

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:

$$\frac{0 + \text{[immagine di un lago]} + \text{[immagine di un lago]}}{2} = 1$$

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

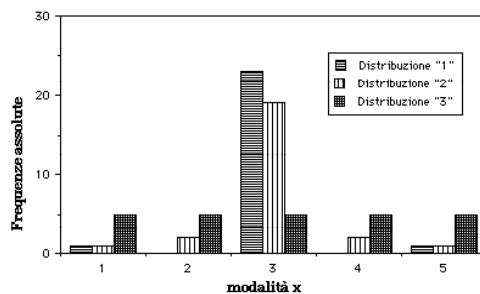
Distribuzione "3"

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.

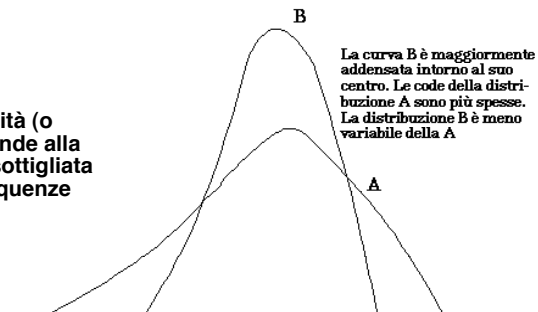
La distribuzione "1" ha minore variabilità perché, a parità di unità considerate, è minore il numero di modalità che presenta.

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.



Spiegazione grafica

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

Studieremo tre classi di indici di variabilità:



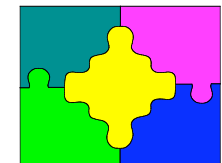
-Indici posizionali di variabilità



-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

- 1) $R = X_{\max} - X_{\min}$;
- 2) $\text{Max}\{X_{(i)} - X_{(j)}\}; i, j = 1, 2, \dots, n\};$
- 3) $\sum_{i=2}^n [X_{(i)} - X_{(i-1)}]$

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

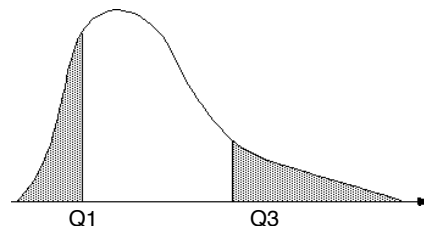
E' simile al campo di variazione solo che invece di includere il 100% delle modalità osservate, ne include la metà ed in particolare quelle centrali:

$$DI = Q_3 - Q_1$$

$$\text{Semi DI} \rightarrow SDI = \frac{Q_3 - Q_1}{2} = \frac{(Q_3 - M_e) + (M_e - Q_1)}{2}$$

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

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

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

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

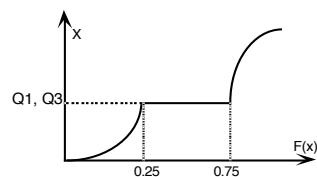
$$SDI = \frac{0.106}{2} = 0.053$$

Caratteristiche 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.

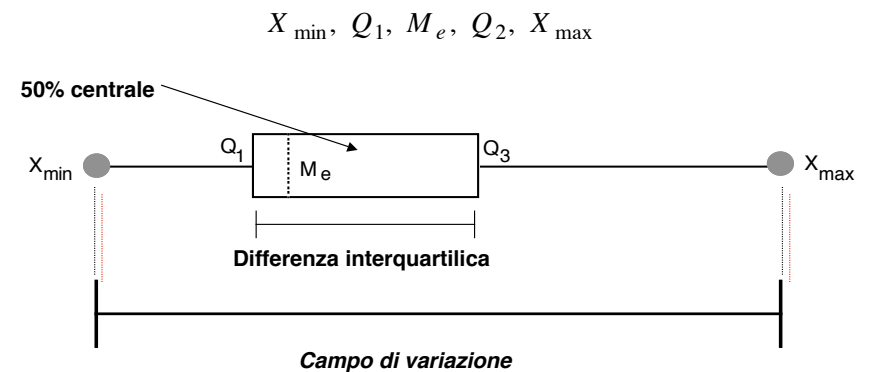
La differenza interquartilica può essere zero anche in presenza di modalità diversificate



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

Box Plot

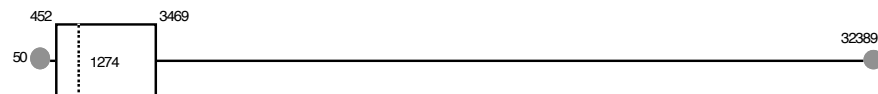
E' la sintesi numerico-grafica di una distribuzione. Si usano 5 numeri



Esempio

Extracomunitari iscritti alle liste di collocamento per Paese di origine

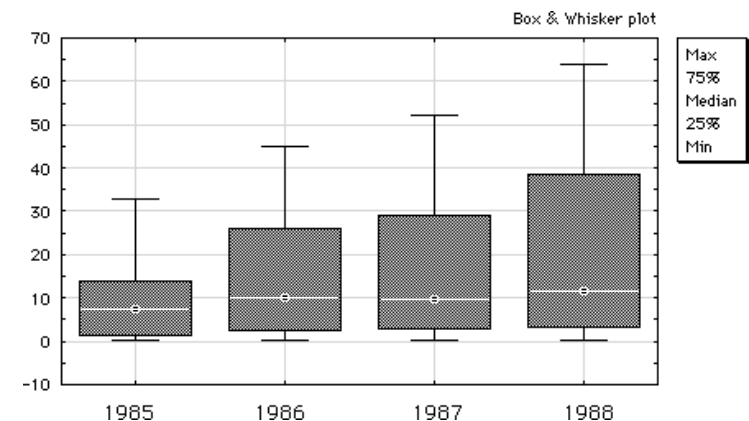
309	425	50	235	3606	4445	644	1696	2142	2863	1032	9596	429	32389	1409
2141	3440	953	11565	2503	1138	11704	452	10142	3469	687	982	71	758	161



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 utilizzano le seguenti barriere:

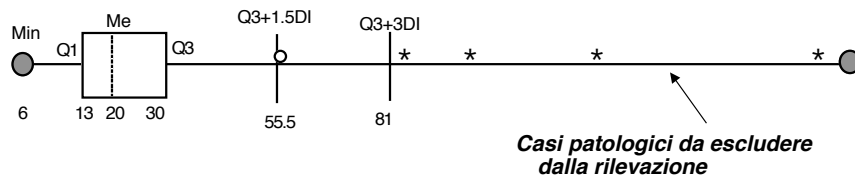
Valori di allerta $X_{\min} - 1.5DI$, $X_{\max} + 1.5DI$

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

$Q_1=13$, $M_0=20$, $Q_3=30$, $DI=17$;
Soglie di allarme: $Q_3+1.5DI=55.5$;
Soglia dei valori remoti: $Q_3+3*DI=81$.



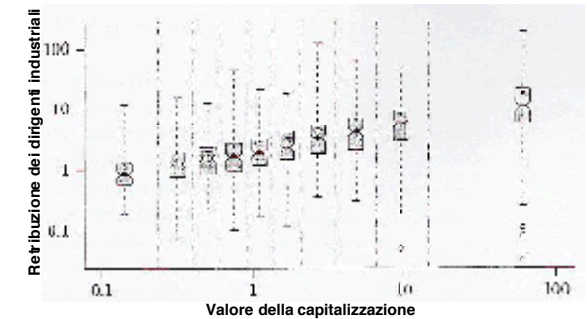
Esempio

Imprese raggruppate per livelli di capitalizzazione. Distribuzione degli stipendi dei dirigenti industriali

+ = media geometrica
* = media aritmetica

Le Linee verticali che partono dalle ascisse dividono la capitalizzazione in decili

Valori in scala logaritmica



E' evidente la tendenza all'aumento degli stipendi all'aumentare del livello di capitalizzazione.

E' anche evidente una tendenza all'aumento della variabilità come testimonia l'ampliamento della differenza interquartilica

Scostamenti tra due serie

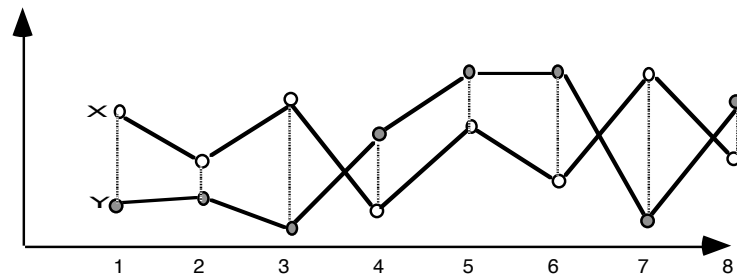
L'idea è di misurare lo scostamento complessivo (devianza) tra le due serie e per far questo possiamo adoperare una

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

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

Metrika di Minkowsky.

Di solito $\alpha=1$ o $\alpha=2$

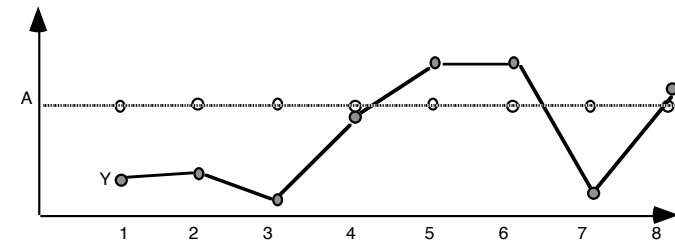


Perché confrontare la "Y" proprio con la "X" ?

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:

$\alpha=1$ - Valore assoluto e Mediana

$\alpha=2$ - Quadrato e Media aritmetica

Scarto quadratico medio

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

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

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		

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

In questo caso le formule di calcolo si semplificano.

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

Scarto quadratico medio (classi)

Un parco macchine è stato suddiviso in classi di percorrenza

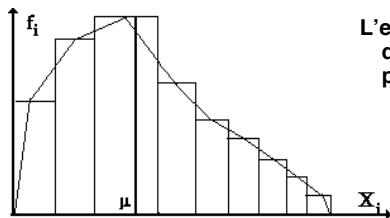
Percorrenza	Auto	c_i	f_i	$f_i \cdot c_i$	$(c_i - \mu)^2 \cdot f_i$	
10	19.9	6	14.95	0.0462	0.6900	10.6367
20	24.9	16	22.45	0.1231	2.7631	7.2613
25	29.9	48	27.45	0.3692	10.1354	2.6539
30	34.9	33	32.45	0.2538	8.2373	1.3651
35	39.9	18	37.45	0.1385	5.1854	7.4171
40	49.9	9	44.95	0.0692	3.1119	15.2033
		130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

La correzione di Sheppard

L'uso dei valori centrali porta ad una valutazione approssimata dell'ammontare effettivo della differenziazione della variabile.



L'entità ed il segno della approssimazione dipendono dalla forma assunta dalla poligonale di frequenza.

Se la variabile è continua, le classi hanno uguale ampiezza, l'istogramma è di tipo campanulare, è disponibile una formula più corretta:

$$\sigma = \sqrt{\sum_{i=1}^k (X_i - \mu)^2 f_i - \frac{d^2}{12}};$$

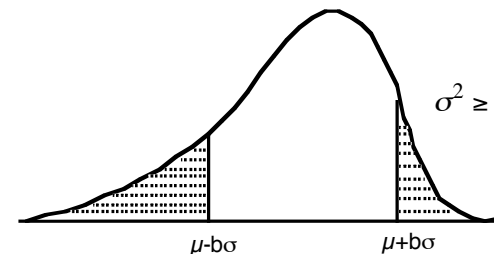
La disuguaglianza di Tchebycheff

L'idea è di approssimare, in base a "μ" e "σ" la frequenza associabile ad intervalli del tipo: |X-μ| ≥ bσ

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

$$\sum_{x \leq \mu - b\sigma} (X_i - \mu)^2 f_i + \sum_{x \geq \mu + b\sigma} (X_i - \mu)^2 f_i + \sum_{\mu - b\sigma < x < \mu + b\sigma} (X_i - \mu)^2 f_i$$

E' positivo!



$$\sigma^2 \geq \sum_{x \leq \mu - b\sigma} (X_i - \mu)^2 f_i + \sum_{x \geq \mu + b\sigma} (X_i - \mu)^2 f_i$$

La disuguaglianza di Tchebycheff/2

Nelle due regioni estreme si ha: $|X - \mu| \geq b\sigma \Rightarrow (X - \mu)^2 \geq (b\sigma)^2$

Quindi:
$$\sigma^2 \geq \sum_{x \leq \mu - b\sigma} (b\sigma)^2 f_i + \sum_{x \geq \mu + b\sigma} (b\sigma)^2 f_i = \sigma^2 b^2 \left[\sum_{x \leq \mu - b\sigma} f_i + \sum_{x \geq \mu + b\sigma} f_i \right]$$

ovvero:
$$\sigma^2 \geq \sigma^2 b^2 \left[\sum_{x \leq \mu - b\sigma} f_i + \sum_{x \geq \mu + b\sigma} f_i \right] \Rightarrow \left[\sum_{x \leq \mu - b\sigma} f_i + \sum_{x \geq \mu + b\sigma} f_i \right] \leq \frac{1}{b^2}$$

In definitiva: $fr. rel. (|X - \mu| < b\sigma) \geq 1 - \frac{1}{b^2}$

la frazione di modalità che ricade in un intervallo simmetrico intorno alla media aritmetica, di ampiezza basata sullo scarto quadratico medio, ha una soglia minima prestabilita

Soglie tipiche

per $b < 1$ non sono informative

b	Soglia	b	Soglia
1.00	0.00	2.75	0.87
1.25	0.36	3.00	0.89
1.50	0.56	3.25	0.91
1.75	0.67	3.50	0.92
2.00	0.75	3.75	0.93
2.25	0.80	4.00	0.94
2.50	0.84	4.25	0.94

La disuguaglianza di Tchebycheff è utile soprattutto dal punto di vista teorico
Poniamo $c = b\sigma$ ovvero $b = c/\sigma$; ne consegue:

$$fr. rel. (|X - \mu| < c) \geq 1 - \left(\frac{\sigma}{c}\right)^2$$

Più grande è σ meno tipica e rappresentativa sarà la media aritmetica;

Più piccolo è σ più corto sarà l'intervallo nel quale ricade una data percentuale di unità e, tanto meno dispersa sarà la distribuzione

Scomposizione della varianza

Supponiamo che la rilevazione totale che ha media μ sia il risultato di un accorpamento di gruppi distinti di numerosità $n_i, i=1,2,...,g$ e che di ciascun gruppo sia nota la media aritmetica:

$$\mu_i = \frac{\sum_{j=1}^{n_i} X_{ij}}{n_i}; \quad i = 1, 2, \dots, g$$

E' agevole verificare che:

$$\sum_{i=1}^g (\mu_i - \mu) n_i = \sum_{i=1}^g \mu_i n_i - \sum_{i=1}^g \mu n_i = n\mu - n\mu = 0$$

La varianza si presta ad una riscrittura che evidenzia quanta parte sia attribuibile alla diversificazione interna ai gruppi (detta "within") e quanta invece sia dovuta a differenze fra i gruppi ("between").

Scomposizione della varianza/2

Quanta parte è attribuibile ad una diversificazione interna ai gruppi (varianza "within") e quanta invece sia ascrivibile a differenze fra i gruppi (varianza "between").

$$\begin{aligned} n\sigma^2 &= \sum_{i=1}^g \sum_{j=1}^{n_i} (X_{ij} - \mu)^2 = \sum_{i=1}^g \sum_{j=1}^{n_i} [(X_{ij} - \mu_i) + (\mu_i - \mu)]^2; \\ &= \sum_{i=1}^g \sum_{j=1}^{n_i} (X_{ij} - \mu_i)^2 + \sum_{i=1}^g \sum_{j=1}^{n_i} (\mu_i - \mu)^2 + 2 \sum_{i=1}^g \sum_{j=1}^{n_i} (\mu_i - \mu)(X_{ij} - \mu_i); \\ &= \sum_{i=1}^g n_i \frac{\sum_{j=1}^{n_i} (X_{ij} - \mu_i)^2}{n_i} + \sum_{i=1}^g n_i (\mu_i - \mu)^2 + 2 \sum_{i=1}^g (\mu_i - \mu) \sum_{j=1}^{n_i} (X_{ij} - \mu_i) \\ &= \sum_{i=1}^g n_i \sigma_i^2 + \sum_{i=1}^g n_i (\mu_i - \mu)^2 \\ &\quad \text{Within} \qquad \text{Between} \end{aligned}$$

Esempio

Emissioni tossiche di tre stabilimenti sono state monitorate per 5 giorni.

Giorni	Stabil. "A"	Stabil. "B"	Stabil. "C"	Totale
1	46.30	48.60	45.10	
2	43.70	52.30	46.70	
3	51.20	50.90	41.80	
4	49.60	53.60	40.40	
5	48.80	55.70	42.60	
Medie	47.92	52.22	43.32	47.82
$(\mu_i - \mu)^2$	0.01	19.36	20.25	
Varianze	6.96	5.77	5.19	19.18
Varianza w.	2.32	1.92	1.73	5.97
Varianza b.	0.00	6.45	6.75	13.21

Se i gruppi avessero la stessa media non ci sarebbe variabilità "between" e la variabilità deriverebbe solo da diversificazioni interne ai gruppi;

Se i gruppi presentassero sempre la stessa modalità sparirebbe la variabilità "within" e conterebbero solo le differenze tra i gruppi.

Analisi della varianza

I risultati della scomposizione sono riprodotti in tabella

Variabilità	Devianze	Gradi di libertà	Scarto medio
Tra gruppi	$\sum_{i=1}^g n_i (\mu_i - \mu)^2$	$g - 1$	$s_2^2 = \frac{\sum_{i=1}^g n_i (\mu_i - \mu)^2}{(g - 1)}$
Nei gruppi	$\sum_{i=1}^g \sum_{j=1}^{n_i} (x_{ij} - \mu_i)^2$	$n - g$	$s_1^2 = \frac{\sum_{i=1}^g \sum_{j=1}^{n_i} (x_{ij} - \mu_i)^2}{(n - g)}$
Totale	$\sum_{i=1}^g \sum_{j=1}^{n_i} (x_{ij} - \mu)^2$	$n - 1$	$F = \frac{s_2^2}{s_1^2}$

"F" è in grado di esprimere l'effetto della divisione in gruppi sulla variabilità.

è dato dal rapporto tra lo scarto medio (rispetto ai gradi di libertà) tra medie dei gruppi rispetto allo scarto medio nei gruppi.

Analisi della varianza : oneway

F=0 significa che non c'è effetto. Ma F>0 significa che l'effetto c'è?

Una certa differenza tra i gruppi dobbiamo aspettarcela non fosse altro che per errori di misurazione e per il fatto che non abbiamo esaminato tutti i candidati possibili.

Nei corsi più avanzati saranno chiariti i meccanismi di questa procedura

Dicendo che "F" è significativo (la classificazione induce in effetti differenze tra le medie) sbagliamo 6.5 volte su 100

Variable	PUNTEGGI
By Variable	LAUREA Tipo di Laurea

Analysis of Variance					
Source	D.F.	Sum of Squares	Mean Squares	F Ratio	F Prob.
Between Groups	2	241.2381	120.6190	3.1929	.0650
Within Groups	18	680.0000	37.7778		
Total	20	921.2381			

I momenti

La media aritmetica è interpretabile in una chiave fisica come "momento" che però cioè tendenza a ruotare intorno ad un centro

Anche lo scarto quadratico medio nasce da una media e quindi può essere considerato un momento:

$$\text{Momenti dall'origine: } \mu_\alpha = \sum_{i=1}^k X_i^\alpha f_i;$$

$$\text{Momenti centrati } \mu'_\alpha = \sum_{i=1}^k (X_i - \mu)^\alpha f_i$$

con α intero positivo.

Per i momenti valgono alcune relazioni:

$$\mu_1 = \mu; \quad \mu'_2 = \mu_2 - \mu^2; \quad \mu'_4 = \mu_4 - 4\mu_1\mu_3 + 6\mu_1^2\mu_2 - 3\mu_1^4$$

I momenti standardizzati

Un ruolo marginale, ma autonomo nella morfologia delle distribuzioni di frequenza è riservato ai momenti 3° e 4°

$$\gamma_1 = \frac{\sum_{i=1}^k (X_i - \mu)^3 f_i}{\sigma^3}; \quad \gamma_2 = \frac{\sum_{i=1}^k (X_i - \mu)^4 f_i}{\sigma^4};$$

In questa versione i momenti sono invarianti per trasformazioni moltiplicative rendendo comparabili fenomeni misurati con unità diverse.

Sotto condizioni molto generali, la conoscenza dei momenti è equivalente alla conoscenza della distribuzione (spesso, basta disporre dei primi due)

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 Comunità europea.

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		

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

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

6					
Semi	Frutti	f	F	IXi-	
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
Impiegati	25.175	Ins. Scuola	26.539	Parlamentari	50.818
Funzionari	44.851	Sottuf.	25.753	Religiosi	18.053
Dirigenti	86.828	Ufficiali	32.115		

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

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

Calciatori	Squadre	f	Xfi	Xi-μ f
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.

Scostamento semplice dalla mediana/2

Lo S_{Me} è misurato nella stessa unità del fenomeno (nell'esempio è espresso in tonnellate perché così è misurata la variabile carico);

Si annulla solo nel caso di distribuzione degenere: tutte le modalità uguali.

E' insensibile a modifiche che diano luogo a compensi negli scarti dalla mediana.

$$|X_i + \delta - M_e| + |X_j - \delta - M_e| = X_i + \delta - M_e + X_j - \delta - M_e = X_i - M_e + X_j - M_e = |X_i - M_e| + |X_j - M_e|$$

La deviazione media

Scostamento semplice medio

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

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
Impiegati	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	X _i - μ 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.

Le differenze medie

La variabilità è tanto più grande quanto maggiore è la diversificazione tra le modalità.

Una misura basata su questa idea è la differenza media di ordine “α”:

$$\Delta_R^{\alpha} = \left[\frac{\sum_{i=1}^n \sum_{j=1}^n |X_i - X_j|^{\alpha}}{n^2} \right]^{\frac{1}{\alpha}}$$

Se dalla formula si escludono gli n confronti nulli (ottenuti quando i=j) si ha la differenza media senza ripetizione

$$\Delta^{\alpha} = \left[\frac{\sum_{i=1}^{n-1} \sum_{j=i+1}^n |X_i - X_j|^{\alpha}}{n(n-1)} \right]^{\frac{1}{\alpha}}$$

Le differenze medie/2

il confronto diretto di tutte le modalità sviluppa un'idea di variabilità più ampia rispetto agli scostamenti da un valore medio:

in questi si quantifica l'ammontare delle modifiche da apportare a tutte le modalità per renderle uguali al valore di riferimento;

nelle differenze medie il riferimento è dato, a turno, da ogni modalità.

$$\begin{aligned} (\Delta_R^2)^2 &= \frac{\sum_{i=1}^n \sum_{j=1}^n |X_i - X_j|^2}{n^2} = \frac{\sum_{i=1}^n \sum_{j=1}^n X_j^2}{n^2} + \frac{\sum_{i=1}^n \sum_{j=1}^n X_i^2}{n^2} - 2 \frac{\sum_{i=1}^n \sum_{j=1}^n X_i X_j}{n^2}; \\ &= \frac{n \sum_{j=1}^n X_j^2}{n^2} + \frac{n \sum_{i=1}^n X_i^2}{n^2} - 2 \frac{\sum_{j=1}^n X_j}{n} * \frac{\sum_{i=1}^n X_i}{n} = 2 \left(\frac{\sum_{i=1}^n X_i^2}{n} - \mu^2 \right) = 2\sigma^2 \end{aligned}$$

Nonostante il diverso approccio, gli indici si somigliano.

La differenza semplice media

L'unica che ha avuto vita autonoma è

$$\Delta_R = \sum_{i=1}^n w_i X(i); \quad w_i = \frac{2*(2i - n - 1)}{n^2} \text{ per } i = 1, 2, \dots, n$$

combinazione lineare delle modalità ordinate con pesi simmetrici per modalità equidistanti della mediana ed a somma zero.

man mano che le modalità si allontanano dalla mediana aumenta il loro contributo alla variabilità.

Da notare che i pesi possono essere negativi e che per ogni negativo c'è un peso di uguale modulo, ma di segno opposto

Essa però ha il vantaggio di fondarsi sul confronto tra tutte le modalità senza far riferimento ad un valore fittizio (le medie) e misura quindi la variabilità intrinseca del collettivo.

Esempio

Paese	Produzione	w_i	i	$X_i w_i$
Spagna	12444	-0.28	1	-3456.67
Regno Unito	18733	-0.17	2	-3122.17
Francia	19061	-0.06	3	-1058.94
Italia	25179	0.06	4	1398.83
Germania	51078	0.17	5	8513.00
USA	97943	0.28	6	27206.39
				29480.44

sono perciò evitati i moltiplicoli connessi al confronto a coppie di tutte le modalità. Da notare il fatto che:

$$\Delta = \frac{n}{n-1} \Delta_R$$

non è perciò necessario impostare due formule diverse per calcolare la differenza media senza ripetizione.

La differenza semplice media è insensibile a modifiche delle modalità collocate in uguali posizioni ai due lati della mediana.

$$\Delta_R = \sum_{i=1}^n w_i |X_{(i)} - M_e| \quad \text{con} \quad w_i = \frac{|i - (n+1-i)|}{n^2}$$

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

In che modo è possibile confrontare valori singoli acquisiti in due diverse rilevazioni?

In che modo utilizzare correttamente la flessibilità delle scale intervallari e proporzionali?



Cambiamenti di scala moltiplicativi

Supponiamo che le modalità $\{X_1, X_2, \dots, X_k\}$ siano tutte moltiplicate per la costante non nulla "c":

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

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

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:

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

Lo stesso accade a moda e mediana

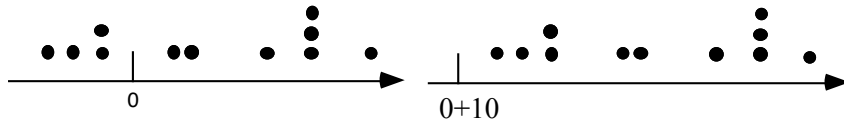
$$\mu(Y) = \sum_{i=1}^k Y_i f_i = \sum_{i=1}^k [c + X_i] f_i = \sum_{i=1}^k c f_i + \sum_{i=1}^k X_i f_i = c + \mu(X)$$

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

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.

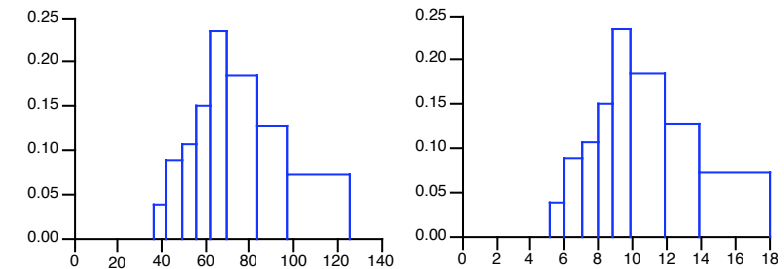
$$R(Y) = Y_{\max} - Y_{\min} = a + bX_{\max} - a - bX_{\min} = b(X_{\max} - X_{\min}) = bR(X)$$

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

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.

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

Esempio.

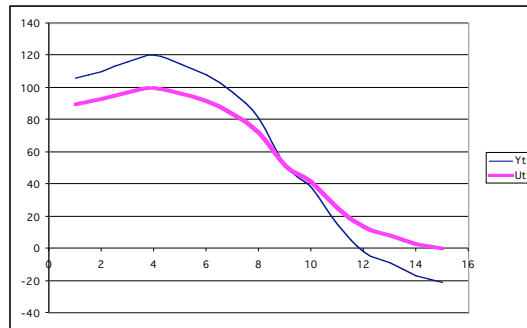
A partire dalla serie $X: \{1, 3, 5, 7\}$ si ottiene la seguente serie:

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

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

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

Attenzione! 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.

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

Ne consegue che:

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

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 1991-92.

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.

Alcuni indici

Tra gli indici di variabilità relativa più diffusi sono da includere

Coefficiente di dispersione

$$DI = \frac{Q_3 - Q_1}{|M_e|}; \quad \sum_{i=1}^k \left| \frac{X_i - M_e}{M_e} \right| f_i = \frac{S_{M_e}}{|M_e|} = CD$$

Indice di Pietra-Ricci

$$Dev. media rel.: \frac{1}{2} \sum_{i=1}^k \left| \frac{X_i - \mu}{|\mu|} \right| f_i = \frac{S_\mu}{2|\mu|}$$

E' facile dimostrare che se si moltiplicano tutte le modalità per la stessa costante (non nulla) gli indici di variabilità relativa restano immutati

Coefficiente di variazione

E' l'indice di variabilità relativa più noto.

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

Esprime la variabilità in termini di unità della media

Consideriamo la trasformazione lineare $Y_i = a + bX_i$

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

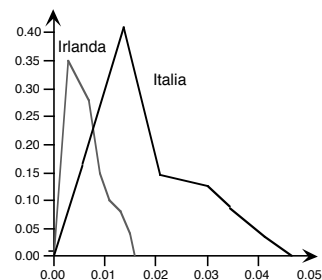
La parte moltiplicativa non ha alcuna influenza su CV, la parte additiva altera la misura della variabilità relativa anche se nulla cambia nella variabilità assoluta.

Esempio

Un'impresa multinazionale ha due stabilimenti di cuscinetti a sfere: uno in Irlanda ed uno in Italia.

Un ispettore ha rilevato i diametri di un lotto di produzione in entrambi gli stabilimenti ottenendo:

Irlanda (in.)			Italia (cm)		
0.000	0.006	35	0.000	0.012	16
0.006	0.008	28	0.012	0.022	42
0.008	0.010	15	0.022	0.028	16
0.010	0.012	10	0.028	0.034	13
0.012	0.014	8	0.034	0.040	9
0.014	0.016	3	0.040	0.046	4
100			100		



CV, CD, DMR Irlanda: {0.04276, 0.48682, 0.41054}

CV, CD, DMR Italia: {0.06893, 0.39512, 0.70514}

Maggiore variabilità implica minore qualità, ma non è chiaro dove intervenire.

Per il coefficiente di variazione e la deviazione media è lo stabilimento italiano, per il coefficiente di dispersione è quello irlandese.

Riflessione sulle scale

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 alterano la misura della variabilità relativa anche se nulla cambia nella variabilità assoluta.

Questo pone in cattiva luce le misure di variabilità relativa che non sono utilizzabili nelle scale intervallari.

Sono valide solo per scale proporzionali

La mutabilità

E' l'attitudine di variabili qualitative ad assumere modalità diverse



ETEOGENEITA'

mutabilità di variabili rilevate su scala nominale



DIVERSITA' (o BIPOLARITA')

mutabilità di variabili rilevate su scala ordinale

Sono concetti molto simili, ma nel secondo si tiene conto della ordinabilità sia delle modalità che degli scarti tra modalità.

Eterogeneità

Una variabile è eterogenea se tutte le sue categorie del dominio hanno uguale frequenza.

E' omogenea se tutte le unità presentano la stessa modalità cioè per la distribuzione degenerare.

Esempio. Stanziamenti statali per le opere pubbliche (Italia, 1990)

Settori	Nord	Centro	Sud	Totale
Trasporti	27.39%	20.78%	51.83%	100.00%
Edilizia	43.60%	19.70%	36.70%	100.00%
Ambiente	25.09%	7.12%	67.80%	100.00%
Reti	23.13%	18.02%	58.84%	100.00%
Varie	1.46%	1.51%	97.04%	100.00%
Totale	22.25%	12.07%	65.68%	100.00%

Se i settori ricevessero finanziamenti eterogenei per comparto territoriale ognuno avrebbe un terzo (33.33%) dell'importo stanziato.

Invece l'omogeneità si avrebbe nel caso un particolare settore venisse finanziato in un solo comparto territoriale

Misure della eterogeneità

1) Debbono essere nulli per la distribuzione degenerare (perfetta omogeneità) ed il valore massimo per la distribuzione in perfetta eterogeneità

$$f_i = 0, \quad i = 1, 2, \dots, k; \quad i \neq m; \quad f_m = 1$$

$$f_i = \frac{1}{k}, \quad i = 1, 2, \dots, k;$$

2) Essere crescente all'aumentare della eterogeneità.

Esempio: Complemento ad uno della frequenza relativa della moda

$$\bar{f} = 1 - f_{Mo} = 1 - \frac{n_{Mo}}{n}$$

È facile controllare che risponde ai requisiti richiesti

Alcuni indici di eterogeneità

Dei vari indici esistenti sembrano utilizzati più spesso:

$$\text{Indice di eterogeneità di Gini: } E_1 = \left(\sum_{i=1}^k f_i (1 - f_i) \right) = \left(1 - \sum_{i=1}^k f_i^2 \right);$$

$$\text{Indice semplice di eterogeneità: } E_5 = \left(\sum_{i=1}^{k-1} \sum_{j=i+1}^k f_i f_j \right)$$

La scelta dipende, da quale caratteristica della mutabilità recepisce l'indice prescelto e dalla sensibilità che mostra nel distinguere le situazioni intermedie tra la minima e la massima eterogeneità

Esempio

Quale item discrimina meglio il livello di soddisfazione dei clienti RC auto?

Item	Deliziati	Soddisfatti	Scontenti	Tot
Cortesia agenti	17.8	78.4	3.80	100.0
Velocità liquidazioni	9.8	61.3	28.90	100.0
Equità liquidazioni	1.7	76.9	21.40	100.0
Competenza agenti	9.2	82.7	8.10	100.0

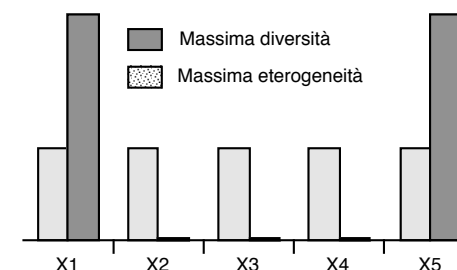
E5: Cortesia : $(17.8 * 78.4 + 17.8 * 3.8 + 78.4 * 3.8) / 100 = 17.61$

Cortesia	17.61	La domanda più eterogenea è quella relativa alla velocità di liquidazione .
Velocità	26.56	
Equità	18.13	Tale domanda è quella da scegliere per classificare le compagnie in gruppi distinti
Competenza	15.05	

La bipolarità

Si applica ai caratteri ordinali ed è analoga alla eterogeneità

Cambia il significato di "differenziazione massima" che si realizza quando la metà dei soggetti si colloca nel primo livello della variabile e l'altra metà nell'ultimo.



Per misurare la dispersione delle variabili ordinali si usano indici basati sulle frequenze cumulate

Misura della bipolarità

Una misura adatta è la seguente:

$$\text{Indice di diversità di Gini: } D_1 = \sum_{i=1}^{k-1} F_i(1 - F_i)$$

ha minimo zero se la variabile è degenere. Se $X=X_j$ per ogni "j" si ha:

$$F_i = 0 \text{ se } i < j \text{ e } F_i = 1 \text{ se } i \geq j \quad D_1 = \frac{k-1}{4}$$

il valore è massimo se le frequenze si bipartiscono tra le due categorie estreme:

per cui

$$F_i = 0.5 \text{ se } i < k \text{ e } F_k = 1 \quad D_1 = \frac{k^2 - 1}{6k}$$

Esempio

Le abitazioni di un certo comune sono state classificate secondo i tipi e le categorie delle rendite catastali.

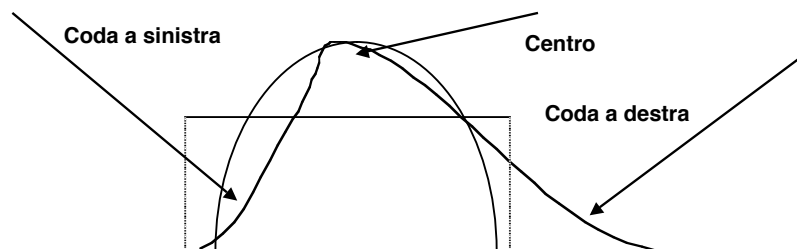
Categorie	Abitazioni	f_i	F_i	D1	D2	D3	D4
A1-Signorili	21	0.0288	0.0288	0.0279	0.0288	0.1379	0.0190
A2-Civili	86	0.1178	0.1466	0.1251	0.1466	0.1868	0.0349
A3-Economiche	127	0.1740	0.3205	0.2178	0.3205	0.1795	0.0322
A4-Popolari	40	0.0548	0.3753	0.2345	0.3753	0.2913	0.0849
A5-Ultra popolari	18	0.0247	0.4000	0.2400	0.4000	0.4333	0.1878
A6-Rurali	438	0.6000	1.0000	0.8453	1.2712	1.2288	0.5990
	730			Max= 1.2500	2.5000	2.0000	0.7454

in grassetto sono riportati i valori degli indici ed in corsivo i valori di massima bipolarità con $k=6$ categorie ordinali.

Gli indici descrivono una situazione intermedia in quanto al polo rurale "A6" non è contrapposto l'estremo "A1" e, d'altra parte, le categorie sono tutte abbastanza presenti.

Esigenze di identificazione

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



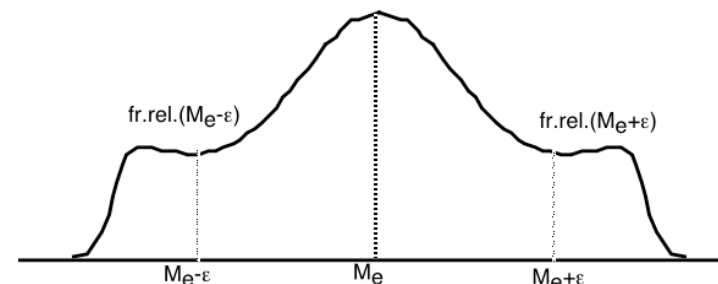
L'uso di soli indici di centralità e variabilità rende equivalenti distribuzioni molto diverse al centro e nelle code.

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

La simmetria statistica

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

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



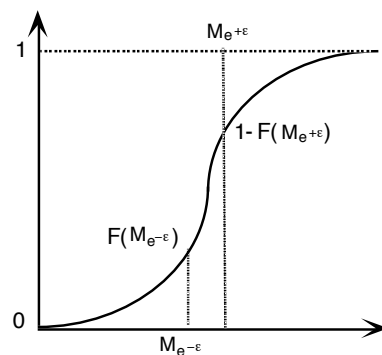
Se si piega la funzione di densità (o l'istogramma) lungo l'asse formato dalla mediana, uno dei due lati si sovrapporrà esattamente all'altro.

Altra definizione di simmetria

La simmetria può anche far riferimento alla funzione di ripartizione. Si dirà simmetrica la distribuzione per la quale:

$$F(M_e - \varepsilon) = 1 - F(M_e + \varepsilon), \quad \forall \varepsilon > 0$$

e due modalità $M_e - \varepsilon$ e $M_e + \varepsilon$ hanno ordinate complementari.



Definizione per serie

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

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

$$\text{ovvero } [X_{(n-i+1)} - M_e] = [M_e - X_{(i)}] \quad \text{per } i = 1, 2, \dots, \left[\frac{n}{2}\right]$$

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:

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

Grafico di Tukey

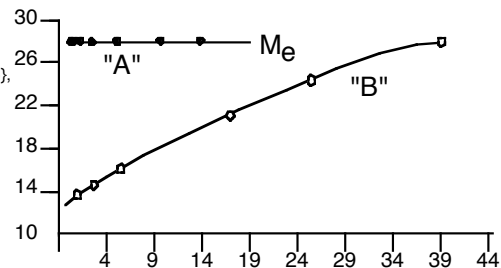
Se si rappresentano graficamente i punti di coordinate

$$\left(\frac{X(n-i+1) - M_e}{4M_e} \right)^2 + \left(\frac{M_e - X(i)}{2} \right)^2; \quad \frac{X(n-i+1) + X(i)}{2}; \quad i = 1, 2, \dots, \lfloor n/2 \rfloor$$

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 $y=M_e=27.5$;

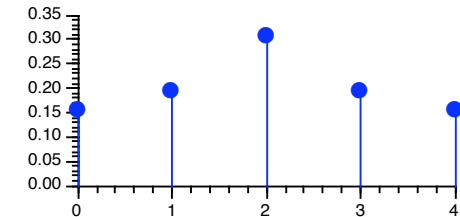
per la B seguono una curva che evidenzia lo sbilanciamento verso i valori grandi.

Definizione per seriazioni

Nel caso di modalità raggruppate dove ogni X_i ha frequenza relativa f_i si avrà simmetria se:

$$\begin{cases} (X_{k-i+1} - M_e) = (M_e - X_i) \\ f_{k-i+1} = f_i \end{cases}; \text{ per } i = 1, 2, \dots, \lfloor k/2 \rfloor$$

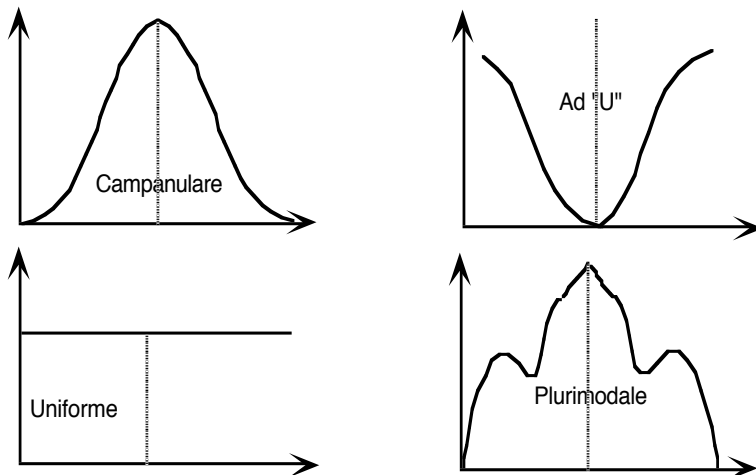
X	n	f _i	Fi	(X-M _e)f _i
0	8	0.1538	0.1538	-0.3077
1	10	0.1923	0.3462	-0.1923
2	16	0.3077	0.6538	0.0000
3	10	0.1923	0.8462	0.1923
4	8	0.1538	1.0000	0.3077
52	1.0000			0.0000



Per modalità in classi è più complessa:

$$\begin{cases} (c_{k-i+1} - M_e) = (M_e - c_i) \\ f_{k-i+1} = f_i \\ (U_{k-i+1} - L_{k-i+1}) = (U_i - L_i) \end{cases}; \text{ per } i = 1, 2, \dots, \lfloor k/2 \rfloor$$

Riconoscibilità della simmetria



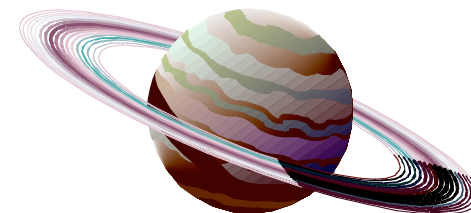
Valutazione

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.

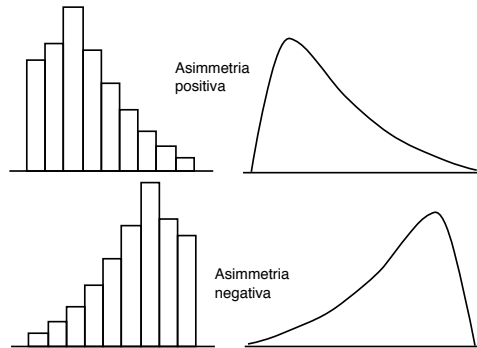
Molto più rilevanti sono quelli concernenti il diverso comportamento nelle due code



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.



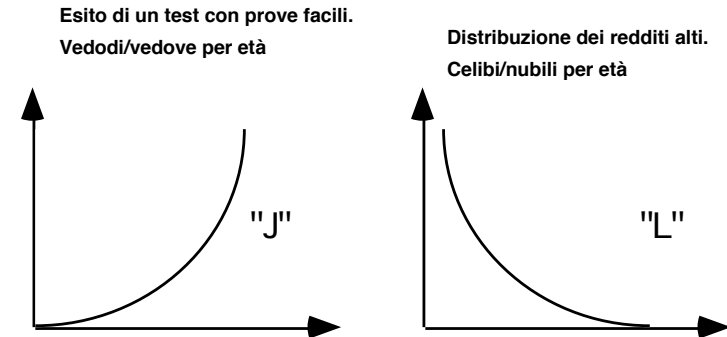
Gli scostamenti possono verificarsi al centro della distribuzione: Moda e mediana non coincidono

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

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.



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.

Indice semplice di asimmetria

Tale indice è basato sull'idea che una distribuzione è simmetrica quando il baricentro fisico ed il "centro ordinale" delle modalità, coincidono.

$$-1 \leq \alpha_1 = \frac{M_a - M_e}{S_{Me}} \leq 1$$

L'indice è standardizzato

Vale 1 quando la mediana coincide con la modalità massima $X(k)$ e vale -1 quando coincide con la minima $X(1)$. Vale zero quando la distribuzione è simmetrica.

x	A: negativa	B: positiva	C: simmetrica
0	5	51	5
1	5	24	20
2	15	15	50
3	24	5	20
4	51	5	5
	100	100	100

	A	B	C
M_e	4	0	2
μ	3.11	0.89	2
S_{Me}	0.89	0.89	0.6
α_1	-1	1	0

A causa delle compensazioni tra scarti positivi e negativi l'indice può essere nullo anche in presenza di asimmetria

Considerazioni

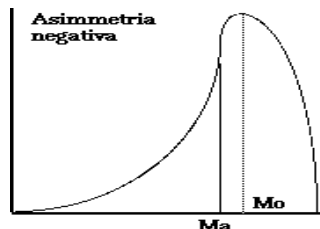
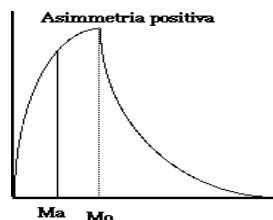
Questo indice si basa sul fatto che nelle distribuzioni unimodali simmetriche si ha

$$M_a = M_o = M_e$$

(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.

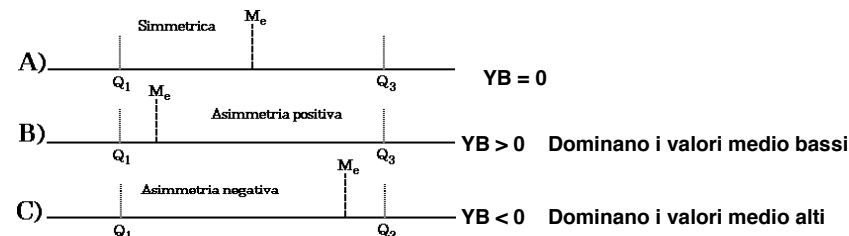
Per ragioni analoghe la media aritmetica è minore della moda per distribuzioni con asimmetria negativa.



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:

$$YB = \frac{(Q_3 - M_e) - (M_e - Q_1)}{(Q_3 - M_e) + (M_e - Q_1)} = \frac{Q_3 + Q_1 - 2M_e}{Q_3 - Q_1}$$



Proprietà dell'indice di Yule-Bowley

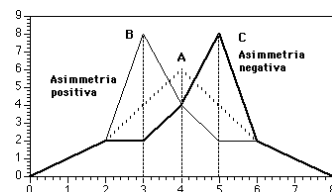
L'indice di YB è relativo $-1 \leq YB \leq 1$ ed è anche standardizzato

Il massimo negativo è ottenuto per le distribuzioni 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%.

X_i	Frequenze assolute			Frequenze rel. Cum.		
	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	1	2	0.95	0.95	0.95
7	1	1	1	1.00	1.00	1.00

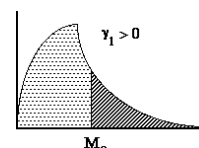
$$YB_A = \frac{3+5-2 \cdot 4}{5-3} = 0 \quad YB_B = \frac{5+3-2 \cdot 3}{5-3} = 1 \quad YB_C = \frac{3+5-2 \cdot 5}{5-3} = -1$$



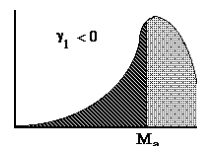
I quartili sono ottenuti con dei processi di approssimazione per cui l'accuratezza dell'indice di Yule-Bowley non può essere superiore a quella adoperata nel calcolo dei quartili.

Indice di Fisher (Skewness)

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



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 $\gamma_1 > 0$



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 $\gamma_1 < 0$.

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

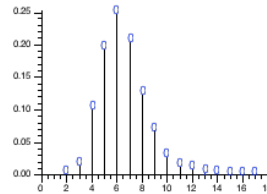
E' possibile stabilire se una distribuzione è più asimmetrica di un'altra, ma non che una distribuzione sia molto o poco asimmetrica.

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.

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