Cinque numeri di sintesi e boxplot

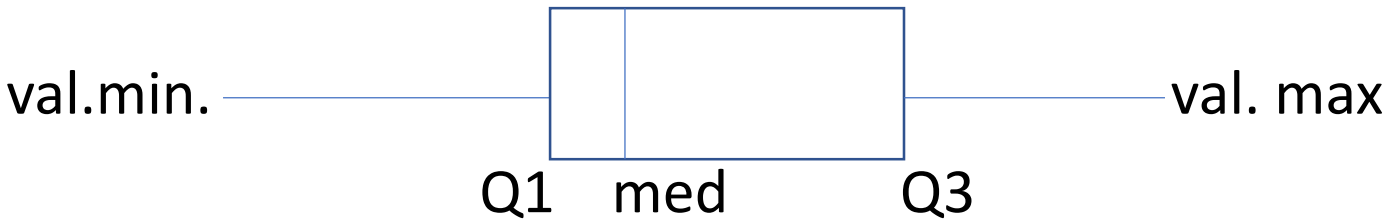
Le 5 quantità: valore minimo, , Q_1, mediana, Q_3, valore massimo permettono una descrizione sintetica dei dati.

La presenza eventuale di valori estremi è rivelata dal range, mentre la mediana e i quartili danno una misura del valore centrale e della dispersione non influenzate da valori estremi.

Il boxplot fornisce una rappresentazione grafica dei dati sulla base di questi cinque numeri.

Boxplot

- Tracciamo un segmento orizzontale dal valore minore al valore maggiore
- Al segmento sovrapponiamo un rettangolo ("box") che si estende dal primo al terzo quartile
- Il rettangolo è diviso in due parti da un segmento verticale in corrispondenza della mediana (secondo quartile)



Esempio

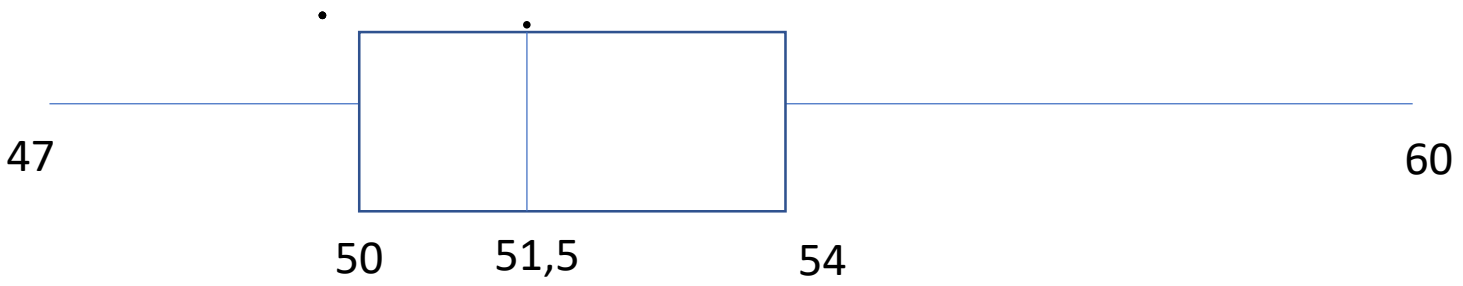
Reddito annuo (in migliaia di euro) di 42 dipendenti di un'azienda informatica

med=(51+52)/2=51,5

Q1=50

Q3=54

reddito	freq.
47	4
48	1
49	3
50	5
51	8
52	10
54	5
56	2
57	3
60	1



Esempio

Voti conseguiti all'esame di MMIB da un CCS di 10 studenti

24 25 25 25 26 | 26 27 28 29 30

Q1 Q2 Q3

Valore min. = 24 valore max. =30

Med=(26 + 26)/2=26

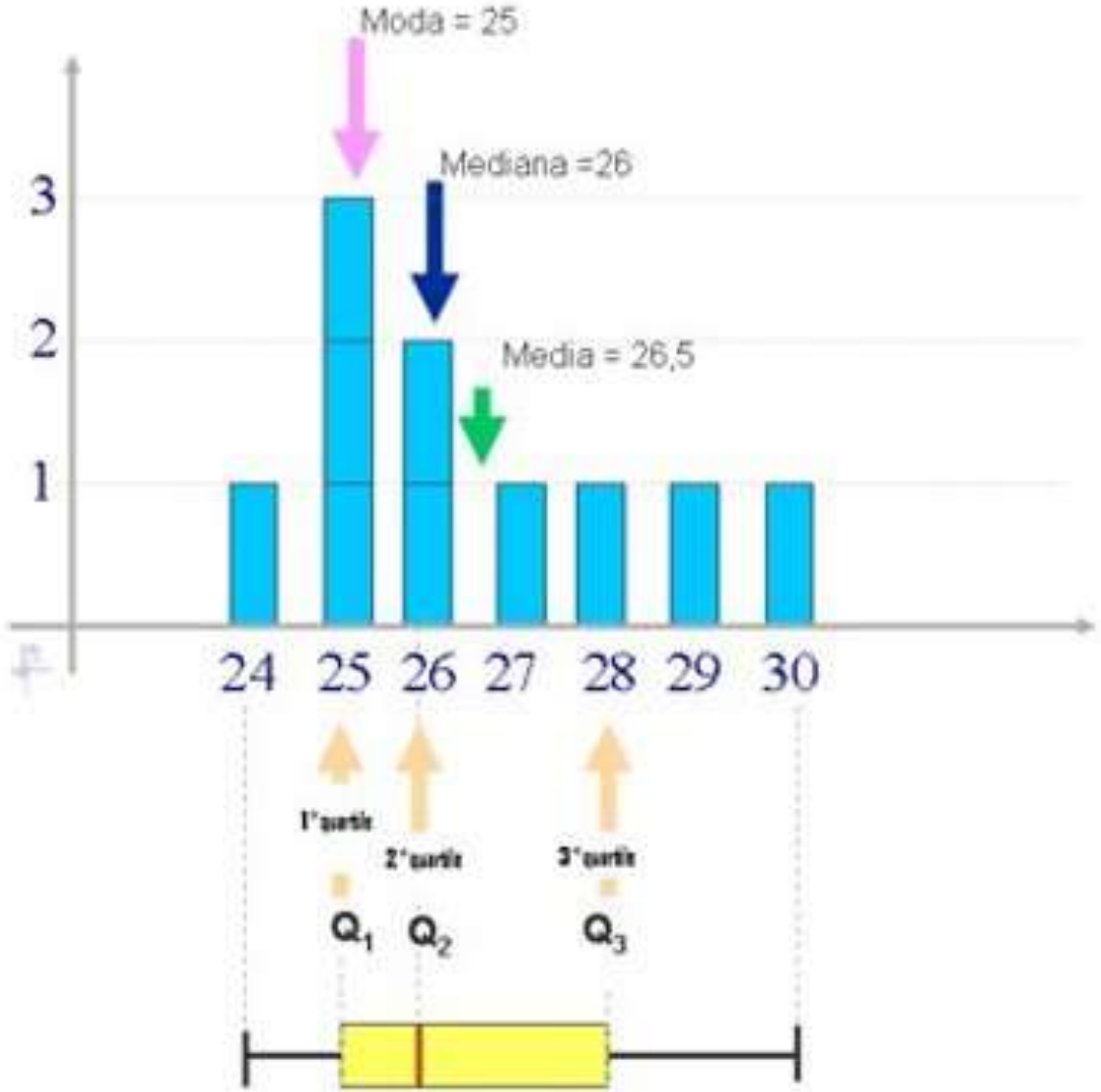
Q1=25

Q3=28

Esempio

Tabella delle frequenze

Voto	frequenza
24	1
25	3
26	2
27	1
28	1
29	1
30	1

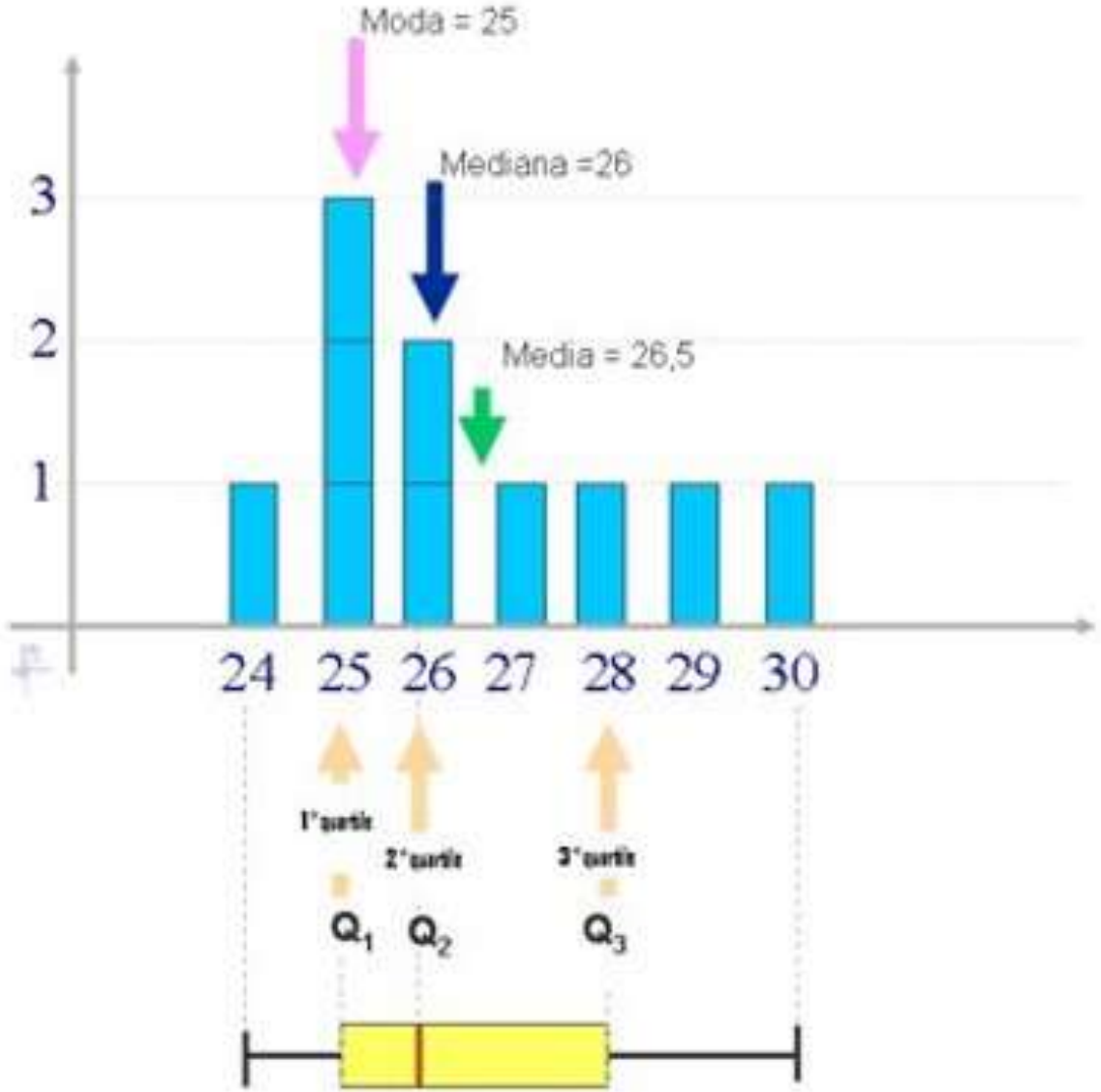


Boxplot

Esempio

Tabella delle frequenze

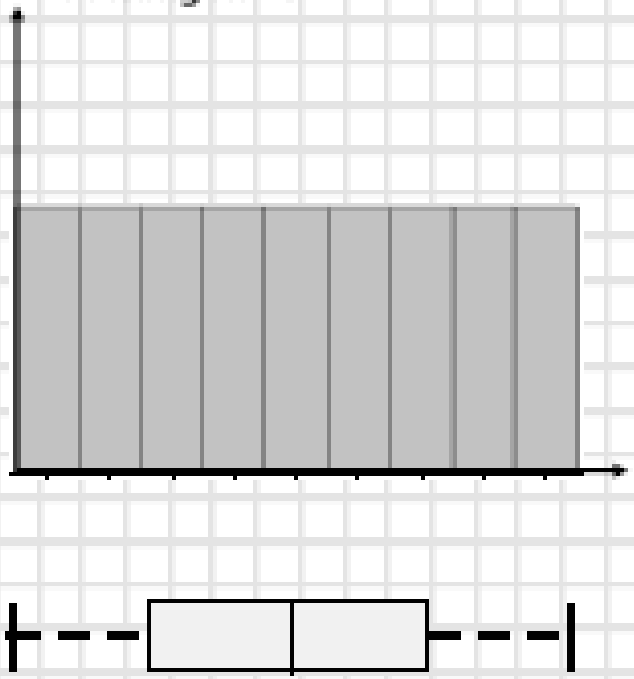
Voto	frequenza
24	1
25	3
26	2
27	1
28	1
29	1
30	1



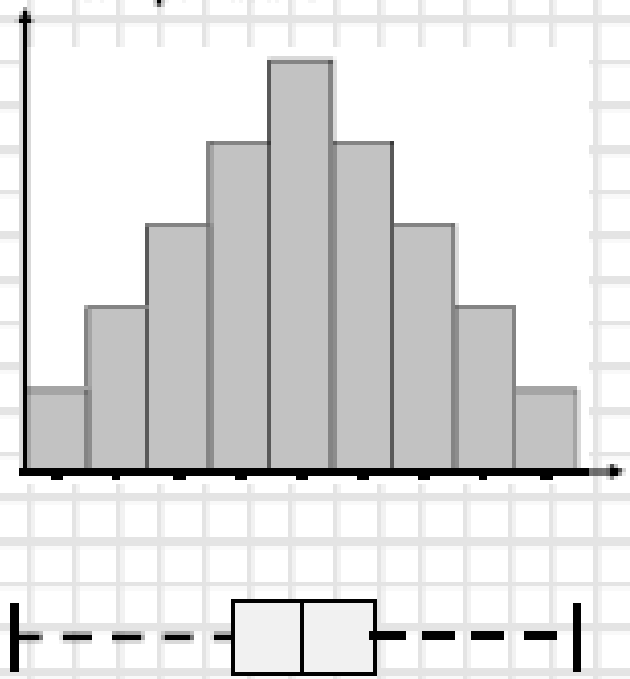
Boxplot

Boxplot e Istogramma

Distribuzione simmetrica rettangolare

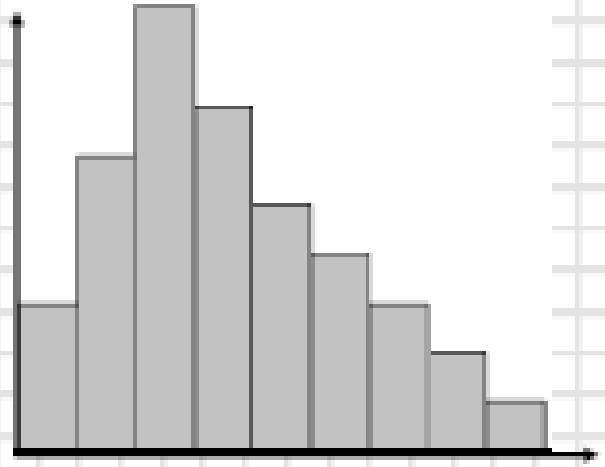


Distribuzione simmetrica campanulare



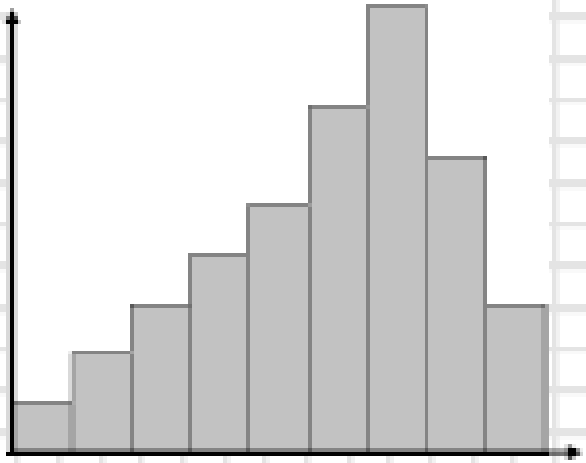
Boxplot e Istogramma

Distribuzione asimmetrica positiva



Me-Q1<Q3-Me

Distribuzione asimmetrica negativa



Me-Q1>Q3-Me

Outlier

Per individuare eventuali outlier la regola più comune è calcolare:

$$1{,}5 \times \Delta$$

dove Δ è la differenza interquartile

$$\Delta = Q_3 - Q_1$$

Un punto viene considerato un possibile outlier se si trova a più di $1.5 \times \Delta$ al di sotto del primo quartile o al di sopra del terzo quartile.

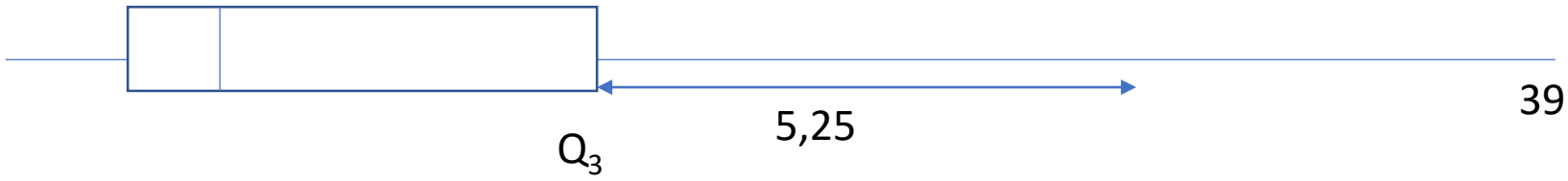
Esempio: età di un campione di 16 partecipanti a un concorso pubblico

età	freq.
24	1
25	5
26	3
27	1
28	1
29	1
30	1
33	1
35	1
39	1

- med=26
- Q_1=25
- Q_3=(29+30)/2=29,5
- $\Delta = Q_3 - Q_1$=29,5-26=3,5
- 1,5 x Δ =1,5 x 3,5=5,25
- 39-Q_3 =39-29,5=9,5

  9,5> 5,25

39 è un outlier



Indice di asimmetria (skewness)

Per un insieme di dati l'indice di asimmetria (skewness) di Pearson è dato da

$$s_k = \frac{\bar{x} - moda}{s}$$

- Se s_k = 0 si ha simmetria
- Se s_k ≥ 1.00 si ha una asimmetria a destra (code più lunghe a destra)
- Se s_k ≤ -1.00 si ha una asimmetria a sinistra