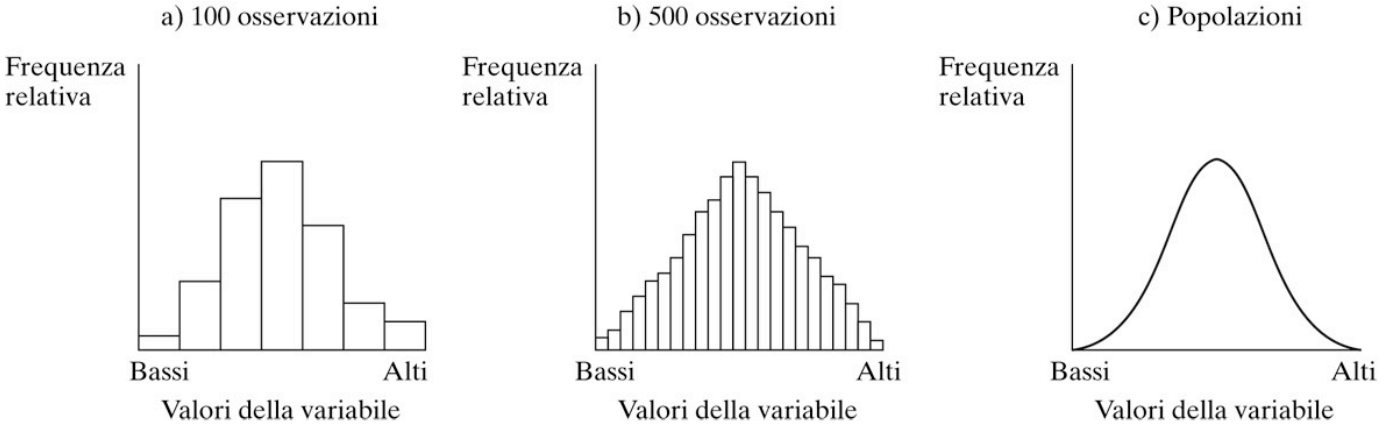
Statistica Sociale

16/03/2022

Distribuzioni di popolazioni e distribuzioni di dati campionari

- La costruzione di distribuzioni di frequenze o di istogrammi può essere fatta sia per i dati di popolazione che per quelli campionari.
- Nel primo caso facciamo riferimento a distribuzioni di dati di popolazione, nel secondo a distribuzioni di dati campionari.
- Si può dire che una distribuzione di dati campionari è una foto sfuocata della distribuzione dei dati di popolazione.
- A mano a mano che l'ampiezza campionaria aumenta le proporzioni campionarie in ciascun intervallo si approssimano sempre di più alle vere proporzioni della popolazione: la distribuzione dei dati campionari diventa sempre più simile alla distribuzione di popolazione.



Nel grafico c) è sovraimpressa una curva liscia che rappresenta la distribuzione nella popolazione (anche se una variabile è discreta, la sua distribuzione a livello di popolazione può essere adeguatamente approssimata da una curva liscia).

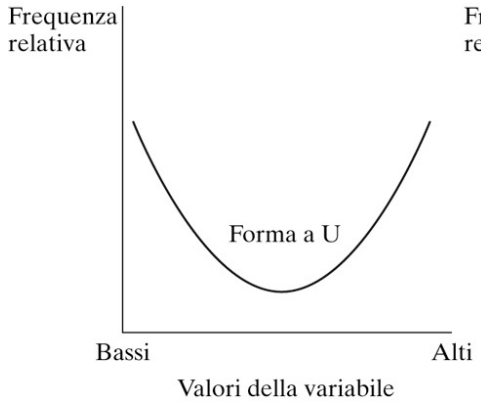
La forma della distribuzione

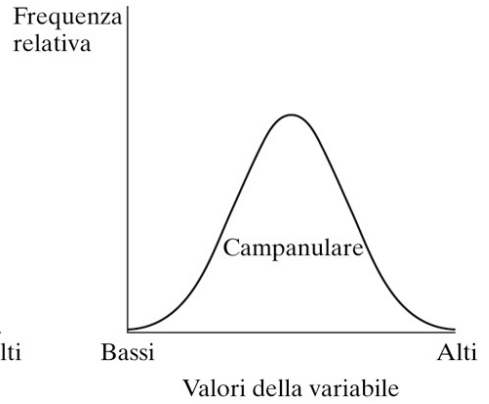
Entrambe le distribuzioni sono simmetriche: il lato della distribuzione a sinistra del valore centrale è l'immagine speculare del lato a destra dello stesso valore.

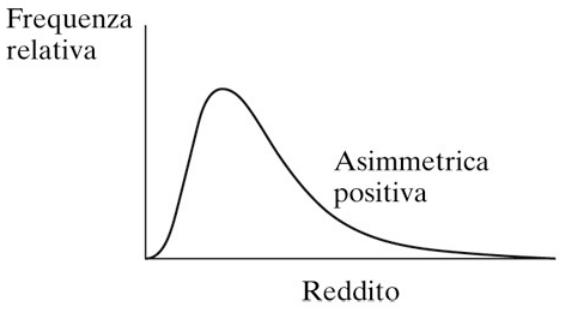
La maggior parte delle distribuzioni che vengono osservate nell'ambito della ricerca sociale non sono simmetriche.

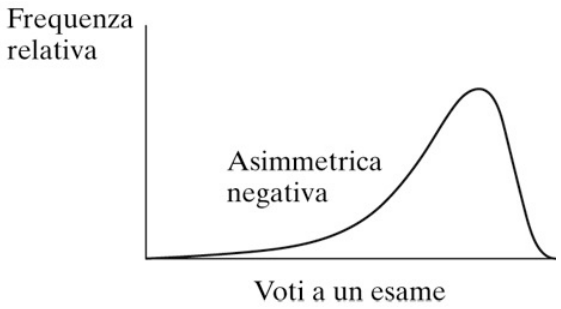
Le parti della figura in corrispondenza dei valori bassi e alti sono chiamate code della distribuzione.

Una distribuzione viene definita asimmetrica positiva o asimmetrica negativa a seconda di quella che è la sua coda più lunga.







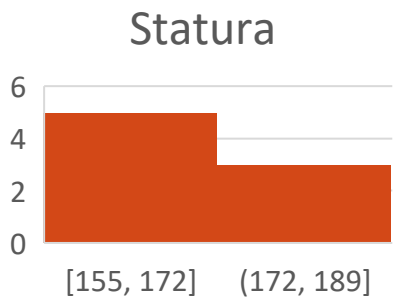


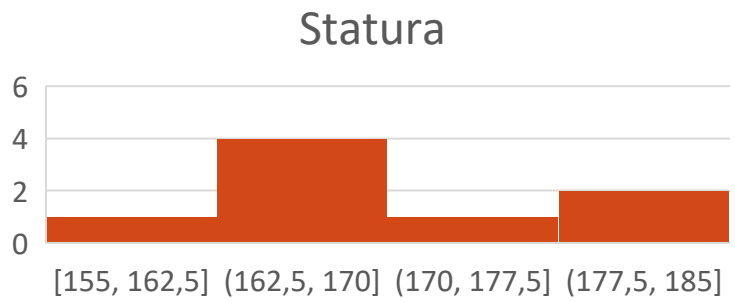
Sintesi di una distribuzione

- Distribuzione di frequenze (assolute, relative, percentuali)
- Rappresentazioni grafiche (grafico a torta, grafico a barre, istogramma)
- Indici per evidenziare con un solo valore alcune caratteristiche essenziali della distribuzione
  - Ad esempio gli indici di posizione sintetizzano con un solo numero la tendenza centrale della distribuzione

Unità	Statura
1	163
2	165
3	185
4	164
5	167
6	175
7	178
8	155

Statura	Numerosità
155	1
163	1
164	1
165	1
167	1
175	1
178	1
185	1

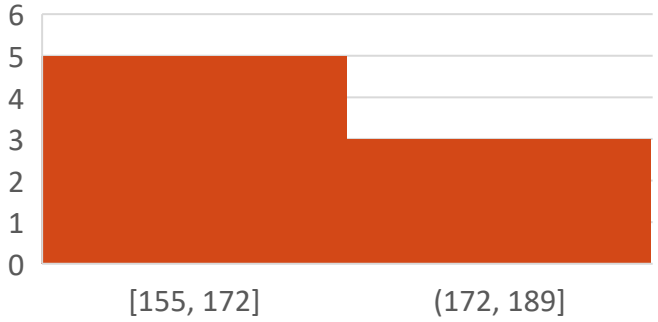




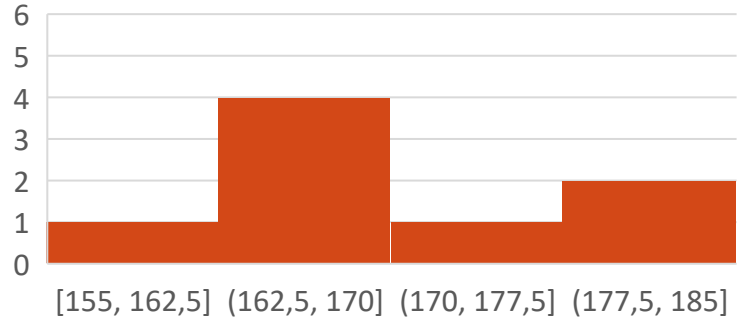
Un campione diverse rappresentazioni

Statura	Numerosità
155	1
163	1
164	1
165	1
167	1
175	1
178	1
185	1

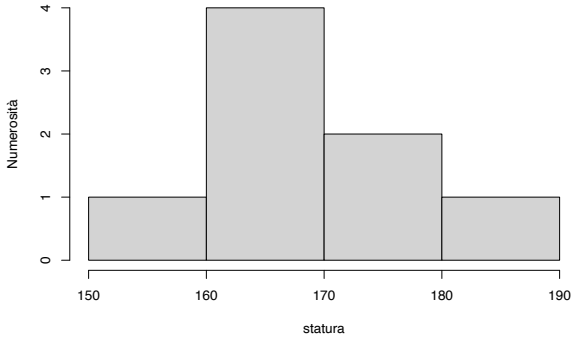
Excel default



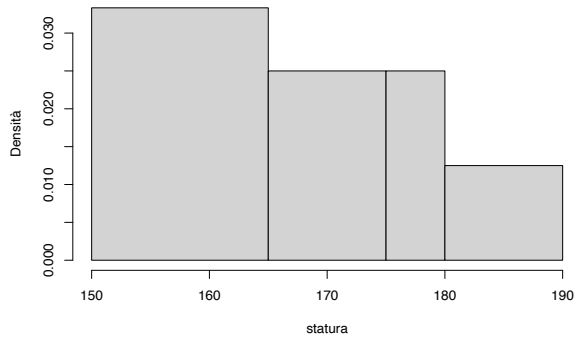
4 classi con stessa ampiezza



R default



4 classi di ampiezza diversa



La media aritmetica

La media è la somma dei valori assunti dalle osservazioni divisa per il totale delle osservazioni

Unità	Statura
1	163
2	165
3	185
4	164
5	167
6	175
7	178
8	155

$$(155 + 163 + 164 + 165 + 167 + 175 + 178 + 185)/8 = 1352/8 = 169$$

Notazione:

- *n:* numerosità campionaria
- x, y, z, … : variabili (ad esempio statura, età, …)
- $x_1, x_2, x_3, \dots, x_n$: osservazioni campionarie (la generica osservazione è x_i)
- $\bar{x}$: media della variabile x

$$\overline{statura} = (statura_1 + \text{statura}_2 + \text{statura}_3 + \text{statura}_4 + \text{statura}_5 + \text{statura}_6 + \text{statura}_7 + \text{statura}_8)/\text{n}$$

$$\overline{statura} = (statura_1 + \cdots + statura_i + \ \dots + \text{statura}_8)/\text{n}$$

$$\overline{statura} = \sum_{i=1}^{n} statura_i\ /\text{n}$$

Equivalentemente

$$\bar{x} = (x_1 + \text{x}_2 + \text{x}_3 + \text{x}_4 + \text{x}_5 + \text{x}_6 + \text{x}_7 + \text{x}_8)/\text{n}$$

$$\bar{x} = (x_1 + \cdots + x_i + \ \dots + \text{x}_8)/\text{n}$$

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

La media aritmetica

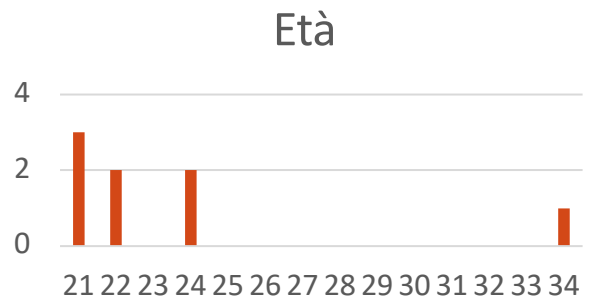
- Per caratteri quantitativi
- La media è un valore sempre compreso tra il minimo ed il massimo valore osservato
- La media può essere influenzata da un'osservazione molto "lontana" (molto più piccola/molto più grande) rispetto alle altre osservazioni -> outlier. In questi casi la media può essere poco rappresentativa dei valori tipici del campione
- La media tende verso la coda più lunga della distribuzione: più forte è l'asimmetria meno rappresentativa sarà la distribuzione
- La media è il baricentro delle osservazioni: la somma delle distanze dalla media delle osservazioni più piccole è uguale alla somma delle distanze di quelle più grandi

Unità	Età
1	24
2	21
3	34
4	22
5	21
6	22
7	24
8	21

Qual è l'età media del campione raccolto nella lezione precedente?

$$\overline{età} = \sum_{i=1}^{n} \frac{età_i}{n} = \sum_{i=1}^{8} \frac{età_i}{8} = \frac{età_1 + età_2 + \cdots (+età_i+) \ldots + età_8}{8} = \frac{24 + 21 + \cdots + 21}{8} =$$

$$= \frac{189}{8} = 23{,}6$$



La media aritmetica

- Se conosciamo la distribuzione di frequenze per le K modalità osservate, la media può essere calcolata come
  - Se conosciamo la distribuzione di frequenze assolute:

$$\bar{x} = \sum_{k=1}^{K} \frac{x_k * numerosità_k}{n} = \frac{x_1 * numerosità_1 + \cdots + x_k * numerosità_k}{n}$$

  - Se conosciamo la distribuzione di frequenze relative:

$$\bar{x} = \sum_{k=1}^{K} x_k * freq\ rel_k = x_1 * freq\ rel_1 + \cdots + x_k * freq\ rel_k$$

k	Età	Numerosità	Freq rel
1	21	3	0,375
2	22	2	0,25
3	24	2	0,25
4	34	1	0,125

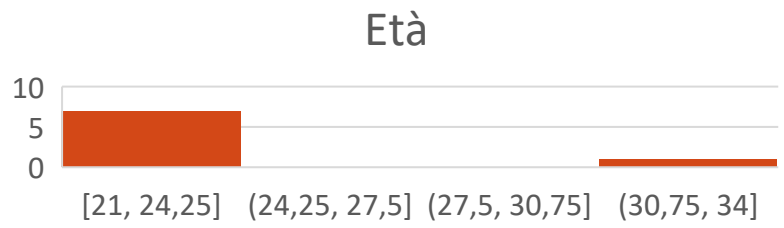
Calcoliamo di nuovo l'età media utilizzando la distribuzione di frequenze

- Con le frequenze assolute

$$\overline{età} = \sum_{k=1}^{4} \frac{età_k * numerosità_k}{8} = \frac{21 * 3 + 22 * 2 + 24 * 2 + 34 * 1}{8} = 23{,}6$$

- Con le frequenze relative

$$\overline{età} = \sum_{k=1}^{4} età_k * freq\ rel_k = 21 * 0{,}375 + 22 * 0{,}25 + 24 * 0{,}25 + 34 * 0{,}125 = \ldots$$



La media aritmetica approssimata

Quando la variabile è suddivisa in classi e non conosciamo i valori osservati ma solo la classe di appartenenza, utilizziamo il valore centrale di ogni classe per ottenere la media aritmetica approssimata

Per K classi, indichiamo il valore centrale della classe con c_k, allora la media aritmetica approssimata si calcola come:

$$\bar{x} \approx \sum_{k=1}^{K} \frac{c_k * numerosità_k}{n} \qquad \text{oppure} \qquad \bar{x} \approx \sum_{k=1}^{K} c_k * freq\ rel_k$$

Età	Numerosità	Freq rel	Valore Centrale
[19-28)	7	0.7	23.5
[28-37)	2	0.2	32.5
[37-46)	1	0.1	41.5
	10	1	

Calcoliamo la media aritmetica approssimata per l'età su un campione raccolto lo scorso anno tra gli studenti del corso di Statistica sociale:

$$\bar{x} \approx \sum_{k=1}^{3} \frac{c_k * n_k}{10} = \frac{23.5 * 7 + 32.5 * 2 + 41.5 * 1}{10} = 23.5 * 0.7 + 32.5 * 0.2 + 41.5 * 0.1 = 27.1$$

- Se il carattere è equidistribuito all'interno della classe la media aritmetica e la media aritmetica approssimata coincidono
- Se una delle classi è non limitata (es: da [37 in su) bisogna scegliere un opportuno valore che possa rappresentare la media interna della classe.

La media aritmetica ponderata

Ad ogni osservazione si attribuisce un peso per aumentare/diminuire la sua importanza nel calcolo della media

Per n osservazioni, indichiamo i pesi con $w_1, \dots, w_n$. Allora la media aritmetica ponderata si calcola come

$$\bar{x}_w = \frac{x_1 * w_1 + x_2 * w_2 + \cdots + x_i * w_i + \cdots + x_N * w_N}{w_1 + w_2 + \cdots + w_i + \cdots + w_N} = \frac{\sum x_i w_i}{\sum w_i}$$

Calcoliamo la media aritmetica ponderata per i seguenti voti considerando come pesi i CFU di ogni insegnamento

Voto	CFU
29	9
25	3
28	6
	18

$$\bar{x}_w = \frac{29 * 9 + 25 * 3 + 28 * 6}{18} = 28$$

$$\bar{x}_w = \frac{29 * 3 + 25 * 9 + 28 * 6}{18} = 26.66 \dots$$

Voto	CFU
29	3
25	9
28	6
	18

E la media aritmetica?

$$\bar{x} = \frac{29 + 25 + 28}{3} = 27.33\dots$$

La mediana

- La variabile deve essere almeno ordinabile

La mediana è la modalità assunta dall'unità centrale del campione ordinato ovvero quell'unità che divide il campione in due parti di uguale numerosità

1. Ordinare le unità secondo le modalità del carattere
2. Se *n* è dispari la mediana è:

$$\text{Me} = x_{(\frac{n+1}{2})}$$

se *n* è pari la mediana è:

$$\text{Me} = \frac{x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}}{2}$$

Le parentesi tonde indicano la posizione nel campione ordinato

Esempio voti:

i valori osservati sono: $x_1 = 29$, $x_2 = 24$, $x_3 = 28$

1. Ordiniamoli: $x_{(1)} = 24$, $x_{(2)} = 28$, $x_{(3)} = 29$
2. *n* è dispari (è uguale a 3) quindi la mediana è

$$\text{Me} = x_{(\frac{3+1}{2})} = x_{(2)} = 28$$

Esempio età:

i valori osservati sono: 22, 21, 21, 21, 21, 27, 31, 44, 23, 32

1. Ordiniamoli : 21, 21, 21, 21, 22, 23, 27, 31, 32, 44
2. *n* è pari (è uguale a 10) quindi la mediana è:

$$\text{Me} = \frac{x_{(\frac{10}{2})} + x_{(\frac{10}{2}+1)}}{2} = \frac{x_{(5)} + x_{(6)}}{2} = \frac{22+23}{2} = 22.5$$

La mediana

Per una variabile qualitativa ordinale o per una variabile quantitativa suddivisa in classi si calcola la classe mediana

1. Ordina le classi/modalità
2. Se *n* è dispari la mediana è:

$$\text{Me} = x_{(\frac{n+1}{2})}$$

se *n* è pari la mediana è:

$$\text{Me} = \frac{x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}}{2}$$

Le parentesi tonde indicano la posizione nel campione ordinato

- La tabella riporta la distribuzione della variabile livello di istruzione per i cittadini statunitensi di età superiore ai 24 anni

T. studio	Freq. (in milioni)	%
Senza titolo di sc. m. sup.	30	15.9%
Sc. m. sup.	56	29.6%
Studi univ. senza .titolo	38	20.1%
Laurea di I livello	14	17.4%
Laurea di II livello	32	16.9%
Master	13	6.9%
Dottorato ricerca	16	13.2%

(Fonte: *American Community Survey*, Bureau of the Census USA)

- *n* è uguale a 199 (30+56+…+16), quindi la posizione mediana corrisponde alla 95-esima osservazione:

$$\text{Me} = x_{(\frac{199+1}{2})} = x_{(100)}$$

  - La prima modalità «Senza titolo» include 30 osservazioni;
  - Le prime due modalità insieme («Senza titolo» e «Sc.m.sup») includono 86 osservazioni $(30 + 56)$;
  - Le prime tre modalità includono 124 osservazioni quindi la mediana ricade in questa classe che è detta classe mediana:

$$\text{Me} = x_{(100)} = \text{«}Studi\ univ.\ senza\ titolo\text{»}$$

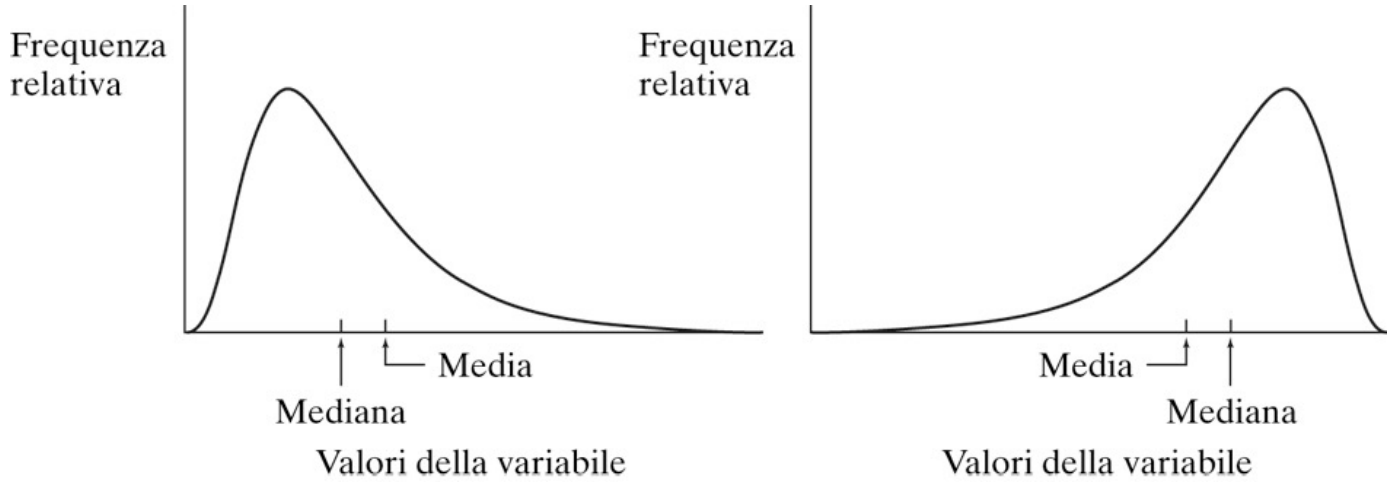
Media vs mediana

Da un'indagine campionaria ISTAT sul reddito delle famiglie italiane nel 2009 è emerso un reddito medio di 2480 euro ed un reddito mediano di 2050 euro.

- Il reddito medio si può interpretare come quel valore che avrebbe dovuto percepire ogni famiglia se l'ammontare totale fosse stato ripartito equamente tra tutte le famiglie
- Il reddito mediano ci dice che il 50% delle famiglie italiane nel 2009 ha percepito un reddito minore o uguale a 2050 euro, mentre l'altro 50% superiore al valore mediano

La mediana è più robusta della media ai valori estremi/code

Quando una distribuzione è simmetrica media e mediana coincidono



La moda

Può essere calcolata per tutti i tipi di variabili (anche per i qualitativi nominali)

La moda è definita come la modalità a cui è associata la frequenza osservata più elevata

Quando il numero di modalità possibili è alto la moda è probabilmente “troppo sintetica”

Se il carattere è suddiviso il classi parliamo di classe modale

- Maggiore approssimazione
- Se le classi sono di ampiezza diversa vanno confrontate le densità (vd istogramma) e non le frequenze!

Esistono distribuzioni unimodali o bimodali

- trova un esempio grafico di distribuzione unimodale e distribuzione bimodale nelle slides precedenti

La moda non è sensibile agli outlier

- Calcola la moda del carattere “colore degli occhi” utilizzando il campione raccolto in classe

Esercizi

Calcola la media, la mediana e la moda per questi valori: 10, 5, 20, 5, 15

- Come cambiano i tre indici se aggiungo 50 al valore più grande osservato?

$\bar{x} = 11$	$\bar{x} = 21$
Me$= 10$	Me$= 10$
Mo$= 5$	Mo$= 5$

Completa la frase:

In una distribuzione bimodale…

- A: Media, moda e mediana coincidono se la distribuzione è simmetrica
- B: Media e mediana coincidono se la distribuzione è simmetrica
- C: Media, moda e mediana non coincidono