Richiamo: un campione $\{x_1, \dots x_n\}$

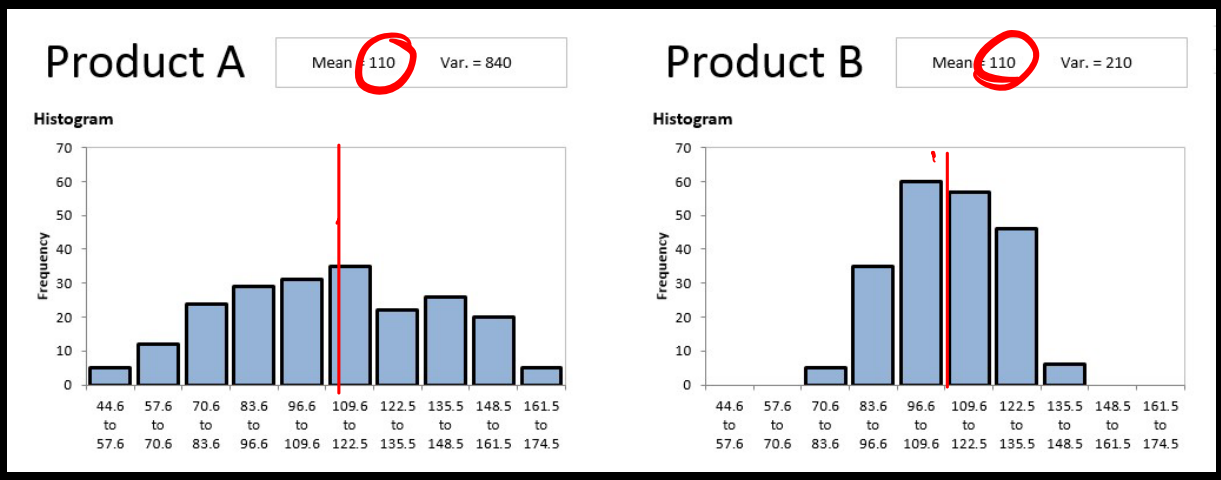
Il valore "tipico" di un campione può essere sintetizzato con le seguenti quantità.

- **media**: $\bar{x} = \frac{1}{n}(x_1 + x_2 + \dots x_n) = \frac{1}{n}\sum_{j=1}^{n} x_j$
- **mediana** ordino $\{x_1 \dots x_n\}$ in maniera crescente

$$x_{(1)}, \dots x_{(n)}$$

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

- **moda**: il valore (i) con freq. più alta.



campionaria
varianza: misura la dispersione attorno alla media

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

Nota: la presenza di $(n-1)$ e non (n) al denominatore è per ragioni tecniche.

Esempio: Consideriamo i campioni $\begin{cases} A = \{3,4,6,7,10\} \\ B = \{-20, 5, 15, 24\} \end{cases}$

$$\bar{x}_A = \frac{1}{5}(3+4+6+7+10) = 6$$

$$\bar{x}_B = \frac{1}{4}(-20+5+15+24) = 6$$

$$s_A^2 = \frac{1}{5-1}\left((3-6)^2+(4-6)^2+(6-6)^2+(7-6)^2+(10-6)^2\right) = 7.5$$

$$s_B^2 = \frac{1}{4-1}\left((-20-6)^2+(5-6)^2+(15-6)^2+(24-6)^2\right) = 360{,}67$$

Richiamo: $\overline{a \cdot x} = \frac{1}{n}\sum_{j=1}^{n} a x_j = a\,\frac{1}{n}\sum_{j=1}^{n} x_j = a \cdot \bar{x}$

$$\overline{x+b} = \frac{1}{n}\sum(x_j+b) = \frac{1}{n}\left(\sum x_j + nb\right) = \frac{1}{n}\sum x_j + \frac{1}{n}nb = \bar{x}+b$$

Proprietà 1: come nel caso della media, l'espressione per la varianza campionaria di dati trasformati con

$$y_i = a x_i + b \qquad a, b \in \mathbb{R}$$

prende una forma semplice:

$$s_y^2 = \frac{1}{n-1}\sum_{j=1}^{n}(y_j-\bar{y})^2 = \frac{1}{n-1}\sum\left(ax_j + b - (a\bar{x}+b)\right)^2$$

$$= \frac{1}{n-1}\sum_{j=1}^{n}(ax_j - a\bar{x})^2 = \frac{1}{n-1}\sum_{j=1}^{n} a^2(x_j-\bar{x})^2 = a^2\,\underbrace{\frac{1}{n-1}\sum_{j=1}^{n}(x_j-\bar{x})^2}_{s_x^2}$$

$$= a^2 s_x^2$$

Definiamo infine la radice quadrata della varianza con il nome deviazione standard.

Def: La deviazione standard campionaria di un insieme di dati $\{x_1 \dots x_n\}$ è

$$s_x := \sqrt{\frac{1}{n-1}\sum_{j=1}^{n}(x_j - \bar{x})^2} = \sqrt{s_x^2}$$

Esempio (Incidenti aerei).

$n = 9$

Year	1997	1998	1999	2000	2001	2002	2003	2004	2005
Accidents	25	20	21	18	13	13	7	9	18

Source: National Safety Council.

$= x_j$

$$\bar{x} = \frac{1}{9}(25 + 20 + 21 + \dots + 18) = 16$$

$$s_x^2 = \frac{1}{9-1}\left((25-16)^2 + (20-16)^2 + \dots\right) = 34.75$$

$$s_x = \sqrt{s_x^2} = \sqrt{34.75} = 5.89$$

Percentili campionari e box plot

Esempio

Talvolta è utile identificare la parte del campione che ha voti nel "miglior X %" (esempio il miglior 5 % della classe)

In statistica identifichiamo il valore che separa il k % più basso (il "peggior" k %) dei dati dal (100-k) % più alto (il "miglior" (100-k) %) come il k-percentile

INT94 TAV. 1

TASSO ALUNNI INTERNI PROMOSSI

Esami di Stato - Anno Scol.: 93-94

Valori per 100 Alunni Scrutinati (In ordine decrescente)

Scuole statali e Non Statali

REGIONI	93-94 Interni
Molise	99,1
Emilia Rom.	98,3
Piemonte	98,3
Lombardia	98,2
Umbria	98,2
Veneto	98,2
Puglia	98,0
Trentino-Alto A.	97,9
Campania	97,9
Calabria	97,8
Friuli Venezia-G.	97,7
Toscana	97,6
Marche	97,4
Basilicata	97,3
Liguria	97,2
Abruzzo	96,6
Sardegna	96,0
Lazio	96,0
Sicilia	95,8
Valle d'Aosta	95,7

$n = 20$

Def: Sia k un numero intero con $0 \leq k \leq 100$.

Il k-percentile campionario è il valore del punto dato x_j tale che (almeno)

$$\begin{cases} k\,\% \text{ del database è minore o uguale a } x_j \\ (100-k)\,\% \text{ del database è maggiore o uguale a } x_j \end{cases}$$

Se due punti dato soddisfano questa condizione allora il k-percentile è la loro media aritmetica.

Esempio: il 93-percentile è: 98.3

In pratica, per un campione di ampiezza n, il k-percentile si ottiene identificando i valori (al massimo 2) che soddisfano

1. Almeno $n \cdot \frac{k}{100}$ valori è minore o uguale ad essi
2. Almeno $n \cdot \frac{100-k}{100}$ valori è maggiore o uguale ad essi

Quindi, per un database $\{x_1, \dots x_n\}$ ordinato in maniera crescente

- se $n\frac{k}{100}$ non è intero esiste un solo tale punto:
  il primo valore intero maggiore di $n\frac{k}{100}$
- se $n\frac{k}{100}$ è intero ne esistono due: $\{n\frac{k}{100}, n\frac{k}{100}+1\}$
  $\rightarrow$ si fa la media

Processo (calcolo del k-percentile)

Calcola $n\frac{k}{100} = m$

- m non intero → k-perc: numero intero maggiore di m
- m intero → k-perc: $\frac{x_m + x_{m+1}}{2}$

Esempio (Popolazione città US)

Si trovi il 75-percentile $n = 25$

$$n\frac{k}{100} = 25 \cdot \frac{75}{100} = 18.75 \to 19$$

75 perc : 1.296 682

Si trovi l'80-percentile.

$$n \cdot \frac{k}{100} = 25 \cdot \frac{80}{100} = 20 \quad \begin{cases} 20 \\ 21 \end{cases}$$

l'80 percentile è la media tra il 20 esimo e 21-esimo elemento del basso

$$\frac{1448394 + 1512986}{2} = 1480690$$

TABLE 2.6 *Population of 25 Largest U.S. Cities, July 2006*

Rank	City	Population
1	New York, NY	8,250,567
2	Los Angeles, CA	3,849,378
3	Chicago, IL	2,833,321
4	Houston, TX	2,144,491
5	Phoenix, AR	1,512,986
6	Philadelphia, PA	1,448,394
7	San Antonio, TX	1,296,682
8	San Diego, CA	1,256,951
9	Dallas, TX	1,232,940
10	San Jose, CA	929,936
11	Detroit, MI	918,849
12	Jacksonville, FL	794,555
13	Indianapolis, IN	785,597
14	San Francisco, CA	744,041
15	Columbus, OH	733,203
16	Austin, TX	709,893
17	Memphis, TN	670,902
18	Fort Worth, TX	653,320
19	Baltimore, MD	640,961
20	Charlotte, NC	630,478
21	El Paso, TX	609,415
22	Milwaukee, WI	602,782
23	Boston, MA	590,763
24	Seattle, WA	582,454
25	Washington, DC	581,530

Def:
- Il 25-percentile campionario è il primo quartile
- Il 50-percentile campionario è il secondo quartile
- Il 75-percentile campionario è il terzo quartile

Nota: il secondo quartile è la mediana

Un modo di rappresentare un database attraverso i suoi quartili è il boxplot:

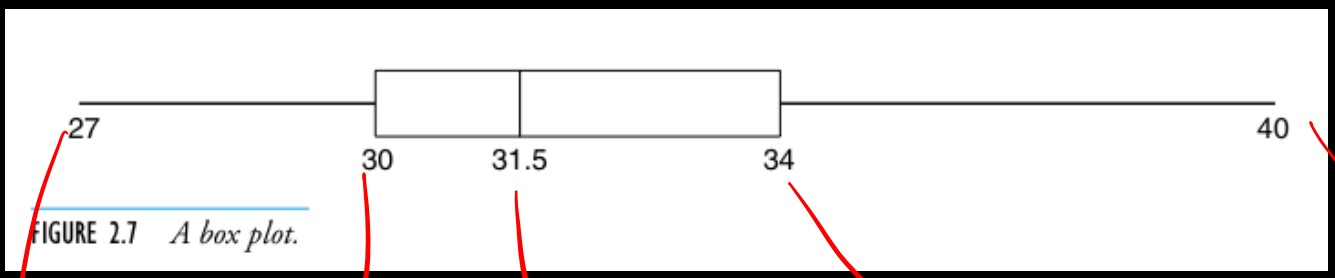


FIGURE 2.7 *A box plot.*

mediana

primo quartile

terzo quartile

valore estremo (minimale)

valore estremo (massimale)

Covarianza e correlazione campionaria

A volte siamo confrontati con database in cui ogni punto dato ha più di una caratteristica.

Esempio: homes.csv contiene 9 caratteristiche.

- prezzo di vendita
- prezzo di listino
- superficie
- # camere
- # letti
- # bagni
- età
- ettari
- tasse.

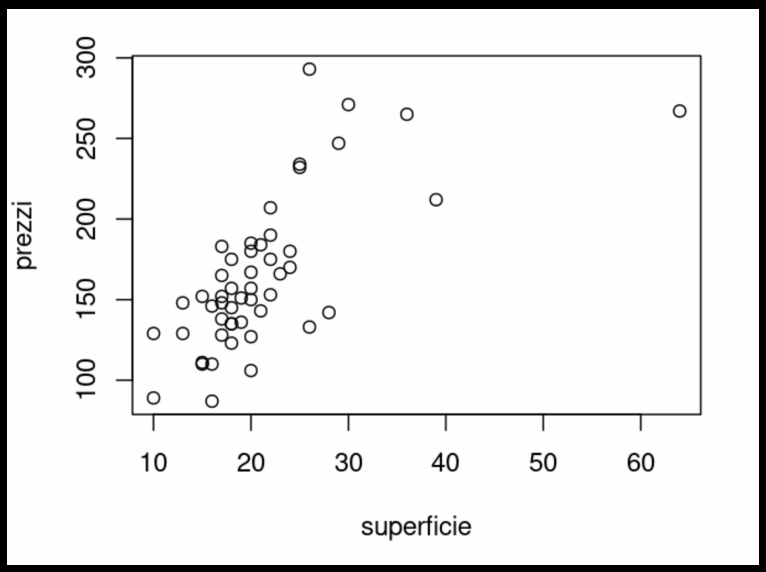
Vorremmo capire se la superficie e il prezzo di una casa sono correlati tra loro. (cioè se valori grandi di una caratteristica corrispondono a valori grandi dell'altra)

Tentativo #1: tabella

x / y																			
Sell	142	175	129	138	232	135	150	207	271	89	153	87	234	106	175	165	166	136	148
Living	28	18	13	17	25	18	20	22	30	10	22	16	25	20	22	17	23	19	17

151	180	293	167	190	184	157	110	135	267	180	183	185	152	148	152	146	170	127	265
19	24	26	20	22	21	20	16	18	64	20	17	20	17	13	15	16	24	20	36

Tentativo #2: grafico a dispersione



Vediamo che valori grandi di superficie corrisp. a valori grandi di prezzo.

Vorremmo catturare questo "associazione" con un numero

Def: Chiamiamo un campione/insieme di dati **bivariato** quando ad ogni osservazione/punto dato corrispondendono due valori accoppiati (x_i, y_i)

L'insieme di dati ha una forma

$$\{(x_1,y_1), (x_2,y_2) \dots (x_n,y_n)\} \quad (*)$$

Def: Dato un campione bivariato (*) definiamo la sua **covarianza campionaria**

$$q = \frac{1}{n-1} \sum_{j=1}^{n} (x_j - \bar{x})(y_j - \bar{y})$$

dove $\bar{x} = \frac{1}{n} \sum x_j \quad \bar{y} = \frac{1}{n} \sum y_j$

Esempio: x_j = superficie casa j $\quad \bar{x} = \frac{1}{n} \sum x_j = 21.12$

y_j = prezzo casa j $\quad \bar{y} = \frac{1}{n} \sum y_j = 164.36$

x \ y								
Sell	142	175	129	138	232	135	150	207
Living	28	18	13	17	25	18	20	22

$$q = \frac{1}{50-1}\left[(28 - 21.12)(142 - 164.36) + \dots\right] = 271.77$$

Notiamo che se trasformiamo $(x_j, y_j) \to (ax_j, by_j)$

la covarianza cambia ...

Esempio