Corso di Modelli per l'analisi statistica

Prof. G. Scandurra

a.a. 2020-2021

Variabilità

- Il calcolo di una media non esaurisce la descrizione sintetica di un fenomeno osservato in un collettivo
- Due insiemi di valori o due distribuzioni di frequenza, pur avendo lo stesso valore medio, possono essere molto differenti tra di loro
- Gli indici di variabilità forniscono informazioni complementari a quelle degli indici medi

Punti vendita	Ricavi	Costi	addetti	ubicazione	Genere respons.	Vendita On-line	R.O
1	350	205	5	centro	maschio	si	145
2	200	100	3	periferia	maschio	si	100
3	600	350	10	semicentro	femmina	no	250
4	500	270	10	periferia	femmina	no	230
5	270	200	6	centro	maschio	no	70
6	180	120	3	centro	maschio	no	60
7	205	105	3	periferia	maschio	no	100
8	340	210	5	semicentro	femmina	no	120
9	280	140	4	centro	femmina	si	140

Variabilità

Distribuzioni teoriche

Ricavi	Ricavi (A)	Ricavi (B)	Ricavi (C)
350	325	300	140
200	325	350	270
600	325	400	830
500	325	200	605
270	325	300	120
180	325	325	200
205	325	300	190
340	325	400	200
280	325	350	370

Distribuzione osservata

Le 3 distribuzioni teoriche hanno la stessa media della distribuzione osservata

$$\bar{x} = 325$$

La sintesi con la media aritmetica porta allo stesso risultato

Eppure le distribuzioni sono molto diverse tra di loro

Distrib. (A)
Variabilità nulla
Tutti i valori uguali

Passando da (A) a (B) e da (B) a (C), la variabilità aumenta

Misure di variabilità come ampiezza di un intervallo

Il range (o campo di variazione) $\text{range} = x_{max} - x_{min}$

è l'ampiezza dell'intervallo che contiene tutti i valori

La differenza interquartile $dQ = Q_3 - Q_1$

è l'ampiezza dell'intervallo che contiene il 50% dei valori (quelli centrali)

La variabilità aumenta al crescere di questi indici

Range e dQ possono non essere misure pienamente affidabili perché dipendono solo da due valori

Calcolo del *range* e della differenza interquartile

Ricavi
350
200
600
500
270
180
205
340
280

x_{min}	180
x_{max}	600
$Range=x_{max}-x_{min}$	**420**

I ricavi di tutti i punti vendita sono compresi in un intervallo di ampiezza 420

Q_1	205
Q_3	350
$dQ=Q_3-Q_1$	**145**

I ricavi del 50% dei punti vendita (quelli che occupano le posizioni centrali) sono compresi in un intervallo di ampiezza 145

Misure di variabilità come dispersione dalla media

La varianza σ^2 è funzione delle differenze (scarti) tra ogni valore x_i e la media $\bar{x}$

$$\sigma^2 = \frac{1}{n}\sum_{1=1}^{n}(x_i - \bar{x})^2 \qquad \sigma^2 \geq 0$$

La devianza è il numeratore della varianza

$$\text{Dev}(X) = \sum_{i=1}^{n}(x_i - \bar{x})^2$$

Misure di variabilità come dispersione dalla media

La deviazione standard (o scarto quadratico medio) è la radice quadrata della varianza

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

Il coefficiente di variazione CV è il rapporto tra la dev. standard e la media moltiplicato per 100

$$CV = \frac{\sigma}{\bar{x}}100 \qquad \bar{x} > 0$$

Confronto tra due distribuzioni in termini di variabilità

CV si calcola per confrontare la variabilità della distribuzione del carattere X con quella del carattere Y quando sono espressi o con diversa unità di misura (ad es. X in migliaia di euro e Y in metri quadri) o con diverso ordine di grandezza (ad es. X e Y hanno medie sensibilmente diverse)

Se $CV_X > CV_y$ allora la variabilità del carattere X è maggiore di quella del carattere Y

Variabilità

Ricavi x_i	Scarti dalla media $(x_i - \bar{x})$	Quadrato degli scarti $(x_i - \bar{x})^2$
350	25	625
200	-125	15625
600	275	75625
500	175	30625
270	-55	3025
180	-145	21025
205	-120	14400
340	15	225
280	-45	2025

media $\bar{x} = 325$

$$\sum_{i=1}^{n}(x_i - \bar{x}) = 0$$

Per la proprietà della media

$$\sum_{i=1}^{n}(x_i - \bar{x})^2 = \text{Dev}(X) = 163200$$

Devianza=163200

$$\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{\text{Dev}(X)}{n} = \sigma^2 =$$

$$= \frac{163200}{9} = 18133{,}3$$

Varianza=18133,3

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

$$= \sqrt{18133{,}3} = 134{,}7$$

Dev.std.=134,7

Variabilità dei ricavi dei punti vendita

- Un basso grado di variabilità indica che i punti vendita realizzano performance simili (i ricavi si discostano poco tra di loro)

- Viceversa un alto grado di variabilità fa capire che c'è una certa eterogeneità nei risultati delle vendite ottenuti nei diversi negozi

Varianza di una distribuzione di frequenza

Addetti (x_j)	Numero punti vendita (n_j)	$(x_j - \bar{x})^2 \cdot n_j$
3	2	19,34
4	1	4,45
6	3	0,04
7	1	0,79
10	2	30,26

$$\bar{x} = 6{,}11$$

$$\sigma^2 = \frac{1}{n}\sum_{j=1}^{K}(x_j - \bar{x})^2 n_j = \frac{54{,}88}{9} = 6{,}10$$

$$\sigma = \sqrt{6{,}10} = 2{,}47 \qquad CV = \frac{2{,}47}{6{,}11}100 = 40{,}43\%$$

Media e varianza da una distribuzione di freq. relative

$$\bar{x} = \frac{\sum_{j=1}^{K} x_j \cdot n_j}{n} = \sum_{j=1}^{k} x_j \cdot \frac{n_j}{n} = \sum_{j=1}^{k} x_j \cdot f_j = x_1 \cdot f_1 + x_2 \cdot f_2 + \ldots + x_k \cdot f_k$$

$$\sigma^2 = \frac{1}{n}\sum_{j=1}^{K}(x_j - \bar{x})^2 n_j = \sum_{j=1}^{K}(x_j - \bar{x})^2 \frac{n_j}{n} = \sum_{j=1}^{K}(x_j - \bar{x})^2 f_j =$$

$$= (x_1 - \bar{x})^2 \cdot f_1 + (x_2 - \bar{x})^2 \cdot f_2 + \ldots + (x_k - \bar{x})^2 \cdot f_k$$

Confronto del rendimento di due investimenti (uguale media)

	F_1	F_2
2003	7,7	6,4
2004	6,1	5,9
2005	0,4	3,2
2006	9,8	7,1
2007	3,5	4,9
media	5,5	5,5
var	10,7	1,8

Negli ultimi cinque anni, due fondi di investimento F_1 e F_2 hanno avuto lo stesso rendimento medio annuo, ma le varianze sono molto diverse $Var(F_1)>Var(F_2)$

Una varianza maggiore indica che rendimenti molto diversi dalla media sono più frequenti

➡ Maggiore volatilità ➡ Maggior rischio

A parità di rendimento medio, il cliente che è disposto ad accettare un rischio più alto sceglierà di investire in F_1

Confronto del rendimento di due investimenti (media diversa)

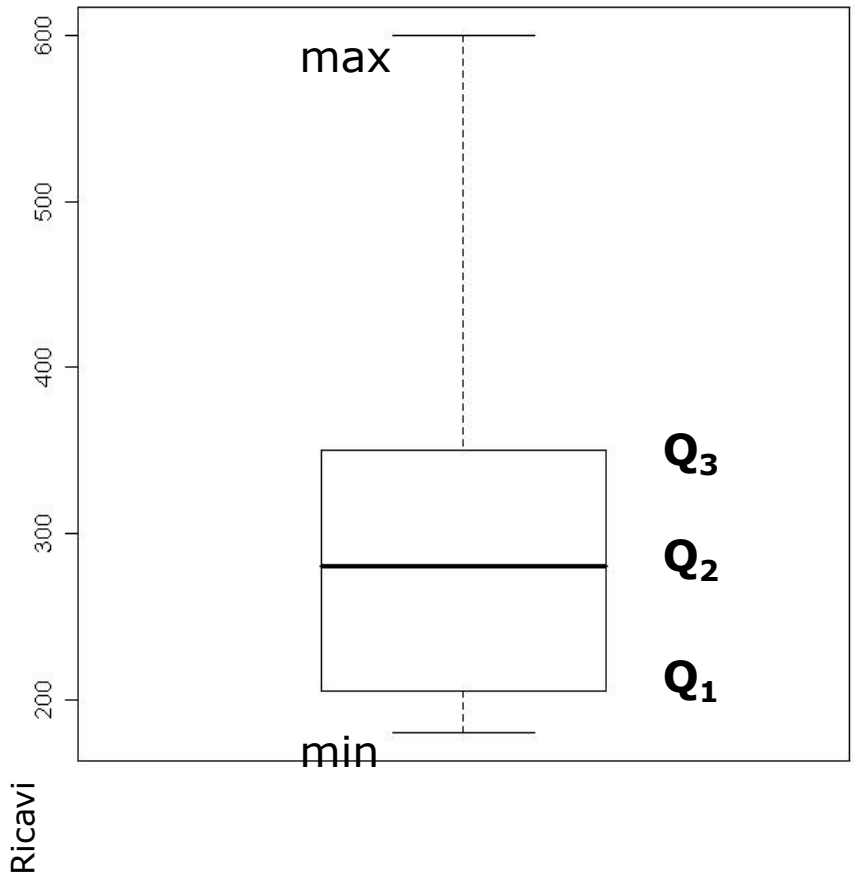
	F_1	F_2
2003	9,7	1,4
2004	7,1	1,9
2005	0,9	2,2
2006	9,9	2,1
2007	7,5	4,9
media	7,0	2,5
var	10,6	1,5
CV	46,5	49,3

Il rendimento di F_1 ha registrato una media e una varianza superiore a quello di F_2
Si può concludere che F_1 rappresenta un investimento più rischioso rispetto a F_2?

Le due medie hanno un ordine di grandezza diverso

⇒ la variabilità si confronta con CV
A F_1 è associata una variabilità (volatilità) più bassa

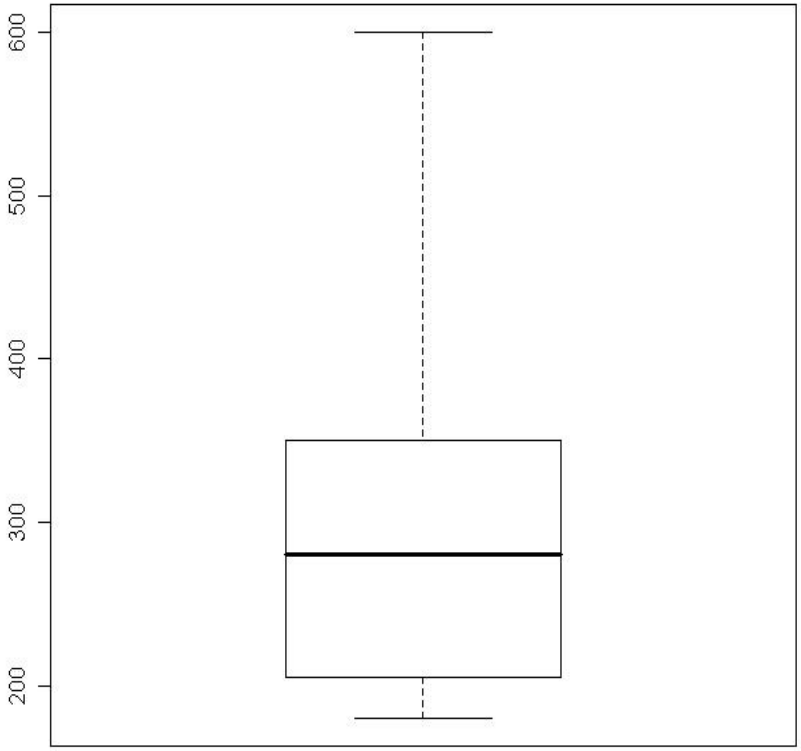
Box plot



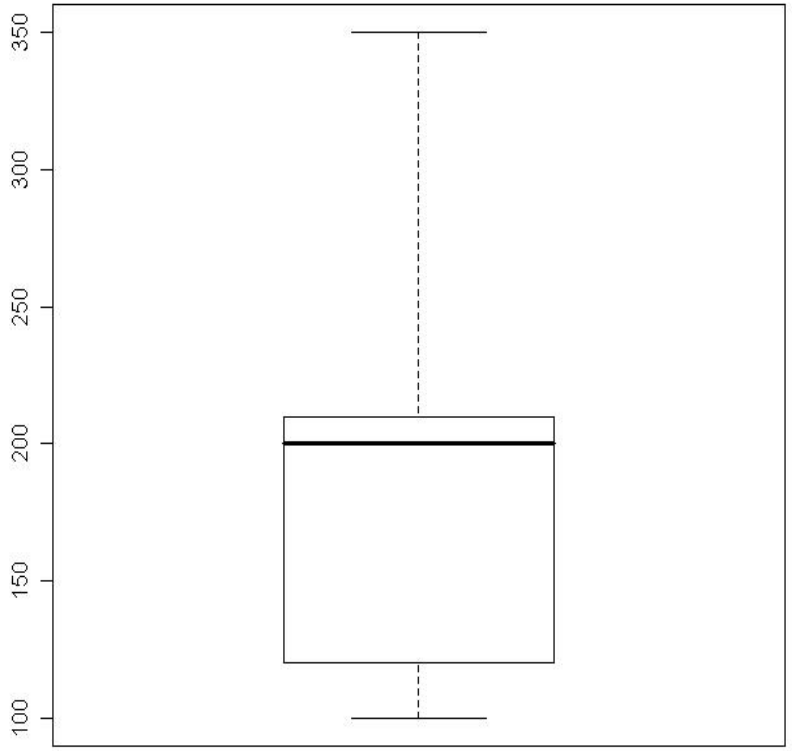
L'altezza del box indica la dispersione del 50% delle osservazioni centrali intorno alla mediana.
Si evidenzia una certa simmetria nella parte centrale, dato che la differenza Q_2-Q_1 non è molto diversa da Q_3-Q_2

I segmenti esterni al box indicano la dispersione dei valori estremi.
Si nota una maggiore dispersione del 25% dei valori più grandi, dato che la differenza max-Q_3 è molto maggiore di Q_1-min

Box plot: ricavi e costi



Ricavi



Costi

Media e deviazione standard

La deviazione standard σ è espressa nella stessa unità di misura della media $\overline{x}$

Media e deviazione standard forniscono, insieme, una sintesi globale di un insieme di valori

Media e deviazione standard costituiscono le caratteristiche chiave (parametri) di molte distribuzioni