

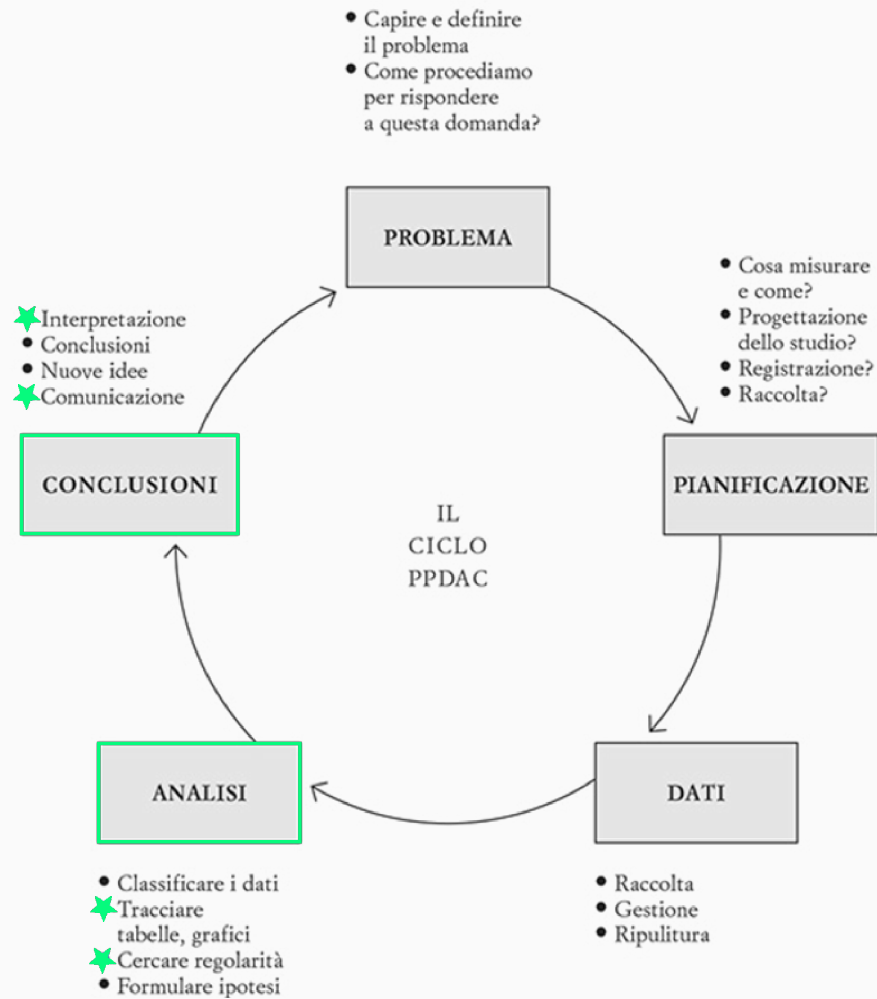
# **La statistica descrittiva**

## **(Parte II: Le variabili numeriche)**

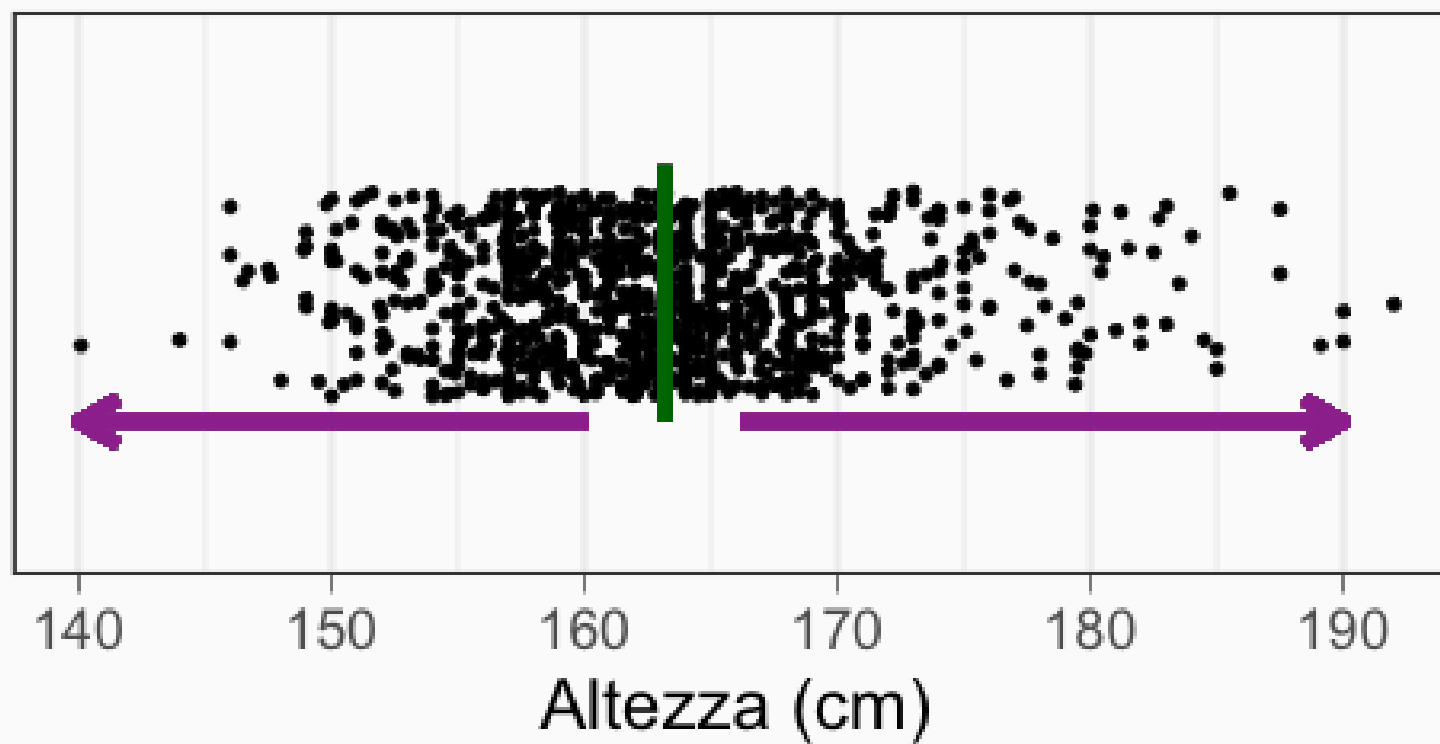
# Obiettivi di apprendimento

- Saper calcolare e interpretare misure di centralità, dispersione e correlazione
- Saper visualizzare dati numerici
- Saper interpretare tabelle e figure in articoli scientifici

# Le fasi della ricerca



# Misure di centralità e dispersione



# Misure di centralità: la moda



L'elemento più frequente

5	18	20	22	24	25	25	26	26	26	27	27	28	29	30
---	----	----	----	----	----	----	----	----	----	----	----	----	----	----

moda = 26

# Esercizio #1

? Qual è la moda dei seguenti insiemi?

$$x = \{1, 1, 1, 3, 4, 4, 4, 7, 8, 8, 9, 9\}$$
$$\text{moda}(x) = ?$$

$$y = \{1, 3, 4, 7, 8, 9, 11, 17, 21, 42\}$$
$$\text{moda}(y) = ?$$

# Esercizio #1 -- Soluzione



Qual è la moda dei seguenti insiemi?

$$x = \{1, 1, 1, 3, 4, 4, 4, 7, 8, 8, 9, 9\}$$

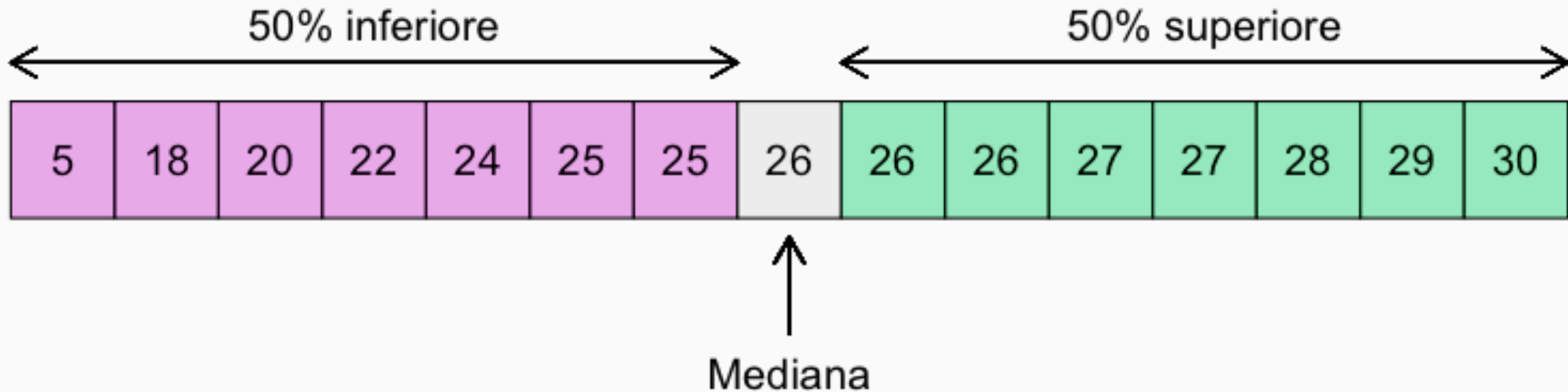
$$\text{moda}(x) = 1 \text{ e } 4$$

$$y = \{1, 3, 4, 7, 8, 9, 11, 17, 21, 42\}$$

$$\text{moda}(y) = \text{Non esiste}$$

# Misure di centralità: la mediana

🎯 Il valore "in mezzo"



⚠️ I dati devono essere ordinati!



## Esercizio #2

? Quali sono le mediane di questi insiemi?

$$x = \{6, 34, 40, 55, 75\}$$

$$\text{mediana}(x) = ?$$

$$y = \{6, 34, 40, 55, 175\}$$

$$\text{mediana}(y) = ?$$

## Esercizio #2 -- Soluzione

? Quali sono le mediane di questi insiemi?

$$x = \{6, 34, 40, 55, 75\}$$
$$\text{mediana}(x) = x_3 = 40$$

$$y = \{6, 34, 40, 55, 175\}$$
$$\text{mediana}(y) = ?$$

! I dati devono essere ordinati!

## Esercizio #2 -- Soluzione

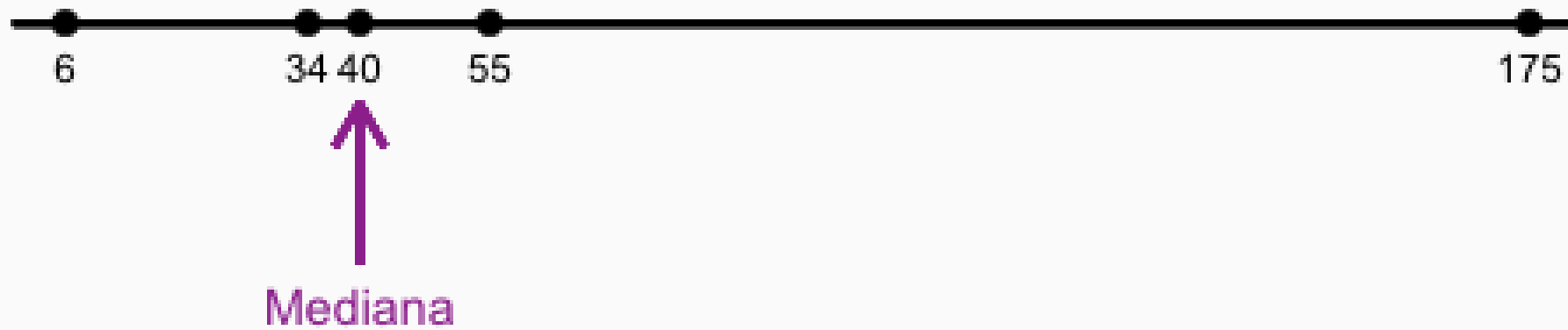
? Quali sono le mediane di questi insiemi?

$$x = \{6, 34, 40, 55, 75\}$$
$$\text{mediana}(x) = x_3 = 40$$

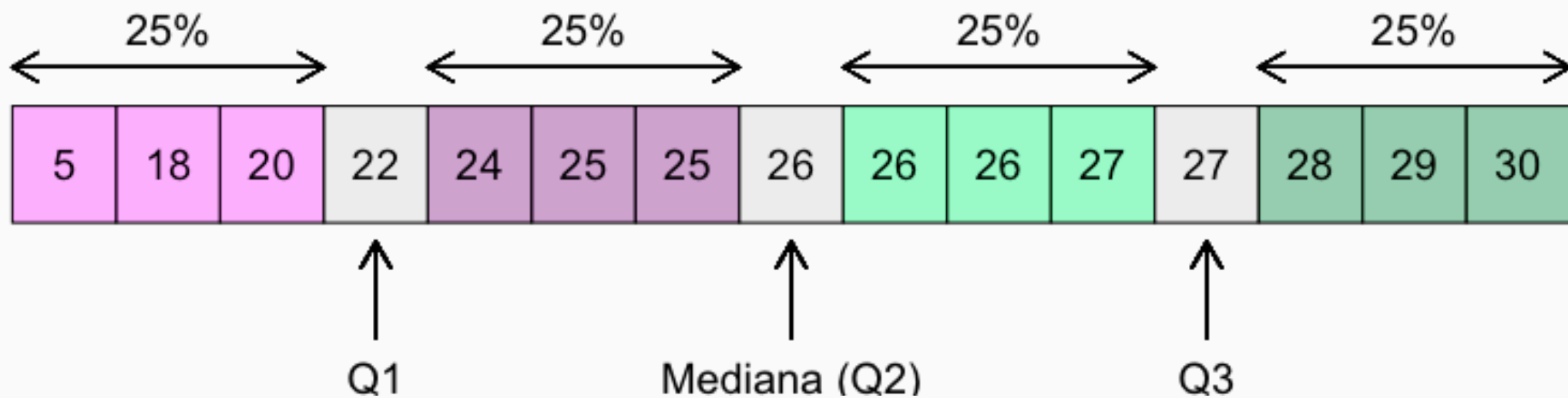
$$y = \{6, 34, 40, 55, 175\}$$
$$\text{mediana}(y) = y_3 = 40$$

! I dati devono essere ordinati!

# Mediana e valori estremi

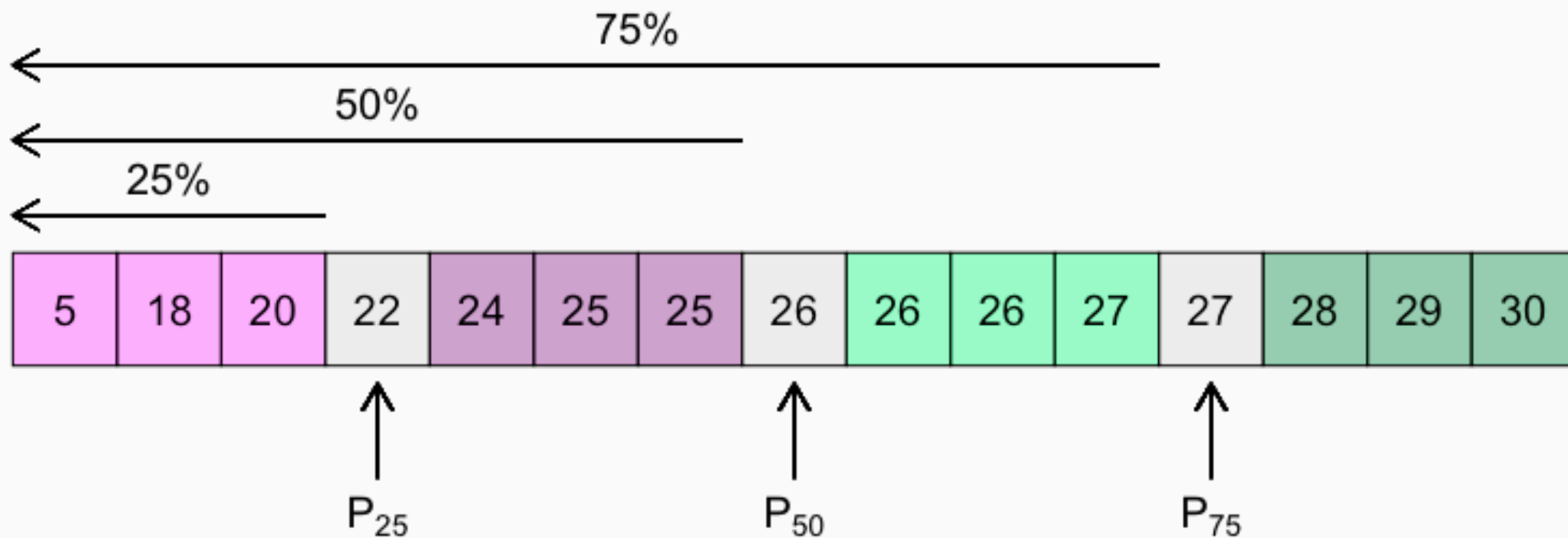


# Quartili



⚠ I dati devono essere ordinati!

# Percentili



⚠ I dati devono essere ordinati!

## Esercizio #3

? Maria ha ricevuto un punteggio di 70 a un esame, posizionandosi nel 45° percentile.


L'esame è andato...

- a) bene: ha ricevuto un voto superiore a più di metà delle persone che hanno dato quell'esame
- b) non benissimo: ha ricevuto un voto inferiore a più di metà delle persone che hanno dato quell'esame
- c) non ho abbastanza elementi per decidere

## Esercizio #3 -- Soluzione

? Maria ha ricevuto un punteggio di 70 a un esame, posizionandosi nel 45° percentile.

L'esame è andato...

- a) bene: ha ricevuto un voto superiore a più di metà delle persone che hanno dato quell'esame
- b) non benissimo: ha ricevuto un voto inferiore a più di metà delle persone che hanno dato quell'esame 
- c) non ho abbastanza elementi per decidere



# Misure di centralità: la media



Media (aritmetica)

$$\bar{x} = \frac{1}{n} \left( \sum_{i=1}^n x_i \right) = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

5	18	20	22	24	25	25	26	26	26	27	27	28	29	30
---	----	----	----	----	----	----	----	----	----	----	----	----	----	----

$$\bar{x} = \frac{5+18+20+22+24+25+25+26+26+26+27+27+28+29+30}{15} = 23.9$$

## Esercizio #4

? Quali sono le medie di questi insiemi?

$$x = \{6, 34, 40, 55, 75\}$$

$$\bar{x} = ?$$

$$y = \{6, 34, 40, 55, 175\}$$

$$\bar{y} = ?$$

## Esercizio #4 -- Soluzione



Quali sono le medie di questi insiemi?

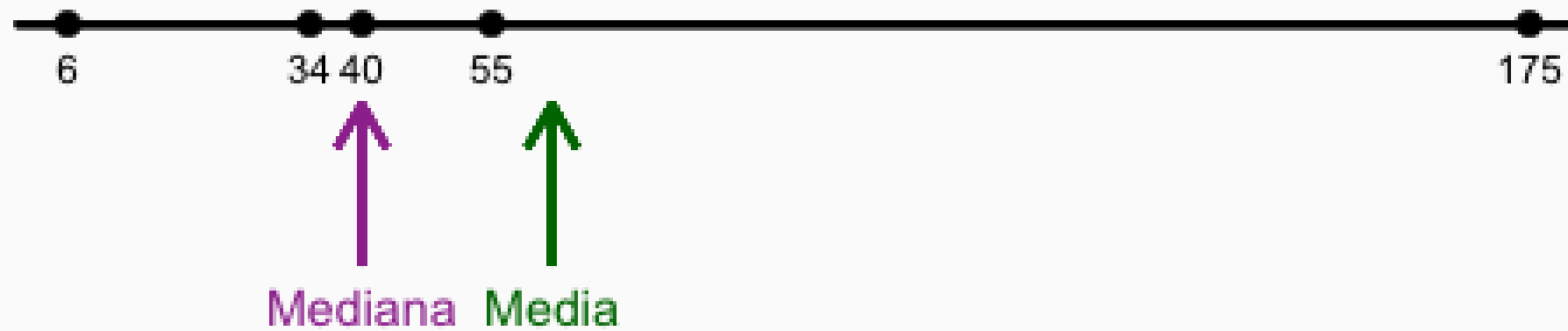
$$x = \{6, 34, 40, 55, 75\}$$

$$\bar{x} = \frac{1}{n} \left( \sum_{i=1}^n x_i \right) = \frac{6+34+40+55+75}{5} = 42$$

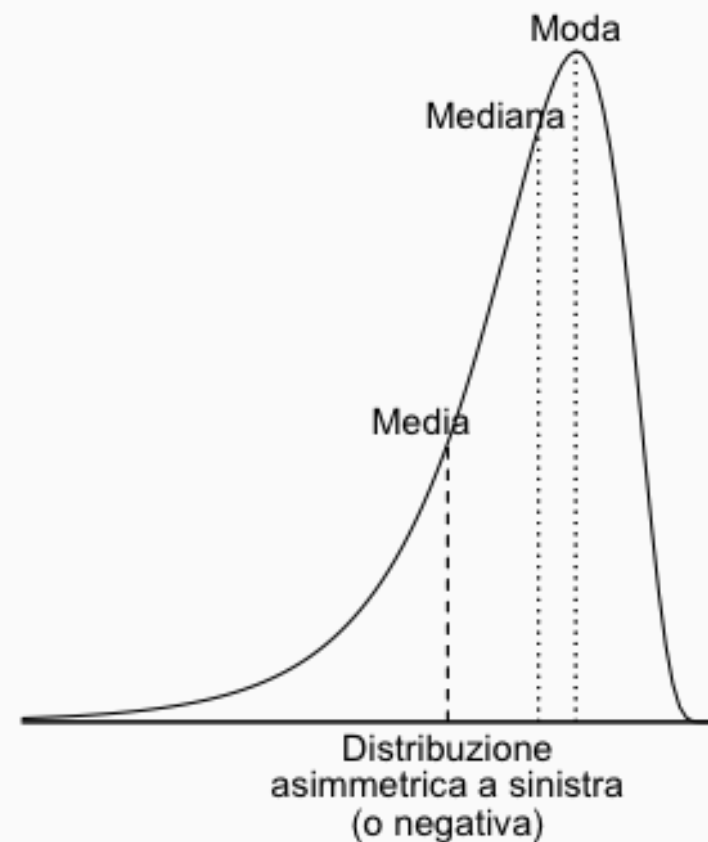
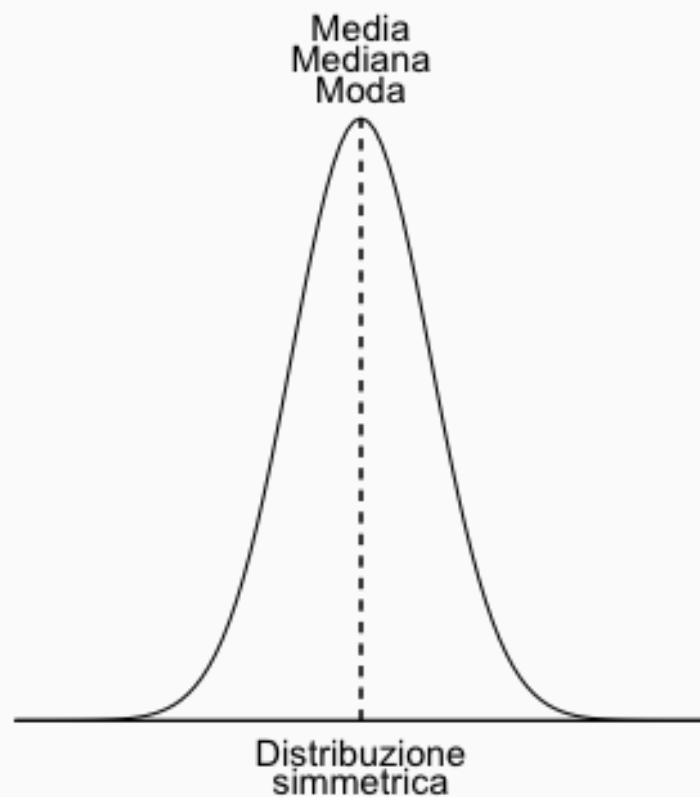
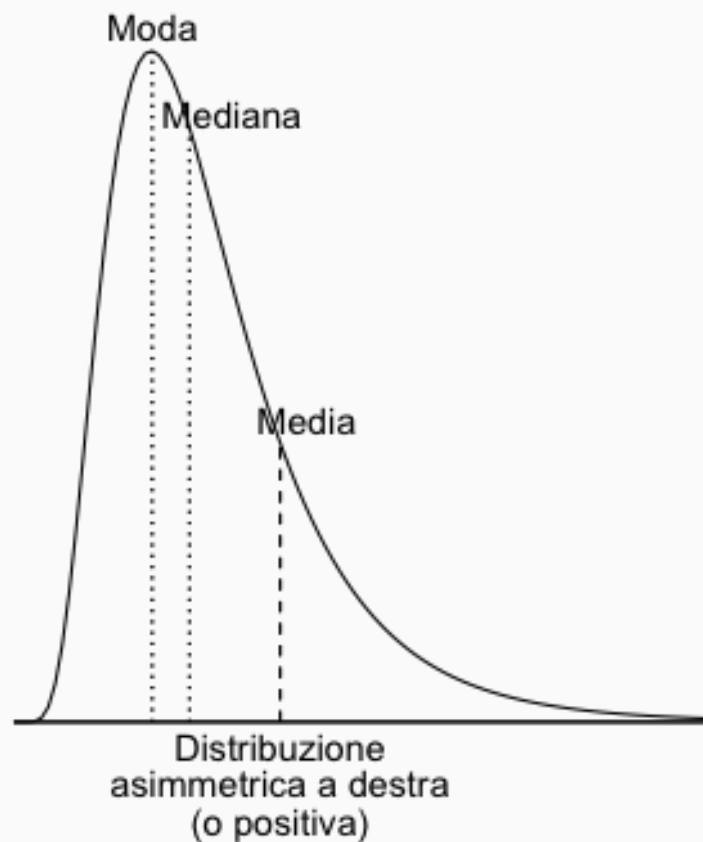
$$y = \{6, 34, 40, 55, 175\}$$

$$\bar{y} = \frac{1}{n} \left( \sum_{i=1}^n y_i \right) = \frac{4+36+45+50+175}{5} = 62$$

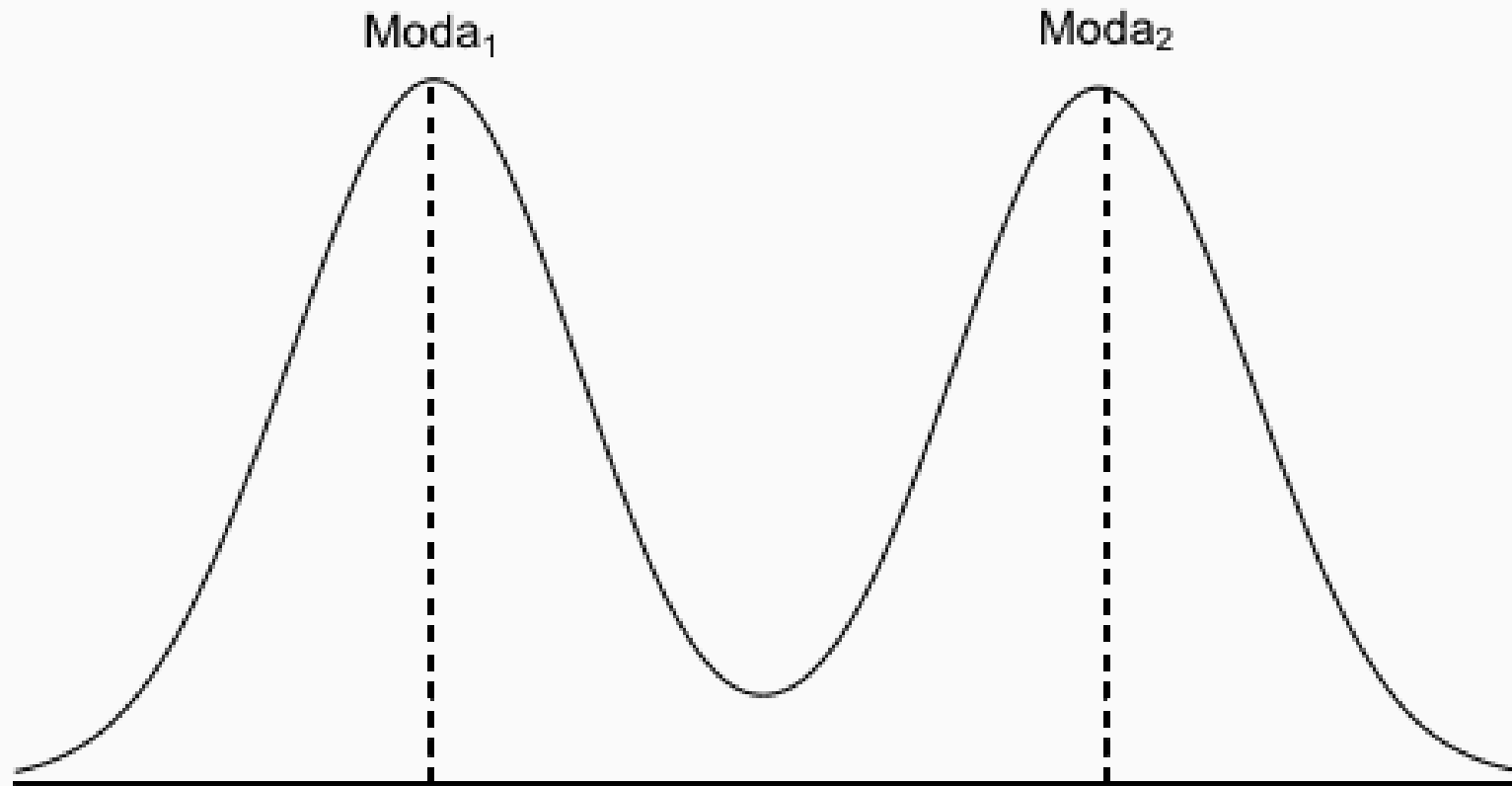
# Media e valori estremi



# La forma delle distribuzioni



# La forma delle distribuzioni



## Esercizio #5

? Nei risultati di uno studio è riportata la seguente frase:

*The mean length of stay was 22.4 days (median: 14 days).*

La distribuzione empirica ha una forma...


- a) simmetrica
- b) asimmetrica a destra
- c) asimmetrica a sinistra
- d) nessuna delle precedenti

## Esercizio 5 -- Soluzione

? Nei risultati di uno studio è riportata la seguente frase:

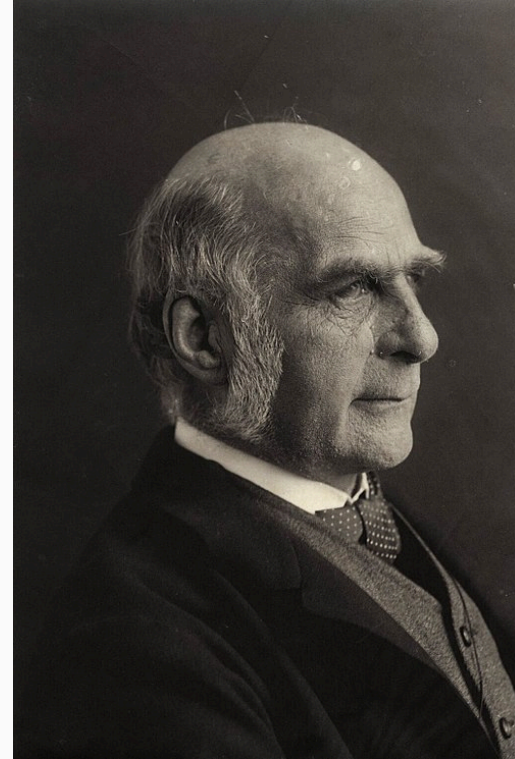
*The mean length of stay was 22.4 days (median: 14 days).*

La distribuzione empirica ha una forma...

- a) simmetrica
- b) asimmetrica a destra 
- c) asimmetrica a sinistra
- d) nessuna delle precedenti



# Vox populi \*



\* Wisdom of Crowds o Saggezza della Folla

Galton, F. *Vox Populi*, Nature, 1907, <https://doi.org/10.1038/075450a0>

# Vox populi \*

- Competizione presso la "Mostra del Pollame e del Bestiame da Macello, Plymouth, 1907
- Obiettivo: indovinare il peso "lavorato" della carne macellata
- Costo: 6 penny
- Partecipanti: 787
- Vincita: premio non specificato



\* Wisdom of Crowds o Saggezza della Folla

# Vox populi \*

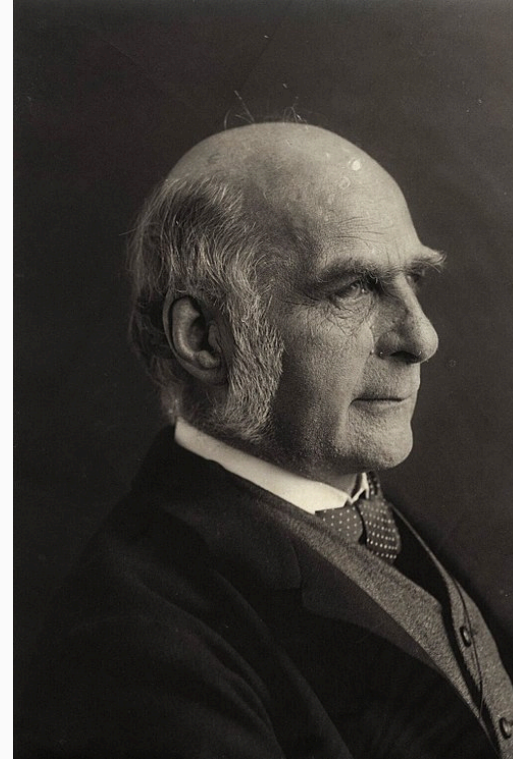
- Giudizio non influenzato da passioni personali e/o proselitismo vario
- Burloni evitati dal costo di ingresso
- Premio incita a fare del proprio meglio
- Partecipano soprattutto "esperti"



\* Wisdom of Crowds o Saggezza della Folla

# Vox populi \*

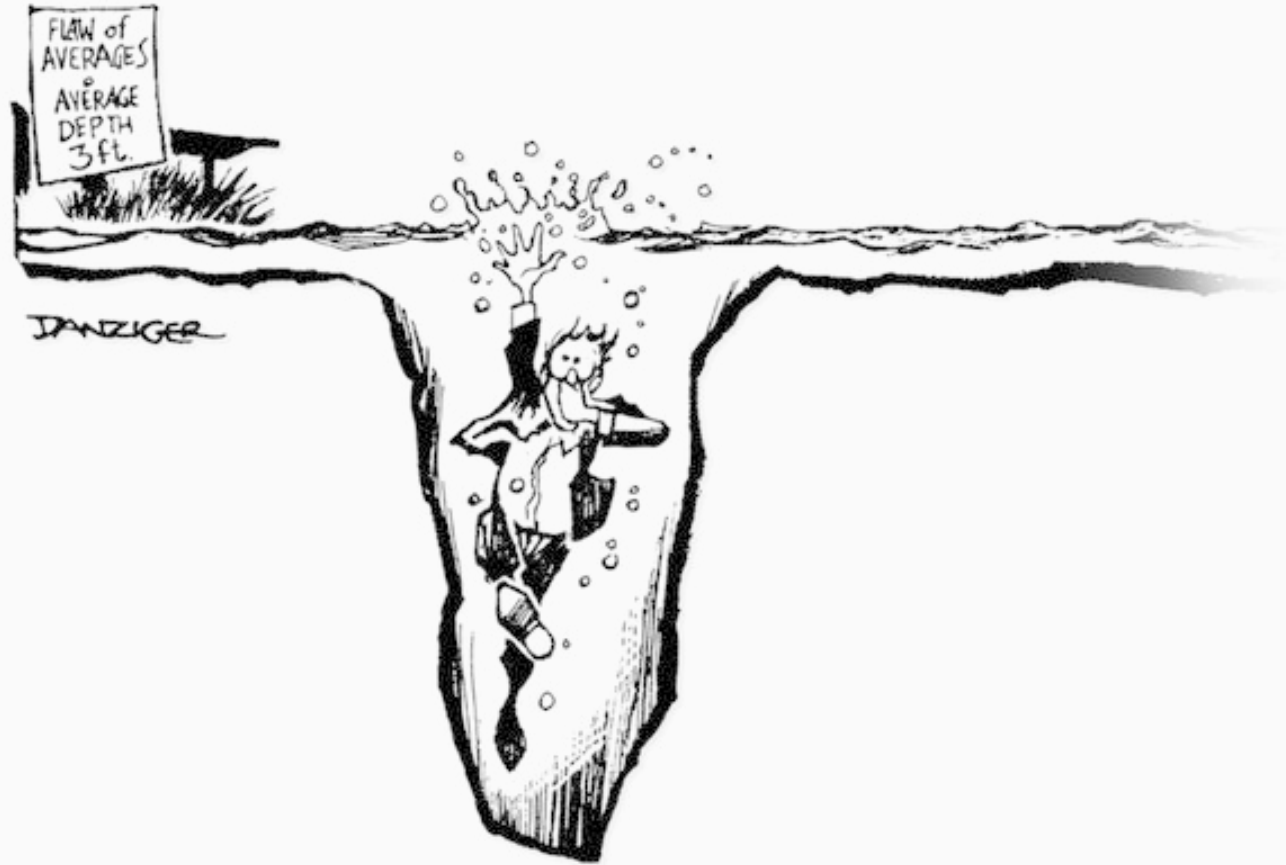
- **Mediana** dei 787 tentativi: 1207 lb (547 kg)
- Peso "lavorato": 1198 lb (543 kg)
- Differenza: 9 lb (4kg, 0.8%)



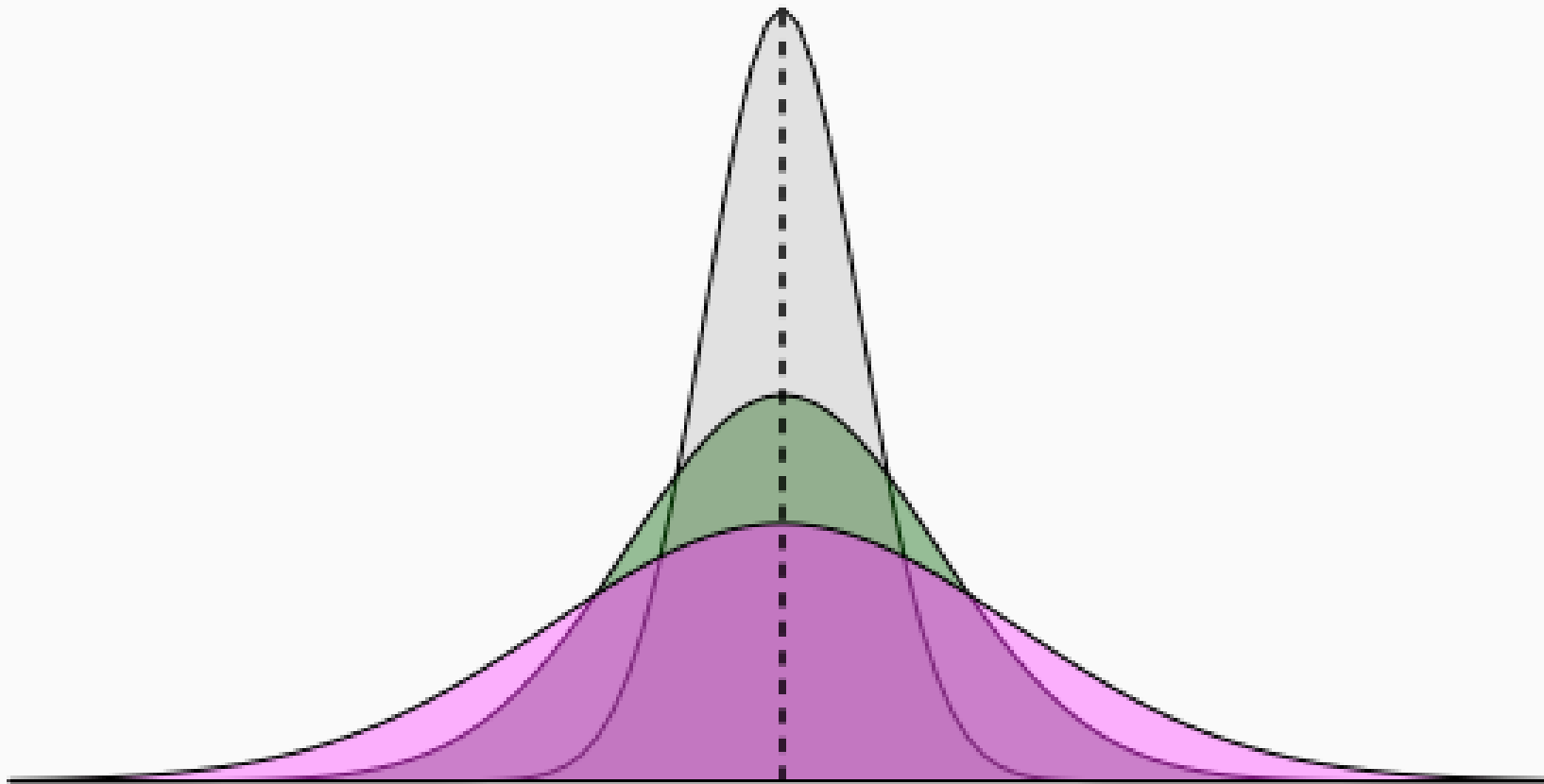
\* Wisdom of Crowds o Saggezza della Folla



# Misure di dispersione



# Misure di dispersione



# Misure di dispersione: range



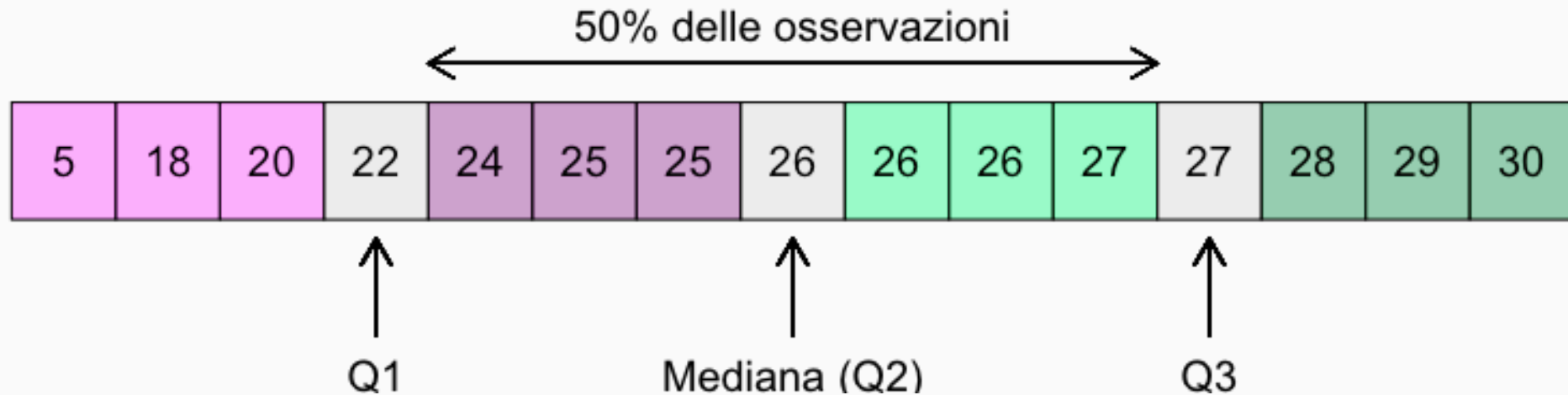
$$\text{range} = \text{max} - \text{min}$$

5	18	20	22	24	25	25	26	26	26	27	27	28	29	30
---	----	----	----	----	----	----	----	----	----	----	----	----	----	----

$$\text{range} = 30 - 5 = 25$$

# Misure di dispersione: range interquantile

🎯  $IQR = Q1 - Q3$

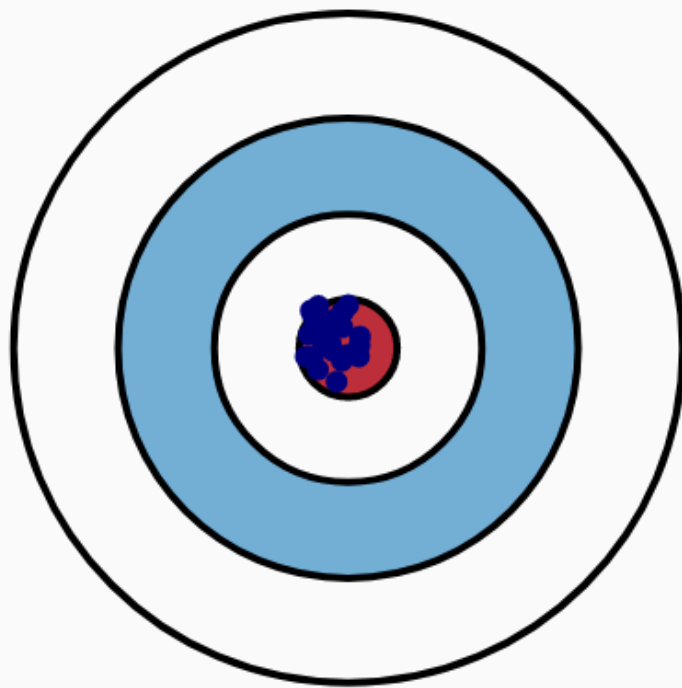


$$IQR = 22 - 27$$

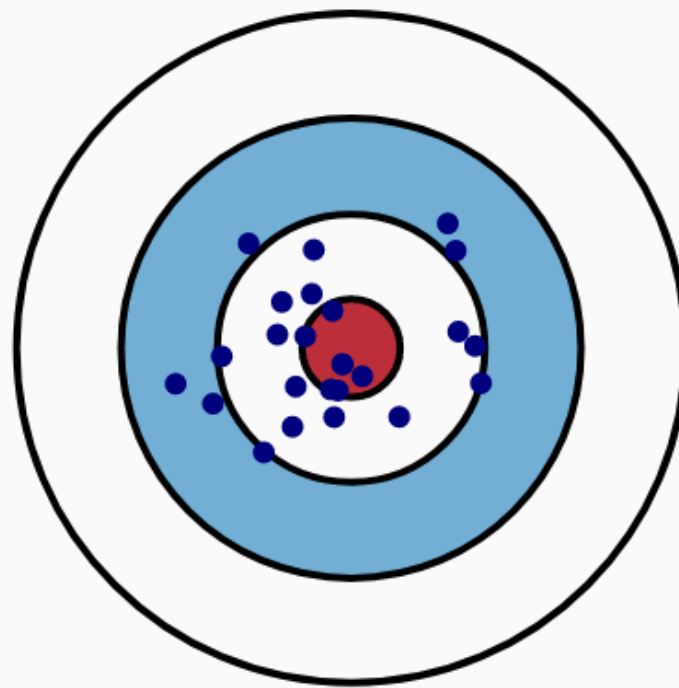


# Misure di dispersione: varianza

Low Variance



High Variance



# Misure di dispersione: varianza


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

5	18	20	22	24	25	25	26	26	26	27	27	28	29	30
---	----	----	----	----	----	----	----	----	----	----	----	----	----	----

$$\bar{x} = 23.9$$

$$s^2 = \frac{(5-23.9)^2 + (18-23.9)^2 + (20-23.9)^2 + \dots + (28-23.9)^2 + (29-23.9)^2 + (30-23.9)^2}{(15-1)} = 37.6$$

# Misure di dispersione: deviazione standard

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

5	18	20	22	24	25	25	26	26	26	27	27	28	29	30
---	----	----	----	----	----	----	----	----	----	----	----	----	----	----

$$\bar{x} = 23.9$$

$$s = \sqrt{\frac{(5-23.9)^2 + (18-23.9)^2 + \dots + (29-23.9)^2 + (30-23.9)^2}{(15-1)}} = \sqrt{37.6} = 6.1$$

# Esercizio #6

- ? Il range è sensibile alla posizione centrale della distribuzione empirica  
a) Vero      b) Falso
- ? La mediana si calcola sommando i valori e dividendoli per il loro numero  
a) Vero      b) Falso
- ? La mediana è il valore che ha metà dei dati inferiori e metà superiori a esso  
a) Vero      b) Falso
- ? La mediana, rispetto alla media, è più sensibile ai valori estremi  
a) Vero      b) Falso
- ? Due distribuzioni con la stessa media hanno la stessa deviazione standard  
a) Vero      b) Falso

# Esercizio #6 -- Soluzione

- ? Il range è sensibile alla posizione centrale della distribuzione empirica  
a) Vero      b) Falso ☒
- ? La mediana si calcola sommando i valori e dividendoli per il loro numero  
a) Vero      b) Falso
- ? La mediana è il valore che ha metà dei dati inferiori e metà superiori a esso  
a) Vero      b) Falso
- ? La mediana, rispetto alla media, è più sensibile ai valori estremi  
a) Vero      b) Falso
- ? Due distribuzioni con la stessa media hanno la stessa deviazione standard  
a) Vero      b) Falso

# Esercizio #6 -- Soluzione

- ? Il range è sensibile alla posizione centrale della distribuzione empirica  
a) Vero      b) Falso      ☒
- ? La mediana si calcola sommando i valori e dividendoli per il loro numero  
a) Vero      b) Falso      ☒
- ? La mediana è il valore che ha metà dei dati inferiori e metà superiori a esso  
a) Vero      b) Falso
- ? La mediana, rispetto alla media, è più sensibile ai valori estremi  
a) Vero      b) Falso
- ? Due distribuzioni con la stessa media hanno la stessa deviazione standard  
a) Vero      b) Falso

# Esercizio #6 -- Soluzione

- ? Il range è sensibile alla posizione centrale della distribuzione empirica  
a) Vero      b) Falso ☒
- ? La mediana si calcola sommando i valori e dividendoli per il loro numero  
a) Vero      b) Falso ☒
- ? La mediana è il valore che ha metà dei dati inferiori e metà superiori a esso  
a) Vero ☒      b) Falso
- ? La mediana, rispetto alla media, è più sensibile ai valori estremi  
a) Vero      b) Falso
- ? Due distribuzioni con la stessa media hanno la stessa deviazione standard  
a) Vero      b) Falso

# Esercizio #6 -- Soluzione

- ? Il range è sensibile alla posizione centrale della distribuzione empirica  
a) Vero      b) Falso      ☒
- ? La mediana si calcola sommando i valori e dividendoli per il loro numero  
a) Vero      b) Falso      ☒
- ? La mediana è il valore che ha metà dei dati inferiori e metà superiori a esso  
a) Vero ☒      b) Falso
- ? La mediana, rispetto alla media, è più sensibile ai valori estremi  
a) Vero      b) Falso      ☒
- ? Due distribuzioni con la stessa media hanno la stessa deviazione standard  
a) Vero      b) Falso



# Esercizio #6 -- Soluzione

- ? Il range è sensibile alla posizione centrale della distribuzione empirica  
a) Vero      b) Falso      ☒
- ? La mediana si calcola sommando i valori e dividendoli per il loro numero  
a) Vero      b) Falso      ☒
- ? La mediana è il valore che ha metà dei dati inferiori e metà superiori a esso  
a) Vero ☒      b) Falso
- ? La mediana, rispetto alla media, è più sensibile ai valori estremi  
a) Vero      b) Falso      ☒
- ? Due distribuzioni con la stessa media hanno la stessa deviazione standard  
a) Vero      b) Falso      ☒

# I valori estremi

TABLE 3. Length of In-Patient Stay, by Surgical Procedure			
Procedure	No. of procedures	Length of stay, d	
		Mean $\pm$ SD	Median (IQR)
Breast surgery	1,338	3.3 $\pm$ 4.4	3 (0-5)
Coronary artery bypass graft	570	9.6 $\pm$ 15.2	8 (7-9)
Cesarean section	4,831	4.9 $\pm$ 6.4	4 (3-5)
Repair of fractured neck of femur	2,303	13.8 $\pm$ 12.2	10 (7-17)
Hip replacement	6,432	8.7 $\pm$ 5.9	7 (6-9)
Abdominal hysterectomy	1,484	5.4 $\pm$ 4.0	5 (4-6)
Knee replacement	4,483	8.2 $\pm$ 5.0	7 (6-9)
Major vascular surgery	269	22.4 $\pm$ 23.1	14 (8-30)
Overall	21,710	7.8 $\pm$ 8.0	6 (4- 9)

The mean length of stay was 7.8 days but was greatly influenced by 2 patients with lengths of stay of almost 1 year. The median length of stay was 6 days, with 90% of patients discharged within 14 days after the procedure. Table 3 displays measures of central tendency (mean and median values) and dispersion (SDs and interquartile ranges) for the length of stay for each type of surgical procedure. .

# Esercizio #7

? Nei risultati di uno studio è riportata la seguente frase:

*The density of calcification in the coronary artery averaged  $68.9 \pm 244.2$  (range 0 to 1526) in patients and  $8.8 \pm 41.8$  (range 0 to 243.4) in controls.*

Come descrivereste in Table 1 questa variabile?


- a) con media e deviazione standard
- b) con mediana e interquantile range
- c) con mediana e deviazione standard
- d) non ho abbastanza elementi per decidere

# Esercizio #7 -- Soluzione

? Nei risultati di uno studio è riportata la seguente frase:

*The density of calcification in the coronary artery averaged  $68.9 \pm 244.2$  (range 0 to 1526) in patients and  $8.8 \pm 41.8$  (range 0 to 243.4) in controls.*

Come descrivereste in Table 1 questa variabile?

- a) con media e deviazione standard
- b) con mediana e interquantile range 
- c) con mediana e deviazione standard
- d) non ho abbastanza elementi per decidere

# Esercizio #8

Table 1. Demographic Characteristics of the Participants		
Characteristic	All Participants (N = 277)	
	Oxytocin (N = 139)	Placebo (N = 138)
Age		
Mean — yr	10.4±4.1	10.4±4.0
Distribution — no. (%)		
3–6 yr	34 (24)	35 (25)
7–11 yr	54 (39)	53 (38)
12–17 yr	51 (37)	50 (36)
Sex — no. (%)		
Male	122 (88)	120 (87)
Female	17 (12)	18 (13)



Qual è la percentuale femmine nel gruppo di intervento?


- a) 13%
- b) 12%
- c) 18%
- d) 17%
- e) Non è possibile capirlo dalla tabella

# Esercizio #8 -- Soluzione

Table 1. Demographic Characteristics of the Participants		
Characteristic	All Participants (N = 277)	
	Oxytocin (N = 139)	Placebo (N = 138)
Age		
Mean — yr	10.4±4.1	10.4±4.0
Distribution — no. (%)		
3–6 yr	34 (24)	35 (25)
7–11 yr	54 (39)	53 (38)
12–17 yr	51 (37)	50 (36)
Sex — no. (%)		
Male	122 (88)	120 (87)
Female	17 (12)	18 (13)



Qual è la percentuale femmine nel gruppo di intervento?

- a) 13%
- b) 12% 
- c) 18%
- d) 17%
- e) Non è possibile capirlo dalla tabella

Sikich, L. et al., *Intranasal Oxytocin in Children and Adolescents with Autism Spectrum Disorder*, NEJM, 2021

# Esercizio #9

Table 1. Demographic Characteristics of the Participants		
Characteristic	All Participants (N = 277)	
	Oxytocin (N = 139)	Placebo (N = 138)
Age		
Mean — yr	10.4±4.1	10.4±4.0
Distribution — no. (%)		
3–6 yr	34 (24)	35 (25)
7–11 yr	54 (39)	53 (38)
12–17 yr	51 (37)	50 (36)
Sex — no. (%)		
Male	122 (88)	120 (87)
Female	17 (12)	18 (13)



In questo studio, l'età è stata raccolta come una variabile...

- a) categorica
- b) ordinale
- c) numerica
- d) non è possibile dirlo

Sikich, L. et al., *Intranasal Oxytocin in Children and Adolescents with Autism Spectrum Disorder*, NEJM, 2021


00:30

# Esercizio #9 -- Soluzione

Table 1. Demographic Characteristics of the Participants		
Characteristic	All Participants (N = 277)	
	Oxytocin (N = 139)	Placebo (N = 138)
Age		
Mean — yr	10.4±4.1	10.4±4.0
Distribution — no. (%)		
3–6 yr	34 (24)	35 (25)
7–11 yr	54 (39)	53 (38)
12–17 yr	51 (37)	50 (36)
Sex — no. (%)		
Male	122 (88)	120 (87)
Female	17 (12)	18 (13)



In questo studio, l'età è stata raccolta come una variabile...

- a) categorica
- b) ordinale
- c) numerica 
- d) non è possibile dirlo

Sikich, L. et al., *Intranasal Oxytocin in Children and Adolescents with Autism Spectrum Disorder*, NEJM, 2021



# Esercizio #10

Table 1. Demographic Characteristics of the Participants		
Characteristic	All Participants (N = 277)	
	Oxytocin (N = 139)	Placebo (N = 138)
Age		
Mean — yr	10.4±4.1	10.4±4.0
Distribution — no. (%)		
3–6 yr	34 (24)	35 (25)
7–11 yr	54 (39)	53 (38)
12–17 yr	51 (37)	50 (36)
Sex — no. (%)		
Male	122 (88)	120 (87)
Female	17 (12)	18 (13)



Qual è l'età media nel gruppo di controllo?

- a) 10.4
- b) 4.1
- c) 4.0
- d) Non è possibile capirlo dalla tabella

Sikich, L. et al., *Intranasal Oxytocin in Children and Adolescents with Autism Spectrum Disorder*, NEJM, 2021


00:30

# Esercizio #10 -- Soluzione

Table 1. Demographic Characteristics of the Participants		
Characteristic	All Participants (N = 277)	
	Oxytocin (N = 139)	Placebo (N = 138)
Age		
Mean — yr	10.4±4.1	10.4±4.0
Distribution — no. (%)		
3–6 yr	34 (24)	35 (25)
7–11 yr	54 (39)	53 (38)
12–17 yr	51 (37)	50 (36)
Sex — no. (%)		
Male	122 (88)	120 (87)
Female	17 (12)	18 (13)



Qual è l'età media nel gruppo di controllo?

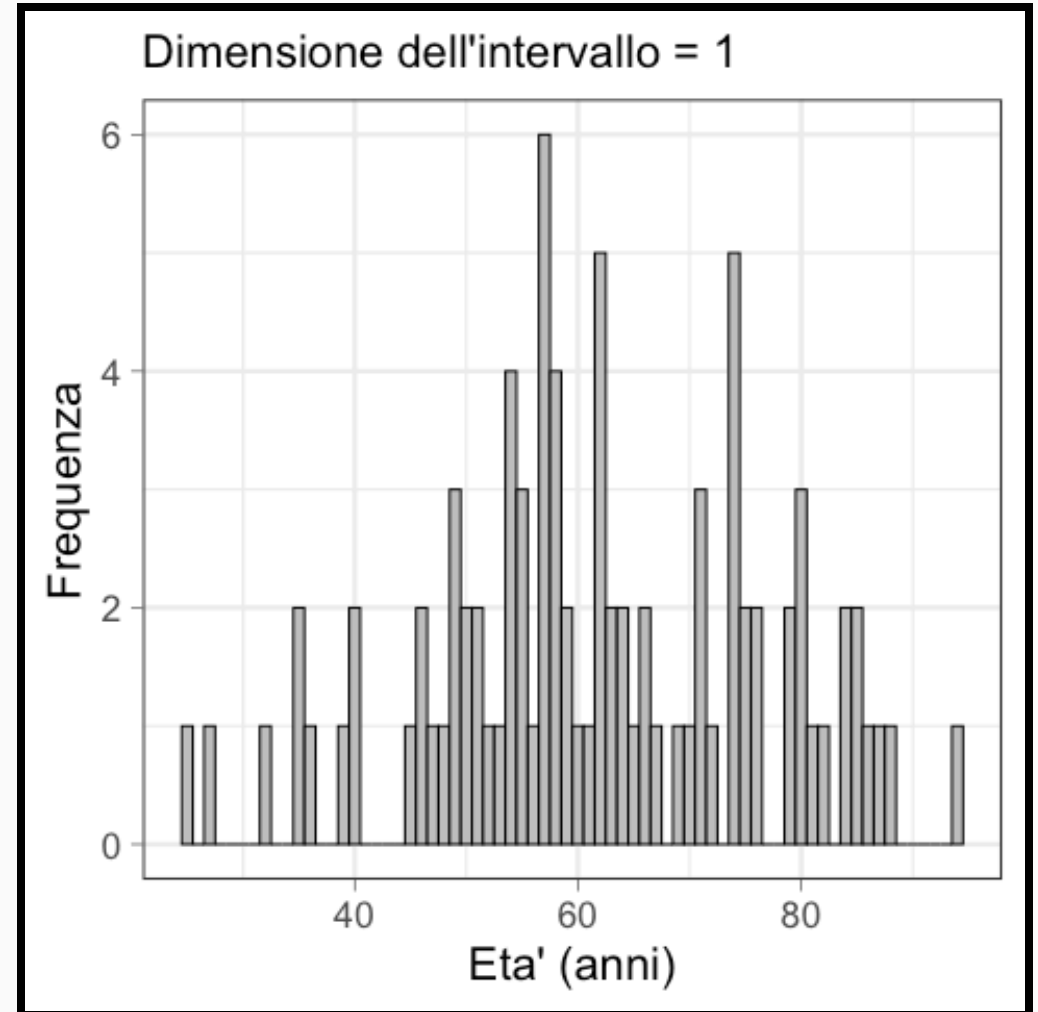
- a) 10.4 
- b) 4.1
- c) 4.0
- d) Non è possibile capirlo dalla tabella

Sikich, L. et al., *Intranasal Oxytocin in Children and Adolescents with Autism Spectrum Disorder*, NEJM, 2021

# **La visualizzazione dei dati numerici**

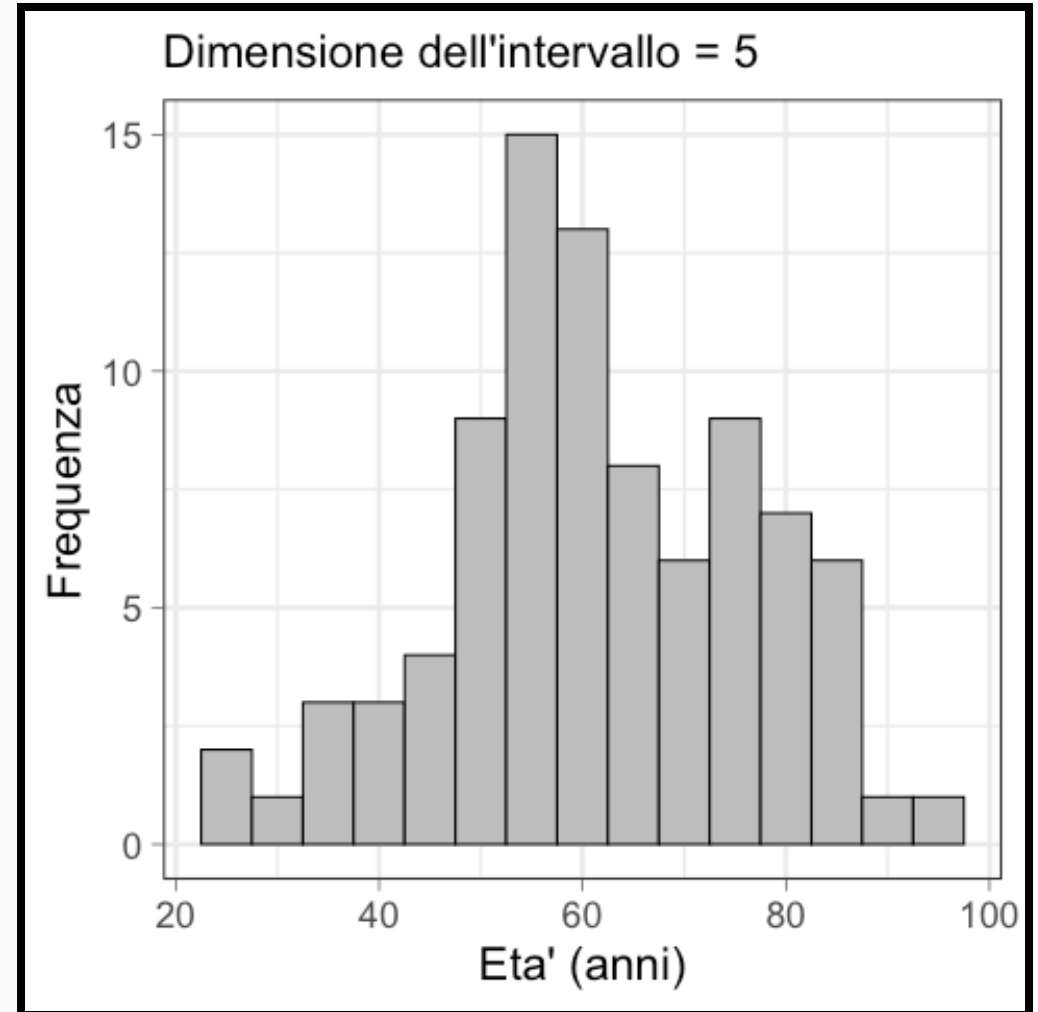
# Istogramma

Visconti A., et al., *Total serum N-glycans associate with response to immune checkpoint inhibition therapy and survival in patients with advanced melanoma*, BMC Cancer, 2023 doi:10.1186/s12885-023-10511-3



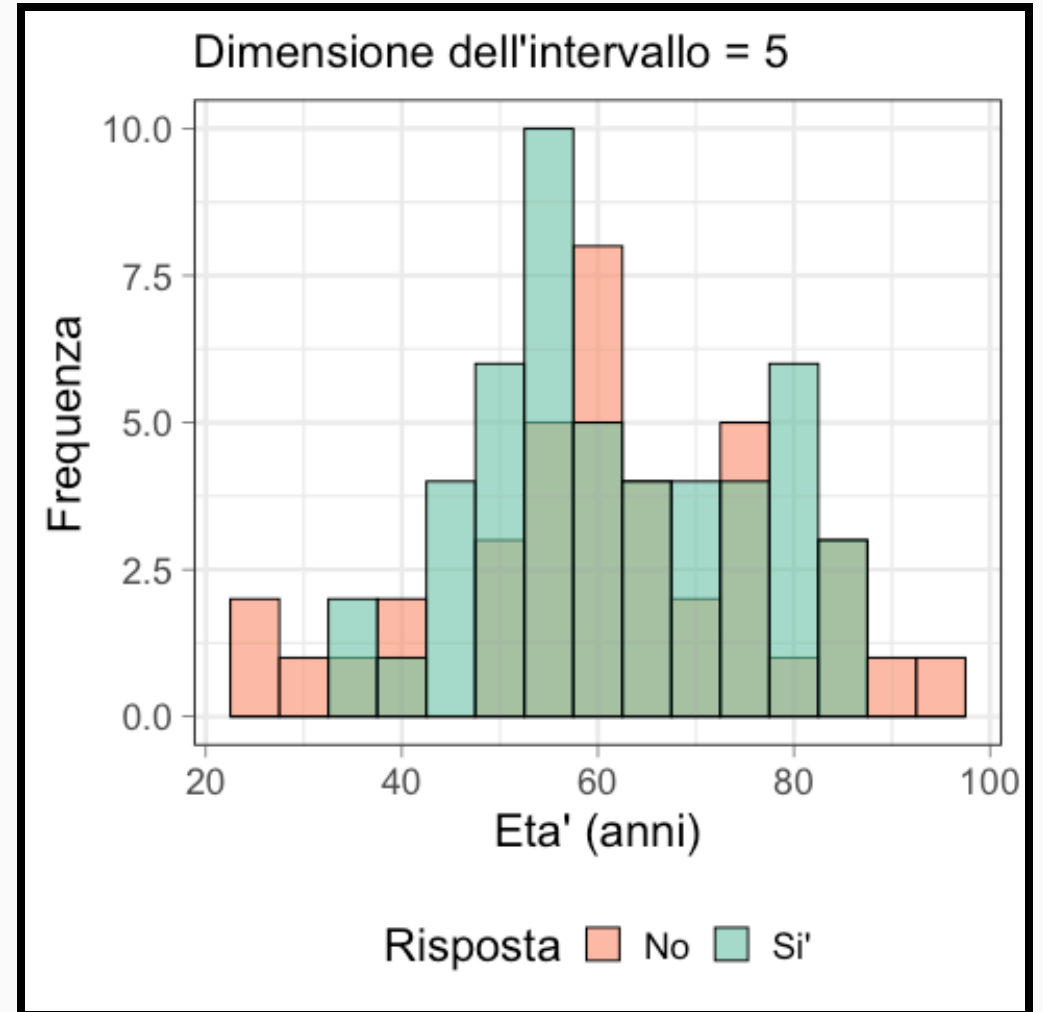
# Istogramma

Visconti A., et al., *Total serum N-glycans associate with response to immune checkpoint inhibition therapy and survival in patients with advanced melanoma*, BMC Cancer, 2023 doi:10.1186/s12885-023-10511-3



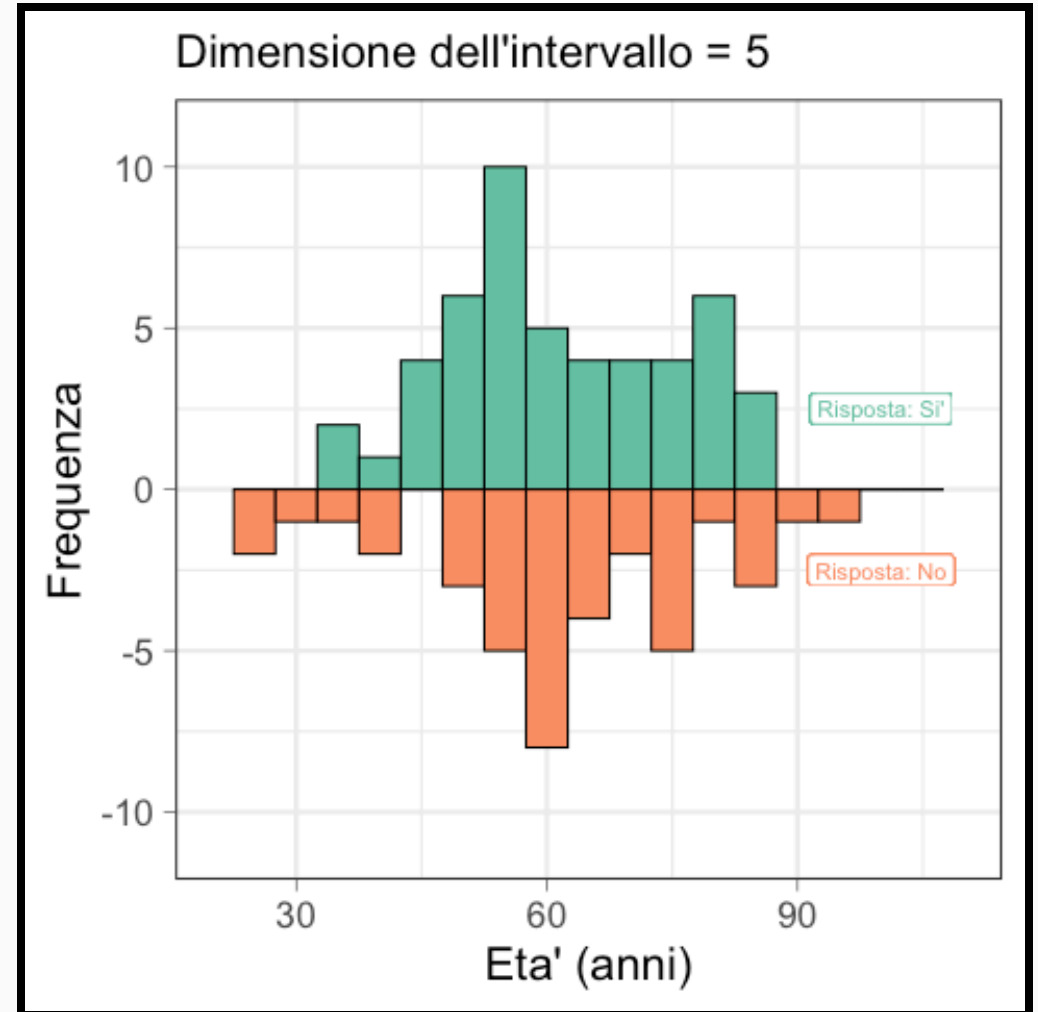
# Istogramma

Visconti A., et al., *Total serum N-glycans associate with response to immune checkpoint inhibition therapy and survival in patients with advanced melanoma*, BMC Cancer, 2023 doi:10.1186/s12885-023-10511-3

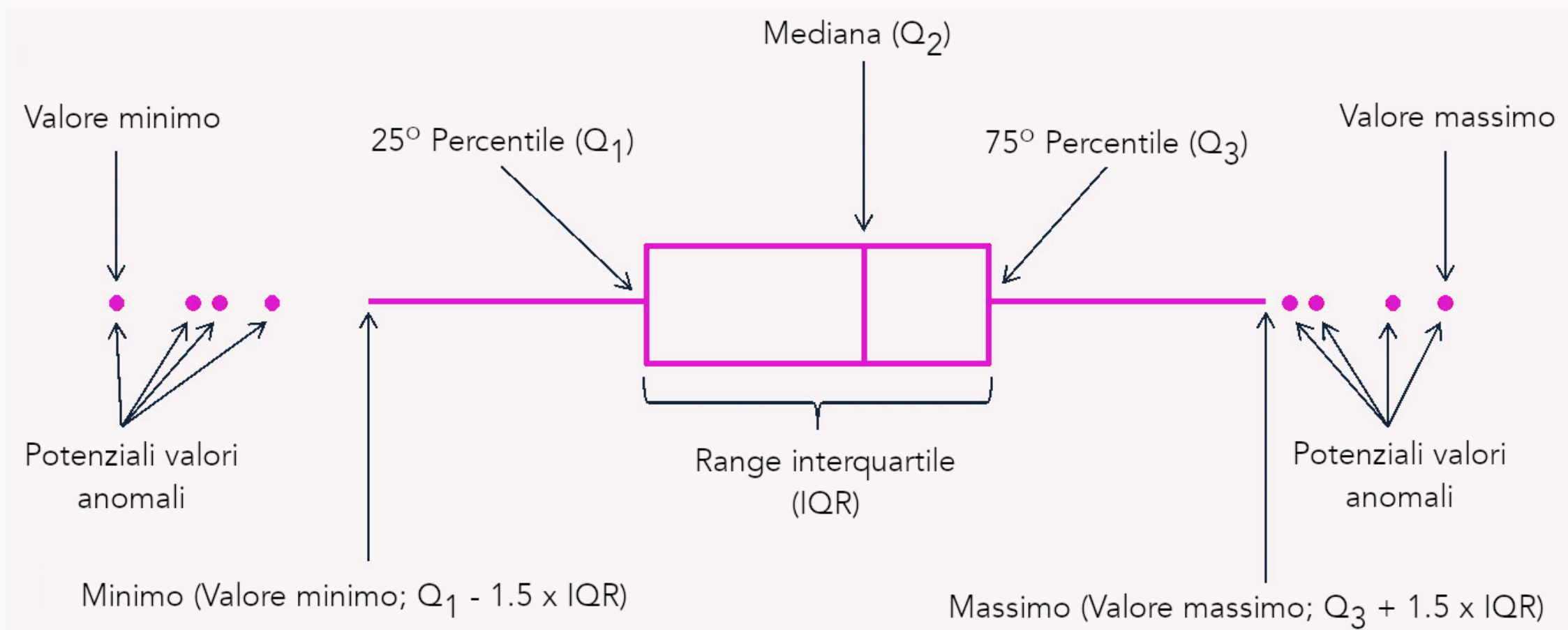


# Miami plot/Mirror histogram

Visconti A., et al., *Total serum N-glycans associate with response to immune checkpoint inhibition therapy and survival in patients with advanced melanoma*, BMC Cancer, 2023 doi:10.1186/s12885-023-10511-3



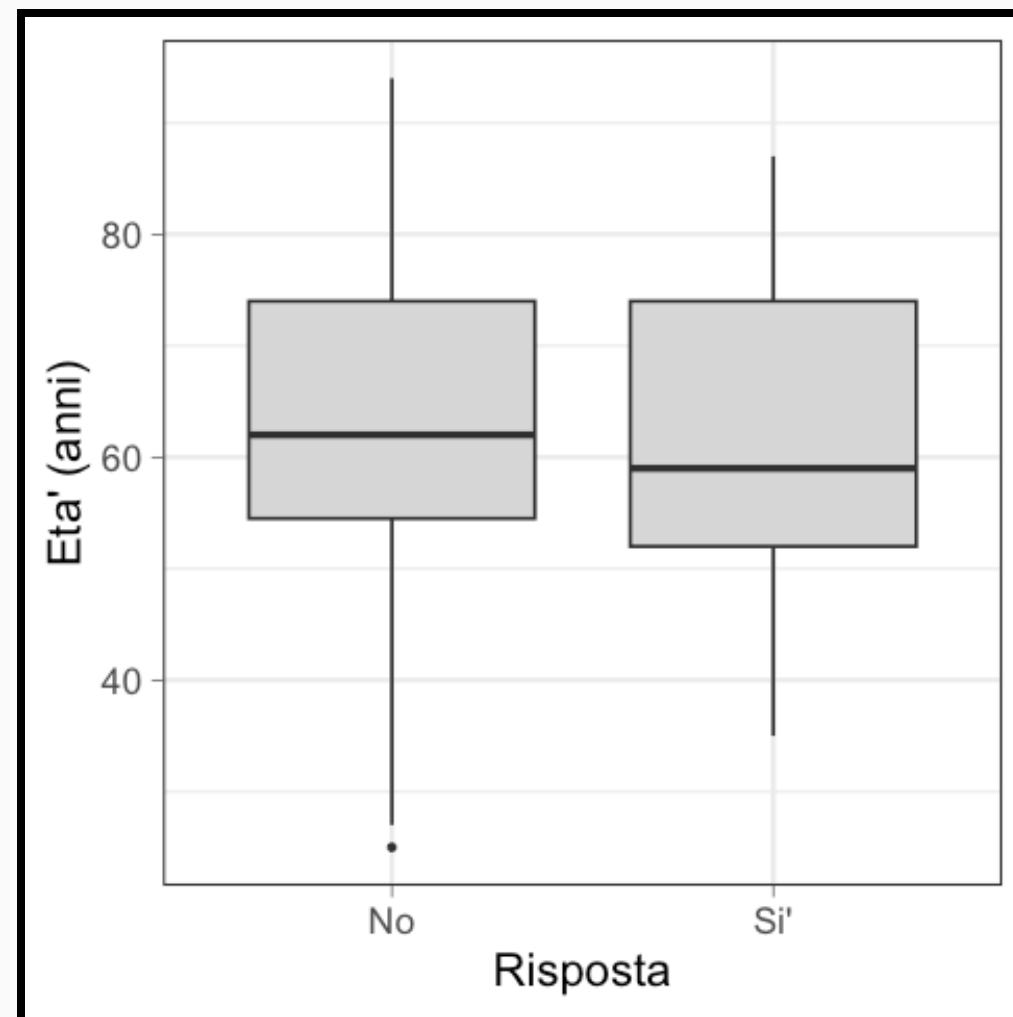
# Boxplot



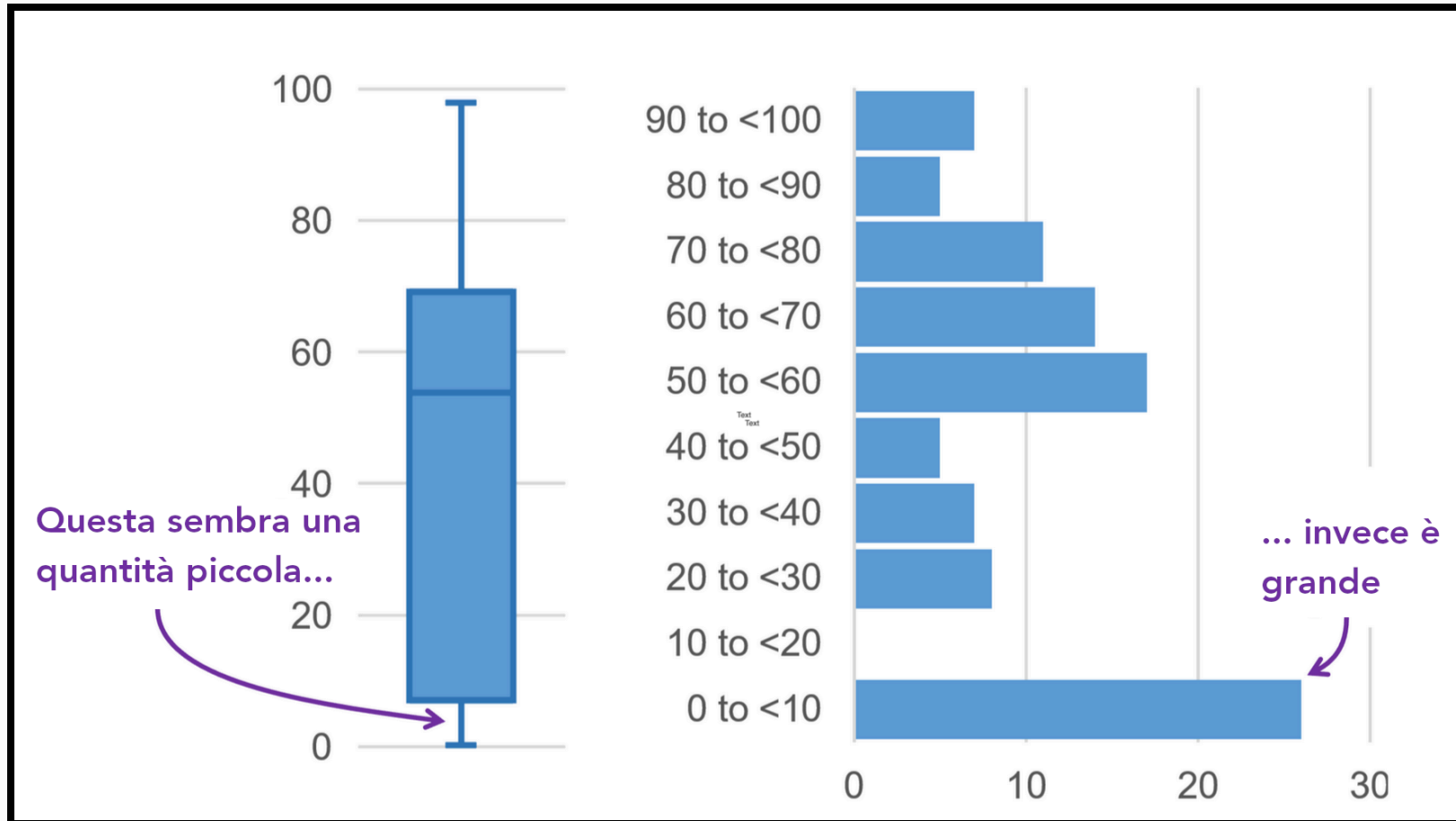


# Boxplot

Visconti A., *et al.*, Total serum *N*-glycans associate with response to immune checkpoint inhibition therapy and survival in patients with advanced melanoma, BMC Cancer, 2023 doi:10.1186/s12885-023-10511-3

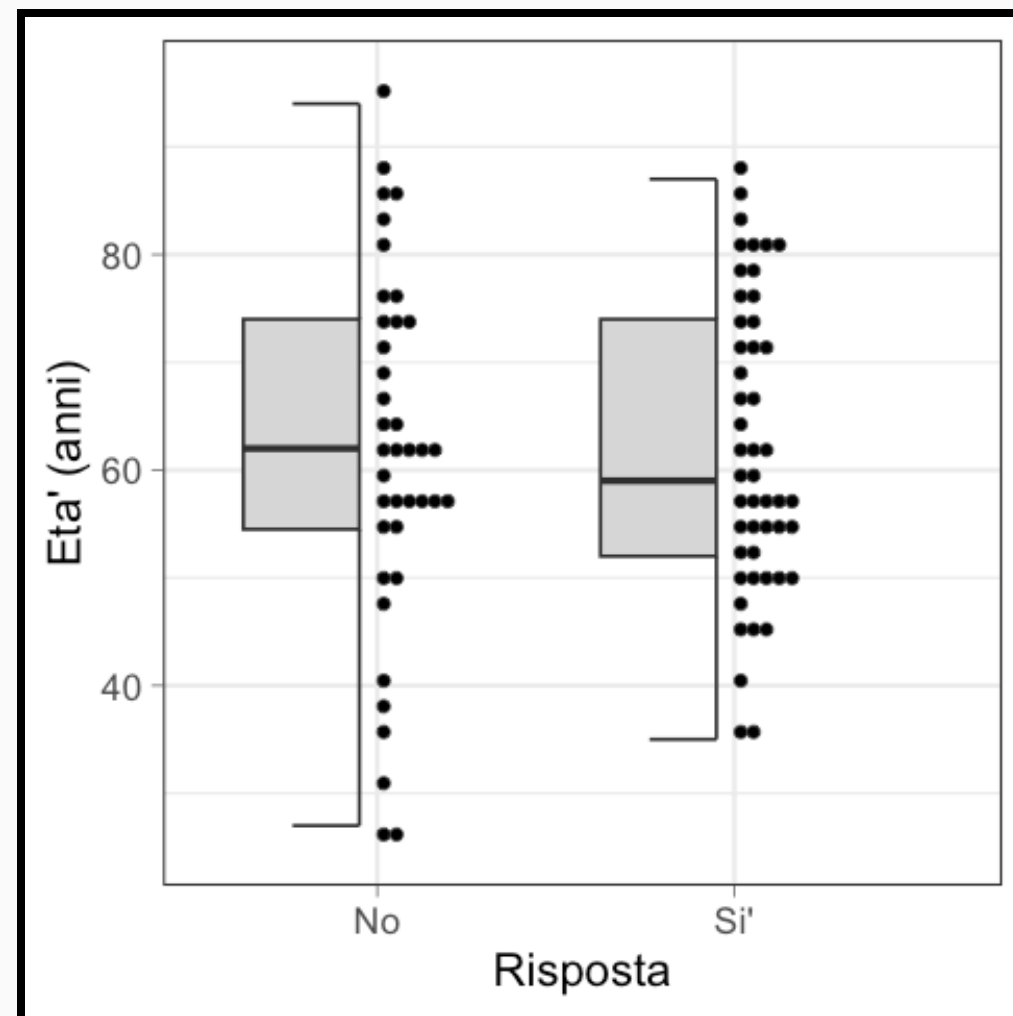


# Boxplot

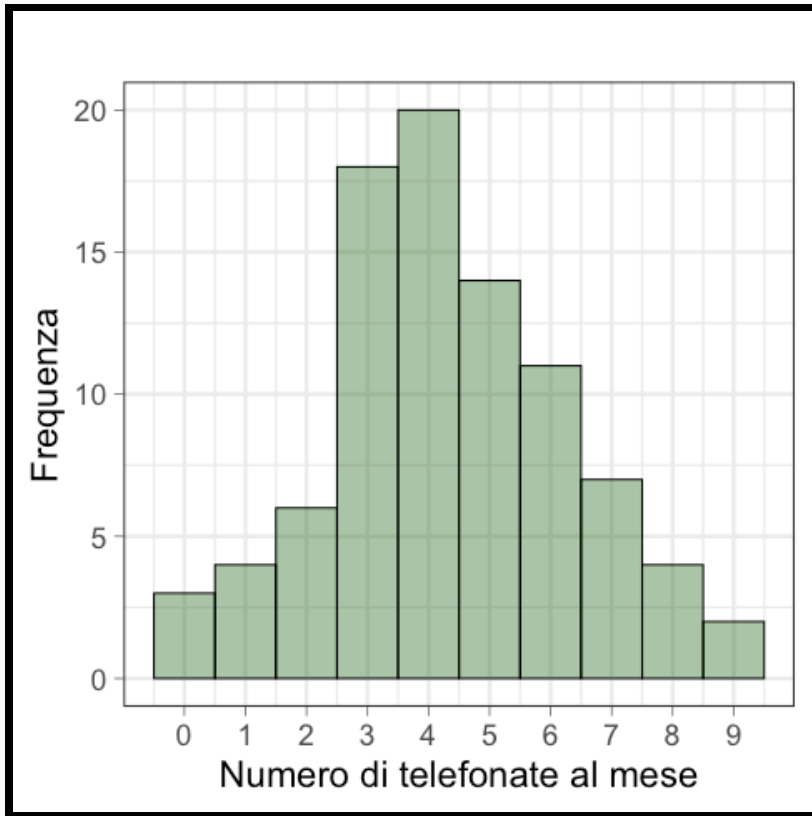


# Boxplot

Visconti A., et al., *Total serum N-glycans associate with response to immune checkpoint inhibition therapy and survival in patients with advanced melanoma*, BMC Cancer, 2023 doi:10.1186/s12885-023-10511-3



# Esercizio #11

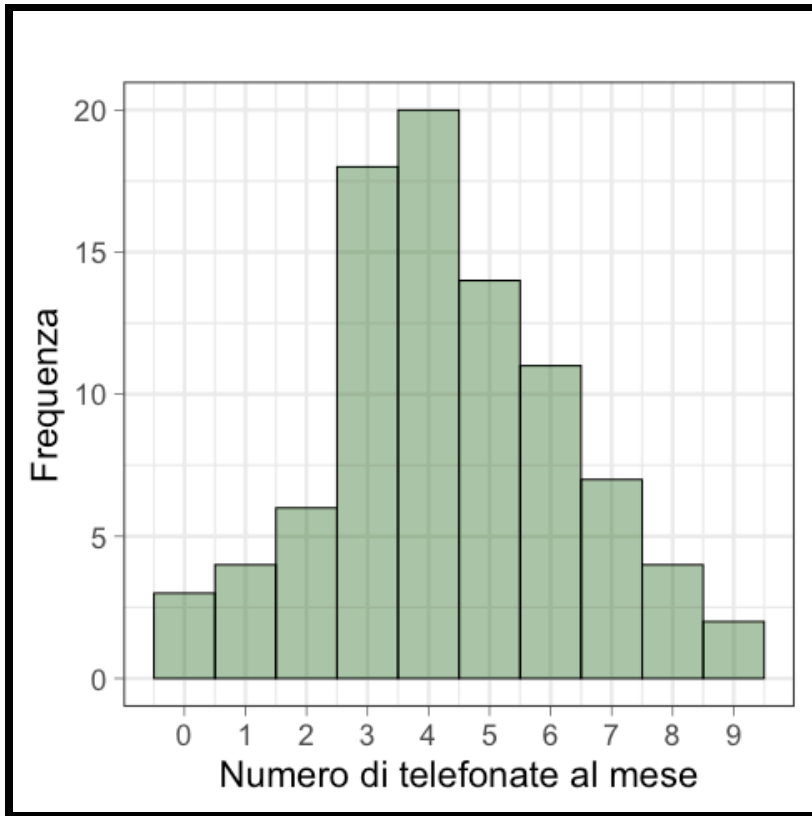


Questi dati sono stati raccolti intervistando 89 studenti universitari fuori sede

Qual è il modo migliore per descriverli?

- a) La media perché i dati sono numerici discreti e la distribuzione è abbastanza simmetrica
- b) La mediana perché i dati sono numerici discreti e la distribuzione è abbastanza simmetrica
- c) La moda perché i dati sono numerici discreti e la distribuzione è abbastanza simmetrica

# Esercizio #11 -- Soluzione

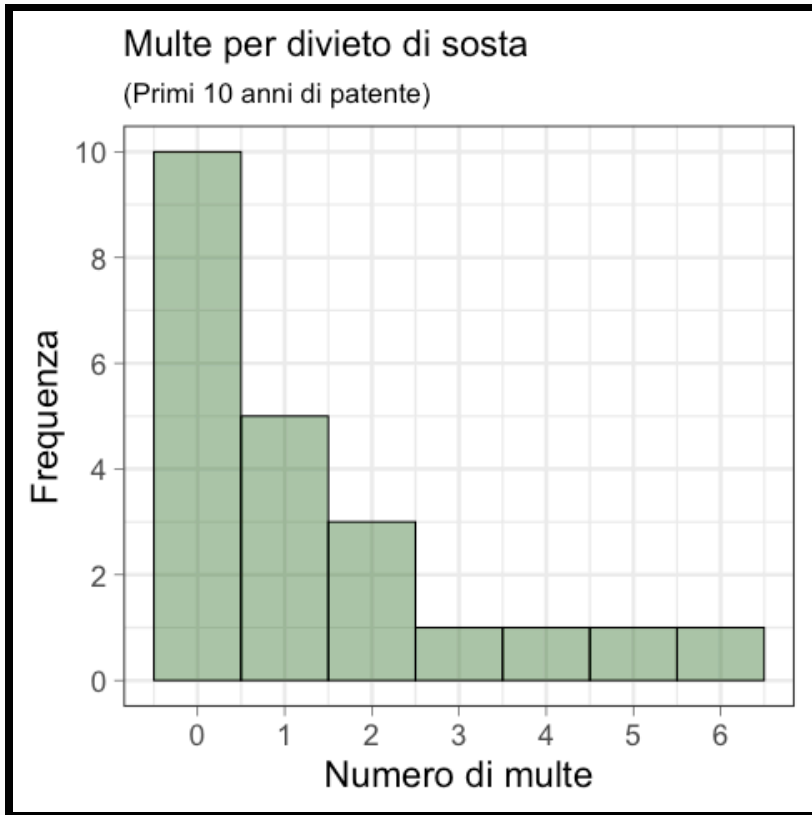


? Questi dati sono stati raccolti intervistando 89 studenti universitari fuori sede

Qual è il modo migliore per descriverli?

- a) La media perché i dati sono numerici discreti e la distribuzione è abbastanza simmetrica ✓
- b) La mediana perché i dati sono numerici discreti e la distribuzione è abbastanza simmetrica
- c) La moda perché i dati sono numerici discreti e la distribuzione è abbastanza simmetrica

# Esercizio #12

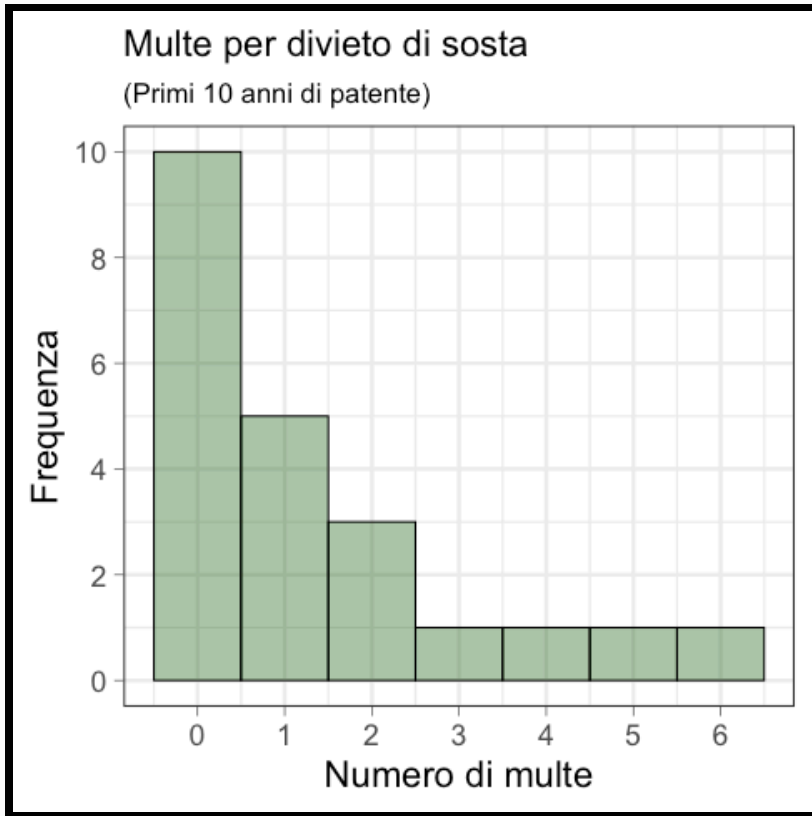


Questi dati sono stati raccolti intervistando 22 automobilisti

Qual è il modo migliore per descriverli?

- a) La media perché i dati sono numerici discreti e la distribuzione è molto asimmetrica
- b) La mediana perché i dati sono numerici discreti e la distribuzione è molto asimmetrica
- c) La moda perché i dati sono numerici discreti e la distribuzione è molto asimmetrica

# Esercizio #12 -- Soluzione

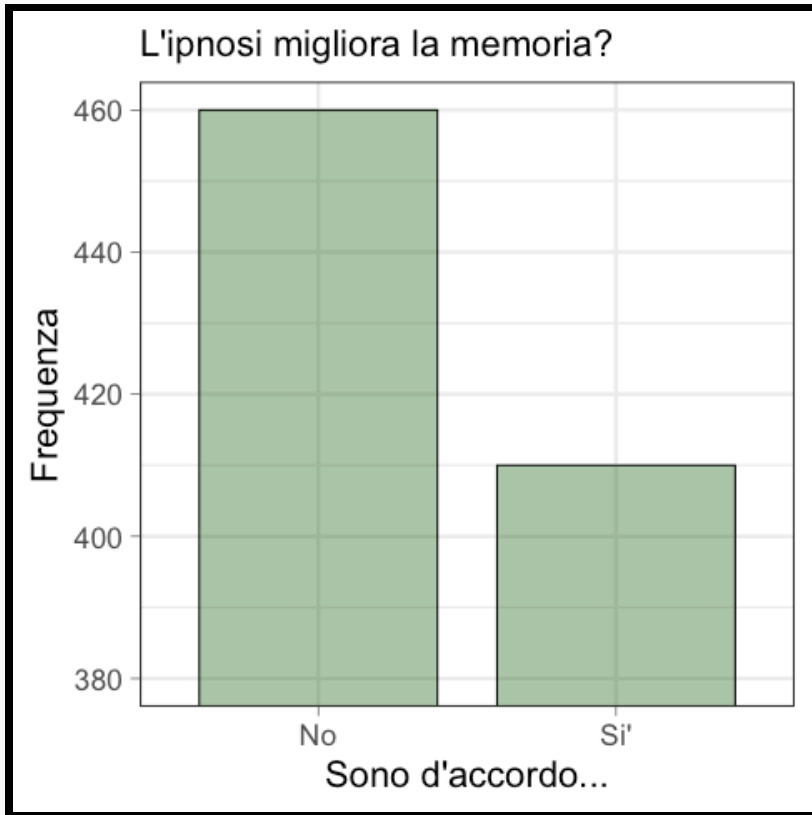


? Questi dati sono stati raccolti intervistando 22 automobilisti

Qual è il modo migliore per descriverli?

- a) La media perché i dati sono numerici discreti e la distribuzione è molto asimmetrica
- b) La mediana perché i dati sono numerici discreti e la distribuzione è molto asimmetrica ✓
- c) La moda perché i dati sono numerici discreti e la distribuzione è molto asimmetrica

# Esercizio #13



Questi dati sono stati raccolti intervistando 870 psicologi

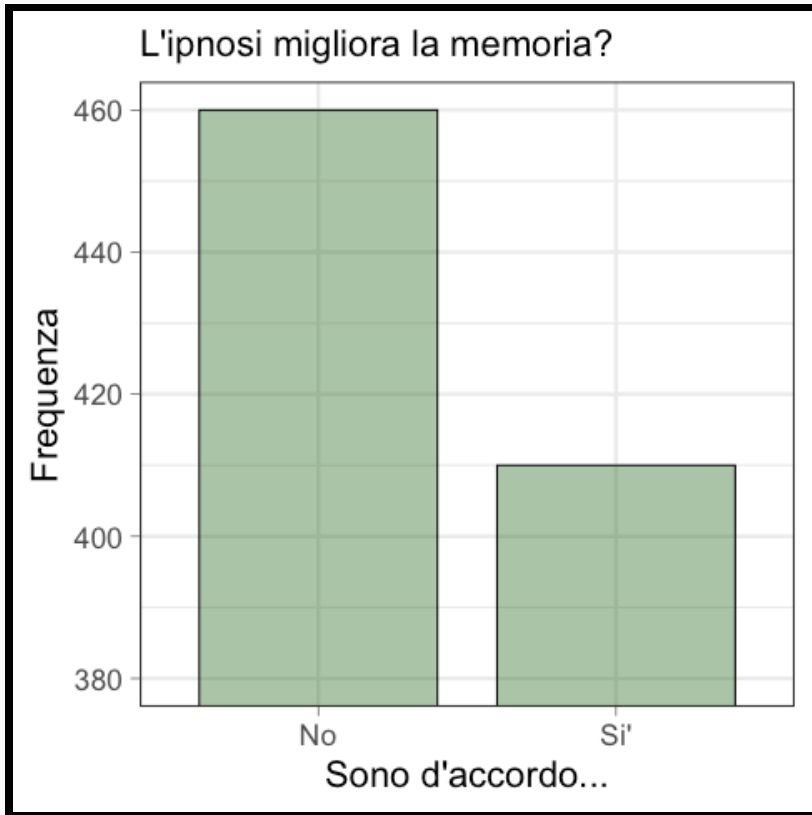
Qual è il modo migliore per descriverli?

- a) La media perché i dati sono categorici e la distribuzione è asimmetrica
- b) La mediana perché i dati sono categorici e la distribuzione è asimmetrica
- c) La moda perché i dati sono categorici

00:30



# Esercizio #13 -- Soluzione

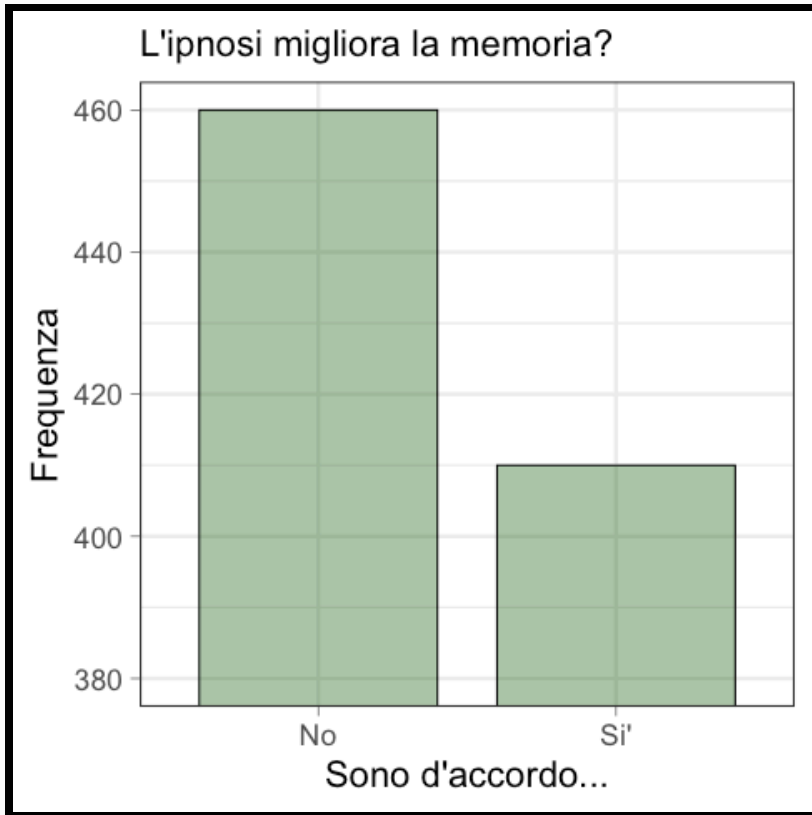


? Questi dati sono stati raccolti intervistando 870 psicologi

Qual è il modo migliore per descriverli?

- a) La media perché i dati sono categorici e la distribuzione è asimmetrica
- b) La mediana perché i dati sono categorici e la distribuzione è asimmetrica
- c) La moda perché i dati sono categorici ☒

# Esercizio #14



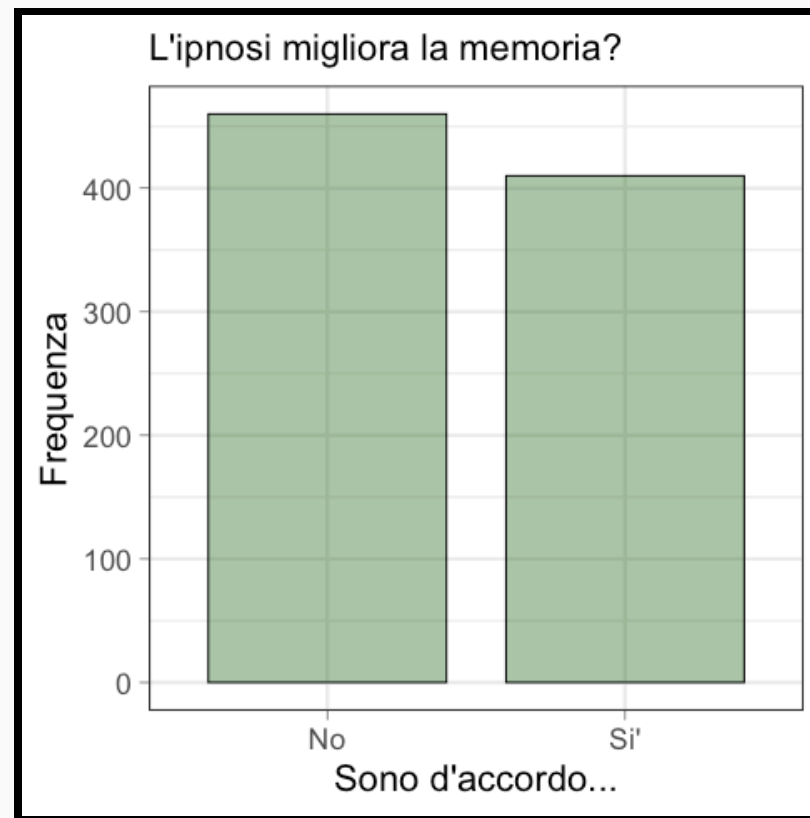
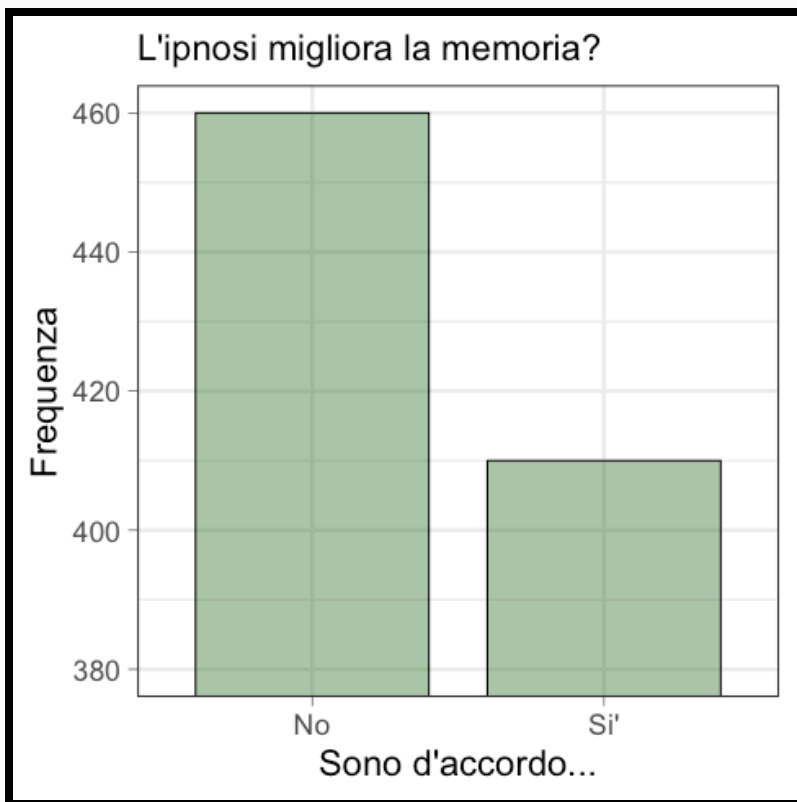
Questi dati sono stati raccolti intervistando 870 psicologi

La rappresentazione usata è corretta?

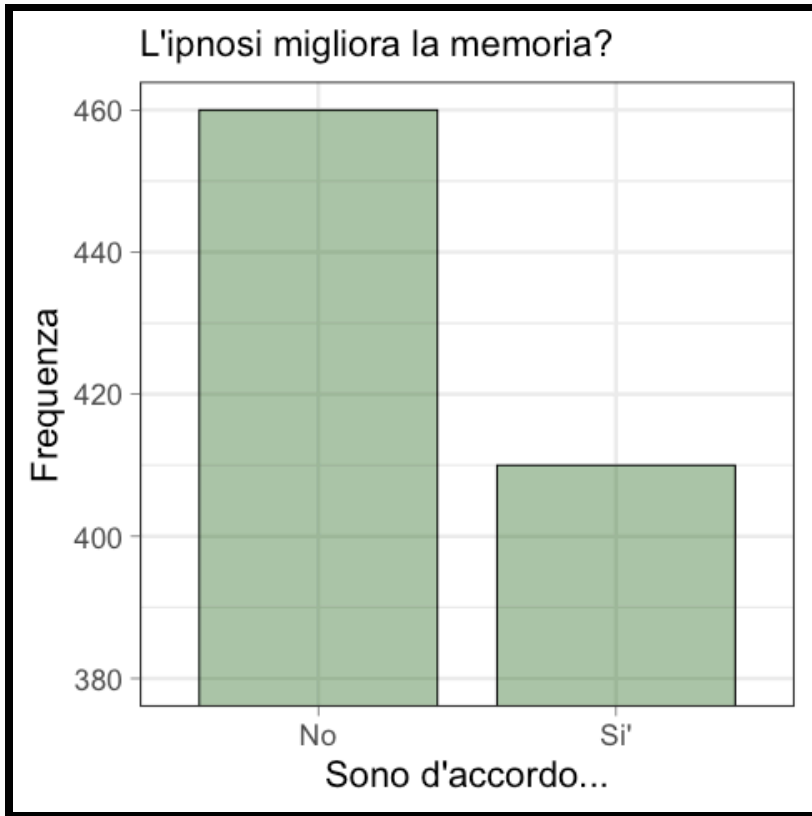
- a) Sì
- b) No

# Esercizio #14

? Quale rappresentazione grafica è corretta?




# Esercizio #14 -- Soluzione



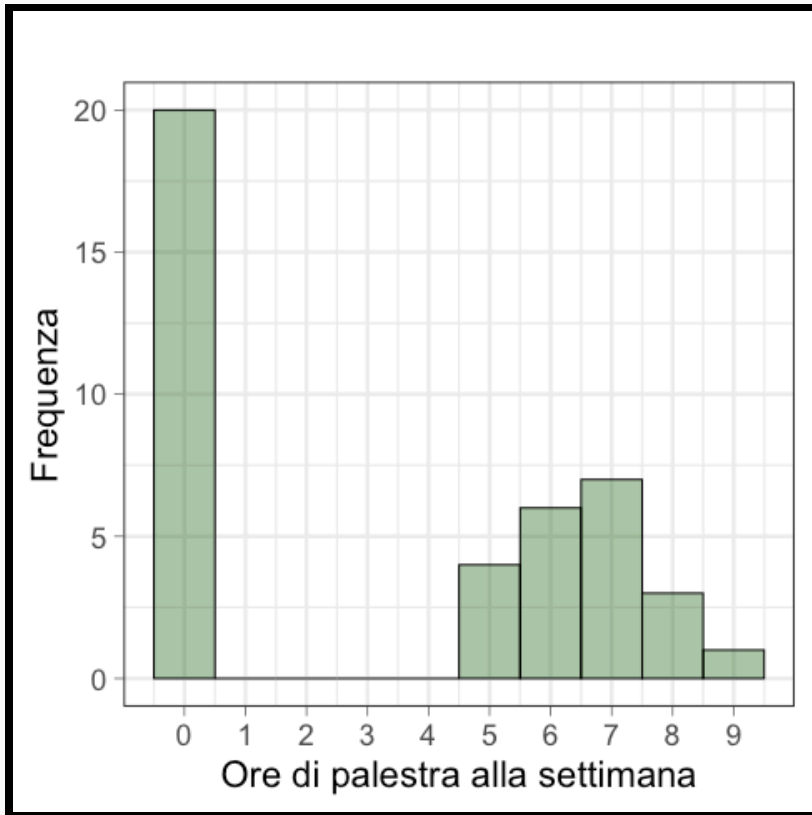
Questi dati sono stati raccolti intervistando 870 psicologi

La rappresentazione usata è corretta?

a) Sì

b) No 

# Esercizio #15

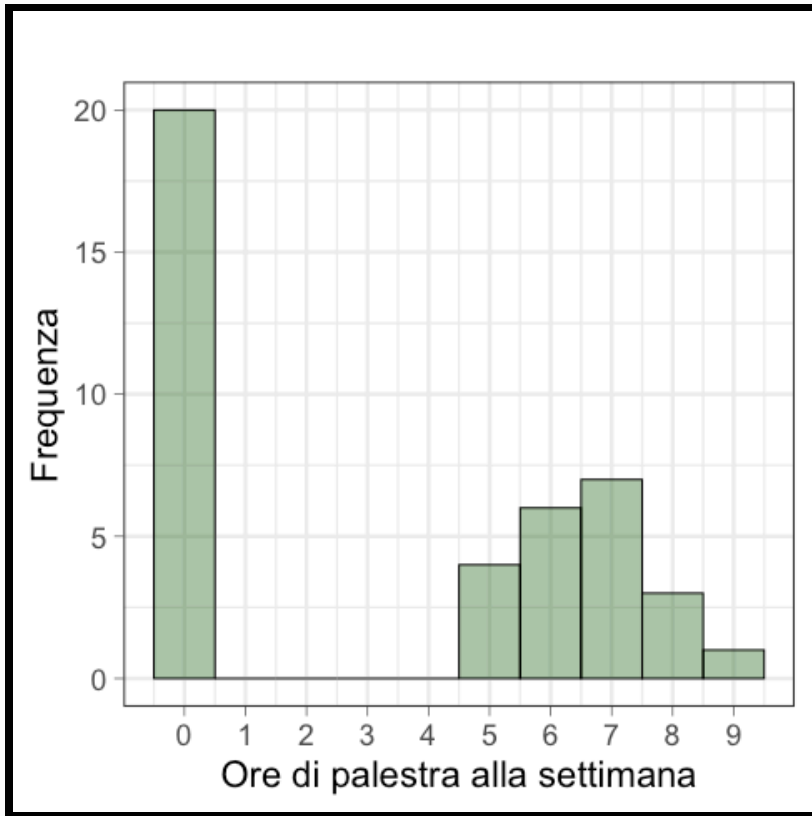


Questi dati sono stati raccolti intervistando 41 genitori in un parco giochi.

Qual è il modo migliore per descriverli?

- a) Gli intervistati spendono tra le 0 e le 9 ore in palestra, con una media di  $3.4 \pm 3.4$  ore (mediana: 5 ore; moda: 0 ore).
- b) Circa la metà degli intervistati ha riportato di non essere andata in palestra. I rimanenti spendono in palestra tra le 5 e le 9 ore, con una media di  $6.6 \pm 1.1$  ore (mediana: 7 ore)

# Esercizio #15 -- Soluzione



? Questi dati sono stati raccolti intervistando 41 genitori in un parco giochi.

Qual è il modo migliore per descriverli?

- a) Gli intervistati spendono tra le 0 e le 9 ore in palestra, con una media di  $3.4 \pm 3.4$  ore (mediana: 5 ore; moda: 0 ore).
- b) Circa la metà degli intervistati ha riportato di non essere andata in palestra. I rimanenti spendono in palestra tra le 5 e le 9 ore, con una media di  $6.6 \pm 1.1$  ore (mediana: 7 ore) ✓

## **Esercizio #16**

Quanti partner (etero)sessuali le persone in Gran Bretagna riferiscono di aver avuto nella loro vita?

# Esercizio #16

Quanti partner (etero)sessuali le persone in Gran Bretagna riferiscono di aver avuto nella loro vita?

? Cosa ci dicono  
queste statistiche?

	Uomini 35-44	Donne 35-44
Moda	1	1
Range	0-500	0-550
Media	14.3	8.5
SD	24.2	19.7
Mediana	8	5
IQR	4-18	3-10

Think

01:00



# Esercizio #16

Quanti partner (etero)sessuali le persone in Gran Bretagna riferiscono di aver avuto nella loro vita?

? Cosa ci dicono  
queste statistiche?

	Uomini 35-44	Donne 35-44
Moda	1	1
Range	0-500	0-550
Media	14.3	8.5
SD	24.2	19.7
Mediana	8	5
IQR	4-18	3-10

Pair

02:00

# Esercizio #16

Quanti partner (etero)sessuali le persone in Gran Bretagna riferiscono di aver avuto nella loro vita?

? Cosa ci dicono queste statistiche?

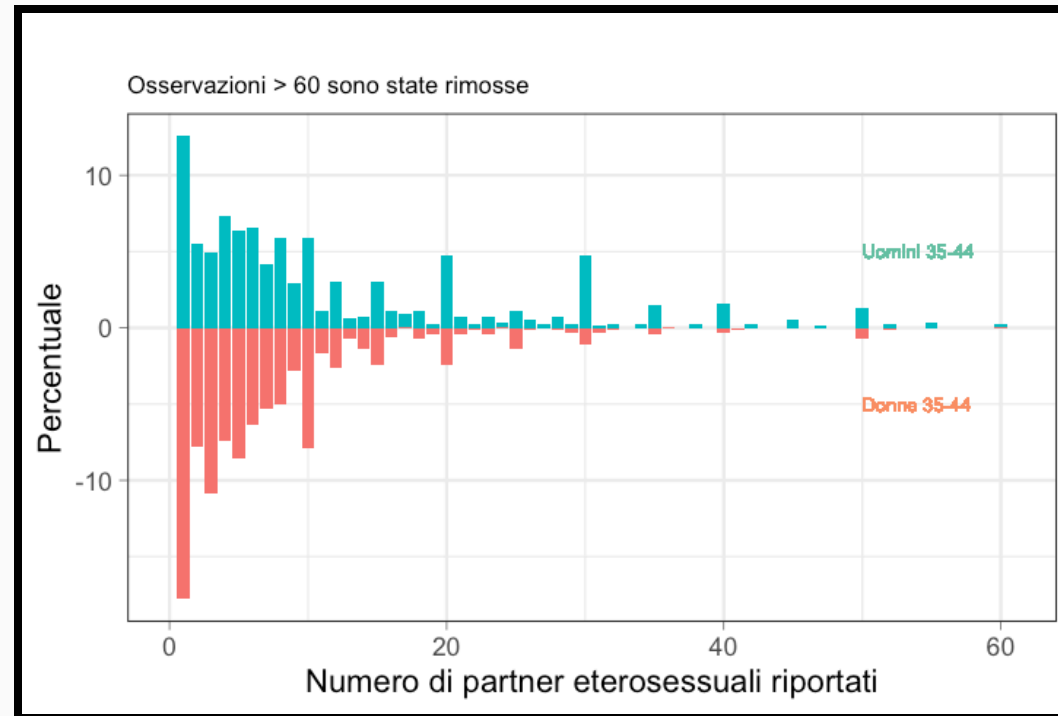
	Uomini 35-44	Donne 35-44
Moda	1	1
Range	0-500	0-550
Media	14.3	8.5
SD	24.2	19.7
Mediana	8	5
IQR	4-18	3-10

Share

05:00

# Esercizio #16 (bis)

? Il grafico della distribuzione conferma quello che abbiamo detto?  
Aggiunge informazione?



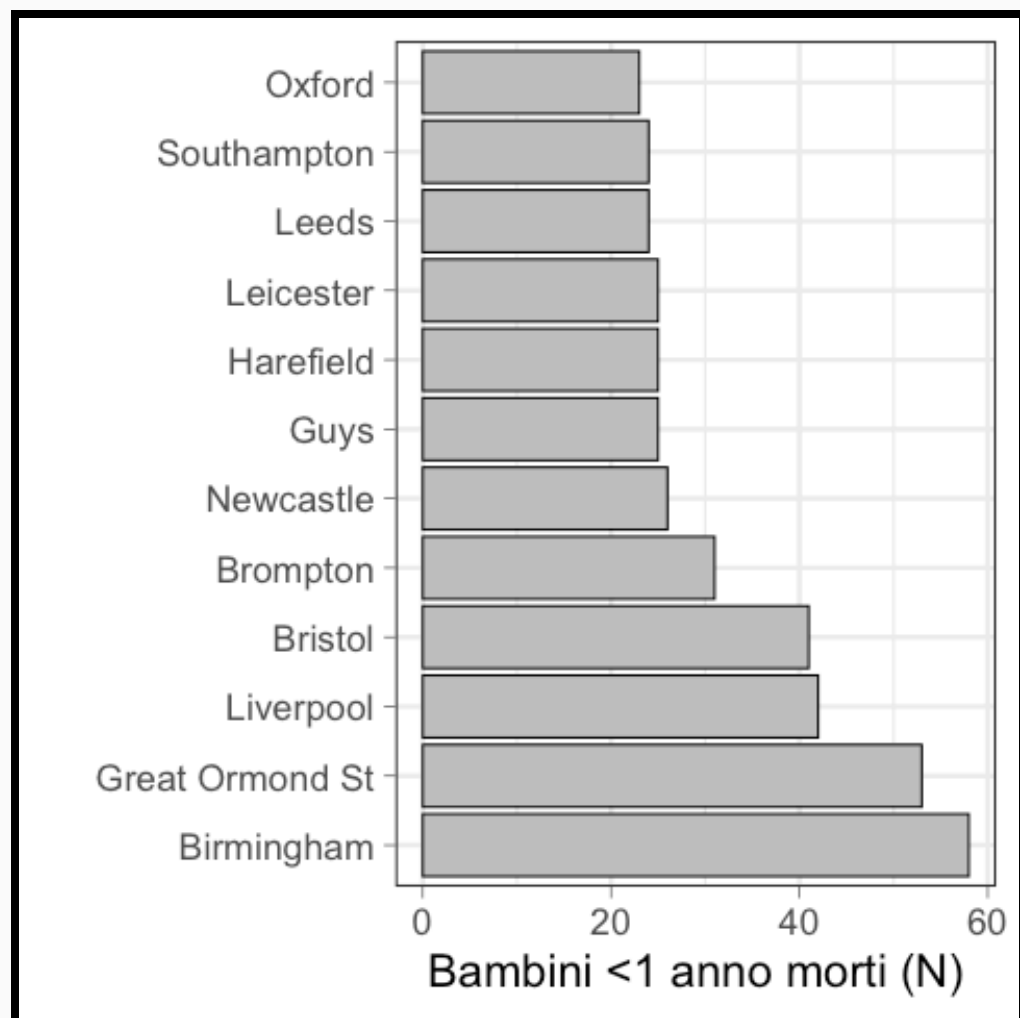
# La relazione (lineare) tra due variabili numeriche

Cosa è successo ai bambini sottoposti a interventi cardiocirurgici in alcuni ospedali britannici tra il 1991 e il 1995?

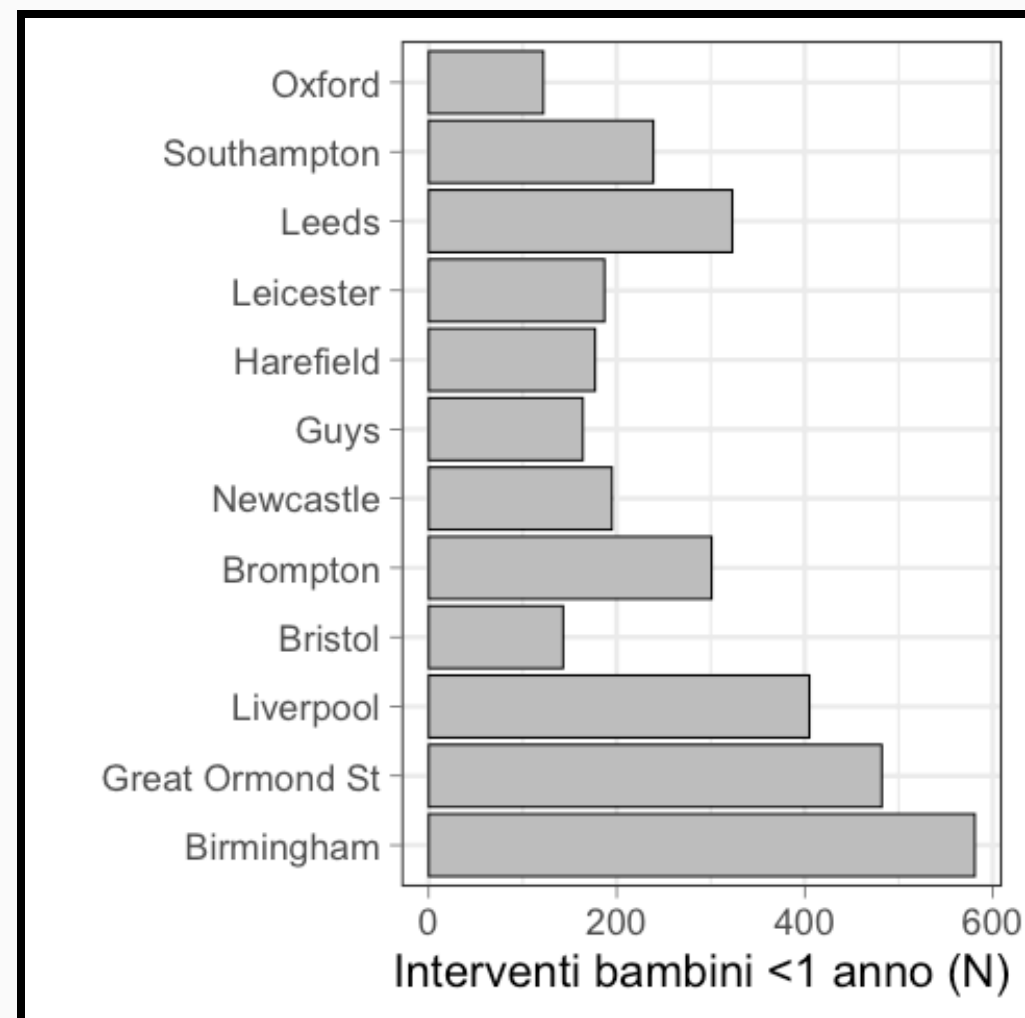
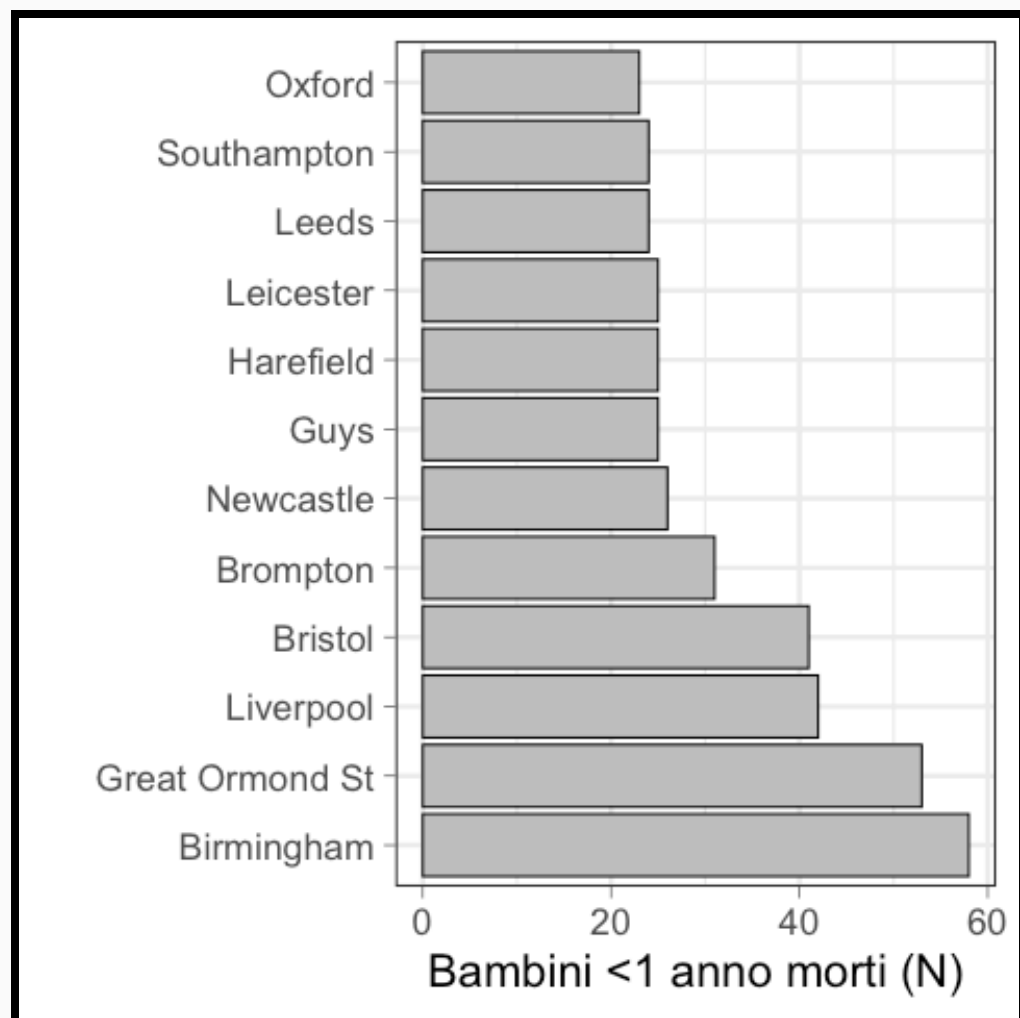
Ospedale	Interventi	Sopravvissuti (N)	Morti (N)	Sopravvissuti (%)	Morti (%)
Bristol	143	102	41	71.3	28.7
Leicester	187	162	25	86.6	13.4
Leeds	323	299	24	92.6	7.4
Oxford	122	99	23	81.1	18.9
Guys	164	139	25	84.8	15.2
Liverpool	405	363	42	89.6	10.4
Southampton	239	215	24	90.0	10.0
Great Ormond St	482	429	53	89.0	11.0
Newcastle	195	169	26	86.7	13.3
Harefield	177	152	25	85.9	14.1
Birmingham	581	523	58	90.0	10.0
Brompton	301	270	31	89.7	10.3

D.J. Spiegelhalter et al., *Commissioned Analysis of Surgical Performance Using Routine Data: Lessons from the Bristol Inquiry*, 2002, Journal of the Royal Statistical Society Series A: Statistics in Society

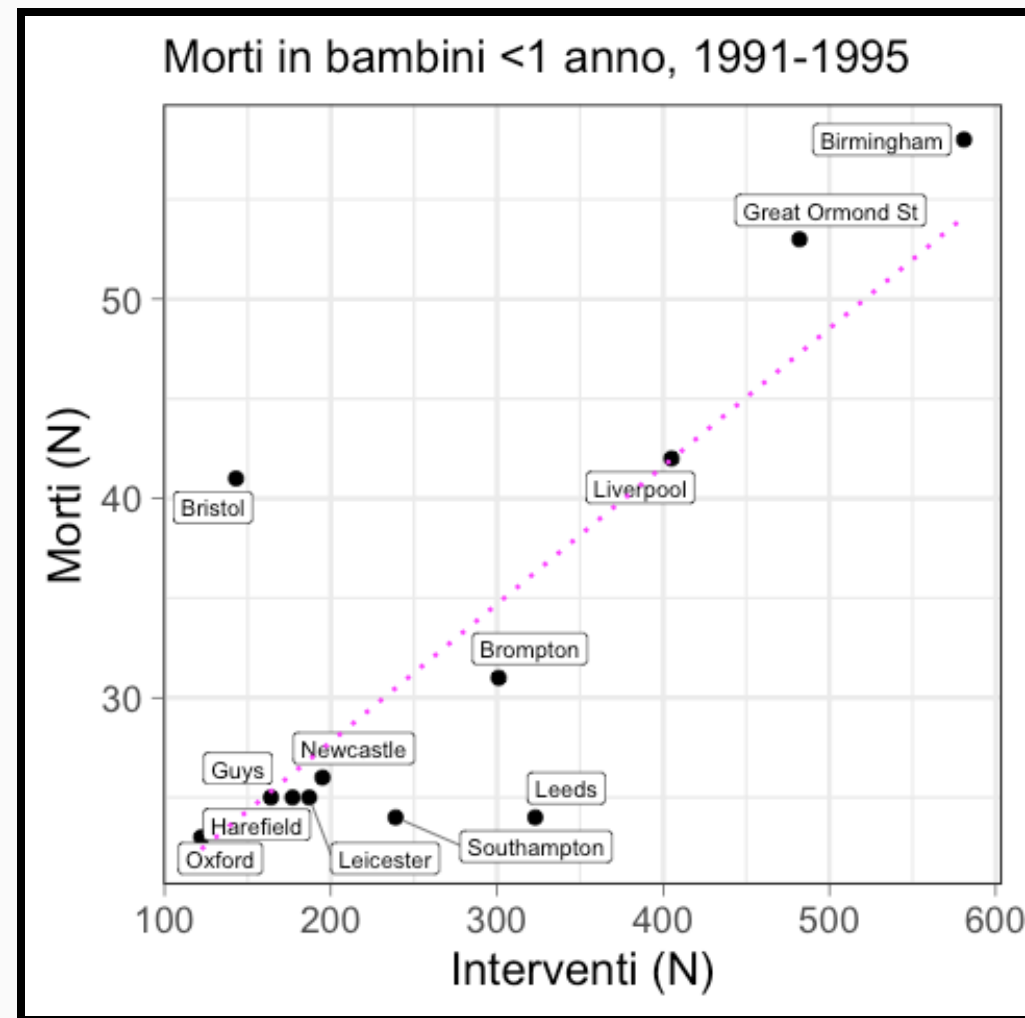
# Visualizziamo i dati



# Visualizziamo i dati



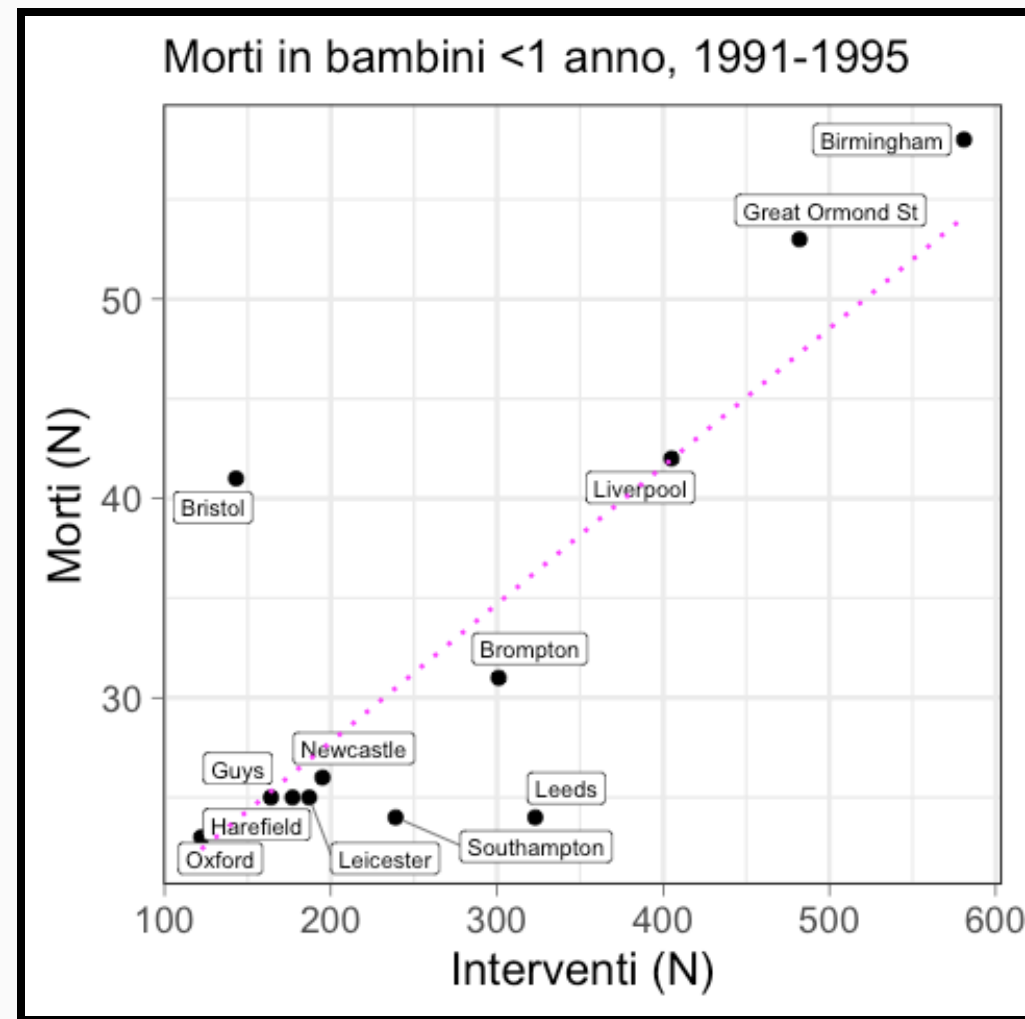
# La relazione (lineare) tra due variabili numeriche



# La relazione (lineare) tra due variabili numeriche

Indice di correlazione<sup>1</sup>

- $r = 0.82$
- $r_{\text{no Bristol}} = 0.93$



<sup>1</sup> In questo caso di Pearson (ma formulato da Galton). Un altro indice di correlazione è quello di Spearman



# Indici di correlazione

- Non indicano causalità
- Hanno un valore compreso tra  $-1$  e  $1$
- Il segno indica la direzione della relazione **lineare**
- $r^2 \times 100 = R^2$  (o coefficiente di determinazione) indica la percentuale di variabilità di una variabile che è predetta dalla variabilità dell'altra variabile  
$$R^2 = r^2 \times 100 = 0.82^2 \times 100 = 0.67 \times 100 \rightarrow 67\% \text{ della variabilità}$$

$ r $	Interpretazione
0-0.25	nessuna o poca correlazione
0.25-0.50	discreta correlazione
0.50-0.75	buona correlazione
0.75-0.99	eccellente correlazione
1	perfetta correlazione

# Esercizio #17

? Una correlazione  $r = -0.7$  indica che al crescere del valore di una variabile, il valore dell'altra variabile...

- a) cresce
- b) decresce
- c) rimane costante
- d) dipende dalle variabili

## Esercizio #17 -- Soluzione

? Una correlazione  $r = -0.7$  indica che al crescere del valore di una variabile, il valore dell'altra variabile...

- a) cresce
- b) decresce ☒
- c) rimane costante
- d) dipende dalle variabili

# Esercizio #18

? Quale dei seguenti valori di  $r$  indica la correlazione più forte?

a)  $-0.2$

b)  $+0.4$

c)  $-0.7$

d)  $+1.1$

## Esercizio #18 -- Soluzione

? Quale dei seguenti valori di  $r$  indica la correlazione più forte?

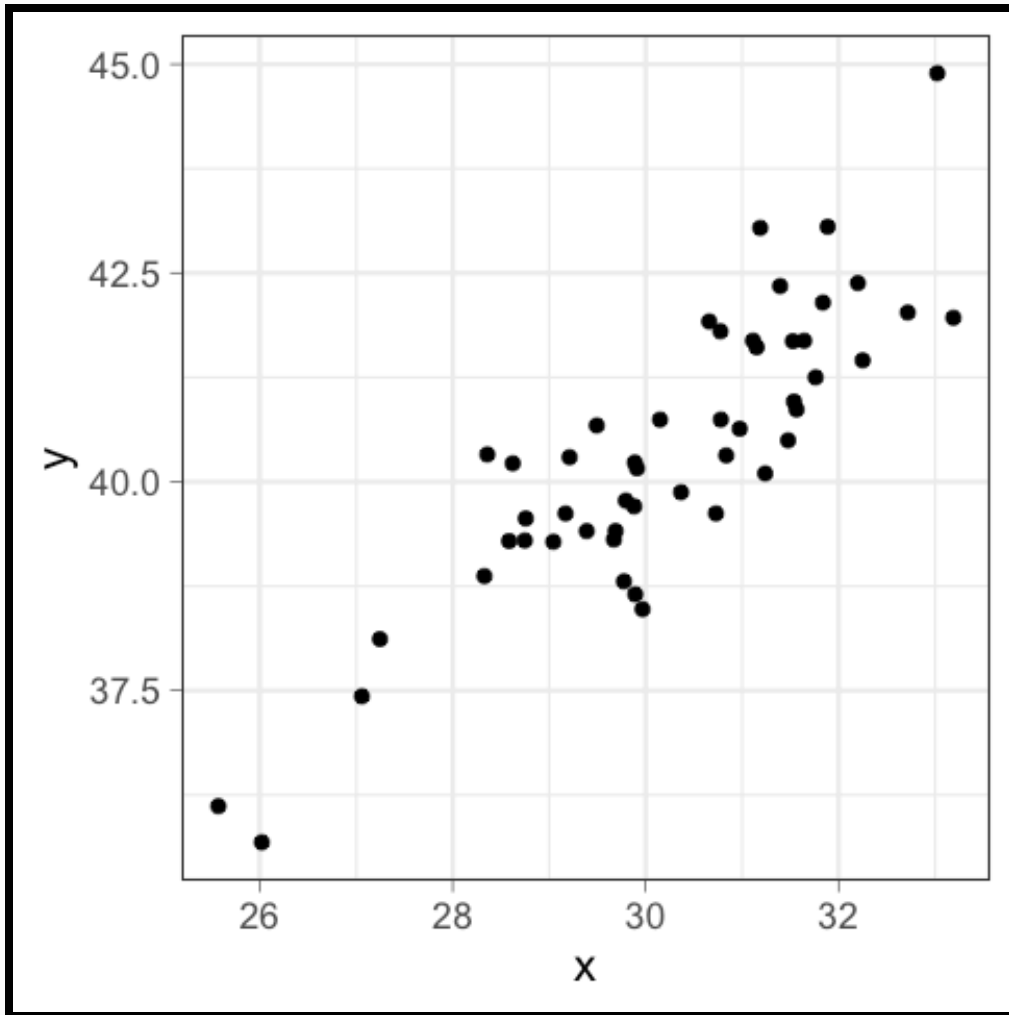
a)  $-0.2$

b)  $+0.4$

c)  $-0.7$  

d)  $+1.1$

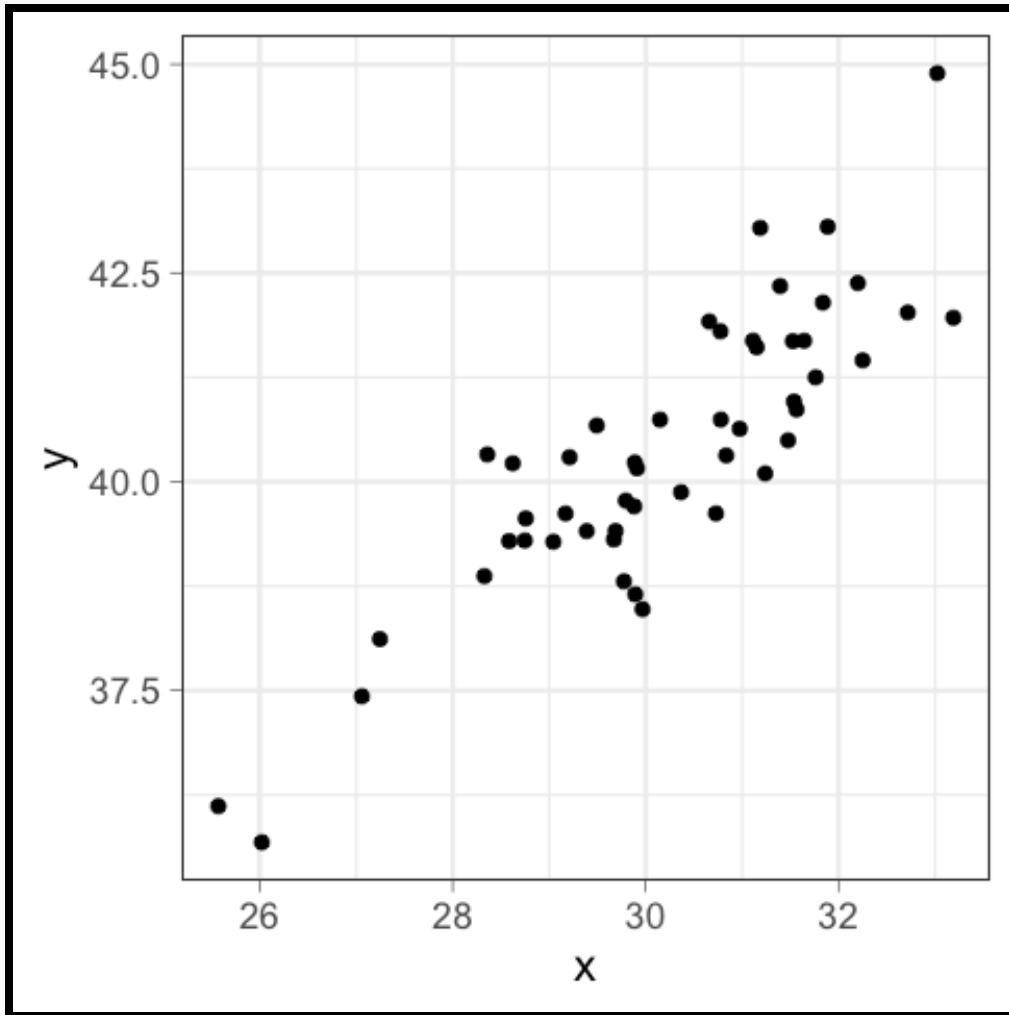
# Esercizio #19



? Osservando lo scatterplot, quale tra questi potrebbe essere un valore plausibile per l'indice di correlazione di Pearson tra le due variabili mostrate?

- a)  $r = +0.1$
- b)  $r = +0.9$
- c)  $r = -0.9$
- d) Non è calcolabile

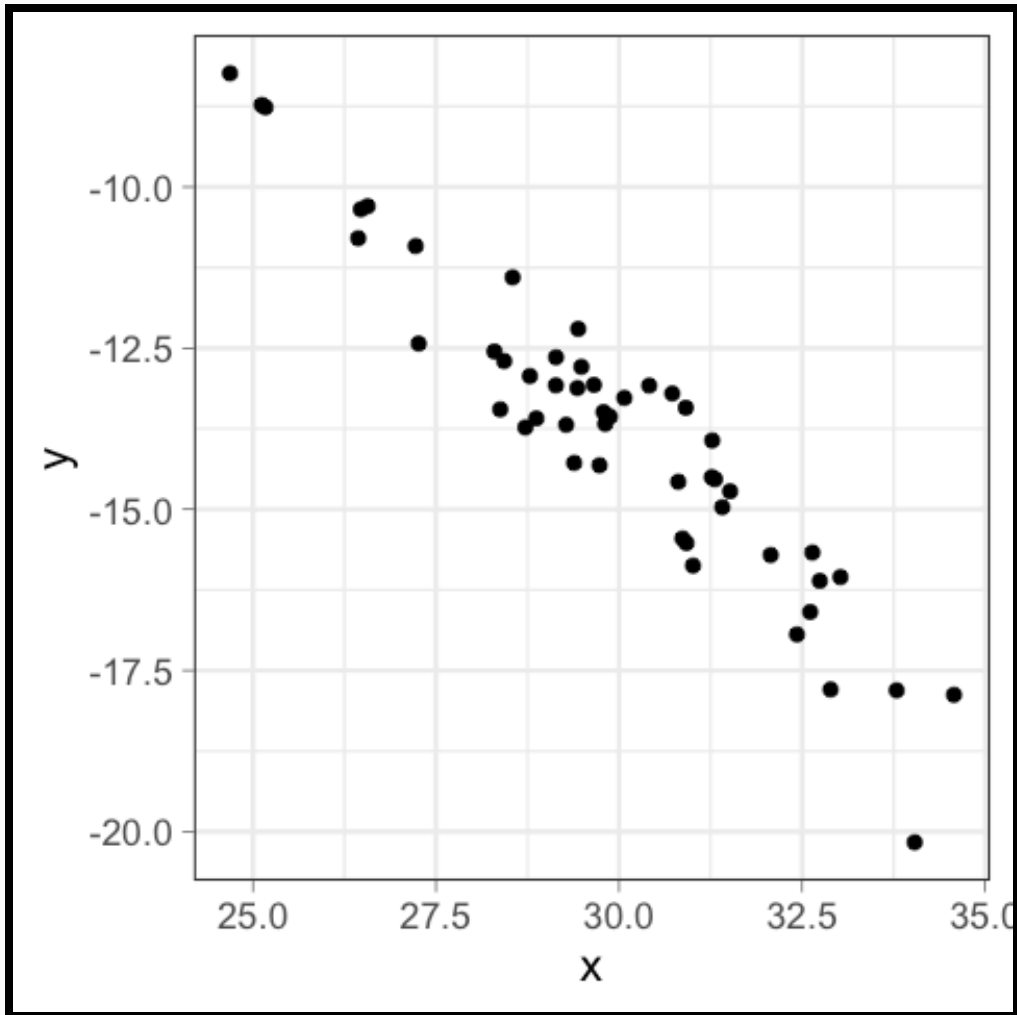
# Esercizio #19 -- Soluzione



? Osservando lo scatterplot, quale tra questi potrebbe essere un valore plausibile per l'indice di correlazione di Pearson tra le due variabili mostrate?

- a)  $r = +0.1$
- b)  $r = +0.9$  ✓
- c)  $r = -0.9$
- d) Non è calcolabile

## Esercizio #20

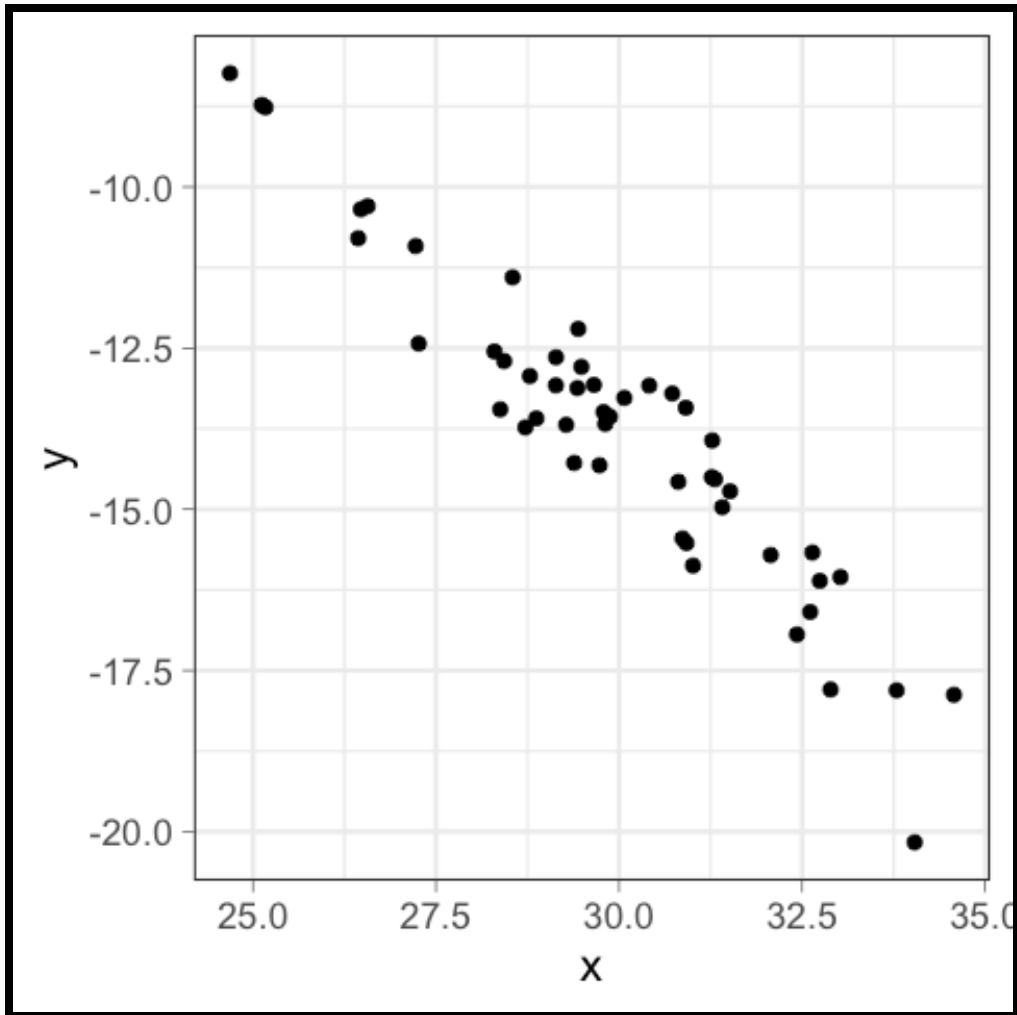


Osservando lo scatterplot, quale tra questi potrebbe essere un valore plausibile per l'indice di correlazione di Pearson tra le due variabili mostrate?

- a)  $r = +0.1$
- b)  $r = +0.9$
- c)  $r = -0.9$
- d) Non è calcolabile



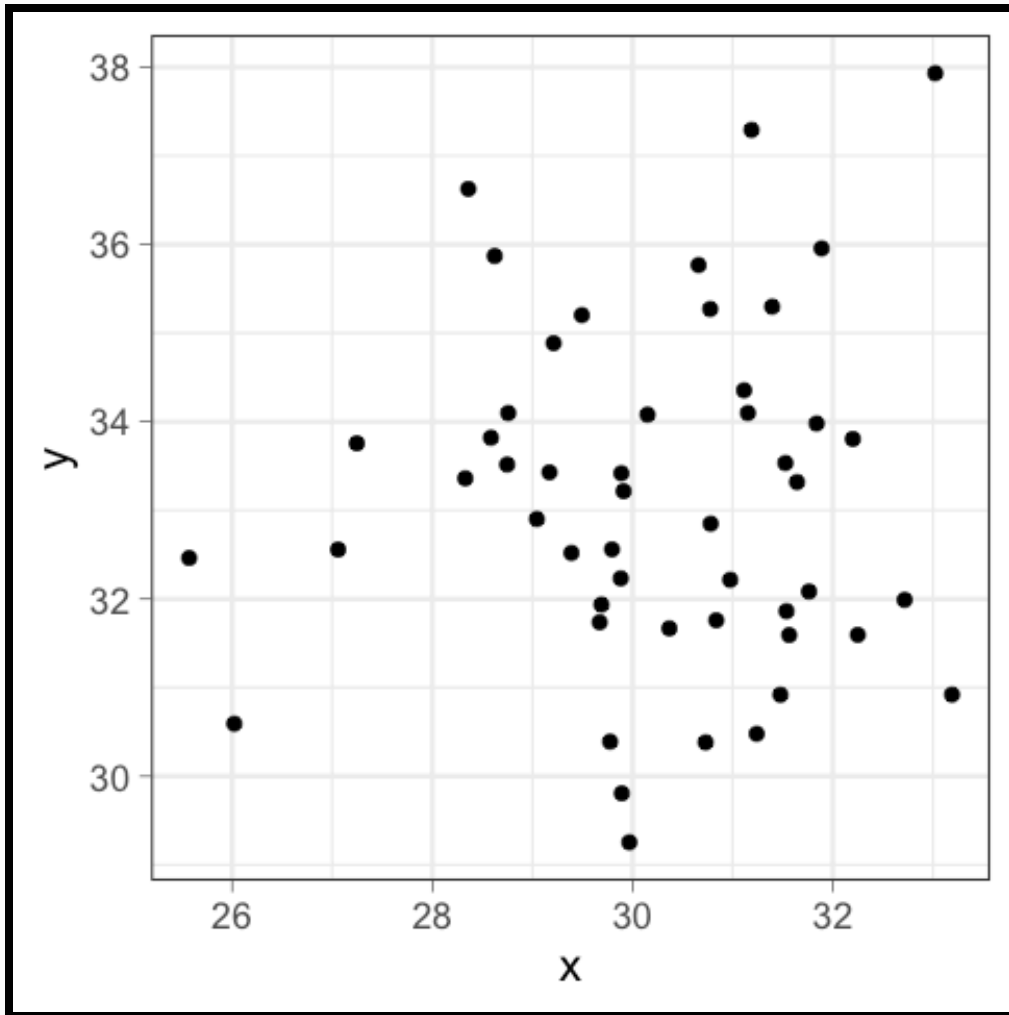
# Esercizio #20 -- Soluzione



? Osservando lo scatterplot, quale tra questi potrebbe essere un valore plausibile per l'indice di correlazione di Pearson tra le due variabili mostrate?

- a)  $r = +0.1$
- b)  $r = +0.9$
- c)  $r = -0.9$  ✓
- d) Non è calcolabile

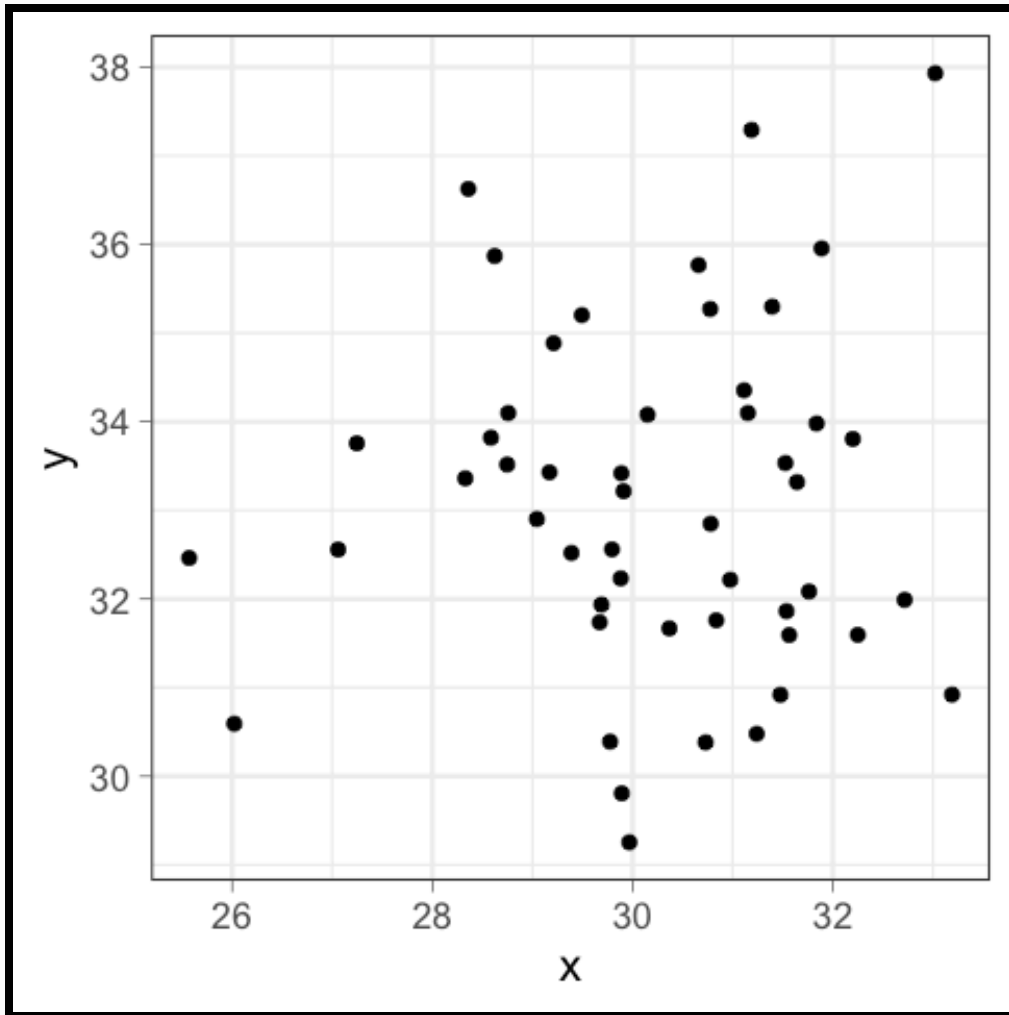
# Esercizio #21



? Osservando lo scatterplot, quale tra questi potrebbe essere un valore plausibile per l'indice di correlazione di Pearson tra le due variabili mostrate?

- a)  $r = +0.1$
- b)  $r = +0.9$
- c)  $r = -0.9$
- d) Non è calcolabile

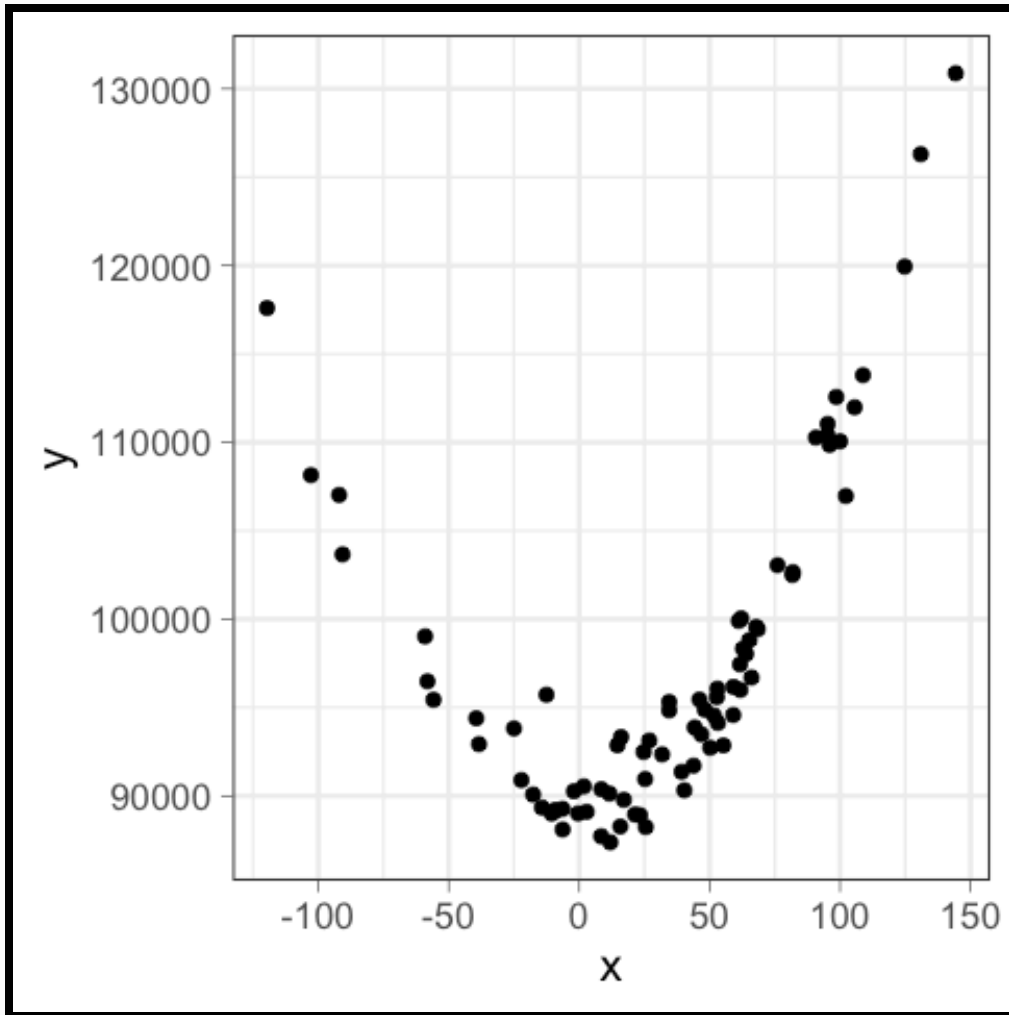
# Esercizio #21 -- Soluzione



? Osservando lo scatterplot, quale tra questi potrebbe essere un valore plausibile per l'indice di correlazione di Pearson tra le due variabili mostrate?

- a)  $r = +0.1$  ✓
- b)  $r = +0.9$
- c)  $r = -0.9$
- d) Non è calcolabile

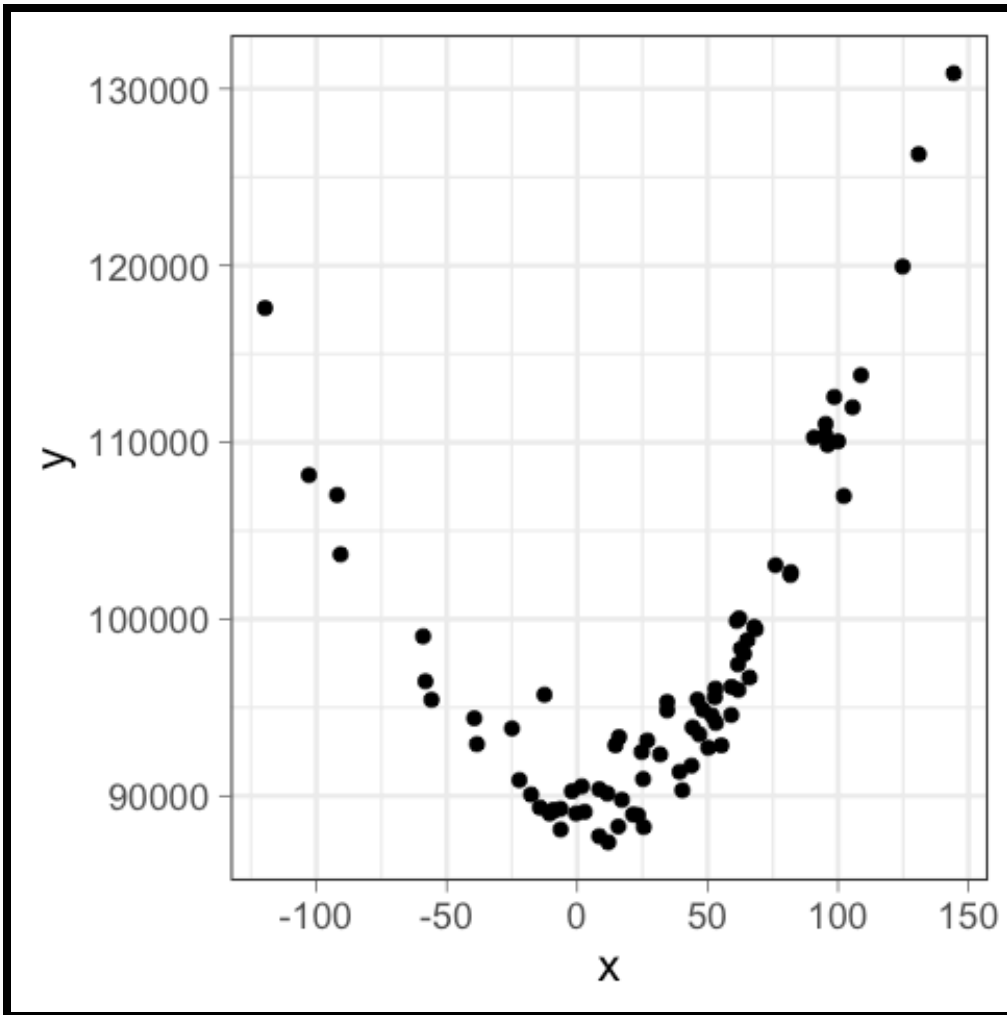
# Esercizio #22



Osservando lo scatterplot, quale tra questi potrebbe essere un valore plausibile per l'indice di correlazione di Pearson tra le due variabili mostrate?

- a)  $r = +0.1$
- b)  $r = +0.9$
- c)  $r = -0.9$
- d) Non è calcolabile

# Esercizio #22 -- Soluzione



? Osservando lo scatterplot, quale tra questi potrebbe essere un valore plausibile per l'indice di correlazione di Pearson tra le due variabili mostrate?

a)  $r = +0.1$

b)  $r = +0.9$

c)  $r = -0.9$

d) Non è calcolabile



# Esercizio #23

- ? Uno studio ha individuato una correlazione lineare  $r = -0.7$  tra le ore di sonno e un indice di irritabilità (scala 0-100; 0: poco irritabile, 100: molto irritabile).


Come interpreto questo valore?

- a) All'aumentare delle ore di sonno aumenta l'irritabilità
- b) All'aumentare delle ore di sonno diminuisce l'irritabilità
- c) La mancanza di sonno causa un aumento dell'irritabilità
- d) La mancanza di sonno causa una diminuzione dell'irritabilità
- e) Nessuna delle precedenti

# Esercizio #23 -- Soluzione

- ? Uno studio ha individuato una correlazione lineare  $r = -0.7$  tra le ore di sonno e un indice di irritabilità (scala 0-100; 0: poco irritabile, 100: molto irritabile).

Come interpreto questo valore?

- a) All'aumentare delle ore di sonno aumenta l'irritabilità
- b) All'aumentare delle ore di sonno diminuisce l'irritabilità 
- c) La mancanza di sonno causa un aumento dell'irritabilità
- d) La mancanza di sonno causa una diminuzione dell'irritabilità
- e) Nessuna delle precedenti

# Esercizio #24



Posso calcolare la correlazione tra...

a) L'indice di irritabilità e le ore dormite

Vero Falso

b) L'indice di irritabilità del primo e del secondo figlio

Vero Falso

c) L'indice di irritabilità prima e dopo un'attività

Vero Falso

d) L'indice di irritabilità in uomini e donne

Vero Falso



# Esercizio #24 -- Soluzione



Posso calcolare la correlazione tra...

a) L'indice di irritabilità e le ore dormite

Vero ☒ Falso

b) L'indice di irritabilità del primo e del secondo figlio

Vero Falso

c) L'indice di irritabilità prima e dopo un'attività

Vero Falso

d) L'indice di irritabilità in uomini e donne

Vero Falso

# Esercizio #24 -- Soluzione



Posso calcolare la correlazione tra...

a) L'indice di irritabilità e le ore dormite

Vero ☒ Falso

b) L'indice di irritabilità del primo e del secondo figlio

Vero ☒ Falso

c) L'indice di irritabilità prima e dopo un'attività

Vero Falso

d) L'indice di irritabilità in uomini e donne

Vero Falso

# Esercizio #24 -- Soluzione



Posso calcolare la correlazione tra...

a) L'indice di irritabilità e le ore dormite

Vero ☒ Falso

b) L'indice di irritabilità del primo e del secondo figlio

Vero ☒ Falso

c) L'indice di irritabilità prima e dopo un'attività

Vero ☒ Falso

d) L'indice di irritabilità in uomini e donne

Vero Falso

# Esercizio #24 -- Soluzione



Posso calcolare la correlazione tra...

a) L'indice di irritabilità e le ore dormite

Vero ☒ Falso

b) L'indice di irritabilità del primo e del secondo figlio

Vero ☒ Falso

c) L'indice di irritabilità prima e dopo un'attività

Vero ☒ Falso

d) L'indice di irritabilità in uomini e donne

Vero Falso ☒

# Correlazione & valori estremi

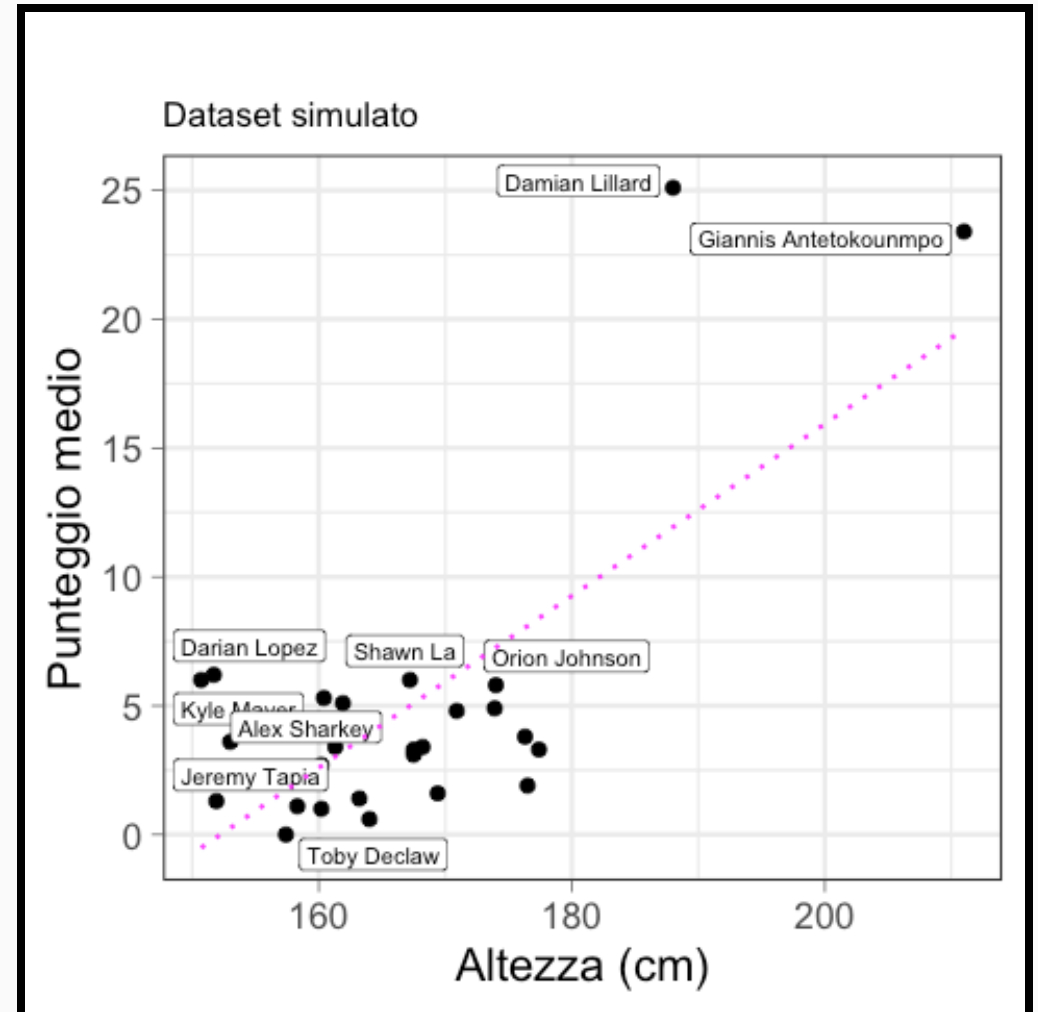
Altezza (cm) e numero di canestri

- $r = 0.72$

# Correlazione & valori estremi

Altezza (cm) e numero di canestri

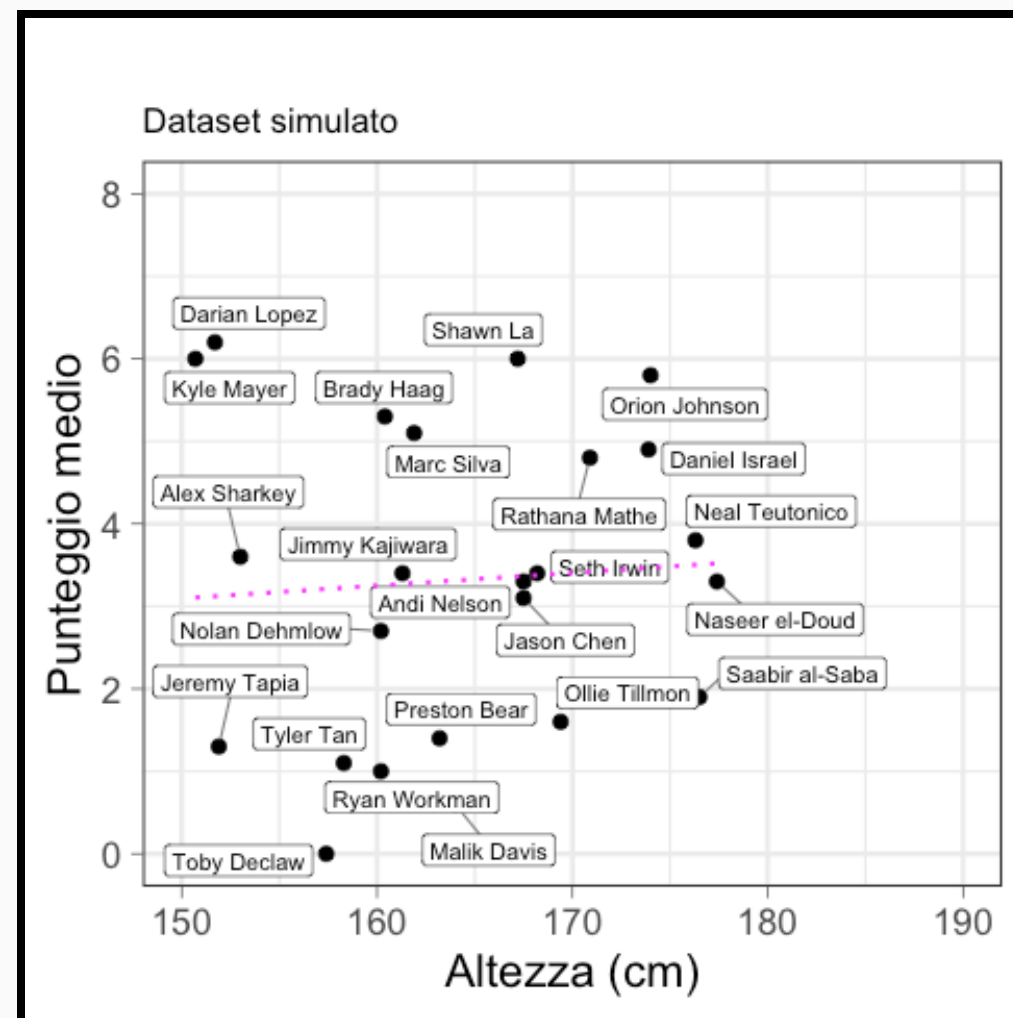
- $r = 0.72$



# Correlazione & valori estremi

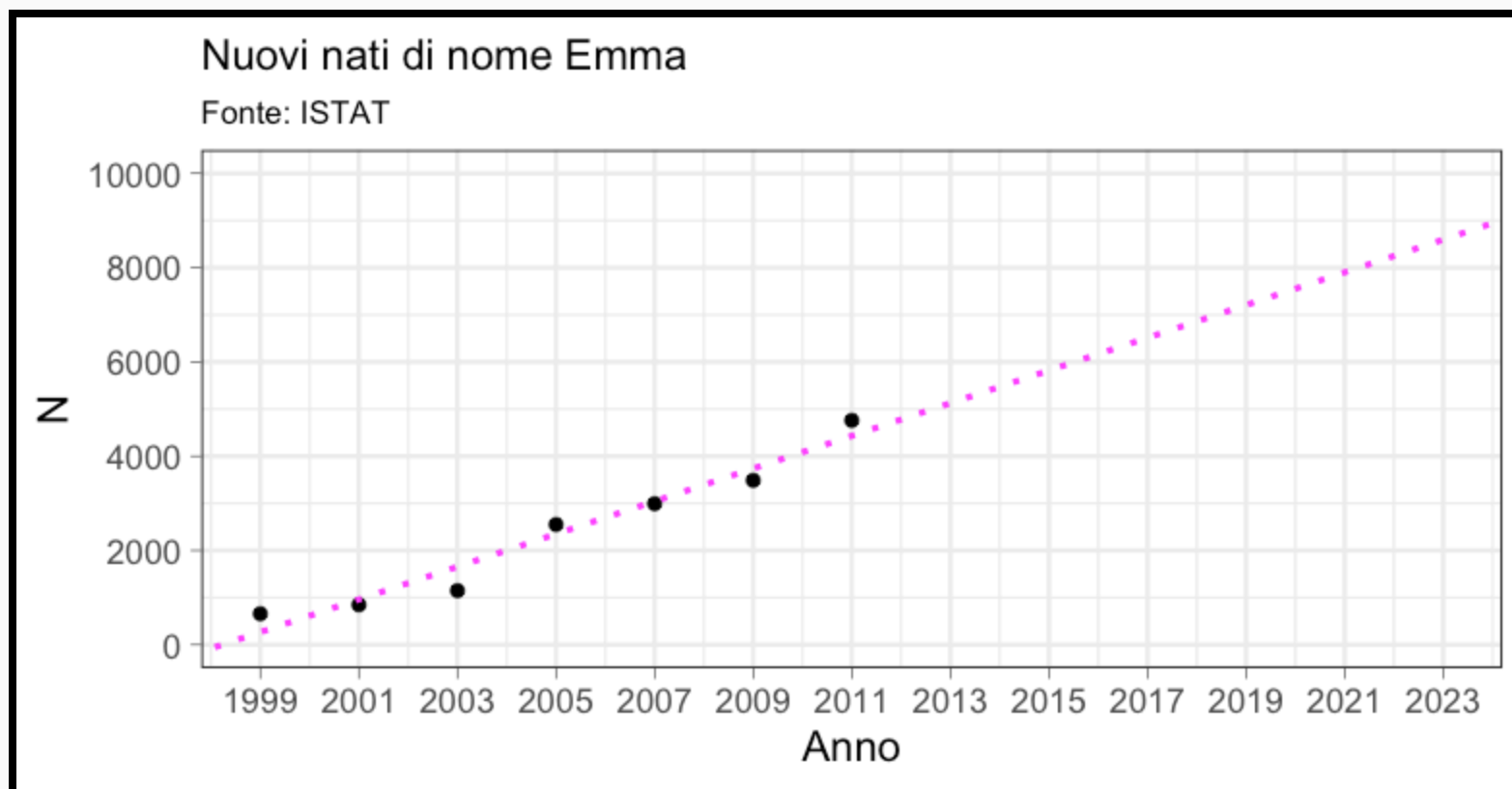
Altezza (cm) e numero di canestri

- $r = 0.72$
- $r_{\text{no outliers}} = 0.07$



# Correlazione: interpolare ed estrapolare

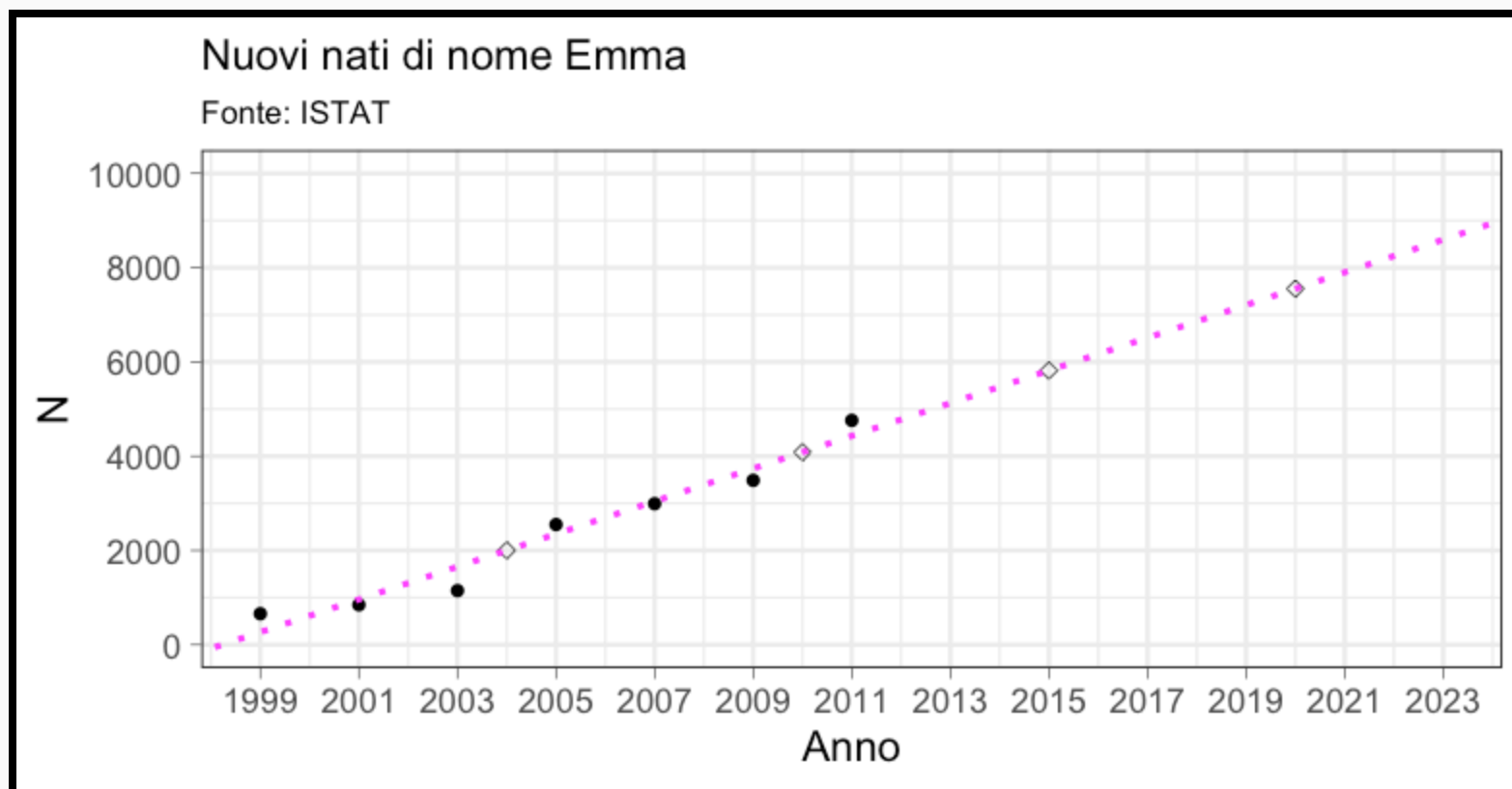
? Quante bambine di nome Emma sono nate nel 2004, 2010, 2015 e 2020?





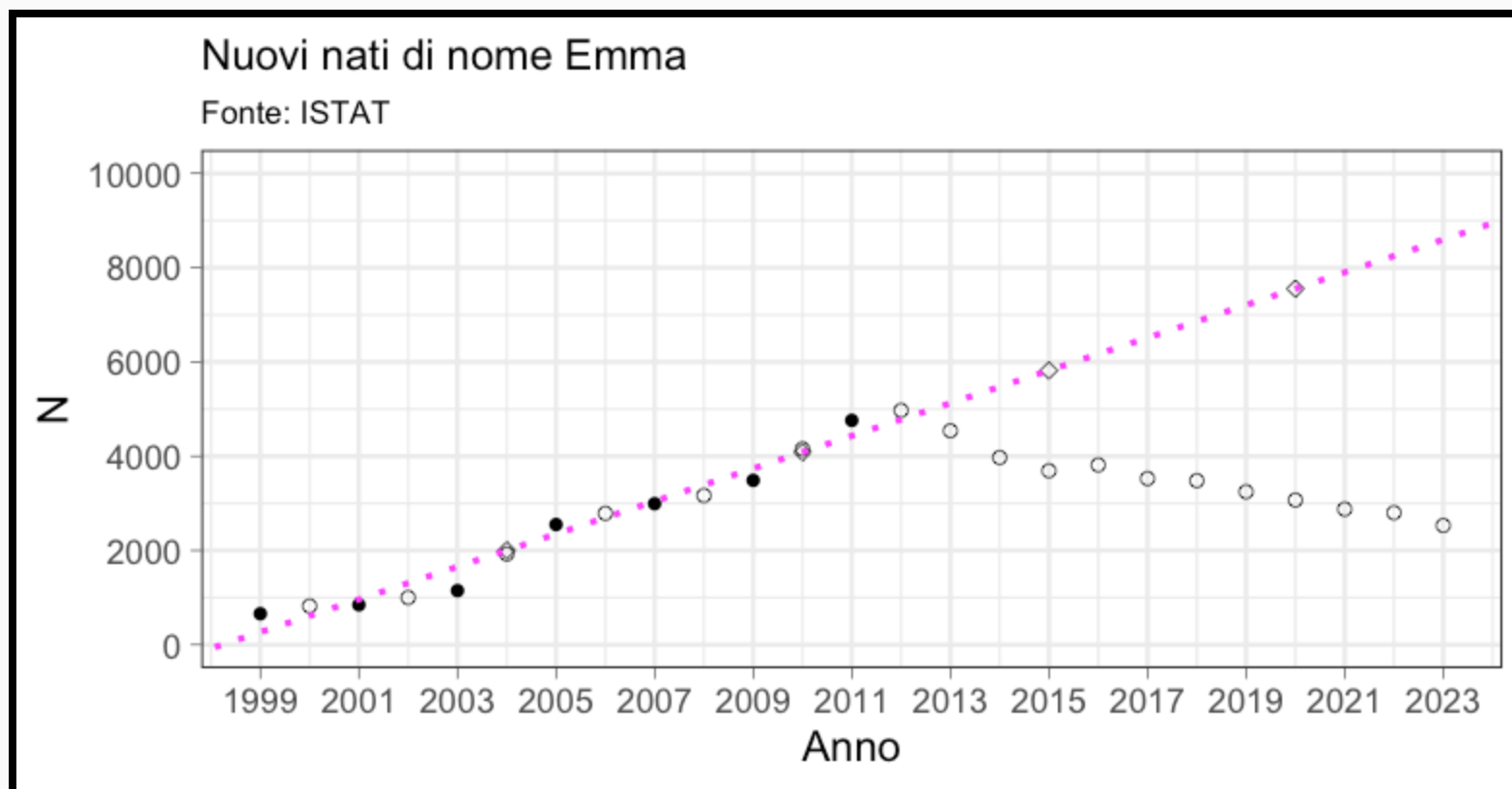
# Correlazione: interpolare ed estrapolare

? Quante bambine di nome Emma sono nate nel 2004, 2010, 2015 e 2020?



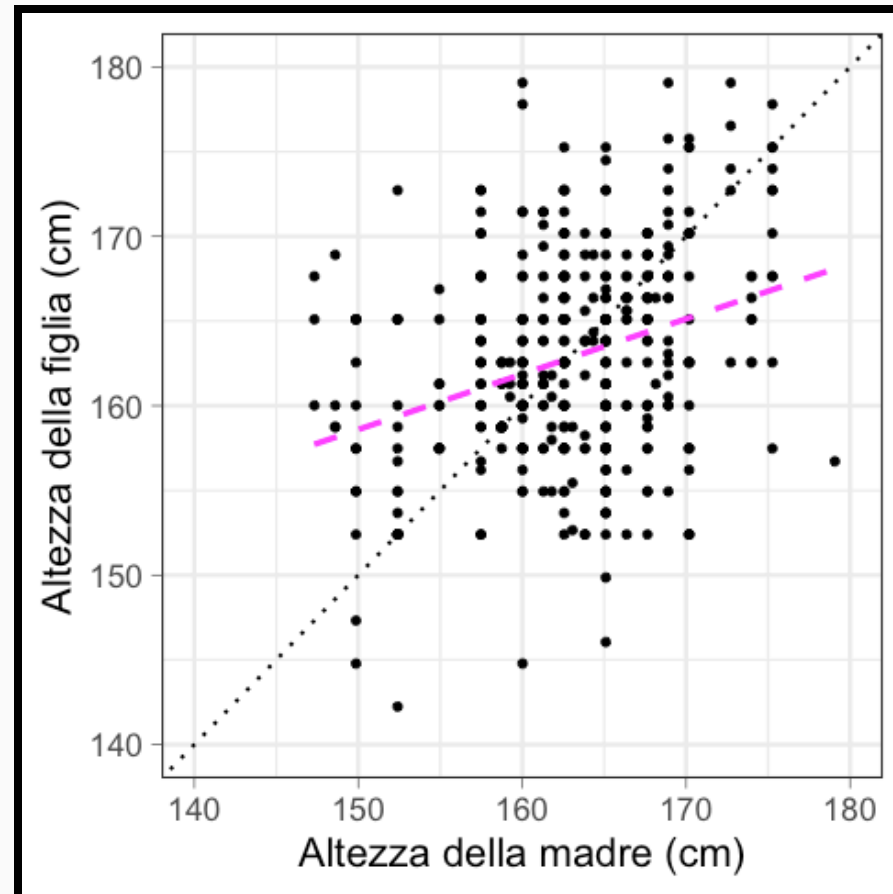
# Correlazione: interpolare ed estrapolare

? Quante bambine di nome Emma sono nate nel 2004, 2010, 2015 e 2020?



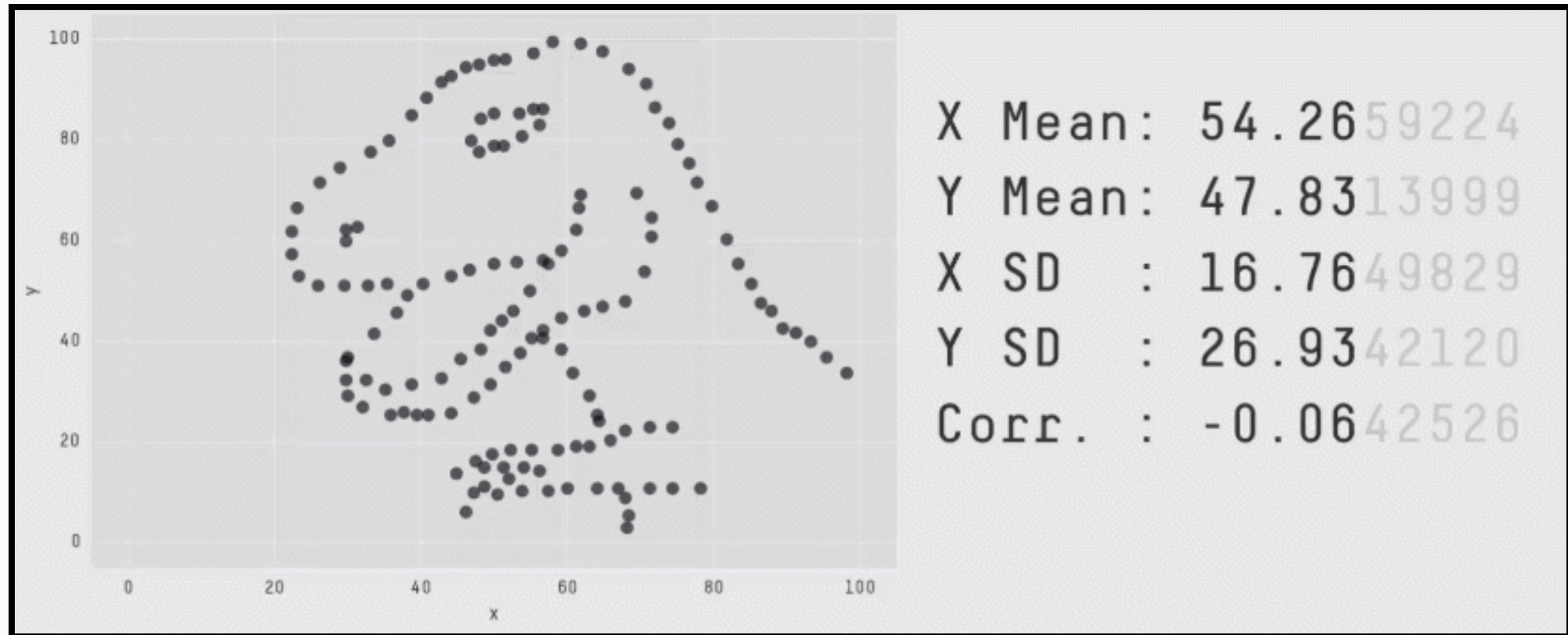
# **La regressione verso la media**

# La regressione verso la media



F. Galton, *Regression Towards Mediocrity in Hereditary Stature*, The Journal of the Anthropological Institute of Great Britain and Ireland, 1886, <https://doi.org/10.2307/2841583>

# Perché visualizzare i dati?



**Datasaurus Dozen**, Matejka, J & Fitzmaurice, G. *Same Stats, Different Graphs: Generating Datasets with Varied Appearance and Identical Statistics through Simulated Annealing*, Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems, doi:10.1145/3025453.3025912

# Parametri vs statistiche

	Parametro	Statistica
Numerosità	$N$	$n$
Media	$\mu$	$\bar{x}$
Deviazione Standard	$\sigma$	$s$
Proporzione	$\pi$	$p$
Correlazione	$\rho$	$r$

# Esercizio #25

? La media nella popolazione viene indicata con...

a)  $M$

b)  $m$

c)  $\mu$

d)  $\bar{x}$

# Esercizio #25 -- Soluzione

? La media nella popolazione viene indicata con...

a)  $M$

b)  $m$

c)  $\mu$  

d)  $\bar{x}$



# Cosa abbiamo imparato in questa lezione?

- Le variabili numeriche possono essere rappresentate con misure di centralità, dispersione e correlazione (statistiche)
- Alcune statistiche sono "falsate" se le distribuzioni empiriche sono asimmetriche e/o includono valori estremi e possono nascondere dettagli importanti dei dati
- Le variabili numeriche possono essere rappresentate graficamente in diversi modi, ma alcune possono nascondere dettagli importanti delle distribuzioni empiriche
- Visualizzare i dati è importante per interpretarli
- Il campione viene rappresentato con le statistiche, la popolazione con i parametri