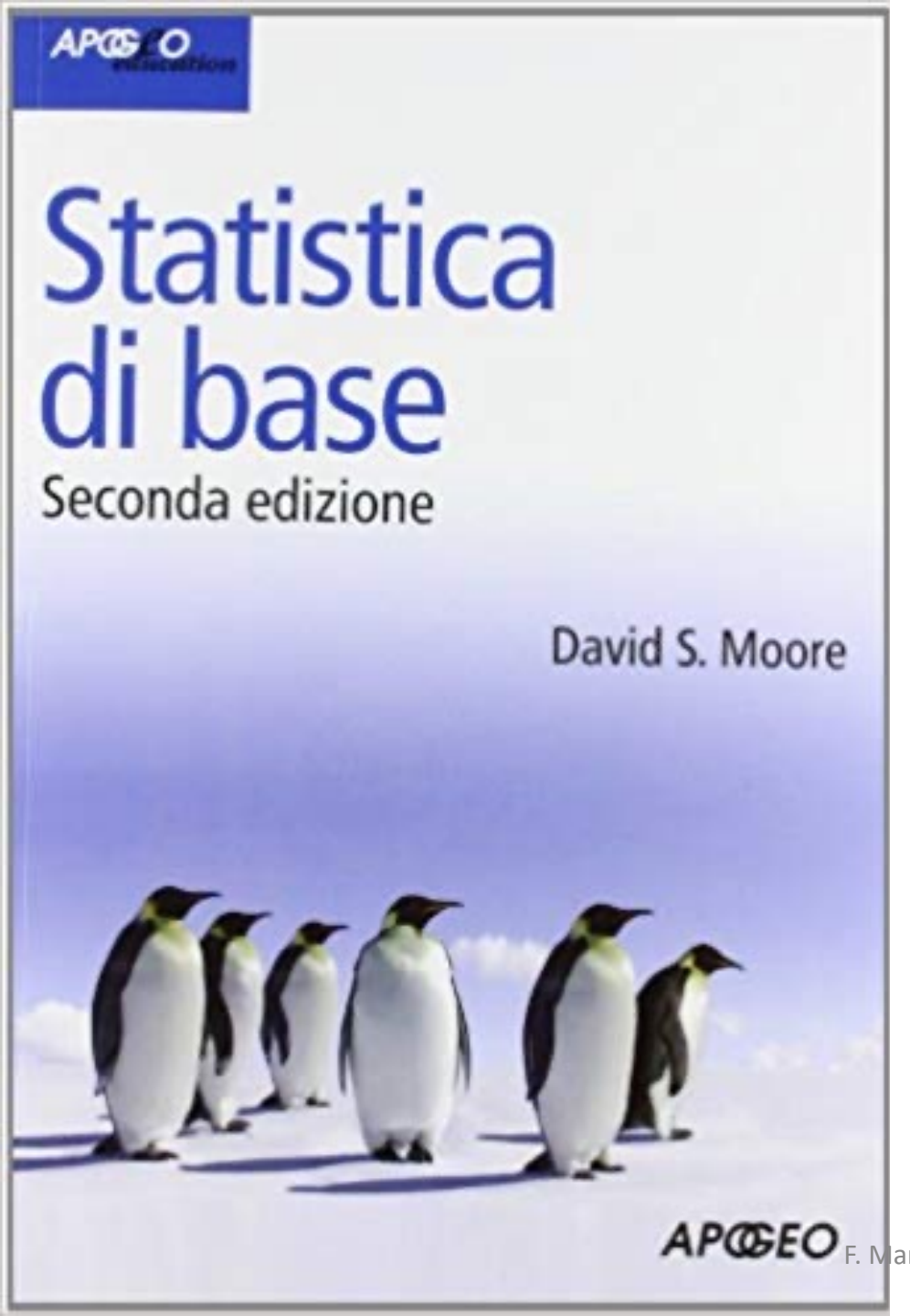
UNIVERSITÀ DEGLI STUDI DI ROMA "FORO ITALICO"

Rappresentazione numerica delle distribuzioni

Corso di Laurea in Scienze Motorie e Sportive (L22)

Insegnamento di Informatica

Prof. Federico Mari



Libro di testo

D.S. Moore. **Statistica di base.** Idee & strumenti. Apogeo, 2013. ISBN: 978-8838786426.

Parte I

Capitolo 2

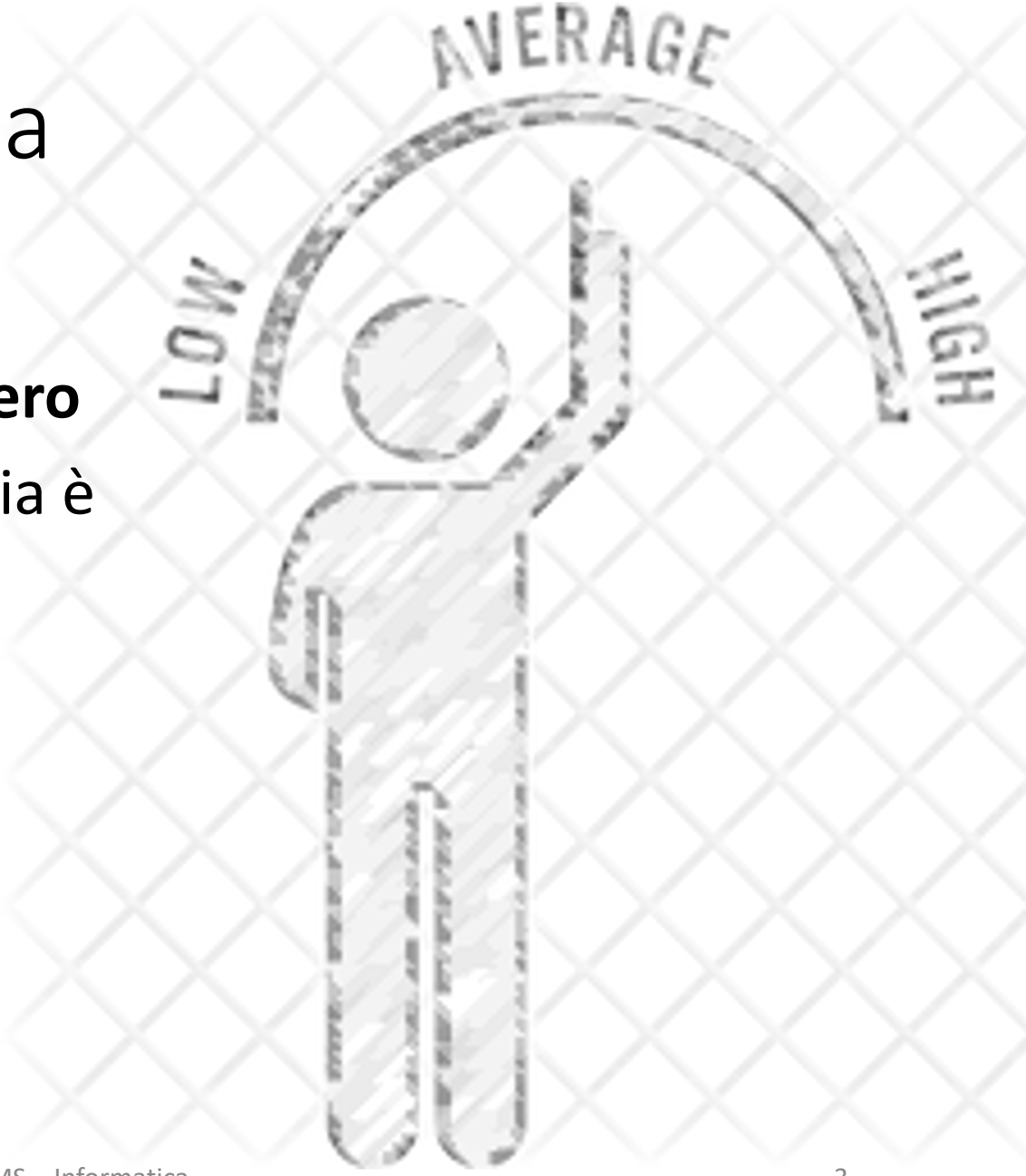
Misurare il centro: la media

- Si devono **sommare i valori** delle osservazioni **e dividere per il loro numero**
- Con n osservazioni $x_1, x_2, \ldots, x_n$ la media è

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

- In modo coinciso

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



Esempio: redditi netti di laureati (BI 2002)

Osservazioni (in migliaia di euro)

14	14	39	7	31	27	14	109	8	19	26	45	40	7	25

0	778
1	4449
2	567
3	19
4	5
5	
6	
7	
8	
9	
10	9

Diagramma ramo-foglia

La **media** è $\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}$

$$= \frac{14 + 14 + \cdots + 25}{15}$$

$$= \frac{425}{15} = 28.333 \text{ (}\mathbf{28\ 333\ euro}\text{)}$$

Outlier

Esempio: redditi netti di laureati (BI 2002)

Osservazioni

14	14	39	7	31	27	14	109	8	19	26	45	40	7	25

Media con le **frequenze assolute**

$$\bar{x} = \frac{2 \cdot 7 + 8 + 3 \cdot 14 + 19 + 25 + 26 + 27 + 31 + 39 + 40 + 45 + 109}{15}$$

Media con le **frequenze relative**

$$\bar{x} = \frac{2}{15}7 + \frac{1}{15}8 + \frac{3}{15}14 + \frac{1}{15}19 + \cdots + \frac{1}{15}109$$

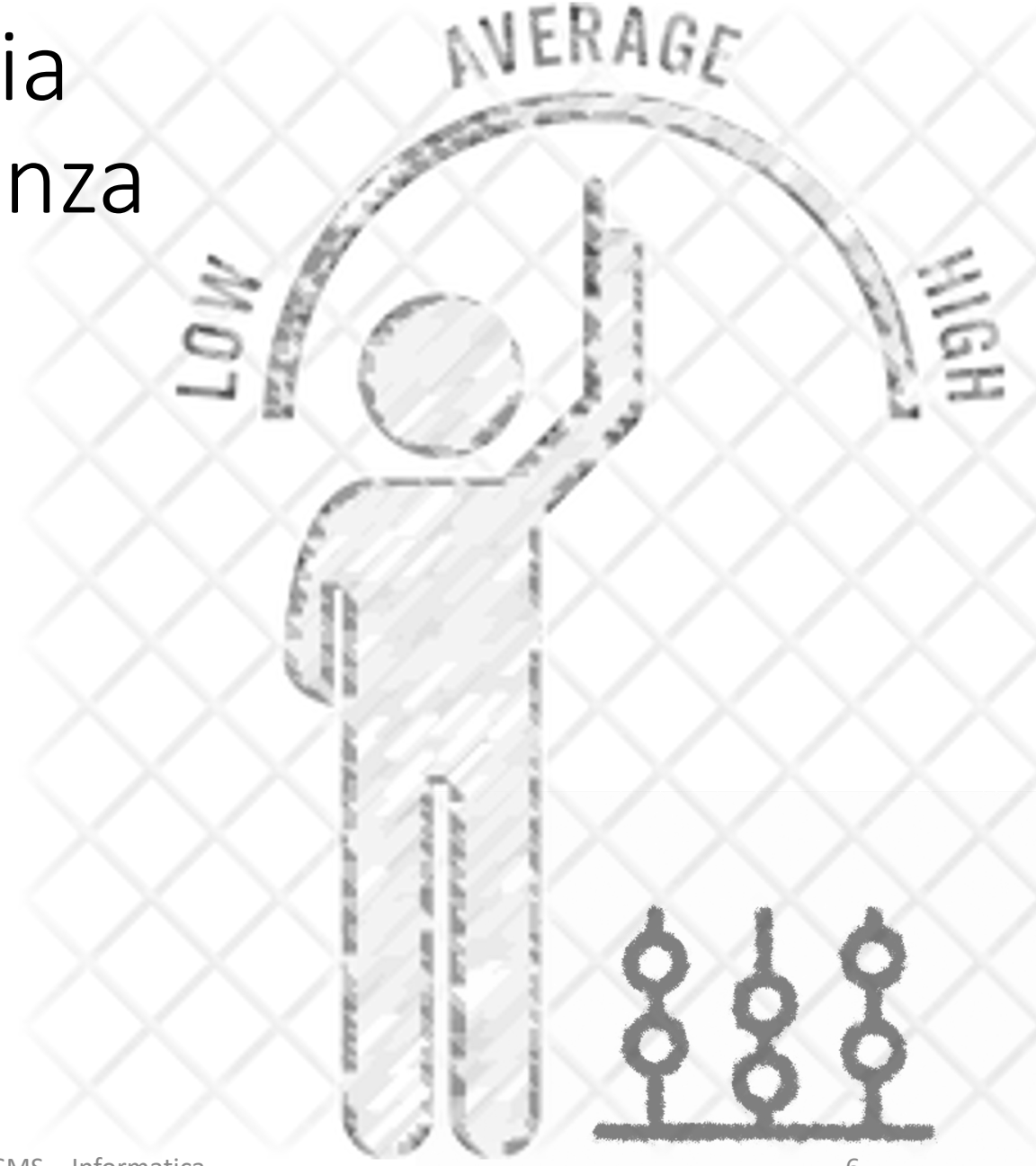
Misurare il centro: la media per distribuzioni di frequenza

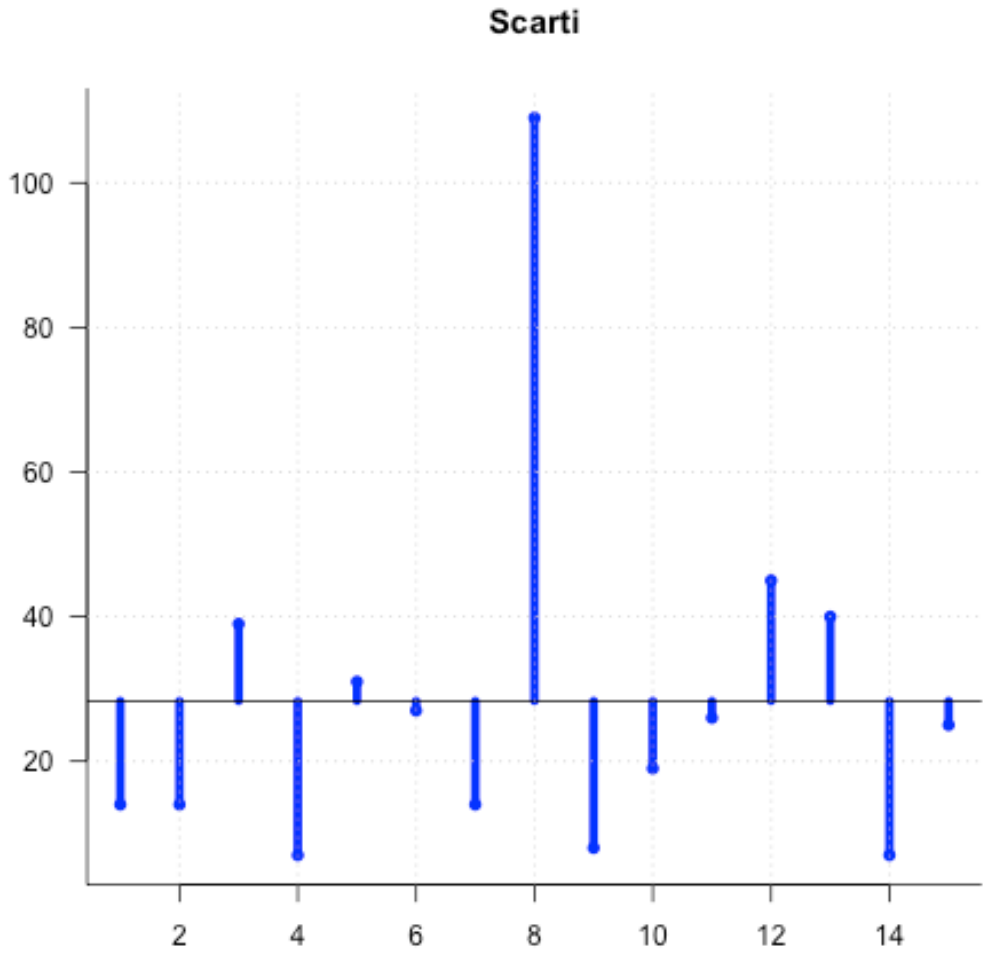
- Frequenze assolute n_i

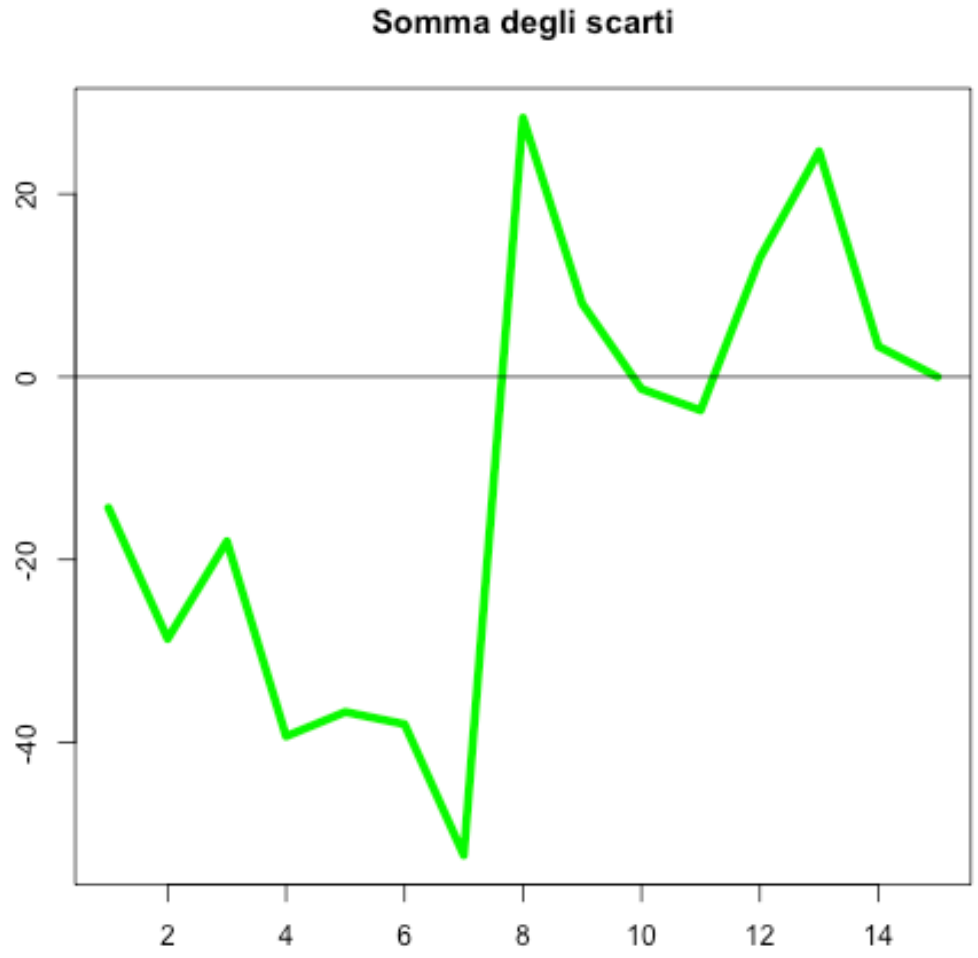
$$\bar{x} = \frac{\sum_{i=1}^{k} n_i x_i}{n}$$

- Frequenze relative $f_i = \frac{n_i}{n}$

$$\bar{x} = \sum_{i=1}^{k} x_i f_i$$







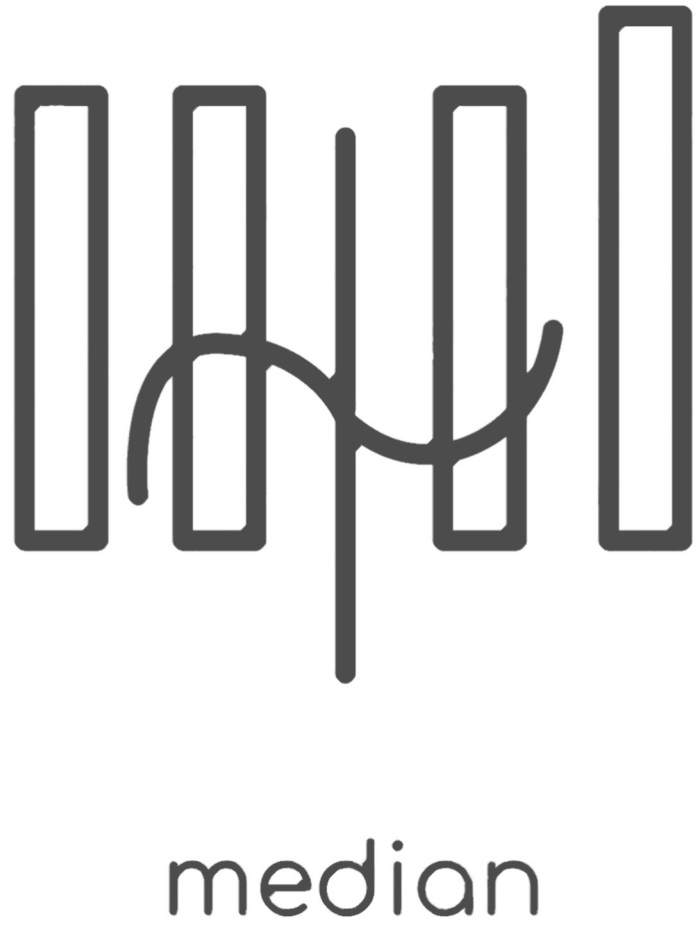
La somma degli scostamenti dei valori osservati dalla media è nulla

$$\sum_{i=1}^{n} (x_i - \overline{x}) = 0$$

La somma degli scostamenti dei valori osservati dalla media è nulla

Misurare il centro: la mediana

- La mediana è M
- Punto centrale di una distribuzione
- Metà osservazioni alla sua sinistra
- Metà osservazioni alla sua destra



Esempio: redditi netti di laureati (BI 2002)

Osservazioni

14	14	39	7	31	27	14	109	8	19	26	45	40	7	25

Ordinare dalla più piccola alla più grande

7	7	8	14	14	14	19	25	26	27	31	39	40	45	109

Se n è dispari allora la posizione di $M = \frac{(n+1)}{2} = \frac{16}{2} = 8$

7	7	8	14	14	14	19	**25**	26	27	31	39	40	45	109

$$M = 25$$

Esempio: redditi netti di laureati (BI 2002)

Numero pari di osservazioni

7	7	8	14	14	14	19	25	26	27	31	39	40	45

Se n è pari allora la mediana è la media delle due osservazioni centrali

7	7	8	14	14	14	**19**	**25**	26	27	31	39	40	45

$$M = \frac{(19 + 25)}{2} = 22$$

Confronto tra media e mediana

7	7	8	14	14	14	19	25	26	27	31	39	40	45

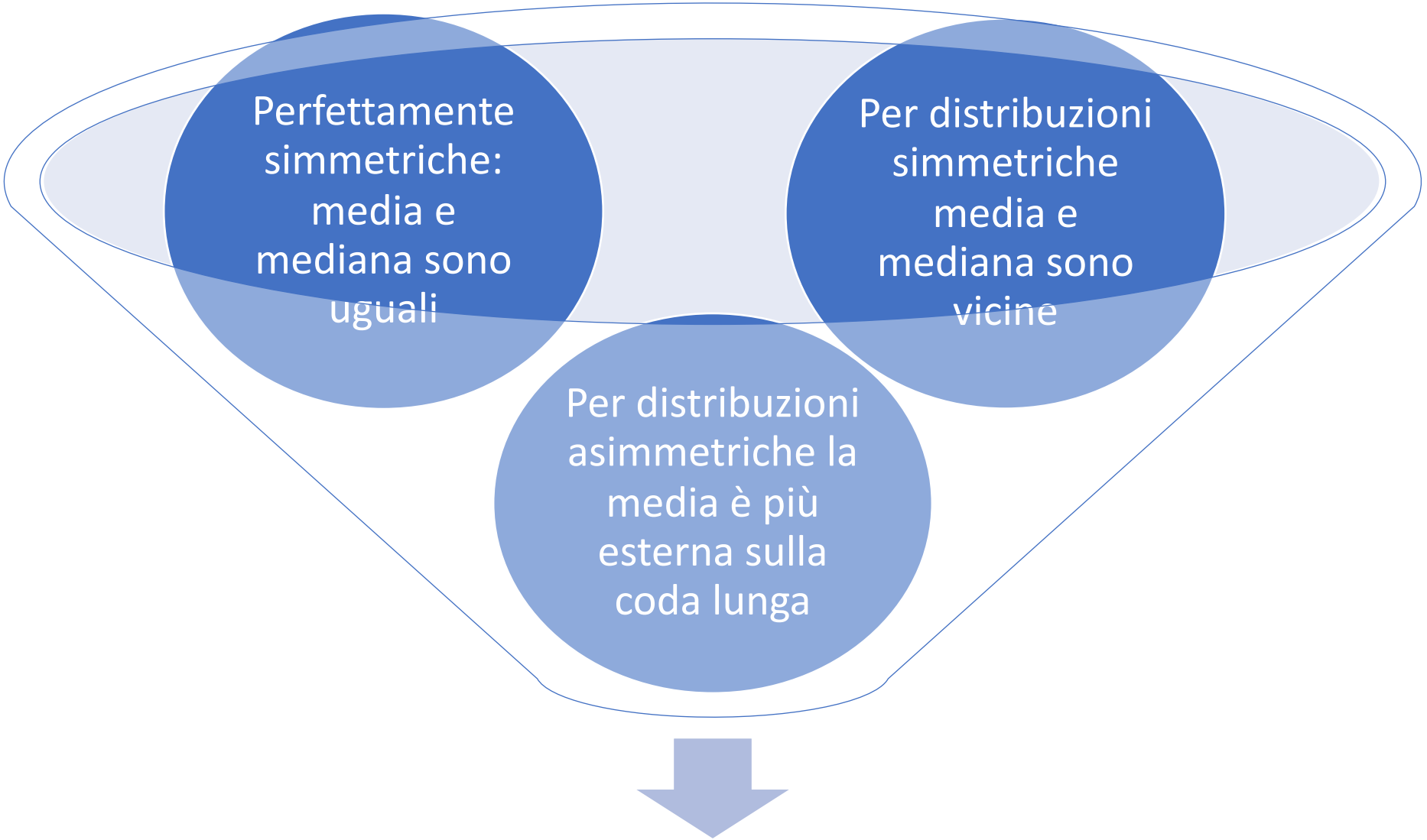
$M = 22$ $\bar{x} = 22.57$

7	7	8	14	14	14	19	25	26	27	31	39	40	45	109

$M = 25$ $\bar{x} = 28.33$

7	7	8	14	14	14	19	25	26	27	31	39	40	45	1090

$M = 25$ $\bar{x} = 93.73$



La mediana è più **robusta** della media

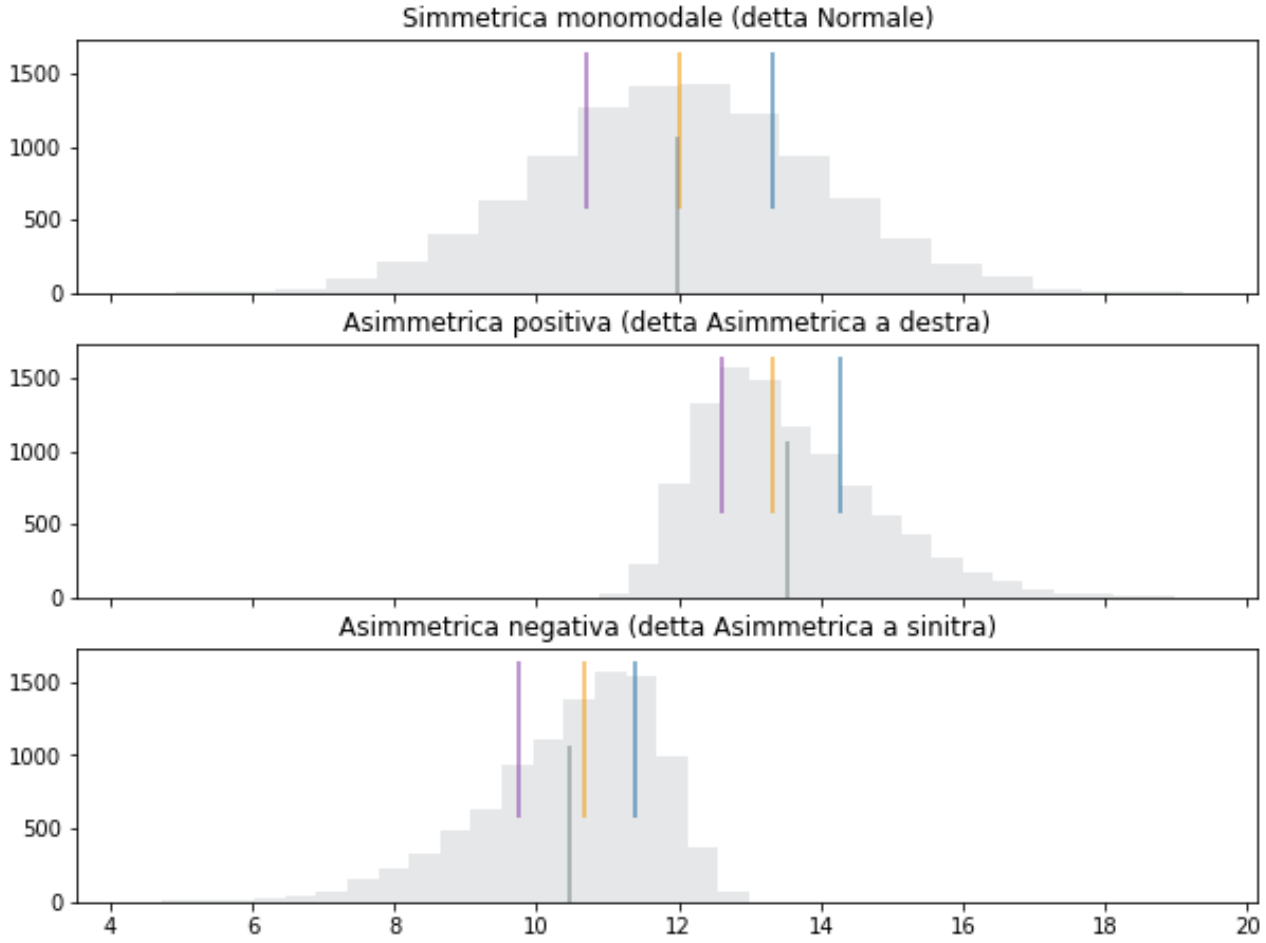
Misurare il centro: la moda

- Il più semplice indicatore del centro della distribuzione
- È la **modalità con maggiore frequenza** (assoluta, relativa o percentuale)
- La moda dei 15 laureati in BI è di 14 (mila euro) con frequenza 3

7	7	8	14	14	14	19	25	26	27	31	39	40	45	109

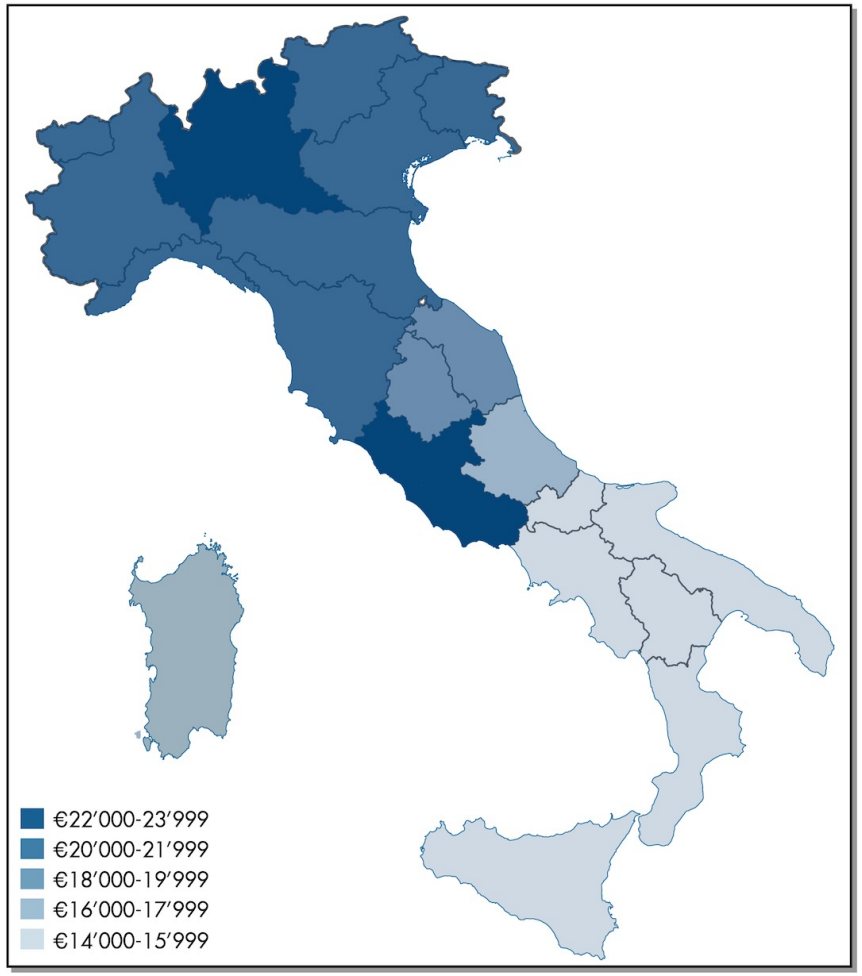
Simmetria delle distribuzioni e indici numerici

- In ogni grafico
  - le linee verticali alte sono Q1, M, Q3
  - la linea verticale bassa è la media
  - La moda è il picco più alto
- *Simmetrica* M = media
  - → Solo per le simmetriche si usa la media
- *Asimmetrica a destra* M < media
- *Asimmetrica a sinistra* M > media



REDDITO ITALIANO MEDIO — REGIONE PER REGIONE

dati Istat 2013 (relativi al 2011)



LOMBARDIA	€ 23'210
LAZIO	€ 22'160
VALLE D'AOSTA	€ 21'260
P.A. BOLZANO	€ 21'200
EMILIA ROMAGNA	€ 21'180
LIGURIA	€ 21'000
PIEMONTE	€ 20'870
P.A. TRENTO	€ 20'300
VENETO	€ 20'270
FRIULI V.G.	€ 20'270
TOSCANA	€ 20'100
MEDIA	€ 19'660
UMBRIA	€ 18'630
MARCHE	€ 18'310
SARDEGNA	€ 16'840
ABRUZZO	€ 16'670
CAMPANIA	€ 16'360
SICILIA	€ 15'600
PUGLIA	€ 15'390
MOLISE	€ 15'200
BASILICATA	€ 14'980
CALABRIA	€ 14'230

Nicolas Lozito

YOU TREND

Misurare la dispersione: i quartili Q_1 e Q_3

- **Media**, moda e **mediana** non raccontano tutta la storia
- Necessaria una misura della dispersione
- **Valore minimo e valore massimo** sono indicativi ma potrebbero essere outlier
- Aumentiamo la rappresentazione della dispersione guardando l'intervallo che copre la **metà intermedia** dei dati contrassegnato dai **quartili**
- Contando nella lista ordinata dalla più piccola osservazione, il **primo quartile Q_1 è a un quarto** della lista e il **terzo quartile Q_3 a tre quarti**

Esempio: redditi netti di laureati (BI 2002)

Osservazioni ordinate dalla più piccola alla più grande

7	7	8	14	14	14	19	25	26	27	31	39	40	45	109

Individuare la mediana $M = 25$

7	7	8	14	14	14	19	**25**	26	27	31	39	40	45	109

Q_1 è la mediana lato sinistro Q_3 è la mediana lato destro

7	7	8	**14**	**14**	14	19	**25**	26	27	**31**	**39**	40	45	109

$Q_1 = 14$ $Q_2 = M$ $Q_3 = 35$

Il sommario a cinque numeri

Minimo	Q_1	M	Q_3	**Massimo**

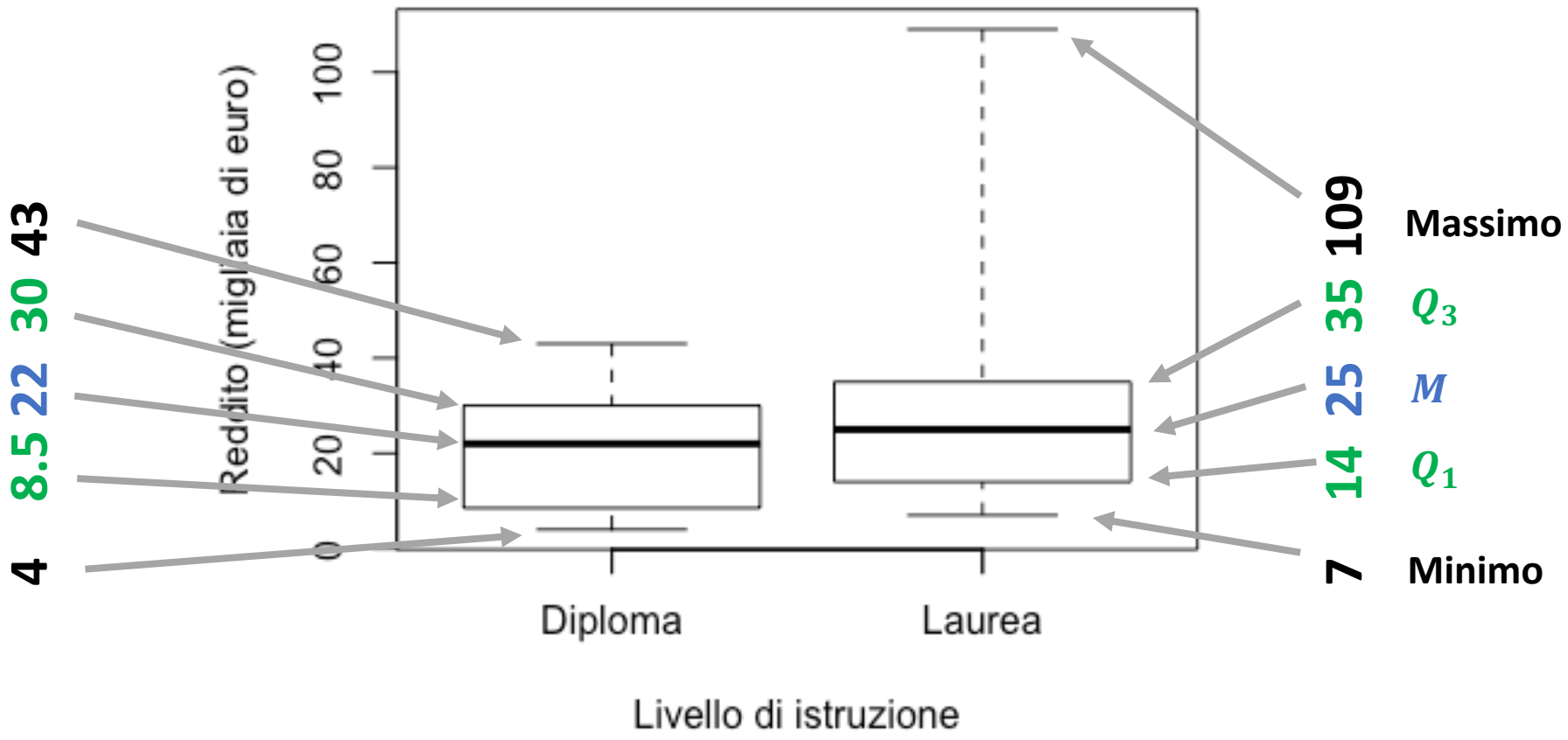
- Redditi medi laureati BI 2002

7	**14**	**25**	**35**	**109**

- Redditi medi diplomati BI 2002

4	**8.5**	**22**	**30**	**43**

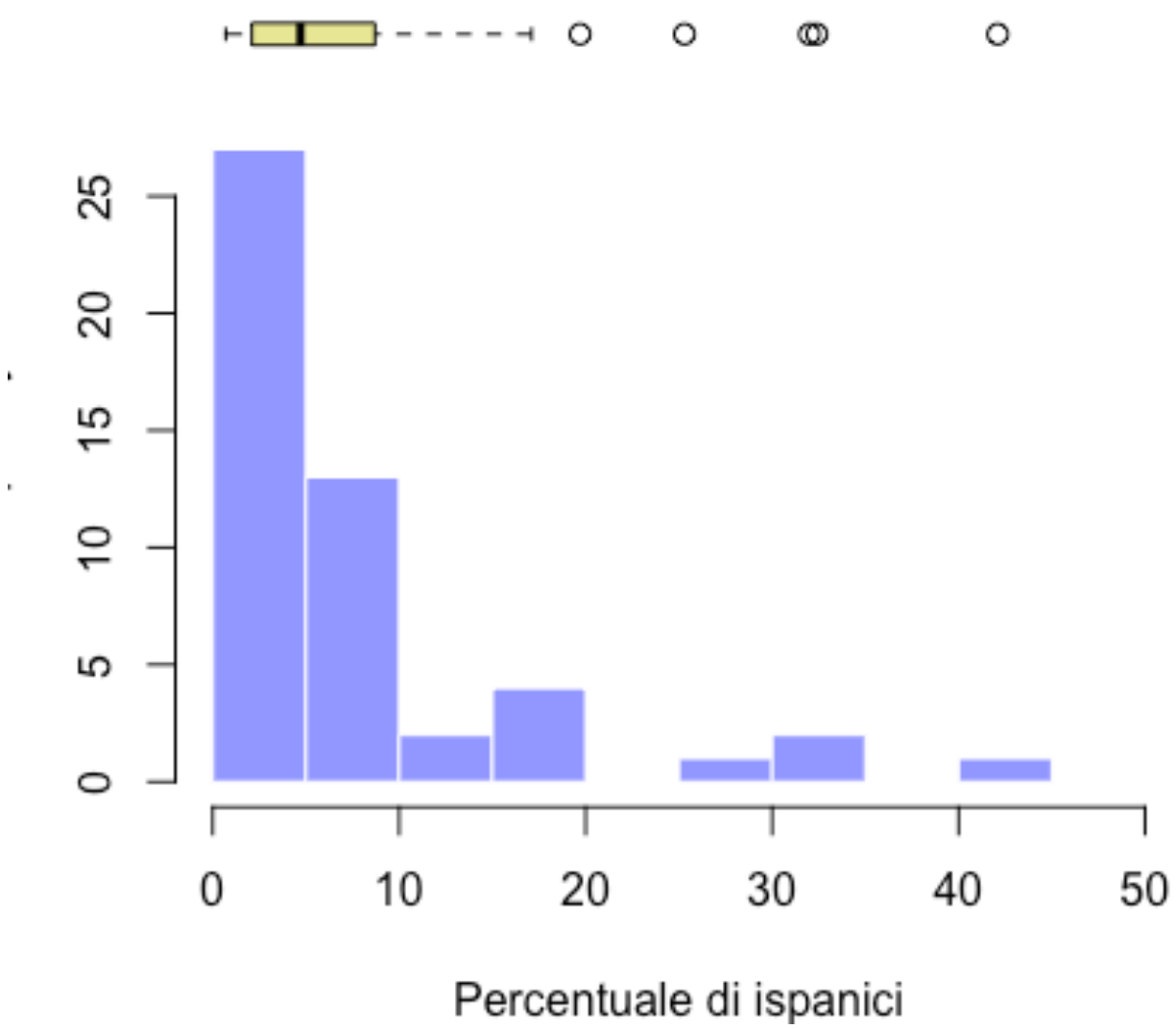
I grafici a scatola e baffi (box and whiskers)



- Dà informazioni chiare sulla simmetria o asimmetria (lunghezza dei baffi)

Boxplot e istogramma

- Versione alternativa di boxplot:
  - **baffi non arrivano a min e max** ma a una percentuale
  - gli **outlier sono punti (cerchi) esterni**
- Sovrapponiamo il boxplot all'istogramma
- Capiamo che nelle prime due colonne dell'istogramma c'è più del 50% della distribuzione (il box è lì sopra) e chi sono gli outlier



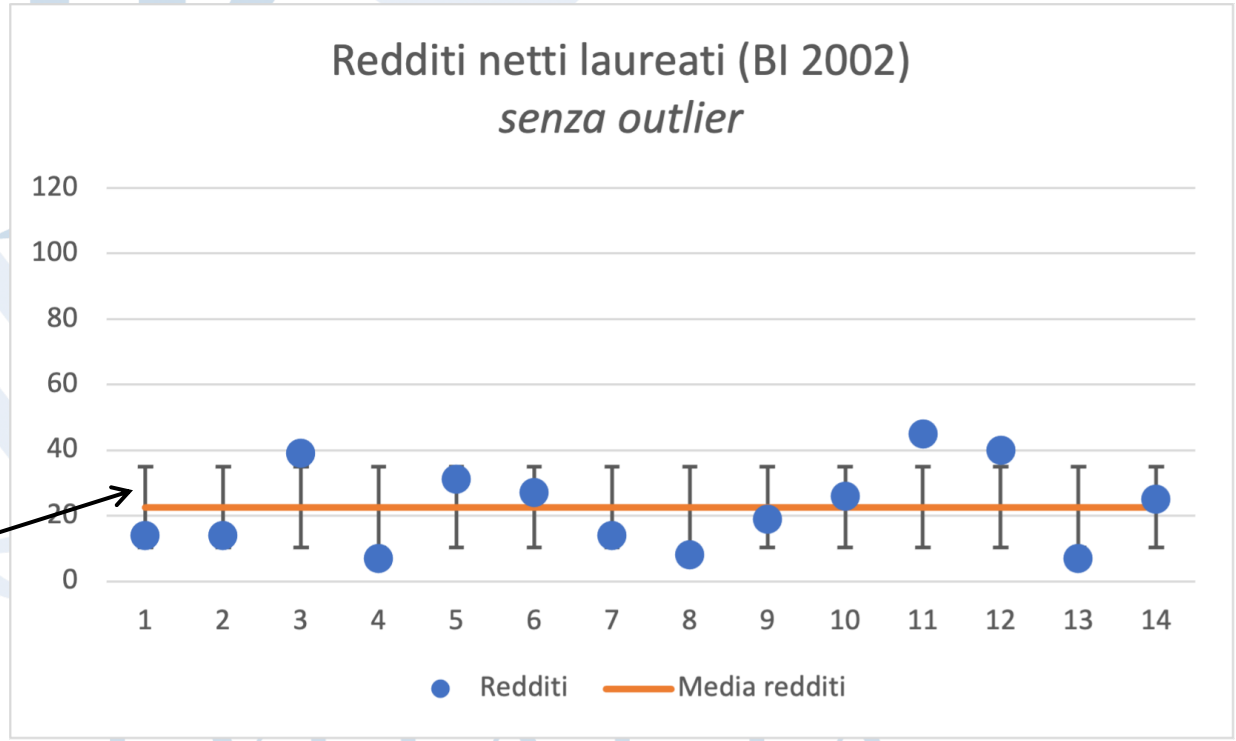
Esempio: redditi netti di laureati (BI 2002)

Osservazioni (tolgo outlier)

14	14	39	7	31	27	14	~~109~~	8	19	26	45	40	7	25

Quanto sono lontane, in media, le osservazioni dalla loro media?

Si usa la **deviazione standard**
Rappresentata dalle barre d'errore (*error bars*) in figura



Misurare la dispersione: deviazione standard

- Misura quanto sono lontane le osservazioni dalla loro media
- La **varianza s^2** è la media dei quadrati degli scarti

$$s^2 = \frac{(x_1 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n-1}$$

- La **deviazione standard s** è la radice quadrata della varianza s^2

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Esempio: redditi netti di laureati (BI 2002)

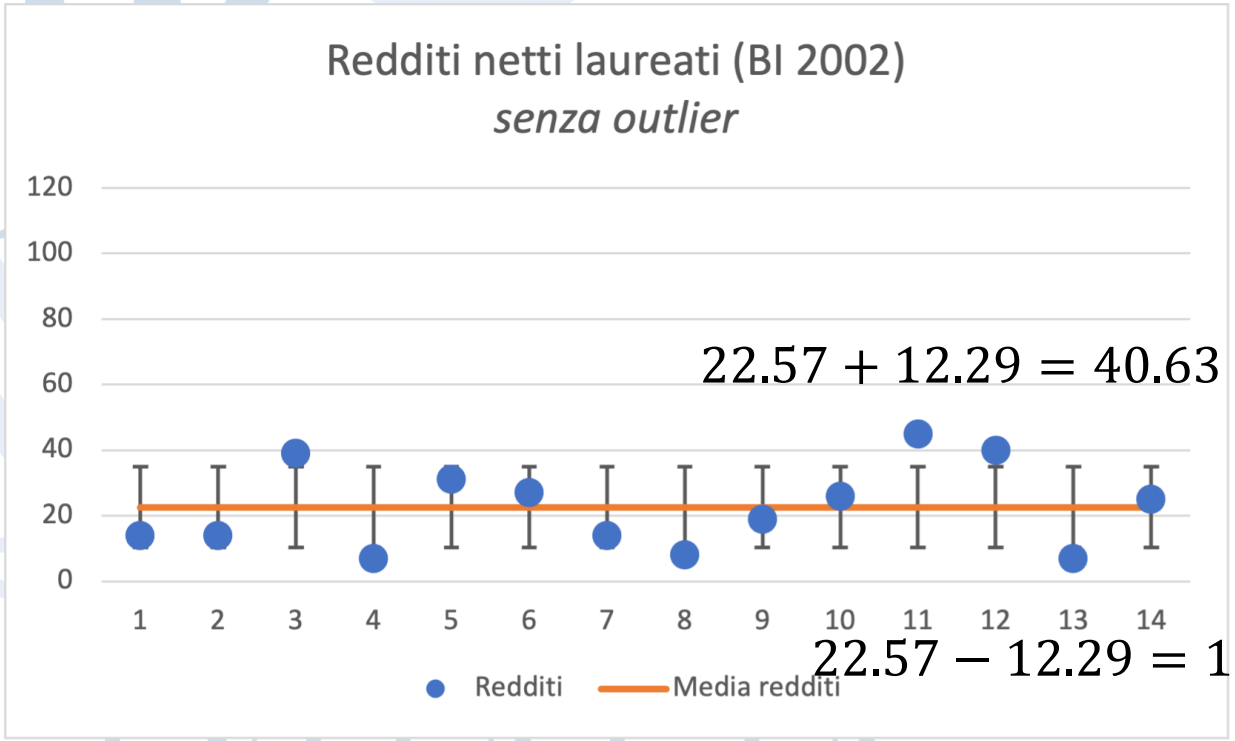
Osservazioni: **TOLGO L'OUTLIER**

14	14	39	7	31	27	14	~~109~~	8	19	26	45	40	7	25

Quanto sono lontane le osservazioni dalla loro media?

$\bar{x} = 22.57$ Linea arancione

$s = 12.29$ Barre verticali *Error bars*



$22.57 + 12.29 = 40.63$

$22.57 - 12.29 = 16.04$

Distr. a forma di campana: percentuale di osservazioni in $(\bar{x} - s, \bar{x} + s)$ è ~68%

Esempio: redditi netti di laureati (BI 2002)

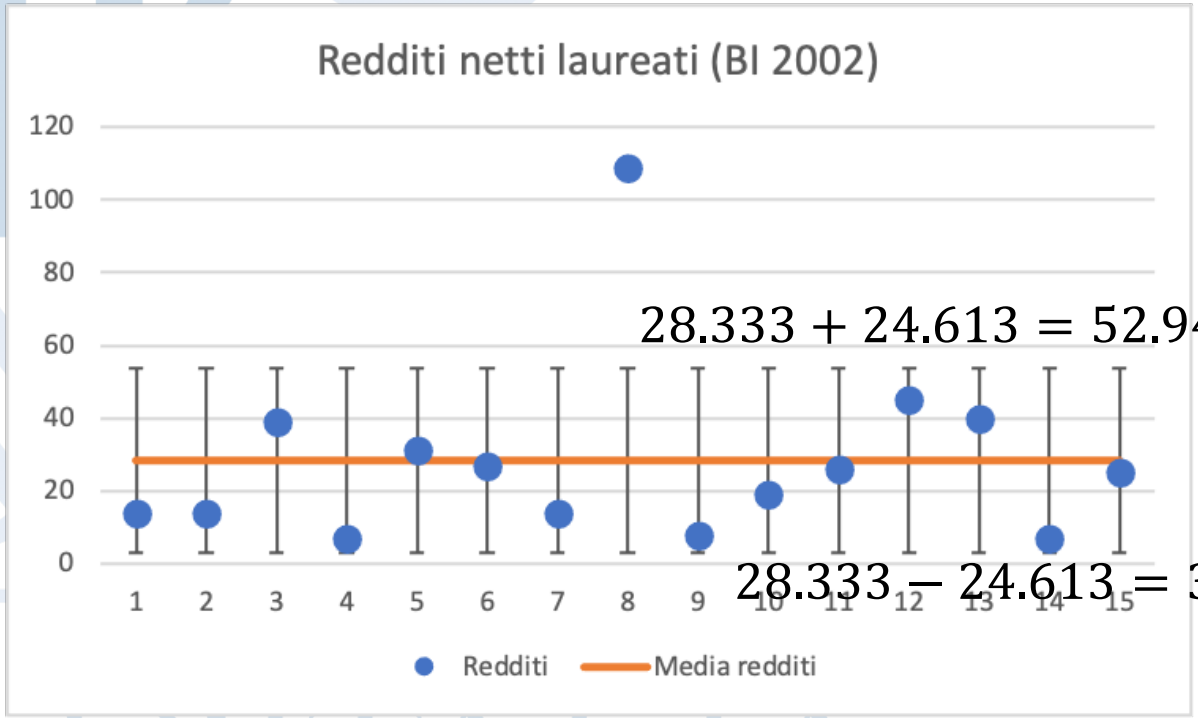
Osservazioni

14	14	39	7	31	27	14	109	8	19	26	45	40	7	25

Quanto sono lontane le osservazioni dalla loro media?

$\bar{x} = 28.333$ Linea arancione

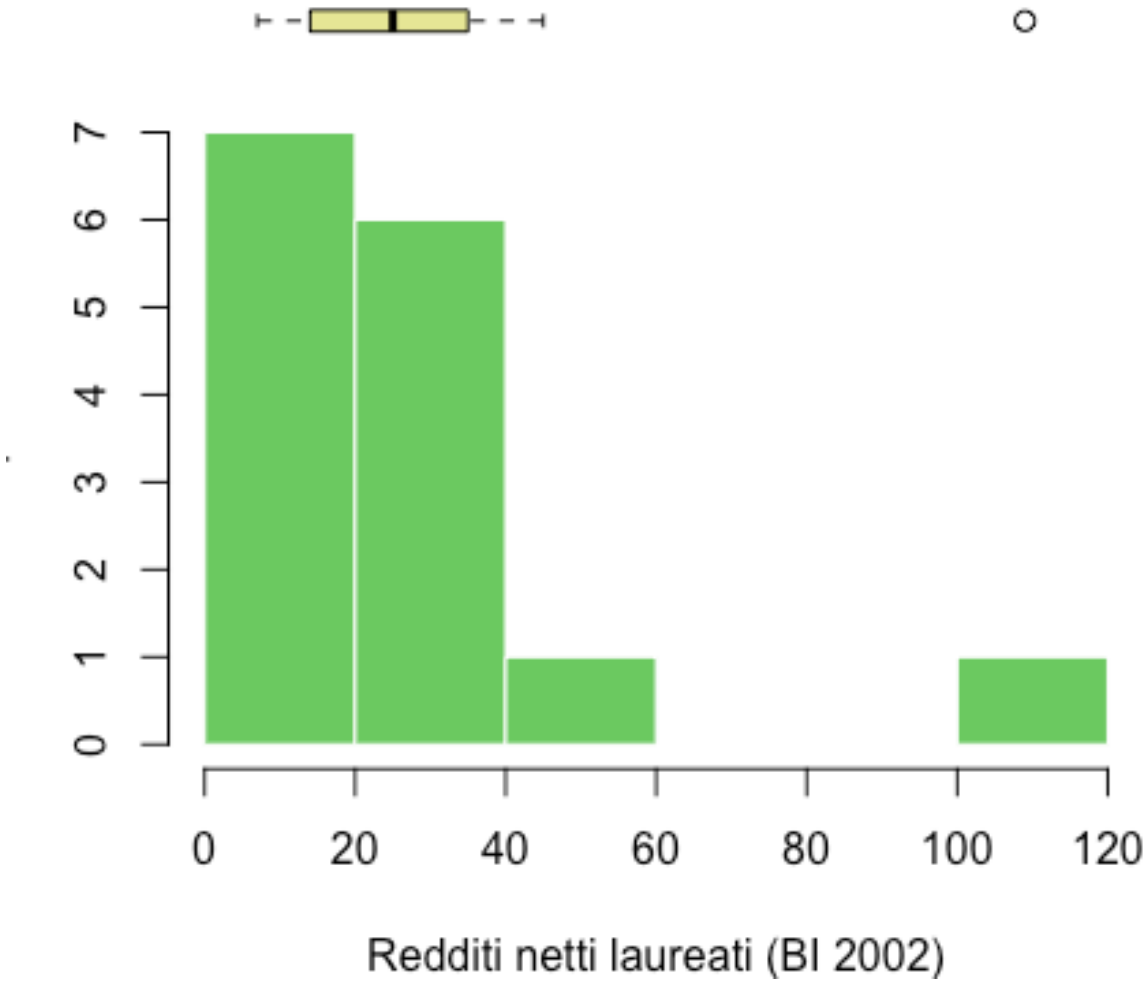
$s = 24.613$ Barre verticali *Error bars*



Per colpa dell'outlier: quasi tutte le osservazioni sono nell'intervallo $(\bar{x} - s, \bar{x} + s)$

Esempio: redditi netti di laureati (BI 2002)

- La distribuzione **non è simmetrica** perché ha outlier
- Quindi **non si usano media e deviazione standard**
- Ma **si usa il sommario (boxplot)**
- Si deve comunque **partire dall'istogramma**
- **Un istogramma con sopra il boxplot può andare**
- Si noti che abbiamo 6 colonne, e ciò è dovuto al fatto che abbiamo pochissime osservazioni



Scegliere le misure di centro e dispersione

- Media e deviazione standard per distribuzione simmetrica
- Sommario a cinque numeri per
  - Distribuzione asimmetrica
  - Distribuzione con molti outlier
- In generale, bisogna sempre rappresentare graficamente i dati