



Cofinanziato
dall'Unione europea



DATA LAB

GUARDA AVANTI

Big Data, nuove competenze
per nuove professioni.



“Anticipare la crescita con le nuove competenze sui Big Data” Operazione Rif. PA 2023-19167/RER approvata con

DGR

The background of the slide is an underwater photograph. On the left, a scuba diver in dark gear and a yellow BCD is swimming towards the right. On the right, a massive, dense school of silver fish, possibly sardines, fills the water column. The water is a deep blue, and bubbles are visible rising from the diver.

ARTIFICIAL INTELLIGENCE & MACHINE LEARNING

Operazione Rif. PA 2023-19167/RER/10/1, “ANTICIPARE LA CRESCITA CON LE NUOVE
COMPETENZE SUI BIG DATA”, approvata dalla Regione Emilia-Romagna con DGR n° 843
del 29/05/2023 e co-finanziata dal Fondo Sociale Europeo Plus 2021-2027



LE MISURE DI TENDENZA CENTRALE

Individuare un indice che
rappresenti significativamente
un

insieme di dati statistici.

Esempio:

Nella tabella seguente sono riportati i valori del tasso glicemico rilevati su 10 pazienti:

Paziente	Glicemia (mg/100cc)
1	$x_1=103$
2	$x_2=97$
3	$x_3=90$
4	$x_4=119$
5	$x_5=107$
6	$x_6=71$
7	$x_7=94$
8	$x_8=81$
9	$x_9=92$
10	$x_{10}=96$
Totale	950

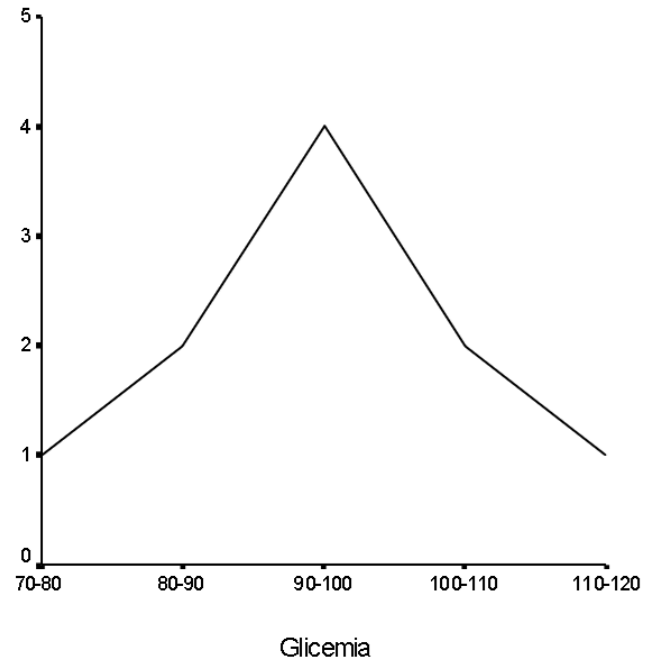
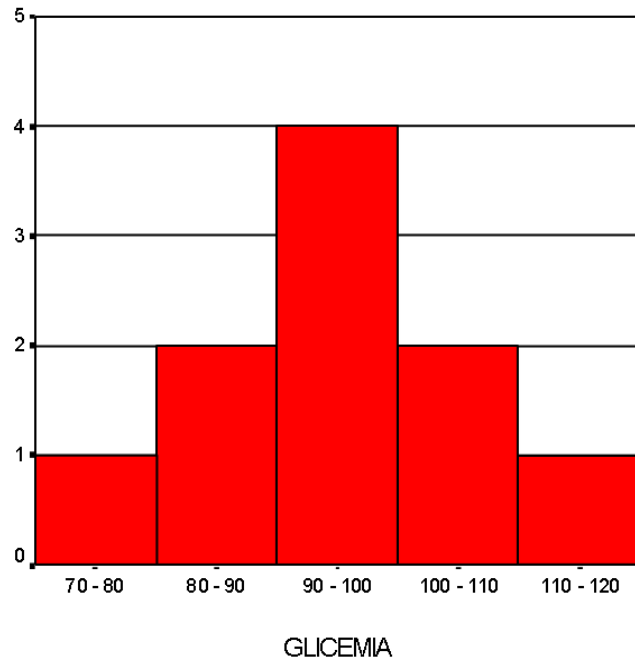
Calcolo delle frequenze di ogni classe: assolute e relative percentuali

Classi di valori di glicemia
70 ——— 80
80 ——— 90
90 ——— 100
100 ——— 110
110 ——— 120
Totale

Frequenza assoluta
1
2
4
2
1
10

Frequenza %
$1 / 10 \cdot 100\% = 10 \%$
$2 / 10 \cdot 100\% = 20 \%$
$4 / 10 \cdot 100\% = 40 \%$
$2 / 10 \cdot 100\% = 20 \%$
$1 / 10 \cdot 100\% = 10 \%$
100 %

Calcolo delle frequenze di ogni classe: assolute e relative percentuali



LE MISURE DI POSIZIONE

- ✓ media aritmetica;
- ✓ mediana;
- ✓ moda;
- ✓ media armonica;
- ✓ media geometrica.

LA MEDIA ARITMETICA

DEFINIZIONE: La media aritmetica è quel valore che avrebbero tutte le osservazioni se non ci fosse la variabilità (casuale o sistematica).

Più precisamente, è quel valore che sostituito a ciascun degli n dati ne fa rimanere costante la somma.

dato un insieme di n elementi $\{x_1, x_2, \dots, x_n\}$

Si dice **media aritmetica semplice** di n numeri il numero che si ottiene dividendo la loro somma per n .

$$\bar{x} = \frac{x_1 + x_2 + \dots x_n}{n}$$

Formalmente possiamo esprimere la media aritmetica semplice attraverso la seguente formula:

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

Nell'Esempio in esame si ha:

Paziente	Glicemia (mg/100cc)
1	$x_1=103$
2	$x_2=97$
3	$x_3=90$
4	$x_4=119$
5	$x_5=107$
6	$x_6=71$
7	$x_7=94$
8	$x_8=81$
9	$x_9=92$
10	$x_{10}=96$
Totale	950

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n} = \frac{950}{10} = 95mg$$

Esempio Riportiamo i tempi di sopravvivenza (mesi) di
19 pazienti con cancro

Mesi di sopravvivenza (x_i)	Frequenza (f_i)
8,5	2
9,2	4
7,3	8
6,8	2
10,1	3
Totale	19

$x_i \cdot f_i$
17
36,8
58,4
13,6
30,3
156,1

MEDIA ARITMETICA PESATA (o PONDERATA)

Si dice media aritmetica pesata di n numeri:

$$\frac{x_1 \cdot p_1 + x_2 \cdot p_2 + \dots + x_m \cdot p_m}{p_1 + p_2 + \dots + p_m}$$

Dove i pesi p_j sono le frequenze assolute di ogni modalità.

$$\bar{x} = \frac{\sum_{i=1}^k x_i f_i}{n} = \frac{156,1}{19} = 8,2 \text{ mesi}$$

PROPRIETA' DELLA MEDIA ARITMETICA

- ✓ compresa tra il minimo dei dati e il massimo dei dati;
- ✓ la somma degli scarti dalla media aritmetica è sempre uguale a zero;
$$\sum (x_i - \bar{x}) = 0$$
- ✓ la somma degli scarti al quadrato dalla media aritmetica assume valore minimo;

$$\sum (x_i - \bar{x})^2 < \sum (x_i - z)^2$$

Esempio: Lunghezza (cm) in un campione di 66 neonati

55.9	51.3	53.0	50.5	54.9	53.4	53.7	50.0	53.8	52.5	55.6
47.9	54.3	56.0	51.8	54.1	55.6	57.6	53.3	51.1	54.3	52.3
55.3	52.4	56.3	53.7	54.4	54.5	52.5	52.7	51.4	55.5	52.7
57.4	51.7	50.8	49.4	52.0	53.7	54.8	53.5	49.5	50.4	56.4
48.5	53.1	49.5	53.2	53.1	52.6	54.3	54.9	53.7	55.2	51.7
51.4	51.0	52.6	52.8	59.3	56.4	51.5	58.9	52.3	54.6	53.8

la media aritmetica dei 66 valori di lunghezza è:

$$=(55.9+51.3+53.0+50.5+54.9+53.4+\dots+53.8)/66$$

$$= 3517.500/66$$

$$= 53.295$$

Media aritmetica per dati raggruppati in classi

Valore centrale della classe X_i	f_i	%	$X_i f_i$
48.0	2	3.03	96.00
49.5	3	4.55	148.50
51.0	12	18.18	612.00
52.5	15	22.73	787.50
54.0	14	21.21	756.00
55.5	10	15.15	555.00
57.0	5	7.58	285.00
58.5	4	6.06	234.00
60.0	1	1.52	60.00
	66	100	3534.00

$$\bar{x} = \frac{48.0 \times 2 + 49.5 \times 3 + \dots 60.0 \times 1}{2 + 3 + \dots 1} = \frac{3534.0}{66} = 53.545$$

La **media aritmetica** è la misura di posizione più usata ma, a volte, altre misure come la **mediana** e la **moda** si dimostrano utili.

Si consideri un campione di valori di velocità media di corsa in km/h misurati in 7 persone

{8, 5, 7, 6, 35, 5, 4}

In questo caso, la media che è = 10 km/h non è un valore tipico della distribuzione: soltanto un valore su 7 è superiore alla media!

Limite della media aritmetica:
è notevolmente influenzata dai valori estremi della distribuzione.

Esempio Età alla morte di 5 soggetti

$$\begin{array}{lll} x_1 = 34 \text{ anni}; & x_2 = 70 \text{ anni}; & x_3 = 74 \text{ anni}; \\ x_4 = 64 \text{ anni}; & & x_5 = 68 \text{ anni}. \end{array}$$

La media aritmetica è pari a:

$$\bar{x} = (34 + 70 + 74 + 64 + 68) / 5 = 62 \text{anni}$$

LA MEDIANA

DEFINIZIONE: La mediana (Me) è quell'osservazione che bipartisce la distribuzione in modo tale da lasciare al “di sotto” lo stesso numero di termini che lascia al “di sopra”.

L'idea che è alla base della **mediana** è di cercare un numero che sia più grande di un 50% delle osservazioni e più piccolo del restante 50%.

Ritornando all'Esempio della Glicemia, per il calcolo della mediana è necessario disporre i dati in ordine crescente:

71, 81, 90, 92, 94, 96, 97, 103, 107, 119

$$\text{Me} = (94+96)/2 = 95 \text{ mg/100 cc}$$

Il fatto che mediana e media aritmetica in questo caso coincidano non è casuale in quanto la distribuzione è **simmetrica**.

Ma, in generale, ciò non avviene.

Vantaggio nell'uso della mediana: non è influenzata dalle osservazioni anomale o estreme.

Fasi operative per il calcolo della mediana

1. ordinamento crescente dei dati;
2. se il numero di dati n è dispari, la mediana corrisponde al dato che occupa la $(n+1)/2$ esima posizione;
3. se il numero di dati n è pari, la mediana è data dalla media aritmetica dei due dati che occupano la posizione $n/2$ e quella $(n/2)+1$.

LA MODA

DEFINIZIONE: La Moda (Mo) è l'osservazione che si verifica con maggior frequenza in una data distribuzione.

Si possono avere anche più valori modali.

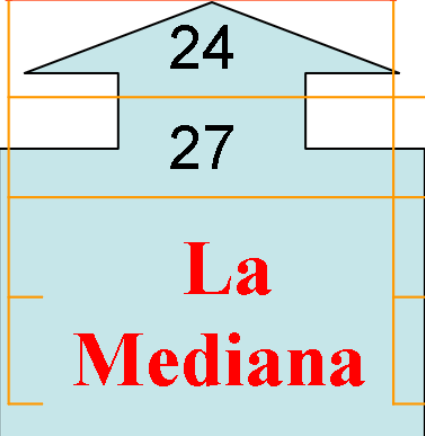
Mesi sopravvivenza (x_i)	Frequenze	Frequenze Cumulate	Cum %
6,8	2	2	10.5
7,3	8	10	52.6
8,5	2	12	63.1
9,2	4	16	84.2
10,1	3	19	100
Totale	19		

Media aritmetica= 8,2 mesi
 Mediana= 7,3 mesi
 Moda=7,3 mesi

In presenza di una distribuzione di frequenze è necessario considerare le frequenze cumulate

Voti ordinati (x_i)	Frequenze (f_i)	Freq. Cum. (F_i)	Freq.Cum. ($F_i\%$)
18	2 (10.5)	2	10.5
20	4 (21.0)	2+4 = 6	31.5
22	8 (42.1)	6+8 = 14	73.6
24	2 (10.5)	14+2 = 16	84.1
27	2 (10.5)	16+2 = 18	94.6
30	1 (5.4)	18+1 = 19	100
Totale	19		

Voti ordinati	Frequenze	Freq.Cum. F_i	Freq.Cum. $F_i\%$
18	2 (10.5)	2	10.5
20	4 (21.0)	6	31.5
22	8 (42.1)	14	73.6
24	2 (10.5)	16	84.1
27	2 (10.5)	18	94.6
	1 (5.4)	19	100
	19 (100.0)		



**La
Mediana**

I QUANTILI

- ✓ Generalizzano la mediana;
- ✓ L'idea alla base di un **quantile- p** dove $p \in [0; 1]$ è di cercare un numero che sia più grande $p\%$ dei dati osservati e più piccolo del restante $(1-p\%)$ dei dati.

I quantili con p uguale a 0,25; 0,50 e 0,75 vengono chiamati rispettivamente il primo, il secondo e il terzo **quartile**.

Dividono la popolazione in quattro parti uguali. Si osservi che il 2 quartile coincide con la mediana.

I quantili con $p = 0,01; \dots ; 0,99$ si chiamano **percentili**.

Quale misura di posizione usare?

A quale misura di tendenza centrale ci riferiamo?

- Il proprietario di una ditta afferma "Lo stipendio mensile nella nostra ditta è **2.700** euro"
 - Il sindacato dei lavoratori dice che "lo stipendio medio è di **1.700** euro".
 - L'agente delle tasse dice che "lo stipendio medio è stato di **2.200** euro".
- Queste risposte diverse sono state ottenute tutte dai dati della tabella.

Media aritmetica	= euro 2.700
Mediana	= euro 2.200
Moda	= euro 1.700

Stipendio mensile	N° di lavoratori
1.300	2
1.700	22
2.200	19
2.600	3
6.500	2
9.400	1
23.000	1

Interpretazione delle misure di posizione

- La **media aritmetica** indica che, se il denaro fosse distribuito in modo che ciascuno ricevesse la stessa somma, ciascun dipendente avrebbe avuto 2.700 euro
- La **moda** ci dice che la paga mensile più comune è di 1.700.euro
- La moda si considera spesso come il valore tipico dell'insieme di dati poiché è quello che si presenta più spesso. **Non tiene però conto degli altri valori** e spesso in un insieme di dati vi è **più di un valore** che corrisponde alla definizione di moda.
- La **mediana** indica che circa metà degli addetti percepiscono meno di 2.200.euro, e metà di più.
- La mediana **non è influenzata dai valori estremi** eventualmente presenti ma solo dal fatto che essi siano sotto o sopra il centro dell'insieme dei dati.

Relazione tra media, mediana e moda

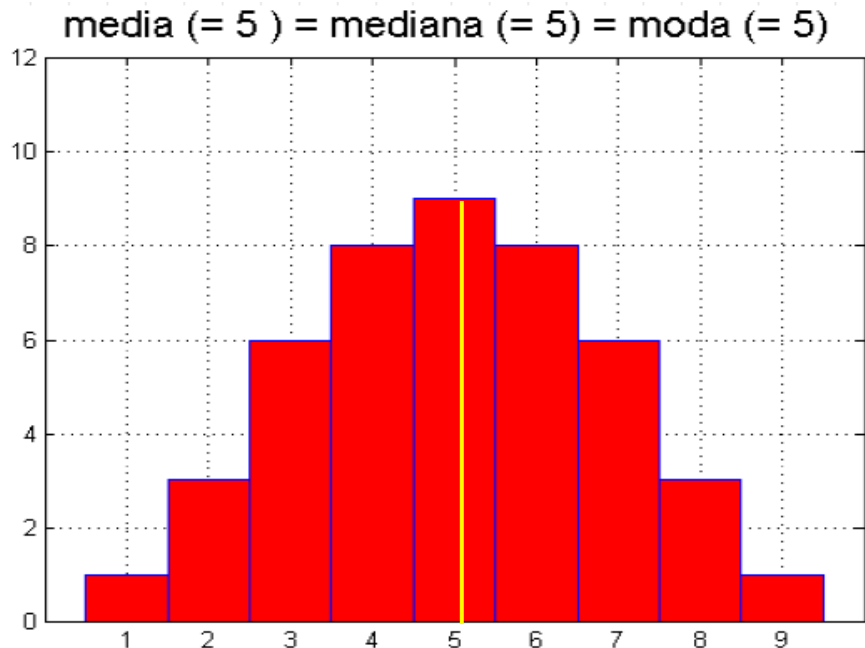
In una distribuzione perfettamente **simmetrica**, la media, la mediana e la moda hanno lo stesso valore. In una distribuzione **asimmetrica**, la media si posiziona nella direzione dell'asimmetria. Nelle distribuzioni di dati biologici, l'asimmetria è quasi sempre verso destra (asimmetria positiva, verso i valori più elevati), e quindi la media è maggiore della mediana o della moda

DISTRIBUZIONE SIMMETRICA

Le osservazioni equidistanti dalla mediana (coincidente in questo caso col massimo centrale) presentano la stessa frequenza relativa

Un esempio importante è fornito dalla ***distribuzione normale***

Media = Mediana = Moda



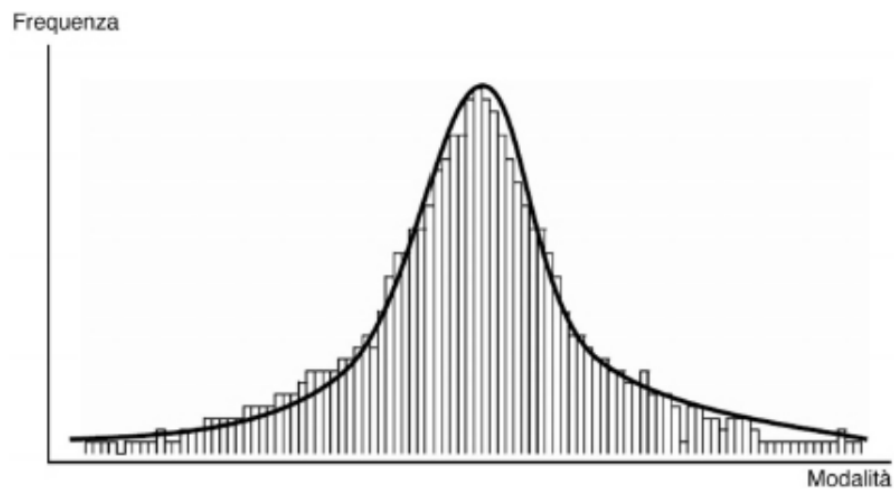
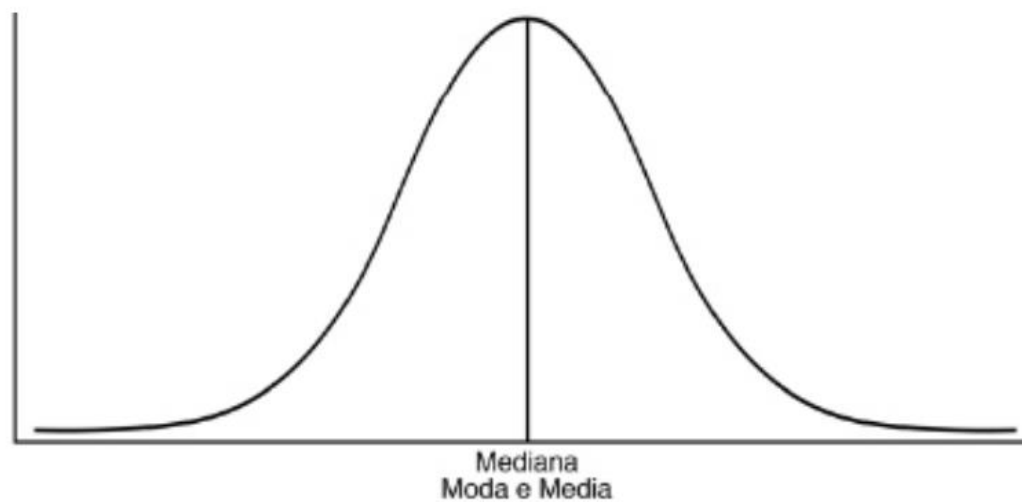


Fig. 3.19 Esempio di approssimazione della rappresentazione della distribuzione di frequenza mediante una curva continua.

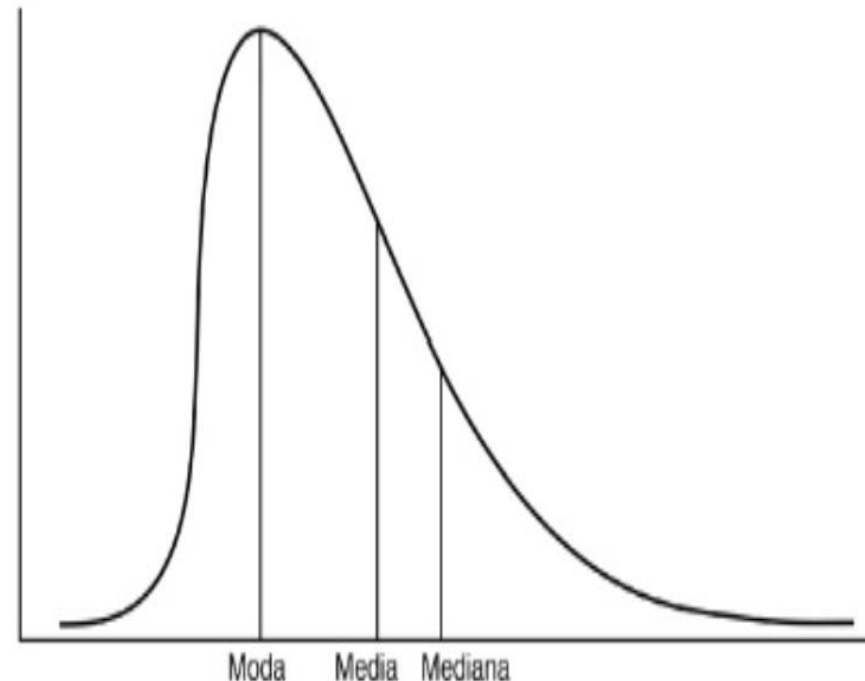
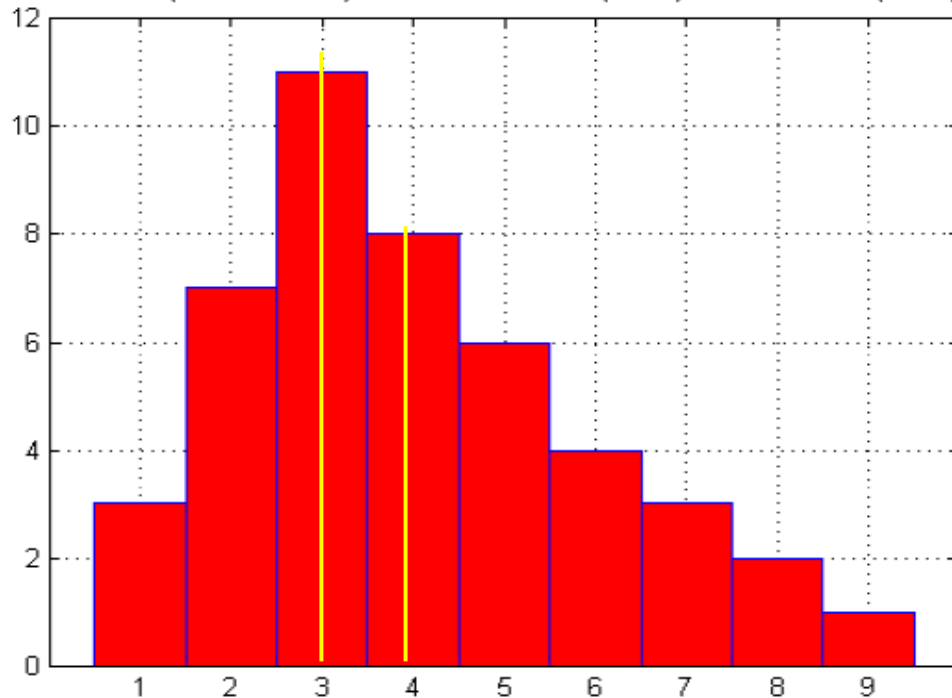


DISTRIBUZIONE ASIMMETRICA POSITIVA

La curva di frequenza ha una coda più lunga a destra
del massimo centrale

Media > Mediana > Moda

media (= 4.044) > mediana (= 4) > moda (= 3)

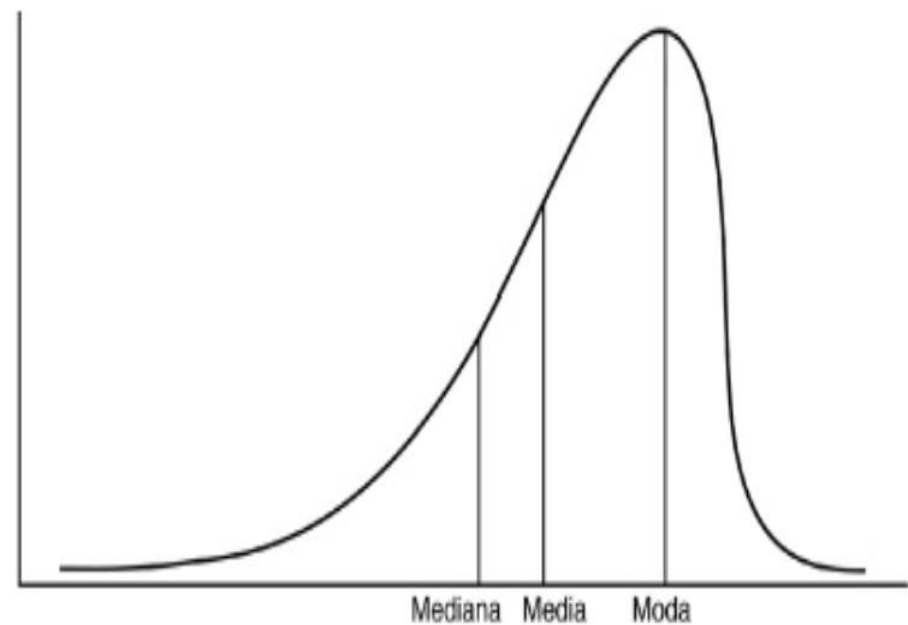
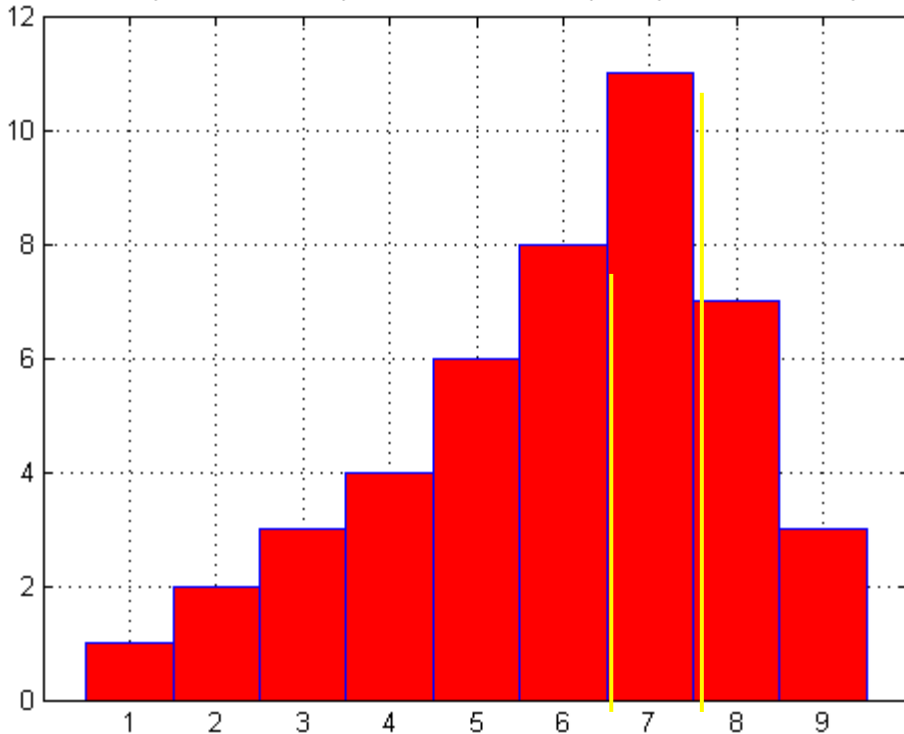


DISTRIBUZIONE ASIMMETRICA NEGATIVA

La curva di frequenza ha una coda più lunga a sinistra
del massimo centrale

Media < Mediana < Moda

media (= 5.9556) < mediana (= 6) < moda (= 7)

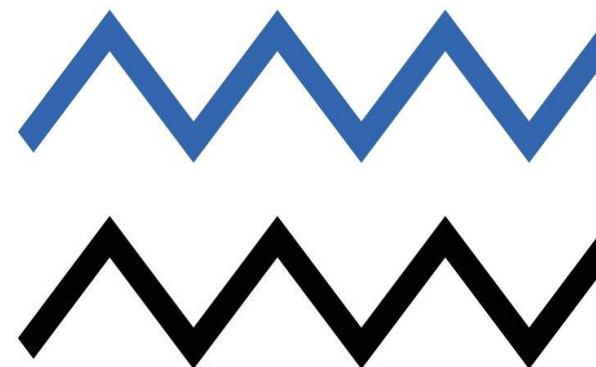




GUARDA AVANTI

Big Data, nuove competenze
per nuove professioni.

www.bigdata-lab.it



Università
degli Studi
di Ferrara



UNIMORE
UNIVERSITÀ DEGLI STUDI DI
MODENA E REGGIO EMILIA



UNIVERSITÀ
DI PARMA



POLITECNICO
MILANO 1863
POLO TERRITORIALE DI
PIACENZA



UNIVERSITÀ
CATTOLICA
del Sacro Cuore