Slide Set 4
Posizione e Trasformazioni dei Dati

Pietro Coretto
pcoretto@unisa.it

Corso di
Analisi e Visualizzazione dei Dati (Parte I)

Corso di Laurea in "Statistica per i Big Data" (L-41)
Università degli Studi di Salerno

Versione: 22 febbraio 2022 (h10:19)

Notes

Misure di posizione

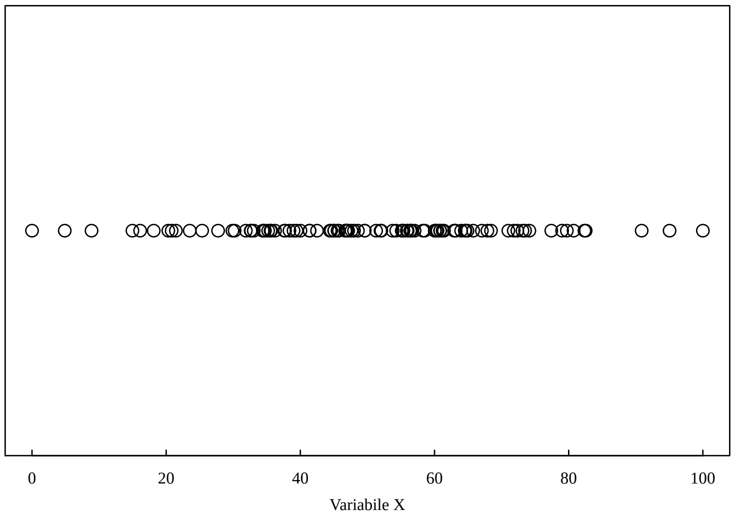
Restringiamo il campo alle variabili quantitative/numeriche

Posizione (o locazione)
Le misure di posizione (o locazione) hanno lo scopo di individuare regioni specifiche del range dei dati osservati

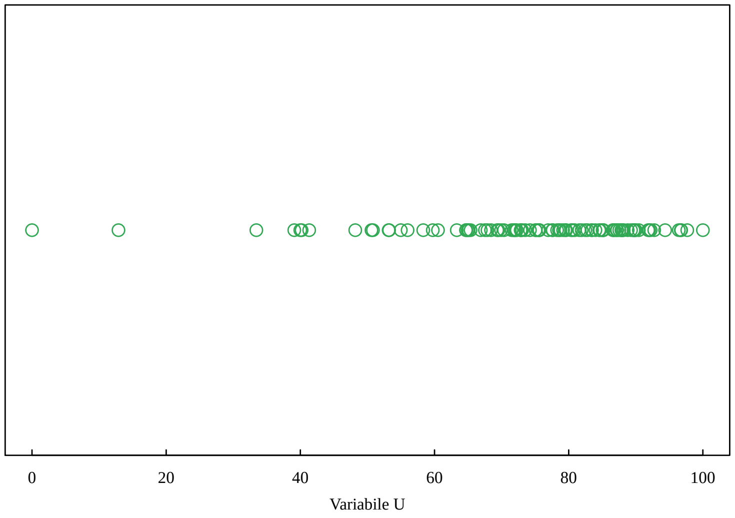
Centralità/tendenza centrale
è una misura di posizione che ha lo scopo di fornire una descrizione numerica della localizzazione del centro di massa dei dati

⚠ Non sempre questo coincide con la nozione di *centro geometrico*. Infatti, ritorniamo all'esempio della lezione precedente . . .

Notes



Notes



Notes

Weighting e *distribuzione empirica*

Il ruolo di $x = 40$ nelle distribuzioni mostrate nei due stripcharts precedenti è la stessa?

Supponiamo di poter associare un peso w_i ad ogni x_i. Ovvero una quantità che esprime l' importanza/influenza relativa di x_i nel campione osservato. Un sistema di pesi deve avere le seguenti caratteristiche

- non-negativi: $w_i \geq 0$
- non tutti nulli: $w_i \neq 0$ per almeno un indice $i = 1, 2, \ldots, n$

Dati i pesi $\{w_1, w_2, \ldots, w_n\}$ otteniamo i pesi normalizzati $\{\pi_1, \pi_2, \ldots, \pi_n\}$

$$\pi_i = \frac{w_i}{\sum_{i=1}^{n} w_i} = \frac{\textit{Peso di } x_i}{\textit{Peso Totale dei Dati}}$$

dove adesso $\pi_i \in [0, 1]$

Notes

Interpretiamo π_i come la misura dell'importanza/influenza attribuita ad x_i nella distribuzione dei dati osservata

Esempio: nel paese ZYX si rilevano i seguenti dati regionali:

Regione	A	B	C
Tasso di disoccupazione (%)	21.8	6.1	8.3
Pololazione ($10^6 \times$abitanti)	5.9	0.126	10

Domanda: quale sarebbe un numero rappresentativo per il tasso di disoccupazione dell'intero paese?

Notes

Dal punto di vista dell'economia nazionale il 6.1% di B *pesa* quanto l'8.3% di C?

La popolazione totale è di 16.026 (milioni), consideriamo i pesi $w_A = 5.9$, $w_B = 0.126$, $w_C = 10$, e la normalizzazione

$$\pi_A = \frac{5.9}{16.026} = 0.368, \quad \pi_B = \frac{0.126}{16.026} = 0.008, \quad \pi_C = \frac{10}{16.026} = 0.624$$

Sia x_i il tasso di disoccupazione della regione i, calcoliamo una sintesi numerica del tasso nazionale come

$$\begin{aligned}\pi_A x_A + \pi_B x_B + \pi_C x_C &= 21.8 \times 0.368 + 6.1 \times 0.008 + 8.3 \times 0.624 \\ &= 13.25\end{aligned}$$

In pratica i pesi consentono di modificare il contributo relativo di ciascuno dei valori di x

Notes

Esempio: uno studente sostiene due prove scritte durante il corso, ed un esame finale. Nelle tre prove consegue i voti: $x_1 = 24, x_2 = 18, x_3 = 27$. Tuttavia, il livello di difficoltà delle tre prove non è costante. Il professore attribuisce alle tre prove i pesi $\pi_1 = \frac{2}{10}, \pi_2 = \frac{3}{10}, \pi_3 = \frac{5}{10}$. Il voto finale è calcolato come

$$\begin{aligned}\pi_1 x_1 + \pi_2 x_2 + \pi_3 x_3 &= 0.2 \times 24 + 0.3 \times 18 + 0.5 \times 27 \\ &= 23.7\end{aligned}$$

In tutti questi casi abbiamo sintetizzato un insieme di numeri con una loro somma pesata

$$\sum_{i=1}^{n} \pi_i x_i \; = \; \pi_1 \times x_1 + \pi_2 \times x_2 + \ldots \pi_n \times x_n$$

in modo da ottenere un valore di x_i rappresentativo dei livelli più *influenti/pesanti*.

Notes

Weighting

moltissimi strumenti statistici costruiti per caratterizzare un qualche aspetto della distribuzione osservata si basano su un sistema di weighting/pesaggio dei dati

Distribuzione

caratterizzare la distribuzione dei dati significa attribuire ad ogni osservazione il suo peso relativo

Definizione (distribuzione empirica dei dati)

Date le osservazioni $\{x_1, x_2, \ldots, x_n\}$, la distribuzione empirica dei dati assegna un peso *uniforme* $\frac{1}{n}$ ad ogni unità $i = 1, 2, \ldots, n$.

Osserva: i pesi $\pi_i = \frac{1}{n}$ sono non tutti, e $\sum_{i=1}^n \pi_i = \sum_{i=1}^n \frac{1}{n} = n\frac{1}{n} = 1$

Notes

Media

Definizione (media empirica)

Dati i livelli osservati $\{x_1, x_2, \ldots, x_n\}$ per una variabile X, si definisce media empirica

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

⚠ possiamo considerare $\bar{x}$ per una variabile non numerica?

La media (empirica) è una somma pesata: infatti, fissato $\pi_i = 1/n$ per ogni $i = 1, 2, \ldots, n$ otteniamo

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

Nota: $\bar{x}$ è anche chiamata "*media campionaria osservata*", in questo corso quando parliamo semplicemente di "*media*" ci riferiamo sempre a $\bar{x}$

Notes

Supponiamo di osservare $\{7, 3, 7, 7, 3\}$:

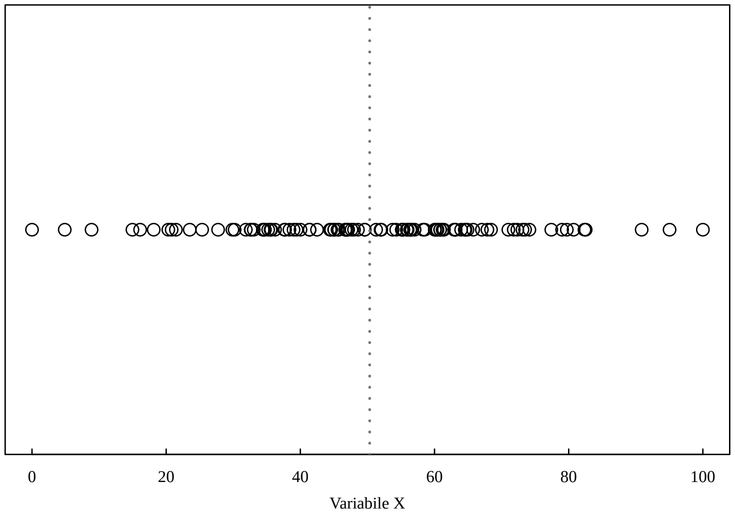
$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n} x_i = & \frac{1}{5}(7+3+7+7+3)\\ =& \frac{1}{5}(3{\times}2+7{\times}3)\\ =& 3{\times}\frac{2}{5}+7{\times}\frac{3}{5}\\ =& 5.4\end{aligned}$$

Dati $K \leq n$ livelli distinti con frequenze assolute $n_1, n_2, \ldots, n_K$, possiamo scrivere

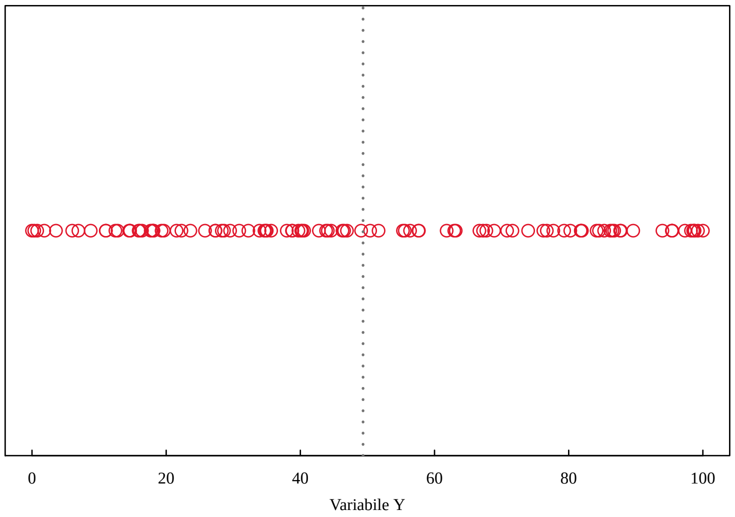
$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}\sum_{k=1}^{K} x_k n_k = \sum_{k=1}^{K} f_k x_k$$

dove per costruzione $\sum_{i=k}^{K} f_k = 1$ e $f_k \in [0, 1]$

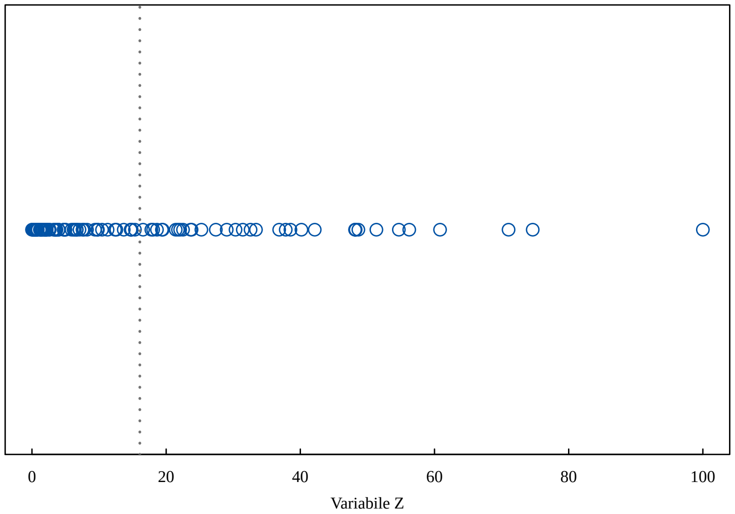
Notes



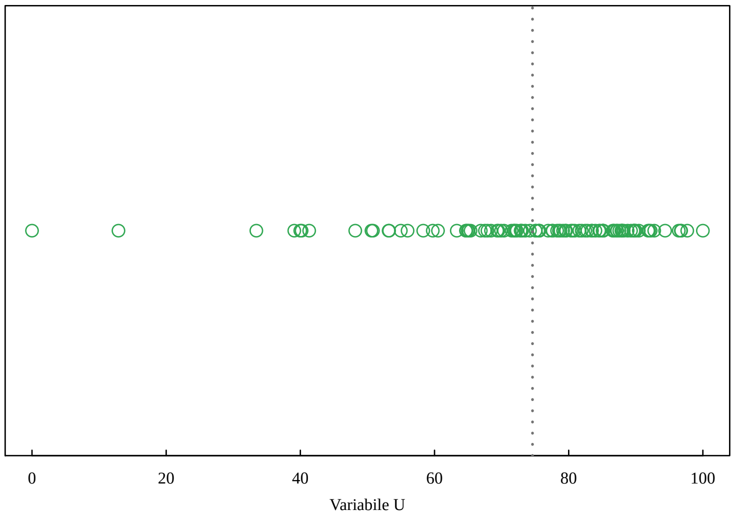
Notes



Notes



Notes



Notes

Quindi la media:

- esprime un valore sintetico rappresentativo della distribuzione di X: ciascun x_i contribuisce alla formazione di $\bar{x}$ con un peso/influenza pari alla sua frequenza
- è una misura di locazione/posizione: $\bar{x}$ tenderà ad assumere valori collocati in una regione del range contente la massa prevalente dei dati osservati
- è una misura di centralità: questo potrà essere chiaro solo dopo aver introdotto le deviazione quadratica totale

Notes

Mediana

Definizione (mediana)

Dati i livelli osservati $\{x_1, x_2, \ldots, x_n\}$ per una variabile X, si definisce mediana, indicata con $\mathrm{Med}(X)$, il quantile al livello $\alpha = 50\%$.

Poiché $\mathrm{Med}(X) = X_{\frac{1}{2}}$

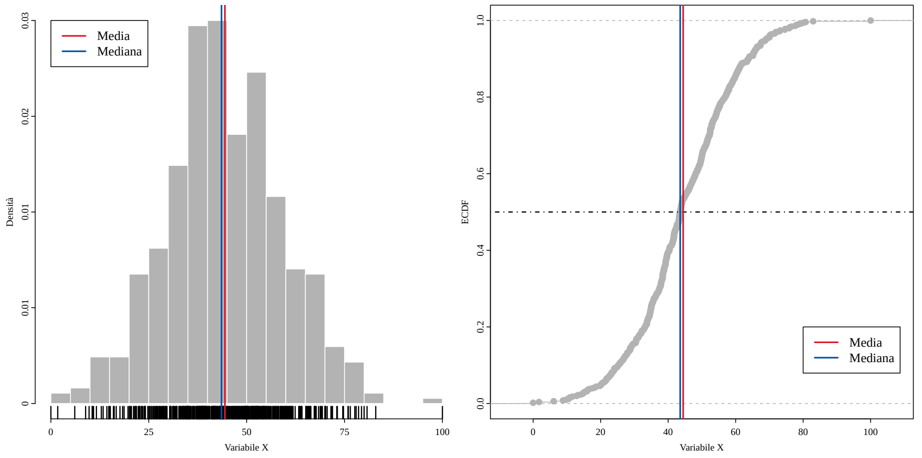
- l'intervallo $[x_{\min}, \mathrm{Med}(X)]$ contiene circa il 50% dei dati
- l'intervallo $(\mathrm{Med}(X), x_{\max}]$ contiene il restante 50% (circa) dei dati

La mediana è una misura di

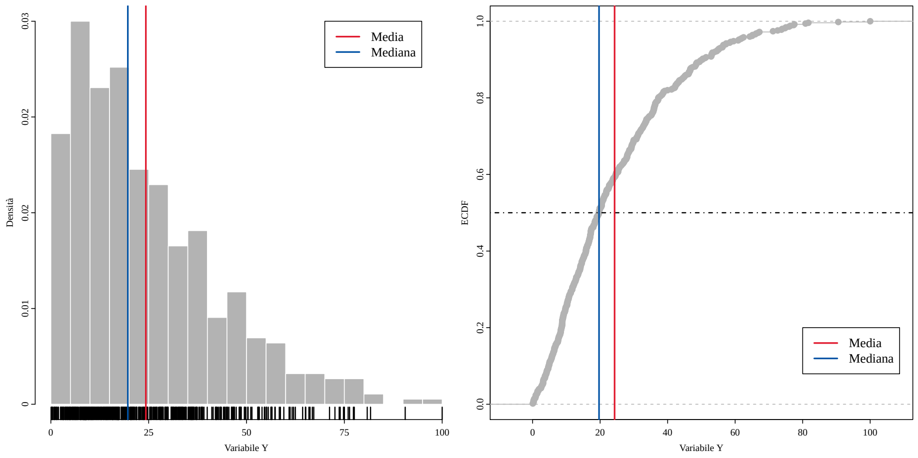
- posizione: individua un punto del range dei dati
- centralità: individua un punto che divide il range in parti contenenti una stessa massa di dati. Tuttavia, torneremo sulla centralità in seguito.

Per il calcolo della mediana useremo le approssimazioni numeriche dei quantili.

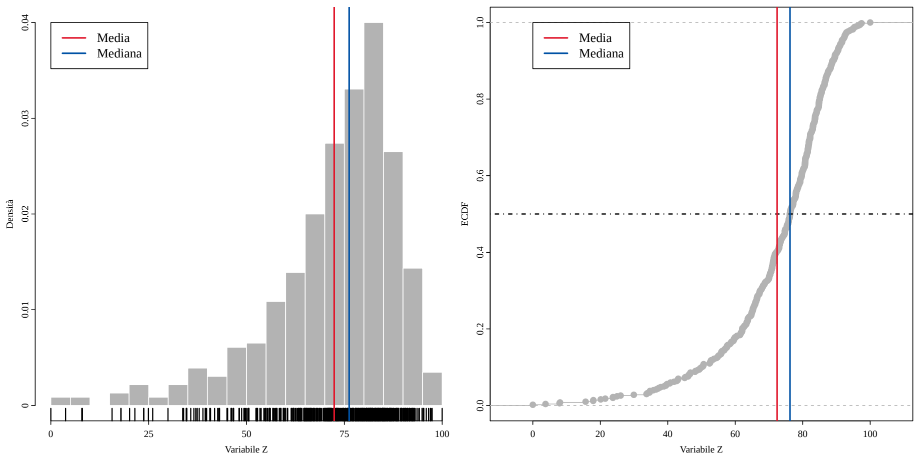
Notes



Notes


Media
Mediana
Densità
Variabile Y
ECDF
Variabile Y
Media
Mediana

Notes


Media
Mediana
Densità
Variabile Z
ECDF
Variabile Z
Media
Mediana

Notes

Moda

A differenza della media e della mediana la *Moda* può essere calcolata per qualsiasi tipo di variabile

Definizione (Moda)

Dati i livelli/labels osservati $\{x_1, x_2, \ldots, x_n\}$ per una variabile X, si definisce moda, si indicata con $\mathrm{Mod}(X)$,

(i) il livello/label osservato più frequente nei dati non continui

(ii) il livello di massima densità nei dati continui

La moda è una misura di

- posizione: individua un punto del range dei dati
- centralità: in un caso molto specifico.

Notes

Esempio: si rileva il $X =$ "*numero di pezzi prodotti (in centinaia)*" in un giorno da $n = 6$ macchinari, dati: $\{1, 2, 3, 3, 2, 2\}$. $\mathrm{Mod}(X) = 2$.

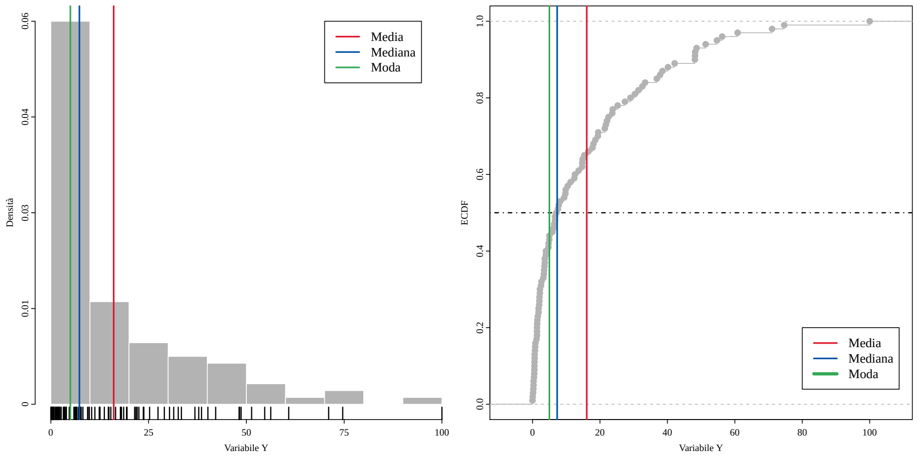
Tuttavia se avessimo osservato $\{3, 2, 3, 3, 2, 2\}$, la moda non avrebbe fornito un *unico valore*, infatti $\mathrm{Mod}(X) = \{2, 3\}$

Esempio: data set `bwght.csv`, X=`parity`="*numero di figli*"

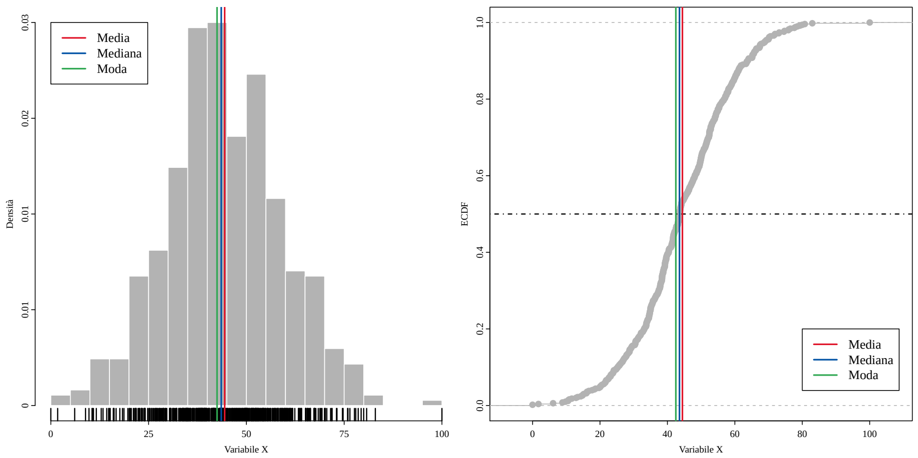
X	n_k
1	795
2	389
3	146
4	39
5	15
6	4

$\mathrm{Mod}(X) = 1$

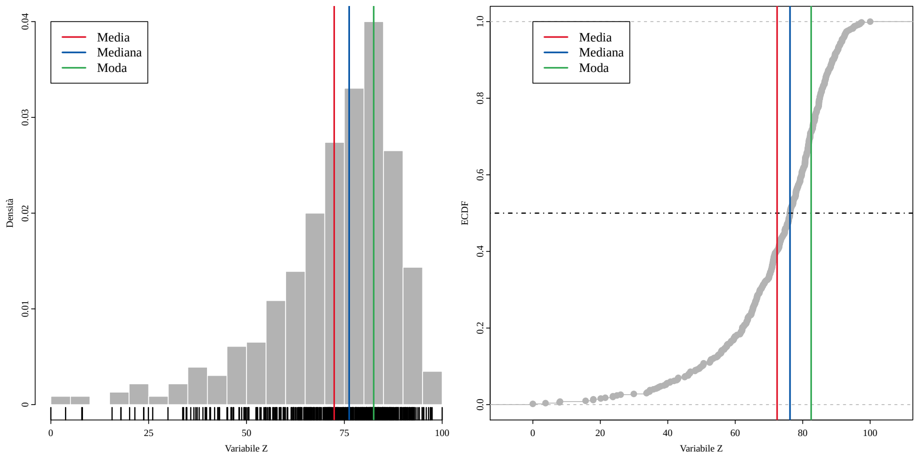
Notes


Media
Mediana
Moda
Densità
Variabile Y
ECDF
Variabile Y
Media
Mediana
Moda

Notes


Media
Mediana
Moda
Densità
Variabile X
ECDF
Variabile X
Media
Mediana
Moda

Notes



Esempi/Applicazioni $\longrightarrow$ `R` **script file**

Notes

Serie di quantili

Le misure di posizione precedenti tendono ad identificare un punto del range con alta densità/massa di dati

Anche i quantili possono essere visti come misure di posizione. Tuttavia, essi seguono un "*principio di localizzazione*" diverso

Per "*serie di quantili*" intendiamo i quantili calcolati su una griglia di valori $\{\alpha_1, \alpha_2, \ldots, \alpha_m\}$ equispaziata

Serie di quantili di uso comune:

- Quartili: $\alpha \in \{25\%, 50\%, 75\%\}$
- Decili: $\alpha \in \{10\%, 20\%, \ldots 90\%\}$
- Percentili: $\alpha \in \{1\%, 2\%, \ldots 99\%\}$

Osserva: $X_{0\%} = x_{\min}$ e $X_{100\%} = x_{\max}$

Notes

Quartili

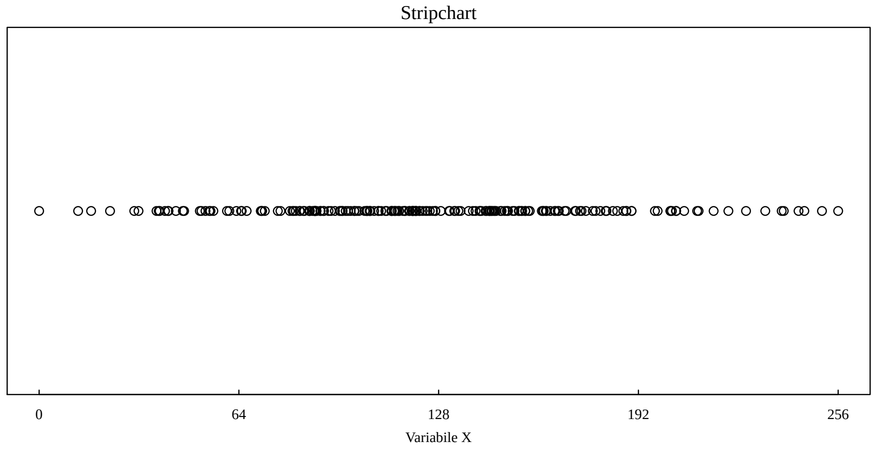
Dato X solitamente indichiamo convenzionalmente

- $Q_1 = X_{0.25}$,
- $Q_2 = X_{0.5} = \mathrm{Med}(X)$
- $Q_3 = X_{0.75}$

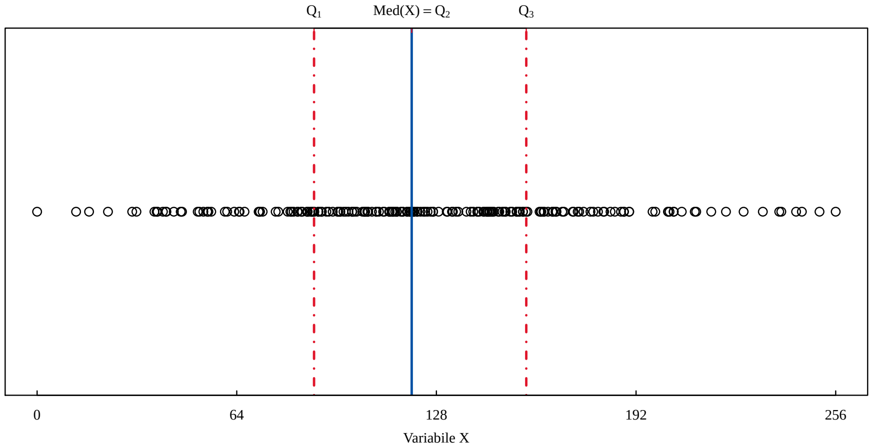
Identificano 3 posizioni sul range dei dati utili ad identificare 4 intervalli contenenti ciascuno il $25\% = \frac{1}{4}$ dei dati, infatti

- I quarto: $x_k \in [x_{\min}, Q_1]$,
- II quarto: $x_k \in (Q_1, Q_2] = (Q_1,\ \mathrm{Med}(X)]$
- III quarto: $x_k \in (Q_2, Q_3] = (\mathrm{Med}(X),\ Q_3]$
- IV quarto: $x_k \in (Q_3, x_{\max}]$

Notes

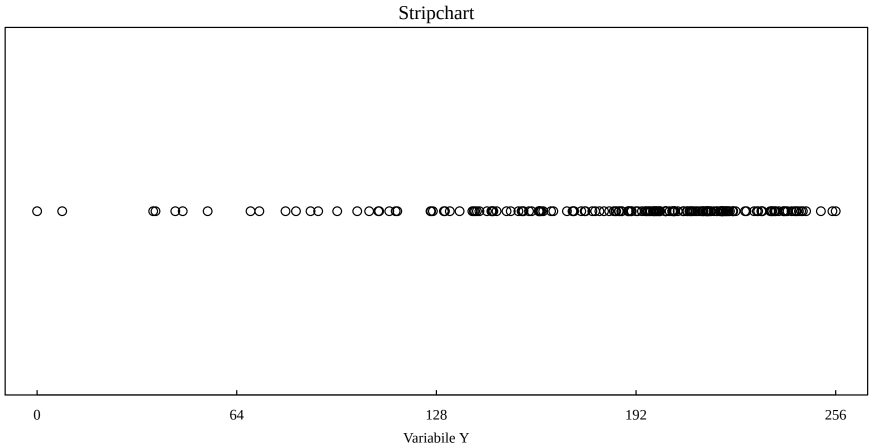


Notes



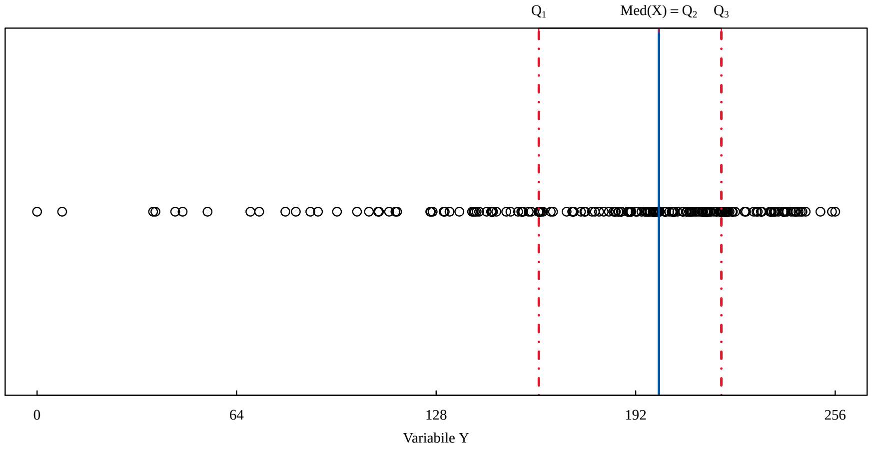
Q1
Med(X) = Q2
Q3
0
64
128
192
256
Variabile X

Notes



Stripchart
0
64
128
192
256
Variabile Y

Notes



Notes

Tukey's 5-number summary

J.W. Tukey (1977): range e quartili sono tutto quello che ci serve per iniziare a capire cosa c'è in una distribuzione univariata

Esempio: il ROE = "*redditività del capitale proprio*". Dal data set balancesheet.RData, consideriamo la variabile X =dat$roe. Nota: X è misurato in scala percentuale

Tukey's 5-number

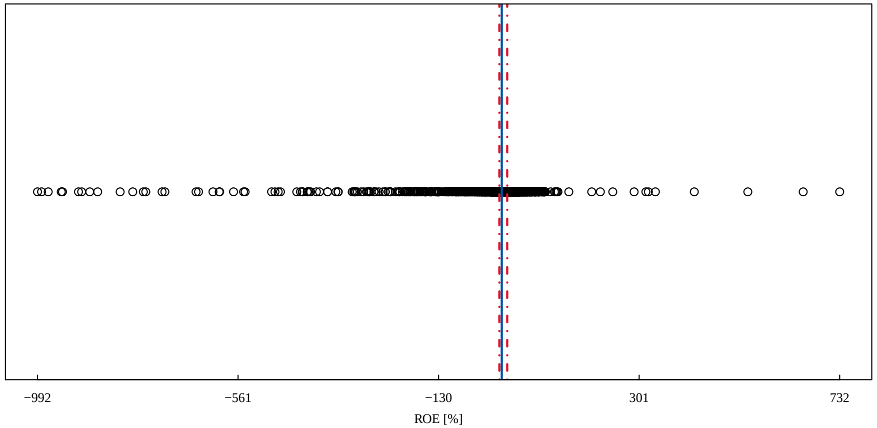
Minimo $x_{\min}$	**Lower-hinge** $Q_1 = X_{25\%}$	**Mediana** $Q_2 = \mathrm{Med}(X)$	**Upper-hinge** $Q_3 = X_{75\%}$	**Massimo** $x_{\max}$
-992.28	0.17	5.39	17.11	731.60

Data bulk: è la massa centrale dei dati, ovvero il 50% dei dati centrali contenuti in $[Q_1,\ Q_3]$. Nota: *Bulk=maggior parte*, *Hinge=cardine, perno, cerniera*

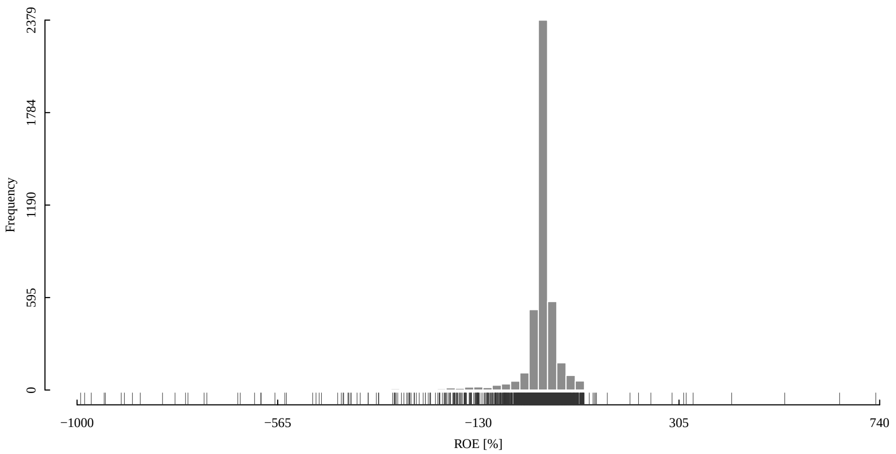
Code: è la parte dei dati esterna al Bulk e che si estende verso $x_{\min}$ e $x_{\max}$

Notes

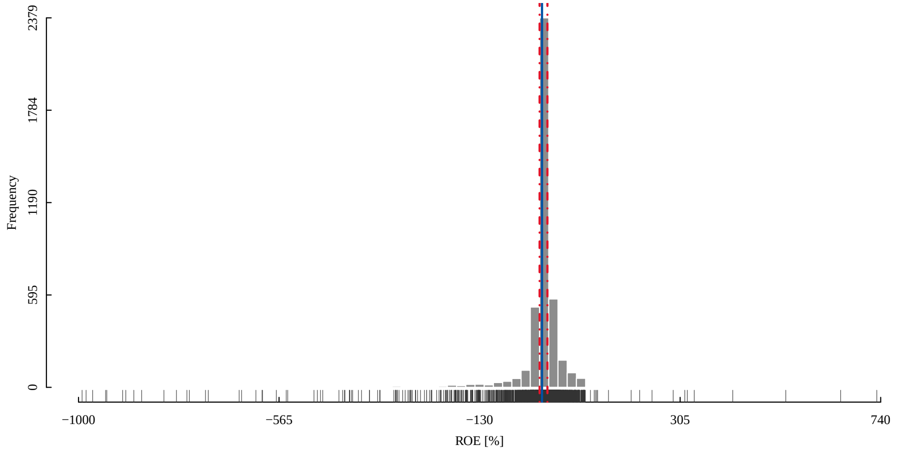
Osserva: in alcuni casi per vedere meglio dentro il *bulk* dei dati abbiamo bisogno di qualche altra rappresentazione grafica (. . . ne parleremo presto!)



Notes



Notes



Notes

Decili

I decili (8 numeri + max e min) sono meno usati dei percentili. Tuttavia, a differenza dei percentili sono facili da *tabellare*

Esempio: decili di X =dat$roe nel data set balancesheet.RData

0%	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
-992.3	-18.0	-0.1	0.7	2.7	5.4	8.9	13.6	21.4	36.1	731.6

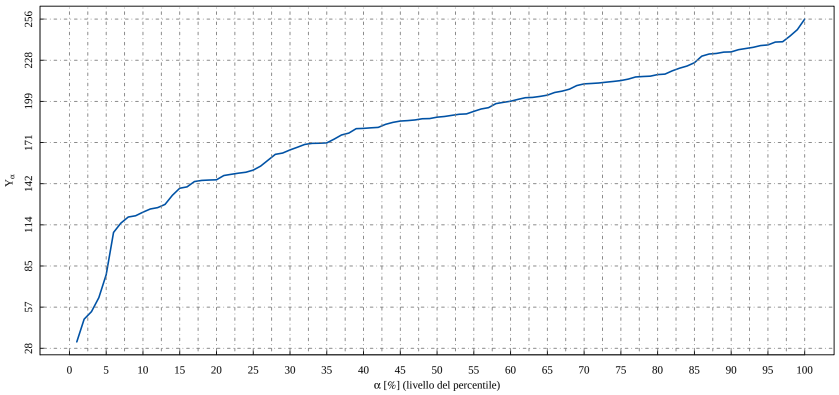
Utilizzo

- lettura più profonda del 5-number summary, soprattutto nelle code della distribuzione
- sulla coda sx guardiamo allo *spread* sul 10% dei dati più piccoli $(X_{10\%} - x_{\min}) = (-18 - (-992.3)) = 974.3$
- sulla coda dx guardiamo allo *spread* sul 10% dei dati più grandi $(x_{\max} - X_{90\%}) = (731.6 - 36.1) = 695.5$

Notes

Percentili

I percentili sono difficili da tabellare. Solitamente li rappresentiamo graficamente. Supponiamo di aver osservato una qualche variabile Y, e di aver calcolato Y_α per $\alpha = 1\%, 2\%, \ldots$



Notes

Esempio: la "*World Health Organization*" (WHO) ed il "*Center For Disease Control*" (CDC) misurano diversi parametri della crescita

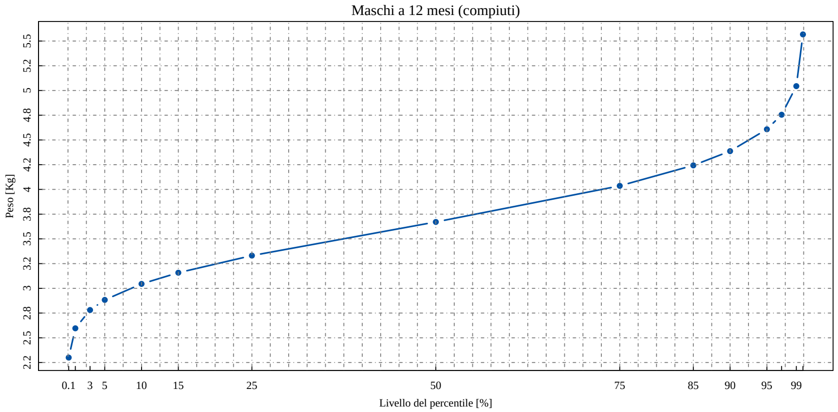
- altezza, peso, bmi, circonferenza del cranio, etc.
- per età≤36mesi, età≤2anni, età $\in [5, 19]$ anni
- separatamente per maschie e femmine

Da questo complesso *survey* si ottengono le "*famigerate*" *curve di crescita* (per alcune mamme moderne sono una vera ossessione!).

I seguenti percentili sono attenuti dall'ultimo data set disponibile su https://www.who.int

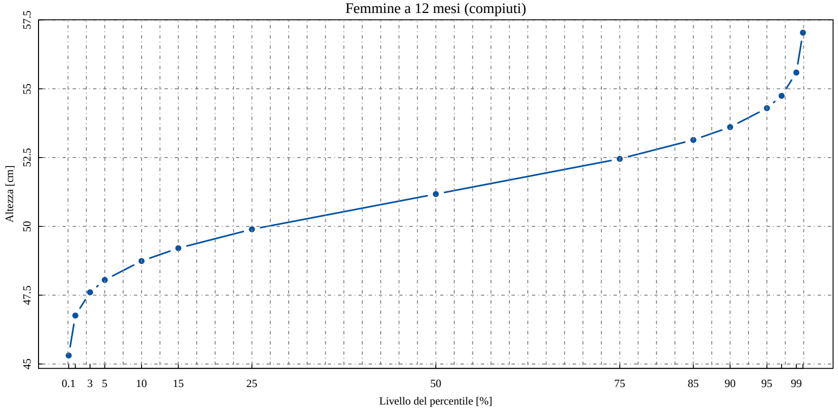
Notes

Esempio: percentili per il peso dei maschi ad 1 anno dalla nascita (fonte: WHO, 2019)



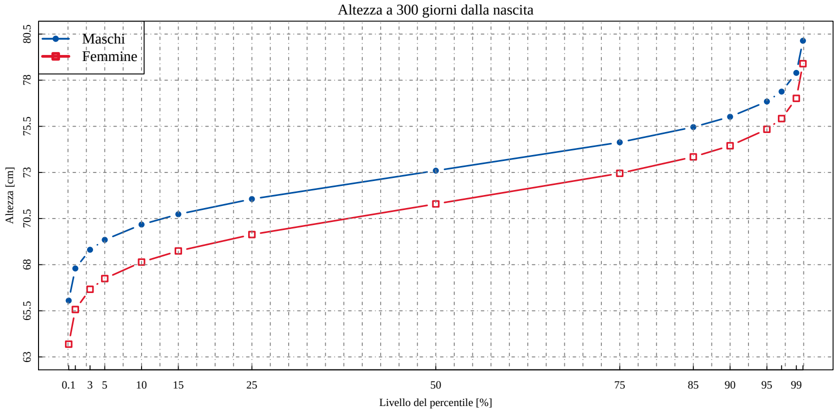
Notes

Esempio: percentili di crescita per l'altezza delle femmine ad 1 anno dalla nascita (fonte: WHO, 2019)



Notes

Esempio: confronto per i percentili dell'altezza dei maschi vs femmine, età = 300 giorni, in pratica siamo verso l'inizio del decimo mese (fonte: WHO, 2019)



Notes

Quelle che il pediatra chiama "*curve di crescita*" (*growth charts*) non sono altro che una *brillante* rappresentazione grafica dove

- fissato il sesso
- si considerano insieme peso e altezza (oppure: altezza e circonferenza del cranio)
- si fa vedere l'evoluzione di un certo percentile nel tempo (età)

Esempi: CDC growth charts

- Curve di crescita peso/altezza: maschi [0, 36] mesi https://www.cdc.gov/growthcharts/data/set1clinical/cj41c017.pdf
- Curve di crescita peso/altezza: femmine [0, 36] mesi https://www.cdc.gov/growthcharts/data/set1clinical/cj41c018.pdf

Esempi/Applicazioni ⟶ `R` script file

Notes

Trasformazioni di variabili

Talvolta invece di studiare la distribuzione di X, studiamo la distribuzione di $Y = g(X)$, dove g è una trasformazione della variabile originaria.

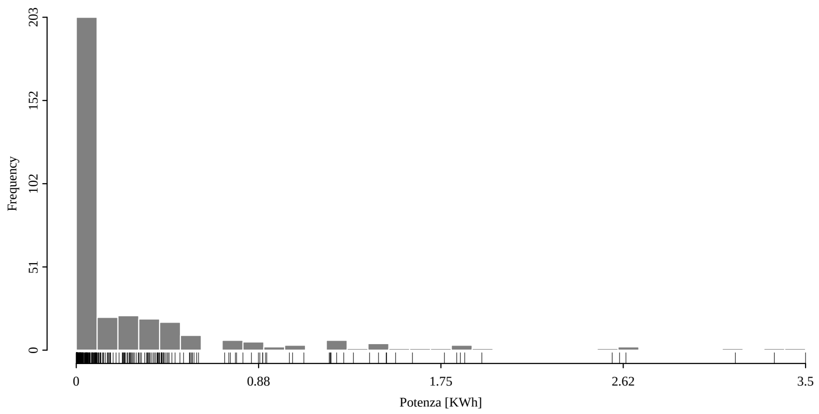
In questi casi non lavoriamo sulle osservazione $\{x_1, x_2, \ldots, x_n\}$ ma sulla loro trasformazione $\{g(x_1), g(x_2), \ldots, g(x_n)\}$

La trasformazione dei dati potrebbe essere necessaria per due motivi principali:

- rendere la descrizione dei dati osservati funzionale al tipo di analisi
- per motivi metodologici

Notes

Esempio: tra le 10:30 e le 11:30 si misura l'assorbimento di potenza elettrica su $n = 329$ contatori scelti a caso tra le utenze domestiche di una grande città. Ogni contatore ha una capacità di picco effettivo non superiore a 3.5KWh



Notes

Dal punto di vista del dimensionamento della rete elettrica, ci interessa la distribuzione rispetto al picco. Solitamente la potenza viene espressa in Decibels: se X è la potenza

$$\text{dB} = 10 \log_{10}\left(\frac{X}{\text{potenza di picco}}\right)$$

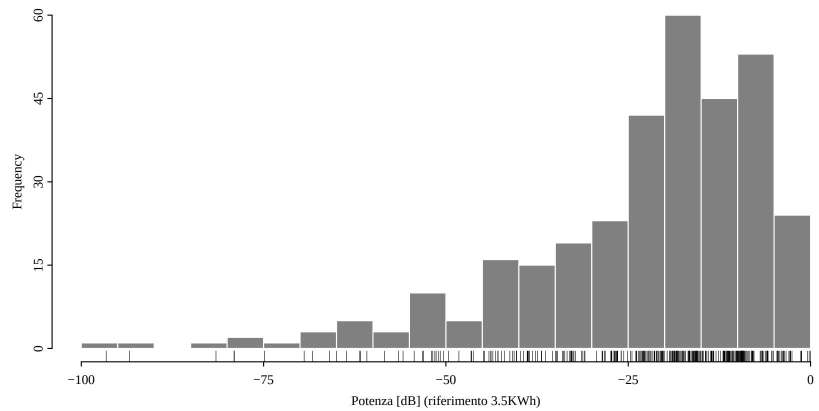
Ad esempio

- -30dB $\Longrightarrow$ X è 1000 volte inferiore al picco
- -20dB $\Longrightarrow$ X è 100 volte inferiore al picco,
- -10dB $\Longrightarrow$ X è 10 volte inferiore al picco,
- -3dB $\Longrightarrow$ X è la metà del picco

Quindi se $\{x_1, x_2, \ldots, x_n\}$ sono i dati osservati sulla potenza, li trasformiamo in

$$y_i \text{ [dB]} = 10 \log_{10}\left(\frac{x_i \text{ [KWh]}}{3.5}\right)$$

Notes



Notes

Traformazioni lineari

Una classe di trasformazioni sui dati di grande rilevanza sono quelle lineari

Fissate due costanti $a, b \in \mathbb{R}$ la trasformazione lineare della variabile X è una nuova variabile

$$Y = a + bX.$$

Quindi i dati osservati originari $\{x_1, x_2, \ldots, x_n\}$ si trasformano in

$$\{y_1, y_2, \ldots, y_n\} = \{a + bx_1,\ a + bx_2, \ldots, a + bx_n\}$$

Perché ci interessano così tanto le trasformazioni lineari?

Notes

Esempio 1: scaling

Gli esami si valutano con voti da 18 a 30, mentre il voto di laurea si valuta con voti da 66 a 110. Nota: $18/3 = 6$, e $66/11 = 6$, mentre 30/3=10 e 110/11=10. Supponiamo di voler uniformare tutti i voti in un range [0,100]

Sia E il voto di un esame in trentesimi, il nuovo voto Y è ottenuto come

$$Y = \frac{E}{3} \times 10 = \frac{10}{3}E = a + bE \qquad \text{per } a = 0, \text{ e } b = 10/3.$$

Sia L il voto di laurea, il nuovo voto Y è ottenuto come

$$Y = \frac{L}{11} \times 10 = \frac{10}{11}L = a + bL \qquad \text{per } a = 0, \text{ e } b = 11/10$$

Osserva: una trasformazione lineare con $a = 0$ e $b \neq 0$ corrisponde ad un cambiamento di scala di misurazione

Notes

Esempio 2: cambiamento di unità di misura

Sia X la velocità in Km/h. Supponiamo di volerla trasformare in Y espressa in m/s

$$Y\left[\frac{\text{m}}{\text{s}}\right] = X\left[\frac{\text{Km}}{\text{h}}\right] \implies Y\left[\frac{\text{m}}{\text{s}}\right] = X\left[\frac{1000\text{m}}{60^2\text{s}}\right]$$

dividendo tutto per m e moltiplicando per s, otteniamo la trasformazione

$$Y = 0.278X = a + bX \qquad \text{per } a = 0, \text{ e } b = 0.278,$$

essa trasforma $X\left[\frac{\text{Km}}{\text{h}}\right] \longrightarrow Y\left[\frac{\text{m}}{\text{s}}\right]$

Osserva: una trasformazione lineare con $a = 0$ e $b \neq 0$ ha prodotto un cambiamento di scala che coincide con una cambiamento di unità di misura

Notes

Esempio 3: cambiamento di unità di misura

Sia X la temperatura misura in °C (Celsius). Supponiamo di volerla trasformare in Y espressa in °F (Fahrenheit). Si noti che

$$Y\ [°\text{F}] = 32 + \frac{9}{5}X\ [°\text{C}]$$

quindi la trasformazione lineare $Y = a + bX$ con $a = 32$ e $b = 9/5$ trasforma $X\ [°\text{C}] \longrightarrow Y\ [°\text{F}]$.

Notes

Alcuni effetti delle trasformazioni lineari

Data la trasformazione lineare $Y = a + bX$, questa produce due effetti: *shift* e *scaling*.

Shift: quando $a \neq 0$
Tutti i punti sono spostati verso sinistra $(a < 0)$ o verso destra $(a > 0)$ di uno stesso ammontare

Scaling: quando $b \neq 0$

- $b > 0 \implies$ l'ampiezza del range viene modificata
- $b < 0 \implies$ l'ampiezza del range viene modificata, inoltre l'ordinamento delle unità statistiche si inverte in modo speculare

In particolare:

- se $|b| < 1 \implies$ l'ampiezza del range si comprime
- se $|b| > 1 \implies$ l'ampiezza del range si espande

Notes

Proprietà (Impatto delle trasformazioni sul range)

Siano $\{x_1, x_2, \ldots, x_n\}$ le osservazioni campionarie per la variabile X. Sia $[x_{\min}, x_{\max}]$ il range di valori osservati la cui ampiezza è $(x_{\max} - x_{\min})$. Data la trasformazione lineare $Y = a + bX$, con $b \neq 0$, allora

$$\text{ampiezza del range di Y} = |b|\,(x_{\max} - x_{\min})$$

Dimostrazione. Prima di tutto notiamo che quando $b > 0$ allora

$$x_{\min} \leq x_{\max} \implies bx_{\min} \leq bx_{\max} \implies \underbrace{a + bx_{\min}}_{y_{\min}} \leq \underbrace{a + bx_{\max}}_{y_{\max}}$$

Notes

Tuttavia per $b < 0$

$$x_{\min} \leq x_{\max} \implies bx_{\min} \geq bx_{\max} \implies \underbrace{a + bx_{\min}}_{y_{\max}} \geq \underbrace{a + bx_{\max}}_{y_{\min}}$$

Quindi possiamo scrivere

$$(y_{\max} - y_{\min}) := \begin{cases} b(x_{\max} - x_{\min}) & \text{se } b > 0 \\ -b(x_{\max} - x_{\min}) & \text{se } b < 0 \end{cases}$$

che in modo compatto scriviamo

$$(y_{\max} - y_{\min}) = |b| \, (x_{\max} - x_{\min})$$

■

Notes

Proprietà (Impatto delle trasformazioni lineari sulla media)

Siano $\{x_1, x_2, \ldots, x_n\}$ le osservazioni campionarie per la variabile X. Sia $\bar{x}$ la media campionaria di X. Data la trasformazione lineare $Y = a + bX$, allora

$$\bar{y} = a + b\bar{x}$$

Dimostrazione.

$$\begin{aligned} \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i &= \frac{1}{n}\sum_{i=1}^{n} a + bx_i \\ &= \frac{1}{n}\sum_{i=1}^{n} a + \frac{1}{n}\sum_{i=1}^{n} bx_i \\ &= \frac{1}{n} na + b\frac{1}{n}\sum_{i=1}^{n} x_i \\ &= a + b\bar{x} \end{aligned}$$

■

Notes

Proprietà (Impatto delle trasformazioni lineari sui quantili)

Siano $\{x_1, x_2, \ldots, x_n\}$ le osservazioni campionarie per la variabile X. Sia X_α il quantile di livello α di X. Data la trasformazione lineare $Y = a + bX$, allora

$$Y_\alpha = \begin{cases} a + bX_\alpha & \text{se } b \geq 0 \\ a + bX_{(1-\alpha)} & \text{se } b < 0 \end{cases}$$

Dimostrazione. Assumiamo $b \geq 0$. Esisterà qualche $j \leq n$ per cui

$$\underbrace{x_{\min}, \ldots, x_{(j)}}_{\alpha\% \text{ osservazioni}} \leq X_\alpha \leq \underbrace{x_{(j+1)}, \ldots, x_{\max}}_{(1-\alpha)\% \text{ osservazioni}}$$

moltiplichiamo tutto per b ed aggiungiamo a, otteniamo oncora

$$\underbrace{a + bx_{\min}, \ldots, a + bx_{(j)}}_{\alpha\% \text{ osservazioni}} \leq a + bX_\alpha \leq \underbrace{a + bx_{(j+1)}, \ldots, a + bx_{\max}}_{(1-\alpha)\% \text{ osservazioni}}$$

Notes

quindi

$$\underbrace{y_{\min}, \ldots, y_{(j)}}_{\alpha\% \text{ osservazioni}} \leq a + bX_\alpha \leq \underbrace{y_{(j+1)}, \ldots, y_{\max}}_{(1-\alpha)\% \text{ osservazioni}}$$

Poichè $a + bX_\alpha$ lascia alla sua sinistra ancora $\alpha\%$ dati, allora allora questo non può che coincidere con Y_α. Quindi deve essere $Y_\alpha = a + bX_\alpha$

Ora assumiamo il caso $b < 0$. Esisterà qualche $j \leq n$ per cui

$$\underbrace{x_{\min}, \ldots, x_{(j)}}_{\alpha\% \text{ osservazioni}} \leq X_\alpha \leq \underbrace{x_{(j+1)}, \ldots, x_{\max}}_{(1-\alpha)\% \text{ osservazioni}}$$

moltiplicando tutto per $b < 0$ le disuguaglianze si invertono

$$\underbrace{bx_{\min}, \ldots, bx_{(j)}}_{\alpha\% \text{ osservazioni}} \geq bX_\alpha \geq \underbrace{bx_{(j+1)}, \ldots, bx_{\max}}_{(1-\alpha)\% \text{ osservazioni}}$$

Notes

e sommando a

$$\underbrace{a+bx_{\min},\ \ldots,\ a+bx_{(j)}}_{\alpha\%\text{ osservazioni}} \geq\ a+bX_\alpha\ \geq \underbrace{a+bx_{(j+1)},\ \ldots,\ a+bx_{\max}}_{(1-\alpha)\%\text{ osservazioni}}$$

Da cui

$$\underbrace{y_{\max},\ \ldots,\ y_{(j)}}_{\alpha\%\text{ osservazioni}} \geq\ a+bX_\alpha\ \geq \underbrace{y_{(j+1)},\ \ldots,\ y_{\min}}_{(1-\alpha)\%\text{ osservazioni}}$$

Poichè $a+bX_\alpha$ lascia alla sua sinistra $(1-\alpha)\%$ valori di Y, allora questo è per definizione il quantile di livello $(1-\alpha)$ ovvero risulta

$$Y_{(1-\alpha)} = a + bX_\alpha$$

il che vuol dire $Y_\alpha = a + bX_{(1-\alpha)}$. ■

Notes

Osserva: Se $\alpha = 0.5$, allora $\alpha = (1-\alpha)$, quindi nel caso della mediana la proprietà puo' essere scritta come

$$\mathrm{Med}(Y) = a + b\,\mathrm{Med}(X)$$

qualunque sia il segno di b.

Osserva: se le x_i sono misurate con una certa unità di misura $[u]$, in quale unità di misura sono espresse media e quantili (e quindi mediana)? È facile verificare (esercizio) che sia media che quantili sono espresse nella stessa unità di misura $[u]$.

Esempi/Applicazioni $\longrightarrow$ R **script file**

Notes