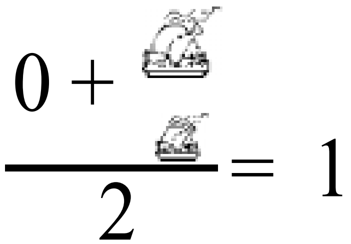
Insufficienza delle medie

Le medie forniscono informazioni sul centro della distribuzione

Le medie assolvono il loro compito di sintesi in modo più o meno efficiente in dipendenza del grado di variabilità del fenomeno

Dobbiamo spiegare:



Essenza della poesia di Trilussa

Si tuffa in un lago dalla profondità media di 10 centimetri

Esempio

La variabilità è più grande se maggiore è la gamma di modalità che può assumere e se minore è la diversificazione tra le frequenze.

Distribuzione "1"

x_i	n_i
1	1
3	23
5	1
	25

Distribuzione "2"

x_i	n_i
1	1
2	2
3	19
4	2
5	1
	25

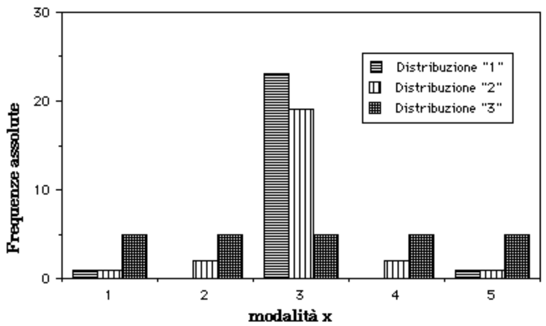
Distribuzione "3"

x_i	n_i
1	5
2	5
3	5
4	5
5	5
	25

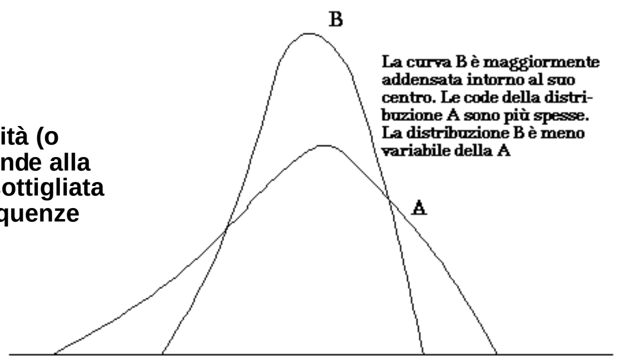
Le tre distribuzioni hanno in comune: moda, mediana e media aritmetica.

La distribuzione "1" ha minore variabilità perché, a parità di unità considerate, è minore il numero di modalità che presenta.

La distribuzione "3" ha maggiore variabilità perché, a parità del numero di modalità e del numero di unità, sono minori le differenze tra le frequenze.



Graficamente la variabilità (o dispersione) corrisponde alla forma più o meno assottigliata del poligono delle frequenze



Gli indici di variabilità

Gli indici di variabilità quantificano l'attitudine a variare della distribuzione.

Qualunque sia lo schema di costruzione, l'indice dovrà essere:

1) **nullo se e solo se le modalità sono tutte uguali.**
2) **crescente all'aumentare della differenziazione tra le modalità.**
3) **non negativo**

Consideriamo tre classi di indici di variabilità:

- **-Indici posizionali di variabilità**
- **-Indici basati sullo scostamento da un valore medio**
- **-Indici di variabilità relativa**

il campo di variazione (Range)

E' il più semplice degli indici di variabilità e si ottiene dalla differenza tra modalità più grande e la più piccola

$$) \; R = X_{\max} - X_{\min};$$

$$2) \; Max\left\{\left|X_{(i)} - X_{(j)}\right|; \; i, j = 1,2,\ldots n;\right\};$$

$$3) \sum_{i=2}^{n}\left[X_{(i)} - X_{(i-1)}\right]$$

Esempio.
Variazioni dell'indice di borsa MIB rispetto al giorno precedente.

2.3%	1.8%	-0.7%	0.2%	1.4%	2.2%	-1.9%	-0.5%	1.9%
1	2	3	4	5	6	7	8	9

Una volta ordinate le modalità si ottiene R=2.3 - (-1.9) = 5.2.

Proprietà del campo di variazione

R usa modalità le estreme perciò risente dei valori anomali.

R è utilizzabile per controllare processi stabili ed in cui un valore al di fuori del range implichi il verificarsi di una situazione atipica.

Il suo maggiore difetto è che resta invariato se alla distribuzione si aggiungono modalità già incluse nel range a prescindere dal grado di diversificazione che esse introducono nella distribuzione.

Un leggero miglioramento lo si ottiene dividendo R per n

$$R^* = \frac{R}{n}$$

dove R* indicherà lo "scarto medio" tra due valori consecutivi della distribuzione.

La differenza interquartilica

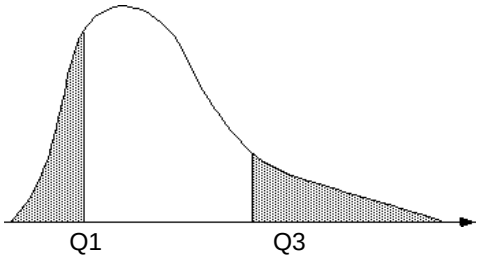
E' simile al campo di variazione solo che invece di includere il 100% delle modalità osservate, ne include la metà ed in particolare quelle centrali:

$$DI = Q_3 - Q_1$$

Semi DI $\longrightarrow$ $$SDI = \frac{Q_3 - Q_1}{2} = \frac{\left(Q_3 - M_e\right) + \left(M_e - Q_1\right)}{2}$$

DI rappresenta l'intervallo più piccolo che include almeno il 50% delle modalità centrali.

L'idea base di DI è che quanto più le modalità sono strette intorno alla mediana tanto più sarà piccolo l'intervallo che ne include metà e quindi minore è la variabilità.



Esempio

istituti di credito per classi di "prime rate" praticati alla clientela.

X_i		n_i	f_i	F_i
9.00	9.05	16	0.0808	0.0808
9.05	9.10	30	0.1515	0.2323
9.10	9.15	44	0.2222	0.4545
9.15	9.20	51	0.2576	0.7121
9.20	9.25	36	0.1818	0.8939
9.25	9.40	14	0.0707	0.9646
9.40	9.50	7	0.0354	1.0000
		198	1.0000	

$$Q_1 = \left[1 - \frac{0.25 - 0.2323}{0.4545 - 0.2323}\right]9.10 + \left[\frac{0.25 - 0.2323}{0.4545 - 0.2323}\right]9.15 = 9.104$$

$$Q_3 = \left[1 - \frac{0.75 - 0.7121}{0.8939 - 0.7121}\right]9.20 + \left[\frac{0.75 - 0.7121}{0.8939 - 0.7121}\right]9.25 = 9.210$$

$$DI = (9.210 - 9.104) = 0.106$$

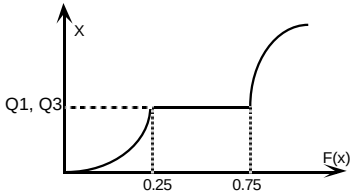
$$SDI = \frac{0.106}{2} = 0.053$$

Caatteristiche della DI

Ignora ogni eventuale aggiunta di coppie di modalità che lascino inalterato il primo ed il terzo quartile

La DI si usa per tenere sotto controllo le modalità intermedie senza troppo curarsi di quello che succede negli estremi.

La differenza interquartilica può essere zero anche in presenza di modalità diversificate

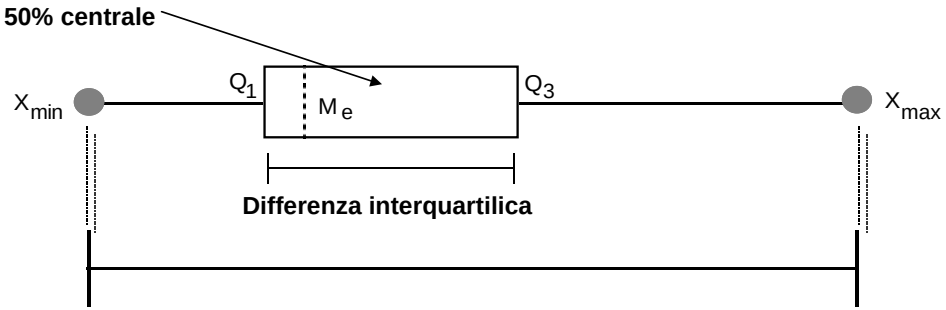


E' sufficiente che sia costante il 50% centrale della distribuzione.

BOXPLOT

E' la sintesi numerico-grafica di una distribuzione. Si usano 5 numeri

$$X_{\min},\ Q_1,\ M_e,\ Q_2,\ X_{\max}$$

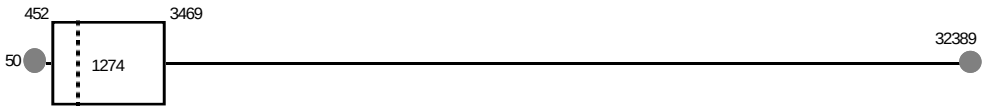


Campo di variazione

ESEMPIO

Extracomunitari iscritti alle liste di collocamento per Paese di origine

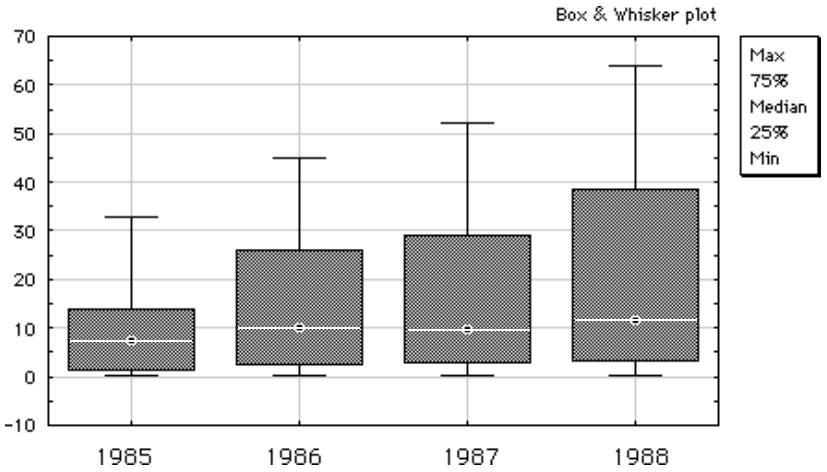
309	425	50	235	3606	4445	644	1696	2142	2863	1032	9596	429	32389	1409
2141	3440	953	11565	2503	1138	11704	452	10142	3469	687	982	71	758	161



il valore anomalo allunga la distribuzione

Confronto di più distribuzioni

Patrimonio netto dei fondi comuni mobiliari



Sebbene la variabilità aumenti sistematicamente, la mediana delle distribuzioni si mantiene stabile

Boxplot e valori anomali

Per individuare i valori atipici si utlizzano le seguenti barriere:

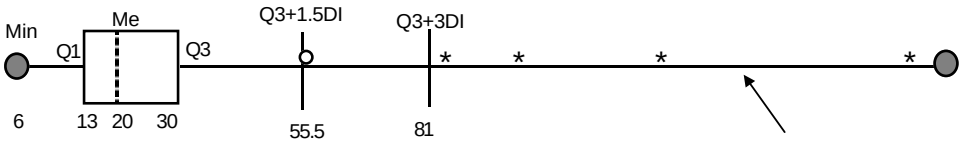
$$Valori\ di\ allerta: X_{\min} - 1.5DI,\ X_{\max} + 1.5DI$$

$$Valori\ anomali: X_{\min} - 3DI,\ X_{\max} + 3DI$$

Assenze dal lavoro:

41	15	6	21	10	21	9	7	44	8	16	7	28	21	14
16	15	7	8	6	22	15	29	36	175	126	27	34	43	41
30	13	8	15	15	20	56	6	21	98	29	14	90	14	28

Q_1=13, M_e= 20, Q_3=30, DI=17;
Soglie di allarme: Q_3+1.5DI=55.5;
Soglia dei valori remoti: Q_3+3*DI=81.



Casi patologici da escludere dalla rilevazione

Scostamenti tra due serie

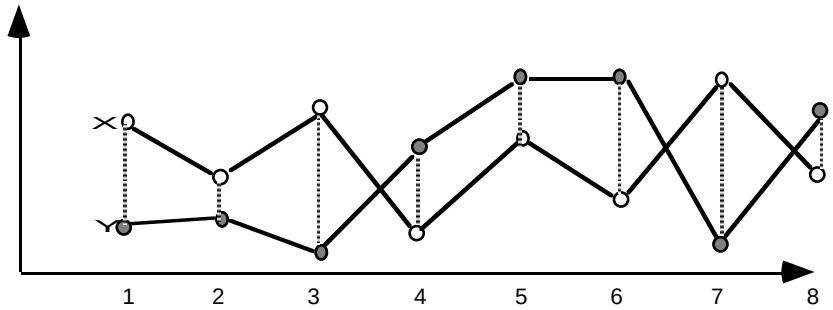
L'idea è di misurare lo scostamento complessivo (devianza) tra le due serie e per far questo possiamo adoperare una

$$S^{\alpha} = \left[\sum_{i=1}^{n} |X_i - Y_i|^{\alpha}\right]^{\frac{1}{\alpha}}$$

La distanza complessiva percorsa se ogni X_i si portasse sulla corrispondente Y_i.

Metrika di Minkowsky.

Di solito α=1 o α=2

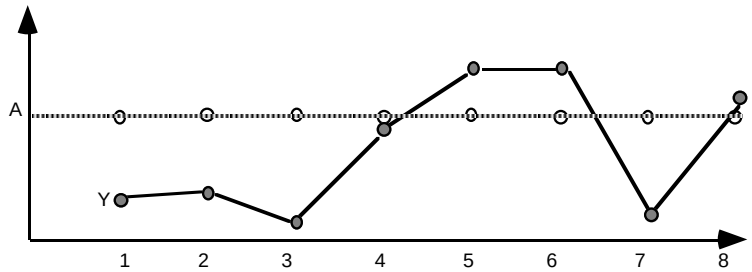


Perché confrontare la "Y" proprio con la "X" ?

Come misurare la variabilità della "Y" se varia anche la "X"?

Scostamenti da un valore medio

Si supponga che la "X" sia costante



Ogni scelta del valore di A -di solito una media- e di α (di solito "1 o "2") comporta la definizione di un indice di variabilità

$$S(A,\alpha) = \left[\sum_{i=1}^{n} |Y_i - A|^{\alpha} f_i\right]^{\frac{1}{\alpha}}$$

Le combinazioni più usuali sono:

α=1 - Valore assoluto e Mediana **α=2 - Quadrato e Media aritmetica**

Scarto quadratico medio

E' la misura più nota di variabilità (detta anche DEVIAZIONE STANDARD)

$$\sigma = \sqrt{\sum_{i=1}^{k} (X_i - \mu)^2 f_i}; \quad \mu = \sum_{i=1}^{k} X_i f_i$$

Paese	Tariffa	Paese	Tariffa	Paese	Tariffa
Gran Bretagna	0.64	Finlandia	1.84	Italia	1.80
Spagna	1.51	Danimarca	0.98	Belgio	2.78
Francia	0.71	Olanda	2.00	Austria	7.61
Germania	1.00	Svezia	1.68		

$\mu = \frac{\sum_{i=1}^{n} X_i}{n} = 2.0501$; $\sigma = \sqrt{\frac{\sum_{i=1}^{n} (X_i - \mu)^2}{n}} = 6.1556$

In questo caso le formule di calcolo si semplificano.

$$Varianza = \sigma^2 = \sum_{i=1}^{k} (X_i - \mu)^2 f_i = \sum_{i=1}^{k} X_i^2 f_i - \mu^2$$

Scarto quadratico medio (classi)

Un parco macchine è stato suddiviso in classi di percorrenza

Percorrenza		Auto	c_i	f_i	$f_i c_i$	$(c_i-\mu)^2 f_i$
10	19.9	6	14.95	0.0462	0.6900	10.6367
20	24.9	16	22.45	0.1231	2.7631	7.2613
25	29.9	48	27.45	0.3692	10.1354	2.6539
30	34.9	33	32.45	0.2538	8.2373	1.3651
35	39.9	18	37.45	0.1385	5.1854	7.4171
40	49.9	9	44.95	0.0692	3.1119	15.2033
		130		1.0000	30.1231	44.5374

σ=6.6736

L'uso dei valori centrali implica che si sta approssimando la variabilità complessiva poiché si fa l'ipotesi -quasi sempre non vera- che all'interno delle classi le modalità si addensino intorno al loro centro

Scostamento semplice dalla mediana

$$S_{Me} = \sum_{i=1}^{k} |X_i - M_e| f_i$$

Spesa in pubblicità per l'olio di oliva in alcuni Paesi della Comnità eureopea.

Paese	Spesa	Paese	Spesa	Paese	Spesa
Italia	512	R.U.	240	Olanda	128
Francia	240	Benelux	128	Irlanda	80
Grecia	240	Spagna	640	Danim.	80
Germ.	384	Portog.	256		

$$M_e = X_{(6)} = 240;\ S_{Me} = \frac{\sum_{i=1}^{n} |X_i - M_e|}{n} = 11.37$$

un campione di frutti di una pianta è stato classificato per numero di semi:

Semi	Frutti	f_i	F_i	$\lvert X_i - M_e \rvert$
0	1	0.0070	0.0070	0.0420
1	4	0.0280	0.0350	0.1399
2	6	0.0420	0.0769	0.1678
3	9	0.0629	0.1399	0.1888
4	16	0.1119	0.2517	0.2238
5	31	0.2168	0.4685	0.2168
6	76	0.5315	1.0000	0.0000
	143	1.0000		0.9790

In questo caso l'applicazione della formula non presenta alcuna difficoltà. L'interpretazione è però ardua data la forma a "J" della distribuzione: in media le modalità differiscono di un seme dalla mediana.

La deviazione media

$$S_\mu = \sum_{i=1}^{k} |X_i - \mu| f_i;$$

Reddito medio unitario (in milioni) per varie categorie di lavoratori dipendenti.

Qualifica	R.M.U.	Qualifica	R.M.U.	Qualifica	R.M.U.
Operai	17.767	Doc. Univ.	61.787	Magistrati	78.754
Impiegai	25.175	Ins. Scuola	26.539	Parlamentari	50.818
Funzionari	44.851	Sottuf.	25.753	Religiosi	18.053
Dirigenti	86.828	Ufficiali	32.115		

$$\mu = 42.59;\ S_\mu = \frac{\sum_{i=1}^{n} |X_i - \mu|}{n} = 20.02$$

Squadre di calcio di serie A e B. Numero di elementi nella rosa

Calciatori	Squadre	f_i	$X_i f_i$	$\lvert X_i - \mu \rvert f_i$
18	4	0.1053	1.8947	0.3158
19	3	0.0789	1.5000	0.1579
20	5	0.1316	2.6316	0.1316
21	8	0.2105	4.4211	0.0000
22	14	0.3684	8.1053	0.3684
23	3	0.0789	1.8158	0.1579
24	1	0.0263	0.6316	0.0789
	38	1.0000	21.0000	1.2105

In media le "rose" differiscono dalla media "21 giocatori" per poco più di un giocatore.

Effetti dell'unità di misura

Le medie e gli indici di variabilità sono espressi nelle medesime unità di misura del fenomeno cui si riferiscono: se si modifica la scala cambiano anche gli indici descrittivi.

Cambiamenti di scala moltiplicativi

Supponiamo che le modalità {X1, X2, ..., Xk} siano tutte moltiplicate per la costante non nulla "c":

$$Y_i = cX_i \ (i=1,2,...,k)$$

$$\mu(Y) = \sum_{i=1}^{k} Y_i f_i = \sum_{i=1}^{k} cX_i f_i = c\sum_{i=1}^{k} X_i f_i = c\mu(X)$$

moda e mediana sono pure riproduttive

$$V(Y) = \sum_{i=1}^{k} [Y_i - \mu(Y)]^2 f_i = \sum_{i=1}^{k} [cX_i - c\mu(X)]^2 f_i = c^2 \sum_{i=1}^{k} [X_i - \mu(X)]^2 f_i = c^2 V(X)$$

da cui $\sigma(Y)=|c|\sigma(X)$. Quindi media e scarto sono moltiplicati per "c"

Cambiamenti di scala additivi

Vediamo ora l'effetto di un cambiamento additivo cioè della somma di una costante "c" a tutte le modalità:

$$Y_i = c + X_i \quad (i=1,2,...,k)$$

Lo stesso accade a moda e mediana

$$\mu(Y) = \sum_{i=1}^{k} Y_i f_i = \sum_{i=1}^{k} [c+X_i] f_i = \sum_{i=1}^{k} c f_i + \sum_{i=1}^{k} X_i f_i = c + \mu(X)$$

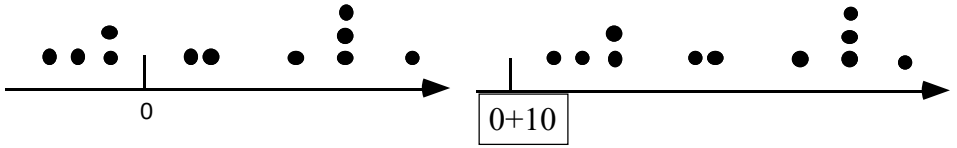
$$V(Y) = \sum_{i=1}^{k} [Y_i - \mu(Y)]^2 f_i = \sum_{i=1}^{k} [[c+X_i] - [c+\mu(X)]]^2 f_i$$

$$= \sum_{i=1}^{k} [c+X_i-c-\mu(X)]^2 f_i = \sum_{i=1}^{k} [X_i - \mu(X)]^2 f_i = V(X)$$

Quindi la media si sposta di un ammontare "c", ma la variabilità rimane inalterata.

Significato

Sia data la serie X : {1, 3, 5, 7} che ha media 4 e varianza 5. Se sommiamo 10 a tutte le modalità avremo:



L'avanzamento dell'origine ha incrementato dello stesso ammontare ciascuna modalità, ma non ha alterato le interdistanze.

Ci si aspetta quindi che per trasformazioni additive gli indici di variabilità rimangano invariati.

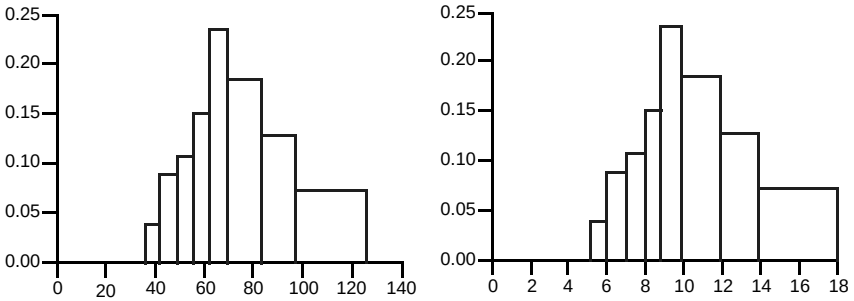
$$R(Y) = Y_{\max} - Y_{\min} = a + bX_{\max} - a - bX_{\min} = b(X_{\max} - X_{\min}) = bR(X)$$

$$DI(Y) = Q_3(Y) - Q_1(Y) = a + Q_3(X) - a - bQ_1(Y) = bDI(X)$$

Esempio

Distribuzione del tempo (in giorni) necessari a completare le pratiche in una Camera di commercio si è costruito l'istogramma delle frequenze; successivamente si sono trasformati i giorni in settimane.

		Pratiche
36	42	12
43	49	28
50	56	34
57	62	48
63	69	75
70	83	59
84	97	41
98	126	23
		320



Le trasformazioni lineari influenzano la collocazione e la scala delle ascisse, ma non alterano la forma dell'istogramma (o del poligono delle frequenze).

Trasformazione dei dati

Le trasformazioni tendono a uniformare il campo di variazione delle variabili

- **Per le rappresentazioni grafiche**
- **Per il confronto di variabili su scale differenti**
- **Per omogeneizzare misure ottenute in condizioni di variabilità ineguale**

Lo schema adottato è:

$$Y_i = \left[\frac{X_i - \text{indice di centralità}}{\text{indice di variabilità}}\right] * \text{costante}$$

che è una trasformazione lineare del tipo: $y = a + bx$

Trasformazione unitaria

Ottenuta sottraendo a tutte le $\{X_i\}$ la modalità più piccola e dividendo poi il risultato per il campo di variazione.

L'effetto è che ora le $\{Y_i\}$ assumono valori nell'intervallo [0,100] con indubbi vantaggi per la costruzione dei grafici.

$$U_i = \left(\frac{X_i - X_{\min}}{X_{max} - X_{\min}} \right) * 100; \quad i = 1,2,\ldots,n$$

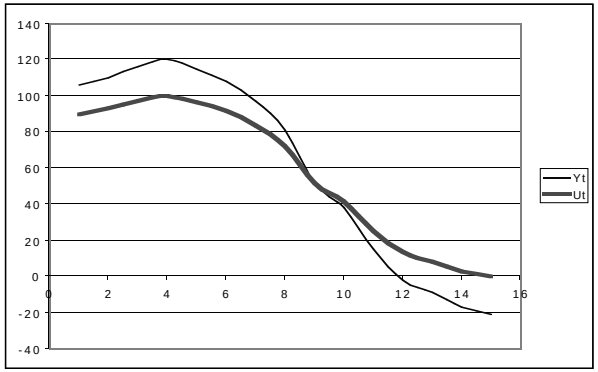
Esempio.
A partire dalla serie X: {1, 3, 5, 7} si ottiene la seguente serie:

$$Y = \left\{0, \left[\frac{1-1}{7-1}\right]*100, \left[\frac{3-1}{7-1}\right]*100, \left[\frac{5-1}{7-1}\right]*100, \left[\frac{7-1}{7-1}\right]*100\right\}$$

$$= \{0,\ 33.3333,\ 66.6667,\ 100\}$$

Esempio

t	Yt	Ut
1	106	90.07
2	110	92.91
3	116	97.16
4	120	100.00
5	115	96.45
6	108	91.49
7	97	83.69
8	81	72.34
9	52	51.77
10	38	41.84
11	15	25.53
12	-2	13.48
13	-9	8.51
14	-17	2.84
15	-21	0.00



L'andmento della serie non è alterato dalla trasformazione unitaria. La linea più spessa Rimane però limitata tra zero e 100.

Atenzione! Se si applica la trasformazione unitaria a due serie, queste avranno comunque in comune due valori (0 e 100) anche se prima ero totalmente diverse.

Variabili standardizzate

E' la trasformazione più nota perché, qualunque sia la variabile di partenza, le cosiddette unità standard hanno media aritmetica zero e deviazione standard uno.

$$Z_i = \left[\frac{X_i - \mu(X)}{\sigma(X)}\right] = -\frac{\mu(X)}{\sigma(X)} + \frac{1}{\sigma(X)} X_i \qquad (i=1,2,\ldots,k)$$

Ne consegue che:

$$\mu(Z) = -\frac{\mu(X)}{\sigma(X)} + \frac{1}{\sigma(X)}\mu(X) = 0; \qquad \sigma(Z) = \frac{1}{\sigma(X)}\sigma(X) = 1$$

Z=0.7 significa che il valore originario è superiore alla media e la supera di un ammontare pari al 70% dello scarto quadratico medio.

il riferimento all'unità di misura è del tutto scomparso.

Esempio

Produzione olearia 91-92.

Regione	Resa	Z_resa	Regione	Resa	Z_resa	Regione	Resa	Z_resa
Puglia	19.40	0.49	Abruzzo	16.60	-1.03	Molise	16.20	-1.25
Calabria	19.90	0.76	Sardegna	19.20	0.38	Veneto	16.20	-1.25
Sicilia	20.00	0.81	Basilicata	20.70	1.19	Lombardia	15.90	-1.41
Campania	18.40	-0.05	Liguria	22.30	2.06	Emilia Rom.	15.50	-1.63
Lazio	18.50	0.00	Marche	18.20	-0.16	Trentino A.A.	19.30	0.43
Toscana	18.20	-0.16	Umbria	20.00	0.81			

La resa in Emilia Romagna è 15.5 quintali di olio per 100 quintali di olive. Questo, a livello di confronto regionale, non è molto informativo.

Il punteggio standard delle altre regioni è Z=(15.5-18.5)/1.84=-1.63 è quindi inferiore alla media.

Non solo, ma è inferiore alla media per il 63% della deviazione standard

Indici di variabilità relativa

Sono indici che permettono il confronto della variabilità tra variabili espresse

- **Con ordini grandezza ineguali**
- **In unità di misura eterogenee**
- **Per campi di variazioni diversi**

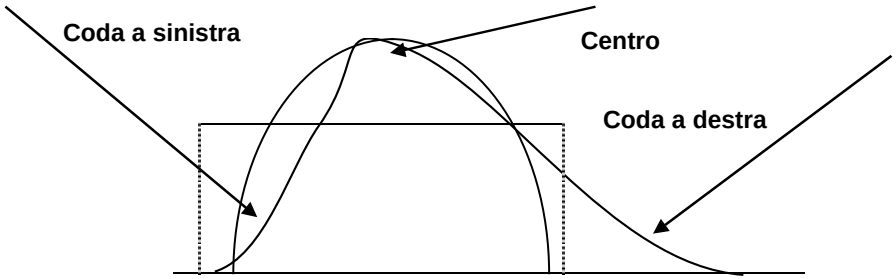
Gli indici di variabilità relativa mirano ad eliminare i riferimenti dimensionali presente negli indici di variabilità con la definizione di "numeri puri", cioè

$$\text{misura di variabilità relativa} = \frac{\text{misura di variabilità assoluta}}{\text{media}}$$

La media al denominatore deve essere positiva o in valore assoluto.

Esigenze di identificazione

Esistono distribuzioni che hanno la stessa tendenza centrale e la stessa dispersione, ma dissimili per altri aspetti importanti?



L'uso di soli indici di centralità e variabilità rende equivalenti distribuzioni molto diverse al centro e nelle code.

E' evidente che occorrono altre informazioni per identificare la distribuzione su questi due aspetti

Coefficiente di variazione

E' l'indice di variabilità relativa più noto.

$$CV = \frac{\sigma}{|\mu|} = \sqrt{\sum_{i=1}^{k}\left(\frac{X_i - \mu_x}{\mu_x}\right)^2}$$

Esprime la variabilità in termini di unità della media

Consideriamo la trasformazione lineare $Y_i = a + bX_i$

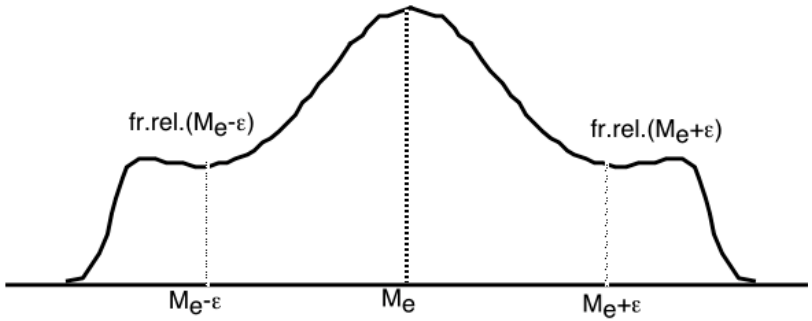
$$CV^2(Y) = \sum_{i=1}^{k}\left(\frac{Y_i - \mu_y}{\mu_y}\right)^2 f_i = \sum_{i=1}^{k}\left(\frac{a + bX_i - a - b\mu_x}{a + b\mu_x}\right)^2 f_i = b\sum_{i=1}^{k}\left(\frac{X_i - \mu_x}{a + b\mu_x}\right)^2 f_i$$

La parte moltiplicativa non ha alcuna influenza su CV, la parte additiva altera la misura della variabilità relativa anche se nulla cambia nella variabilità assoluta.

La simmetria statistica

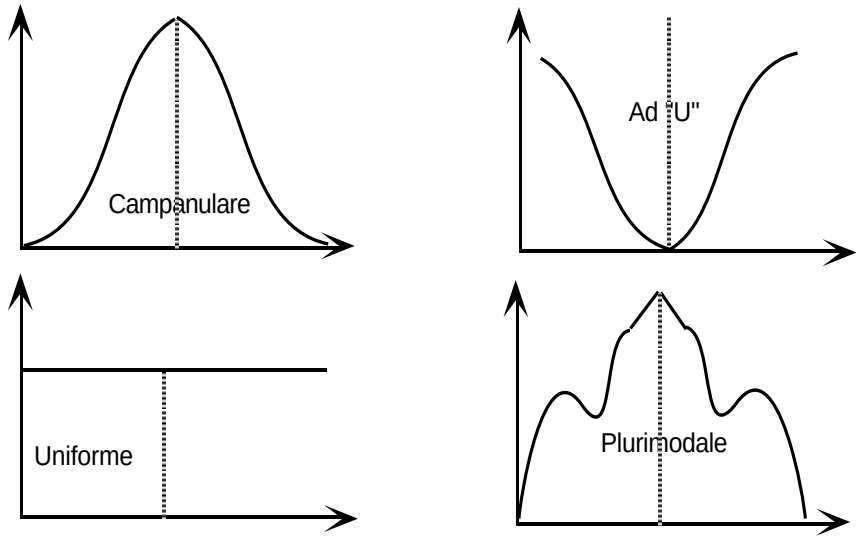
Una distribuzione di frequenza è simmetrica intorno ad un polo se genera un istogramma o un poligono di frequenza simmetrico.

$$fr.rel.(M_e - \varepsilon) = fr.rel.(M_e + \varepsilon), \quad \forall \varepsilon > 0$$



Se si piega la funzione di densità (o l'istogramma) lungo l'asse formato dalla mediana, uno dei due lati si sovrapporrà esattamente all'altro.

Riconoscibilità della simmetria



Definizione per serie

Si supponga che le modalità siano disposte in ordine crescente di grandezza. Perché la distribuzione sia simmetrica è necessario che

$$\frac{X_{(i)}+X_{(n-i+1)}}{2}=M_e$$

$$ovvero\ \left[X_{(n-i+1)}-M_e\right]=\left[M_e-X_{(i)}\right]\ \ per\ \ i=1,2,\ldots,\left[n/2\right]$$

E' simmetrica la distribuzione per cui gli scarti negativi dalla mediana sono uguali in numero ed in grandezza (tranne il segno) a quelli positivi.

La distribuzione: {2, 3, 4, 5, 6, 7} è simmetrica. Infatti:

$$\frac{2+7}{2}=4.5;\ \frac{3+6}{2}=4.5;\ \frac{4+5}{2}=4.5$$

Grafico di Tukey

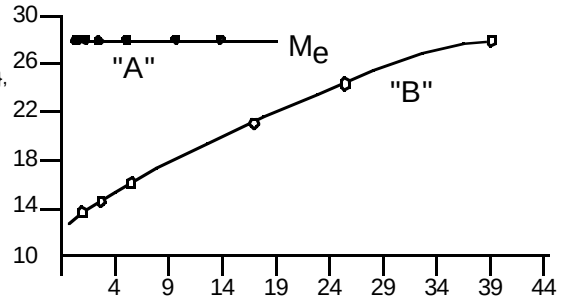
Se si rappresentano graficamente i punti di coordinate

$$\frac{\left(X_{(n-i+1)}-M_e\right)^2+\left(M_e-X_{(i)}\right)^2}{4M_e};\quad \frac{X_{(n-i+1)}+X_{(i)}}{2};\quad i=1,2,\ldots,\left[n/2\right]$$

Se la distribuzione è simmetrica correranno paralleli all'asse delle ascisse lungo la mediana

Saranno crescenti se c'è uno sbilanciamento sia verso i valori alti che verso i valori bassi

A: {0, 5, 10, 15, 20, 25, 30, 35, 40, 45, 50, 55},
B={0, 1, 2, 4, 6, 10, 15, 21, 28, 36, 45, 55}



Per la A i punti sono allineati lungo la retta $y=M_e=27.5$;

per la B seguono una curva che evidenzia lo sbilanciamento verso i valori grandi.

Valutazione

La mancanza di simmetria (presenza di asimmetria) è frutto di scompensi tra scarti per modalità ricadenti su di un lato della mediana rispetto a scarti per modalità ricadenti sul lato opposto.

La asimmetria è tanto più spiccata quanto maggiori sono le differenze tra scarti negativi e scarti positivi dalla mediana.

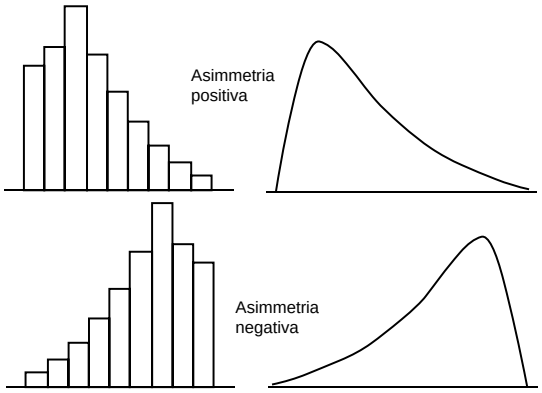
In genere, gli squilibri vicino al centro non sono considerati preoccupanti e sono di fatto ignorati.

Molto più rilevanti sono quelli concernenti il diverso comportamento nelle due code

Asimmetria con segno

Si parlerà di asimmetria positiva se gli scarti dovuti a valori minori della mediana hanno più peso (grandezza e frequenza) degli scarti per valori superiori alla mediana.

Si parlerà di asimmetria negativa nel caso opposto.



Gli scostamenti possono verificarsi al centro della distribuzione: Moda e mediana non coincidono

e/o nelle code per la presenza di valori remoti solo su di un lato della distribuzione.

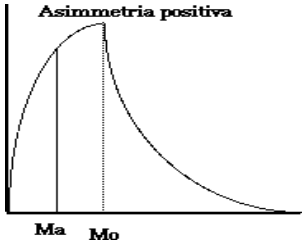
Considerazioni

Questo indice si basa sul fatto che nelle distribuzioni unimodali simmetriche si ha

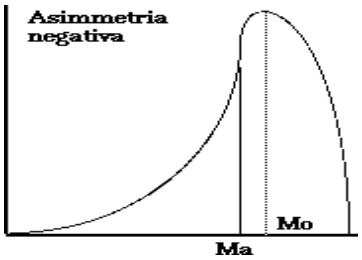
$$M_a = M_o = M_e$$

(la presenza di tale uguaglianza è un indizio di simmetria, ma non una certezza).

Nelle distribuzioni asimmetriche positive la media aritmetica è maggiore della moda a causa della "coda" allungata (effetto di "skewness") verso i valori grandi.



Per ragioni analoghe la media aritmetica è minore della moda per distribuzioni con asimmetria negativa.



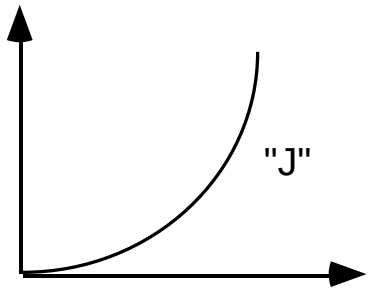
Significato della asimmetria

L'asimmetria aiuta ad interpretare il fenomeno.

La negativa può essere l'esito di una "accelerazione" del fenomeno che esaurisce la sua spinta dopo un livello molto elevato.

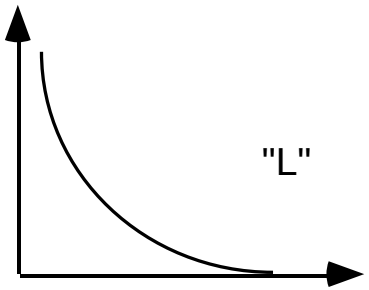
Esito di un test con prove facili.

Vedodi/vedove per età



Distribuzione dei redditi alti.

Celibi/nubili per età



L'asimmetria positiva può derivare dalla presenza di un "freno" che si attiva dopo un livello piuttosto basso.

Misura della asimmetria

La misura della asimmetria mira a quantificare lo scostamento da una situazione di simmetria.

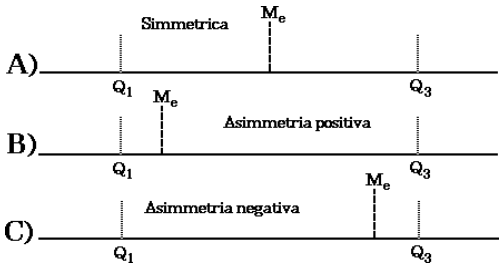
I requisiti minimi per un indice di asimmetria $\alpha(X)$ sono:

- **1) $\alpha(X)=0$ se la distribuzione è simmetrica;**
- **2) $\alpha(X)$ aumenta all'aumentare dello scostamento dalla situazione di simmetria;**
- **3) Nel caso di distribuzioni unimodali si deve avere $\alpha(X)<0$ se c'è un allungamento verso i valori piccoli e $\alpha(X)>0$ se l'allungamento è verso i valori grandi.**

L'indice di Yule-Bowley

E' basato sul confronto tra quartili e si concentra sugli sbilanciamenti che si verificano fra le modalità comprese nel 50% centrale della distribuzione:

$$YB = \frac{(Q3 - Me) - (Me - Q1)}{(Q3 - Me) + (Me - Q1)} = \frac{Q3 + Q1 - 2Me}{Q3 - Q1}$$



YB = 0

YB > 0 Dominano i valori medio bassi

YB < 0 Dominano i valori medio alti

Proprietà dell'indice di Yule-Bowley

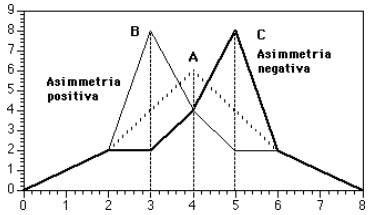
L'indice di YB è relativo -1≤ YB ≤ 1 ed è anche standardizzato

Il massimo negativo è ottenuto per le distribuzione asimmetriche a destra (asimmetria negativa) in cui almeno la metà del primo 50% delle unità ha la modalità pari alla mediana;

Il massimo positivo è raggiunto da distribuzioni in cui la mediana è portata almeno dalla prima metà dell'ultimo 50%.

	Frequenze assolute			Frequenze rel. Cum.		
X_i	A	B	C	A	B	C
1	1	1	1	0.05	0.05	0.05
2	2	2	2	0.15	0.45	0.15
3	4	8	2	0.30	0.65	0.25
4	6	4	4	0.65	0.80	0.40
5	4	2	8	0.85	0.90	0.60
6	2	2	2	0.95	0.95	0.95
7	1	1	1	1.00	1.00	1.00
	20	20	20			

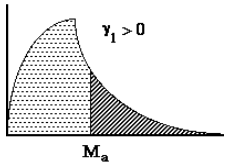
$YB_a = \frac{3+5-2*4}{5-3} = 0$ $YB_b = \frac{5+3-2*3}{5-3} = 1$ $YB_c = \frac{3+5-2*5}{5-3} = -1$



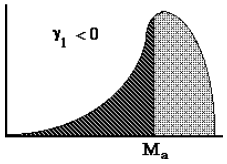
I quartili sono ottenuti con dei processi di approssimazione per cui l'accuratezza dell'indice di Yule-Bowley non può essere superiore a quella adoperata nel calcolo dei quartili.

Indice di Fisher (Skewness)

$$\gamma_1 = \sum_{i=1}^{k} \left(\frac{X_i - \mu}{\sigma} \right)^3 f_i$$



Nelle distribuzioni con asimmetria positiva ovvero con coda distesa verso i valori grandi gli scarti positivi saranno, in modulo , più grandi di quelli negativi per cui γ1>0



Nelle distribuzioni con asimmetria negativa gli scarti negativi (relativi cioè a modalità inferiori alla media) prevarranno su quelli positivi (sono più distanti) e si ha γ1 <0.

γ_1 è invariante rispetto a trasformazioni lineari, ma non varia in un intervallo definito.

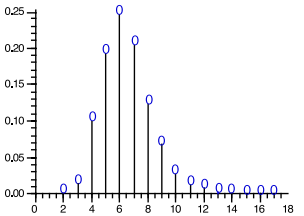
E' possibile stabilire se una distribuzione è più asimmetrica di un'altra, ma non che una distribuzione sia molto o poco asimmetrica.

Trattandosi di una media può succedere che l'indice si annulli anche in presenza di sostanziale asimmetria.

Esempio

Distribuzione per numero di lettere nel cognome di un campione di residenti nel Regno Unito.nti nel Regno Unito.

Lettere	Cognomi	$(Z_i)^3$
2	3	-0.017
3	40	-0.106
4	279	-0.267
5	532	-0.107
6	687	-0.005
7	563	0.005
8	342	0.078
9	189	0.194
10	77	0.215
11	35	0.207
12	20	0.215
13	6	0.106
14	4	0.108
15	2	0.079
16	1	0.055
17	1	0.074
	2781	0.834



L'indice positivo evidenzia la distensione verso i cognomi più lunghi.

L'indice positivo evidenzia la distensione verso i cognomi più lunghi.

I valori remoti hanno su quest'indice effetti esasperati: sia attraverso la media aritmetica che attraverso gli scarti al cubo.

Tuttavia, la presenza dello scarto quadratico medio attenua l'esplosione dei valori dell'indice.