Statistica per le scienze sociali

Seconda edizione

Enrica Amaturo, Biagio Aragona, Maria Gabriella Grassia, Carlo Natale Lauro, Marina Marino

UTET UNIVERSITÀ

4 Sintetizzare e confrontare le distribuzioni

- I valori di disuguaglianza
- Omogeneità (equilibrio) ed eterogeneità (squilibrio)
- Dispersione
- Variabilità rispetto a un centro
- Altri indici di variabilità
- Rappresentare graficamente la variabilità: il box plot
- La forma di una distribuzione: asimmetria e curtosi
- Concentrazione di una variabile trasferibile
- La standardizzazione
- Confronti basati sulle differenze
- I numeri indice

Capitolo a cura di B. Aragona e V. Simonacci

Statistica per le scienze sociali

Seconda edizione

Enrica Amaturo, Biagio Aragona, Maria Gabriella Grassia, Carlo Natale Lauro, Marina Marino

UTET

4 Sintetizzare e confrontare le distribuzioni

- I valori di disuguaglianza
- Omogeneità (equilibrio) ed eterogeneità (squilibrio)
- Dispersione
- Variabilità rispetto a un centro
- Altri indici di variabilità
- **Rappresentare graficamente la variabilità: il box plot**
- **La forma di una distribuzione: asimmetria e curtosi**
- Concentrazione di una variabile trasferibile
- La standardizzazione
- Confronti basati sulle differenze
- I numeri indice

Capitolo a cura di B. Aragona e V. Simonacci

Il grafico per rappresentare la variabilità: il Box-plot

Un modo per rappresentare graficamente la variabilità di una distribuzione è dato dal box-plot

Il box-plot è un grafico caratterizzato da tre elementi:

- una linea o punto, che indicano la posizione valore centrale della distribuzione
- un rettangolo (box) la cui altezza indica la variabilità dei valori "prossimi" alla media
- due segmenti che partono dal rettangolo e i cui estremi sono determinati in base ai valori estremi della distribuzione

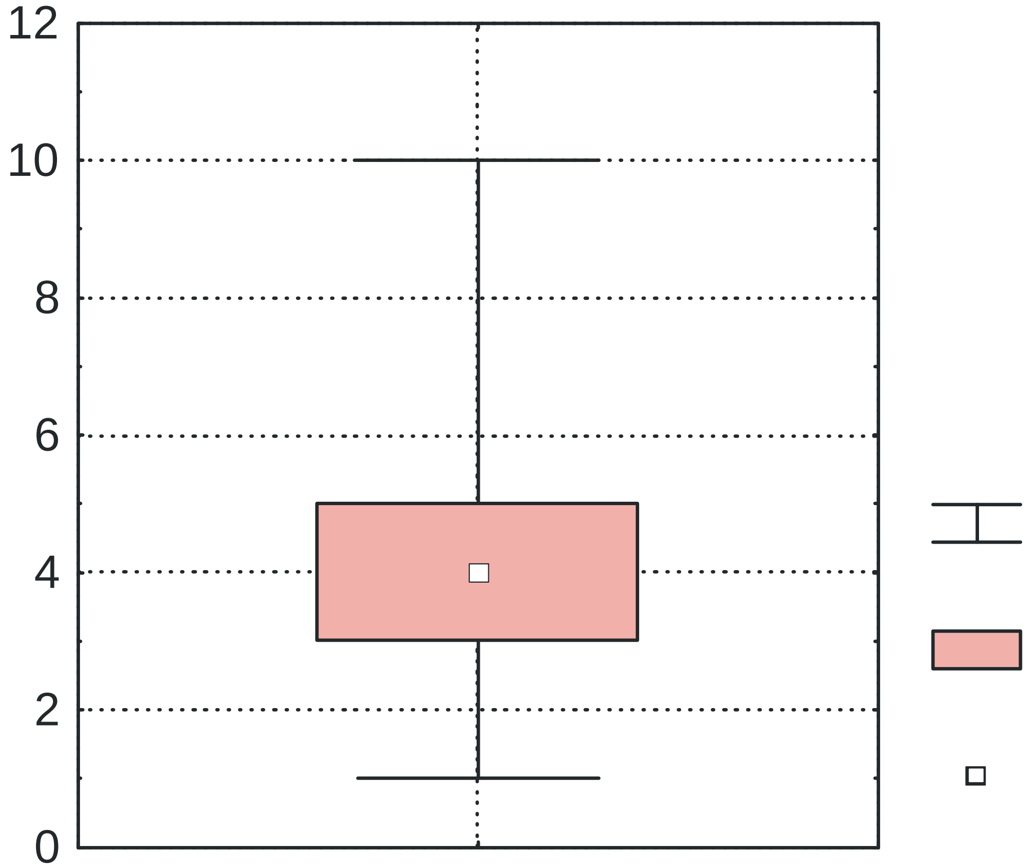
Ad esempio, come valore centrale si può prendere la mediana, come altezza del box la distanza interquartile e come estremi dei segmenti il valore minimo e massimo osservati.

Box-plot Esempio

N° atti aggressivi	1	2	3	4	5	6	7	8	9	10
frequenza	3	8	30	45	22	12	10	5	2	1
Frequenza cumulata	3	11	41	86	108	120	130	135	137	138

N=138

posizione mediana=138/2=69 e 70



Max = 10
Min = 1

Q3=5
Q1=3

Valore mediano:
Me=4

Box-plot
Elementi

L - Il valore più grande

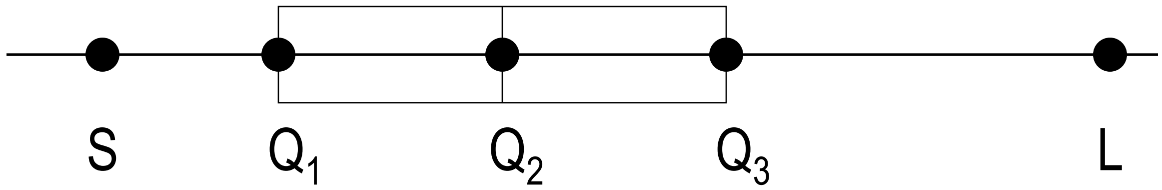
Q_3 - Il terzo quartile

Q_2 – La mediana

Q_1 – Il primo quartile

S - Il valore più piccolo

Un box plot è necessario in presenza di outliers (valori anomali).

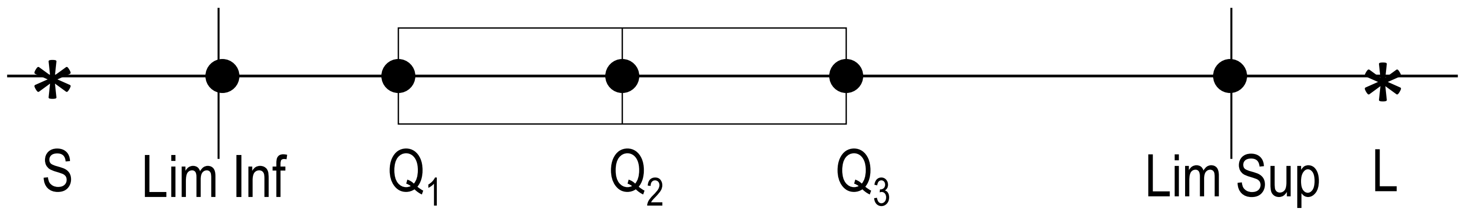


Box-plot
Elementi

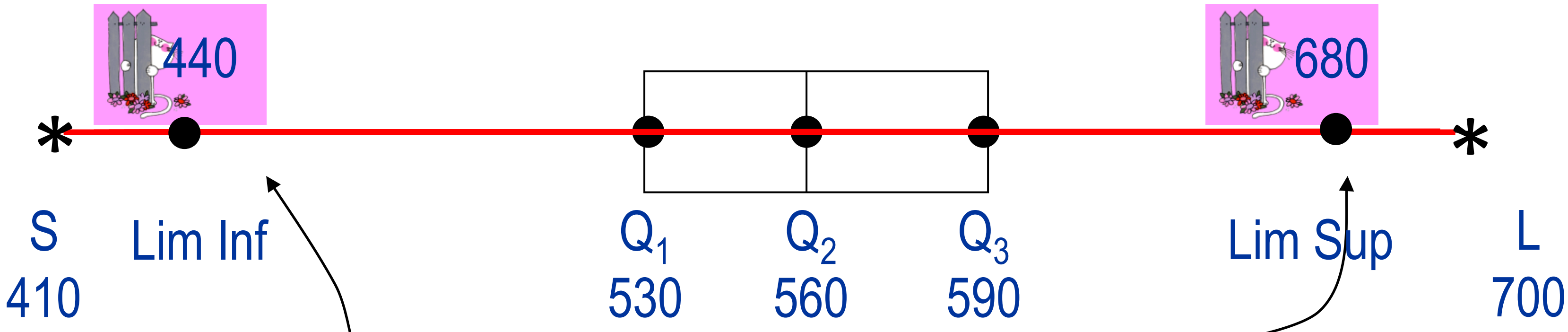
Sono considerati ANOMALI quei valori x_i per i quali si verifica una delle seguenti condizioni:

$$\boxed{X < Q_1 - 1.5\,(Q_3 - Q_1)} \quad \text{o} \quad \boxed{X > Q_3 + 1.5\,(Q_3 - Q_1)}$$

In presenza di valori anomali, i segmenti del rettangolo non si fermano al valore massimo o minimo della distribuzione, ma al valore interno più vicino al limite. I valori anomali, invece, vengono segnalati con un asterisco *



Creare a box plot per i punteggi di intelligenza di 200 allievi



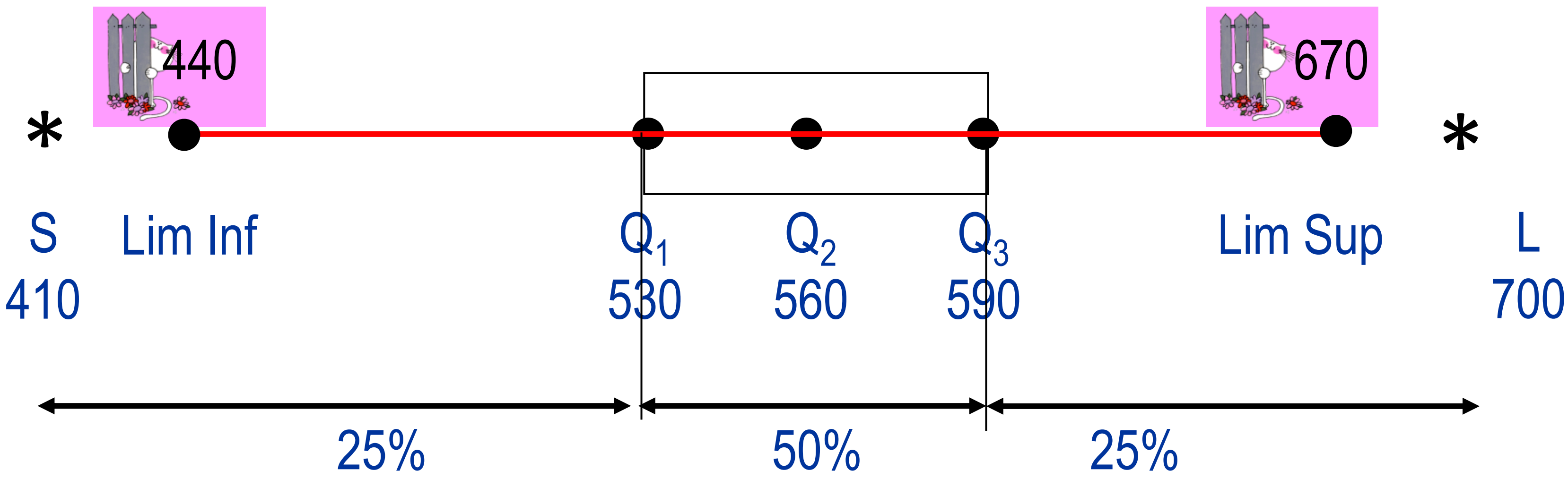
Differenza Interquartile (IQR) = $Q_3 - Q_1$ = 590 - 530 = 60

Intervallo ={Q_1-1.5(IQR), Q_3+1.5(IQR} = {440, 680}

Gli outliers sono 700, e 410.

I baffi (whiskers) saranno tracciati fino ai due valori estremi che non sono outliers (440 e 680)

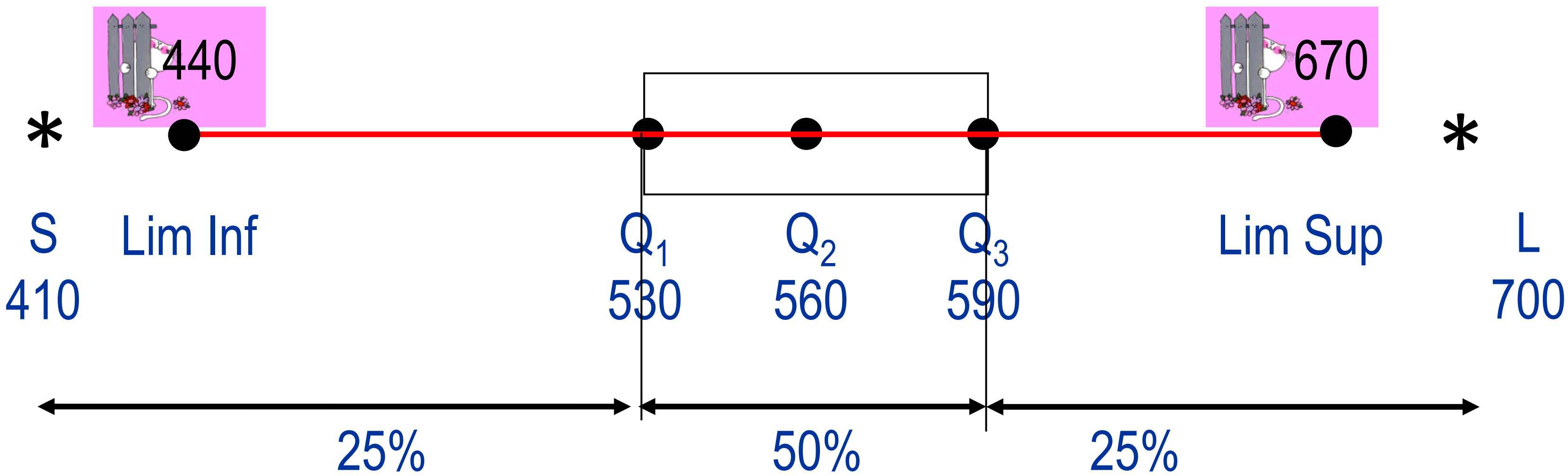
Box-plot Esempio



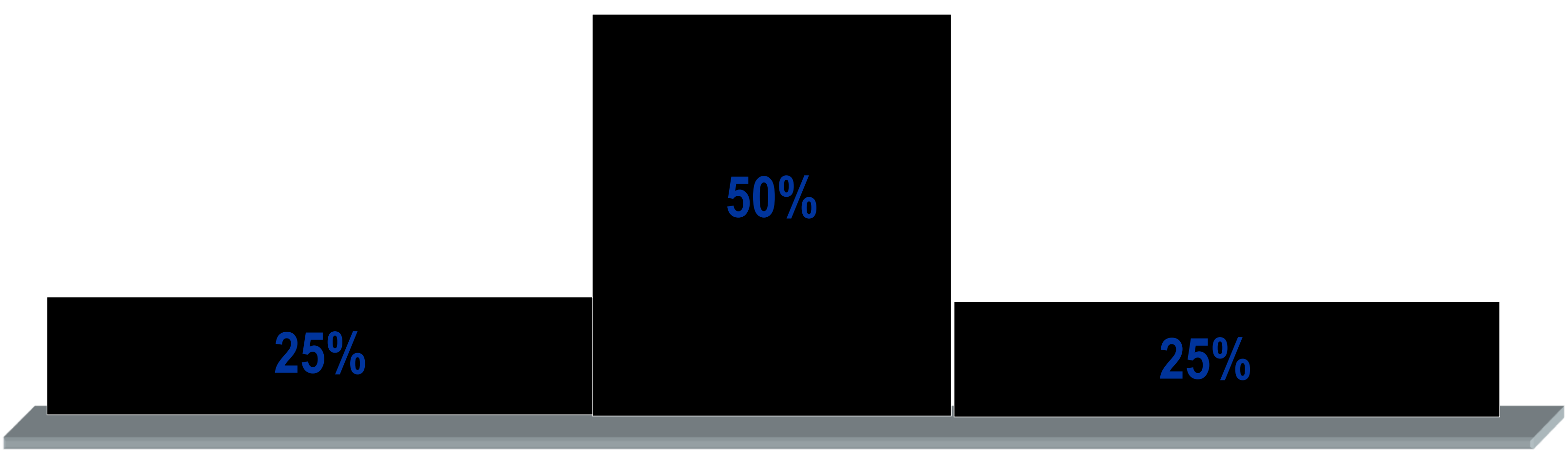
Interpretazione del box plot

- Il range dei punteggi è tra 410 e 700.
- La metà dei valori è inferiore a 560, e l'atra metà è superiore a 560.
- La metà dei valori cade nell'intervallo tra 530 e 590.
- ¼ cade prima di 530 ed ¼ dopo 590.

Box-plot
Esempio



La distribuzione è perfettamente simmetrica



Costruzione di un Box-plot
Esempio

Età degli studenti ad un appello di esame del Corso di Statistica:

18 29 20 23 19 25 26 61 28 30 27 19

Costruzione di un Box-plot
Esempio

Età degli studenti ad un appello di esame del Corso di Statistica:

18 29 20 23 19 25 26 61 28 30 27 19

Si ordinano le modalità

↓

18 19 19 20 23 25 26 27 28 29 30 61

Costruzione di un Box-plot
Esempio

Età degli studenti ad un appello di esame del Corso di Statistica:

18 29 20 23 19 25 26 61 28 30 27 19

Si ordinano le modalità

18 19 19 20 23 25 26 27 28 29 30 61

Q_1 $Q_2 = Me$ Q_3

Costruzione di un Box-plot
Esempio

Età degli studenti ad un appello di esame del Corso di Statistica:

18 29 20 23 19 25 26 61 28 30 27 19

Si ordinano le modalità

18 19 19 20 23 25 26 27 28 29 30 61

Q_1 Q_2=Me Q_3

$Q_1 = 19{,}5$

$Q_3 = 28{,}5$

Costruzione di un Box-plot
Esempio

Età degli studenti ad un appello di esame del Corso di Statistica:

18 29 20 23 19 25 26 61 28 30 27 19

Si ordinano le modalità

18 19 19 20 23 25 26 27 28 29 30 61

Q_1 $\quad$ Q_2=Me $\quad$ Q_3

$Q_1 = 19{,}5$

$Q_3 = 28{,}5$

$IQR = Q_3 - Q_1 = 28{,}5 - 19{,}5 = 9$

Dati anomali:

a) $Q_3 + 1{,}5\,(Q_3 - Q_1) =$

b) $Q_1 - 1{,}5\,(Q_3 - Q_1) =$

Costruzione di un Box-plot
Esempio

Età degli studenti ad un appello di esame del Corso di Statistica:

18 29 20 23 19 25 26 61 28 30 27 19

Si ordinano le modalità

18 19 19 20 23 25 26 27 28 29 30 61

Q_1 Q_2=Me Q_3

$Q_1 = 19{,}5$

$Q_3 = 28{,}5$

$IQR = Q_3 - Q_1 = 28{,}5 - 19{,}5 = 9$

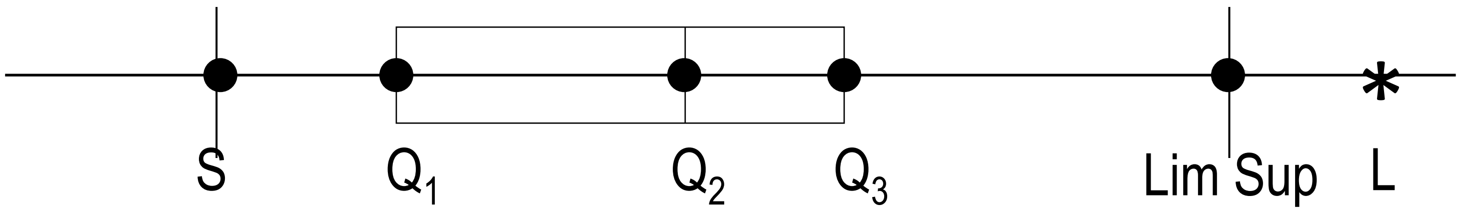
Dati anomali:

a) $Q_3 + 1{,}5\,(Q_3 - Q_1) = 28{,}5 + 1{,}5 \times 9 = 42$

b) $Q_1 - 1{,}5\,(Q_3 - Q_1) = 19{,}5 - 1{,}5 \times 9 = 6$

Costruzione di un Box-plot
Esempio

18 19 19 20 23 25 26 27 28 29 30 61



18 19,5 25,5 28,5 42 61

Dati anomali:

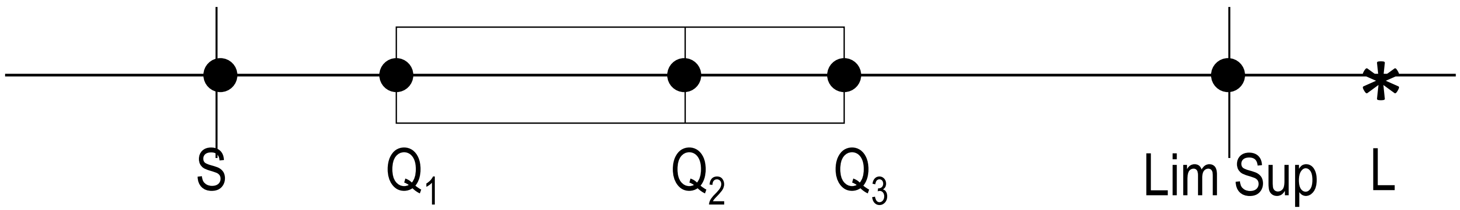
a) $Q_3 + 1{,}5\,(Q_3 - Q_1) = 28{,}5 + 1{,}5 \times 9 = 42$

b) $Q_1 - 1{,}5\,(Q_3 - Q_1) = 19{,}5 - 1{,}5 \times 9 = 6$

$IQR = Q_3 - Q_1 = 28{,}5 - 19{,}5 = 9$

Costruzione di un Box-plot Esempio

18 19 19 20 23 25 26 27 28 29 30 61 ← valore anomalo



18 19,5 25,5 28,5 42 61

Dati anomali:

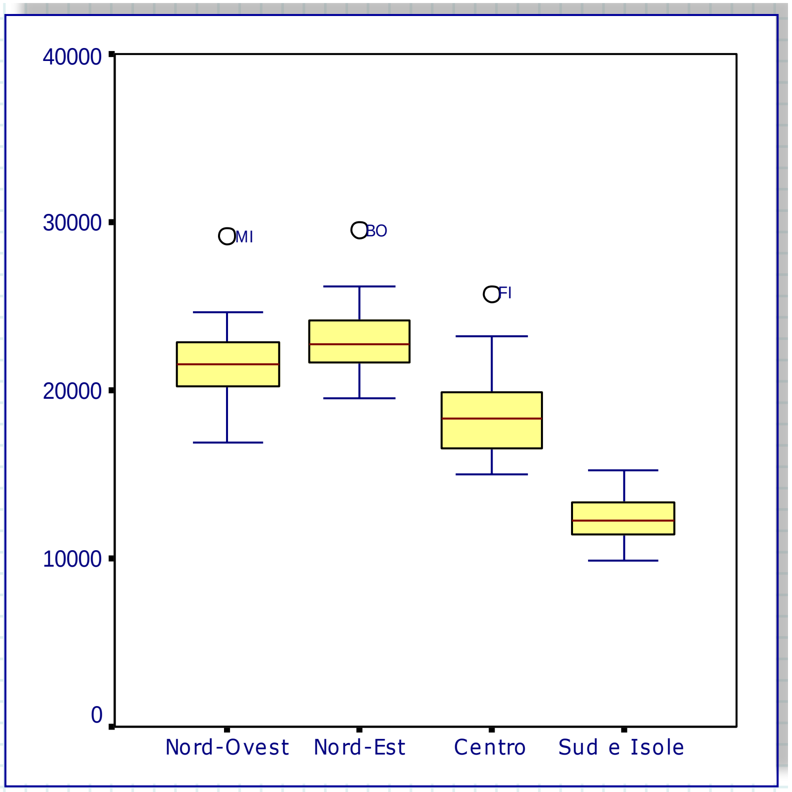
a) $Q_3 + 1{,}5\,(Q_3 - Q_1) = 28{,}5 + 1{,}5 \times 9 = 42$

b) $Q_1 - 1{,}5\,(Q_3 - Q_1) = 19{,}5 - 1{,}5 \times 9 = 6$

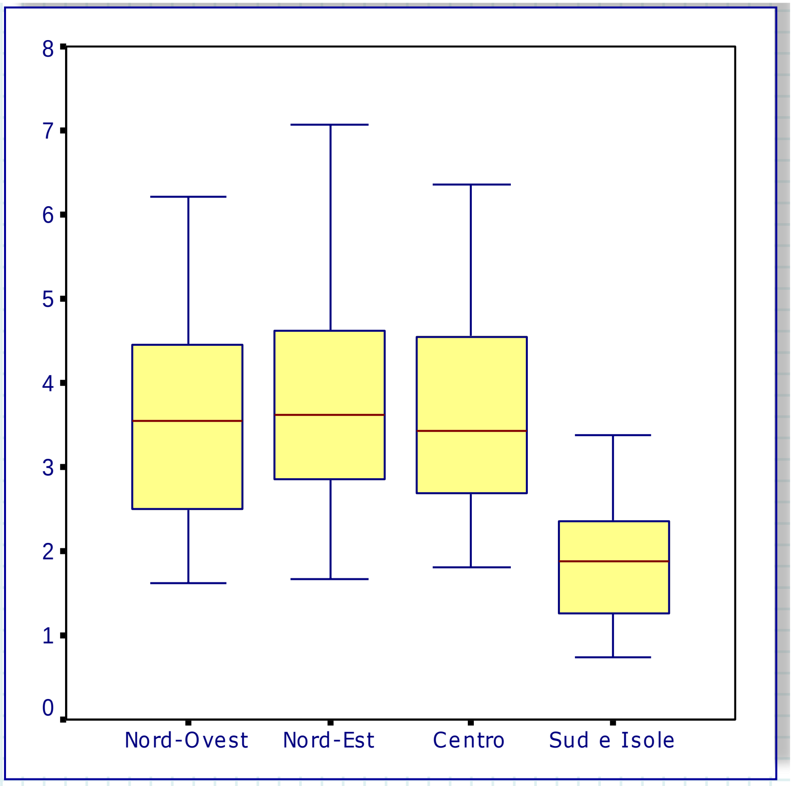
$IQR = Q_3 - Q_1 = 28{,}5 - 19{,}5 = 9$

Box-plot multipli

Reddito p.c.
(in €)

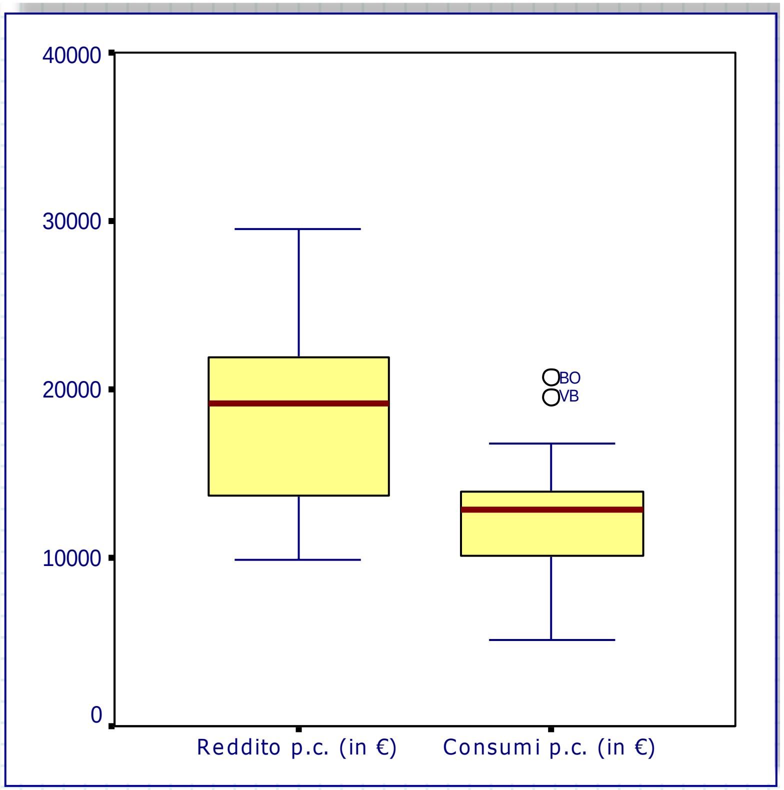


Num. sale cinematografiche
(per 100mila ab.)

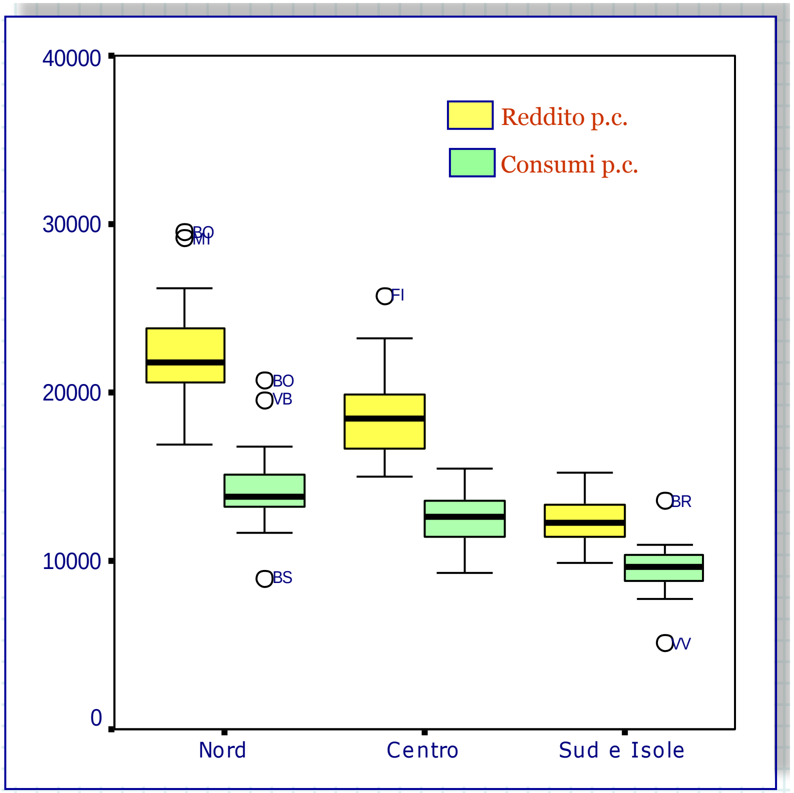


Box-plot multipli

Reddito p.c. e consumi p.c.
(in €)



Reddito p.c. e consumi p.c.
(in €, per zona geografica)



Forma della distribuzione: Asimmetria e Curtosi

Malgrado i centri e gli indici di variabilità ci dicano molto di come si distribuisce una variabile, non esauriscono l'insieme di informazioni che i dati possono racchiudere

Due variabili possono avere la stessa variabilità e lo stesso centro, ma essere molto diverse per il comportamento delle code della distribuzione

Esistono principalmente due misure che studiano qual è la forma di una distribuzione: **l'asimmetria e la curtosi**.

Forma della distribuzione: Asimmetria

Una distribuzione è asimmetrica quando **non** è possibile individuare un asse verticale che suddivida la distribuzione in <u>due parti specularmente uguali</u>

La nozione di asimmetria ha senso solo quindi solo se il carattere è almeno ordinabile

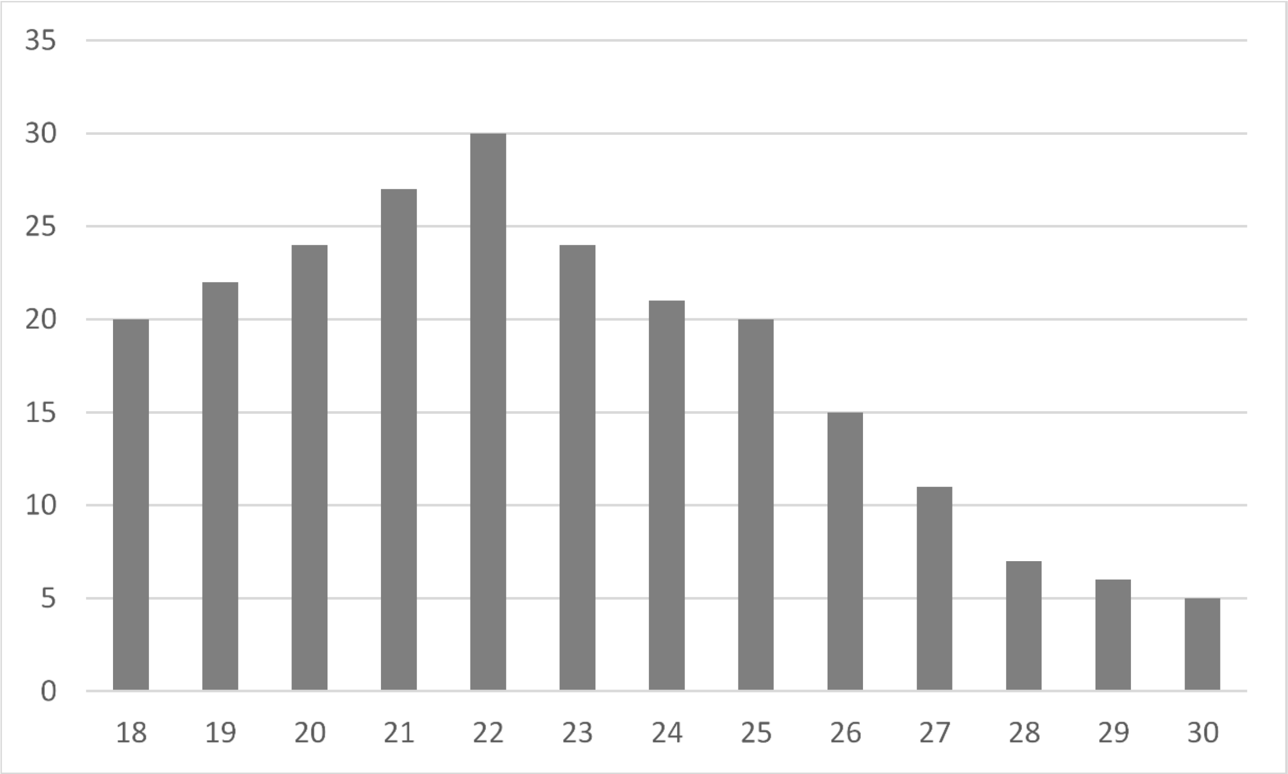
Essa si misura confrontando gli indici di posizione più comuni: moda, mediana e media

Se la media supera la mediana significa che la distribuzione presenta valori più alti verso il semiasse positivo delle x cioè verso destra. In questo caso si avrà **asimmetria positiva (a destra)**

Viceversa, se la media è inferiore della mediana avremo una coda verso sinistra e quindi **asimmetria negativa (a sinistra)**

Forma della distribuzione: Asimmetria

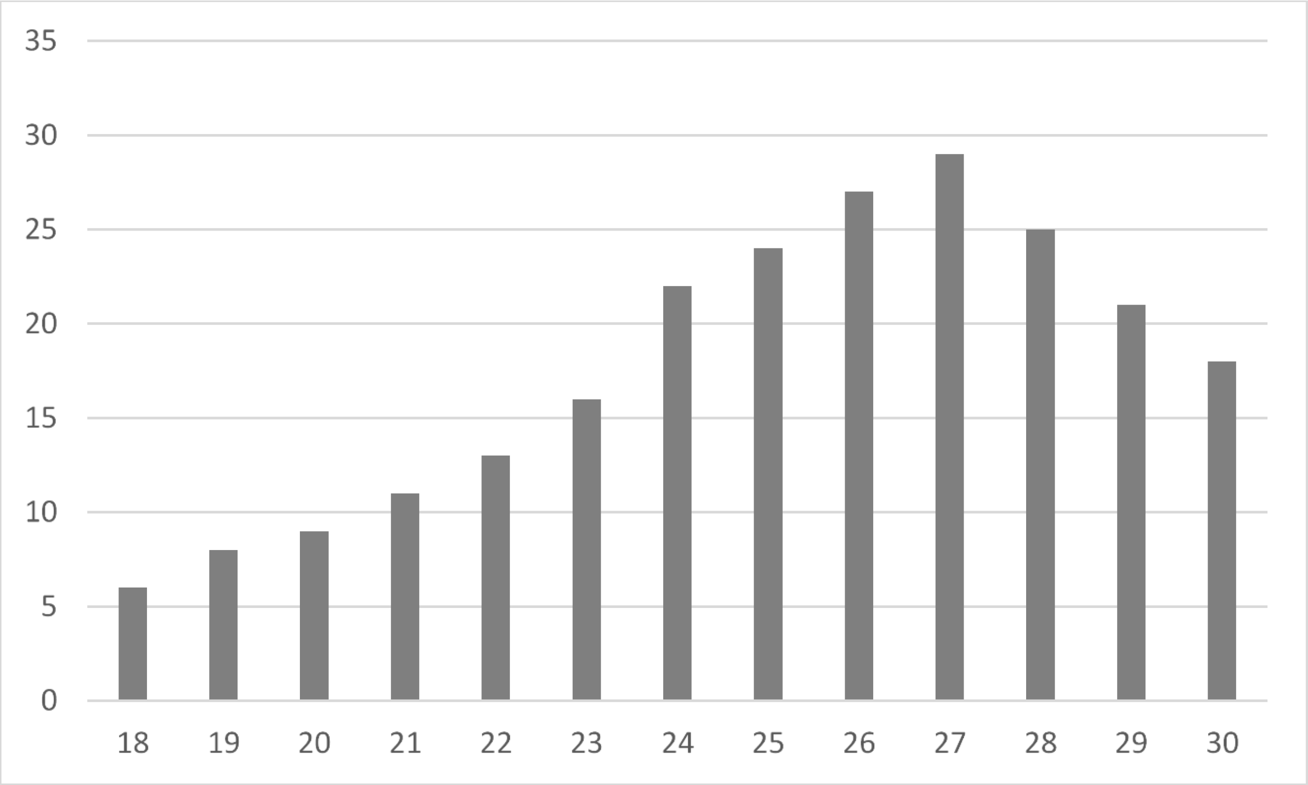
Se la media supera la mediana significa che la distribuzione presenta valori più alti verso il semiasse positivo delle x cioè verso destra. In questo caso si avrà **asimmetria positiva (a destra)**



Asimmetria positiva

Forma della distribuzione: Asimmetria

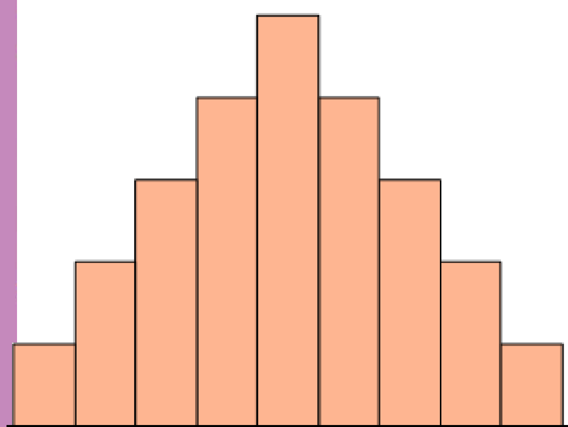
Se la media è inferiore della mediana avremo una coda verso sinistra e quindi **asimmetria negativa (a sinistra)**



Asimmetria negativa

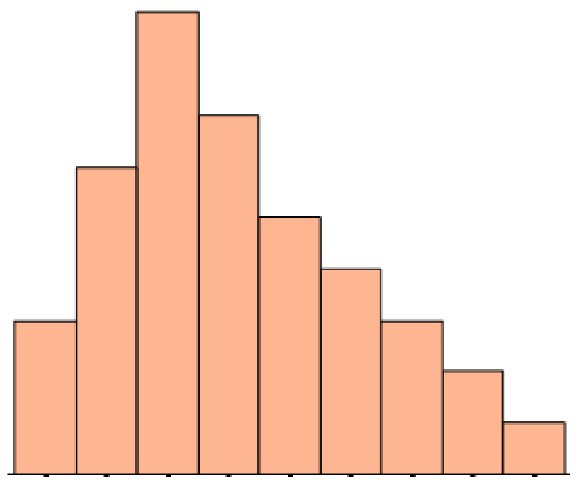
Forma della distribuzione: Asimmetria

Distribuzione simmetrica campanulare



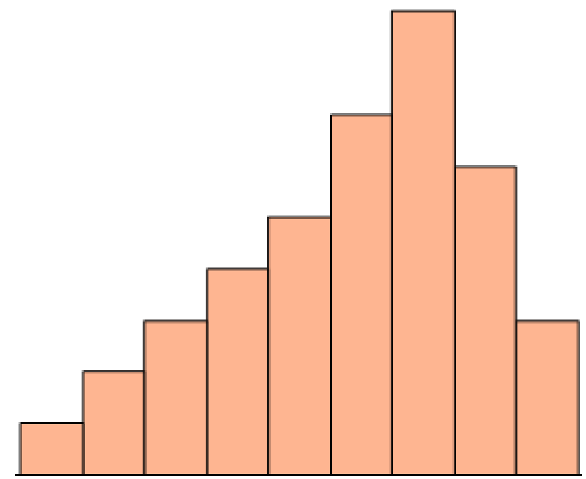
Moda=Mediana=Media

Distribuzione asimmetrica positiva



Moda<Mediana<Media

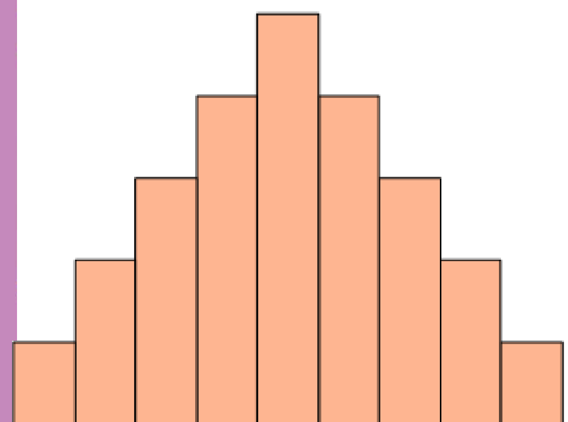
Distribuzione asimmetrica negativa



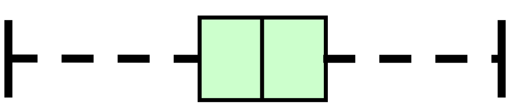
Moda>Mediana>Media

Forma della distribuzione: Istogramma e Box-plot

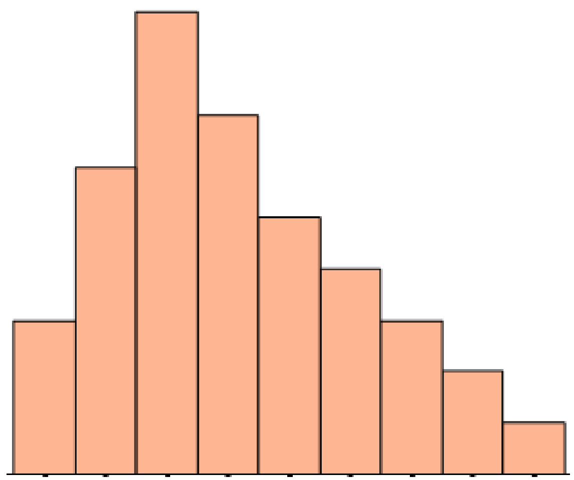
Distribuzione simmetrica campanulare



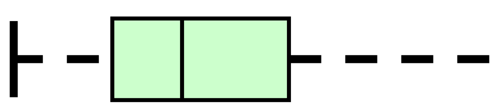
Moda=Mediana=Media



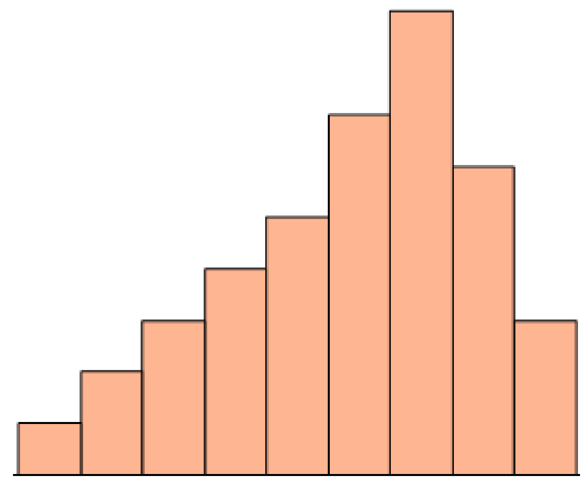
Distribuzione asimmetrica positiva



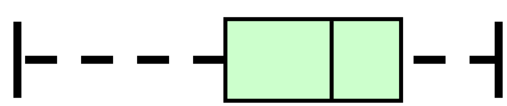
Moda<Mediana<Media



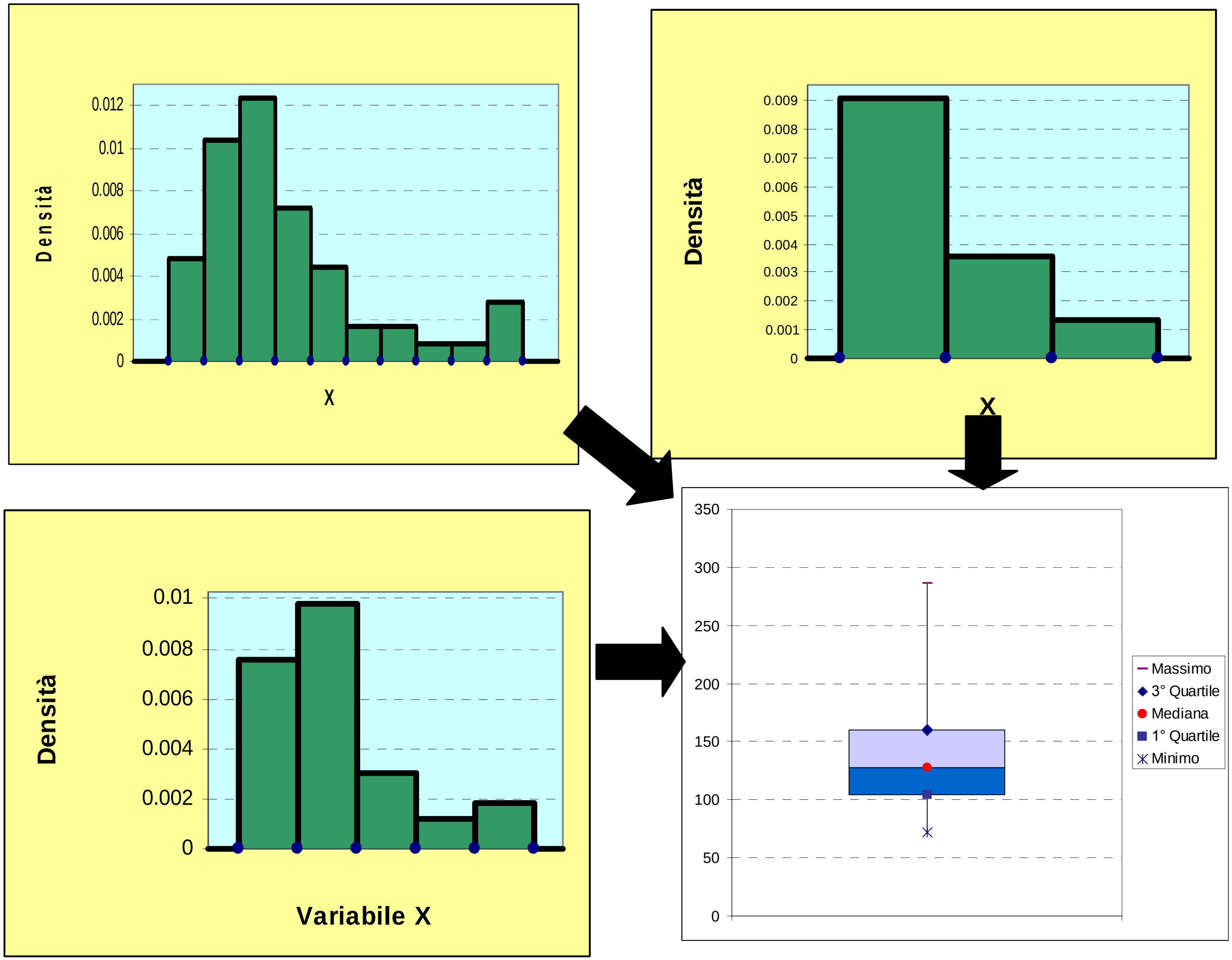
Distribuzione asimmetrica negativa



Moda>Mediana>Media

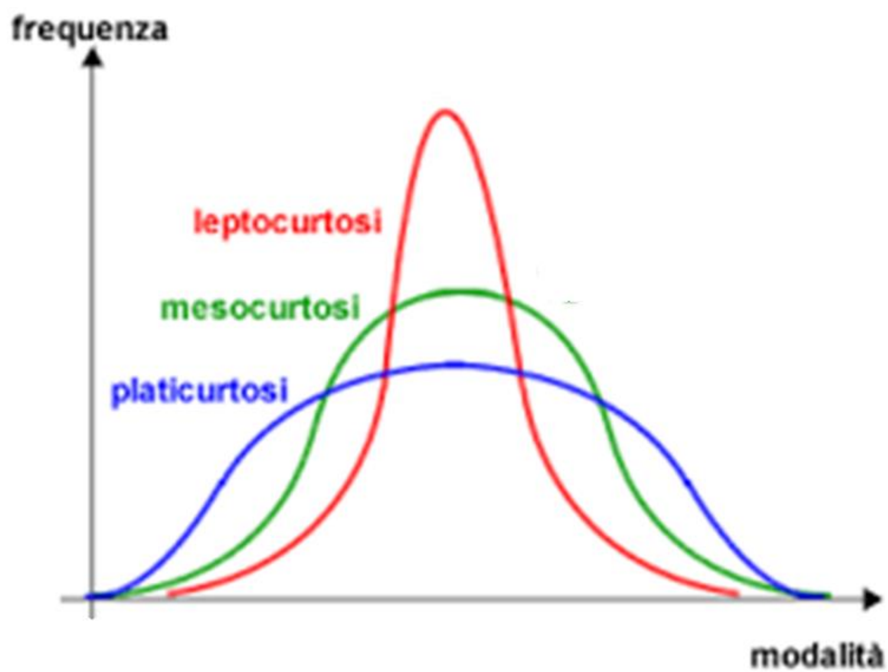


Forma della distribuzione: Istogramma e Box-plot



Forma della distribuzione: Curtosi

Un altro aspetto della forma di una distribuzione di frequenza riguarda il suo appiattimento e quindi il peso più o meno accentuato delle code rispetto alla parte centrale della distribuzione. Il rapporto tra la parte centrale della distribuzione e le code si definisce **curtosi**



Esempio
Individuare i valori anomali

Età	n_i
25	3
26	6
27	8
28	5
29	5
30	3
61	1

Età studenti ad un appello dell'esame di Statistica

Esempio
Individuare i valori anomali

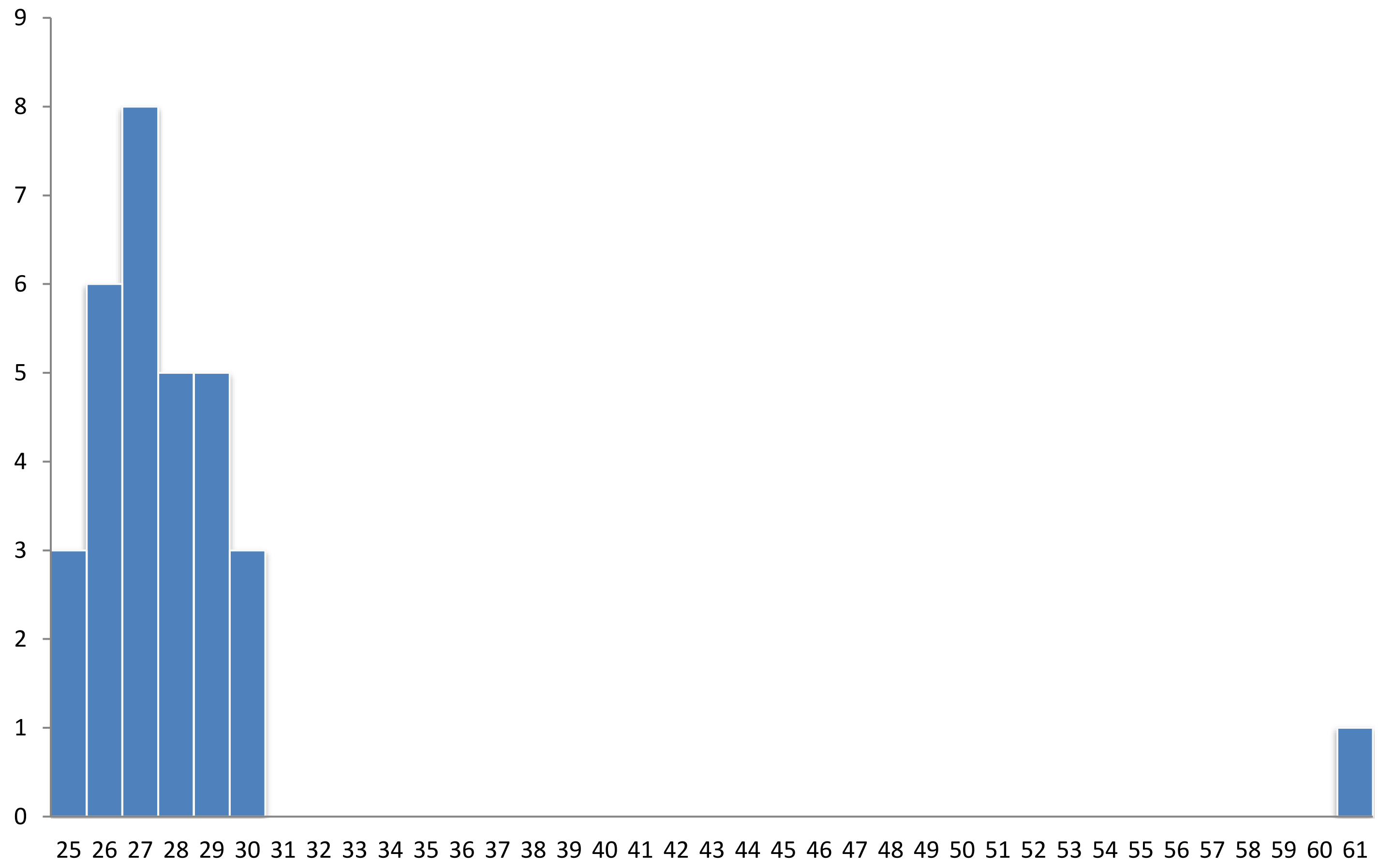
Età	n_i
25	3
26	6
27	8
28	5
29	5
30	3
61	1

Età studenti ad un appello dell'esame di Statistica

Esempio
Individuare i valori anomali

Età	n_i
25	3
26	6
27	8
28	5
29	5
30	3
61	1

Età studenti ad un appello dell'esame di Statistica



Esempio
Individuare i valori anomali

Età	n_i	N_i	
25	3	3	
26	6	9	Q_1
27	8	17	Me
28	5	22	
29	5	27	Q_3
30	3	30	
61	1	31	

Mediana e Quartili

$$\overbrace{25\ 25\ 25}^{3}\quad \overbrace{26\ 26\ 26\ 26\ 26\ 26}^{6}\quad \overbrace{27\ 27\ 27\ 27\ 27\ 27\ 27\ 27}^{8}\quad \overbrace{28\ 28\ 28\ 28\ 28}^{5}\quad \overbrace{29\ 29\ 29\ 29\ 29}^{5}\quad \overbrace{30\ 30\ 30}^{3}\quad \overbrace{61}^{1}$$

Q_1 $\quad$ Q_2=Me $\quad$ Q_3

valore anomalo

$Q_1 = 26$

Me= 27

$Q_3 = 29$

$Q_3 - Q_1 = 29-26 = 3$

Dati anomali:

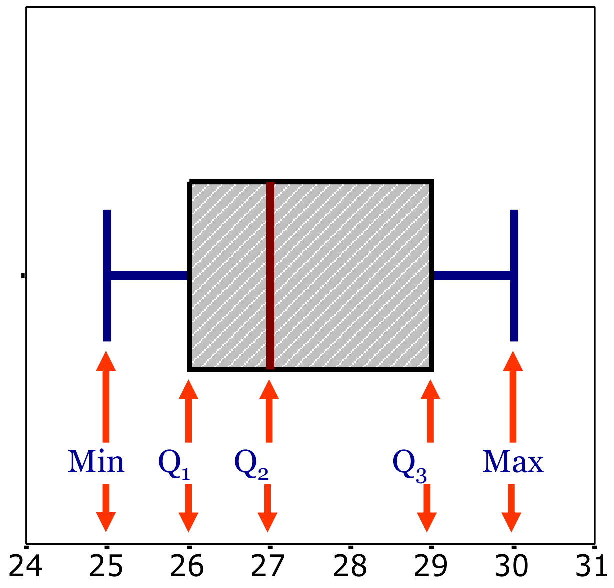
a) $Q_3 + 1{,}5\,(Q_3-Q_1) = 29 + 1{,}5\times 3 = 33{,}5$

b) $Q_1 - 1{,}5\,(Q_3-Q_1) = 26 - 1{,}5\times 3 = 21{,}5$

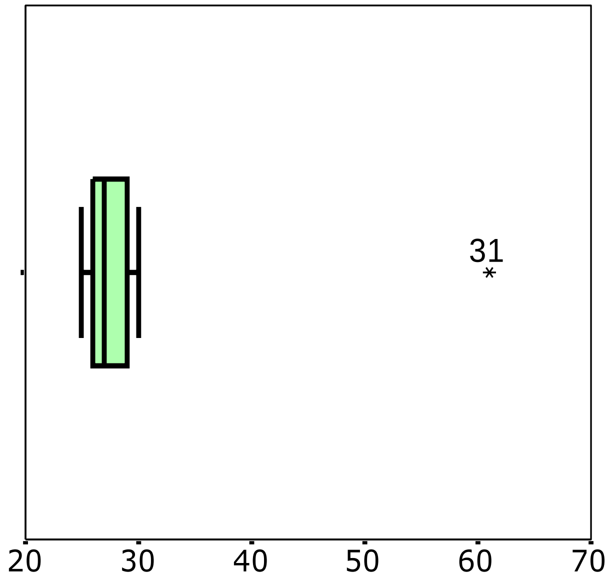
Esempio
Individuare i valori anomali

Età	n_i	N_i	
25	3	3	
26	6	9	Q_1
27	8	17	Me
28	5	22	
29	5	27	Q_3
30	3	30	
61	1	31	

a) Senza il dato anomalo

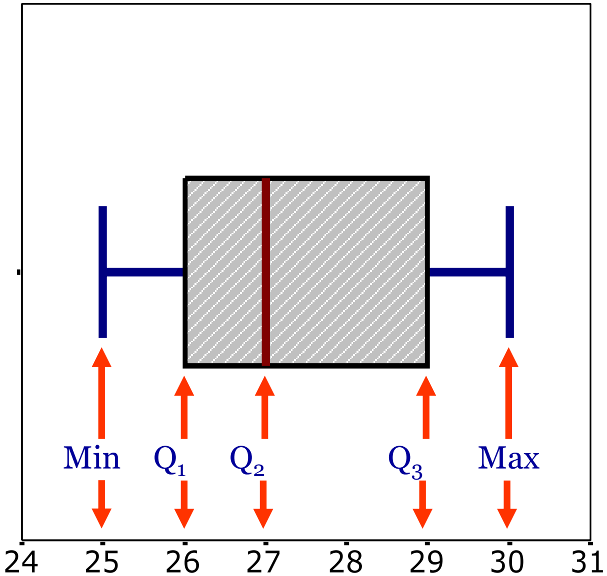


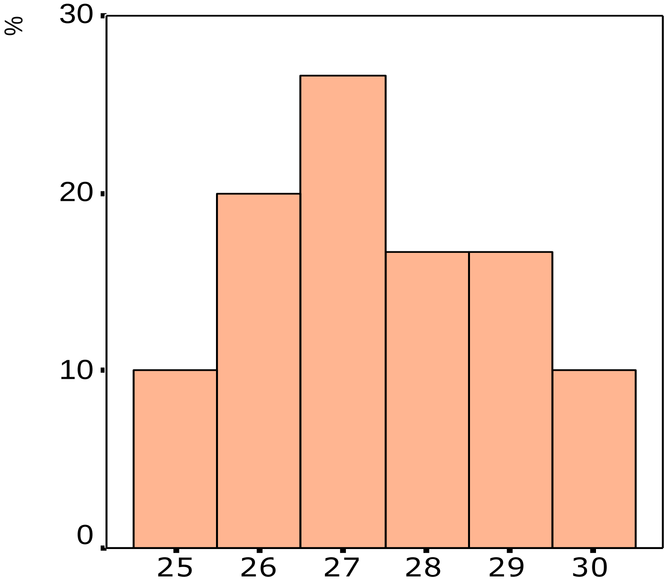
b) Con il dato anomalo

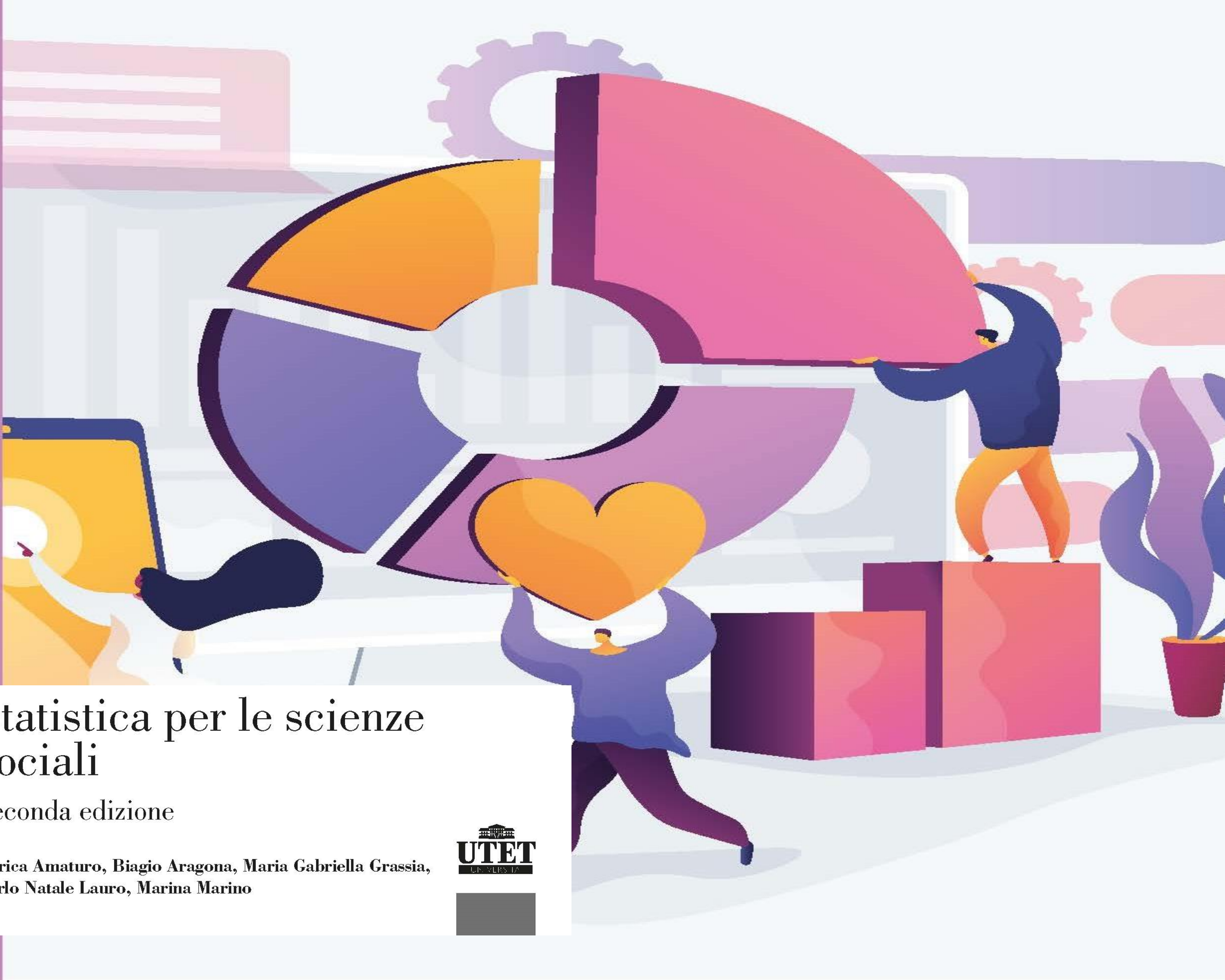


Esempio
Individuare i valori anomali

Età	n_i	N_i
25	3	3
26	6	9
27	8	17
28	5	22
29	5	27
30	3	30
~~61~~	~~1~~	~~31~~







Statistica per le scienze sociali

Seconda edizione

Enrica Amaturo, Biagio Aragona, Maria Gabriella Grassia, Carlo Natale Lauro, Marina Marino

UTET