Slide Set 5
Dispersione

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 (h10:20)

Notes

Dispersione (scatter, spread)

Dispersione: nozione intuitiva
fissata una *posizione* di riferimento, ci chiediamo:

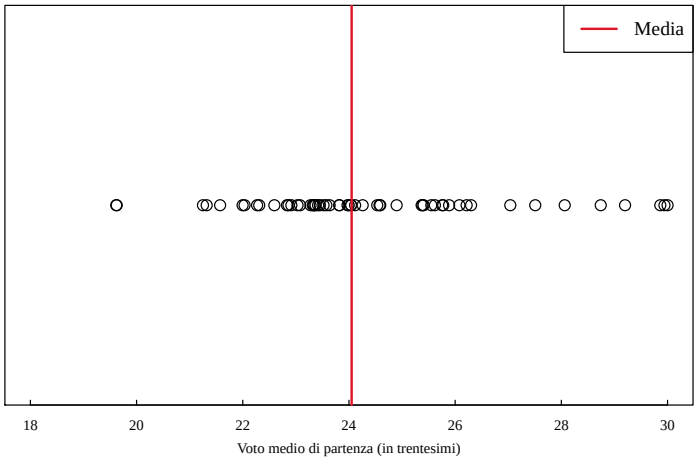
- *in media* i punti osservati quanto si *discostano* da questa posizione?
- i dati osservati, *complessivamente*, quanto sono diversi da questo riferimento?
- *quanta variazione* osserviamo rispetto a questo riferimento?

Nella letteratura spesso troviamo i sinonimi di

- scatter (dall'inglese: *spargere*, *sparpagliare*, *disperdere*, etc)
- spread (dall'inglese: *diffondere*, *dilagare*, etc)

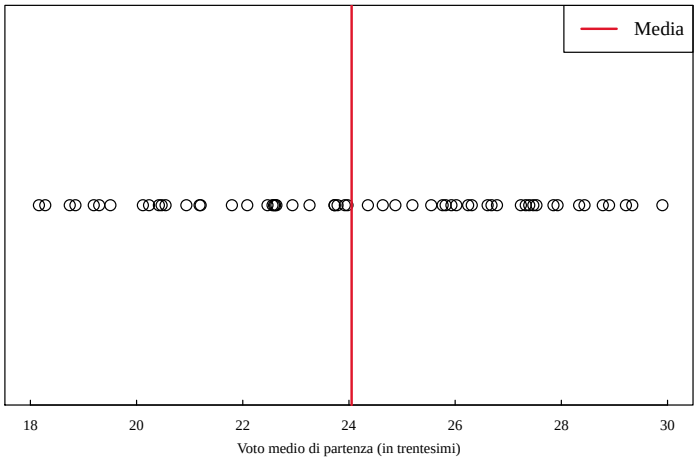
Notes

Esempio: nella sessione del giorno 14/01/2019, arrivano in seduta di laurea $n = 60$ studenti con il seguente punteggio di partenza



Notes

Esempio: nella sessione del giorno 15/01/2019, arrivano in seduta di laurea $n = 73$ studenti con il seguente punteggio di partenza



Notes

Deviazione da un punto e deviazioni totali

Supponiamo di osservare il campione $\{x_1, x_2, \ldots, x_n\}$, ad esempio

x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
-2	1	2	3	2	1	-1	2

Fissiamo un punto di riferimento $x' = 0$. Domanda: quanto devia ogni x_i da x'?

Osserva

- $x_3 = 2$ devia da $x' = 0$ tanto quanto $x_1 = -2$
- $x_7 = -1$ devia da $x' = 0$ tanto quanto $x_2 = 1$

Notes

Possiamo misurare la deviazione da x' (anche detta scostamento) usando diverse *metriche*. Le principali sono

Deviazioni quadratiche: $(x_i - x')^2$

- $(x_3 - x')^2 = (2 - 0)^2 = 4$
- $(x_1 - x')^2 = (-2 - 0)^2 = 4$
- $(x_7 - x')^2 = (-1 - 0)^2 = 1$
- $(x_2 - x')^2 = (1 - 0)^2 = 1$

Deviazioni assolute: $|x_i - x'|$

- $|x_3 - x'| = |2 - 0| = 2$
- $|x_1 - x'| = |-2 - 0| = 2$
- $|x_7 - x'| = |-1 - 0| = 1$
- $|x_2 - x'| = |1 - 0| = 1$

Notes

Nell'esempio precedente, prendendo a riferimento sempre $x' = 0$ calcoliamo tutti gli scostamenti usando le due metriche:

Osservazione: x_i	-2	1	2	3	2	1	-1	2
Dev. quadratica: $(x_i - x')^2$	4	1	4	9	4	1	1	4
Dev. assoluta: $\lvert x_i - x' \rvert$	2	1	2	3	2	1	1	2

Domande:

- globalmente, il campione osservato quanto devia dal punto $x' = 0$?
- il punto x' quanto *rappresenta* il campione osservato?

Notes

Poiché ogni deviazione è non-negativa, possiamo facilmente aggregare le deviazioni sommandole

- Deviazione totale quadratica:

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

- Deviazione totale assoluta:

$$\sum_{i=1}^{n} \lvert x_i - x' \rvert$$

Osserva: nei due casi

- deviazione totale=0 se e solo se $x_1 = x_2 =, \ldots, x_n = x'$
- *ceteris paribus*, se avviciniamo un qualche x_i a x' la deviazione totale si riduce, se allontaniamo un qualche x_i da x' la deviazione totale aumenta

Notes

Se la deviazione totale rispetto a x' è grande

- il campione osservato devia *fortemente* da x'
- x' è una posizione poco rappresentativa, o poco *centrale*, rispetto all'intera distribuzione

Se la deviazione totale rispetto a x' è piccola

- il campione osservato devia *poco* da x'
- x' è una posizione molto rappresentativa , o molto *centrale*, rispetto all'intera distribuzione

Qualunque sia il tipo di deviazione totale (quadratica o assoluta), essa misura due cose:

1. il grado di rappresentatività/centralità di x'
2. la *dispersione* complessiva rispetto a x'

Esploriamo i due punti

Notes

Centralità

Media e mediana sono misure di posizione e di *centralità*. Perché?

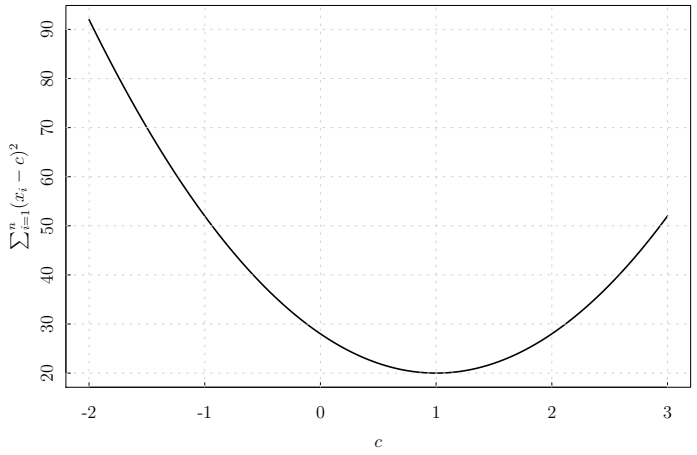
Continuiamo sempre con lo stesso esempio numerico. Facciamo variare c tra $x_{\min}$ e $x_{\max}$ e ogni volta calcoliamo le deviazioni totali:

- $\sum_{i=1}^{n}(x_i - c)^2$
- $\sum_{i=1}^{n}|x_i - c|$

Il valore c sarà tanto più rappresentativo/centrale per la distribuzione quanto più sarà piccolo il corrispondente valore della deviazione totale

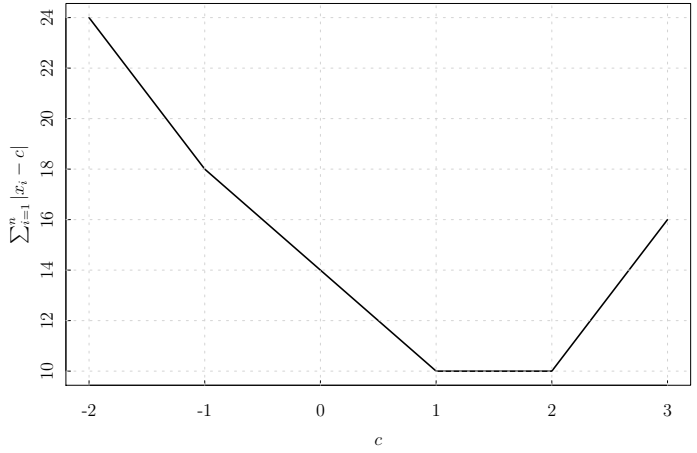
Notes

Deviazioni totali quadratiche in funzione di $c \in [x_{\min}, x_{\max}] = [-2, 3]$:



Notes

Deviazioni totali assolute in funzione di $c \in [x_{\min}, x_{\max}] = [-2, 3]$:



Notes

Proprietà (Centralità della media vs deviazione totale quadratica)

Sia $\{x_1, x_2, \ldots, x_n\}$ un campione osservato con media $\bar{x}$. Per ogni $c \in \mathbb{R}$

$$\sum_{i=1}^{n}(x_i - \bar{x})^2 \;\leq\; \sum_{i=1}^{n}(x_i - c)^2$$

Ovvero $\bar{x}$ è il punto più centrale rispetto alle deviazioni quadratiche

Proprietà (Centralità della mediana vs deviazione totale assoluta)

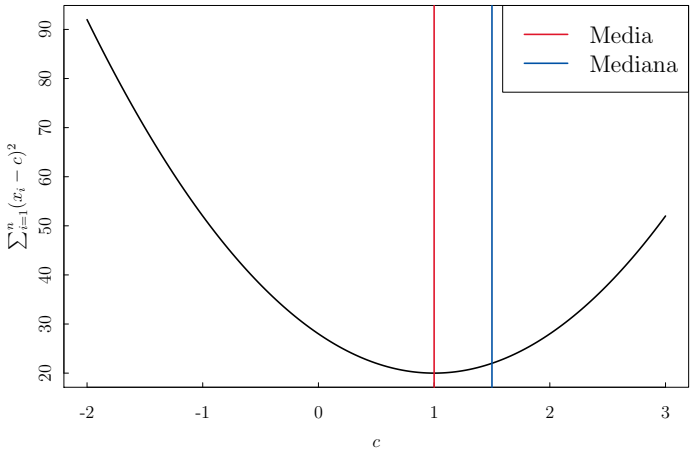
Sia $\{x_1, x_2, \ldots, x_n\}$ un campione osservato con mediana $\mathrm{Med}(x)$. Per ogni $c \in \mathbb{R}$

$$\sum_{i=1}^{n}|x_i - \mathrm{Med}(x)| \;\leq\; \sum_{i=1}^{n}|x_i - c|$$

Ovvero $\mathrm{Med}(x)$ è il punto più centrale rispetto alle deviazioni assolute

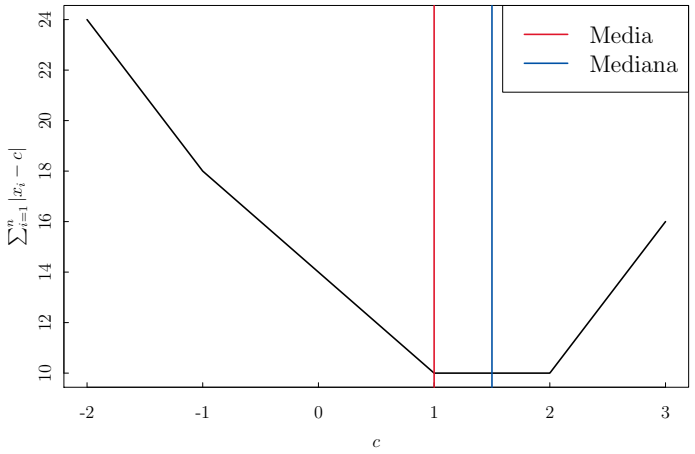
Notes

Deviazioni totali quadratiche in funzione di $c \in [x_{\min}, x_{\max}] = [-2, 3]$:



Notes

Deviazioni totali assolute in funzione di $c \in [x_{\min}, x_{\max}] = [-2, 3]$:



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

Notes

Misure di variabilità

Variabilità
è la dispersione rispetto al centro della distribuzione

Definizione (Varianza empirica)

Sia $\{x_1, x_2, \ldots, x_n\}$ un campione osservato con media $\bar{x}$. Si definisce varianza empirica la quantità

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2 \tag{5.1}$$

Nota: $\hat{\sigma}^2$ è anche chiamata "*varianza campionaria osservata*". In questo corso quando parleremo di varianza ci riferiamo sempre alla (5.1) oppure alla seguente (5.2)

Notes

Osserva

- $\hat{\sigma}^2 =$ media degli scostamenti quadratici da $\bar{x}$ pesati per la distribuzione empirica (peso $1/n$ per tutte le unità)
- $\hat{\sigma}^2$ è grande(piccola) se la dispersione rispetto a $\bar{x}$ è grande(piccola)
- $\hat{\sigma}^2$ è tanto più piccola quanto più $\bar{x}$ è rappresentativa/centrale per la distribuzione

Tuttavia, per piccoli campioni la (5.1) introduce delle distorsioni. Per valori moderati di n si preferisce utilizzare la seguente variante

Notes

Definizione (Varianza non distorta)

Sia $\{x_1, x_2, \ldots, x_n\}$ un campione osservato con media $\bar{x}$. Si definisce varianza non distorta la quantità

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \tag{5.2}$$

Osserva:

$$s^2 = \frac{n}{n-1}\hat{\sigma}^2$$

per n sufficientemente grande le due varianze sono equivalenti, infatti $\lim_{n\to+\infty} s^2 = \hat{\sigma}^2$.

Faremo sempre riferimento a s^2, ma tutte le proprietà che seguono valgono per entrambe le espressioni

Notes

Premessa
date le osservazioni $\{x_1, x_2, \ldots, x_n\}$ si chiama momento empirico (anche detto momento campionario) di ordine r la quantità

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

Osserva

- la media $\bar{x}$ coincide con il momento primo, infatti $m_1 = \bar{x}$
- Inoltre il momento secondo di X è generalmente denotato come

$$m_{xx} = m_2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2$$

Esiste una connessione importante tra deviazioni quadratiche, varianza ed i primi due momenti

Notes

Proprietà (Decomposizione della varianza)

Sia $\{x_1, x_2, \ldots, x_n\}$ un campione osservato con media $\bar{x}$, allora la deviazione totale quadratica dalla media può essere decomposta come

$$\sum_{i=1}^{n}(x_i - \bar{x})^2 = n(m_{xx} - \bar{x}^2) \tag{5.3}$$

Da cui la decomposizione

$$\hat{\sigma}^2 = m_{xx} - \bar{x}^2 \tag{5.4}$$

$$s^2 = \frac{n}{n-1}(m_{xx} - \bar{x}^2) \tag{5.5}$$

Dimostrazione. Iniziamo a provare la (5.3). Sviluppando il quadrato

$$\sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n}\left(x_i^2 + \bar{x}^2 - 2x_i\bar{x}\right),$$

da cui

Notes

$$\begin{aligned}\sum_{i=1}^{n}(x_i-\bar{x})^2 &= \sum_{i=1}^{n}x_i^2+\sum_{i=1}^{n}\bar{x}^2-\sum_{i=1}^{n}2\bar{x}x_i\\ &= nm_{xx}+n\bar{x}^2-(2\bar{x})\sum_{i=1}^{n}x_i\\ &= nm_{xx}+n\bar{x}^2-(2\bar{x})n\bar{x}\\ &= n(m_{xx}+\bar{x}^2-2\bar{x}^2)\\ &= n(m_{xx}-\bar{x}^2)\end{aligned}$$

Dalla precedente si ottiene immediatamente

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

La (5.4) e (5.4) seguono facilmente. ■

Notes

Proprietà (Varianza vs trasformazioni lineari)

Sia $\{x_1, x_2, \ldots, x_n\}$ un campione osservato con varianza $\hat{\sigma}^2_X$. Fissati i coefficienti $a, b \in \mathbb{R}$, si consideri la trasformazione lineare $Y = a + bX$. Sia $\hat{\sigma}^2_Y$ la varianza di Y, allora

$$\hat{\sigma}^2_Y = b^2\hat{\sigma}^2_X \tag{5.6}$$

e quindi $s^2_Y = b^2 s^2_X$.

Dimostrazione.

$$\begin{aligned}\hat{\sigma}^2_Y = \frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})^2 &= \frac{1}{n}\sum_{i=1}^{n}(a+bx_i-(a+b\bar{x}))^2\\ &= \frac{1}{n}\sum_{i=1}^{n}b^2(x_i-\bar{x})^2\\ &= b^2\hat{\sigma}^2_X\end{aligned}$$

Per s^2 la dimostrazione è analoga ■

Notes

Conseguenza: la varianza è espressa con il quadrato dell'unità di misura. Infatti supponiamo x_i sia espressa nell'unità di misura u, allora definito $y_i = x_i u$, applicando la precedente otteniamo

$$s_Y^2 = s_X^2 u^2$$

Per esprimere la variabilità nella stessa unità di misura ricorriamo ad una misura di variabilità ottenuta trasformando monotonamente la varianza

Definizione (Deviazione standard)

Sia $\{x_1, x_2, \ldots, x_n\}$ un campione osservato con media $\bar{x}$ e varianza s^2 (oppure $\hat{\sigma}^2$). Si definisce *deviazione standard* (anche detta *scarto quadratico medio*) la quantità

$$s = \sqrt{s^2} \qquad (\text{oppure} \quad \sqrt{\hat{\sigma}^2}) \tag{5.7}$$

Osserva: la deviazione standard è espressa nella stessa unità di misura, infatti nel caso precedente

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

Notes

Osservazioni finali sulla varianza

Assenza di variabilità

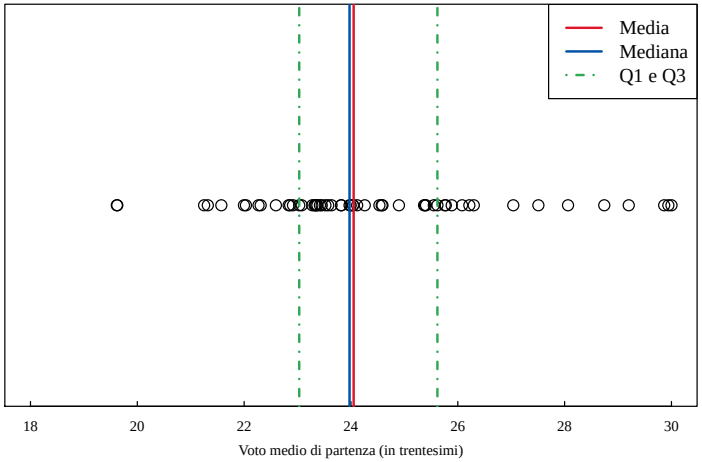
- sia la varianza che la devizione standard sono misure non negative (≥ 0)
- $s^2 = \hat{\sigma}^2 = s = \hat{\sigma}^2 = 0$ **se e solo se** tutte le x_i sono uguali alla media $\bar{x}$. Questo può accadere solo se tutte le osservazioni hanno un livello di espressione costante.

Variabilità = nozione relativa

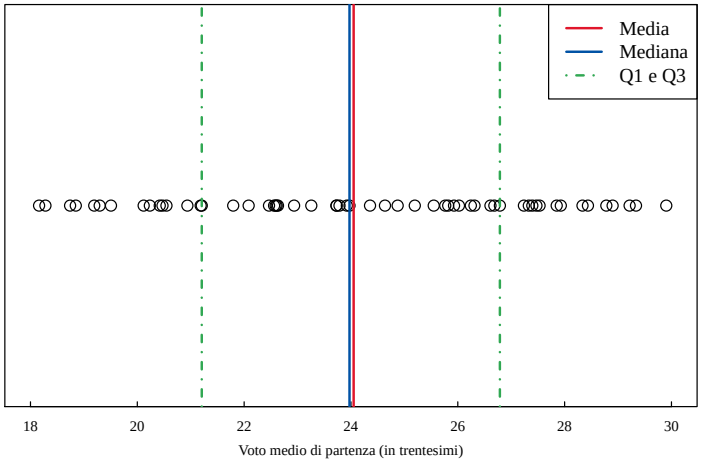
- la varianza non ha un massimo, quindi non è possibile fare valutazioni assolute
- è possibile confrontare le varianze di una stessa variabile misurata su campioni diversi, posto che i livelli siano espressi nella stessa unità di misura

Notes

Ora consideriamo come misura di centralità= mediana. E consideriamo alcuni esempi precedenti



Notes



Notes

Osserva

- all'aumentare(diminuire) della dispersione a destra della mediana, la differenza $Q_3 - \mathrm{Med}(X)$ aumenta(diminuisce)
- all'aumentare(diminuire) della dispersione a sinistra della mediana, la differenza $\mathrm{Med}(X) - Q_2$ aumenta(diminuisce)

Quindi l'aumentare(diminuire) della dispersione intorno alla mediana $\Longrightarrow$ $(Q_3 - Q_1)$ aumenta(diminuisce)

Definizione (IQR)

Sia $\{x_1, x_2, \ldots, x_n\}$ un campione osservato e siano Q_1, Q_2, Q_3 i suoi quartili. Si definisce range interquartile la differenza

$$\mathrm{IQR} = Q_3 - Q_1 = X_{75\%} - X_{25\%} \tag{5.8}$$

Notes

Quindi l'IQR è una misura di variabilità rispetto alla centralità misurata in termini di mediana

Proprietà (IQR vs trasformazioni lineari)

Sia $\{x_1, x_2, \ldots, x_n\}$ un campione osservato range interquartile pari a $\mathrm{IQR}(X)$. Fisati i coefficienti $a, b \in \mathbb{R}$, si consideri la trasformazione lineare $Y = a + bX$, allora

$$\mathrm{IQR}(Y) = |b| \, \mathrm{IQR}(X) \tag{5.9}$$

Dimostrazione. a cura dello studente! Basta usare le trasformazioni lineari per i quantili. ■

Osserva: la conseguenza della (5.9), è che l'IQR è espresso nella stessa unità di misura delle x_i

Notes

Osservazioni finali su IQR

Assenza di variabilità

- $\mathrm{IQR} \geq 0$
- $\mathrm{IQR} = 0$ **se e solo se** il *bulk* dei dati (50% delle osservazioni centrali intorno alla mediana) sono tutte pari allo stesso valore $=$ $Q_1 = \mathrm{Med}(X) = Q_3$

Variabilità $=$ nozione relativa

valgono le stesse considerazioni fatte a proposito della varianza

Notes

È facile provare che

- $x_{\min} \leq \bar{x} \leq x_{\max}$
- $x_{\min} \leq \mathrm{Med}(X) \leq x_{\max}$
- $x_{\max} - x_{\min} \geq 0$, ma $x_{\max} - x_{\min} = 0$ se e solo se non esiste alcuna variabilità nei dati

All'aumentare della variabilità ci aspettiamo che gli estremi osservati si allontanino dal centro. Di conseguenza una misura di variabilità utilizzata in alcune applicazioni (es: idrologia, finanza) è la seguente

Definizione (Ampiezza del range)

Sia $\{x_1, x_2, \ldots, x_n\}$ un campione osservato. Si definisce ampiezza del range la differenza

$$R = x_{\max} - x_{\min} \tag{5.10}$$

Come si comporta R rispetto alle trasformazioni lineari? Risposta a cura dello studente!

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

Notes

Standardizzazione dei dati

Definizione (Trasformazione standard)

Sia $\{x_1, x_2, \ldots, x_n\}$ un campione osservato con media $\bar{x}$ e varianza s_X^2, la trasformazione

$$z_i = \frac{x_i - \bar{x}}{s_X} \qquad \text{con } 1, 2, \ldots, n \tag{5.11}$$

è detta *trasformazione standard*, $\{z_1, z_2, \ldots, z_n\}$ sono detti *dati standardizzati*.

Esempio: osserviamo la variabile X con $\bar{x} = 2.71$ e $s_X = 4.86$

Dati osservati (x_i)	10	-4	8	3	0	0	2
Dati standardizzati (z_i)	1.5	-1.4	1.1	0.1	-0.6	-0.6	-0.1

⚠ Nella definizione abbiamo usato s^2, usando $\hat{\sigma}^2$ otteniamo comunque una standardizzazione dei dati

Notes

Proprietà (Trasformazione standard)

Siano $\{z_1, z_2, \ldots, z_n\}$ i dati standardizzati ottenuti da un campione $\{x_1, x_2, \ldots, x_n\}$

1. la trasformazione standard è una trasformazione lineare
2. i dati standardizzati hanno media=0 e varianza=1 indipendentemente dalla media/varianza della X
3. i dati standardizzati non dipendono dall'unità di misura di X

Dimostrazione. Proviamo prima la 1):

$$z_i = \frac{x_i - \bar{x}}{s_X} = \left(-\frac{\bar{x}}{s_X}\right) + \left(\frac{1}{s_X}\right) x_i = a + b x_i \qquad \text{con } a = -\frac{\bar{x}}{s_X},\ b = \frac{1}{s_X}$$

Notes

Proviamo la 2). Scriviamo le $z_i = a + bx_i$, applichiamo le proprietà delle trasformazioni lineari sulla media

$$\bar{z}_i \;=\; a + b\bar{x} \;=\; -\frac{\bar{x}}{s_X} + \frac{1}{s_X}\bar{x} \;=\; 0$$

Applichiamo le proprietà delle trasformazioni lineari sulla media

$$s_Z^2 \;=\; b^2 s_X^2 \;=\; \left(\frac{1}{s_X}\right)^2 s_X^2 \;=\; 1$$

Per provare la 3), supponiamo che X sia stata misurata in termini di u, sappiamo che $\bar{x}$ e s_X sono espresse in termini di u, quindi

$$\frac{x_i[u] - \bar{x}[u]}{s_X[u]} \;=\; \frac{(x_i - \bar{x})[u]}{s_X[u]} \;=\; z_i$$

■

Notes

Domanda: a cosa serve la trasformazione standard?

- centrare la distribuzione su $\bar{x}$ e *normalizzare* la sua varianza. In tal modo possiamo preparare i dati per studiare aspetti ulteriori della distribuzione (es: simmetria, spessore delle code, valori anomali, etc)
- depurare le variabili dall'unità di misura
- *normalizzare* il range dei dati. Si noti che

$$z_{\max} - z_{\min} \;=\; |b|\,(x_{\max} - x_{\min}) \;=\; \frac{1}{s_X}\,(x_{\max} - x_{\min})$$

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

Notes