dati: tra $x = 1$ e $x = 3$ la funzione di ripartizione empirica resta costante e pari a 0.5.

Esempio: chiamate perse per turno

Esempio sui dati "chiamate perse" (frequenze relative cumulate)



$\hat{F}(x) = \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\{x_i \leq x\}}$. Qui $\hat{F}(x) = 0.5$ per ogni $x \in [1, 3]$: qualunque valore $Me \in [1, 3]$ soddisfa la definizione di mediana.

- ▶ Calcolare Me per un dataset pari e individuare l'intervallo delle mediane equivalenti.
- ▶ Verificare $\arg \min_a \sum_i |X_i - a|$ per un piccolo insieme di dati.
- ▶ Studiare l'effetto di una trasformazione monotona/non monotona sulla mediana.

3

Quantili

e rappresentazione grafica dei dati tramite boxplot

Idea: i quantili sono una **generalizzazione della mediana**. Abbiamo visto che la mediana è un valore Me tale che circa il 50% delle osservazioni è $\leq Me$ e circa il 50% è $\geq Me$.

I quantili fanno lo stesso ma per **percentuali anche diverse dal 50%**.

Definizione

Il **quantile di livello p** , che indichiamo con q_p , dove $p \in (0, 1)$, è il più piccolo valore per cui il $(p \times 100)\%$ delle osservazioni ha modalità uguale o minore di q_p .

- ▶ I quantili con $p = \{0.25, 0.50, 0.75\}$ sono chiamati **quartili**: $(q_{0.25}, q_{0.5}, q_{0.75})$
- ▶ I quantili con $p = \{0.1, 0.2, 0.3, \dots, 0.9\}$ sono chiamati **decili**: $(q_{0.1}, q_{0.2}, q_{0.3}, \dots, q_{0.9})$
- ▶ I quantili con $p = \{0.01, 0.02, \dots, 0.98, 0.99\}$ sono chiamati **percentili**: $(q_{0.01}, q_{0.02}, \dots, q_{0.99},)$

Calcolo dei quantili

Per calcolare i quantili, abbiamo bisogno di introdurre la **funzione “ceiling”** (parte intera superiore).

Definizione: Per ogni numero reale x , $\lceil x \rceil$ è il più piccolo intero non minore di x .

Esempi: $\lceil 3.2 \rceil = 4$, $\lceil 5.0 \rceil = 5$, $\lceil 18.1 \rceil = 19$.

Consideriamo $x_1, \dots, x_n$ i dati osservati di una variabile X . Indichiamo con $x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$ i dati ordinati in ordine crescente (così che $x_{(i)}$ indica l' i -esimo valore ordinato).

Definizione

Per calcolare il quantile q_p di livello p , con $p \in (0, 1)$, poniamo $k = n \times p$

$$q_p = \begin{cases} x_{(\lceil k \rceil)} & \text{se } k \text{ non è un numero intero} \\ \frac{x_{(k)} + x_{(k+1)}}{2} & \text{se } k \text{ è un numero intero.} \end{cases}$$

Note:

- ▶ nel primo caso, stiamo prendendo il valore osservato in posizione k (arrotondata per eccesso); nel secondo, prendiamo la media dei due valori osservati intorno alla posizione k .
- ▶ per $p = 0.5$ questa regola coincide con la formula della mediana vista in precedenza.

Dati ordinati ($n = 20$):

2, 3, 3, 4, 5, 5, 6, 7, 7, 8, 8, 9, 9, 10, 11, 12, 13, 14, 15

► **Quantile di livello $p = 0.30$:**

Calcoliamo $k = n \times p = 20 \times 0.30 = 6$ ($\implies$ intero)

Il quantile è dato dalla media dei valori in posizione 6 e 7:

$$q_{0.30} = \frac{x_{(6)} + x_{(7)}}{2} = \frac{5 + 6}{2} = 5.5$$

► **Quantile di livello $p = 0.32$:**

Calcoliamo $k = n \times p = 20 \times 0.32 = 6.4$ ($\implies$ non intero)

Applichiamo la funzione ceiling: $\lceil k \rceil = \lceil 6.4 \rceil = 7$,

Il quantile è dato semplicemente dal valore in posizione 7:

$$q_{0.32} = x_{(\lceil 6.4 \rceil)} = x_{(7)} = 6$$

Esercizio: calcolo dei quartili

Dati ordinati ($n = 10$):

50, 60, 65, 70, 75, 80, 90, 190, 200, 250

Vogliamo calcolare i quartili, ovvero, $(q_{0.25}, q_{0.5}, q_{0.75})$

- **Primo quartile:** $p = 0.25$

Calcoliamo $k = n \times p = 10 \times 0.25 = 2.5$ ($\implies$ non intero)

Applichiamo la funzione ceiling: $\lceil k \rceil = \lceil 2.5 \rceil = 3$,

Il primo quartile è $q_{0.25} = x_{(\lceil 2.5 \rceil)} = x_{(3)} = 65$

- **Secondo quartile:** $p = 0.5$ (= mediana)

Calcoliamo $k = n \times p = 10 \times 0.5 = 5$ ($\implies$ intero)

Il secondo quartile (mediana) è la media tra $x_{(5)}$ e $x_{(6)}$: $q_{0.5} = Me = \frac{x_{(5)} + x_{(6)}}{2} = \frac{75 + 80}{2} = 77.5$

- **Terzo quartile:** $p = 0.75$

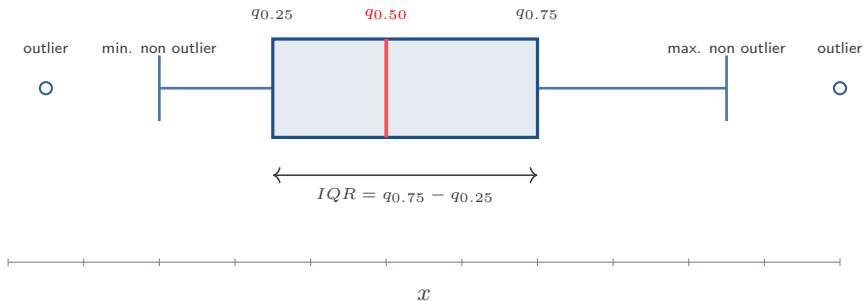
Calcoliamo $k = n \times p = 10 \times 0.75 = 7.5$ ($\implies$ non intero)

Applichiamo la funzione ceiling: $\lceil k \rceil = \lceil 7.5 \rceil = 8$,

Il terzo quartile è $q_{0.75} = x_{(\lceil 7.5 \rceil)} = x_{(8)} = 190$

Il boxplot

Il **boxplot** (o diagramma a scatola con baffi) è una rappresentazione grafica sintetica della distribuzione dei dati, basata sui quantili.



I baffi si estendono fino al valore osservato più estremo entro $1.5 \cdot IQR$ dai quartili; oltre tale soglia i punti sono considerati outlier.

1) Si disegna la **scatola**:

- ▶ il **primo quartile** $q_{0.25}$ e il **terzo quartile** $q_{0.75}$ delimitano la **scatola** (box);
- ▶ la **mediana** $q_{0.50}$, rappresentata da una linea all'interno della scatola;

2) Si calcolano:

- il **range interquartile** $IQR = q_{0.75} - q_{0.25}$, che misura la variabilità del 50% centrale dei dati;
- le **barriere**, che delimitano la regione dei valori “non anomali”:

$$[q_{0.25} - 1.5 \cdot IQR, q_{0.75} + 1.5 \cdot IQR]$$

3) Si disegnano i **baffi** e, eventualmente, gli **outlier**:

- ▶ i **baffi** si estendono non fino alle barriere, ma fino al **valore osservato più estremo che rientra ancora entro le barriere**:

$$\text{baffo inferiore} = \min\{x_i : x_i \geq q_{0.25} - 1.5 \cdot IQR\}$$

$$\text{baffo superiore} = \max\{x_i : x_i \leq q_{0.75} + 1.5 \cdot IQR\}$$

- ▶ eventuali **valori anomali** (outlier): i dati osservati che cadono **oltre le barriere**, rappresentati come punti isolati.