Misure di variabilità

Che cos'è la variabilità?

È l'attitudine delle unità di un collettivo ad assumere differenti modalità di un carattere.

È necessario verificare se le unità statistiche assumono modalità molto diverse (alta variabilità) o, viceversa, se presentano modalità simili (bassa variabilità)

Esempio, prendendo l'aula come collettivo di riferimento:

- *Abbiamo tutti la stessa statura?*
- *Percepiamo lo stesso stipendio?*
- *Ci rechiamo nella stessa località di villeggiatura?*
- *Scegliamo le stesse auto?*

Che cos'è la variabilità?

Se le stature fossero tutte uguali fra loro si direbbe che non c'è variabilità.

La variabilità si può misurare con diversi tipi di indicatori, ognuno dei quali deve rispettare alcuni requisiti:

- **se la variabilità è nulla l'indicatore deve assumere valore 0;**
- **l'indicatore deve crescere al crescere della variabilità;**
- **l'indicatore può assumere soltanto valori positivi.**

Variabilità: dispersione e disuguaglianza

Se si vuole misurare la variabilità dei dati rispetto ad un valore di tendenza centrale (ad esempio la media) allora si utilizzeranno misure di dispersione dei dati intorno a tale valore... se, viceversa, s'intende misurare la disuguaglianza tra le modalità assunte da un carattere si parlerà di eterogeneità e di concentrazione.

Variabilità: dispersione

Quantificare la dispersione può essere utile come misura del rischio che deve essere affrontato nella formulazione dei processi decisionali (investimenti, introduzione in nuovi mercati, analisi dei concorrenti)…nell'ipotesi, ad esempio, che le medie di due distribuzioni siano simili ma la variabilità/dispersione dei dati intorno alla media sia molto diversa!

Vediamo in che modo…

Preferireste sostenere un esame di statistica con un professore, sapendo che il voto medio preso al suo esame è 21,5/30, con una dispersione dei voti rispetto alla media bassa oppure preferireste sostenere lo stesso esame con un altro prof. la cui votazione media è 22,5/30 ma la variabilità è molto elevata???

Distribuzione dei voti del 1° prof:
19 22 25 20 20 23 25 25 18 24 18 19

Distribuzione dei voti del 2° prof:
17 19 21 17 19 18 30 30 17 19 30 30

Misure di dispersione o variabilità

1) Campo di variazione o *range*
2) Differenza interquartilica
3) Varianza
4) Scarto quadratico medio
5) Coefficiente di variazione

CAMPO DI VARIAZIONE:
distribuzione di unità

È la differenza tra il valore massimo e minimo della distribuzione:

$$Range = X_{MAX} - X_{min}$$

azienda	vendite (in migliaia)
RENA	600
FOR	550
TOYO	720
FER	100
ALF	1000
KI	800

Range = 1000 – 100 = 900

Le vendite di tutte le aziende presentano una variabilità di 900,000 euro

CAMPO DI VARIAZIONE:
distribuzione di frequenze

Il numero di autovetture vendute (variabile) da 19 concessionarie:

azienda	vendite	N. concessionarie
RENA	50	2
FOR	150	5
TOYO	100	1
FER	10	1
ALF	30	4
KI	80	6
Totale		19

Range = 150 – 10 = 140

Le vendite delle 19 concessionarie presentano una variabilità di 140,000 euro

Differenza interquartilica

La differenza interquartilica (D.I.) indica l'ampiezza dell'intervallo centrale nel quale è compreso il 50% del collettivo statistico.

$$D.I. = Q_3 - Q_1$$

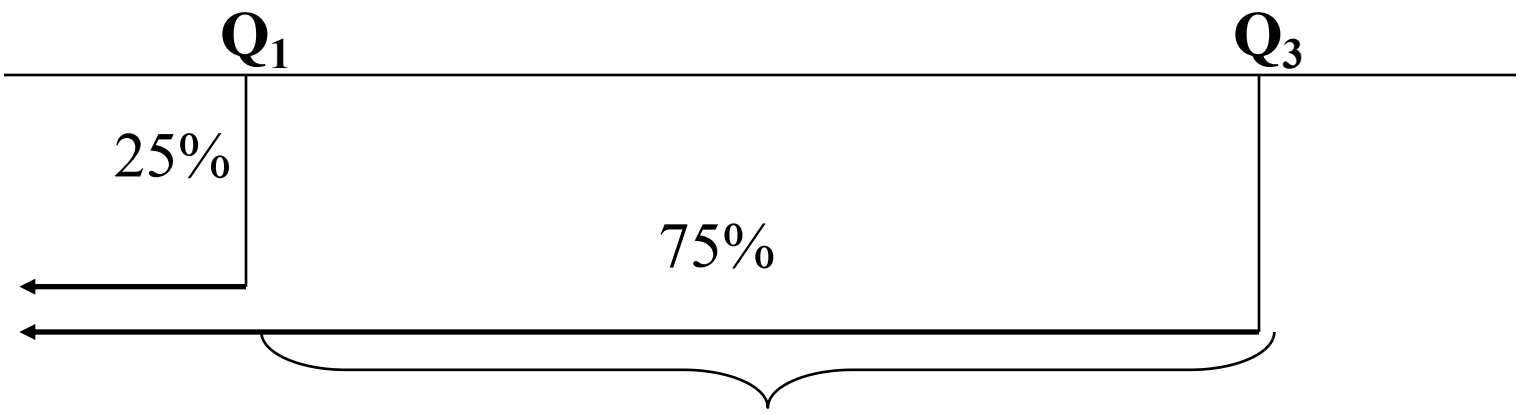
Conseguenza: più il fenomeno è "accentrato" intorno alla mediana tanto più la D.I. è piccola.

ATTENZIONE

D.I. può essere nulla anche in presenza di variabilità.

Ad es. quando il primo e il terzo quartile ricadono nella stessa posizione!!

DIFFERENZA INTERQUARTILICA



75%-25% = 50%

VARIANZA: distribuzioni di unità

È la somma dei quadrati delle differenze tra ciascuna modalità e la media aritmetica, divisa per N.
È definita, anche, come la media aritmetica degli scarti di ciascuna modalità dalla media aritmetica, elevati al quadrato

$$\mathrm{Var}(X) = \sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \overline{X})^2}{N}$$

DEVIAZIONE STANDARD O SCARTO QUADRATICO MEDIO (s.q.m.): distribuzioni di unità

Radice quadrata della media aritmetica del quadrato degli scarti di ciascuna modalità dalla media aritmetica

$$s.q.m. = \sigma = \sqrt{\sigma^2} = \sqrt{\frac{\sum_{i=1}^{N}\left(x_i - \overline{X}\right)^2}{N}}$$

Ridimensiona l'unita di misura della variabile

Interpretazione attesa: il valore trovato evidenzia un'oscillazione nell'intervallo compreso tra $\overline{X} \pm \sigma$

SI UTILIZZANO GLI S.Q.M. PER CONFRONTARE LE VARIABILITA' ASSOLUTE DI DUE DISTRIBUZIONI, QUANDO:

- le medie aritmetiche delle distribuzioni hanno una grandezza simile
- i caratteri sono della stessa natura e sono espressi nella stessa unità di misura

Esempio

La paga settimanale (in euro) di 5 operatori di un *call-center* è la seguente 500, 100, 400, 300, 50

Calcolare:

1. Varianza
2. s.q.m.

Esempio: calcolo di var(x)

	x_i	$(x_i-\overline{X})^2$
	50	48,400
	100	28,900
	300	900
	400	16,900
	500	52,900
Totale	1,350	148,000
media	270	29,600

La variabilità media delle paghe, rispetto alla paga media di 270 euro, è di circa 29,600

$$Var(X) = \sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \overline{X})^2}{N} = \frac{148,000}{5} = 29,600$$

Esempio: calcolo di s.q.m.(x)

x_i	$(x_i - \bar{X})^2$
50	48,400
100	28,900
300	900
400	16,900
500	52,900
1,350	148,000
	29,600

Lo scarto quadratico medio è:

$$\text{s.q.m.}(x) = \sigma = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \bar{X})^2}{N}} = \sqrt{29{,}600} = 172.05$$

INTERPRETAZIONE

Lo scarto delle paghe rispetto alla paga media di 270 euro è, in media, di 172 euro…

… ciò significa che la paga può variare tra 98 euro (270-172) e 442 euro (270+172)

Nell'intervallo 98-442 ($\bar{X} \pm \sigma$) euro, ricade il 66% delle osservazioni (ricadono infatti 3 osservazioni, 3 paghe su 5 =3/5 = 0,66=66%)

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

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

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

$$\sigma = \sqrt{6{,}10} = 2{,}47$$

Media e varianza da una distribuzione di freq. relative

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

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

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

Passi per il calcolo della varianza

1) Definire il tipo di distribuzione (unità – frequenze- in classi)
2) Aggiungere nuove colonne per:
1) Calcolare la media aritmetica semplice o ponderata (per frequenze non unitarie)
2) Calcolare gli scarti dalla media ed elevarli al quadrato
3) Moltiplicare gli scarti al quadrato per le rispettive frequenze assolute, se diverse dall'unità
3) Sommare gli scarti al quadrato (devianza) e dividere per il totale delle frequenze

Misura relativa di dispersione

COEFFICIENTE DI VARIAZIONE (CV)

Misura il confronto, in termini di dispersione dei dati intorno alla media, tra due o più fenomeni che possono essere caratterizzati da differenti:

- unità di misura
- circostanze o aree di misurazione

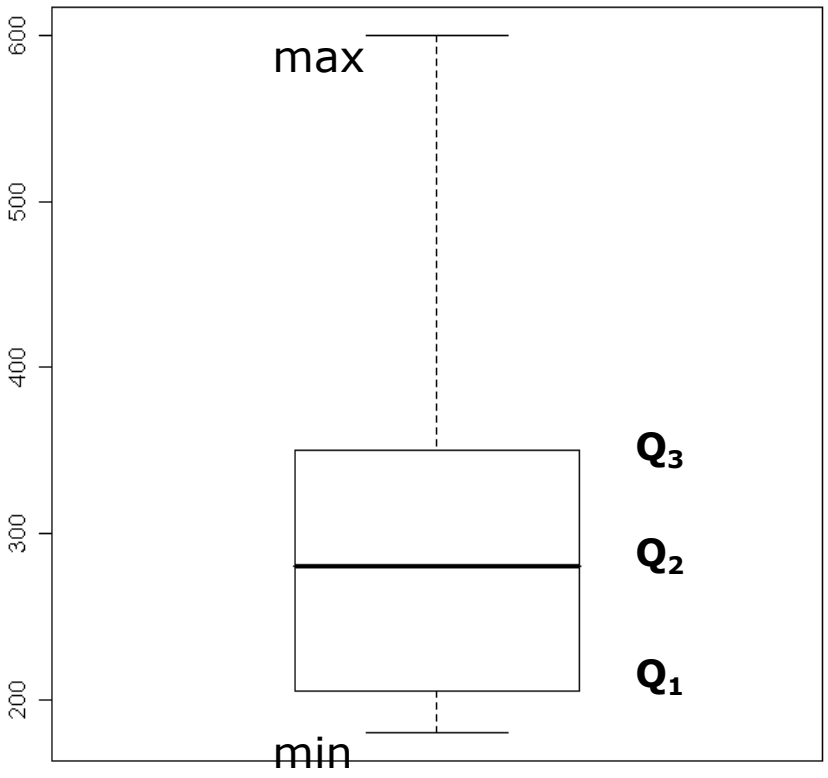
$$CV = \frac{\sigma}{\bar{x}} \times 100$$

N.B. è quasi sempre espresso in termini percentuali (%)

Appart. ti	Centro (X)	Periferia (y)	$(x_i - \bar{X})^2$	$(y_i - \bar{y})^2$
1	940	575	5,852.3	23,531.56
2	955	690	3,782.3	1,474.56
3	965	694	2,652.3	1,183.36
4	975	705	1,722.3	547.56
5	980	725	1,332.3	11.56
6	985	725	992.25	11.56
7	999	745	306.25	275.56
8	1,000	750	272.25	466.56
9	1,119	775	10,506.25	2,171.56
10	1,247	900	53,130.25	29,446.56
totali	10,165	7,284	80,549	59,120.4
media	1,016.5	728.4		
VAR	8,055	5,912		
s.q.m.	89.75	76.89		
CV (%)	8.8	10.6		

Centro	Periferia
89.75/1,016.5 =0.088	< 76.89/728.4 =0.106

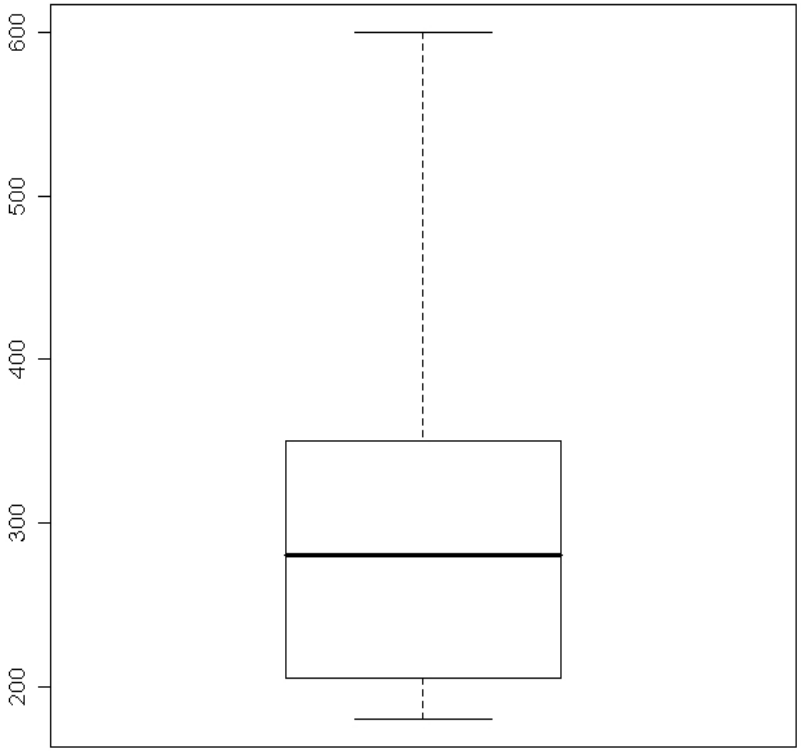
Box plot



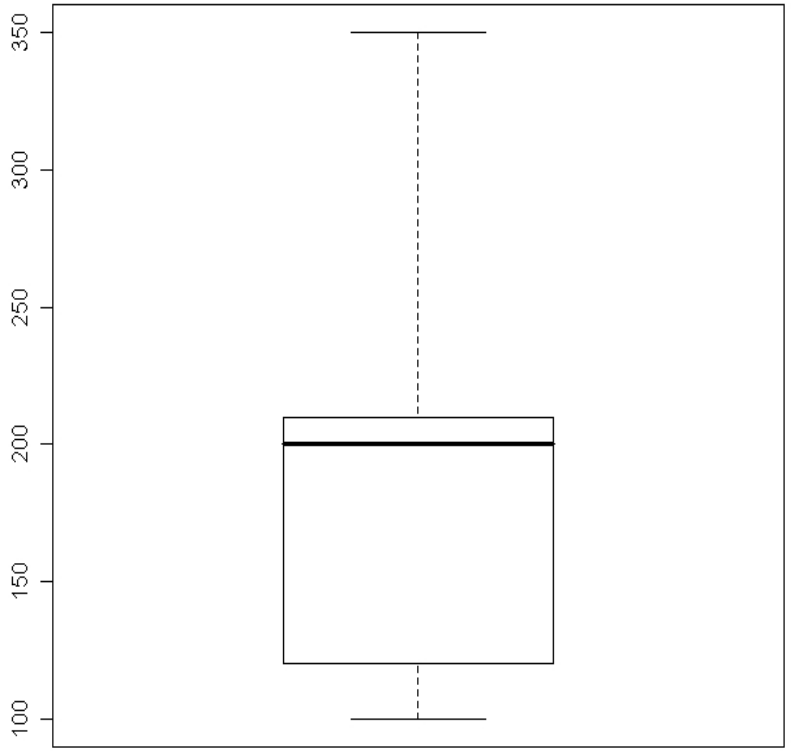
L'altezza del box indica la dispersione del 50% delle osservazioni centrali intorno alla mediana.
Si evidenzia una certa simmetria nella parte centrale, dato che la differenza Q_2-Q_1 non è molto diversa da Q_3-Q_2

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

Teorema di Chebyshev

Conoscendo solo media $\bar{X}$ e deviaz. standard σ di una distribuzione qualunque di valori, possiamo ricavare la proporzione minima di osservazioni comprese in un intervallo centrato intorno alla media, di ampiezza pari ad un multiplo di σ

Se abbiamo ulteriori informazioni sulla forma della distribuzione (se sappiamo che i valori seguono una nota distribuzione chiamata gaussiana o normale), possiamo ricavare la proporzione esatta di osservazioni comprese in un intervallo dello stesso tipo

Teorema di Chebyshev

Nell'intervallo $(\bar{x} - k\sigma, \bar{x} + k\sigma)$ cade una proporzione (freq. rel.) di valori almeno pari a $1 - \frac{1}{k^2}$

Qualunque sia la forma della distribuzione!

k	Proporzione minima di osservazioni che cadono tra $\bar{x} - k\sigma$ e $\bar{x} + k\sigma$
2	$1-1/2^2=0,75$
3	$1-1/3^2=0,89$
4	$1-1/4^2=0,94$

Almeno il 75% dei valori è compreso nell'intervallo $(\bar{x} - 2\sigma, \bar{x} + 2\sigma)$

Intervallo centrato intorno alla media, di ampiezza pari a 2σ

Applicazione del Teorema di Chebyshev

Riguardo alle spese sostenute per la protezione dell'ambiente nell'ultimo trimestre da un gruppo di imprese estrattive, si è osservato

$$\bar{x} = 2390€ \quad e \quad \sigma = 780€$$

Senza avere altre informazioni su come sono distribuite le spese, posso concludere che almeno il 75% delle imprese ha speso un ammontare compreso tra 830€ e 3950€ (si applica il Teorema con k=2)

Alternativamente concludo che non più del 25% delle imprese ha sostenuto spese ambientali inferiori a 830€ o superiori a 3950€

Omogeneità ed eterogeneità

Sono aspetti della variabilità di un carattere qualitativo

- Eterogeneità nulla (o massima omogeneità) → Il carattere assume un'unica modalità (tutte le unità del collettivo presentano quella modalità)
- Eterogeneità massima (o minima omogeneità) → Il carattere presenta tutte le modalità e a ciascuna di esse è associata la stessa frequenza

Eterogeneità

Eterogeneità nulla

(tutte le unità hanno la stessa modalità)

Mod.	Freq. rel.
a	1

Eterogeneità massima

(a ciascuna modalità è associata la stessa frequenza)

Mod.	Freq. rel.
a_1	1/k
a_2	1/k
…	
a_k	1/k

Indice di eterogeneità

Mod.	Freq.	Freq. rel.
a_1	n_1	f_1
a_2	n_2	f_2
a_j	n_j	f_j
a_K	n_K	f_K

Indice di eterogeneità di Gini

$$E_1 = 1 - \sum_{j=1}^{K} f_j^2$$

$$0 \le E_1 \le \frac{K-1}{K}$$

Indice relativo di eterogeneità di Gini

$$e_1 = \frac{E_1}{\frac{K-1}{K}} = E_1 \frac{K}{K-1}$$

$$0 \le e_1 \le 1$$

Eterogeneità dell'ubicazione dei punti vendita

Ubicazione del p.v.	F. ass. (n_j)	F. rel. (f_j)	f_j^2
Centro	4	0,45	0,20
Semicentro	2	0,22	0,05
Periferia	3	0,33	0,11
Totale	*9*	*1,00*	0,36

$$E_1 = 1 - \sum_{j=1}^{K} f_j^2 = 1 - 0{,}36 = 0{,}64$$

$$0 \le E_1 \le \frac{K-1}{K} = \frac{2}{3} = 0{,}67$$

$$e_1 = \frac{0{,}64}{0{,}67} = 0{,}96$$

Eterogeneità dell'ubicazione dei punti vendita

C'è un elevato grado di eterogeneità

La distribuzione osservata si avvicina a quella che si avrebbe nella situazione di massima eterogeneità

	Distr. osservata		Distr. con la max eterog.	
Ubicazione del p.v.	F. ass. (n_j)	F. rel. (f_j)	F. ass. (n_j)	F. rel. (f_j)
Centro	4	0,45	3	0,33
Semicentro	2	0,22	3	0,33
Periferia	3	0,33	3	0,33
Totale	*9*	*1,00*	*9*	*1,00*