

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. Definiremo ora alcune **misure statistiche** che permettono di quantificare determinate proprietà salienti della variabile x .

Media $\left\{ \begin{array}{l} \bar{x} \text{ si preferisce per la m. campionaria} \\ \mu_x \text{ o semplicemente } \mu \text{ per la m. di una popolazione} \end{array} \right.$

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

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

Varianza

$$\begin{aligned} \sigma_x^2 &= \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \dots + (x_n - \bar{x})^2}{n} \\ &= \frac{x_1^2 + x_2^2 + \dots + x_n^2}{n} - \bar{x}^2 \end{aligned}$$

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

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 .

Es. 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:

Per almeno il 75% degli individui

$$\mu_x - 2\sigma_x < x < \mu_x + 2\sigma_x$$

Varianza campionaria

$$s_x^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \dots + (x_n - \bar{x})^2}{n-1}$$

Questa formula differisce da quella della varianza

per una piccola correzione al denominatore, che diviene trascurabile per n grandi.

Deviazione standard campionaria

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

La varianza, la deviazione standard e la media campionaria sono *stimatori*, ossia quantità pensate per stimare, usando un campione, le misure corrispondenti relative alla popolazione.

Vedremo come si può quantificare l'incertezza in questa stima dovuta alla casualità del campione.

Una descriviamo una categoria di misura 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

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

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

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.

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

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

Quantili

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.

Dato un numero reale p con $0 < p < 1$ calcoliamo il p -esimo quantile q_p come segue. Dapprima calcoliamo il numero

$$k = 1 + p \cdot (n-1)$$

Quindi dividiamo k nella sua parte intera $\lfloor k \rfloor$ e

la sua parte frazionaria $\text{frac}(k)$.

$$\begin{array}{lll} \underline{\text{Es.}} & k_1 = 4.742 & \lfloor k_1 \rfloor = 4 \quad \text{frac}(k_1) = 0.742 \\ & k_2 = 6 & \lfloor k_2 \rfloor = 6 \quad \text{frac}(k_2) = 0 \end{array}$$

Abbiamo ora

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

Es. Cerchiamo il $\frac{3}{4}$ -esimo quantile di

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

Abbiamo

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

$$k = 1 + \frac{3}{4} (6 - 1) = 4.75 \quad \lfloor k \rfloor = 4 \quad \text{frac}(k) = 0.75$$

$$q_{3/4} = x_{(4)} + 0.75 \cdot (x_{(5)} - x_{(4)}) = 5.225$$

Nota: lo 0.5-esimo quantile $q_{0.5}$ coincide esattamente con la mediana $\tilde{x}$.

QUARTILI $\neq$ QUANTILI

Quartili & differenza interquartile

Per $n=1,2,3$ si chiama n -esimo quartile, denotato con Q_n , il quantile $g_{n/4}$. Per cui:

$Q_1 = g_{0.25}$ si dice primo quartile

$Q_2 = g_{0.5} = \tilde{x}$ si dice secondo quartile

$Q_3 = g_{0.75}$ si dice terzo quartile

La differenza interquartile è una misura di dispersione e vale

$$IQR = Q_3 - Q_1 \quad (\text{INTER-QUARTILE RANGE})$$

Nota: 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 $Q_3 < x$.

Decili & centili o percentili

Per $n=1,2,\dots,9$ si chiama n -esimo decile il valore

$$D_n = g_{n/10}$$

Per $n=1,2,\dots,99$ si chiama n -esimo centile o anche

n-esimo percentile il valore

$$P_n = q_{\frac{n}{100}}$$

Chiaramente si ha

$$P_{25} = Q_1 = q_{0.25}$$

$$P_{50} = Q_2 = q_{0.5} = \tilde{x} \leftarrow \text{la mediana}$$

$$P_{75} = Q_3 = q_{0.75}$$

inoltre $D_n = P_{10 \cdot n}$

E₅ La variabile statistica x assume sui 10 individui di un campione i valori

$$10.4^{\circ}\text{C} \quad 8.5^{\circ}\text{C} \quad 9.1^{\circ}\text{C} \quad 9.0^{\circ}\text{C} \quad 10.0^{\circ}\text{C}$$

$$11.5^{\circ}\text{C} \quad 9.2^{\circ}\text{C} \quad 9.8^{\circ}\text{C} \quad 10.2^{\circ}\text{C} \quad 9.7^{\circ}\text{C}$$

calcola $\bar{x}$, σ_x , s_x , $\tilde{x}$, IQR

Soluzione

$$\underline{\bar{x} = 9.64^{\circ}\text{C}}$$

$$\sigma_x^2 = \frac{(10.4^{\circ}\text{C} - 9.64^{\circ}\text{C})^2 + (8.5^{\circ}\text{C} - 9.64^{\circ}\text{C})^2 + \dots}{10}$$

$$\sigma_x^2 = \frac{(0.76 \text{ K})^2 + (-1.14 \text{ K})^2 + \dots}{10} = 0.7384 \text{ K}^2$$

$$\sigma_x = \sqrt{\sigma_x^2} = 0.86 \text{ K}$$

Nota: la variabile x è una temperatura, misurata sulla scala Celsius. La media $\bar{x}$ è misurata nella medesima scala. Perché, allora, la deviazione standard σ_x è misurata in Kelvin? Per capirlo occorre risalire alle differenze $10.4^\circ\text{C} - 9.64^\circ\text{C}$ etc. a numeratore nel calcolo di σ_x^2 . Quale fra queste è corretta:

A $27^\circ\text{C} - 20^\circ\text{C} = 7 \text{ K}$ oppure

B $27^\circ\text{C} - 20^\circ\text{C} = 7^\circ\text{C}$?

A "Fra le 6:30 e le 8:00 c'è un'ora e mezza" oppure

B "Fra le 6:30 e le 8:00 c'è l'una e mezza" ?

Chiaramente A, perché fra 27°C e 20°C ci sono "sette gradi della scala" = 7 K e non "la temperatura situata sette gradi sopra lo zero Celsius" = 7°C .

$$s_x^2 = \frac{(10.4^\circ\text{C} - 9.64^\circ\text{C})^2 + (8.5^\circ\text{C} - 9.64^\circ\text{C})^2 + \dots}{9}$$

$$= \frac{(0.76 \text{ K})^2 + (-1.14 \text{ K})^2 + \dots}{9} = 0.8204 \text{ K}^2$$

$$\underline{s_x = \sqrt{s_x^2} = 0.91 \text{ K}}$$

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	9.8 °C	10.2 °C	9.7 °C
$x_{(10)}$	$x_{(5)}$	$x_{(2)}$	$x_{(8)}$	$x_{(6)}$

Quindi

$$\underline{\tilde{x} = \frac{1}{2} (x_{(5)} + x_{(6)}) = 9.45 \text{ °C}}$$

$$Q_1 = q_{0.25} = x_{(3)} + 0.25 (x_{(4)} - x_{(3)}) = 9.025 \text{ °C}$$

$$Q_3 = q_{0.75} = x_{(7)} + 0.75 (x_{(8)} - x_{(7)}) = 10.15 \text{ °C}$$

$$\underline{IQR = Q_3 - Q_1 = 1.13 \text{ K}}$$

Istogrammi

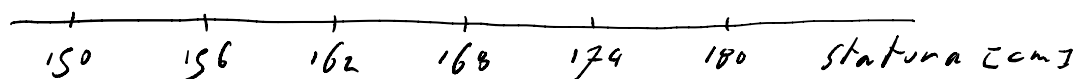
Gli istogrammi sono una maniera conveniente per rappresentare la distribuzione di variabili numeriche. Vediamo un esempio: vogliamo rappresentare la distribuzione della variabile *statura* che assume i valori seguenti, in *cm*, su un campione di 36 individui

15		0468899	← Equivale alla lista: 150, 154, 156, 158, 158, 159, 159, 160, ...
16		000013334455555778889	
17		00115588	
			
		Unità	
		Centinaia e decine	

Tutti i valori ricadono nell'intervallo 150-180. Cominciamo col dividere questo intervallo in un insieme ragionevole di intervalli più piccoli, per esempio.

150-156 156-162 162-168 168-174 174-180

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



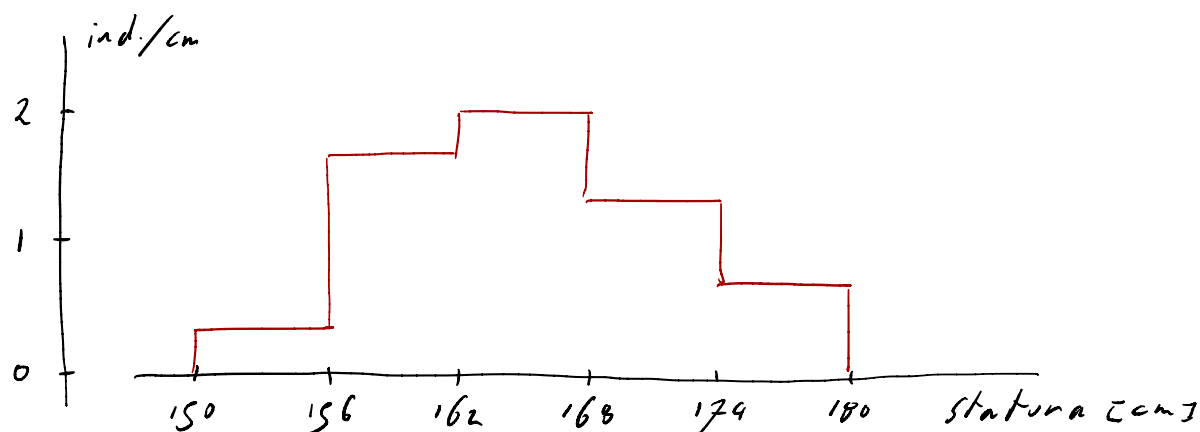
Ora, su ciascun intervallo, soleviamo una colonna la cui **area** rappresenta il numero di individui in quello intervallo.

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

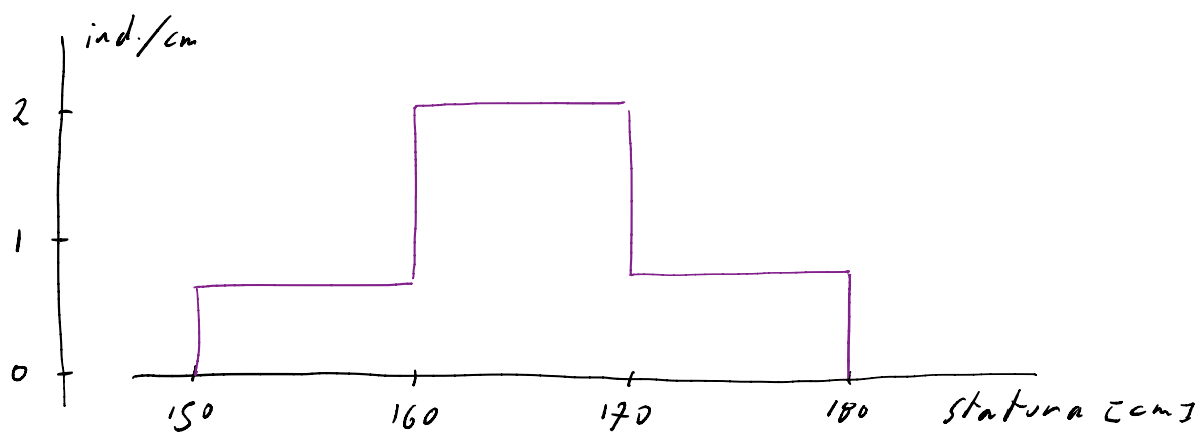
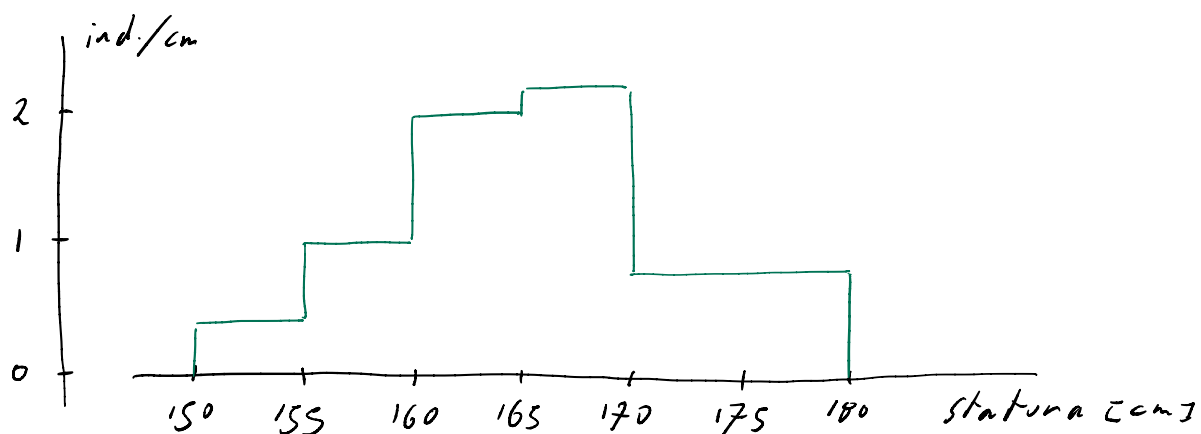
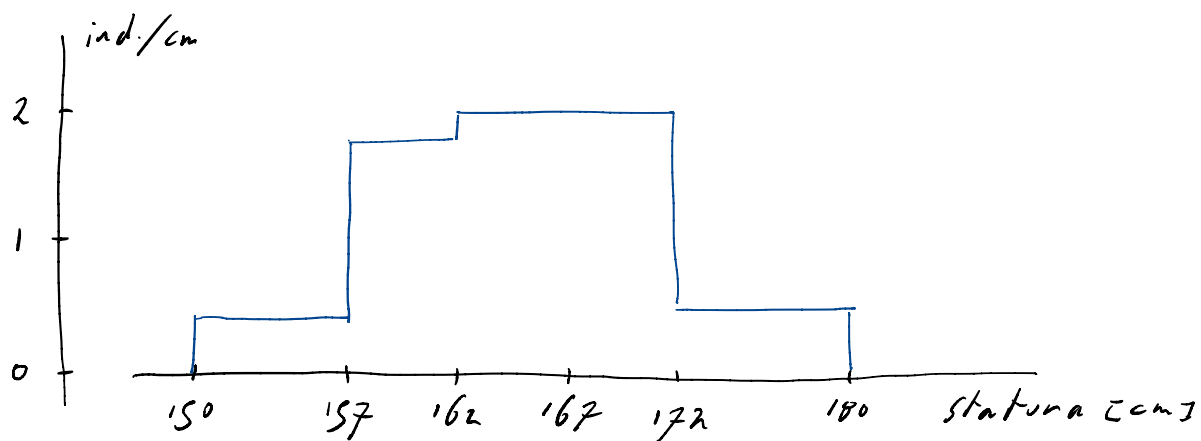
Siccome i lati orizzontali della colonna sono misurati in cm, le altezze devono essere misurate in individui/cm. Queste quantità si chiamano **densità di frequenza**.

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

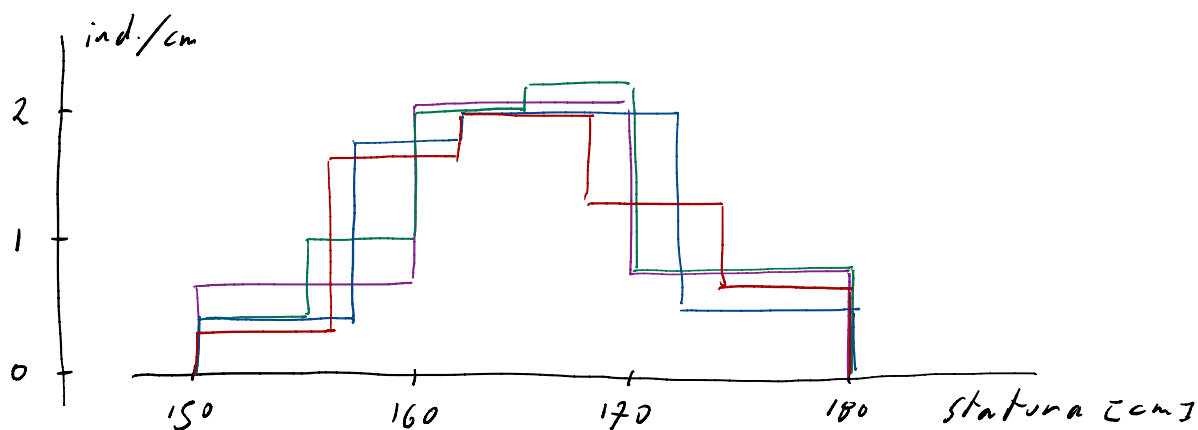
L'istogramma risultante è come segue.



La scelta degli intervalli è arbitraria: ecco altri esempi



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



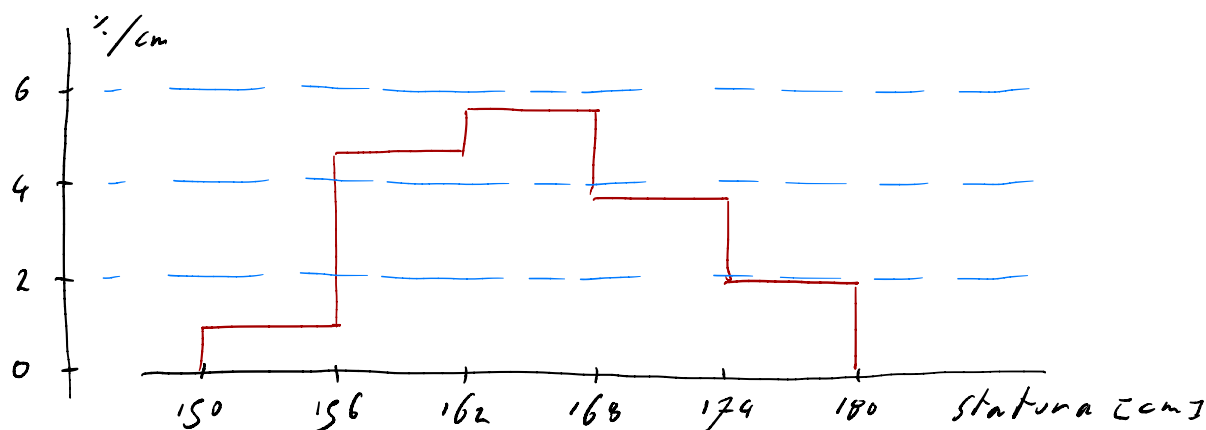
Se avessimo rappresentato il numero di individui tramite l'altezza anziché l'area delle colonne, i diversi grafici non sarebbero stati in accordo fra loro.

La densità di frequenza è stata definita in maniera tale che l'area di ciascuna colonna sia il numero di individui nell'intervallo 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 $\frac{\text{individui}}{\text{unità delle ascisse}}$ quella della densità di probabilità è $\frac{1}{\text{unità delle ascisse}}$. La densità di probabilità si può altresì esprimere

in percentuale, in questo caso l'unità sarà $\%$ delle ascisse e l'area dell'istogramma 100% .

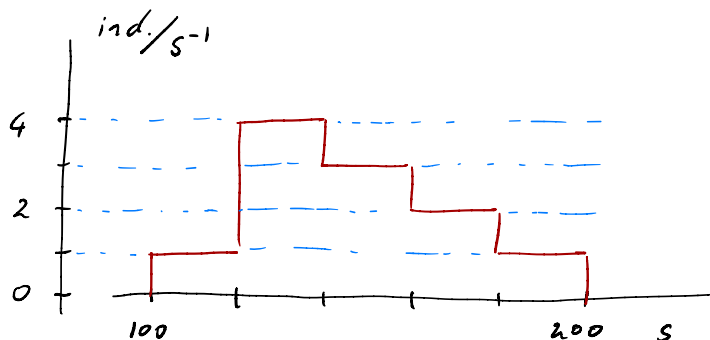
Nel nostro esempio abbiamo

150-156	156-162	162-168	168-174	174-180
2 ind.	10 ind.	12 ind.	8 ind.	4 ind.
$0.33 \frac{\text{ind.}}{\text{cm}}$	$1.67 \frac{\text{ind.}}{\text{cm}}$	$2.00 \frac{\text{ind.}}{\text{cm}}$	$1.33 \frac{\text{ind.}}{\text{cm}}$	$0.67 \frac{\text{ind.}}{\text{cm}}$
$0.3 \frac{\%}{\text{cm}}$	$4.6 \frac{\%}{\text{cm}}$	$5.6 \frac{\%}{\text{cm}}$	$3.7 \frac{\%}{\text{cm}}$	$1.9 \frac{\%}{\text{cm}}$

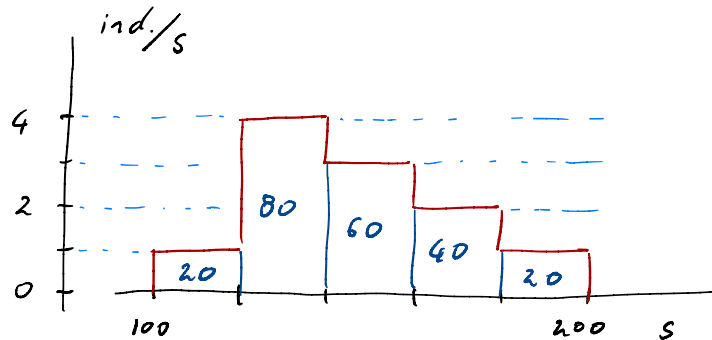


Nota: la forma dell'istogramma non cambia: è solo la scala della ordinate a cambiare.

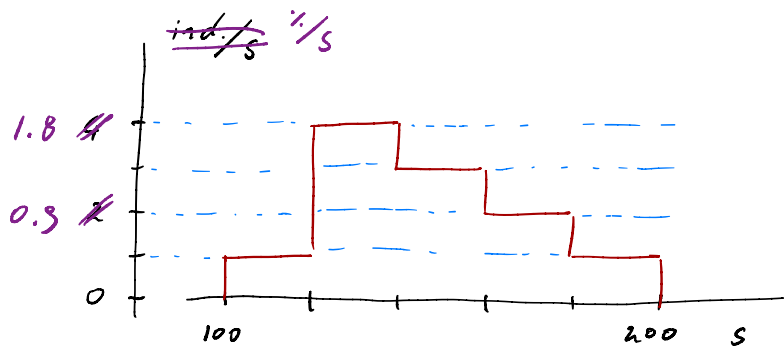
Eg. Converti la scala della ordinata da densità di frequenza a densità di probabilità nel seguente istogramma



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



Abbiamo quindi $n = 220$. Potremmo ora calcolare che $2 \frac{\text{ind.}}{\text{s}}$ equivalgono a $\frac{2}{220} \cdot 100\% \cdot \text{s}^{-1} = 0.9\% \cdot \text{s}^{-1}$ e similmente $4 \frac{\text{ind.}}{\text{s}}$ sono $1.8\% \cdot \text{s}^{-1}$. Ottenendo il grafico seguente



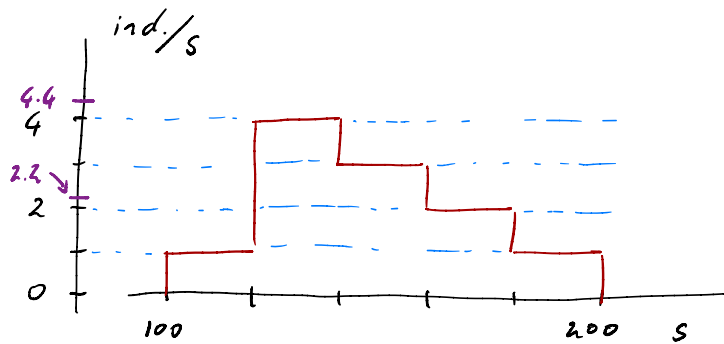
Questo risultato è però insoddisfacente, perché sarebbe meglio avere le cifre tonde $1\% \cdot \text{s}^{-1}$ e $2\% \cdot \text{s}^{-1}$ sull'asse delle ordinate. Calcoliamo dunque a quali densità di probabilità corrispondono queste densità di frequenza.

$$1\% \cdot \text{s}^{-1} \text{ corrisponde a } \frac{1\%}{100\%} \cdot 220 \frac{\text{ind.}}{\text{s}} = 2.2 \frac{\text{ind.}}{\text{s}}$$

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

Usiamo ora questa equivalenza per trovare la posizione

di $1 \times s^{-1}$ e $2 \times s^{-1}$ sull'asse delle ordinate.



Per cui questo è l'istogramma richiesto:

