§17 Statistica descrittiva per variabili numeriche

Consideriamo una variabile statistica X di tipo numerico. Supponiamo che X assuma i valori

$$x_1 \quad x_2 \quad x_3 \quad \dots \quad x_n$$

sugli n individui di una popolazione, o di un campione, che intendiamo studiare (usiamo le lettere maiuscole per le variabili, le minuscole per i loro valori). Definiremo ora alcune misure statistiche che permettono di quantificare determinate proprietà salienti della variabile X.

§17.1 Media, varianza e deviazione standard

Definizione (media)

$$\mu_X = \frac{x_1 + x_2 + \dots + x_n}{n}$$

(il pedice dice di quale variabile si tratta: quando non c'è rischio di confusione, scriviamo semplicemente μ).

La media è una misura di tendenza centrale, ossia intende dare un'idea del valore tipico o centrale di X.

Definizione (varianza)

$$\sigma_X^2 = \frac{(x_1 - \mu_X)^2 + (x_2 - \mu_X)^2 + \dots + (x_n - \mu_X)^2}{n}$$

La varianza è una misura di dispersione, ossia vorrebbe misurare quanto sono sparsi i valori di X.

Definizione (deviazione standard)

$$\sigma_X = \sqrt{\sigma_X^2}$$

Anche la deviazione standard è una misura di dispersione. Ha il vantaggio, rispetto alla varianza, di essere commensurabile con X e μ_X.

Esempio
Se X è misurata in cm, la varianza σ_X^2 è misurata in cm^2, mentre μ_X e σ_X in cm.

Grossolanamente, σ_X è la distanza tipica dei valori di X dalla loro media μ_X. Per esempio, si può dimostrare il fatto seguente.

Fatto
Per almeno il 75% degli individui

$$\mu_X - 2\sigma_X < X < \mu_X + 2\sigma_X$$

Definizione (varianza campionaria)

$$s_X^2 = \frac{(x_1 - \mu_X)^2 + (x_2 - \mu_X)^2 + \ldots + (x_n - \mu_X)^2}{n-1}$$

Questa formula differisce da quella della varianza solo per una piccola correzione al denominatore, che diviene trascurabile per n grandi.

Definizione (deviazione standard campionaria)

$$s_X = \sqrt{s_X^2}$$

La varianza e la deviazione standard campionarie, come la media calcolata su un campione, servono da stimatori, ossia quantità pensate per stimare, usando un campione, le misure corrispondenti relative alla popolazione. Vedremo come si può quantificare l'incertezza in queste stime dovuta alla casualità del campione.

Notazione
Le formule con n e con $n-1$ al denominatore si possono applicare a qualunque insieme di individui, che sia una popolazione o un campione: σ indica sempre la prima, s la seconda. Quando occorre precisare l'insieme su cui una misura è calcolata, lo scriviamo fra parentesi: per esempio $\mu_X(\text{campione})$, $\sigma_X(\text{popolazione})$.

Osservazione
Valgono le seguenti identità algebriche che possono semplificare il calcolo di varianza e varianza campionaria.

$$\sigma_X^2 = \frac{x_1^2 + x_2^2 + \ldots + x_n^2}{n} - \mu_X^2$$
$$s_X^2 = \sigma_X^2 \cdot \frac{n}{n-1}$$

§17.2 Mediana e quantili

Ora descriviamo una categoria di misure che dipendono dai valori di X disposti in ordine crescente. Riordiniamo quindi i numeri

$$x_1 \quad x_2 \quad x_3 \quad \dots \quad x_n$$

ottenendo la lista, identificata dalle parentesi a pedice,

$$x_{(1)} \leqslant x_{(2)} \leqslant x_{(3)} \leqslant \dots \leqslant x_{(n)}$$

Esempio
Se $x_1 = 7$, $x_2 = 3$, $x_3 = 1$, $x_4 = 7$, $x_5 = 3$, allora $x_{(1)} = 1$, $x_{(2)} = 3$, $x_{(3)} = 3$, $x_{(4)} = 7$, $x_{(5)} = 7$.

Definizione (mediana)
La mediana è il valore centrale della lista ordinata. Questo è ben definito se n è dispari, altrimenti si prende la media aritmetica dei due valori centrali.

$$\tilde{X} = \begin{cases} x_{\left(\frac{n+1}{2}\right)} & \text{se } n \text{ è dispari} \\ \frac{1}{2}\left(x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}\right) & \text{se } n \text{ è pari} \end{cases}$$

Come la media, anche la mediana è una misura di tendenza centrale.

Idealmente, la mediana è il valore centrale della lista ordinata. Se siamo interessati al valore che si colloca, per esempio, a $\frac{2}{3}$ della lista, allora calcoliamo il $\frac{2}{3}$-esimo quantile.

Definizione (quantili)
Dato un numero p fra 0 e 1, il p-esimo quantile di X è il valore q_p per cui la frazione di individui sul totale della popolazione che hanno $X < q_p$ è p.

In pratica, se la popolazione è una lista finita di valori, come i dati di un campione, ci troviamo spesso di fronte alla medesima difficoltà della mediana nel caso pari: che la frazione p della lista non cade esattamente su un individuo, ma nell'intervallo fra due individui consecutivi. Possiamo applicare la regola seguente.

Regola (calcolo dei quantili)
Dato un numero reale p con $0 < p < 1$, calcoliamo il p-esimo quantile q_p come segue.

1. Calcoliamo il numero
$$k = 1 + p \cdot (n - 1)$$
2. Dividiamo k nella sua parte intera $\lfloor k \rfloor$ e nella sua parte frazionaria $\mathrm{frac}(k)$.

3. Poniamo

$$\begin{aligned} q_p &= x_{(\lfloor k \rfloor)} \cdot \big(1 - \mathrm{frac}(k)\big) + x_{(\lfloor k \rfloor + 1)} \cdot \mathrm{frac}(k) \\ &= x_{(\lfloor k \rfloor)} + \mathrm{frac}(k) \cdot \big(x_{(\lfloor k \rfloor + 1)} - x_{(\lfloor k \rfloor)}\big) \end{aligned}$$

Esempio
Per esempio, se $k_1 = 4.742$, allora $\lfloor k_1 \rfloor = 4$ e $\mathrm{frac}(k_1) = 0.742$; se $k_2 = 6$, allora $\lfloor k_2 \rfloor = 6$ e $\mathrm{frac}(k_2) = 0$.

Esercizio
Calcola lo 0.75-esimo quantile di

$$x_1 = 4.1 \quad x_2 = 6.2 \quad x_3 = 1.0 \quad x_4 = 2.7 \quad x_5 = 5.6 \quad x_6 = 4.1$$

Soluzione. Abbiamo

$$x_{(1)} = 1.0 \quad x_{(2)} = 2.7 \quad x_{(3)} = 4.1 \quad x_{(4)} = 4.1 \quad x_{(5)} = 5.6 \quad x_{(6)} = 6.2$$

$$k = 1 + 0.75 \cdot (6 - 1) = 4.75 \qquad \lfloor k \rfloor = 4 \qquad \mathrm{frac}(k) = 0.75$$

$$q_{0.75} = x_{(4)} + 0.75 \cdot \big(x_{(5)} - x_{(4)}\big) = 4.1 + 0.75 \cdot 1.5 = 5.225$$

□

Osservazione
Lo 0.5-esimo quantile $q_{0.5}$ coincide esattamente con la mediana $\tilde{X}$.

Nota (altre definizioni)
Per i quantili *di un campione* esistono diverse definizioni. La nostra è quella di R e di Excel; molte calcolatrici prendono invece come 0.75-esimo quantile la mediana della metà superiore dei dati: nell'esercizio, 5.6 anziché 5.225. In questo corso vale la regola qui sopra.

§17.3 Quartili, decili, percentili

(Attenzione: quartili ≠ quantili.) Per $j = 1, 2, 3$ si chiama j-esimo quartile, denotato con Q_j, il quantile $q_{j/4}$. Per cui:

$Q_1 = q_{0.25}$	si dice primo quartile
$Q_2 = q_{0.5} = \tilde{X}$	si dice secondo quartile
$Q_3 = q_{0.75}$	si dice terzo quartile

La differenza interquartile è una misura di dispersione e vale

$$\mathrm{IQR} = Q_3 - Q_1 \qquad \textit{(inter-quartile range)}$$

Osservazione
L'intervallo di valori di X fra Q_1 e Q_3 contiene circa il 50% degli individui. Circa il 25% ha $X < Q_1$, e il restante 25% ha $Q_3 < X$.

Per $j = 1, 2, \ldots, 9$ si chiama j-esimo decile il valore

$$D_j = q_{j/10}$$

Per $j = 1, 2, \ldots, 99$ si chiama j-esimo centile, o anche j-esimo percentile, il valore

$$P_j = q_{j/100}$$

Chiaramente si ha

$$P_{25} = Q_1 = q_{0.25} \qquad P_{50} = Q_2 = q_{0.5} = \tilde{X} \quad \text{(la mediana)} \qquad P_{75} = Q_3 = q_{0.75}$$

inoltre $D_j = P_{10 \cdot j}$.

Esercizio
La variabile statistica X assume sui 10 individui di un campione i valori

$$\begin{array}{lllll} 10.4\,^\circ\mathrm{C} & 8.5\,^\circ\mathrm{C} & 9.1\,^\circ\mathrm{C} & 9.0\,^\circ\mathrm{C} & 10.0\,^\circ\mathrm{C} \\ 11.5\,^\circ\mathrm{C} & 9.2\,^\circ\mathrm{C} & 8.8\,^\circ\mathrm{C} & 10.2\,^\circ\mathrm{C} & 9.7\,^\circ\mathrm{C} \end{array}$$

Calcola μ_X, σ_X, s_X, $\tilde{X}$, IQR.

Soluzione.

$$\begin{aligned} \mu_X &= 9.64\,^\circ\mathrm{C} \\ \sigma_X^2 &= \frac{(10.4\,^\circ\mathrm{C} - 9.64\,^\circ\mathrm{C})^2 + (8.5\,^\circ\mathrm{C} - 9.64\,^\circ\mathrm{C})^2 + \ldots}{10} \\ &= \frac{(0.76\,\mathrm{K})^2 + (-1.14\,\mathrm{K})^2 + \ldots}{10} = 0.7384\,\mathrm{K}^2 \\ \sigma_X &= \sqrt{\sigma_X^2} \approx 0.86\,\mathrm{K} \end{aligned}$$

Nota
La variabile X è una temperatura, misurata sulla scala Celsius. La media μ_X è misurata nella medesima scala. Perché, allora, la deviazione standard σ_X è misurata in kelvin? Per capirlo occorre risalire alle differenze $10.4\,^\circ\mathrm{C} - 9.64\,^\circ\mathrm{C}$ &c., a numeratore nel calcolo di σ_X^2.

Burocraticamente, il sistema SI permette di scrivere le differenze in entrambe le forme:

A. $27\,^\circ\mathrm{C} - 20\,^\circ\mathrm{C} = 7\,\mathrm{K}$ e

B. $27\,^\circ\mathrm{C} - 20\,^\circ\mathrm{C} = 7\,^\circ\mathrm{C}$

sono entrambe corrette. Personalemnte, tuttavia, preferisco la forma A. Questa scrittura ha il vantaggio di sottolineare che il risultato 7 K è una differenza di

temperature — *sette gradi della scala* — e non una temperatura — *il settimo grado della scala*. Temperature e differenze di temperature sono dati di tipo diverso. La situazione è evidente se abbiamo appena calcolato — appunto come nell'esempio di A e B — una differenza, ma c'è rischio di confusione quando la grandezza in questione esce da cacloli più complessi, come per esempio la σ_X appena vista.

$$\begin{aligned} s_X^2 &= \frac{(10.4\,°\mathrm{C} - 9.64\,°\mathrm{C})^2 + (8.5\,°\mathrm{C} - 9.64\,°\mathrm{C})^2 + \dots}{9} \\ &= \frac{(0.76\,\mathrm{K})^2 + (-1.14\,\mathrm{K})^2 + \dots}{9} = 0.8204\,\mathrm{K}^2 \\ s_X &= \sqrt{s_X^2} \approx 0.91\,\mathrm{K} \end{aligned}$$

Mettendo in ordine i dati otteniamo

$x_{(9)}$	$x_{(1)}$	$x_{(4)}$	$x_{(3)}$	$x_{(7)}$
10.4 °C	8.5 °C	9.1 °C	9.0 °C	10.0 °C
11.5 °C	9.2 °C	8.8 °C	10.2 °C	9.7 °C
$x_{(10)}$	$x_{(5)}$	$x_{(2)}$	$x_{(8)}$	$x_{(6)}$

Quindi

$$\begin{aligned} \tilde{X} &= \frac{1}{2}\left(x_{(5)} + x_{(6)}\right) = 9.45\,°\mathrm{C} \\ Q_1 &= q_{0.25} = x_{(3)} + 0.25\left(x_{(4)} - x_{(3)}\right) = 9.025\,°\mathrm{C} \\ Q_3 &= q_{0.75} = x_{(7)} + 0.75\left(x_{(8)} - x_{(7)}\right) = 10.15\,°\mathrm{C} \\ \mathrm{IQR} &= Q_3 - Q_1 \approx 1.13\,\mathrm{K} \end{aligned}$$

□

§17.4 Istogrammi

Gli istogrammi sono una maniera conveniente per rappresentare la distribuzione di variabili numeriche. Vediamo un esempio: vogliamo rappresentare la distribuzione della variabile statura che assume i valori seguenti, in cm, su un campione di 36 individui.

15	0468899
16	000013334455555778889
17	00115588

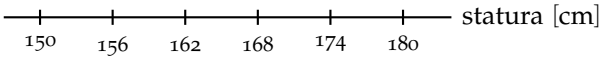
Equivale alla lista 150, 154, 156, 158, 158, 159, 159, 160, ...: a sinistra della barra ci sono centinaia e decine, a destra le unità.

Nota
Questa rappresentazione dei dati si chiama diagramma ramo-foglia (*stem-and-leaf plot*). I dati sono le stature delle studentesse del corso, raccolte a lezione.

Tutti i valori ricadono nell'intervallo 150–180. Cominciamo col dividere questo intervallo in un numero ragionevole di classi, per esempio

150–156 156–162 162–168 168–174 174–180

Per convenzione, intendiamo le classi con l'estremo sinistro incluso e il destro escluso. Non è necessario che tutte le classi della suddivisione abbiano la medesima larghezza. Rappresentiamo la suddivisione così ottenuta sull'asse delle ascisse.



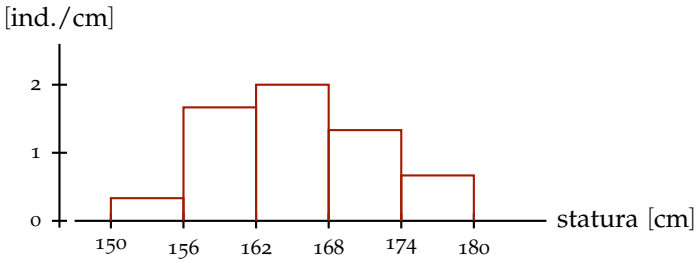
Ora, su ciascuna classe, solleviamo una colonna la cui area rappresenta il numero di individui in quella classe.

150–156	156–162	162–168	168–174	174–180
2 ind.	10 ind.	12 ind.	8 ind.	4 ind.

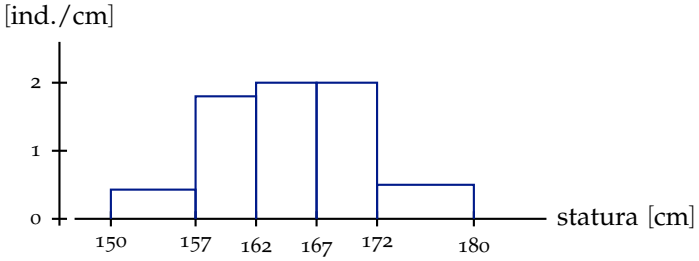
Siccome i lati orizzontali delle colonne sono misurati in cm, affinché le aree risultino misurate in individui, le altezze devono essere misurate in individui/cm. Queste quantità si chiamano densità di frequenza, e si calcolano dividendo il numero di individui in ciascuna colonna per la larghezza della colonna stessa.

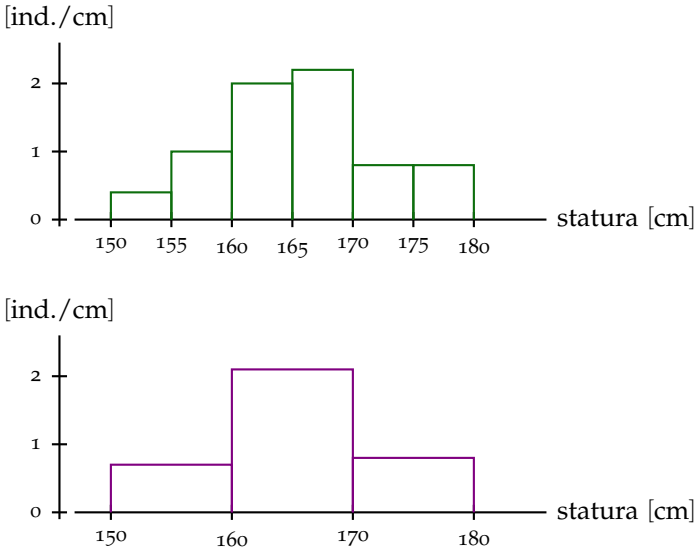
150–156	156–162	162–168	168–174	174–180
2 ind.	10 ind.	12 ind.	8 ind.	4 ind.
$0.33\ \frac{\text{ind.}}{\text{cm}}$	$1.67\ \frac{\text{ind.}}{\text{cm}}$	$2.00\ \frac{\text{ind.}}{\text{cm}}$	$1.33\ \frac{\text{ind.}}{\text{cm}}$	$0.67\ \frac{\text{ind.}}{\text{cm}}$

L'istogramma risultante è come segue.

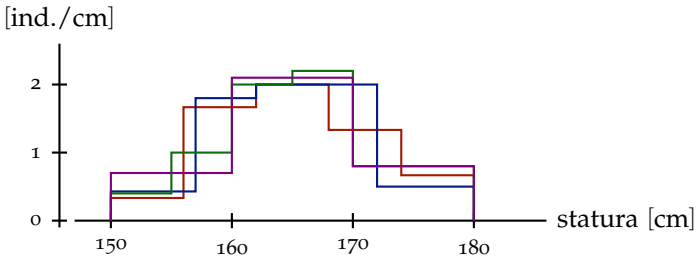


La scelta delle classi è arbitraria: ecco altri esempi.





Sovrapponendo questi grafici possiamo osservare che, grossolanamente, la forma dell'istogramma non dipende dalla scelta delle classi, purché questa sia ragionevole.



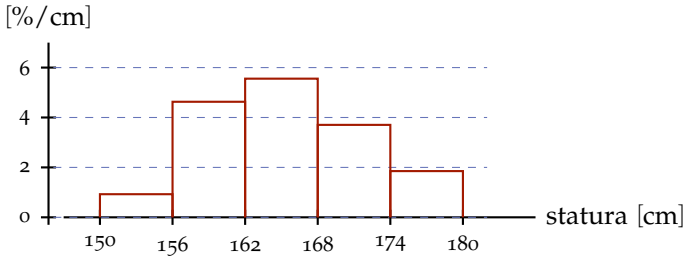
Se avessimo rappresentato il numero di individui tramite l'altezza anziché l'area delle colonne, i diversi grafici non sarebbero stati in accordo fra loro.

La densità di frequenza è stata definita in maniera tale che l'area di ciascuna colonna sia il numero di individui nella classe corrispondente. Di conseguenza, l'area dell'intero istogramma è il numero totale degli individui.

C'è una seconda maniera di graduare l'asse delle ordinate, detta densità di probabilità. La densità di probabilità è semplicemente la densità di frequenza divisa per il totale degli individui (la numerosità del campione — o della popolazione). In questo modo, l'area dell'istogramma è 1. L'unità di misura della densità di frequenza è individui/unità delle ascisse; quella della densità di probabilità è 1/unità delle ascisse. La densità di probabilità si può altresì esprimere in percentuale: in questo caso l'unità sarà %/unità delle ascisse e l'area dell'istogramma 100%.

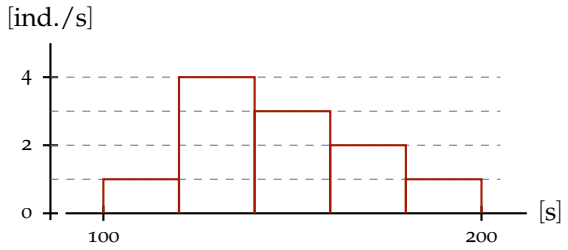
Nel nostro esempio abbiamo

150–156	156–162	162–168	168–174	174–180
2 ind.	10 ind.	12 ind.	8 ind.	4 ind.
$0.33\ \frac{\text{ind.}}{\text{cm}}$	$1.67\ \frac{\text{ind.}}{\text{cm}}$	$2.00\ \frac{\text{ind.}}{\text{cm}}$	$1.33\ \frac{\text{ind.}}{\text{cm}}$	$0.67\ \frac{\text{ind.}}{\text{cm}}$
$0.9\ \frac{\%}{\text{cm}}$	$4.6\ \frac{\%}{\text{cm}}$	$5.6\ \frac{\%}{\text{cm}}$	$3.7\ \frac{\%}{\text{cm}}$	$1.9\ \frac{\%}{\text{cm}}$

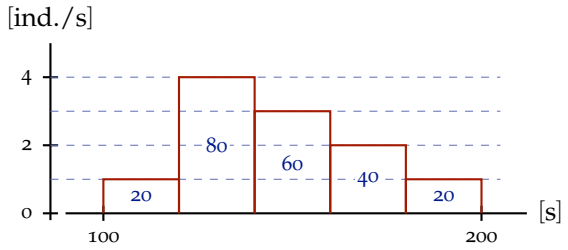


Osservazione
La forma dell'istogramma non cambia: è solo la scala delle ordinate a cambiare.

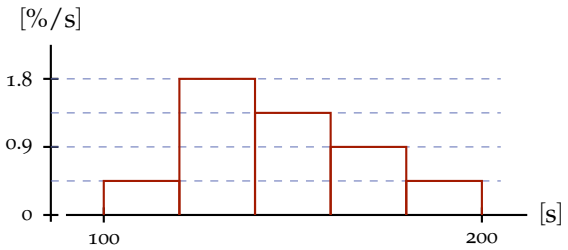
Esercizio
Converti la scala delle ordinate da densità di frequenza a densità di probabilità nel seguente istogramma.



Soluzione. Dobbiamo determinare la numerosità del campione rappresentato. Calcoliamo quindi l'area di ciascuna colonna.



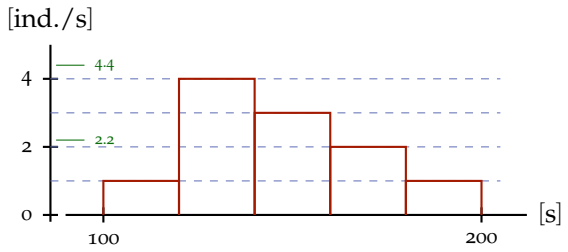
Abbiamo quindi $n = 220$. Potremmo ora calcolare che 2 ind./s equivalgono a $\frac{2}{220} \cdot 100\,\%/s = 0.9\,\%/s$ e similmente 4 ind./s sono 1.8 %/s. Otterremmo così il grafico seguente.



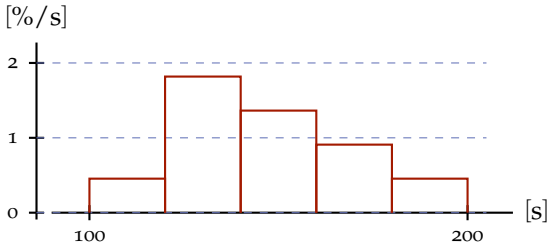
Questo risultato è però insoddisfacente, perché sarebbe meglio avere le cifre tonde 1 %/s e 2 %/s sull'asse delle ordinate. Calcoliamo dunque a quali densità di frequenza corrispondono queste densità di probabilità:

$$1\,\frac{\%}{\text{s}} \text{ corrisponde a } \frac{1\%}{100\%} \cdot 220\,\frac{\text{ind.}}{\text{s}} = 2.2\,\frac{\text{ind.}}{\text{s}}$$
$$2\,\frac{\%}{\text{s}} \text{ corrisponde a } \frac{2\%}{100\%} \cdot 220\,\frac{\text{ind.}}{\text{s}} = 4.4\,\frac{\text{ind.}}{\text{s}}$$

Usiamo ora queste equivalenze per trovare la posizione di 1 %/s e 2 %/s sull'asse delle ordinate.



Per cui questo è l'istogramma richiesto.



□

§17.5 Diagrammi a scatola

Un modo compatto di rappresentare la distribuzione di una variabile numerica, molto usato negli articoli scientifici, è il diagramma a scatola (*box plot*, o *box-and-whisker plot*: diagramma a scatola e baffi). Si costruisce a partire dai quartili.

- Si disegna una scatola che va da Q_1 a Q_3: la sua lunghezza è la differenza interquartile, e contiene circa metà degli individui.
- Dentro la scatola si segna la mediana.
- Dai due lati della scatola partono due baffi, che arrivano fino al valore più piccolo e al valore più grande dei dati che non distano dalla scatola più di $1.5 \cdot \text{IQR}$, cioè che stanno fra
$$Q_1 - 1.5 \cdot \text{IQR} \qquad \text{e} \qquad Q_3 + 1.5 \cdot \text{IQR}$$
- I dati che cadono fuori da questo intervallo si disegnano uno per uno, come punti: sono gli outlier (in italiano: valori anomali).

Esercizio
Disegna il diagramma a scatola delle stature delle studentesse ($n = 36$). La lista ordinata è quella del diagramma ramo-foglia.

Soluzione. Per Q_1, $\tilde{X}$ e Q_3 troviamo rispettivamente

$$k = 1 + 0.25 \cdot 35 = 9.75 \qquad k = 1 + 0.5 \cdot 35 = 18.5 \qquad k = 1 + 0.75 \cdot 35 = 27.25$$

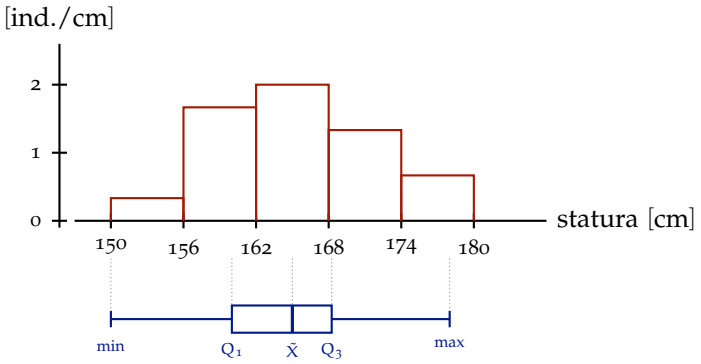
quindi

$$\begin{aligned} Q_1 &= x_{(9)} + 0.75\,(x_{(10)} - x_{(9)}) = 160\,\text{cm} + 0.75 \cdot 0\,\text{cm} = 160\,\text{cm} \\ \tilde{X} &= x_{(18)} + 0.5\,(x_{(19)} - x_{(18)}) = 165\,\text{cm} + 0.5 \cdot 0\,\text{cm} = 165\,\text{cm} \\ Q_3 &= x_{(27)} + 0.25\,(x_{(28)} - x_{(27)}) = 168\,\text{cm} + 0.25 \cdot 1\,\text{cm} = 168.25\,\text{cm} \end{aligned}$$

Dunque $\text{IQR} = 8.25\,\text{cm}$, e i baffi possono arrivare fino a

$$Q_1 - 1.5 \cdot \text{IQR} = 147.625\,\text{cm} \qquad \text{e} \qquad Q_3 + 1.5 \cdot \text{IQR} = 180.625\,\text{cm}$$

Tutti i dati cadono in questo intervallo: i baffi arrivano quindi al minimo, 150 cm, e al massimo, 178 cm, e non ci sono outlier. Ecco il diagramma, sotto l'istogramma degli stessi dati.

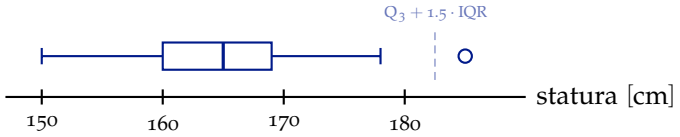


□

Esempio
Aggiungiamo, fittiziamente, ai dati una studentessa alta 185 cm. Abbiamo allora $n = 37$ e

$$Q_1 = x_{(10)} = 160\,\text{cm} \qquad \tilde{X} = x_{(19)} = 165\,\text{cm} \qquad Q_3 = x_{(28)} = 169\,\text{cm}$$

quindi $\text{IQR} = 9\,\text{cm}$, e i baffi possono arrivare fino a $160\,\text{cm} - 13.5\,\text{cm} = 146.5\,\text{cm}$ e a $169\,\text{cm} + 13.5\,\text{cm} = 182.5\,\text{cm}$. Il valore 185 cm è fuori: è un outlier, e si disegna a parte. Il baffo di destra arriva al più grande dei valori rimasti, 178 cm.



Un outlier non è necessariamente un errore: può essere un individuo davvero fuori dall'ordinario, oppure un errore di misura o di trascrizione (una statura di 1.85 m scritta come 185 m, per esempio). Il diagramma a scatola lo mette in evidenza, e conviene andare a vedere da dove viene. Non si scartano gli outlier soltanto perché sono lontani dagli altri dati.

Nota
Il diagramma a scatola dice meno dell'istogramma: per esempio, da un diagramma a scatola non si può capire se la distribuzione ha una sola gobba, due, o più — rispettivamente distribuzioni unimodali, bimodali, multimodali. In compenso è molto compatto, quindi si possono affiancare i diagrammi a scatola di diverse distribuzioni, e ciò permette di confrontare a colpo d'occhio molti gruppi.

Nota importante
La regola dell'$1.5 \cdot \mathrm{IQR}$ (dovuta a John Tukey) è la più diffusa, ma non l'unica: a volte i baffi arrivano semplicemente al minimo e al massimo dei dati. Inoltre, programmi e calcolatrici diversi calcolano i quartili in modi leggermente diversi. Un articolo fatto bene dice quale convenzione ha usato.

§17.6 Esercizi

Test 89 (soluzione a p. 304)
Calcola la deviazione standard di 0, 1, 1.5, 2, 3.

Test 90 (soluzione a p. 304)
Calcola la deviazione standard di 3, 3, 6, 6, 6, 12.

Test 91 (soluzione a p. 305)
Quanto vale la mediana dei numeri 7, −7, 1, 2, −8, 15, 0?

Test 92 (soluzione a p. 305)
A che percentile equivale il primo quartile?

Test 93 (soluzione a p. 305)
A che percentile equivale la mediana?

Esercizio 94 (soluzione a p. 305)
In un campione di fumatori, il numero di sigarette consumate al giorno si distribuisce come segue.

sigarette	fumatori
$0 \leqslant \ldots < 10$	225
$10 \leqslant \ldots < 20$	525
$20 \leqslant \ldots < 40$	450
$40 \leqslant \ldots < 80$	300
$80 \leqslant \ldots$	0

Traccia l'istogramma, con in ordinata la percentuale di fumatori per sigaretta. Qual è il numero mediano di sigarette fumate al giorno? Approssimativamente, che percentuale del totale dei fumatori consuma più di 30 sigarette al giorno? Dai una stima del numero medio di sigarette al giorno spiegando il tuo ragionamento.

Test 89 (p. 124)
Calcola la deviazione standard di 0, 1, 1.5, 2, 3.

Soluzione. Le formule che si possono usare sono due, entrambe conducono alla risposta corretta:

$$\sigma = \sqrt{\text{media dei quadrati degli scarti}} \qquad \sigma = \sqrt{\text{media dei quadrati} - \text{media}^2}$$

Con la prima: la media è $\mu = \frac{0+1+1.5+2+3}{5} = 1.5$, gli scarti sono -1.5, -0.5, 0, 0.5, 1.5, quindi

$$\sigma = \sqrt{\frac{(-1.5)^2 + (-0.5)^2 + 0^2 + 0.5^2 + 1.5^2}{5}} = \sqrt{1} = 1$$

Con la seconda:

$$\sigma = \sqrt{\frac{0^2 + 1^2 + 1.5^2 + 2^2 + 3^2}{5} - 1.5^2} = \sqrt{3.25 - 2.25} = 1$$

Test 90 (p. 124)
Calcola la deviazione standard di 3, 3, 6, 6, 6, 12.

Soluzione. Direttamente: $\mu = \frac{3+3+6+6+6+12}{6} = 6$ e

$$\sigma = \sqrt{\frac{(3-6)^2 \cdot 2 + (6-6)^2 \cdot 3 + (12-6)^2}{6}} = \sqrt{\frac{18+0+36}{6}} = 3$$

(il $\cdot 2$ e il $\cdot 3$ contano le ripetizioni).

A volte, calcoli come questi possono essere abbreviati dalle seguenti osservazioni.

- Sommando una costante A a tutti i numeri di un certo insieme, si somma A alla loro media, mentre la deviazione standard non cambia.
- Moltiplicando tutti i numeri di un certo insieme per la medesima costante positiva B, sia la loro media sia la loro deviazione standard si moltiplicano per B.

Sommiamo -6 ai numeri dati, ottenendo -3, -3, 0, 0, 0, 6: la deviazione standard non cambia. Ora moltiplichiamo per $\frac{1}{3}$, ottenendo -1, -1, 0, 0, 0, 2: la deviazione standard si divide per 3. Quindi

$$\frac{\sigma}{3} = \sqrt{\frac{(-1)^2 \cdot 2 + 0^2 \cdot 3 + 2^2}{6} - \left(\frac{(-1) \cdot 2 + 0 \cdot 3 + 2}{6}\right)^2} = \sqrt{1 - 0} = 1$$

da cui $\sigma = 3$.

Test 91 (p. 124)
Quanto vale la mediana dei numeri 7, −7, 1, 2, −8, 15, 0?

Soluzione. Basta mettere i numeri in ordine e prendere quello centrale:

$$-8 \quad -7 \quad 0 \quad 1 \quad 2 \quad 7 \quad 15$$

Test 92 (p. 124)
A che percentile equivale il primo quartile?

Soluzione. Il primo quartile di una variabile statistica è il valore di quella variabile al di sotto del quale cade un quarto della popolazione. Quindi è il 25-esimo percentile, ossia il valore sotto cui cade il 25% della popolazione. (Con le definizioni del capitolo 17: $Q_1 = q_{0.25} = P_{25}$.)

Test 93 (p. 125)
A che percentile equivale la mediana?

Soluzione. Sotto la mediana cade metà della popolazione. Quindi la mediana è il 50-esimo percentile. (Con le definizioni del capitolo 17: $\tilde{X} = q_{0.5} = P_{50}$.)

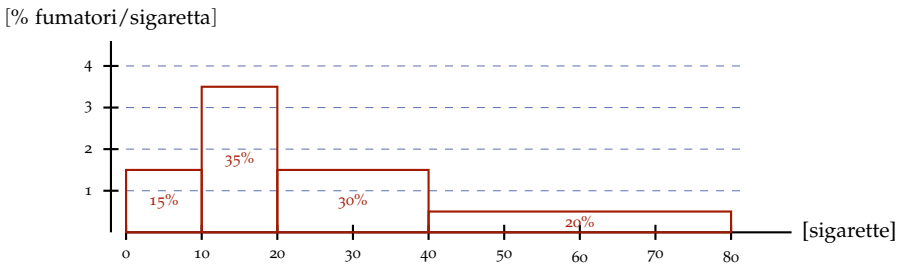
Esercizio 94 (p. 125)
In un campione di fumatori, il numero di sigarette consumate al giorno si distribuisce come segue.

sigarette	fumatori
$0 \leqslant \ldots < 10$	225
$10 \leqslant \ldots < 20$	525
$20 \leqslant \ldots < 40$	450
$40 \leqslant \ldots < 80$	300
$80 \leqslant \ldots$	0

Traccia l'istogramma, con in ordinata la percentuale di fumatori per sigaretta. Qual è il numero mediano di sigarette fumate al giorno? Approssimativamente, che percentuale del totale dei fumatori consuma più di 30 sigarette al giorno? Dai una stima del numero medio di sigarette al giorno spiegando il tuo ragionamento.

Soluzione. L'asse delle ordinate è in percentuale di fumatori per sigaretta, quindi le aree delle colonne rappresentano percentuali di fumatori. Il totale è $225 + 525 + 450 + 300$, cioè 1500 individui, e le percentuali sono

sigarette	fumatori	% fumatori	altezza
0–10	225	15%	$\frac{15\%}{10\text{ sig.}} = 1.5\,\%/\text{sig.}$
10–20	525	35%	$\frac{35\%}{10\text{ sig.}} = 3.5\,\%/\text{sig.}$
20–40	450	30%	$\frac{30\%}{20\text{ sig.}} = 1.5\,\%/\text{sig.}$
40–80	300	20%	$\frac{20\%}{40\text{ sig.}} = 0.5\,\%/\text{sig.}$



La mediana è la soglia che divide a metà l'area dell'istogramma: il $15\% + 35\% = 50\%$ dei fumatori sta sotto le 20 sigarette al giorno, quindi la mediana è 20 sigarette.

Più di 30 sigarette: è l'area a destra di 30, cioè metà della colonna del 30% più tutta quella del 20%: $\frac{1}{2} \cdot 30\% + 20\% = 35\%$.

Per la media, il meglio che possiamo fare è supporre che i fumatori di ciascuna colonna siano equidistribuiti, quindi fumino in media quanto il centro della colonna. Su 100 fumatori:

$$15 \cdot \frac{0+10}{2} + 35 \cdot \frac{10+20}{2} + 30 \cdot \frac{20+40}{2} + 20 \cdot \frac{40+80}{2} = 2700$$

sigarette al giorno, quindi in media $\frac{2700}{100} = 27$ sigarette al giorno a testa.

Nota: la media, 27, è maggiore della mediana, 20. Questo era atteso, perché l'istogramma ha una *coda* verso destra.

Test 95 (p. 133)
In una distribuzione gaussiana, che percentuale della popolazione ricade

1. entro una deviazione standard dalla media?
2. entro due deviazioni standard dalla media?
3. fra la media e la media più una deviazione standard?
4. fra la media meno due deviazioni standard e la media?

Soluzione. A queste domande possiamo rispondere ricordando due fatti mnemonici: il 68% della popolazione ricade entro una deviazione standard dalla media, il 95% entro due. Questi due numeri rispondono alle prime due domande: 68% e 95%. Sottraendo le due aree fra loro, e dal totale del 100%, otteniamo

$$95\% - 68\% = 27\% \qquad 100\% - 95\% = 5\%$$

e sfruttando la simmetria della gaussiana

