STATISTICA 1 - Posizione

Riccardo Corradin, Andrea Gilardi

Introduzione

Una volta che abbiamo raccolto delle **osservazioni**, che sia l'intera popolazione o un singolo campione, ed abbiamo prodotto una **distribuzione di frequenze**, siamo interessati a costruire degli **indici sintetici** di tale distribuzione.

Le cose principali a cui siamo interessati sono:

- dove è **centrata** la distribuzione nel suo supporto;
- quanto è **dispersa** la distribuzione intorno al suo centro.

Chiaramente, entrambe le precedenti sono fortemente dipendenti dal tipo di dato osservato, se di natura **qualitativa** o **quantitativa**, discreto o continuo.

Nelle prossime slides vedremo come costruire opportuni **indici di posizione**, come utilizzarli e come interpretarli.

Valori medi

Introduzione

Media: sintesi **puntuale** dei dati relativa ad un determinato carattere, con lo scopo di rappresentare ed evidenziare una determinata caratteristica del carattere stesso.

→ Per **puntuale**, intendiamo che individua una singola modalità del carattere o un singolo valore numerico, che utilizziamo come rappresentativo di tutte le modalità presenti nella distribuzione di frequenze del carattere.

Spesso con **media** viene intesa la media aritmetica, indice che vedremo nelle slide successive. Qui, intendiamo una famiglia ampia di indici di sintesi per distribuzioni di frequenze.

Il Pollo di Trilussa[1]

Sai ched'è la statistica? È na' cosa
che serve pe fà un conto in generale
de la gente che nasce, che sta male,
che more, che va in carcere e che spósa.

Ma pè me la statistica curiosa
è dove c'entra la percentuale,
pè via che, lì, la media è sempre eguale
puro co' la persona bisognosa.

Me spiego: 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.

[1] La Statistica, Carlo Alberto Salustri, conosciuto con lo pseudonimo di Trilussa

Moda

Moda

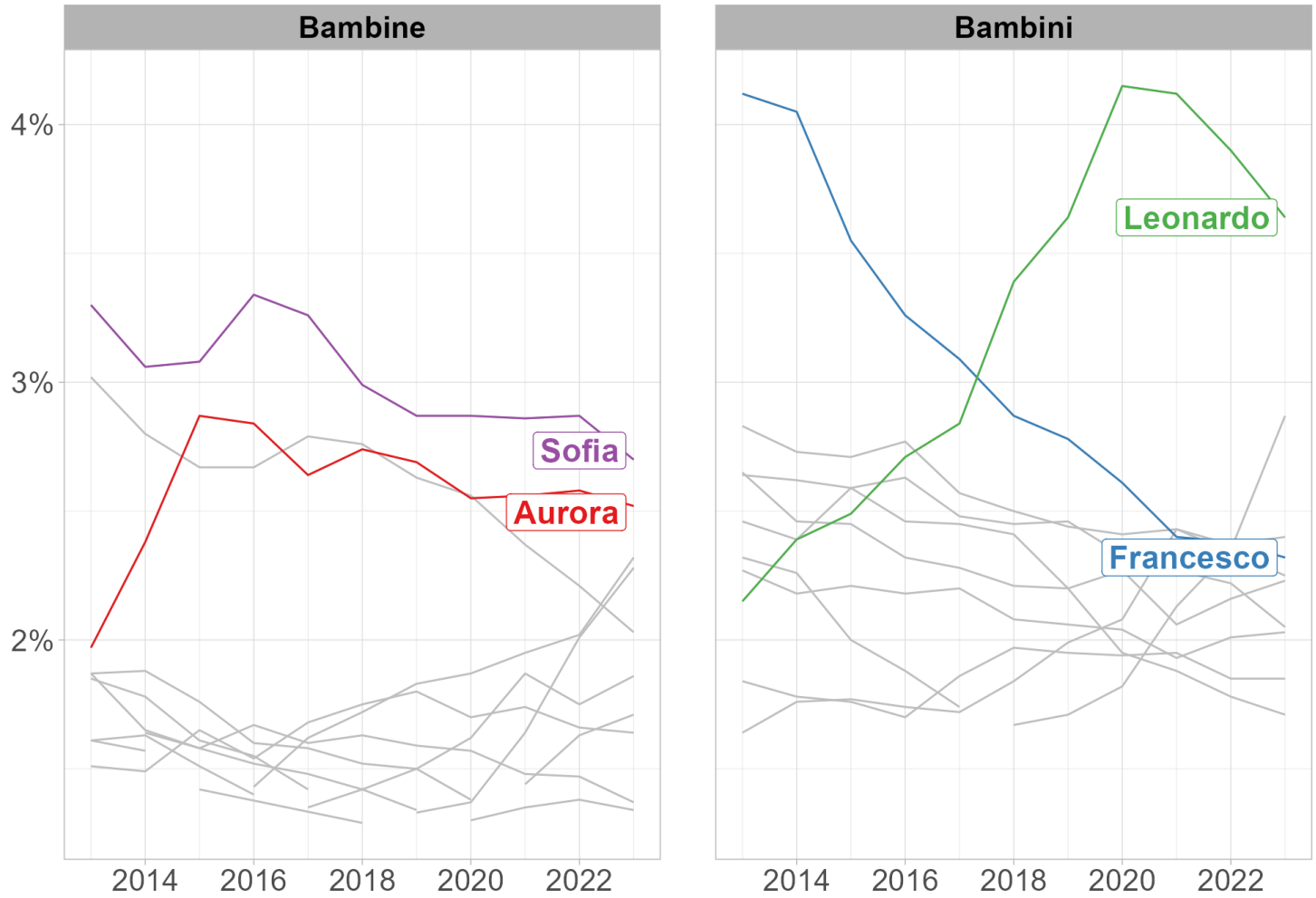
Moda: rappresenta la modalità del carattere che si osserva più frequentemente all'interno della popolazione o del campione considerato.

- All'interno di un campione o dell'intera popolazione, a seconda di cosa stiamo analizzando, selezioniamo la modalità con il maggior addensamento di frequenze, la modalità più comune.

- La moda viene considerata **rappresentativa**, ossia che fornisce una buona sintesi dell'informazione presente nel campione o nella popolazione, quando la frequenza relativa associata alla modalità modale è **maggiore di** 0.5.

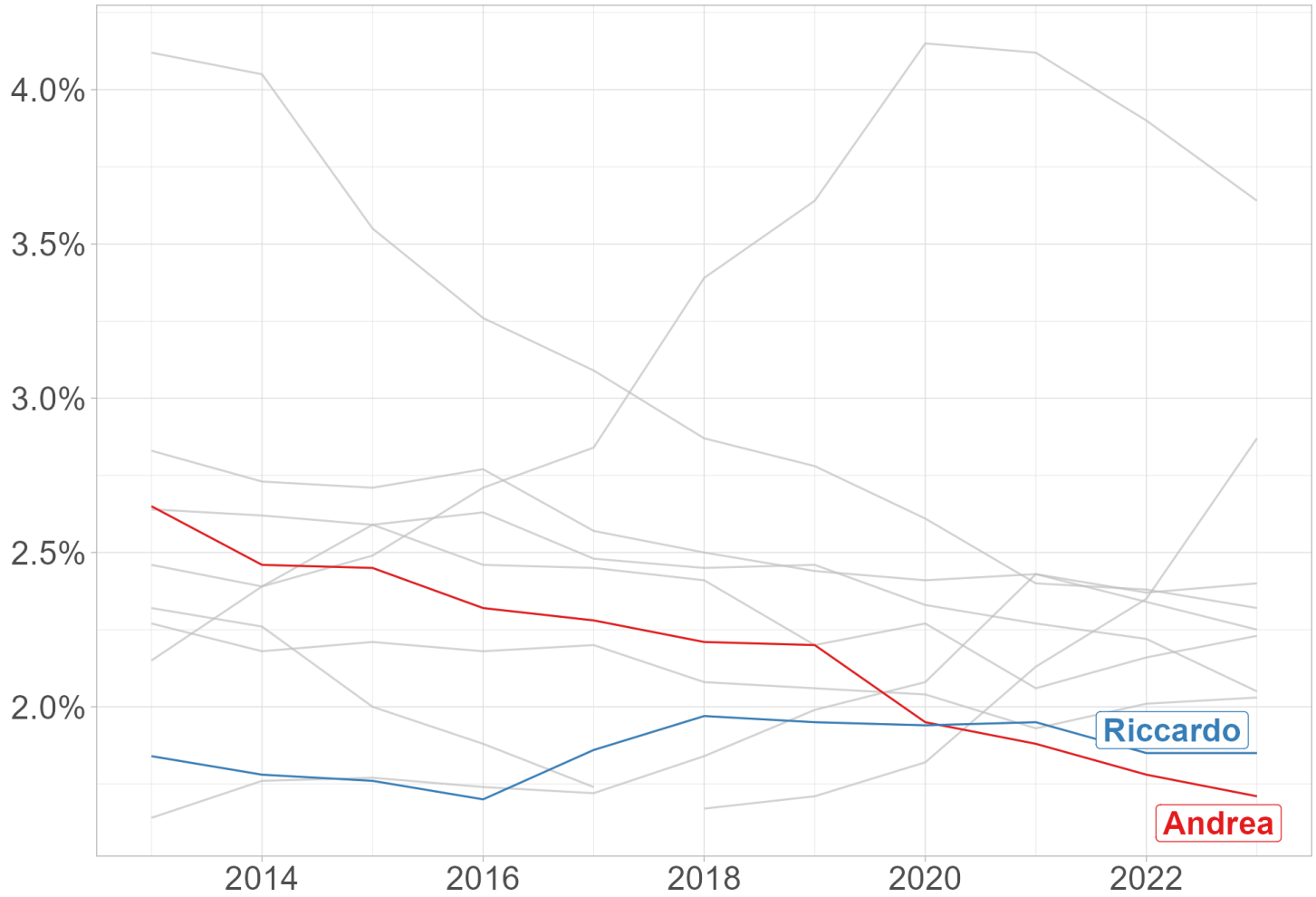
- Quando abbiamo una distribuzione di frequenze con **modalità in classi**, la classe modale è data dalla classe in cui osservo la **frequenza specifica più elevata**.
  $\rightarrow$ Rappresenta la classe in cui verosimilmente ricade la moda.

I 10 nomi più diffusi fra i neonati in Italia (2013-2023)[2]



[2]Fonte: `https://www.istat.it/dati/calcolatori/contanomi/`

E noi?



Moda - Esercizio[3]

Example

Durante l'anno solare 2022, nell'Università Bicocca si sono laureati $N = 7301$ studenti. La seguente riporta le frequenze relative suddivisa per Area Disciplinare:

Area Disciplinare	Frequenza	Frequenza relativa
Scienze economiche e statistiche		0.24436
Scienze giuridiche		0.06684
Scienze mediche		0.07396
Scienze psicologiche		0.12861
Scienze della formazione		0.15464
Scienze MMFFNN		0.25544
Scienze sociologiche		0.07615

- Completare la tabella aggiungendo le frequenze assolute.
- Determinare la classe modale per Area Disciplinare e produrre una opportuna rappresentazione grafica.

[3] Fonte: `https://trasparenza.unimib.it/contenuto602_dati-sugli-studenti_746.html`

Moda - Esercizio

Soluzioni: Per definizione di **Frequenze relative** si ha

Area Disciplinare	Frequenza	Frequenza relativa
Scienze econ. e stat.	$7301 \cdot 0.24436 = 1784$	0.24436
Scienze giuridiche	$7301 \cdot 0.06684 = 488$	0.06684
Scienze mediche	$7301 \cdot 0.07396 = 540$	0.07396
Scienze psicologiche	$7301 \cdot 0.12861 = 939$	0.12861
Scienze della formazione	$7301 \cdot 0.15464 = 1129$	0.15464
Scienze MMFFNN	$7301 \cdot 0.25544 = 1865$	0.25544
Scienze sociologiche	$7301 \cdot = 0.07615 = 556$	0.07615

La classe modale è *Scienze MMFFNN*.

Moda - Distribuzioni numeriche in classi

- Nel caso di distribuzioni di frequenze con **dati numerici raggruppati in classi**, la determinazione della moda richiede alcune considerazioni extra.
- Le frequenze assolute **non sono immediatamente confrontabili se le classi non hanno la stessa ampiezza**.
- E' necessario utilizzare le **frequenze specifiche**. Vediamo un esempio.

Example

Una azienda vuole studiare la distribuzione del tempo di permanenza (in minuti) degli utenti su una pagina di un loro prodotto. Commissiona tale studio ad una agenzia di marketing che le restituisce la seguente tabella:

Intervallo di tempo	$0 \dashv 1$	$1 \dashv 3$	$3 \dashv 7$	$7 \dashv 10$	$10 \dashv 15$
Frequenze	50	80	120	70	40

Moda - Distribuzioni numeriche in classi

Soluzioni: La frequenza maggiore si registra per la classe $3 \dashv 7$. Tuttavia le classi hanno ampiezze diverse. Calcoliamo le **frequenze specifiche**:

Intervallo di tempo	$0 \dashv 1$	$1 \dashv 3$	$3 \dashv 7$	$7 \dashv 10$	$10 \dashv 15$
Frequenze assolute	50	80	**120**	70	40
Ampiezza	1	2	4	3	5
Frequenze specifiche	**50**	40	30	23.33	8

La **classe modale** è quindi $0 \dashv 1$ e possiamo scegliere come moda il valore centrale, cioè 0.5.

Medie di posizione

Medie di posizione

- Le **medie di posizione** sono indici calcolabili per distribuzioni di frequenza di caratteri rilevati su scala **almeno ordinale**.
  - → **Qualitativi su scala ordinale**.
  - → **Quantitativi discreti e continui**.

- Coincidono con una modalità specifica del carattere che occupa una **determinata posizione** nelle modalità ordinate della distribuzione di frequenza.

- Se $x_1, x_2 \ldots, x_N$ sono le nostre osservazioni, indichiamo con

$$x_{(1)}, x_{(2)}, \ldots, x_{(N)},$$

  le osservazioni ordinate crescentemente, cioè tali per cui $x_{(i)} \leq x_{(i+1)}$.

- Sono individuate riferendosi alla quota o al numero di unità statistiche nella popolazione che presentano:
  - → modalità minori o uguali a quella individuata dalla media di posizione;
  - → modalità maggiori o uguali a quella individuata dalla media di posizione.

Mediana

La **mediana** (o quantile di livello 0.5) è la modalità m che rappresenta la posizione centrale all'interno della distribuzione di frequenze.

- → **Circa** la metà (50%) delle osservazioni presenta una modalità del carattere minore o uguale a m.
- → **Circa** la metà (50%) delle osservazioni presenta una modalità del carattere maggiore o uguale a m.

Example

Supponiamo di avere le seguenti osservazioni

$$\{1, 2, 1, 1, 3, 4, 2\}.$$

La mediana rappresenta il valore centrale della distribuzione ordinata dei dati:

$$\{\underbrace{1, 1, 1}_{\text{3 unità}}, 2, \underbrace{2, 3, 4}_{\text{3 unità}}\}.$$

In questo caso la mediana è data dal valore 2.

Mediana

Calcolo della mediana: distribuzione di unità con N osservazioni.

Distinguiamo due casi.

- **Caso 1**: N dispari. La posizione centrale è data da $\frac{N+1}{2}$. La mediana corrisponde quindi a

$$Me = x_{\left(\frac{N+1}{2}\right)}.$$

- **Caso 2**: N pari. Abbiamo due posizioni centrali, $\frac{N}{2}$ e $\frac{N}{2}+1$. In questo caso, possiamo avere che
  - $\rightarrow$ $x_{\left(\frac{N}{2}\right)} = x_{\left(\frac{N}{2}+1\right)}$, allora la mediana sarà data da

$$Me = x_{\left(\frac{N}{2}\right)} = x_{\left(\frac{N}{2}+1\right)}.$$

  - $\rightarrow$ $x_{\left(\frac{N}{2}\right)} \neq x_{\left(\frac{N}{2}+1\right)}$, allora:
    - Me è indeterminata se il carattere è qualitativo.
    - Se il carattere è quantitativo,

$$Me = \frac{x_{\left(\frac{N}{2}\right)} + x_{\left(\frac{N}{2}+1\right)}}{2}.$$

In caso di caratteri discreti, la mediana potrebbe non essere una possibile modalità. In tal caso, prestare attenzione a come commentare il risultato (vedi esempi successivi).

Mediana

Example

Supponiamo di avere le seguenti osservazioni continue

$$x_1 = 1.1,\ x_2 = 0.5,\ x_3, = 1.4,\ x_4 = 1.8,\ x_5 = 0.9,$$
$$x_6 = 1.3,\ x_7 = 0.6,\ x_8 = 1.2.$$

Calcolare la mediana e commentare il risultato ottenuto.

Mediana

Example

Supponiamo di avere le seguenti osservazioni discrete

$$x_1 = 1,\ x_2 = 5,\ x_3, = 4,\ x_4 = 8,\ x_5 = 5,$$
$$x_6 = 2,\ x_7 = 4,\ x_8 = 7.$$

Calcolare la mediana e commentare il risultato ottenuto.

Mediana

Nel caso di **distribuzioni di frequenza**, dobbiamo individuare la modalità o le modalità che occupano le posizioni centrali. La **modalità in posizione centrale** è la modalità del carattere che soddisfa contemporaneamente le due condizioni seguenti

→ La sua frequenza cumulata relativa è ≥ 0.5.
→ La sua frequenza retrocumulata relativa è ≥ 0.5.

Per individuare la mediana con distribuzioni di frequenza, procediamo calcolando le frequenze cumulate e retrocumulate relative. Individuiamo quindi la modalità o le modalità del carattere che soddisfano le precedenti condizioni.

- Se dalle condizioni viene individuata **una sola modalità** m, tale modalità sarà la nostra mediana $Me = m$.
- Se dalle condizioni vengono individuate **due modalità**, allora:
  → se il carattere è **qualitativo**, la mediana è indeterminata;
  → se il carattere è **quantitativo**, allora con modalità m_1 e m_2,

$$Me = \frac{m_1 + m_2}{2}.$$

Attenzione ai commenti nel caso di caratteri discreti.

Mediana

Example

Consideriamo le seguenti osservazioni

$$x_1 = 1,\ x_2 = 5,\ x_3, = 4,\ x_4 = 8,\ x_5 = 5,$$
$$x_6 = 2,\ x_7 = 4,\ x_8 = 7.$$

Calcolare le frequenze cumulate e retrocumulate relative. Trovare la/le modalità mediana/e ed il corrispondente valore della mediana.

Mediana

Nel caso di **distribuzioni di frequenza con modalità in classi** per un carattere **quantitativo discreto**, dobbiamo imputare un valore opportuno della classe.

La distribuzione osservata è coerente con la distribuzione ipotetica di frequenza non in classi. Operativamente, procediamo secondo le seguenti fasi.

- **Fase 1**, calcoliamo le frequenze relative cumulate F_j e retrocumulate Q_j. Individuiamo la classe o le classi che soddisfano contemporaneamente le condizioni

$$F_j \geq \frac{1}{2}, \qquad Q_j \geq \frac{1}{2}.$$

  $\rightarrow$ se individuiamo due classi, $\ell_j - u_j$ e $\ell_{j+1} - u_{j+1}$, allora la mediana è data da

$$Me = \frac{u_j + \ell_{j+1}}{2}.$$

  Attenzione al commento: la mediana non sarà uno dei valori plausibili del supporto.

  $\rightarrow$ se individuiamo una sola classe $\ell_j - u_j$, la mediana sarà all'interno della classe e proseguiamo con la fase 2.

- **Fase 2** Opero all'interno della classe mediana come fosse una normale distribuzione di frequenze.

Mediana

Example

Consideriamo le seguenti osservazioni

$$x_1 = 1,\ x_2 = 5,\ x_3, = 4,\ x_4 = 8,\ x_5 = 5,$$
$$x_6 = 2,\ x_7 = 4,\ x_8 = 7.$$

Dividere le osservazioni nelle seguenti classi: $0 - 3$, $4 - 5$, e $6 - 8$. Calcolare la mediana delle osservazioni.

Mediana

Example

Consideriamo le seguenti osservazioni

$$x_1 = 1,\ x_2 = 5,\ x_3, = 4,\ x_4 = 8,\ x_5 = 5,$$
$$x_6 = 2,\ x_7 = 4,\ x_8 = 7.$$

Dividere le osservazioni nelle seguenti classi: $0-2$, $3-4$, $5-6$ e $7-8$. Calcolare la mediana delle osservazioni.

Mediana

Nel caso di **distribuzioni di frequenza con modalità in classi** per un carattere **quantitativo continuo**, dobbiamo trovare un valore specifico rappresentativo per la mediana.

Operativamente, procediamo come segue.

- Calcoliamo i valori delle frequenze relative cumulate F_j e retrocumulate Q_j, ed individuiamo la classe o le classi che soddisfano contemporaneamente

$$F_j \geq \frac{1}{2}, \qquad Q_j \geq \frac{1}{2}.$$

  → Se vengono individuate due classi, $\ell_j \dashv u_j$ e $\ell_{j+1} \dashv u_{j+1}$, allora

$$Me = u_j = \ell_{j+1}.$$

  → Se viene identificata una sola classe, allora applichiamo la formula seguente

$$Me = \ell_j + (0.5 - F_{j-1}) \times \frac{1}{d_j},$$

  dove:
  - ℓ_j è il limite inferiore della classe mediana,
  - F_{j-1} il valore della frequenza cumulata relativa della classe antecedente alla classe mediana,
  - d_j è la frequenza relativa specifica della classe mediana.

Mediana

Derivazione della formula

$$Me = \ell_j + (0.5 - F_{j-1}) \times \frac{1}{d_j}.$$

$$F_{j-1} + (Me - \ell_j) \times d_j = 0.5$$

$$(Me - \ell_j) \times d_j = 0.5 - F_{j-1}$$

$$(Me - \ell_j) = (0.5 - F_{j-1}) \times \frac{1}{d_j}$$

$$Me = \ell_j + (0.5 - F_{j-1}) \times \frac{1}{d_j}$$

Mediana

Example

Supponiamo di avere le seguenti osservazioni continue

$$x_1 = 1.1,\ x_2 = 0.5,\ x_3, = 1.4,\ x_4 = 1.8,\ x_5 = 0.9,$$
$$x_6 = 1.3,\ x_7 = 0.6,\ x_8 = 1.2.$$

Dividere le osservazioni nelle classi $0 \dashv 1$, $1 \dashv 1.5$ e $1.5 \dashv 2$. Calcolare la mediana e commentare il risultato ottenuto.

Mediana

Example

Il giorno 23/02/2025 vengono riportate su un famoso sito di immobiliare $N = 650$ abitazioni in vendita nel comune di Sesto San Giovanni. Di queste, 105 hanno un prezzo di vendita tra 50 e 120; 169 un prezzo tra 120 e 200; 129 tra 200 e 260; 149 tra 260 e 360. Sapendo che l'ultima classe ha come estremo inferiore 360 ed ampiezza 3040:

1. Costruire una tabella di frequenze in cui vengono riportate le classi, le frequenze assolute, relative, cumulate, e retrocumulate;
2. Rappresentare i dati osservati tramite un istogramma;
3. Determinare la classe modale;
4. Stimare la mediana del prezzo di vendita di una casa.

NB: Tutti i prezzi sono riportati in migliaia di euro.

Mediana

Quantili

Un **quantile di livello** p, con $0 < p < 1$, di una distribuzione di frequenza di un carattere rilevato su **scala almeno ordinale** è quella modalità m tale che

$\rightarrow$ La frequenza cumulata relativa in corrispondenza di m è $\geq p$.

$\rightarrow$ La frequenza retrocumulata relativa in corrispondenza di m è $\geq (1 - p)$.

Interpretazione del quantile di livello p.

- Almeno il $(100 \times p)\%$ delle osservazioni esprime una modalità del carattere minore o uguale a m.
- Almeno il $(100 \times (1 - p))\%$ delle osservazioni esprime una modalità del carattere maggiore o uguale a m.

Per $p = 0.5$, otteniamo la mediana.

Quantili

Example

Consideriamo le seguenti osservazioni

$$x_1 = 1,\ x_2 = 5,\ x_3, = 4,\ x_4 = 8,\ x_5 = 5,$$
$$x_6 = 2,\ x_7 = 4,\ x_8 = 7.$$

Calcolare il quantile di livello $p = 0.25$.

Quantili

Generalmente per i quantili valgono procedure analoghe a quelle utilizzate per il calcolo della mediana, che riportiamo brevemente in seguito per il calcolo di un generico quantile q_p di livello p.

- **Quantile per distribuzioni di unità**
  - $\rightarrow$ Ordiniamo le modalità osservate in senso crescente.
  - $\rightarrow$ Calcoliamo le frequenze relative cumulate F_j e retrocumulate Q_j.
  - $\rightarrow$ Individuiamo la modalità o le modalità per cui $F_j \geq p$ e $Q_j \geq (1-p)$.
    - Con una modalità m o con due modalità uguali a m, $q_p = m$.
    - Con due modalità $m_1 \neq m_2$, se il carattere è qualitativo, il quantile è indeterminato.
    - Con due modalità $m_1 \neq m_2$, se il carattere è quantitativo,

$$q_p = p \times m_1 + (1-p)m_2$$

Attenzione all'interpretazione se il carattere è quantitativo discreto.

Quantili

- **Quantile per distribuzioni di frequenza**
  - $\rightarrow$ Calcoliamo le frequenze relative cumulate F_j e retrocumulate Q_j.
  - $\rightarrow$ Individuiamo la modalità o le modalità per cui $F_j \geq p$ e $Q_j \geq (1-p)$.
    - Se identifichiamo una sola modalità m, allora $q_p = m$.
    - Con due modalità $m_1 < m_2$, se il carattere è qualitativo, il quantile è indeterminato.
    - Con due modalità $m_1 < m_2$, se il carattere è quantitativo,

$$q_p = p \times m_1 + (1-p)m_2$$

Attenzione all'interpretazione se il carattere è quantitativo discreto.

- **Quantile per distribuzioni di frequenza con modalità in classi, carattere quantitativo discreto**
  - $\rightarrow$ Calcoliamo le frequenze relative cumulate F_j e retrocumulate Q_j.
  - $\rightarrow$ Individuiamo la classe o le classi per cui $F_j \geq p$ e $Q_j \geq (1-p)$.
    - Se identifichiamo due classi distinte $\ell_j - u_j$ e $\ell_{j+1} - u_{j+1}$, allora

$$q_p = u_j \times p + \ell_{j+1} \times (1-p).$$

Attenzione all'interpretazione se il carattere è quantitativo discreto.

    - Se identifichiamo una sola classe $\ell_j - u_j$, il quantile è sicuramente al suo interno. Esplodiamo la classe e la trattiamo come distribuzione di unità.

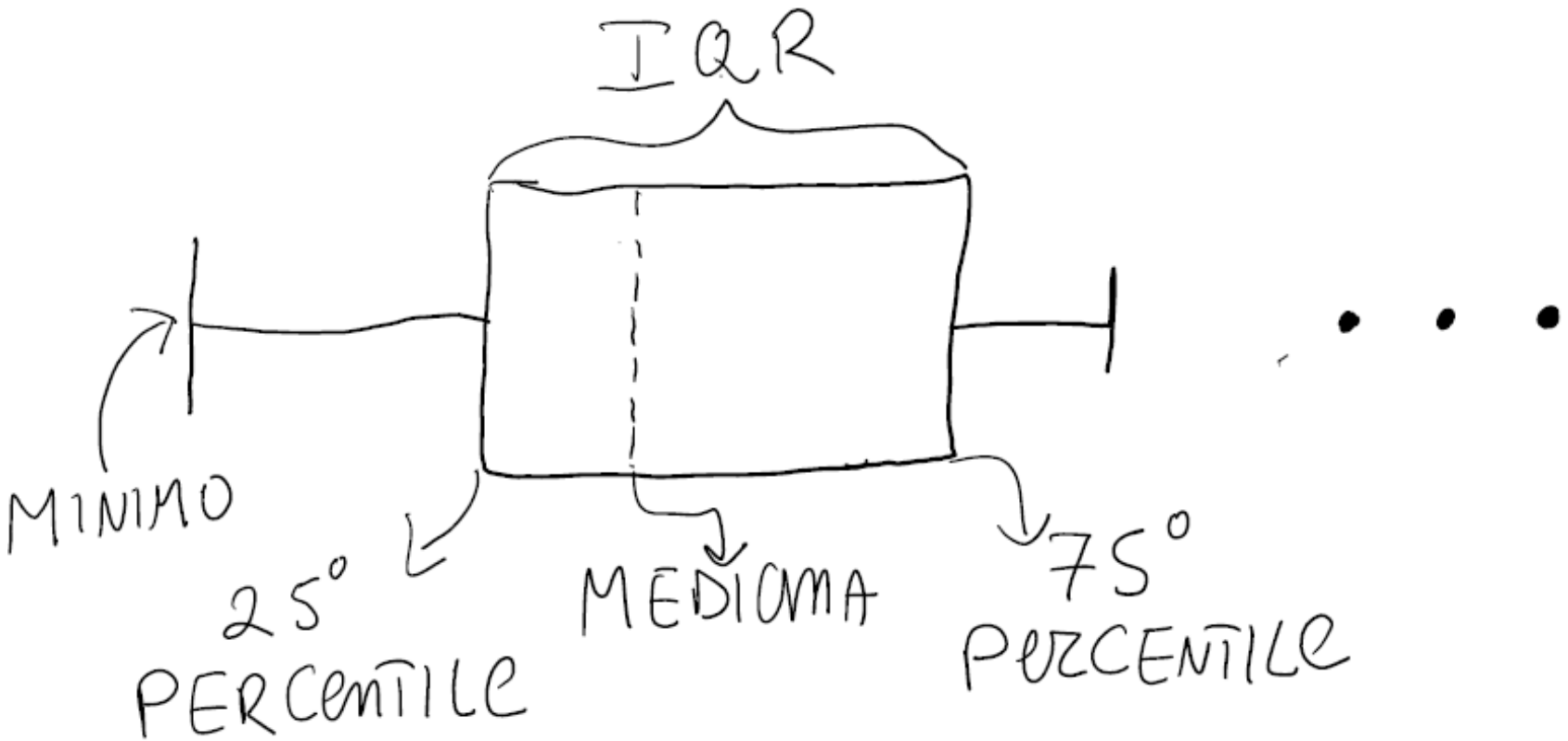
Quantili

- **Quantile per distribuzioni di frequenza con modalità in classi, carattere quantitativo continuo**
  - $\rightarrow$ Calcoliamo le frequenze relative cumulate F_j e retrocumulate Q_j.
  - $\rightarrow$ Individuiamo la classe o le classi per cui $F_j \geq p$ e $Q_j \geq (1-p)$.
    - Se identifichiamo due classi distinte $\ell_j \dashv u_j$ e $\ell_{j+1} \dashv u_{j+1}$, allora
    $$q_p = u_j = \ell_{j+1}.$$
    - Se identifichiamo una sola classe $\ell_j \dashv u_j$, il quantile è sicuramente al suo interno. Applichiamo la formula
    $$q_p = \ell_j + (p - F_{j-1}) \times \frac{1}{d_j}$$
    dove ℓ_j è il limite inferiore della classe del quantile, F_{j-1} il valore della frequenza cumulata relativa della classe antecedente alla classe del quantile, d_j è la frequenza relativa specifica della classe del quantile.

Quantili notevoli:

- **Quartili**, $p = 0.25$, $p = 0.5$ (mediana), $p = 0.75$.
- **Decili**, $p = \frac{i}{10}$, $i = 1, \ldots, 9$, dove $i = 5$ corrisponde alla mediana.
- **Centili**, $p = \frac{i}{100}$, $i = 1, \ldots, 99$, dove $i = 50$ corrisponde alla mediana.

Boxplot[4]



[4]Credits: Andrea Gilardi. Perdonatemi per la bassa qualità del disegno...

Esercizi

Example

Basandosi sulle indagini campionarie relative alle abitudini di partecipazione ad eventi culturali nell'anno 2022 per la città di Roma, è stato ricavato il seguente prospetto relativo agli individui classificati in base alla *Frequenza di Partecipazione*. I valori sono espressi in *migliaia di unità*.

Frequenza di Partecipazione	Residenti (R)	Non Residenti (NR)
Mai	27	8.9
Occasionalmente	56.3	31
Regolarmente	29.4	21.6
Frequentemente	7.1	3.8
Totale	119.8	65.3

1. Confrontare le due distribuzioni in termini di frequenze relative e **commentare il risultato**;
2. Determinare la moda per entrambe le distribuzioni (R e NR);
3. Calcolare il primo quartile del carattere *Frequenza* per la popolazione (R).

Esercizi

Medie analitiche

Medie analitiche

Le **medie analitiche** sono **indici di posizione** solo per caratteri **quantitativi, discreti e continui**.

Rappresentano un valore del supporto che sintetizza, in un punto, la distribuzione di frequenze che abbiamo osservato, secondo le funzioni definite in seguito.

In particolare, nelle prossime slides vedremo

- Media aritmetica.
- Media geometrica.
- Media quadratica.
- Media armonica.

Media aritmetica

La **media aritmetica** è l'indice di posizione più comunemente usato, e spesso viene chiamata semplicemente **media**.

Supponiamo di aver osservato i seguenti valori

$$\{x_1, \ldots, x_N\}.$$

La media aritmetica è devinita come la somma di tutte le osservazioni, divisa per il numero di valori osservati N, ovvero

$$M_1 = \bar{x} = \frac{x_1 + \cdots + x_N}{N} = \frac{1}{N} \sum_{i=1}^{N} x_i.$$

Viene indicata con M_1 o $\bar{x}$ per un insieme di osservazioni $\{x_1, \ldots, x_n\}$.

La somma di tutti i valori $\sum_{i=1}^{n} x_i$ corrisponde al **totale** osservato. La media aritmetica corrisponde alla **equa ripartizione** del totale tra tutte le N unità statistiche.

Media aritmetica

Di seguito, vediamo come calcolare la media aritmetica nel caso di diverse tipologie di distribuzioni.

- **Distribuzione di unità**, abbiamo quindi un campione o una popolazione osservata $\{x_1, \ldots, x_N\}$. La media aritmetica è semplicemente calcolata seguendo la definizione, ovvero

$$M_1 = \frac{1}{N} \sum_{i=1}^{n} x_i.$$

- **Distribuzione di frequenze**, abbiamo quindi

valore	frequenza
x_1	n_1
x_2	n_2
$\vdots$	$\vdots$
x_j	n_j
$\vdots$	$\vdots$
x_k	n_k
totale	N

La media aritmetica è quindi data da

$$\begin{aligned} M_1 &= \frac{\text{totale}}{\text{numero di osservazioni}} \\ &= \frac{1}{N} \sum_{j=1}^{k} n_j x_j = \sum_{j=1}^{k} \frac{n_j}{N} x_j \\ &= \sum_{j=1}^{k} f_j x_j. \end{aligned}$$

Media aritmetica

- **Distribuzione di frequenze in classi**, valido sia per k classi discrete che continue. Per semplicità, riportiamo solo un caso. Per calcolare la media, troviamo il valore centrale di ogni classe, e successivamente procediamo come per le distribuzioni di frequenze.

valore	frequenza	centro
$\ell_1 - u_1$	n_1	x_1^c
$\ell_2 - u_2$	n_2	x_2^c
$\vdots$	$\vdots$	$\vdots$
$\ell_j - u_j$	n_j	x_j^c
$\vdots$	$\vdots$	$\vdots$
$\ell_k - u_k$	n_k	x_k^c
totale	N	

Abbiamo che

$$x_j^c = \frac{\ell_j + u_j}{2}.$$

La media aritmetica è quindi data da

$$M_1 = \frac{1}{N}\sum_{j=1}^{k} n_j x_j^c = \sum_{j=1}^{k} \frac{n_j}{N} x_j^c$$

$$= \sum_{j=1}^{k} f_j x_j^c.$$

Proprietà della media aritmetica

Di seguito verranno presentate alcune **proprietà della media aritmetica**. Seppur valide in generale, per semplicità verranno trattate per il caso di distribuzione di unità.

Teorema

La somma degli scarti delle nostre osservazioni da un valore generico A è uguale a zero se e solo se A è la media aritmetica. In formule, abbiamo

$$\sum_{i=1}^{N}(x_i - A) = 0, \qquad \text{se e solo se} \qquad A = M_1.$$

Il fatto che sia presente "se e solo se" vuol dire che dobbiamo dimostrare entrambe le direzioni.

Proprietà della media aritmetica

$\Rightarrow$ Se $\sum_{i=1}^{N}(x_i - A) = 0$ allora $A = M_1$.

$$0 = \sum_{i=1}^{N}(x_i - A) = \sum_{i=1}^{N} x_i - \sum_{i=1}^{N} A = \sum_{i=1}^{N} x_i - NA,$$

quindi $\sum_{i=1}^{N} x_i - NA = 0$, allora

$$\sum_{i=1}^{N} x_i = NA \quad \Rightarrow \quad \frac{1}{N}\sum_{i=1}^{N} x_i = \frac{NA}{N}, \quad \text{e quindi} \quad A = \frac{1}{N}\sum_{i=1}^{N} x_i = M_1.$$

$\Leftarrow$ Se $A = M_1$ allora $\sum_{i=1}^{N}(x_i - A) = 0$. Infatti, abbiamo che

$$\sum_{i=1}^{N}(x_i - M_1) = \sum_{i=1}^{N} x_i - \sum_{i=1}^{N} M_1 = \sum_{i=1}^{N} x_i - NM_1$$
$$= \sum_{i=1}^{N} x_i - N\frac{1}{N}\sum_{i=1}^{N} x_i = \sum_{i=1}^{N} x_i - \sum_{i=1}^{N} x_i = 0.$$

Proprietà della media aritmetica

1) La somma degli scarti di $x_1, \ldots, x_N$ dalla media aritmetica è uguale a 0,

$$\sum_{i=1}^{N}(x_i - M_1) = 0.$$

2) **Proprietà di internalità della media**, ovvero la media aritmetica assume un valore compreso nel range dei dati osservati,

$$x_{(1)} \leq M_1 \leq x_{(N)},$$

dove $x_{(1)} = \min\{x_1, \ldots, x_N\}$ e $x_{(N)} = \max\{x_1, \ldots, x_N\}$.

Dimostrazione proprietà 2)

$\rightarrow$ Se $x_1 = x_2 = \cdots = x_N = x^*$, allora

$$M_1 = \frac{1}{N}\sum_{i=1}^{N} x_i = \frac{1}{N}\sum_{i=1}^{N} x^* = \frac{1}{N} N x^* = x^*,$$

quindi $x^* = x_{(1)} = M_1 = x_{(N)}$.

Proprietà della media aritmetica

→ Se i valori $x_1, \ldots, x_N$ non sono tutti uguali, allora

$$x_{(1)} < x_{(N)}.$$

Sappiamo che $\sum_{i=1}^{N}(x_i - M_1) = 0$. Allora, alcuni scarti $(x_i - M_1)$ saranno positivi e altri negativi.

- $(x_{(1)} - M_1)$ è lo scarto più piccolo, e sicuramente minore di 0.
- $(x_{(N)} - M_1)$ è lo scarto più grande, e sicuramente maggiore di 0.

Mettendo insieme le due condizioni, abbiamo

$$\begin{cases} (x_{(1)} - M_1) < 0 \\ (x_{(N)} - M_1) > 0 \end{cases} \Rightarrow \begin{cases} x_{(1)} < M_1 \\ x_{(N)} > M_1 \end{cases} \Rightarrow \quad x_{(1)} < M_1 < x_{(N)}.$$

Che conclude la dimostrazione.

Proprietà della media aritmetica

3) Proprietà di minimo di M_1, la somma degli scarti al quadrato da un generico valore A è minima quando $A = M_1$. In formule

$$\sum_{i=1}^{N}(x_i - A)^2 \geq \sum_{i=1}^{N}(x_i - M_1)^2,$$

dove l'uguaglianza vale solo se $A = M_1$.

Dimostrazione proprietà 3)

$$\begin{aligned}
\sum_{i=1}^{N}(x_i - A)^2 &= \sum_{i=1}^{N}((x_i - M_1) + (M_1 - A))^2 \\
&= \sum_{i=1}^{N}\Big[(x_i - M_1)^2 + (M_1 - A)^2 + 2(x_i - M_1)(M_1 - A)\Big] \\
&= \sum_{i=1}^{N}(x_i - M_1)^2 + \sum_{i=1}^{N}(M_1 - A)^2 + 2\sum_{i=1}^{N}(x_i - M_1)(M_1 - A) \\
&= \sum_{i=1}^{N}(x_i - M_1)^2 + N(M_1 - A)^2 + 2(M_1 - A)\underbrace{\sum_{i=1}^{N}(x_i - M_1)}_{=0}
\end{aligned}$$

Proprietà della media aritmetica

Quindi

$$\sum_{i=1}^{N}(x_i - A)^2 = \sum_{i=1}^{N}(x_i - M_1)^2 + N(M_1 - A)^2,$$

dove $(M_1 - A)^2 > 0$ se $A \neq M_1$, e $(M_1 - A)^2 = 0$ se $A = M_1$. Allora

$$\sum_{i=1}^{N}(x_i - A)^2 \geq \sum_{i=1}^{N}(x_i - M_1)^2,$$

con uguaglianza solo se $A = M_1$, che conclude la dimostrazione.

Proprietà della media aritmetica

4) **Proprietà associativa della media aritmetica**. Supponiamo di avere in totale N unità statistiche divise in k gruppi, dove il generico gruppo j di osservazioni è definito come $\{x_{1,j}, \ldots, x_{N_j,j}\}$. I gruppi hanno numerosità $N_1, \ldots, N_k$, dove

$$N = \sum_{j=1}^{k} N_j.$$

Siano inoltre $M_{1,1}, \ldots, M_{1,k}$ le medie aritmetiche specifiche di ogni gruppo, cioè

$$M_{1,j} = \frac{1}{N_j} \sum_{i=1}^{N_j} x_{i,j}.$$

La media aritmetica di tutte le osservazioni M_1 può essere espressa come media ponderata delle medie aritmetiche dei singoli gruppi $M_{1,1}, \ldots, M_{1,k}$, come

$$M_1 = \frac{1}{N} \sum_{j=1}^{k} N_j M_{1,j} = \sum_{j=1}^{k} \frac{N_j}{N} M_{1,j}.$$

Proprietà della media aritmetica

Dimostrazione proprietà 4) Definiamo

- $M_1 = \sum_{i=1}^{N} x_i = \frac{\text{somma delle osservazioni}}{\text{numero delle osservazioni}}$,
- $S_j = \sum_{i=1}^{N_j} x_{i,j} =$ somma delle osservazioni nel gruppo j-esimo,
- $S = \sum_{i=1}^{N} x_i =$ somma delle osservazioni.

dove

$$S = \sum_{j=1}^{k} S_j, \qquad M_{1,j} = \frac{S_j}{N_j}.$$

Allora, possiamo vedere che

$$M_1 = \frac{S}{N} = \frac{\sum_{j=1}^{k} S_j}{N} = \frac{1}{N} \sum_{j=1}^{k} S_j \times \frac{N_j}{N_j}$$

$$= \frac{1}{N} \sum_{j=1}^{k} \frac{S_j}{N_j} N_j = \frac{1}{N} \sum_{j=1}^{k} M_{1,j} N_j,$$

che conclude la dimostrazione.

Proprietà della media aritmetica

5) Proprietà di linearità della media aritmetica. Supponiamo di avere due caratteri X e Y legati dalla relazione funzionale

$$Y = a + bX, \qquad a, b \in \mathbb{R}.$$

Allora, la media aritmetica di Y può essere espressa come funzione della media media aritmetica di X,

$$M_1(Y) = a + bM_1(X).$$

Dimostrazione proprietà 5) Abbiamo due sequenze $\{y_1, \ldots, y_N\}$ e $\{x_1, \ldots, x_N\}$, dove

$$y_i = a + bx_i, \qquad i = 1, \ldots, N.$$

Allora

$$M_1(Y) = \frac{1}{N}\sum_{i=1}^{N} y_i = \frac{1}{N}\sum_{i=1}^{N}(a + bx_i) = \frac{1}{N}\left(\sum_{i=1}^{N} a + b\sum_{i=1}^{N} x_i\right)$$

$$= \frac{1}{N}\left(Na + b\sum_{i=1}^{N} x_i\right) = \frac{Na}{N} + b\frac{1}{N}\sum_{i=1}^{N} x_i = a + bM_1(X).$$

Proprietà della media aritmetica

6) Media aritmetica della somma di due caratteri. Supponiamo di avere due caratteri X e Y. Allora, la media aritmetica della somma

$$Z = X + Y$$

può essere scritta come somma delle medie aritmetiche,

$$M_1(Z) = M_1(X) + M_1(Y)).$$

Dimostrazione proprietà 6) Abbiamo due sequenze di partenza $\{x_1, \ldots, x_N\}$ e $\{y_1, \ldots, y_N\}$, con cui possiamo costruire una terza sequenza

$$z_i = x_i + y_i, \qquad i = 1, \ldots, N.$$

Allora

$$M_1(Z) = \frac{1}{N}\sum_{i=1}^{N} z_i = \frac{1}{N}\sum_{i=1}^{N}(x_i + y_i) = \frac{1}{N}\left(\sum_{i=1}^{N} x_i + \sum_{i=1}^{N} y_i\right)$$

$$= \frac{1}{N}\sum_{i=1}^{N} x_i + \frac{1}{N}\sum_{i=1}^{N} y_i = M_1(X) + M_1(Y),$$

che conclude la dimostrazione.

Esercizi

Example

La seguente tavola riporta le rilevazioni effettuate per 50 giorni feriali del numero di clienti in attesa alla cassa di un supermercato:

Classe	0 - 2	3 - 4	5 - 8	9 - 10	11 - 15	16 - 20
Frequenze	20	10	60	70	35	25

1. Determinare moda, mediana, quartili, e media aritmetica **commentando i risultati ottenuti**;
2. Verificare empiricamente che la somma degli scarti dalla media aritmetica sia pari a 0;

Esercizi

Esercizi

Esercizi

Example

Nel corso di Marketing Applicato il voto finale di ciascuno studente viene calcolato mediante la seguente formula:

$$\text{Voto Finale} = 2 + 0.5 \cdot (\text{Voto Intermedio A} + 1) + 0.5 \cdot (\text{Voto Intermedio B} + 1).$$

I voti ottenuti da 5 studenti nelle due prove intermedie sono:

$$\text{Prova A} = \{27, 22, 30, 33, 18\}$$
$$\text{Prova B} = \{21, 28, 30, 30, 25\}$$

Si determini il *Voto Medio Finale* sfruttando le proprietà della media aritmetica.

Esercizi

Esercizi[5]

Example

Una società sportiva ha raccolto i dati relativi a 100 ragazzi, riguardanti l'età (in anni) e l'altezza (in cm). È noto che l'altezza media complessiva del gruppo è pari a 178 cm. I dati, suddivisi per età, sono riassunti nella seguente tabella:

Età	Numero di Ragazzi	Altezza Media (cm)
16	40	175
17	40	???
18	20	180

Si richiede di:

1. Determinare l'altezza media dei ragazzi di 17 anni;
2. Calcolare l'età media complessiva del gruppo.

[5] Fonte: `https://tommasorigon.github.io/StatI/exe/exe_2_bis.html`

Esercizi

Altre tipologie di media

Media geometrica M_0

Definiamo la **media geometrica** come segue.

- Per **distribuzioni di unità** $\{x_1, \ldots, x_N\}$,

$$M_0 = \left(\prod_{i=1}^{N} x_i \right)^{\frac{1}{N}} = \sqrt[N]{\prod_{i=1}^{N} x_i}.$$

- Per **distribuzioni di frequenze**

$$M_0 = \left(\prod_{i=1}^{k} x_i^{n_i} \right)^{\frac{1}{N}} = \sqrt[N]{\prod_{i=1}^{k} x_i^{n_i}}.$$

- Per **distribuzioni di frequenze con modalità in classi**

$$M_0 = \left(\prod_{i=1}^{k} (x_i^c)^{n_i} \right)^{\frac{1}{N}} = \sqrt[N]{\prod_{i=1}^{k} (x_i^c)^{n_i}},$$

dove x_i^c è il valore centrale della classe.

Media geometrica M_0

Proprietà della media geometrica

1) Il logaritmo della media geometrica di un carattere è uguale alla media aritmetica del logaritmo dello stesso carattere, ovvero

$$\log(M_0(X)) = M_1(\log(X)).$$

Dimostrazione: ricordiamo che

- $\log(ab) = \log(a) + \log(b)$;
- $\log(a^k) = k \log(a)$.

Allora abbiamo che

$$\log(M_0(X)) = \log\left(\left(\prod_{i=1}^{N} x_i\right)^{\frac{1}{N}}\right) = \frac{1}{N}\log\left(\prod_{i=1}^{N} x_i\right)$$

$$= \frac{1}{N}\sum_{i=1}^{N}\log(x_i) = M_1(\log(X)).$$

Di conseguenza,

$$\log(M_0(X)) = M_1(\log(X)) \quad \Rightarrow \quad M_0(X) = \mathrm{e}^{M_1(\log(X))}$$

Media geometrica M_0

2) La media geometrica è la media dei numeri indice a base mobile che lascia invariato l'indice a base fissa calcolato sull'intero periodo di osservazione.

Dimostrazione: supponiamo di avere osservato $x_0, x_1, \ldots, x_N$ a diversi istanti temporali $t_0, t_1, \ldots, t_N$. Costruiamo il generico indice a base mobile all'istante temporale i-esimo come

$$\mathrm{I}_{i,i-1} = \frac{x_i}{x_{i-1}}, \qquad i = 1, \ldots, N.$$

Definiamo il numero indice a base fissa sull'intero arco temporale come $\mathrm{I} = \frac{x_N}{x_0}$.

Abbiamo quindi che

$$M_0 = \left(\prod_{i=1}^{N} \mathrm{I}_{i,i-1} \right)^{\frac{1}{N}} = \left(\frac{x_1}{x_0} \times \frac{x_2}{x_1} \times \cdots \times \frac{x_N}{x_{N-1}} \right)^{\frac{1}{N}} = \sqrt[n]{\frac{x_N}{x_0}},$$

che conclude la dimostrazione.

Media armonica M_{-1}

Definiamo la **media armonica** come segue.

- Per **distribuzioni di unità** $\{x_1, \ldots, x_N\}$,

$$M_{-1} = \frac{N}{\sum_{i=1}^{N} \frac{1}{x_i}}.$$

- Per **distribuzioni di frequenze**

$$M_{-1} = \frac{N}{\sum_{i=1}^{k} \frac{n_i}{x_i}}.$$

- Per **distribuzioni di frequenze con modalità in classi**

$$M_{-1} = \frac{N}{\sum_{i=1}^{k} \frac{n_i}{x_i^c}},$$

dove x_i^c è il valore centrale della classe.

Media quadratica M_2

Definiamo la **media quadratica** come segue.

- Per **distribuzioni di unità** $\{x_1, \ldots, x_N\}$,

$$M_2 = \left(\frac{1}{N} \sum_{i=1}^{N} x_i^2 \right)^{\frac{1}{2}}.$$

- Per **distribuzioni di frequenze**

$$M_2 = \left(\frac{1}{N} \sum_{i=1}^{k} n_i x_i^2 \right)^{\frac{1}{2}}.$$

- Per **distribuzioni di frequenze con modalità in classi**

$$M_2 = \left(\frac{1}{N} \sum_{i=1}^{k} n_i (x_i^c)^2 \right)^{\frac{1}{2}},$$

dove x_i^c è il valore centrale della classe.

Esercizi

Example

Si calcolino la media aritmetica, armonica e geometrica della seguente distribuzione di frequenze:

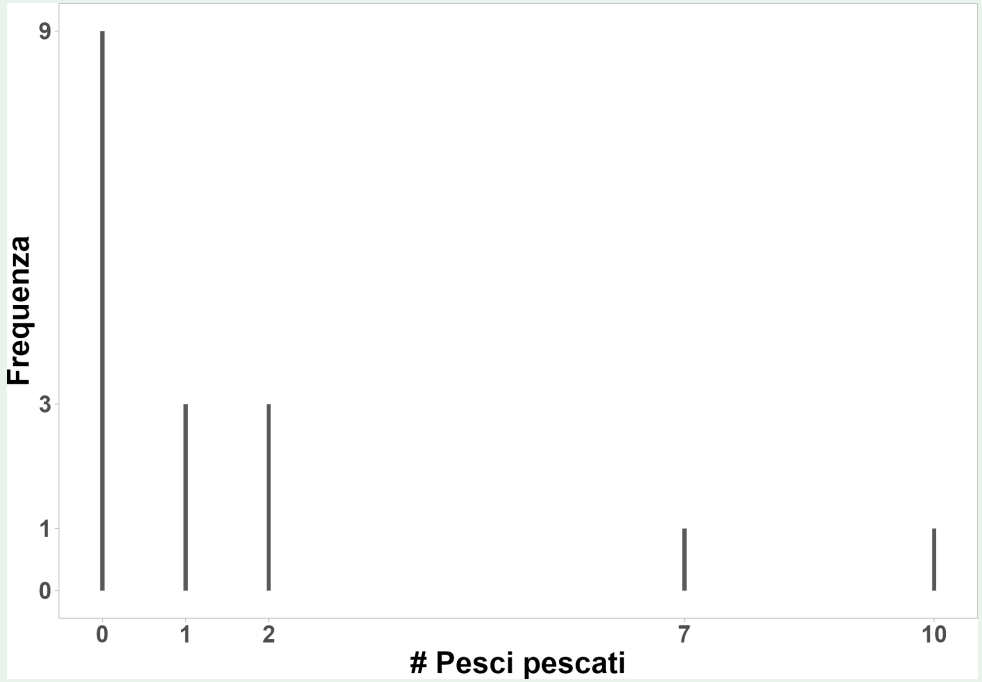
Modalità	Freq. Assolute
5	2
10	3
15	3
20	1
25	4

Esercizi

Esercizi

Example

Il seguente grafico riassume mostra la distribuzione del numero di pesci che Andrea ha pescato durante le sue ultime gite al lago:



Si richiede di calcolare la **moda**, la **moda aritmetica**, la **mediana**, ed il **massimo** di catture. Quale fra questi indici di posizione vi sembra il più adatto a riassumere la distribuzione dei dati? Perchè Andrea non sarebbe contento nel caso venisse usata la moda o la mediana?

Esercizi

Esercizi

Example

Un'industria metallurgica ha registrato la lunghezza (in mm) dei bulloni prodotti in un determinato periodo. I dati, raggruppati in classi, sono riportati nella seguente tabella:

Classe di Lunghezza	$45 \dashv 46$	$46 \dashv 47$	$47 \dashv 48$	$48 \dashv 49$	$49 \dashv 50$
Frequenza	8	15	20	10	5

Si richiede di:

1. Determinare la **moda** del campione, individuando la classe modale;
2. Calcolare la **mediana** del campione;
3. Determinare i **quantili** di ordine $p = 0.47$ e $p = 0.93$. E' possibile che il quantile di ordine $p = 0.33$ sia pari a 45.975? Rispondere alla domanda senza eseguire ulteriori calcoli.
4. Calcolare la **media aritmetica** e **geometrica** della lunghezza;
5. Quanto vale la **media aritmetica** delle lunghezze supponendo che essere vengano adesso misurate in **cm**?

Esercizi