

Principi di Inferenza

Statistica descrittiva: Distribuzioni di frequenza

Laura D'Angelo
Università Milano Bicocca

Una volta osservate le quantità di interesse, vengono organizzate in tabelle opportune, a seconda della loro natura.

- Consideriamo $x_1, \dots, x_n$ i dati osservati di una variabile X .
- n : numero di unità statistiche.
- x_i : valore (o modalità) assunto dalla variabile X per l'unità i -esima.

Distribuzione di unità: annotiamo direttamente i dati osservati.

Osservazione	Valore	Riportiamo l'indice dell'osservazione corrispondente (colonna di sinistra) e il valore (colonna di destra).
1	x_1	Non sintetizziamo in alcun modo quanto osservato. Esempio: per ogni persona in aula, misuriamo la sua altezza. Il valore generico rappresenta l'altezza della i -esima persona.
2	x_2	
3	x_3	
$\vdots$	$\vdots$	
$\vdots$	$\vdots$	
n	x_n	

1

Distribuzioni di frequenza

Variabili qualitative o quantitative discrete

Distribuzione di frequenze: quando vogliamo riportare un **valore osservato** ed **il numero di volte** che osserviamo tale valore (frequenza).

- ▶ Consideriamo variabili **qualitative** o **quantitative discrete**.
- ▶ Indichiamo con $c_1, \dots, c_k$ le modalità possibili della variabile.

Definizione

La **frequenza assoluta** n_j della modalità c_j è il **numero** di unità statistiche per cui $x_i = c_j$.

In formule:

$$n_j = \sum_{i=1}^n \mathbb{1}_{\{x_i = c_j\}}, \quad \text{per } j = 1, \dots, k$$

$$\text{dove } \mathbb{1}_{\{A\}} = \begin{cases} 1 & \text{se } A \text{ è vera} \\ 0 & \text{altrimenti} \end{cases}$$

Proprietà:

- ▶ $n_j \in \mathbb{N}_0$ (numeri interi non negativi)
- ▶ $\sum_{j=1}^k n_j = n$ (numerosità campionaria)

Definizione

La **frequenza relativa** f_j della modalità c_j è la **proporzione** di unità statistiche per cui $x_i = c_j$.

In formule:

$$f_j = \frac{n_j}{n}, \quad \text{per } j = 1, \dots, k$$

Proprietà:

- ▶ Le frequenze relative sono numeri reali in $[0, 1]$.
- ▶ $\sum_{j=1}^k f_j = 1$

Definizione

La **frequenza percentuale** è $p_j = f_j \times 100$

- ▶ $\sum_{j=1}^k p_j = 100\%$

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

Obiettivo

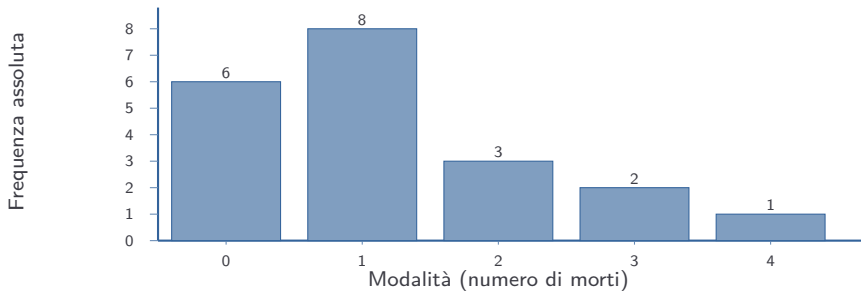
Calcolare le frequenze assolute, relative e percentuali.

Morti	Frequenza assoluta	Frequenza relativa	Percentuale
0	6	0.30	30%
1	8	0.40	40%
2	3	0.15	15%
3	2	0.10	10%
4	1	0.05	5%
Totale	20	1.00	100%

$n = 20$ anni considerati $\Rightarrow$ 20 unità statistiche.

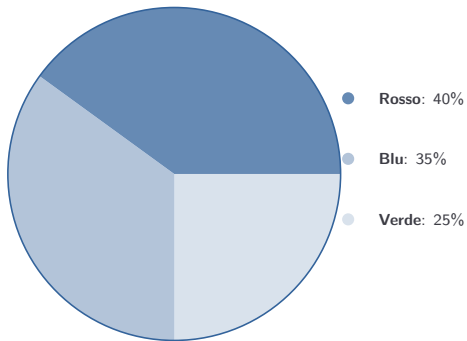
- ▶ **Grafico a barre** — variabili qualitative o quantitative discrete.
 - ▶ Altezza della barra = frequenza assoluta o relativa.
 - ▶ Le modalità vanno sull'asse x , ordinate se necessario.
- ▶ **Grafico a torta** — variabili qualitative nominali (poche modalità).
 - ▶ Ogni fetta ha area proporzionale alla frequenza della modalità.
 - ▶ Poco leggibile se le modalità sono molte.

Esempio: morti da calcio di cavallo (Prussia, 1875–1894) — frequenze assolute



Nota: per frequenze relative si possono scalare le altezze dividendo per n o cambiare l'etichetta dell'asse y .

Esempio qualitativo nominale (pochi livelli): *Colore preferito*



- ✓ Ideale per **qualitative nominali** con poche modalità.
- ✗ Difficile confrontare molte fette o fette simili.
- Alternative consigliate: **barre** ordinate per frequenza.

2

Distribuzioni di frequenza

Variabili quantitative continue

Distribuzioni di frequenza:

Variabili quantitative continue

Idea

Per le variabili quantitative continue, le modalità vengono raggruppate in **intervalli di classe**.

Distribuzione di frequenze in classi (continuo): quando vogliamo riportare un **intervallo di valori osservati** ed il **numero di volte che osserviamo un valore all'interno** dell'intervallo.

- Si definiscono intervalli disgiunti (ovvero le classi)

$$C_1 = [\ell_1, u_1), C_2 = [\ell_2, u_2), \dots, C_k = [\ell_k, u_k)$$

In genere, gli intervalli sono contigui: $u_j = \ell_{j+1}$, ovvero l'estremo superiore di un intervallo coincide con l'estremo inferiore del successivo.

Esempio: Tempo per completare una maratona $C_1 = [1h, 2h), C_2 = [2h, 3h), \dots, C_7 = [7h, 8h)$

Per ogni classe, le frequenze si calcolano in modo analogo al caso discreto.

La **frequenza assoluta** n_j della classe C_j è **il numero** di unità statistiche per cui $x_i \in C_j$:

$$n_j = \sum_{i=1}^n \mathbb{1}_{\{x_i \in C_j\}}, \quad \text{per } j = 1, \dots, k.$$

La **frequenza relativa** f_j della modalità c_j è **la proporzione** di unità statistiche per cui $x_i \in C_j$:

$$f_j = \frac{n_j}{n}, \quad \text{per } j = 1, \dots, k.$$

La **frequenza percentuale** è $p_j = f_j \times 100$.

Valgono le stesse proprietà viste per variabili qualitative o quantitative discrete.

Perché suddividere in classi le variabili continue?

- ▶ Permette una rappresentazione **più sintetica** dei dati e di calcolarne le distribuzioni di frequenza (che per le variabili quantitative continue non possono essere calcolate sui dati originali perchè—atmeno in teoria—tutti diversi uno dall'altro).
- ▶ La raccolta può essere più **facile o economica** (non serve misurare con precisione ogni valore).
- ▶ Favorisce la **privacy**: conosciamo solo l'intervallo di appartenenza del valore, non il valore preciso.

La scelta degli intervalli è cruciale:

- ▶ troppi → dati frammentati;
- ▶ pochi → perdita di informazione.

Come scegliere il **numero** di classi

- ▶ Indichiamo con k **il numero di classi**.
- ▶ Se k è troppo basso → **perdita di informazione**.
- ▶ Se k è troppo alto → rappresentazione non sintetica, costi elevati, scarsa privacy.

In generale

Più grande è n (il numero di dati), più alto può essere k .

In pratica: la scelta dipende dal dataset e dagli obiettivi dell'analisi.

Come scegliere l'**ampiezza** delle classi

- ▶ Le classi devono coprire tutto l'intervallo $[x_{\min}, x_{\max}]$.
- ▶ Devono essere **disgiunte**: ogni valore appartiene a una sola classe.
- ▶ In genere si preferiscono **classi di uguale ampiezza**.

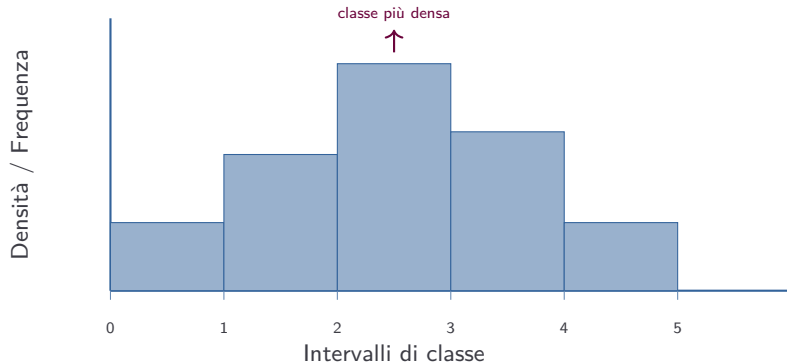
Quando usare ampiezze diverse

- ▶ Se i dati sono fortemente concentrati in una zona e rari altrove.
- ▶ Esempio: reddito annuale (in migliaia di €) di 500 individui:
 - ▶ 80% tra 10 e 50 → classi strette
 - ▶ 1% tra 100 e 300 → classi più ampie

In questi casi si parla di **classi di ampiezza disuguale**.

Rappresentazione: istogramma

Per rappresentare una distribuzione di frequenza di una variabile quantitativa continua (in classi) usiamo gli **istogrammi**.



Distribuzione di una variabile quantitativa continua (classi di uguale ampiezza)

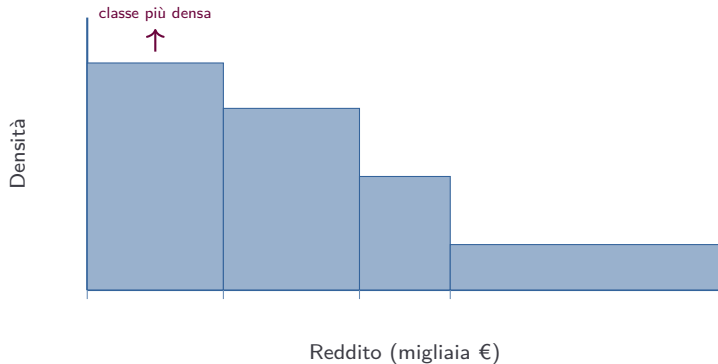
- ▶ Definiamo le classi disgiunte $C_1, C_2, \dots, C_k$, con $C_j = [\ell_j, u_j)$;
- ▶ Per ogni classe si calcolano le frequenze assolute n_j o relative f_j , $j = 1, \dots, k$.
- ▶ Per ogni classe C_j disegniamo un rettangolo:
 - La base del rettangolo corrispondente alla classe C_j , ovvero a ciascun intervallo $[\ell_j, u_j)$;
 - L'**area** A_j del rettangolo è proporzionale alla **frequenza assoluta/relativa** della classe;
 - L'**altezza** h_j del rettangolo rappresenta la **densità della classe**.

Formula

$$A_j = (\text{base})_j \times (\text{altezza})_j = (u_j - \ell_j) \times h_j$$
$$\Rightarrow h_j = \frac{A_j}{(u_j - \ell_j)} = \frac{f_j}{(u_j - \ell_j)}$$

Istogramma con classi di ampiezza diversa

Esempio: reddito annuale (in migliaia di €)



Le classi di ampiezza diversa si rappresentano mantenendo proporzionale l'**area**, non l'altezza.

Obiettivo

Rappresentare la distribuzione di una variabile **quantitativa continua** mediante rettangoli la cui area è proporzionale alla frequenza (assoluta o relativa).

Scelta del numero di classi

- ▶ Troppo poche → perdita di informazione.
- ▶ Troppe → grafico frammentato e poco leggibile.

Scelta dell'ampiezza

- ▶ In genere classi **di uguale ampiezza**.
- ▶ Ampiezze diverse se i dati sono molto concentrati in un intervallo.

Importante

- ▶ L'**area** del rettangolo rappresenta la frequenza della classe.
- ▶ Le **altezze** riflettono le *densità* delle classi.
- ▶ Con ampiezze uguali $\Rightarrow$ altezza $\propto$ frequenza.

3

Frequenze cumulate e la funzione di ripartizione empirica

Solo per variabili con modalità ordinate (ovvero, qualitative ordinali o quantitative).

- ▶ Indichiamo con $c_1, c_2, \dots, c_k$ i valori distinti osservati (o le modalità).
- ▶ n_j è la frequenza assoluta della modalità c_j , mentre f_j è la frequenza relativa, $j = 1 \dots, k$.

Definizione

- ▶ **Frequenza assoluta cumulata:** $N_j = \sum_{h=1}^j n_h$, per $j = 1, \dots, k$
- ▶ **Frequenza relativa cumulata:** $F_j = \sum_{h=1}^j f_h$, per $j = 1, \dots, k$

Proprietà:

- ▶ $N_k = n$, $F_k = 1$
- ▶ N_j indica quante unità statistiche hanno una modalità $\leq c_j$

Definizione

Data una variabile X e i dati $x_1, \dots, x_n$, la **funzione di ripartizione empirica** è:

$$\widehat{F}(x) = \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\{X_i \leq x\}}$$

- ▶ $\widehat{F}(x)$ rappresenta la proporzione di unità statistiche con $x_i \leq x$.
- ▶ È una funzione non decrescente, con:

$$\widehat{F}(\min x_i) = 0, \quad \widehat{F}(\max x_i) = 1$$

Esempio. Variabile: età $\Rightarrow \widehat{F}(30) =$ proporzione di individui nel campione con età ≤ 30 .

Dati (Prussia, 1875–1894):

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

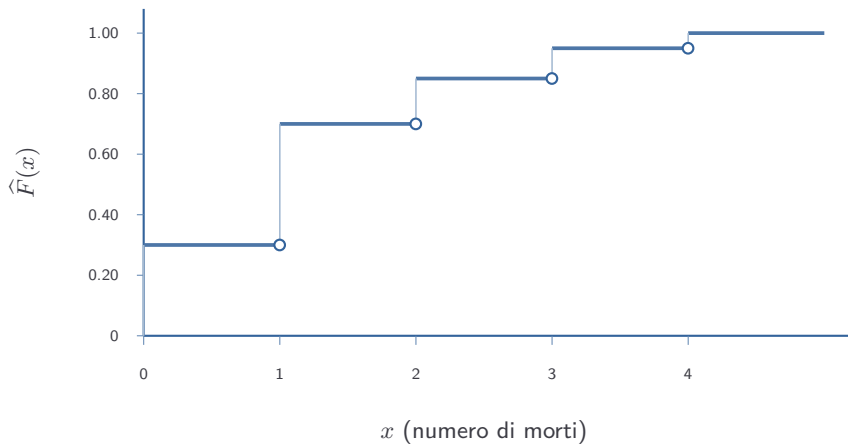
Dati ordinati:

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

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

Funzione di ripartizione empirica

Esempio sui dati Prussia (frequenze relative cumulate)



$$\hat{F}(x) = \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\{x_i \leq x\}} \text{ è non decrescente, con } \hat{F}(\max X) = 1.$$