

Lezione 5

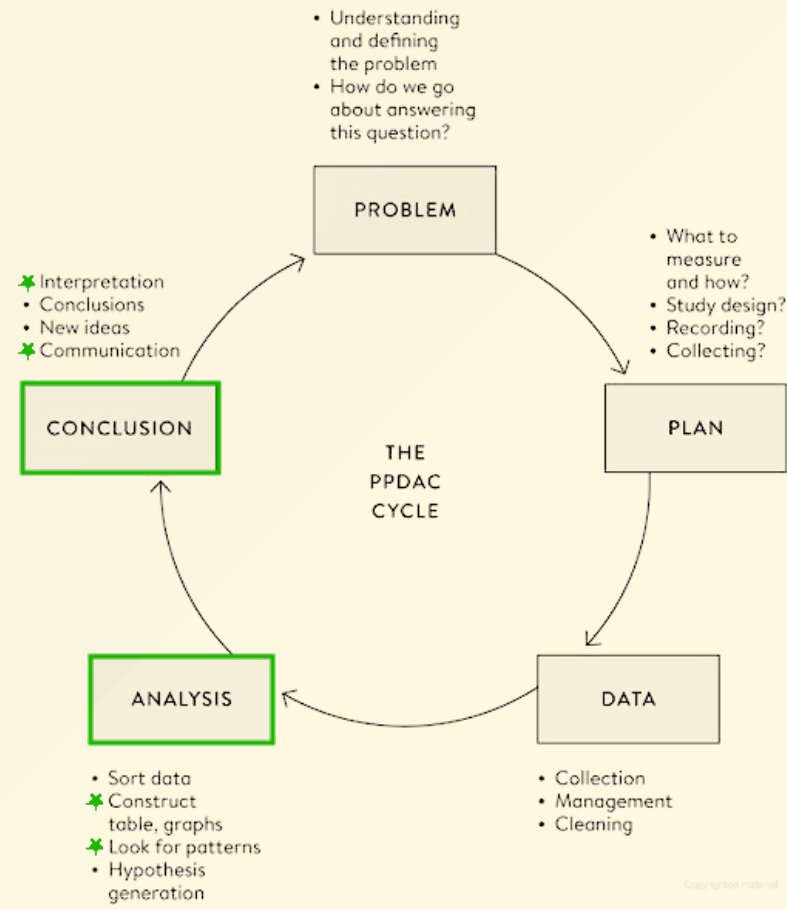
La statistica descrittiva

(Parte II: Le variabili numeriche)

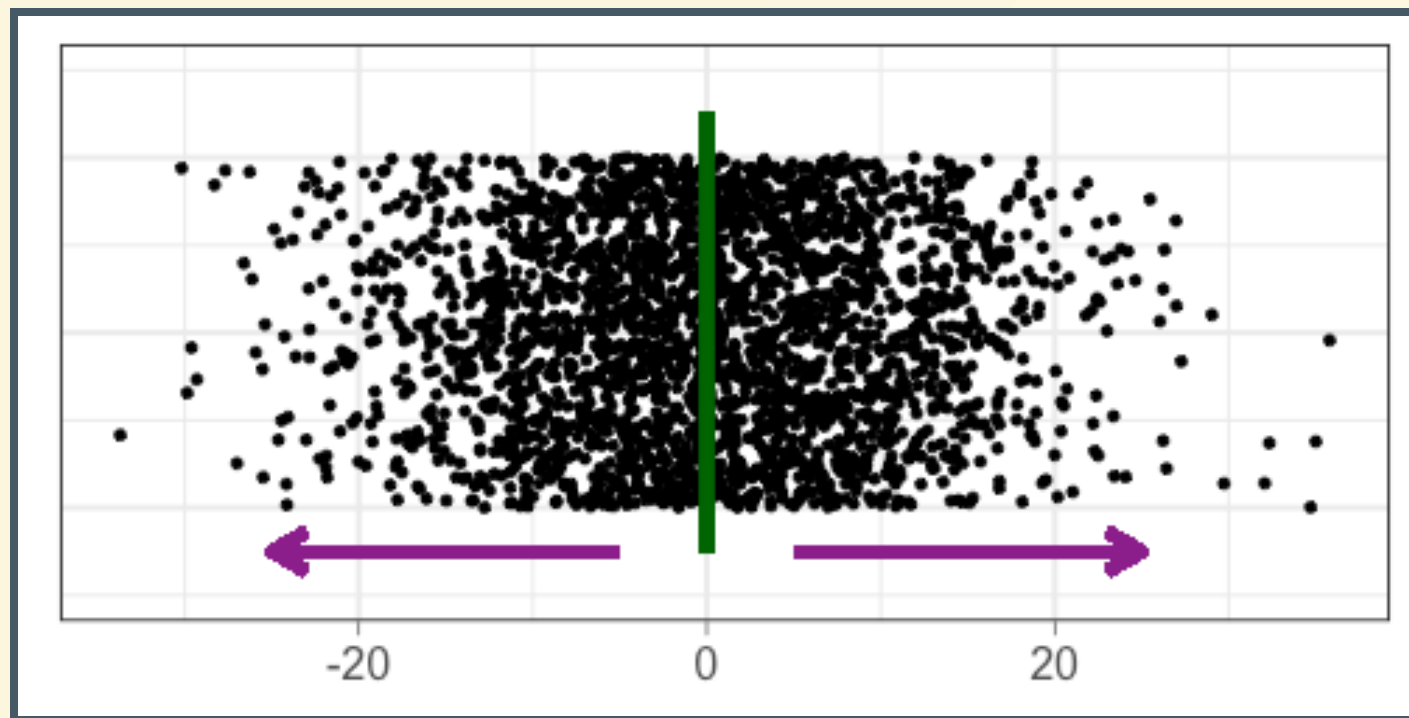
Obiettivi di apprendimento

- Saper calcolare e interpretare diverse 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



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

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

Esercizio #1

? Qual è la moda dei seguenti insiemi?

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

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

Esercizio #1 -- Soluzione

? Qual è la moda dei seguenti insiemi?

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

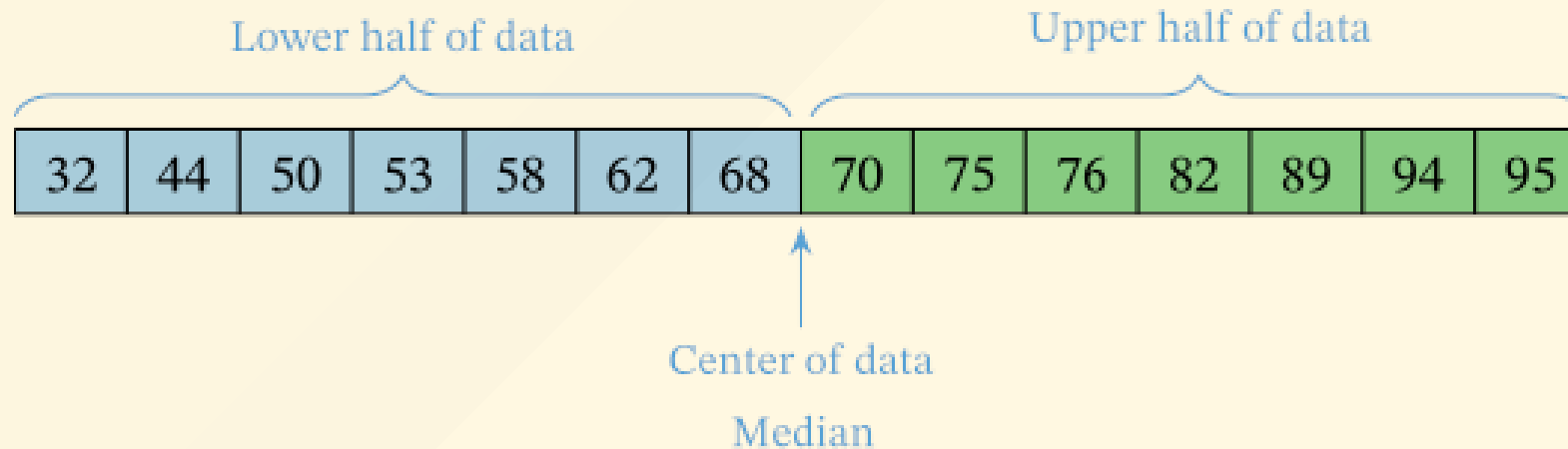
$$\text{moda}(y) = 1 \wedge 4$$

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

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

Misure di centralità: la mediana

 Il valore "in mezzo"



 I dati devono essere ordinati!

Esercizio #2

? Quali sono le mediane di questi insiemi?

$n = 7, y = \{1, 3, 3, 6, 7, 8, 9\}$
 $\text{mediana}(y) = ?$

$n = 7, z = \{1, 3, 3, 6, 7, 8, 109\}$
 $\text{mediana}(z) = ?$

Esercizio #2 -- Soluzione

? Quali sono le mediane di questi insiemi?

$$n = 7, y = \{1, 3, 3, 6, 7, 8, 9\}$$

$$\text{mediana}(y) = y_4 = 6$$

$$n = 7, z = \{1, 3, 3, 6, 7, 8, 109\}$$

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

! I dati devono essere ordinati!

Esercizio #2 -- Soluzione

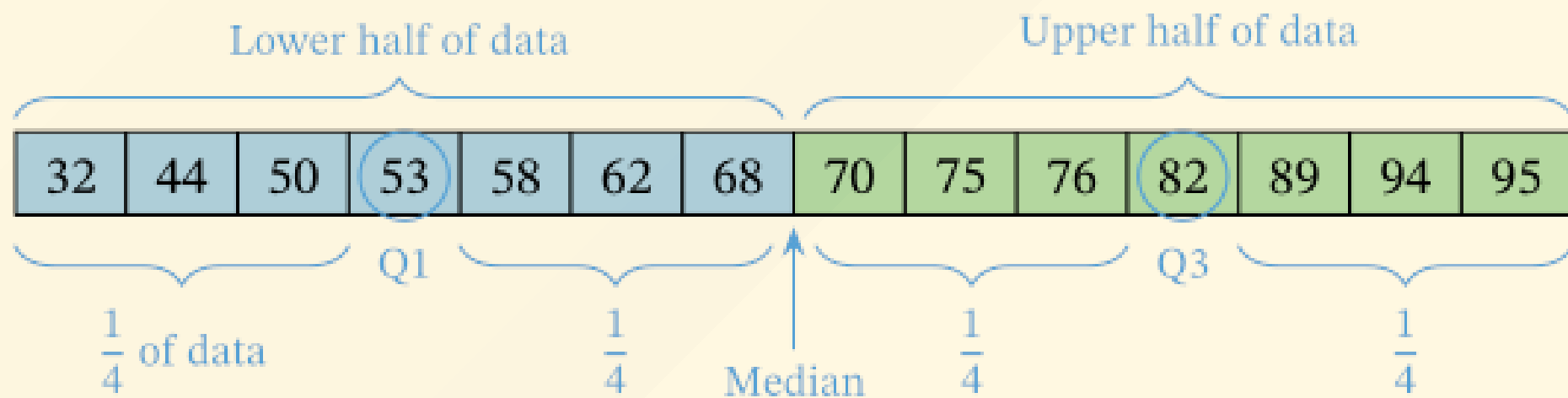
? Quali sono le mediane di questi insiemi?

$$n = 7, y = \{1, 3, 3, 6, 7, 8, 9\}$$
$$\text{mediana}(y) = y_4 = 6$$

$$n = 7, z = \{1, 3, 3, 6, 7, 8, 109\}$$
$$\text{mediana}(z) = z_4 = 6$$

! I dati devono essere ordinati!

Quartili



⚠ I dati devono essere ordinati!

Misure di centralità: la media



Media aritmetica

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



$$x = \{4, 36, 45, 50, 75\}$$

$$\bar{x} = \frac{1}{n} \left(\sum_{i=1}^n x_i \right) = \frac{4+36+45+50+75}{5} = 42$$

Esercizio #3

? Quali sono le medie di questi insiemi?

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

$$\bar{y} = ?$$

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

$$\bar{z} = ?$$

Esercizio #3 -- Soluzione

? Quali sono le medie di questi insiemi?

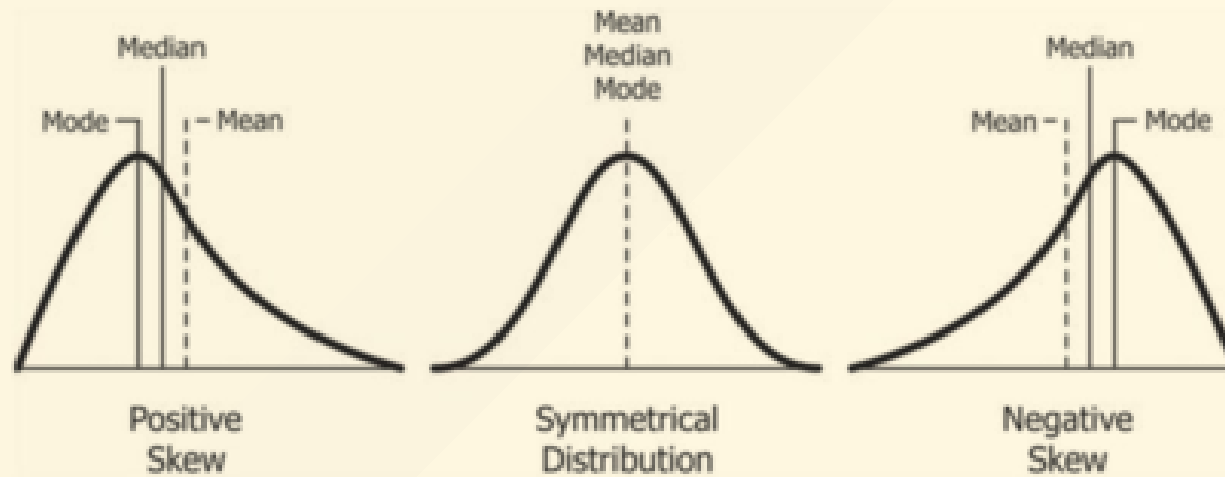
$$y = \{6, 34, 40, 55, 75\}$$

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

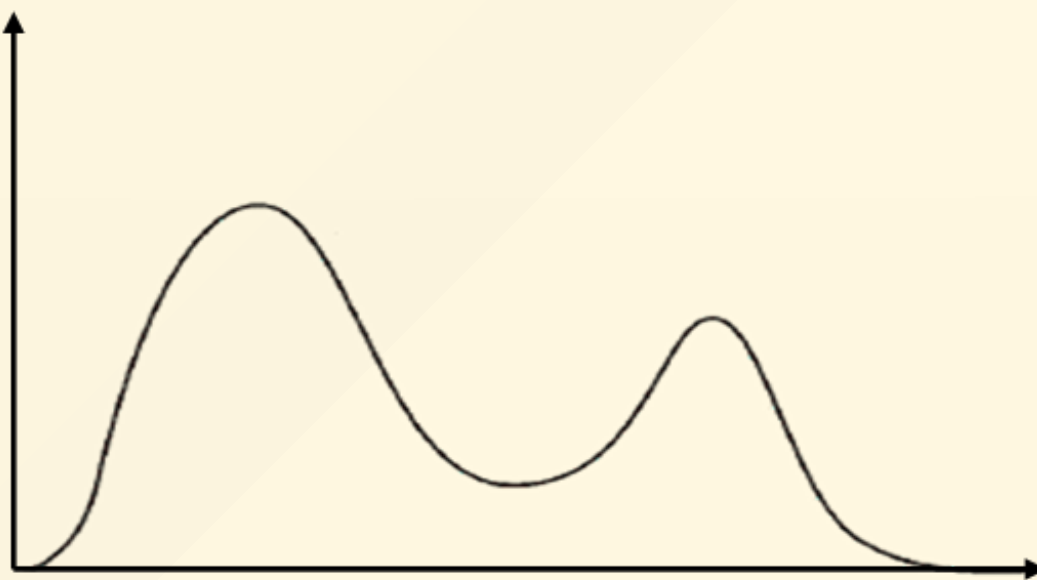
$$z = \{6, 34, 40, 55, 175\}$$

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

La forma delle distribuzioni



La forma delle distribuzioni



Esercizio 4

? 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 4 -- Soluzione

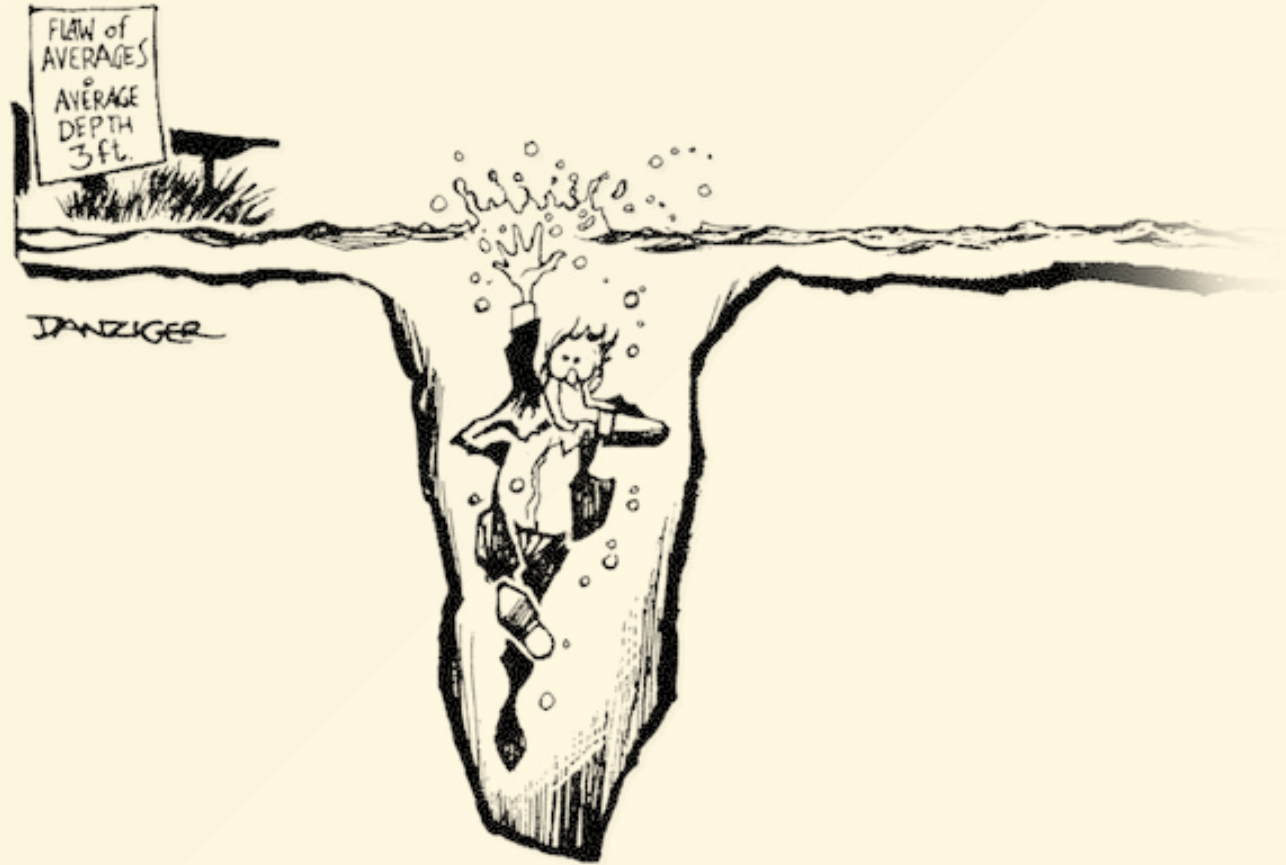
? Nei risultati di uno studio è riportata la seguente frase:

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

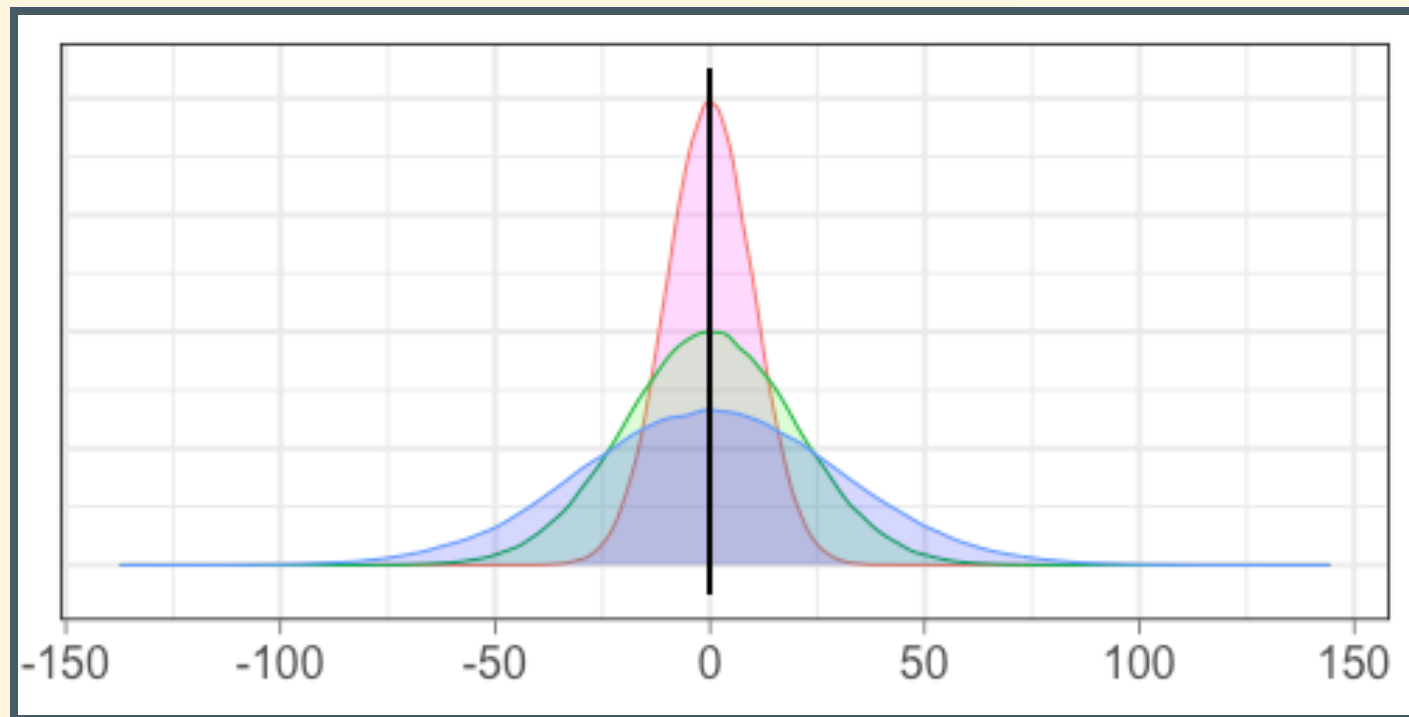
La distribuzione empirica ha una forma...

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

Misure di dispersione

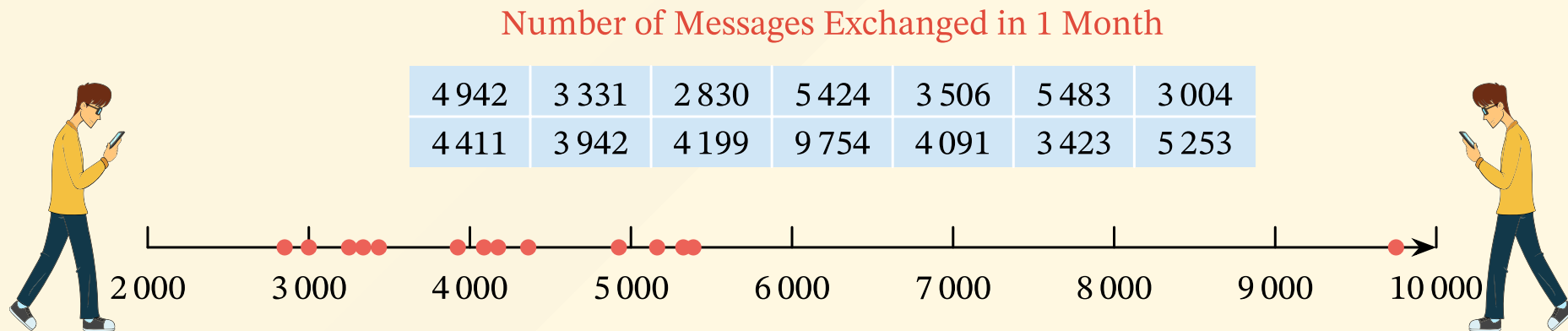


Misure di dispersione



Misure di dispersione: range

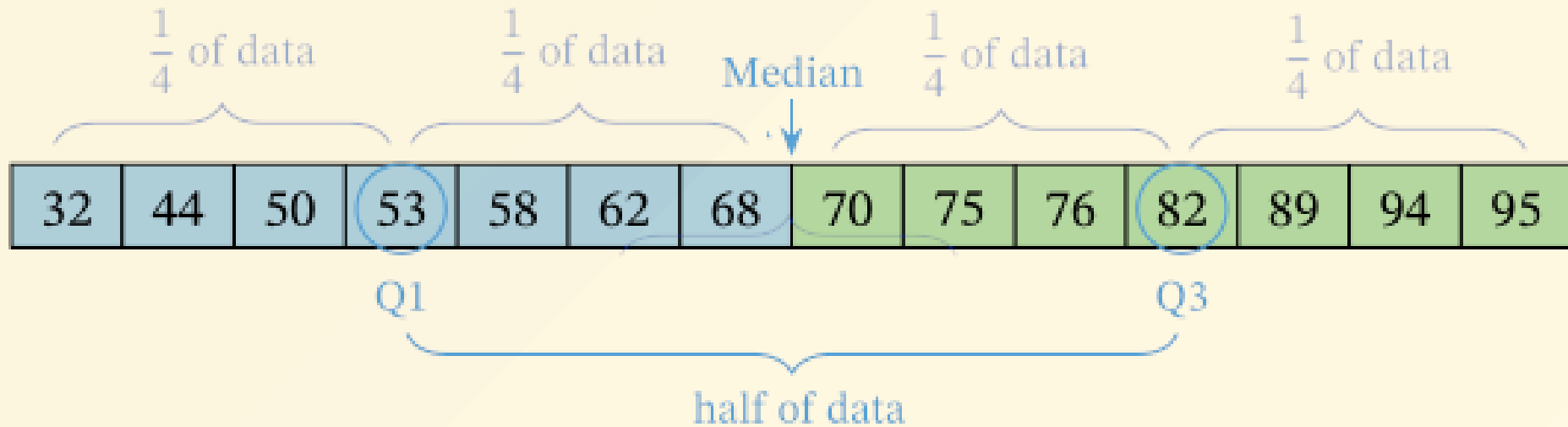
 $\text{range}(x) = \max(x) - \min(x)$



$$\text{range}(x) = 9754 - 2830 = 6924$$

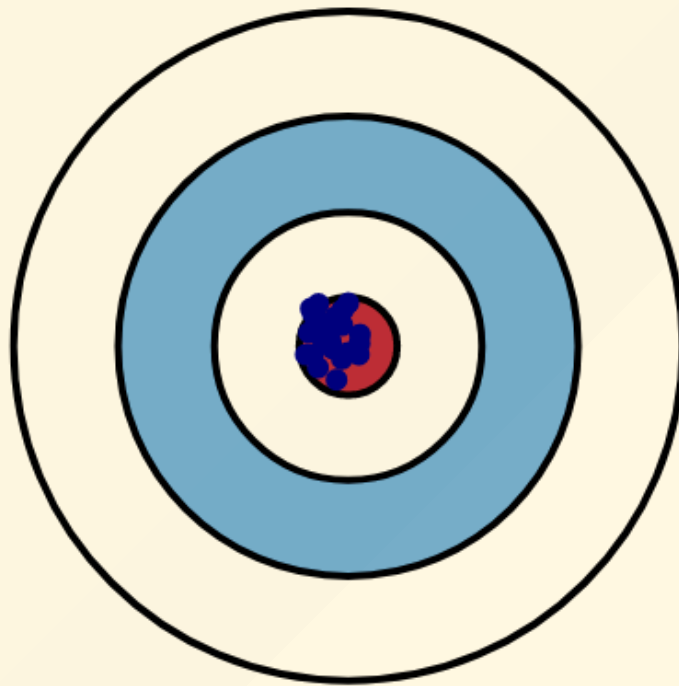
Misure di dispersione: range interquantile

🎯 $\text{IQR}(x) = Q3(x) - Q1(x)$

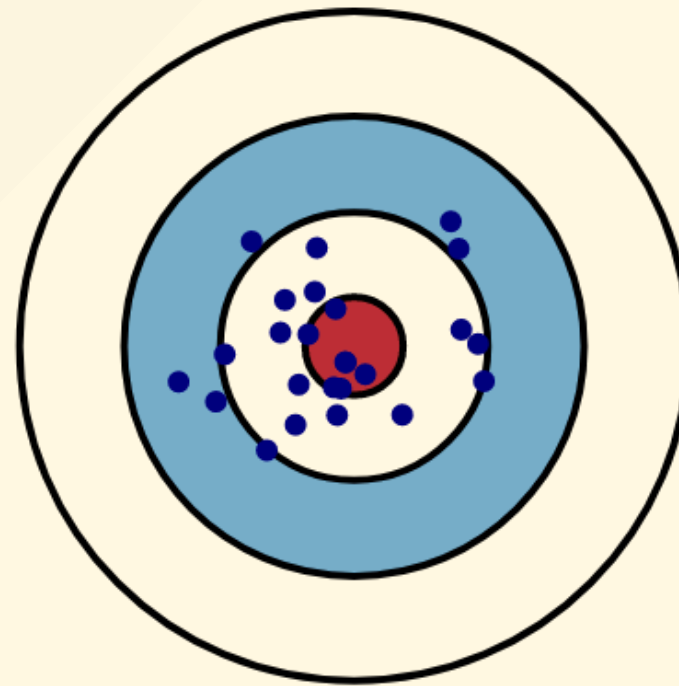


Misure di dispersione: varianza


Low Variance



High Variance



Misure di dispersione: varianza


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

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

 $x = \{1, 2, 3\} \quad \bar{x} = \frac{1+2+3}{3} = 2$

$$\begin{aligned} s &= \frac{1}{3-1} \times [(1-2)^2 + (2-2)^2 + (3-2)^2] = \\ &= \frac{1}{2} \times [1^2 + 0^2 + 1^2] = \frac{1}{2} \times 2 = 1 \end{aligned}$$

Misure di dispersione: deviazione standard

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

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

 $x = \{1, 2, 3\} \quad \bar{x} = \frac{1+2+3}{3} = 2$

$$\begin{aligned} s &= \sqrt{\frac{1}{3-1} \times [(1-2)^2 + (2-2)^2 + (3-2)^2]} = \\ &= \sqrt{\frac{1}{2} \times [1^2 + 0^2 + 1^2]} = \sqrt{\frac{1}{2} \times 2} = \sqrt{1} = 1 \end{aligned}$$

Esercizio #5

- ? La deviazione standard è un indice di dispersione?
a) Vero b) Falso
- ? La moda è una misura di tendenza centrale?
a) Vero b) Falso
- ? La mediana, rispetto alla media, è più sensibile ai valori estremi?
a) Vero b) Falso

Esercizio #5 -- Soluzione

- ? La deviazione standard è un indice di dispersione?
a) Vero ☒ b) Falso
- ? La moda è una misura di tendenza centrale?
a) Vero b) Falso
- ? La mediana, rispetto alla media, è più sensibile ai valori estremi?
a) Vero b) Falso

Esercizio #5 -- Soluzione

- ? La deviazione standard è un indice di dispersione?
a) Vero ☒ b) Falso
- ? La moda è una misura di tendenza centrale?
a) Vero ☒ b) Falso
- ? La mediana, rispetto alla media, è più sensibile ai valori estremi?
a) Vero b) Falso

Esercizio #5 -- Soluzione

- ? La deviazione standard è un indice di dispersione?
a) Vero ☒ b) Falso
- ? La moda è una misura di tendenza centrale?
a) Vero ☒ b) Falso
- ? La mediana, rispetto alla media, è più sensibile ai valori estremi?
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 6

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

Coronary-artery calcium scores averaged 68.9 ± 244.2 (range 0 to 1526) in patients and 8.8 ± 41.8 (range 0 to 243.4) in controls.

Come descrivereste in Table 1 questa variabile?


- a) con media e standard deviation
- b) con mediana e interquantile range
- c) non ho abbastanza elementi per decidere
- d) nessuna delle precedenti

Esercizio 6 -- Soluzione

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

Coronary-artery calcium scores averaged 68.9 ± 244.2 (range 0 to 1526) in patients and 8.8 ± 41.8 (range 0 to 243.4) in controls

Come descrivereste in Table 1 questa variabile?

- a) con media e standard deviation
- b) con mediana e interquantile range 
- c) non ho abbastanza elementi per decidere
- d) nessuna delle precedenti

Esercizio #7

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 di bambine e ragazze nel gruppo di intervento?

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

Esercizio #7 -- 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 di bambine e ragazze nel gruppo di intervento?

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

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

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 l'età media dei pazienti nel gruppo di controllo?

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

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


02:00

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 l'età media dei pazienti nel gruppo di controllo?

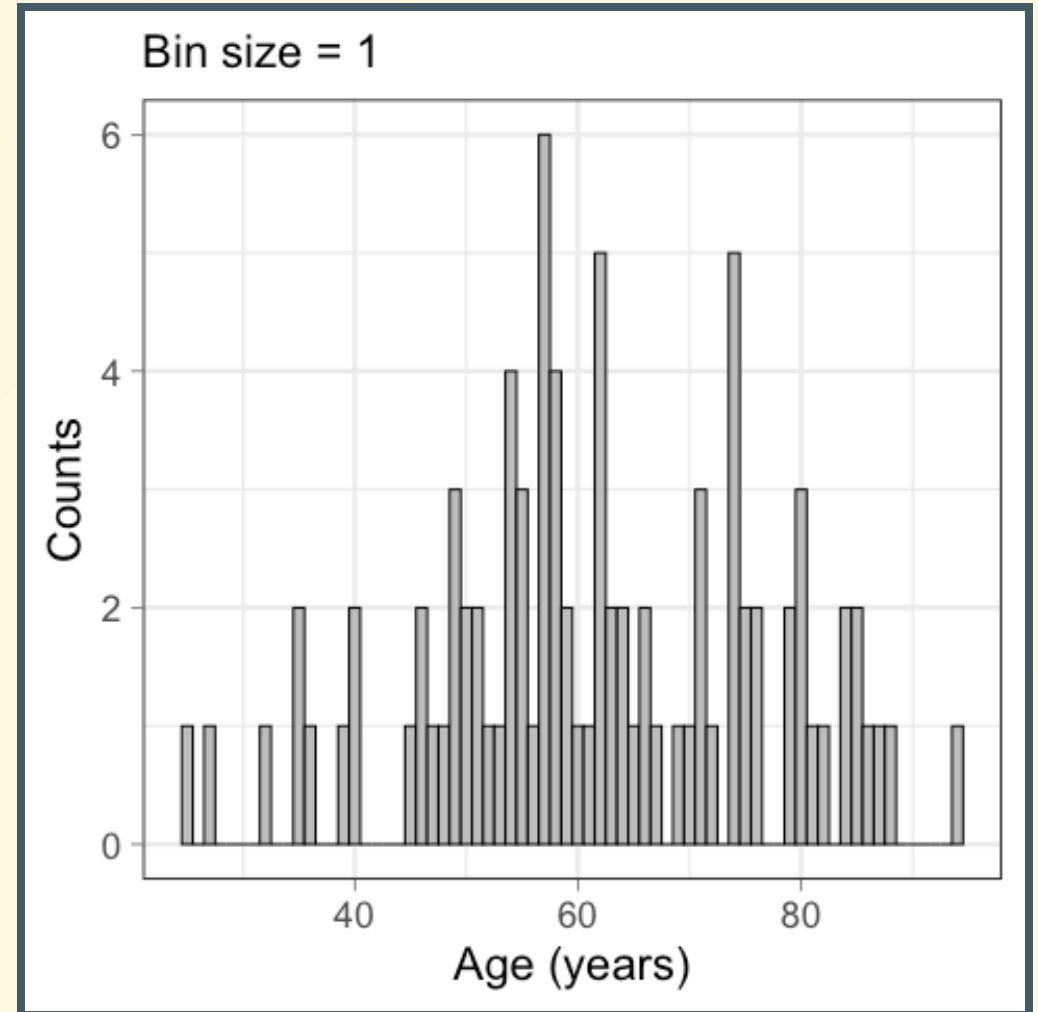
- a) 10.4 
- b) 4.1
- c) 4.0
- d) Non è possibile desumerlo dalla tabella

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

La visualizzazione dei dati numerici

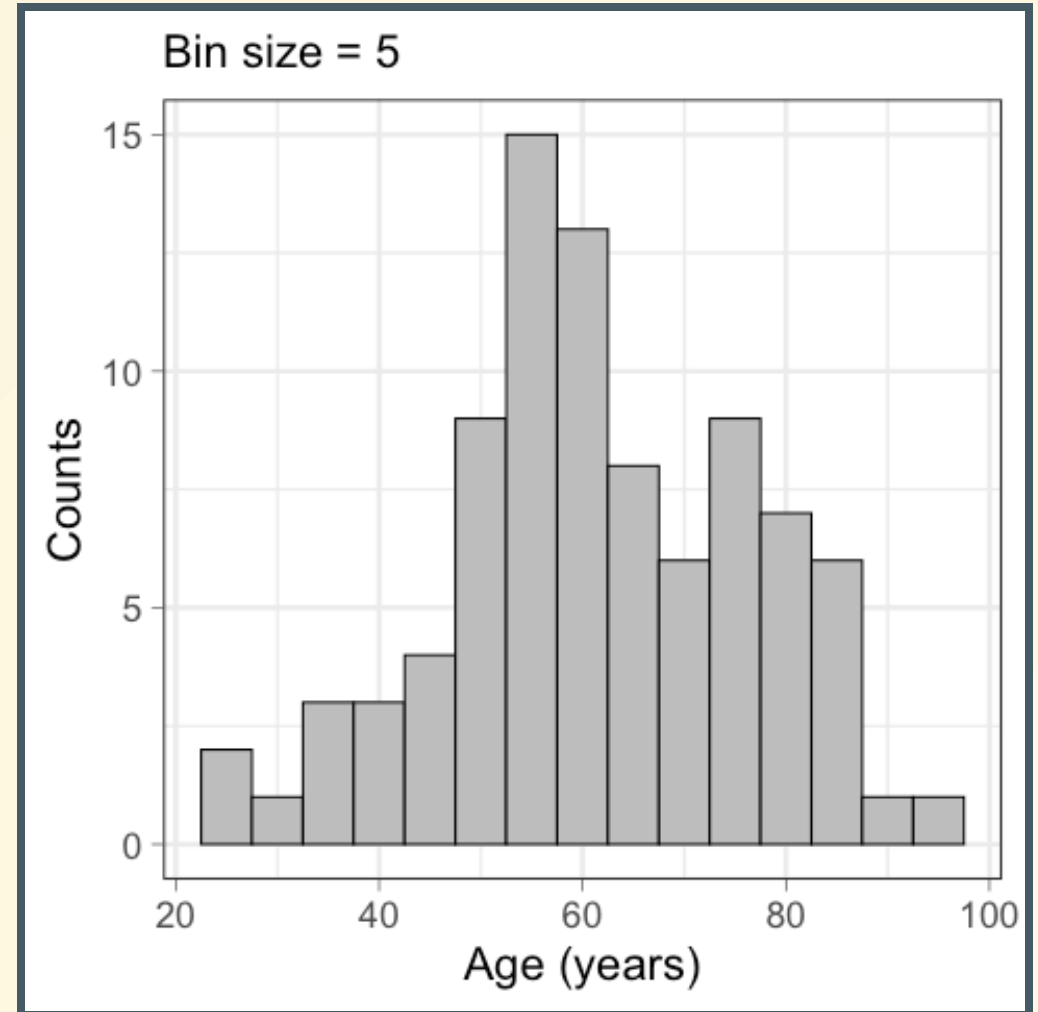
Istogramma

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



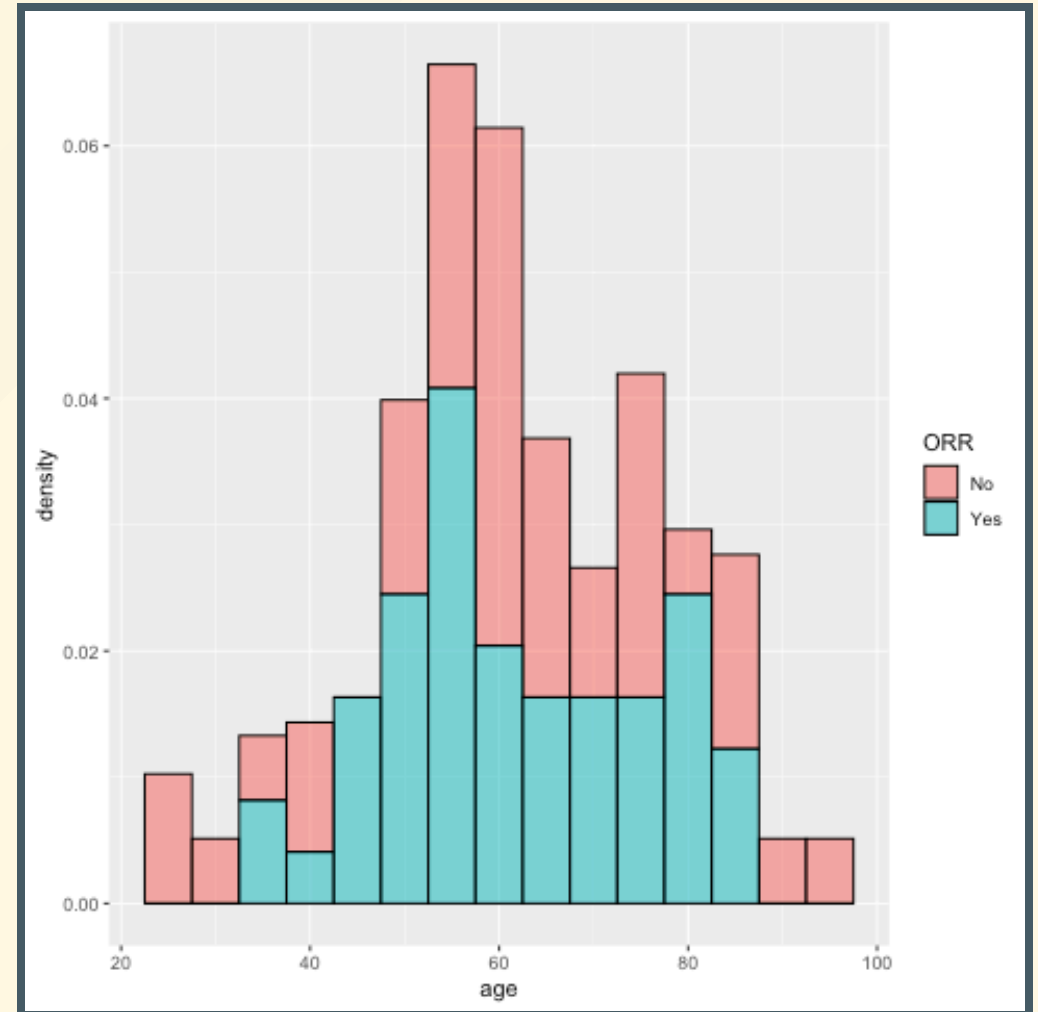
Istogramma

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



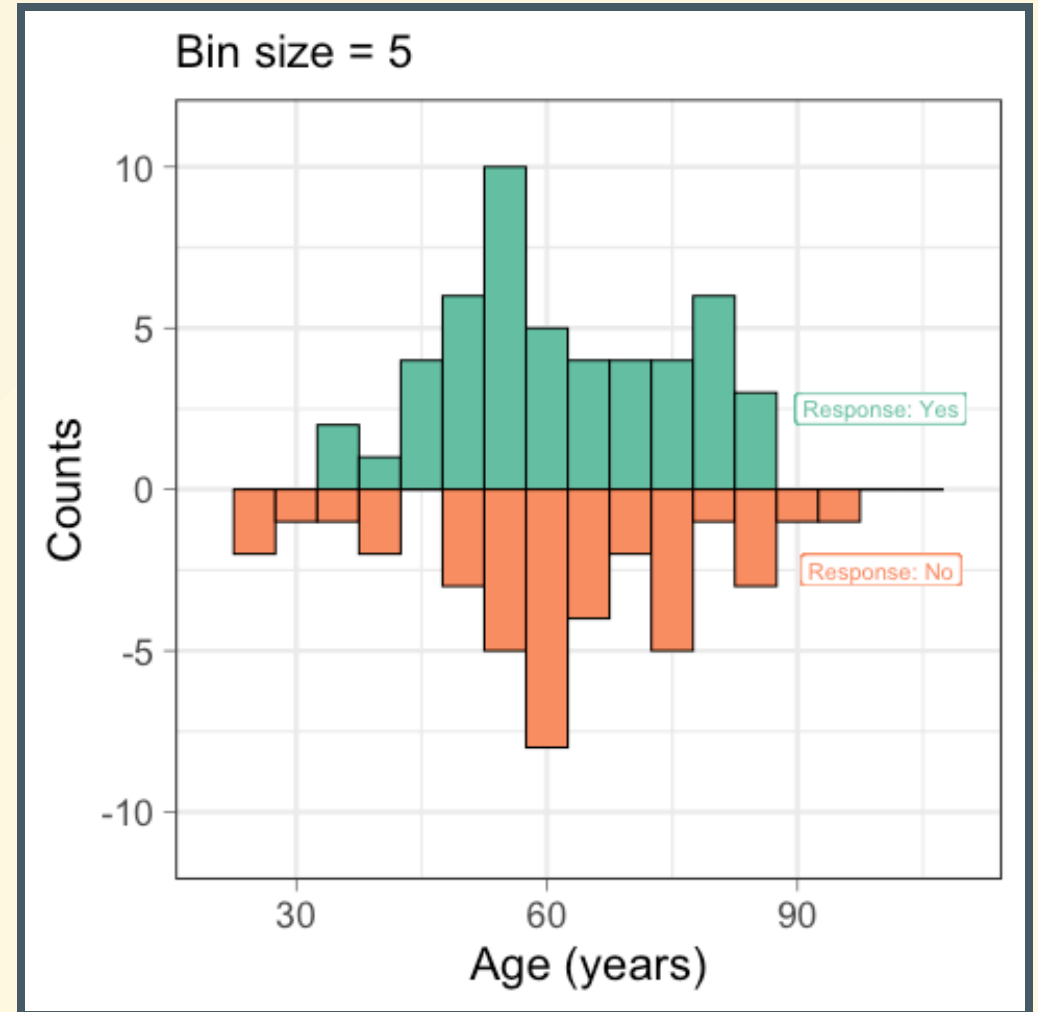
Istogramma

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



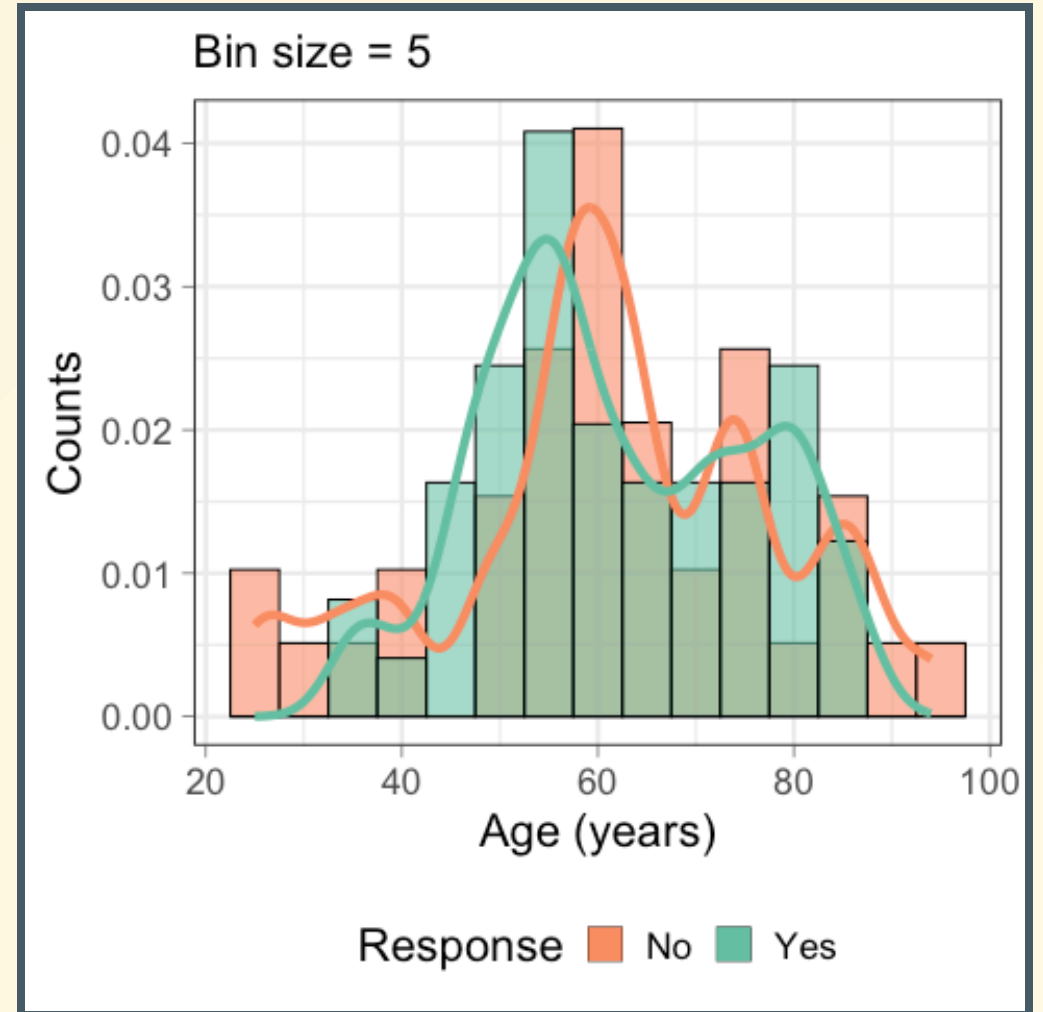
Miami plot/Mirror histogram

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

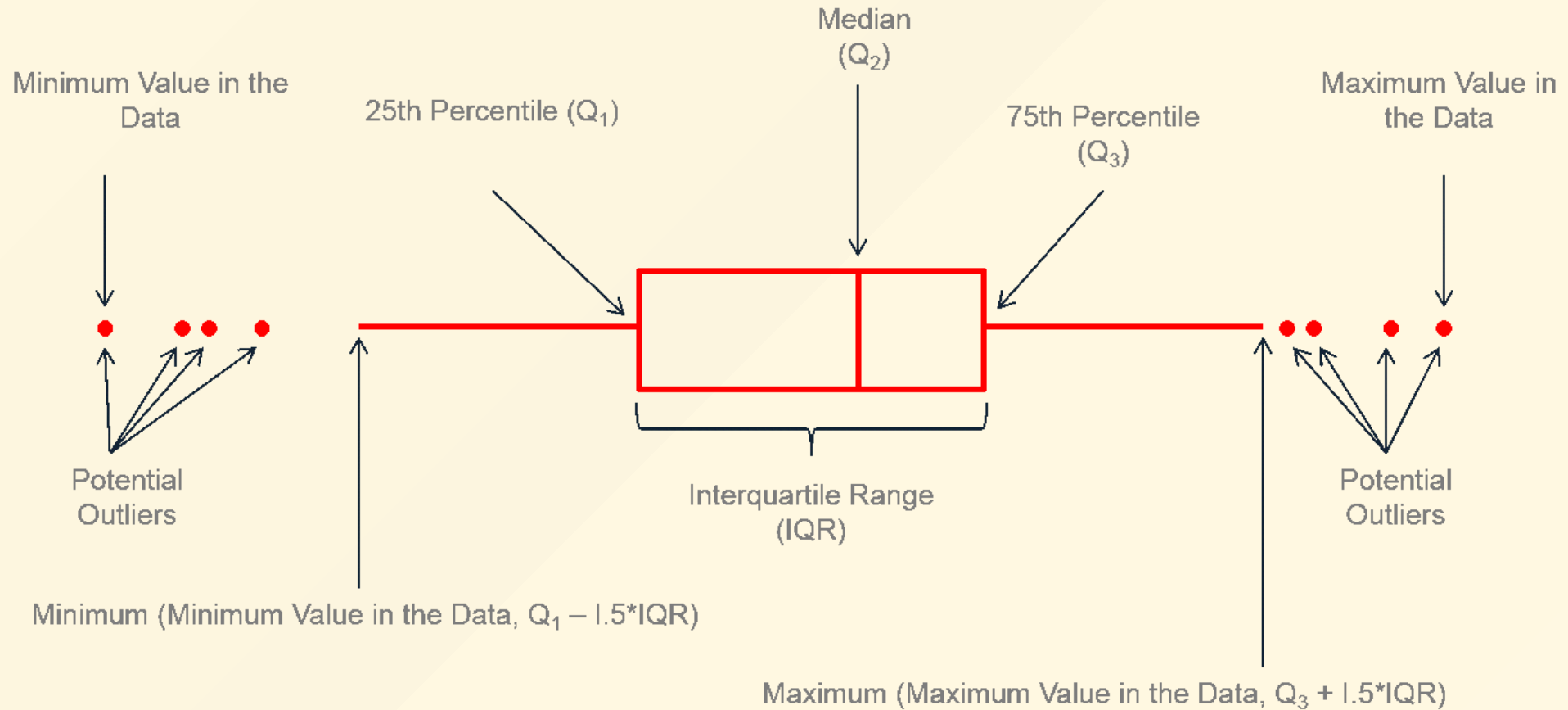


Density plot

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

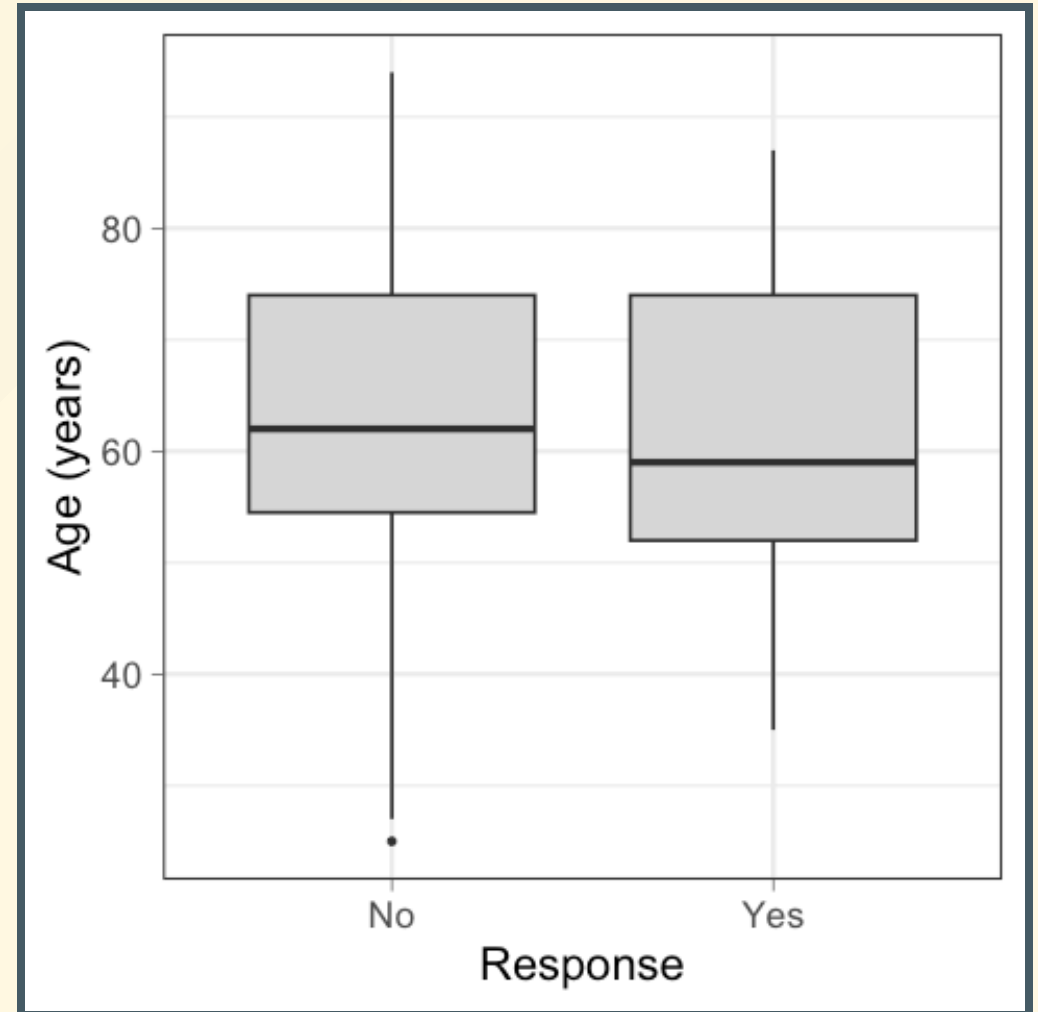


Boxplot

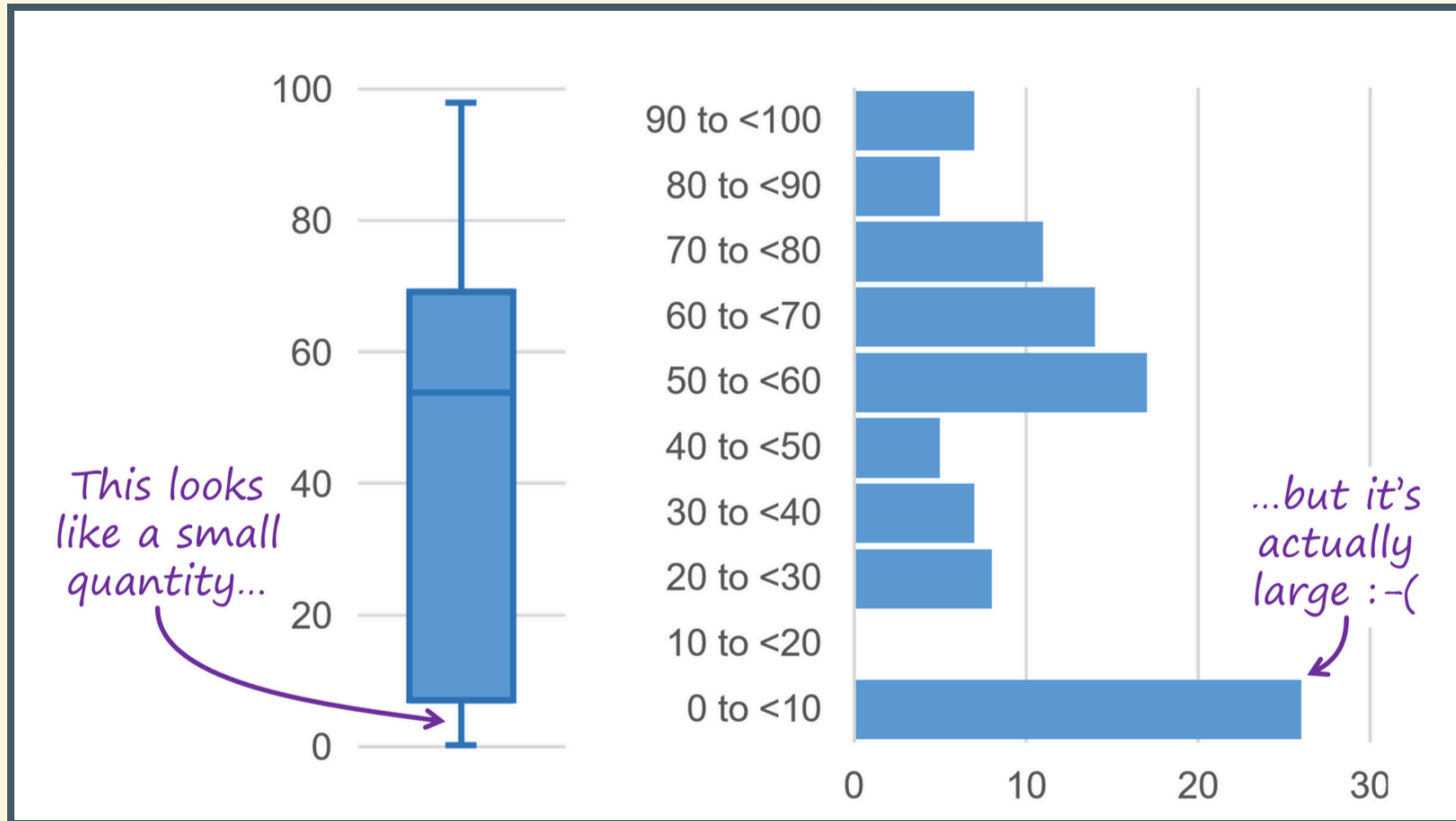


Boxplot

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

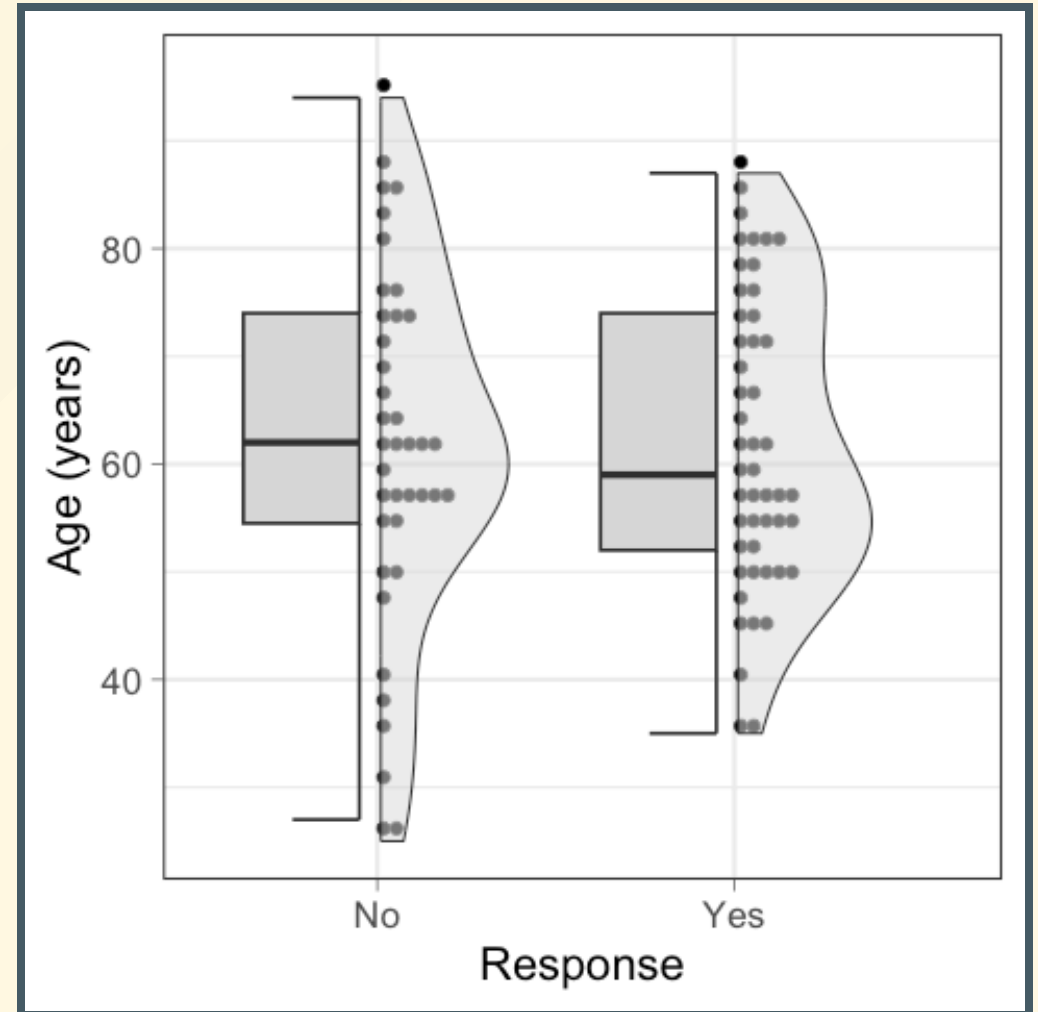


Boxplot



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 #9

“ Quanti partner sessuali le persone in Gran Bretagna riferiscono di aver avuto nella loro vita? ”

Esercizio #9

“ Quanti partner 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 (14)	3-10 (7)

Think

02:00

Esercizio #9

“ Quanti partner 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 (14)	3-10 (7)

Pair

03:00

Esercizio #9

“ Quanti partner sessuali le persone in Gran Bretagna riferiscono di aver avuto nella loro vita? ”

? Cosa ci dicono queste statistiche?

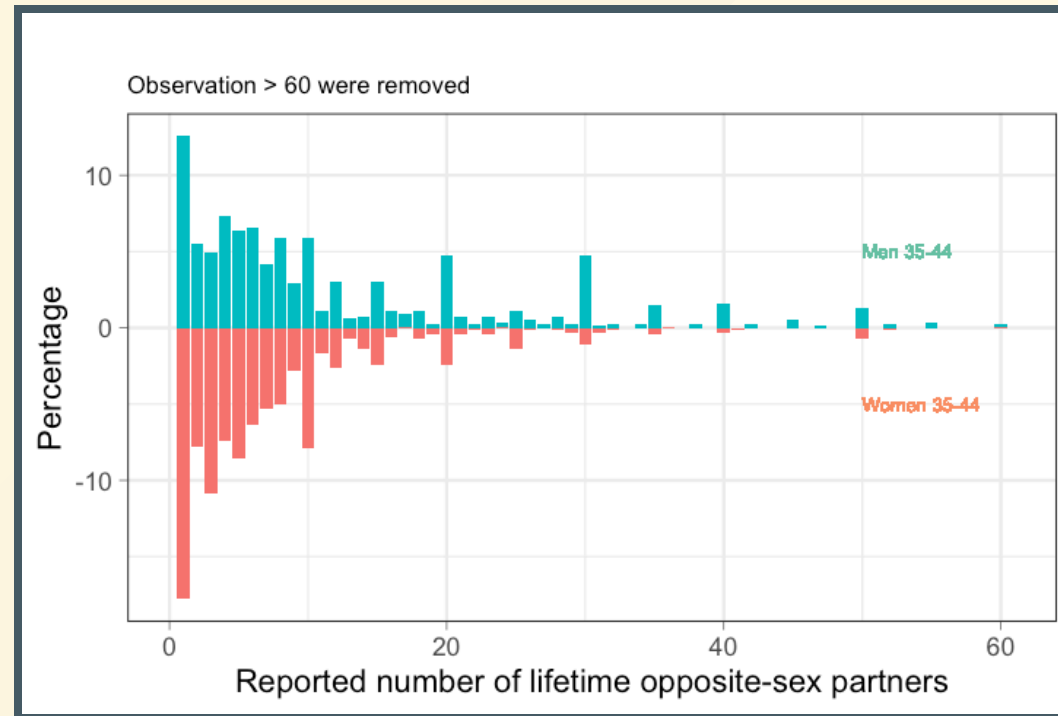
	Uomini 35-44	Donne 35-44
Moda	1	1
Range	0-500	0-550
Media	14.3	8.5
SD	24.2	19.7
Mediana	8	5
IQR	4-18 (14)	3-10 (7)

Share

05:00

Esercizio #9 (bis)

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



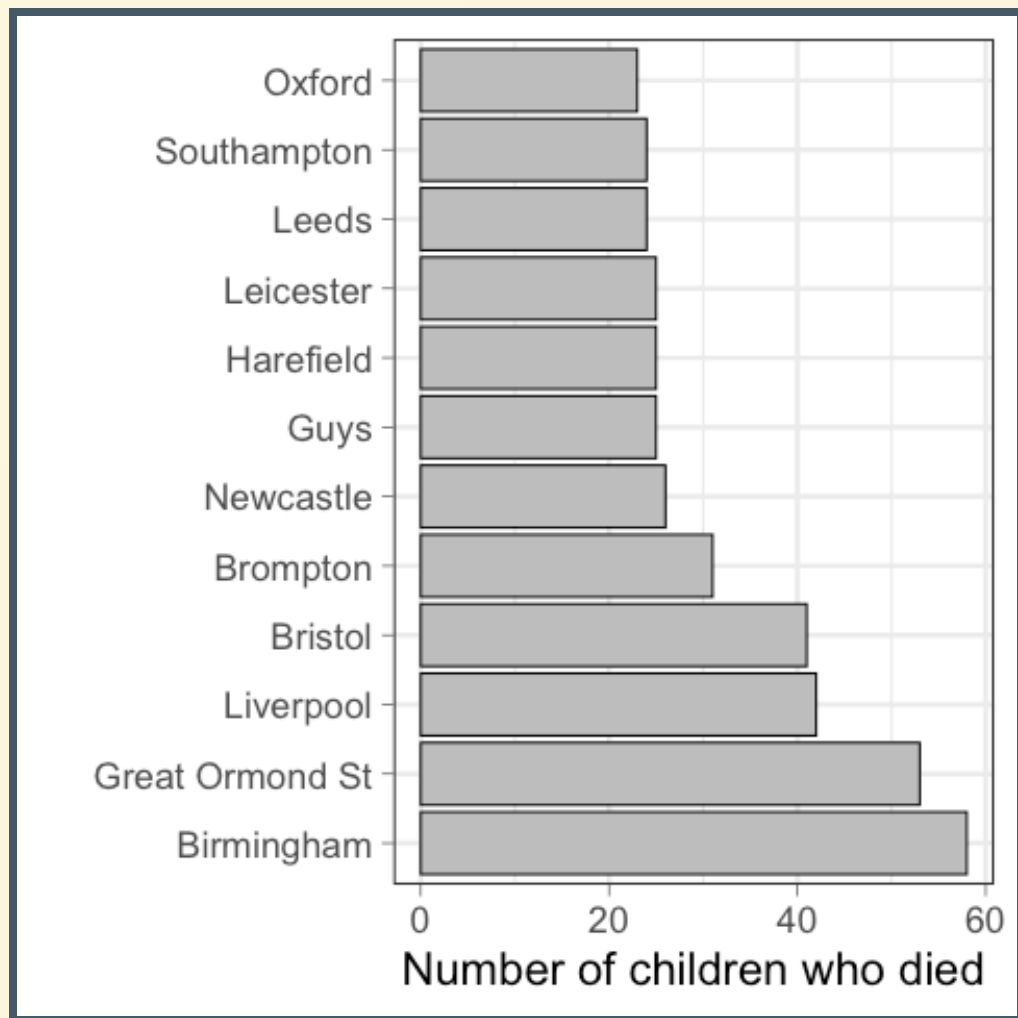
La relazione tra due variabili numeriche

“ Cosa è successo ai bambini sottoposti a interventi cardiocirurugici in alcuni ospedali britannici tra il 1984 e il 1995? ”

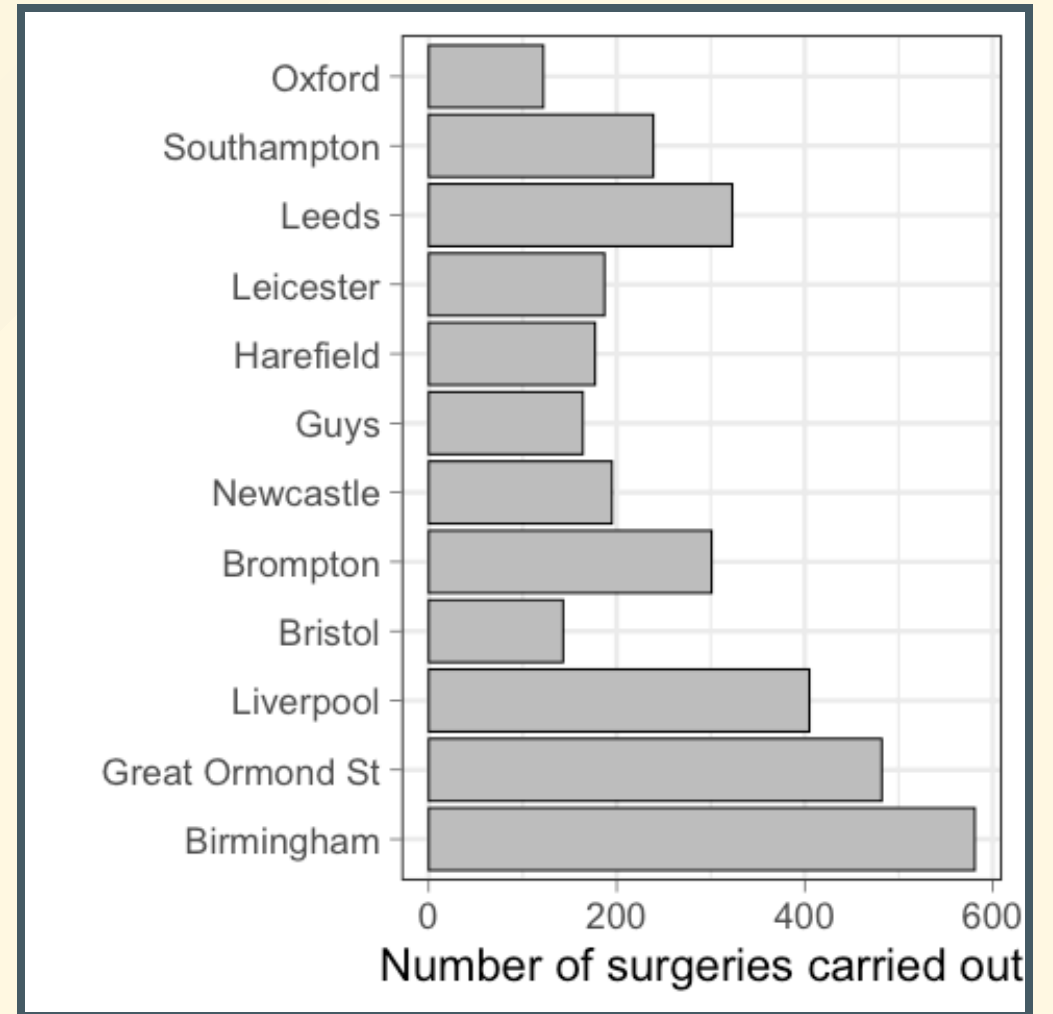
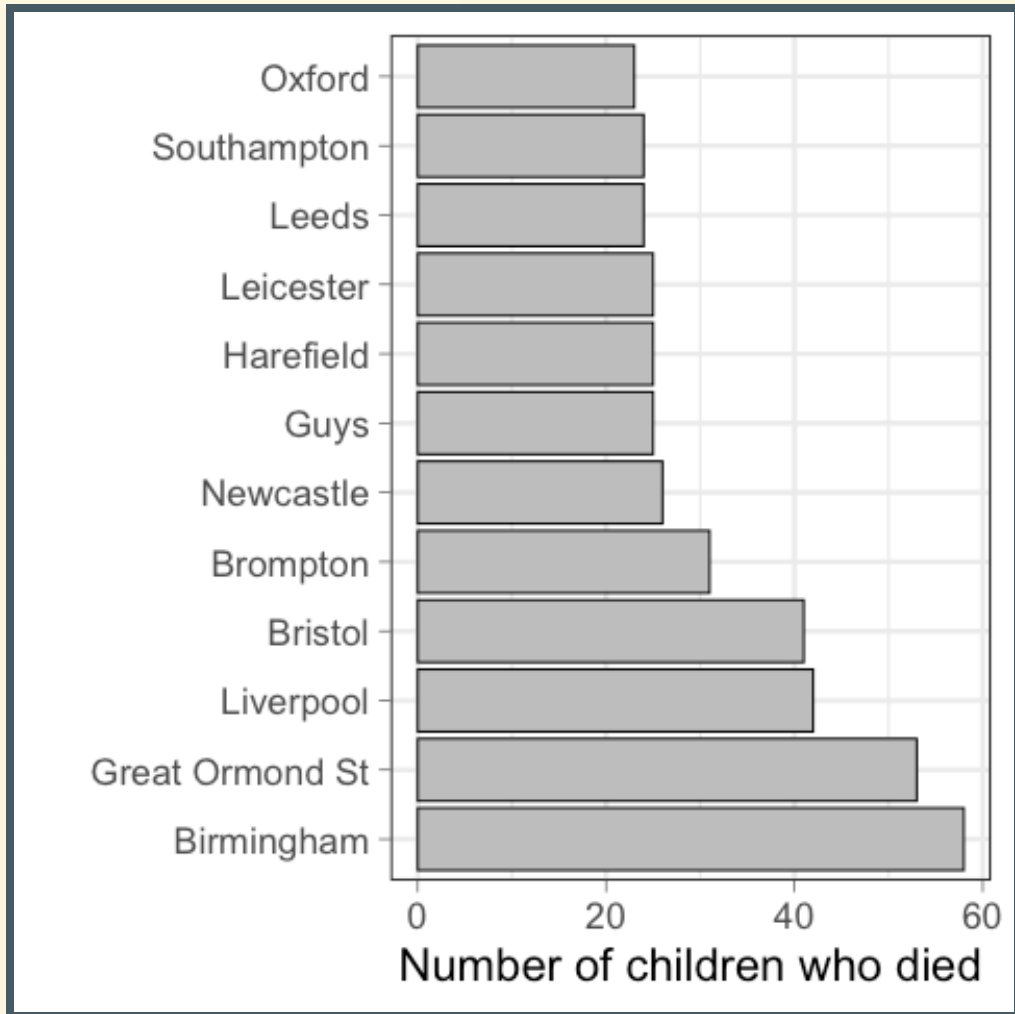
Hospital	Operations	Survivors	Deaths	30-day Survival (%)	Dying (%)
Bristol	143	102	41	71.3	28.7
Leicester	187	162	25	86.6	13.4
Leeds	323	299	24	92.6	7.4
Oxford	122	99	23	81.1	18.9
Guys	164	139	25	84.8	15.2
Liverpool	405	363	42	89.6	10.4
Southampton	239	215	24	90.0	10.0
Great Ormond St	482	429	53	89.0	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 di dati



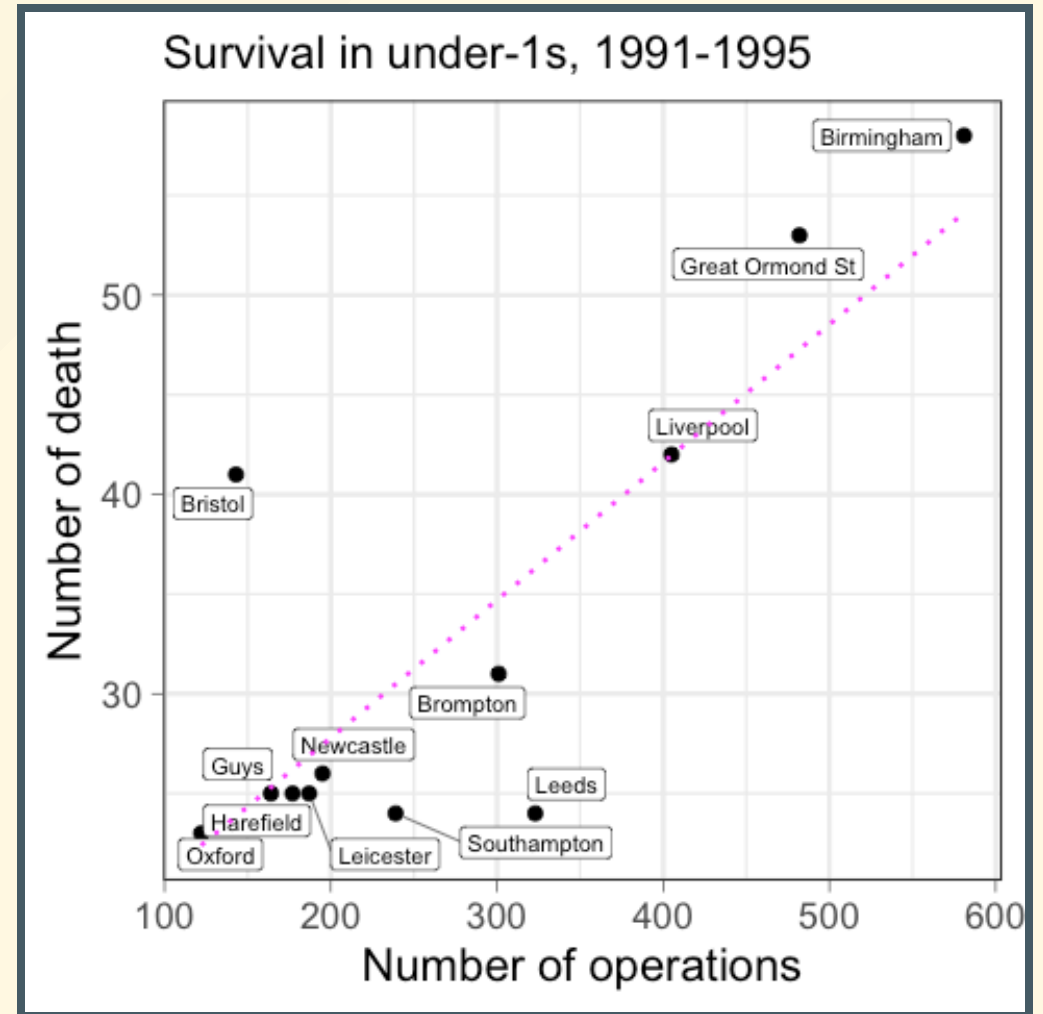
Visualizziamo di dati



La relazione tra due variabili numeriche

Pearson's correlation coefficient

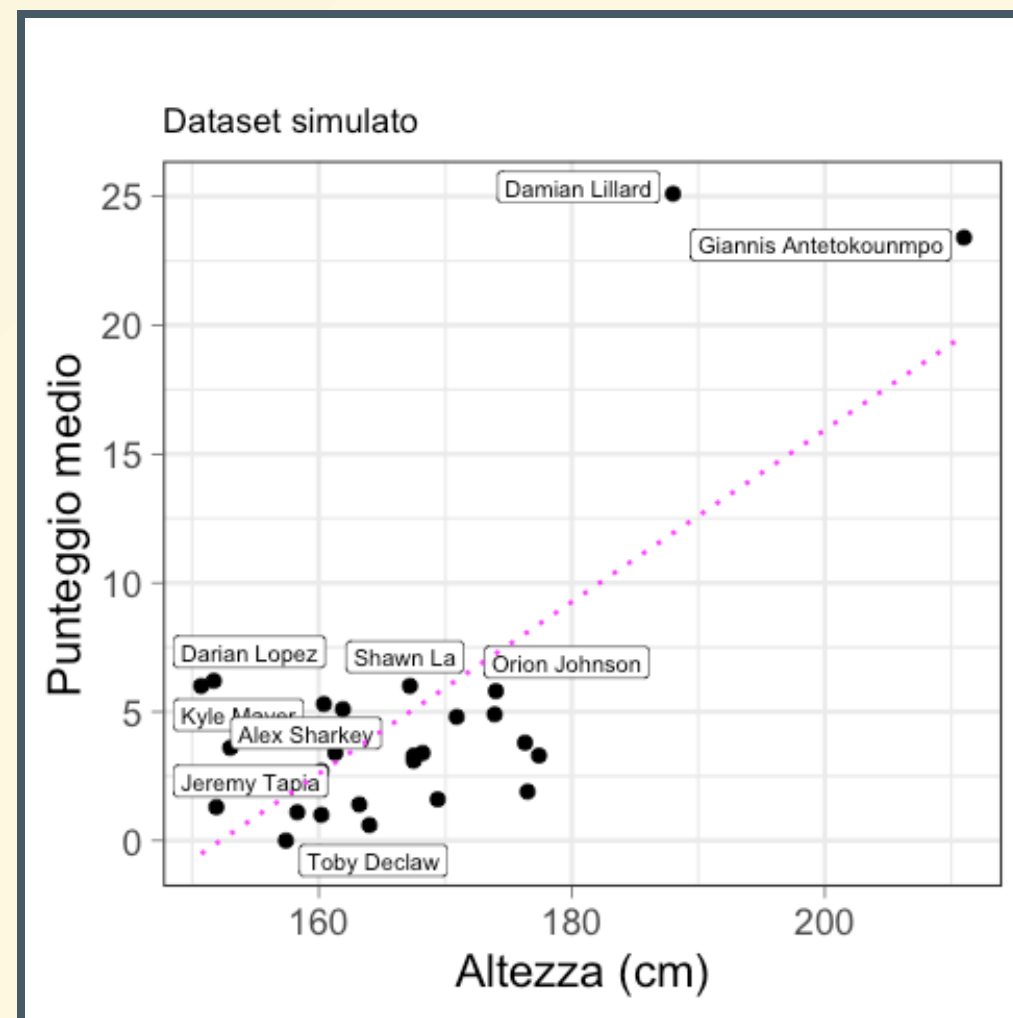
- $\rho = 0.82$
- $\rho_{\text{no Bristol}} = 0.93$



Correlazione & valori estremi

Pearson's correlation coefficient

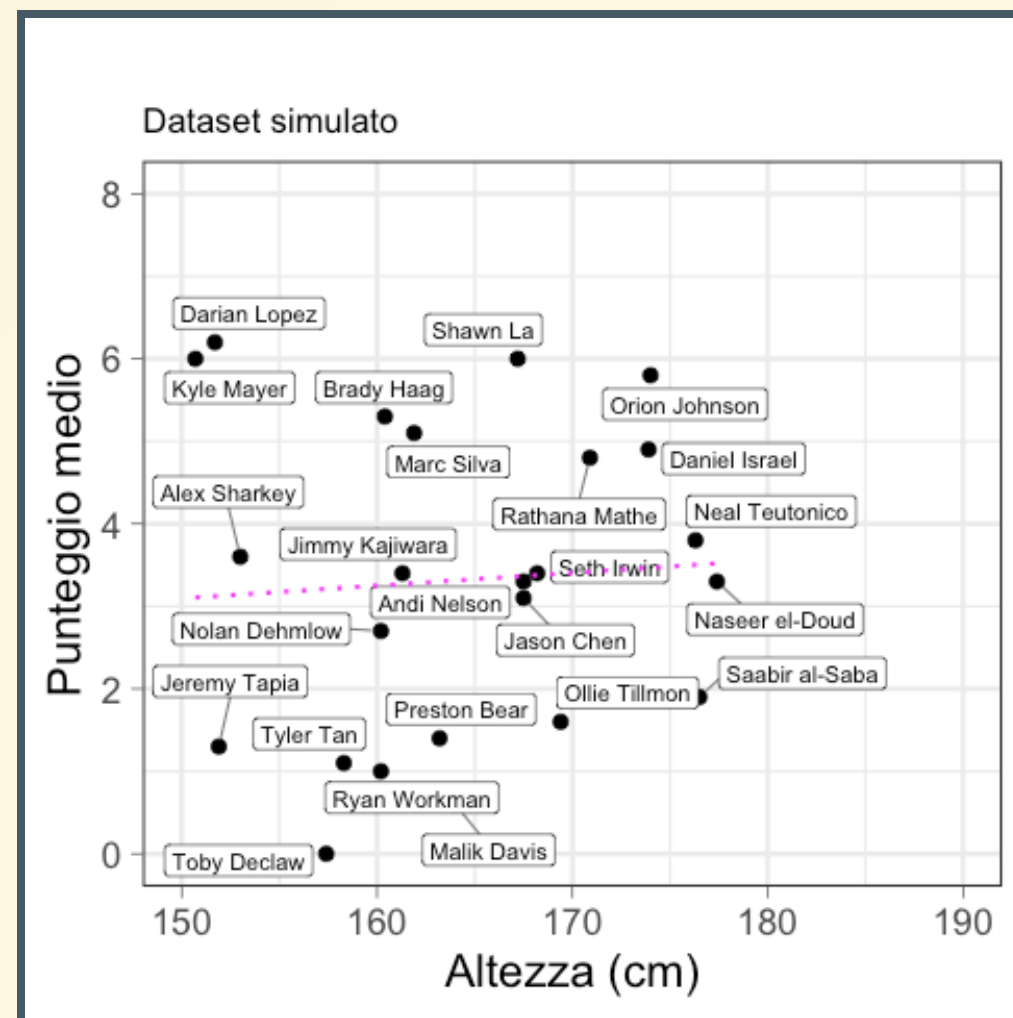
- $\rho = 0.72$



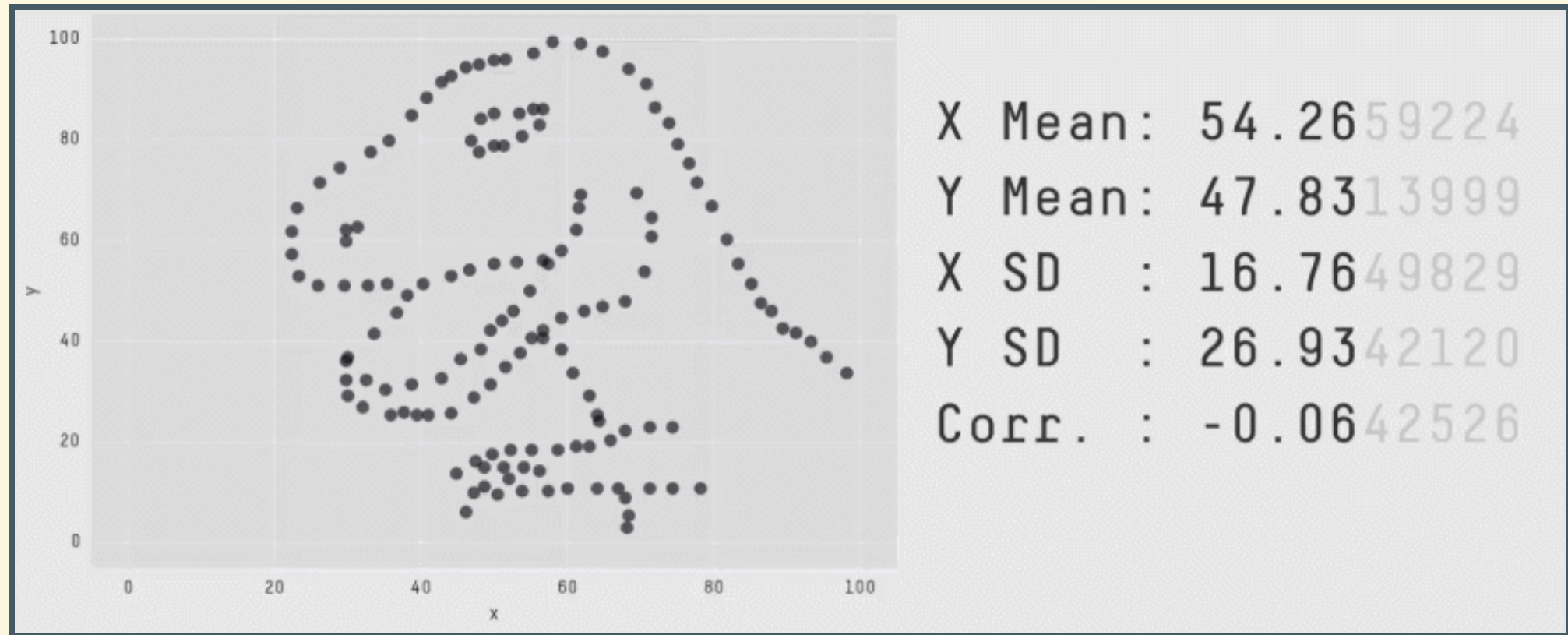
Correlazione & valori estremi

Pearson's correlation coefficient

- $\rho = 0.72$
- $\rho_{\text{no outliers}} = 0.07$



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

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
- Le statistiche possono nascondere dettagli importanti dei dati
- Le variabili numeriche possono essere rappresentate graficamente in diversi modi, ma alcune rappresentazioni possono nascondere dettagli importanti delle distribuzioni sottostanti
- Visualizzare i dati è importante per interpretarli