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:

$$\frac{0 + \text{🔥}}{2} = 1$$

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

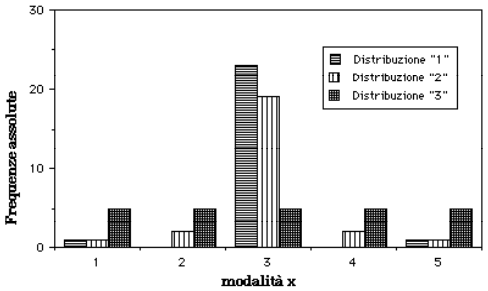
Distribuzione "3"

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 meno variabilità perché, a parità di unità considerate, è minore il numero di modalità che presenta.

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

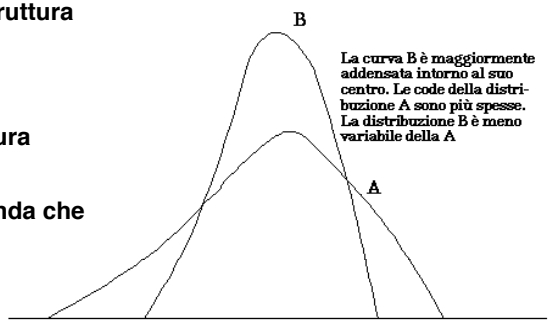


Spiegazione grafica

La media definisce la struttura del fenomeno

La variabilità esprime le deviazioni dalla struttura

Quasi sempre è la seconda che interessa studiare

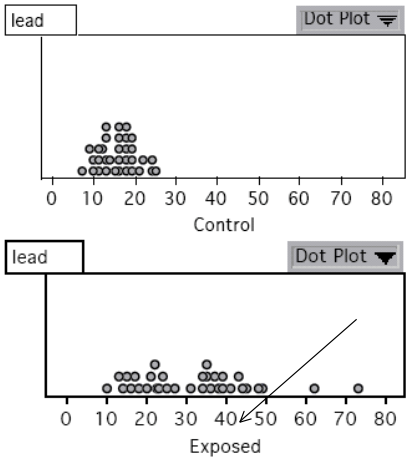


Studio della variabilità

La statistica ha un atteggiamento misto con la variabilità.

A volte cerca di minimizzarla, altre volte di massimizzarla oppure di analizzarla o stimarla.

E' l'aspetto più importante di una distribuzione anche se per comodità didattica è presentata dopo la centralità



Presenza di piombo nel sangue di 33 bambini esposti comparato con quella di un gruppo di controllo non esposto.

Il piombo in quantità elevata può ostacolare la crescita.

Superare la soglia di 40 è pericoloso. Quali sono le cause della 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 per convenzione

Studieremo tre classi di indici di variabilità:

- **-Indici posizionali di variabilità**
- **-Indici basati sugli scostamenti tra modalità**
- **-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

$$1)\ R = X_{\max} - X_{\min};$$

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

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

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

E' 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 intermedie 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

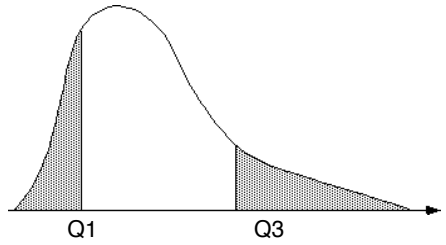
E' simile al campo di variazione solo che invece di includere il 100% delle modalità, ne include la metà ed in particolare quelle centrali:

$$DI = Q_3 - Q_1$$

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

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

Quanto più le modalità sono strette intorno al centro 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	

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

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

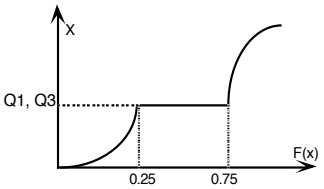
$$DI = (9.210 - 9.104) = 0.106$$

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

Caatteristiche della DI

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.



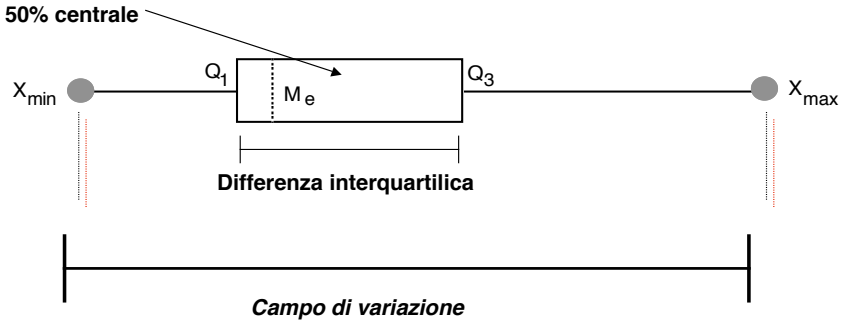
E' sufficiente che sia costante il 50% centrale della distribuzione.

Questo non succede al campo di variazione

Box Plot

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

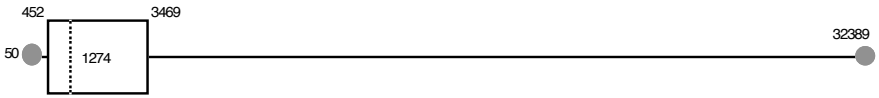
$$X_{\min},\ Q_1,\ M_e,\ Q_2,\ X_{\max}$$



Esempio

Extracomunitari iscritti alle liste di collocamento per Paese di origine

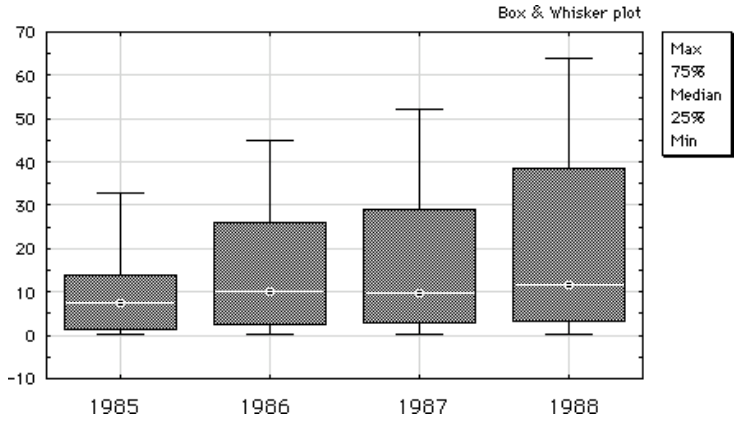
309	425	50	235	3606	4445	644	1696	2142	2863	1032	9596	429	32389	1409
2141	3440	953	11565	2503	1138	11704	452	10142	3469	687	982	71	758	161



il valore anomalo allunga la distribuzione verso i valori grandi

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.

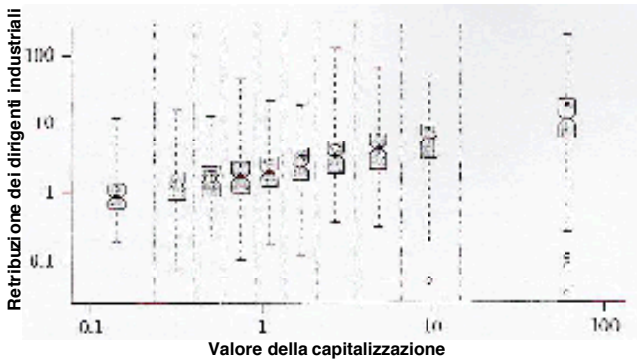
Casi patologici da escludere dalla rilevazione

Esempio: Distribuzione degli stipendi dei dirigenti industriali

+ = media geometrica
*** = media aritmetica**

Le linee verrticali 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

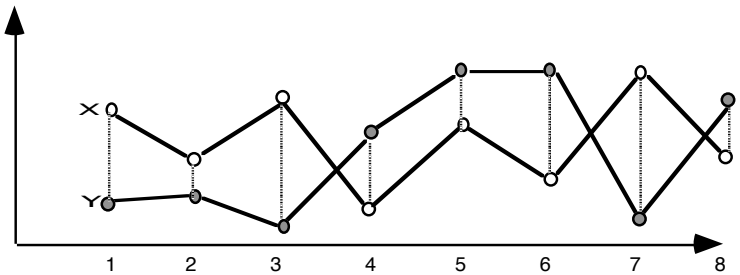
L'idea è di misurare lo scostamento complessivo (devianza) tra le due serie e per far questo possiamo adoperare una

Metrika di Minkowsky.

Di solito α=1 o α=2

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

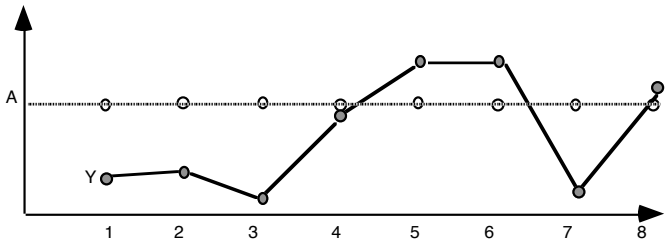
La distanza complessiva percorsa se ogni X_i si portasse sulla corrispondente Y_i.



Perché confrontare la "Y" proprio con la "X" ?
Come misurare la variabilità della "Y" se varia anche la "X"?

Scostamenti da un valore medio

(La X è 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

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

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 (o 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		

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

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 una approssimazione della variabilità poiché si fa l'ipotesi -quasi sempre non vera- che all'interno delle classi le modalità si addensino intorno al loro centro.

In certe condizioni è possibile migliorare il valore (correzione di Sheppard), ma la correzione migliore è di usare più classi o i microdati.

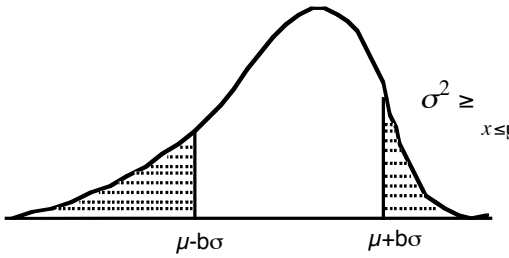
La disuguaglianza di Tchebycheff

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

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

E' positivo!

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



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

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 a *μ*, ha una soglia minima prestabilita.

Se si conoscono media e deviazione standard si può avere un'idea della frequenza

Soglie tipiche

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=bσ ovvero b=c/σ; ne consegue:

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

Più grande è σ meno tipica sarà la media aritmetica.

Più piccolo è σ più corto sarà l'intervallo nel quale ricade una data percentuale di unità e, tanto minore sarà la variabilità.

Scomposizione della varianza

Supponiamo che una rilevazione con media *μ* sia il risultato di un accorpamento di gruppi distinti di numerosità n_i ,i=1,2,…,g.

I gruppi hanno media aritmetica: $\mu_i = \frac{\sum_{j=1}^{n_i} X_{ij}}{n_i};\ i = 1,2,\ldots,g$

E' agevole verificare che:

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

Quanta parte della varianza è attribuibile alla diversificazione interna ai gruppi (detta "within") e quanta invece è dovuta a differenze fra i gruppi ("between").?

Scomposizione della varianza/2

$$n\sigma^2 = \sum_{i=1}^{g}\sum_{j=1}^{n_i} (X_{ij} - \mu)^2 = \sum_{i=1}^{g}\sum_{j=1}^{n_i} \left[(X_{ij} - \mu_i) + (\mu_i - \mu) \right]^2;$$

$$= \sum_{i=1}^{g}\sum_{j=1}^{n_i} (X_{ij} - \mu_i)^2 + \sum_{i=1}^{g}\sum_{j=1}^{n_i} (\mu_i - \mu)^2 + 2\sum_{i=1}^{g}\sum_{j=1}^{n_i} (\mu_i - \mu)(X_{ij} - \mu_i);$$

$$= \sum_{i=1}^{g} n_i \frac{\sum_{j=1}^{n_i} (X_{ij} - \mu_i)^2}{n_i} + \sum_{i=1}^{g} n_i (\mu_i - \mu)^2 + 2\sum_{i=1}^{g} (\mu_i - \mu) \sum_{j=1}^{n_i} (X_{ij} - \mu_i)$$

$$= \sum_{i=1}^{g} n_i \sigma_i^2 + \sum_{i=1}^{g} n_i (\mu_i - \mu)^2$$

Within Between

Esempio

Emissioni tossiche di 3 stabilimenti 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	19.2

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" ed esisterebbero 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 esprime l'effetto della divisione in gruppi sulla variabilità.

E' dato dal rapporto tra lo scarto medio tra le medie dei gruppi rispetto allo scarto medio nei gruppi.

Esempio

I punteggi di alcuni candidati a posti di responsabile della comunicazione di impresa sono stati suddivisi in tre gruppi: laurea scientifica, laurea umanistica, laurea in economia.

Ec	Sc	Um
86	75	92
77	77	90
84	74	87
87	78	82
91	80	75
89	89	72
92	73	
82		

				Totali
n_i	8	7	6	21
T_i	688	546	498	1732
μ_i	86	78	83	μ=82.48
$(\mu_i-\mu)^2 n_i$	99.3373	140.2546	1.6462	241.2381
$\sum(X_{ij}-\mu_i)^2$	172	176	332	680

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 anche per errori di misurazione e per il fatto che non abbiamo esaminato tutti i candidati possibili.

Analisi della varianza : oneway

p_value<0.01

```
- - - - - O N E W A Y - - - - -

Variable  PUNTEGGI
By Variable  LAUREA     Tipo di Laurea

                    Analysis of Variance

                                                        F        F
Source                                                Ratio    Prob.

Between Groups          2       241.2381       120.6190    3.1929   .0650
Within Groups          18       680.0000        37.7778
Total                  20       921.2381
```

Nei corsi più avanzati saranno chiariti I meccanismi inferenziali di questa procedura

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

I momenti

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

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

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

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

con α intero positivo.

Per i momenti valgono alcune relazioni:

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

I momenti standardizzati

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

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

Questi momenti sono invarianti per trasformazioni moltiplicative rendendo comparabili fenomeni misurati con unità diverse.

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

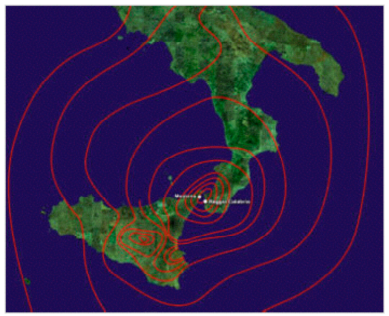
Esempio

I maggiori terremoti dell'ultimo secolo in Italia. Magnitudo in scala Richter. Calcolo dei momenti centrati.

Regione	Magnitudo	2	3	4	5	6
Calabria	6.8	46.2	314.4	2138.1	14539.3	98867.5
Calabria	5.9	34.8	205.4	1211.7	7149.2	42180.5
Sicilia	7.5	56.3	421.9	3164.1	23730.5	177978.5
Campania	6.0	36.0	216.0	1296.0	7776.0	46656.0
Sicilia	4.3	18.5	79.5	341.9	1470.1	6321.4
Abruzzo	5.0	25.0	125.0	625.0	3125.0	15625.0
Friuli	5.9	34.8	205.4	1211.7	7149.2	42180.5
Campania	6.5	42.3	274.6	1785.1	11602.9	75418.9
Sicilia	6.0	36.0	216.0	1296.0	7776.0	46656.0
Friuli	6.5	42.3	274.6	1785.1	11602.9	75418.9
Campania	7.2	51.8	373.2	2687.4	19349.2	139314.1
Sicilia	7.0	49.0	343.0	2401.0	16807.0	117649.0
Umbria	6.3	39.7	250.0	1575.3	9924.4	62523.5
	6.2	39.4	253.8	1655.3	10923.2	72830.0

Da notare l'aumento dei momenti all'aumentare dell'ordine con il conseguente aumento dei calcoli e difficoltà con gli errori di arrotondamento.

L'importanza delle modalità più grandi cresce ed assume un risalto che va spesso oltre quello loro attribuibile in base alla frequenza.



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		

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

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

Semi	Frutti	f_i	F	$\lvert Xi- \rvert f_i$
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

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
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.

Le differenze medie

La variabilità aumenta con 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.

$$\left(\Delta_R^2\right)^2 = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} |X_i - X_j|^2}{n^2} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} X_j^2}{n^2} + \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} X_i^2}{n^2} - 2\frac{\sum_{i=1}^{n}\sum_{j=1}^{n} X_i X_j}{n^2};$$

$$= \frac{n\sum_{j=1}^{n} X_j^2}{n^2} + \frac{n\sum_{i=1}^{n} X_i^2}{n^2} - 2\frac{\sum_{j=1}^{n} X_j}{n} * \frac{\sum_{i=1}^{n} X_i}{n} = 2\left(\frac{\sum_{i=1}^{n} X_i^2}{n} - \mu^2 \right) = 2\sigma^2$$

Nonostante il diverso approccio, gli indici si somigliano.

La differenza semplice media

L'unica che ha avuto vita autonoma è

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

Somma ponderata 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

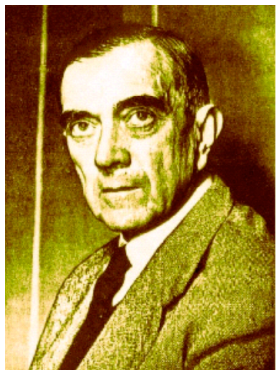
Esempio

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 eivitati i molticalcoli connessi al confronto a coppie di tutte le modalità. Da notare iltro che:

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

non è perciò necessario impostare due formule diverse pe calcolare al differenza media senza ripetizione.



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

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

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

Come 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 {X1, X2, ..., Xk} siano tutte moltiplicate per la costante non nulla "c":

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

moda e mediana sono pure riproduttive

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

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

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:

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

Lo stesso accade a moda e mediana

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

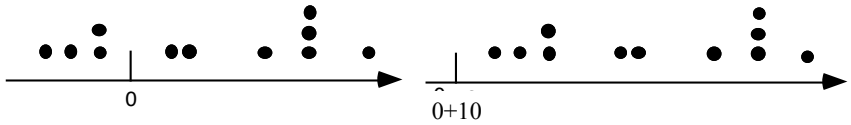
$$V(Y) = \sum_{i=1}^{k} [Y_i - \mu(Y)]^2 f_i = \sum_{i=1}^{k} [[c+X_i] - [c+\mu(X)]]^2 f_i$$

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

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.

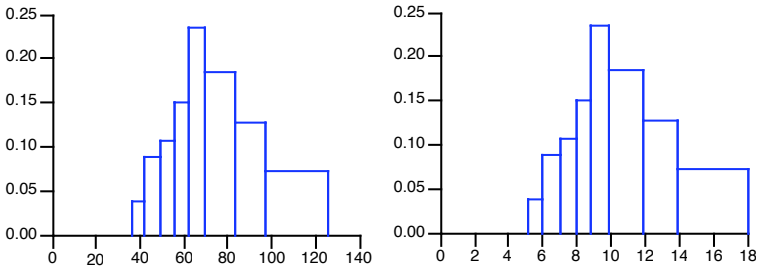
$$R(Y) = Y_{\max} - Y_{\min} = a + bX_{\max} - a - bX_{\min} = b\left(X_{\max} - X_{\min}\right) = bR(X)$$

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

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 la variabilità

- **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 $\{Y_i\}$ hanno valori nell'intervallo [0,100] con indubbi vantaggi per la costruzione dei grafici.

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

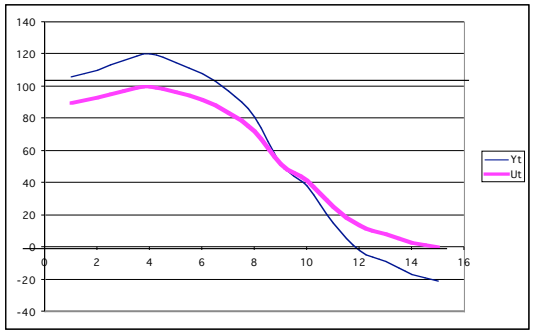
Esempio.
A partire dalla serie X: {1, 3, 5, 7} si ottiene la seguente serie:

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

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

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

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 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

Gli indici di variabilità relativa mirano ad eliminare i riferimenti dimensionali

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

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

Tali indici servono a comparare la variabilità di fenomeni espressi

Con ordini grandezza ineguali

In unità di misura eterogenee

Per campi di variazioni diversi

Alcuni indici

Tra gli indici di variabilità relativa più diffusi sono da includere

Coefficiente di dispersione

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

Indice di Pietra-Ricci

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

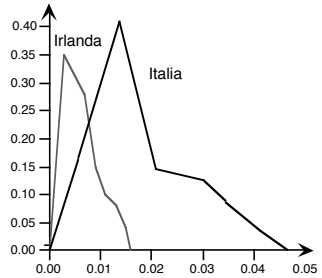
E' facile dimostrare che se si moltiplicano tutte le modalità per la stessa costante (non nulla) gli indici di variabilità relativa restano immutati

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.

Coefficiente di variazione

Esprime la variabilità in termini di unità della media

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

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 in quella assoluta.

La mutabilità

E' l'attidudine 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 degenere.

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

Alcuni indici di eterogeneità

Dei vari indici esistenti sembrano utilizzati più spesso:

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

$$\textit{Entropia della distribuzione}: \quad E_2 = -\sum_{i=1}^{k} f_i Ln\left(f_i\right) \qquad \textbf{\textit{Shannon}}$$

$$\textit{Indice semplice di eterogeneità}: \; E_4 = \left(\sum_{i=1}^{k-1} \sum_{j=i+1}^{k} f_i f_j \right)$$

La scelta dipende da quale caratteristica della mutabilità recepisce l'indice prescelto

E' anche importante e la sensibilità che mostra nel distinguere le situazioni intermedie tra la minima e la massima eterogeneità

Misure della eterogeneità

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

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

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

Debbono crescere all'aumentare della eterogeneità.

Esempio: **Complemento ad uno della frequenza relativa della moda**

$$\bar{f} = 1 - f_{Mo} = 1 - \frac{n_{Mo}}{n} \qquad \textit{Variance ratio}$$

È facile controllare che risponde ai requisiti richiesti

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: $\textit{Cortesia}: \; (17.8 * 78.4 + 17.8 * 3.8 + 78.4 * 3.8)/100 = 17.61$

Cortesia	17.61
Velocità	26.56
Equità	18.13
Competenza	15.05

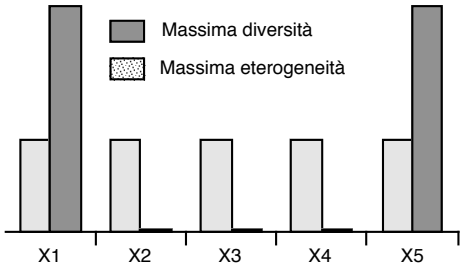
La domanda più eterogenea è quella relativa alla velocità di liquidazione .

Tale domanda è quella da scegliere per classificare le compagnie in gruppi distinti

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:

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

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

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

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

per cui

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

Esempio

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

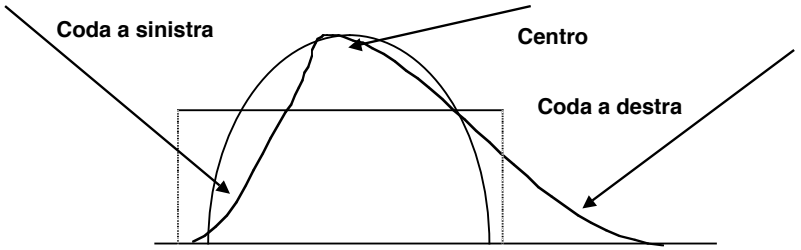
Categorie	Abitazioni	f_i	F_i	D1	D2	D3	D4
A1-Signorili	21	0.0288	0.0288	0.0279	0.0288	0.1379	0.0190
A2-Civili	86	0.1178	0.1466	0.1251	0.1466	0.1868	0.0349
A3-Economiche	127	0.1740	0.3205	0.2178	0.3205	0.1795	0.0322
A4-Popolari	40	0.0548	0.3753	0.2345	0.3753	0.2913	0.0849
A5-Ultra popolari	18	0.0247	0.4000	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

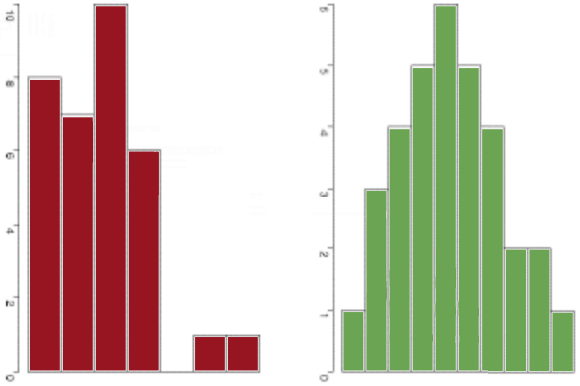
Esistono distribuzioni che hanno la stessa tendenza centrale e la stessa dispersione, ma dissimili per altri aspetti importanti?



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

Confronto di istogrammi



Presenza di piombo nel sangue di 33 bambini esposti comparato con quella di un gruppo di controllo non esposto formato da altri 33 bambini.

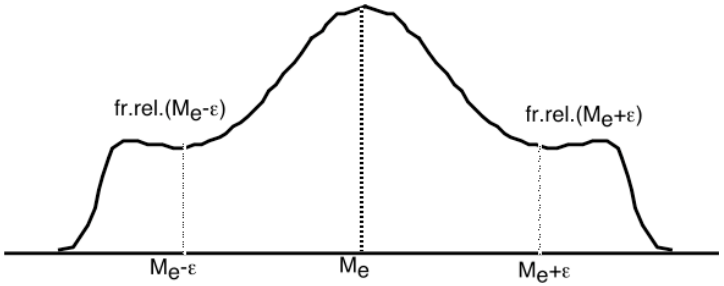
Il piombo in quantità elevata può ostacolare la crescita.

Quale dei due istogrammi è relativo al gruppo esposto?

La simmetria statistica

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

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



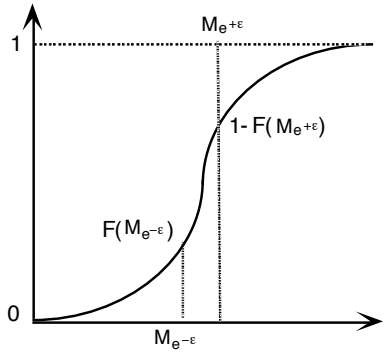
Se si piega la funzione di densità (o l'istogramma) lungo l'asse formato dalla mediana, uno dei due lati si sovrapporrà esattamente all'altro.

Altra definizione di simmetria

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

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

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

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

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

E' simmetrica la distribuzione per cui gli scarti negativi dalla mediana sono uguali (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

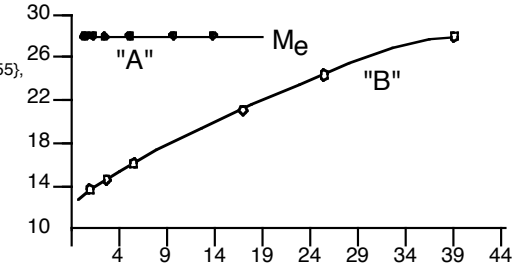
$$\frac{\left(X(n-i+1)-Me\right)^2+\left(Me-X(i)\right)^2}{4Me};\quad \frac{X(n-i+1)+X(i)}{2};\quad i=1,2,\ldots,[n/2]$$

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},

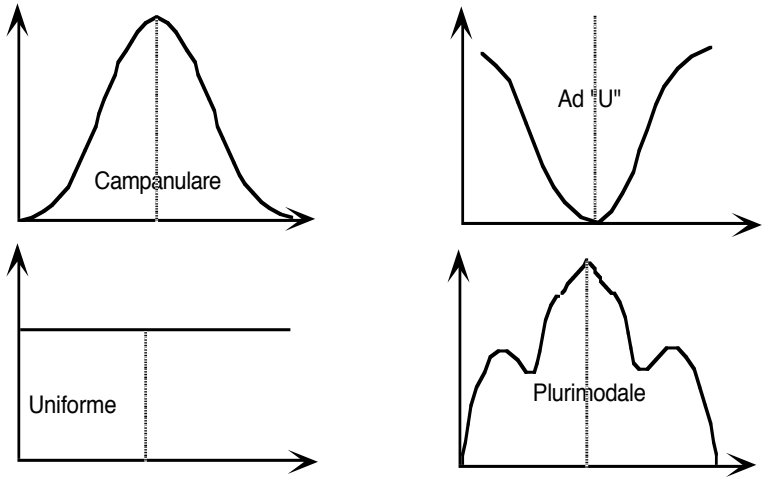
B={0, 1, 2, 4, 6, 10, 15, 21, 28, 36, 45, 55}



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.

Riconoscibilità della simmetria



Riflessione

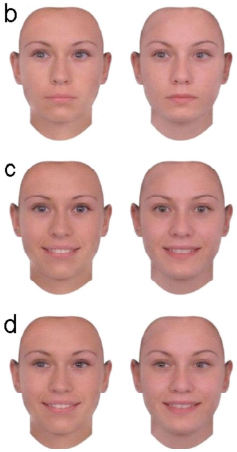
a

I volti simmetrici sono giudicati in genere più attrattivi rispetto alle facce asimmetriche.

In questo caso ci sono più aspetti da considerare: facce sorridenti o che cercano il contatto visivo.

Tuttavia è un fatto che la simmetria dei visi tende ad attirare e l'asimmetria tende a respingere. Perché?

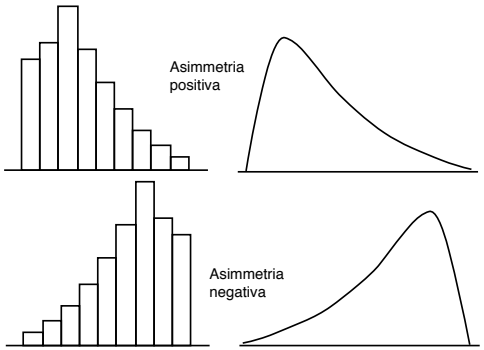
La simmetria deve avere una ragione perché più difficile da ottenere per caso



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

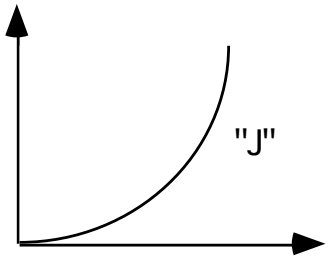
e/o nelle code per la presenza di valori remoti solo su di un lato della distribuzione.

Significato della asimmetria

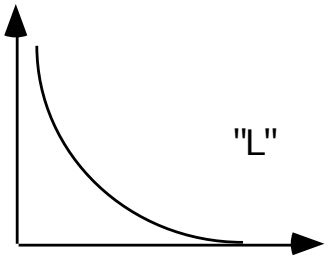
L'asimmetria aiuta ad interpretare il fenomeno.

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

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.

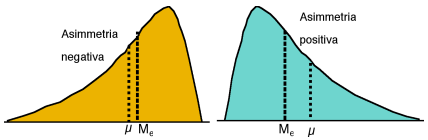
I requisiti minimi per un indice di asimmetria $\alpha(X)$ sono:

- **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.**

Indice di Bonferroni

E' basato sull'idea che una distribuzione è simmetrica se baricentro fisico e "centro ordinale" coincidono.

$$-1 \le \alpha_1 = \frac{\mu - M_e}{S_{M_e}} \le 1$$



Vale 1 se la mediana coincide con X_{max} e vale -1 se coincide con X_{min} . Vale zero se la distribuzione è simmetrica.

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

Esempio

Classificazione di due allevamenti: A e B, di mucche da latte per i giorni-mucca.

Latte		All. A
7	9	123
10	12	875
13	17	1572
18	22	2399
23	27	1777
28	32	439
33	35	2
		7187

Latte		All.B
7	10	178
10	12	1659
13	15	2624
15	20	1061
20	25	784
25	30	280
30	35	12
		6598

α_1 = -0.05, 0.163;

L'accorpamento o la divisione delle classi può modificare il valore degli indici di asimmetria.

In questo caso le differenze sono minime e presumibilmente ininfluenti.

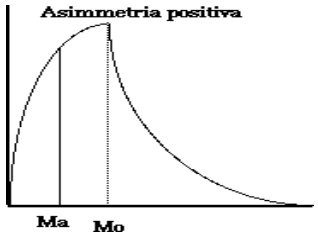
Considerazioni

A_1 si basa sul fatto che nelle distribuzioni unimodali simmetriche si ha

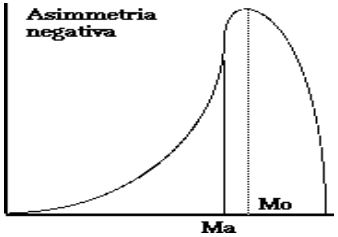
$$M_a = M_o = M_e$$

Tale condizione è un indizio di simmetria, non una certezza.

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



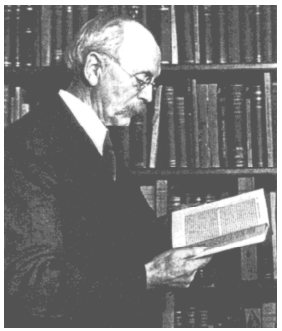
Per ragioni analoghe la media aritmetica è minore della moda per distribuzioni con asimmetria negativa.



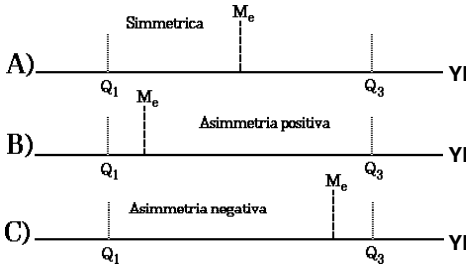
L'indice di Yule-Bowley

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

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



Sir Arthur Bowley



YB = 0. Simmetria?

YB > 0 Dominano i valori medio bassi

YB < 0 Dominano i valori medio alti

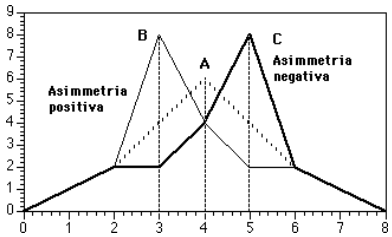
Proprietà dell'indice di Yule-Bowley

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			



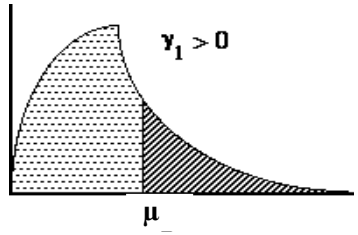
$$YB_a = \frac{3+5-2(4)}{5-3} = 0;\ YB_b = \frac{3+5-2(3)}{5-3} = 1;\ YB_c = \frac{3+5-2(5)}{5-3} = -1;$$

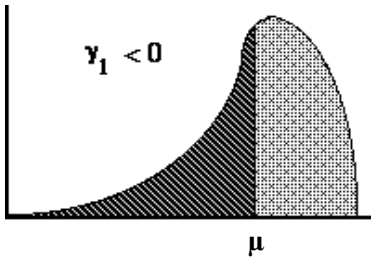
I quartili sono di solito approssimati per cui l'accuratezza dell'indice di Yule-Bowley non può essere superiore a quella adoperata nel calcolo dei quartili.

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 più grandi di quelli negativi per cui γ_1>0





Nelle distribuzioni con asimmetria negativa, gli scarti negativi (cioè per modalità inferiori alla media) prevarranno su quelli positivi (sono più distanti) e si ha γ_1 <0.

Indice di Fisher /2

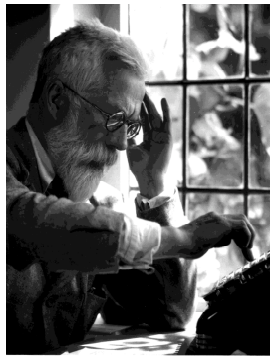
γ_1 **non varia in un intervallo definito.**

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.

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

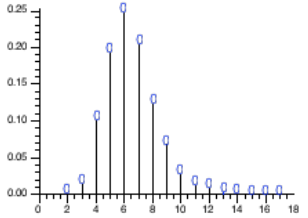
La presenza di pochi valori grandi rende l'indice elevato a causa dei cubi nella sua epsressione



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
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.

Le tre regole dell'analisi statistica

Costruire un grafico

Costruire un grafico

Costruire un grafico

Significato dei grafici

La rappresentazione grafica può essere molto utile se rispetta alcune regole

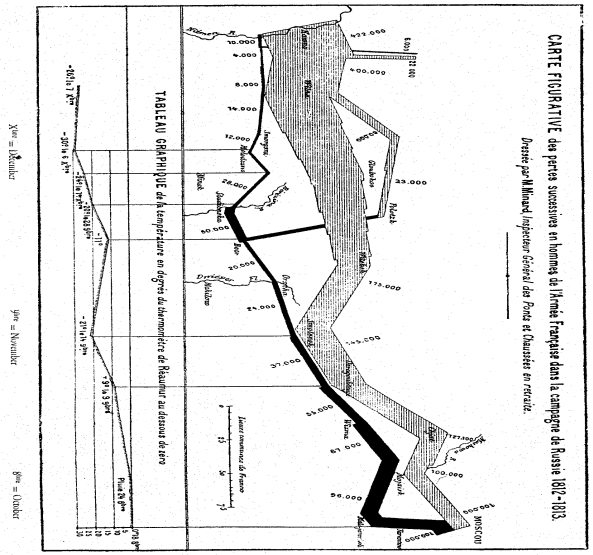
1) L'aspetto informativo deve prevalere rispetto al descrittivo;

2) I grafici portano con la stessa disinvoltura a valutazioni corrette o deviate. Solo la professionalità di chi lo compila tutela l'utente.

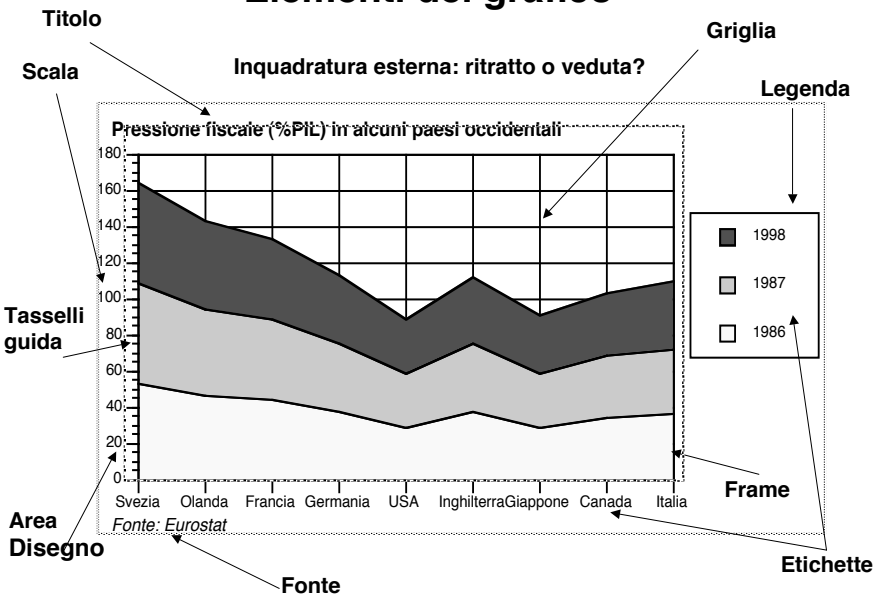
3) Deve essere improntato alla massima semplicità.
Il grafico non deve avere nulla di superfluo, astratto o misterioso.

4) Il grafico deve magnetizzare l'attenzione di chi osserva e convincere de validità dei dati e delle conclusioni presentate

Un grafico vale più di mille parole



Elementi del grafico



L'inquadratura esterna

E' la superficie che il grafico occupa all'interno della pagina

La dimensione dipende dalla disponibilità di spazio, dalla capacità visiva presunta in chi legge e dalla quantità di informazioni che vi deve trovare posto.

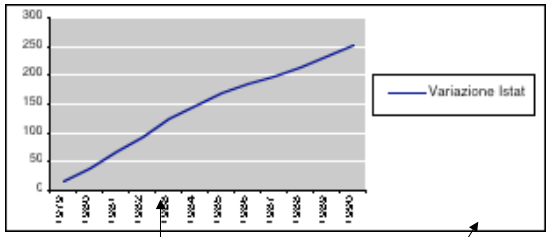
Nelle presentazioni in pubblico è bene dare i dettagli su tabelle, formule, dimostrazioni, risultati intermedi, bibliografia a voce o su trascritti.

Maggiore è lo spazio più ricca di dettagli potrà essere la figura; ma più è densa di elementi, minore è la sua leggibilità.

Ogni elemento inserito deve avere un ruolo ed una posizione conveniente

L'inquadratura esterna/2

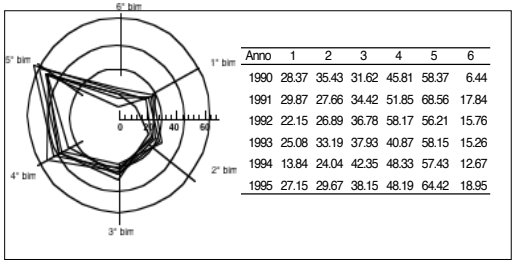
Aggiornamento annuale degli affitti per le abitazioni ultimate entro il 1975.



Area del disegno

Inquadratura esterna

Consumo di energia elettrica



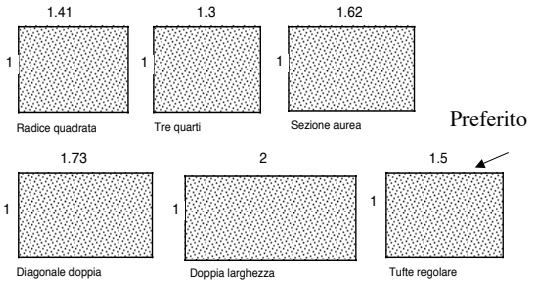
Anno	1	2	3	4	5	6
1990	28.37	35.43	31.62	45.81	58.37	6.44
1991	29.87	27.66	34.42	51.85	68.56	17.84
1992	22.15	26.89	36.78	58.17	56.21	15.76
1993	25.08	33.19	37.93	40.87	58.15	15.26
1994	13.84	24.04	42.35	48.33	57.43	12.67
1995	27.15	29.67	38.15	48.19	64.42	18.95

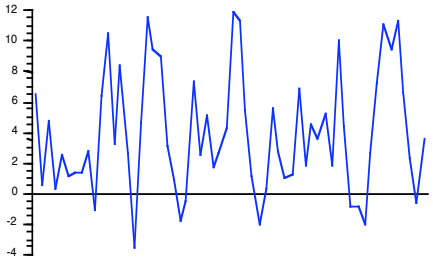
La grafica stellare offre una immagine vivida della ciclicità della serie, ma per valutazioni più puntuali è necessaria la tabella.

Quoziente immagine

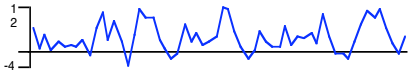
il rapporto tra l'altezza del riquadro e la sua ampiezza

Il quoziente immagine è determinante per la decodifica del messaggio da parte di chi guarda soprattutto al fine di cogliere l'andamento ascendente o discendente delle linee di tendenza nelle serie storiche.





Rimpicciolire talvolta significa migliorare



Densità dei dati

Poiché l'attenzione dell'osservatore cade al centro dell'immagine, le dimensioni dovranno essere tali che curve e simboli siano dominanti e centrali.

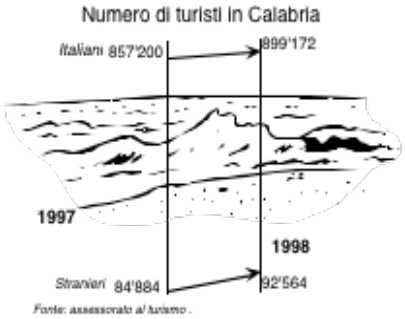
Una guida è costituita dalla densità dei dati

$$DD = \frac{unità \ * \ variabili}{superficie\ del\ grafico}$$

Se il grafico ha altezza 2.5 cm, base 5.33 e se si devono visualizzare 2 variabili per 10 unità, la densità sarà 10*2/(2.5*5.33)=1.5 cm².

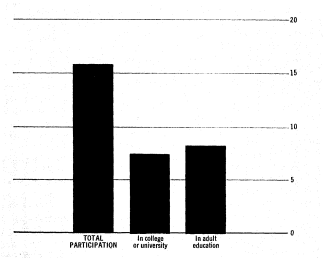
Se ad ogni dato fosse riservata la stessa porzione di grafico ciascuno occuperebbe un cm²

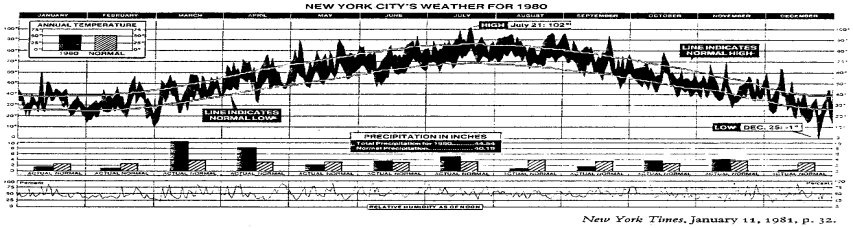
Esempio:diagramma barometricc



Esempio

In questo caso si consuma troppo toner per tre sole informazioni





New York Times, January 11, 1981, p. 32.

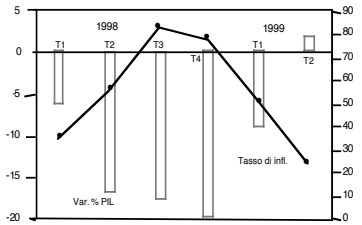
Densità del toner

Esprime il quelle parti che non possono essere soppresse senza una significativa menomazione del messaggio.

$$DT = \frac{Toner\ usato\ per\ gli\ elementi\ essenziali}{Totale\ toner}$$

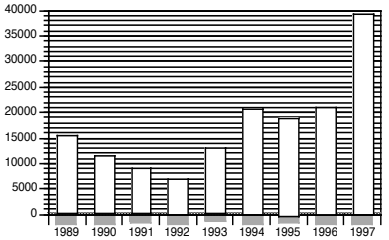
Se DT=1 ogni granello di *toner* fuso sulla pagina è necessario alla presentazione e questo è un indice di eccellenza;
DT-->0 segnalerà la presenza di decorazioni ed altri elementi non indispensabili per la comprensione dei dati.

Un buon esempio



Pil e inflazione in Indonesia

Un cattivo esempio



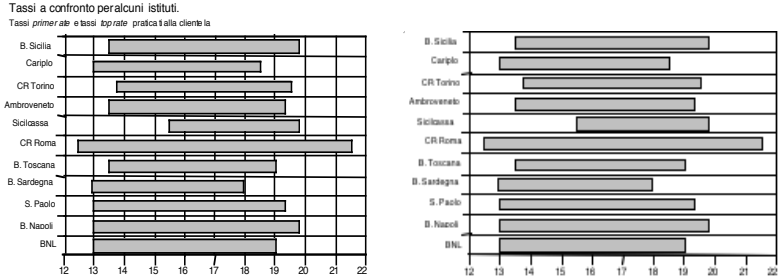
Peso dei fondi in borsa

La griglia

Lo sfondo può essere attraversato da un fascio di linee perpendicolari per dare un'idea sommaria dei valori.

Per semplificare la lettura di particolari si possono inserire anche delle linee -ortogonali alle prime- di separazione tra le unità o categorie

L'effetto è la formazione di una comoda rete che asseconda gli abbinamenti.

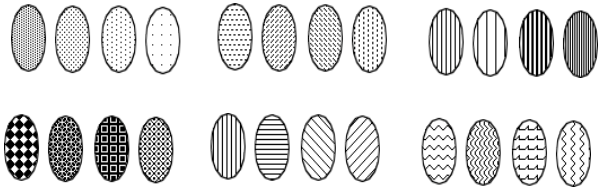


Le linee di griglia dovrebbero essere poche rispetto all'area del grafico, disegnate con un
tratto lieve che non invade l'area del grafico.

La campitura

La trama degli oggetti deve esaltare la contrapposizione tra gli elementi e, se necessario, procedere dal chiaro allo scuro (o viceversa) secondo l'ordine dei valori rappresentati.

La diversa tonalità di grigio può essere ottenuta sia con delle microstrutture *(pattern)* che con la densità di presenza di uno stesso elemento *(shading).*



Il *software* offre varie opzioni : linee e punti sono una scelta comoda e riposante, ma vanno evitate linee con diverso orientamento e altre disarmonie.

Il bianco, di solito, serve ad indicare un valore nullo o mancante; il grigio chiaro sconfina rapidamente sul bianco non appena il *toner* comincia a scarseggiare.

Le tonalità scure sono meno apprezzate delle chiare perché si appropriano di un ruolo dominante non sempre corretto; inoltre la loro stampa si sfrangia nei contorni (peggiorando in qualità) e assorbe molto *toner* (e ciò è costoso e insalubre).

La campitura/2

Suddivisione degli introiti di un concerto all'aperto

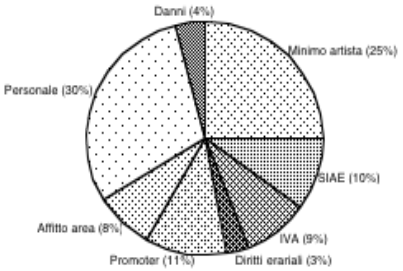
Qui si è scelta una trama tenue per i settori grandi ed una più fitta per i settori piccoli nell'idea che più grande è l'oggetto, meno contrastante può essere la sua veste; più scura è la tonalità, minore deve essere la superficie che impegna.

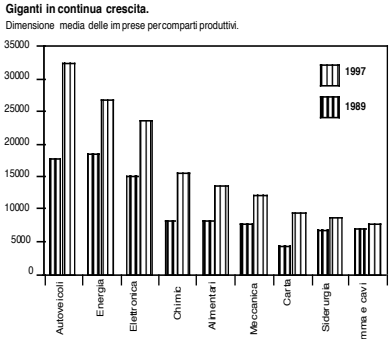
La buona scelta della campitura è fondamentale per risvegliare la concentrazione in chi sta sfogliando distrattamente le pagine di un rapporto..

La cattiva scelta della trama è evidente.

Per queste situazioni è meglio rinunciare alla campitura e presentare i rettangoli in bianco oppure con varie ed uniformi tonalità di grigio.

Peraltro è bene contentarsi delle trame più usuali, senza cercare coperture sofisticate perché è troppo alto il rischio di sovraccaricare il grafico..





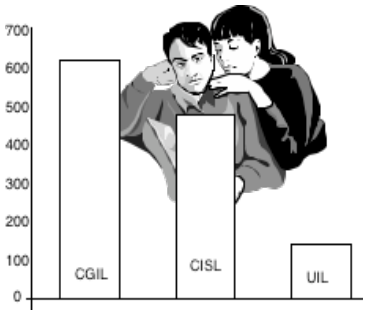
Figurazioni speciali

Si può impreziosire l'apparenza del grafico con simboli e figure che richiamino l'oggetto dei dati

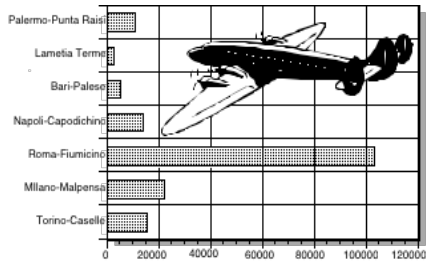
Le figure possono stare sullo sfondo del disegno ovvero decentrate, ma si deve agire con gusto e prudenza: figure con tonalità debordate e contorni a gradini hanno un pessimo effetto.

La scelta delle figure NON è neutra rispetto al messaggio che si vuole mandare.

Disoccupati iscritti ai sindacati confederali:



Aeroporti nazionali per aerei arrivati.



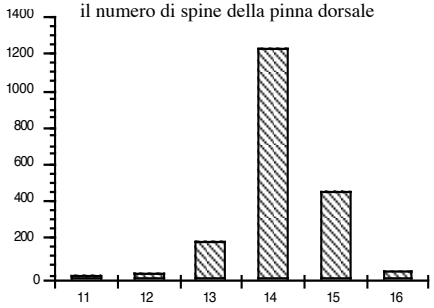
Ortogrammi

Sono grafici costruiti in un sistema di assi di cui uno quantitativo su cui si riportano i valori ed un altro qualitativo su cui sono indicati i soggetti.

Tra le figure è lasciato un certo spazio per mostrare le differenze

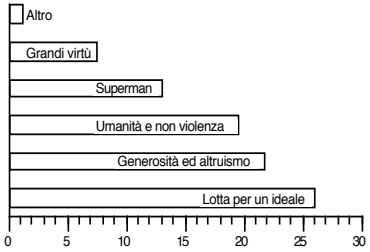
A colonne

Classificazione di 1900 cernie secondo il numero di spine della pinna dorsale



A barre

Interviste ad un campione di giovani e adolescenti "Eroe è chi…

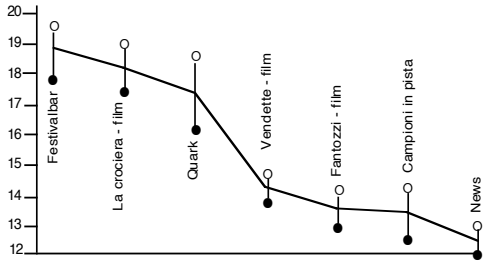


Ortogrammi Min/Max

Se i rettangoli sembrano ingombranti ci si può limitare al solo valore minimo e massimo riportati su segmenti delimitati da due simboli che rimarcano l'estensione.

***Share* dei programmi di prima serata (20:30-22:30) punte minime e massime. Dati Auditel. Estate 1999.**

Programma	Min	Max
Festivalbar	18.02	19.40
La crociera - film	17.45	18.75
Quark	16.30	18.30
Vendette - film	13.91	14.43
Fantozzi - film	13.04	13.94
Campioni in pista	12.65	14.05
News	12.02	12.74



La linea di raccordo centrale è fittizia, ma serve per dare unità al grafico.

Opzionalmente si può far risaltare, con un apposito simbolo, il valore centrale del campo di variazione.

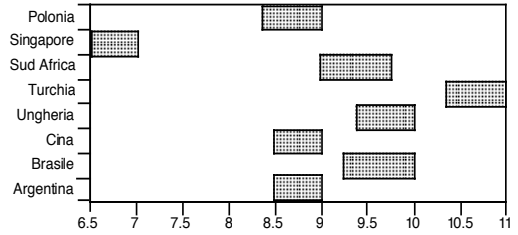
Ortogrammi fluttuanti

Se l'obiettivo è la rappresentazione del solo campo di variazione sono disponibili diverse tipologie di ortogrammi.

Gli ortogrammi fluttuanti sono costituiti da rettangoli di base fissa e di altezza proporzionale al campo di variazione.

Tassi di sconto per operazioni di medio importo in dollari USA (5 anni)

Paese	Minimo	Massimo
Argentina	8.500	9.000
Brasile	9.250	10.000
Cina	8.500	9.000
Ungheria	9.375	10.000
Turchia	10.375	11.000
Sud Africa	9.000	9.750
Singapore	6.500	7.000
Polonia	8.375	9.000

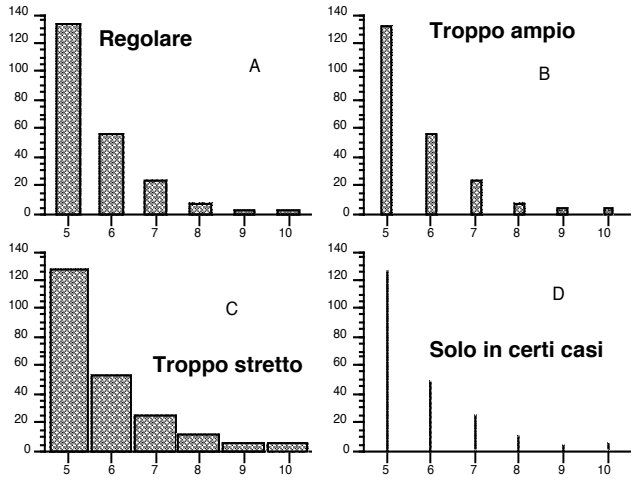


La peculiarità è la collocazione dei rettangoli che poggiano sulla linea del loro valore minimo.

Interspazio

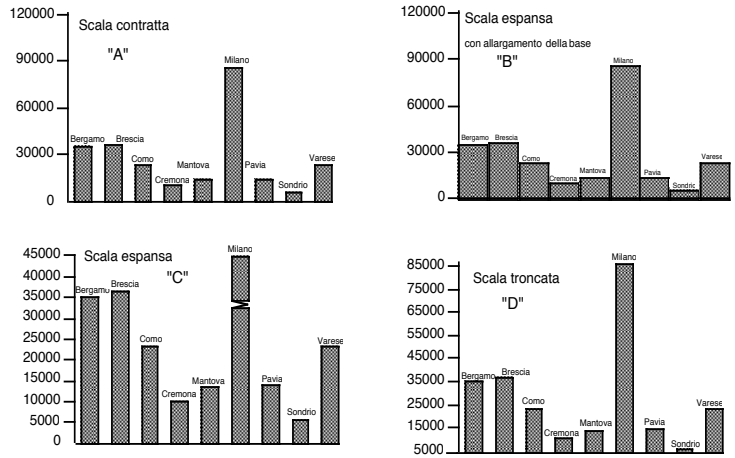
L'interspazio tra le colonne o le barre non dovrebbe essere inferiore alla comune ampiezza dei rettangoli, ma la misura precisa è una scelta personale:

Petali	Fiori
5	133
6	55
7	23
8	7
9	2
10	2
	222



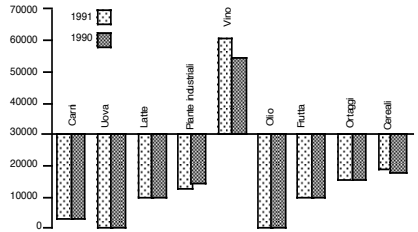
Scelta della scala per gli ortogrammi

La scala dell'asse quantitativo deve essere scelta con equilibrio e trasparenza

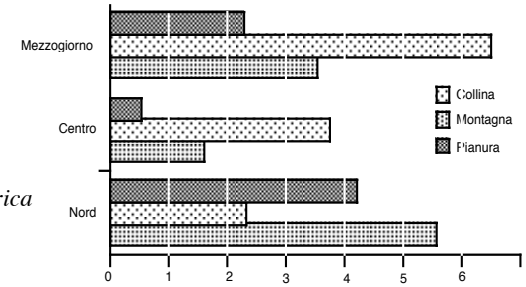


Ortogrammi multipli

Gli ortogrammi di due o più variabili sono presentati in forma congiunta per far notare il loro andamento parallelo e per meglio sfruttare l'area del disegno.

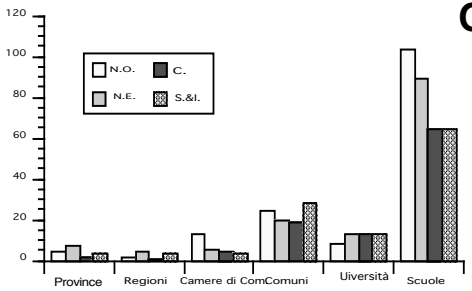


Andamento delle principali produzioni agricole e agro-alimentari



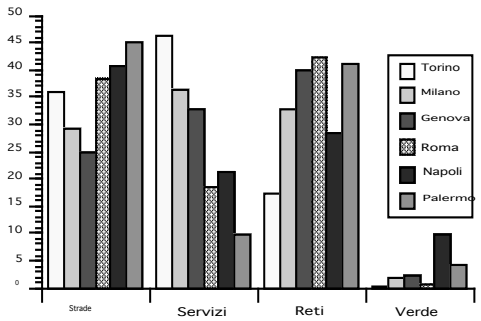
Superficie territoriale (in milioni di ettari) per comparto e zona altimetrica

Ortogrammi multipli/2



Tutte le unità per ogni variabili

Tutte le variabili per ogni unità

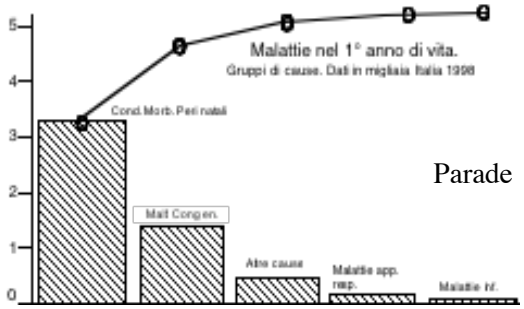


Ortogrammi paretiani

Talvolta è utile presentare i valori secondo il loro ordine di grandezza. Il fine è di esaltare le unità o le modalità dominanti

L'ogiva indica l'importanza congiunta delle modalità aggregate per ordine di rilevanza.

L'inclinazione dei segmenti riflette quello che la nuova modalità aggiunge a ciò che le altre più importanti hanno già realizzato.



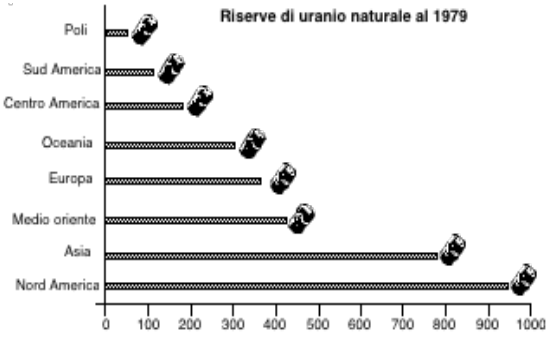
Valutazione integrata di aspetti separati

L'andamento dell'ortogramma ricorda il modello di Pareto dei poligoni di frequenza da cui segue il termine "paretiano".

Ortogrammi a punti

Il rettangolo è portatore di due informazioni visive: per la lunghezza e per l'area. Una sembra ridondante.

In questo ortogramma i rettangoli sono compressi fino a formare una linea.

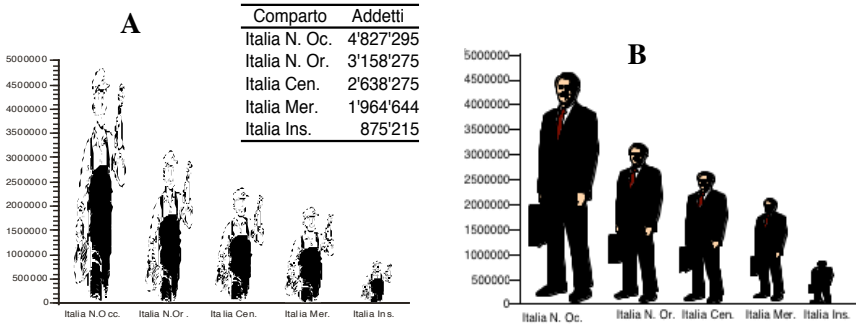


Sul punto terminale si pone un cerchietto o un altro simbolo che mette in risalto la distanza dalla base.

L'ortogramma a punti è il più efficace per comunicare dati che includono poche modalità.

Pittogrammi

La loro particolarità risiede nel sostituire i rettangoli dell'ortogramma con disegni o figure mnemoniche così da attrarre l'osservatore più distratto

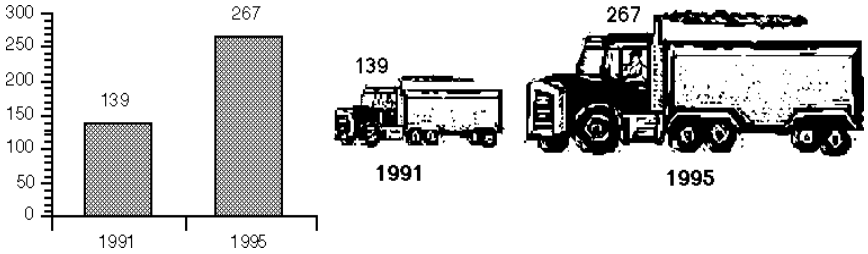


In "A" la figura si adatta al rettangolo di un ortogramma stirandola verso l'alto (o allungandola, in caso di barre), ma ignorando le proporzioni interne.

In "B" si cerca di mantenere l'armonia tra le parti.

Difetto dei pittogrammi

L'altezza della figura è proporzionale ai valori, ma deve esserlo anche la larghezza per ragioni di armonia. Se il rapporto tra due valori è 1:2, le figurepotrebbero farlo sembrare 1:4



Tra i due periodi si è realizzato un incremento del 100%: il valore del '91 è la metà del valore del 1995.

La rappresentazione ideogrammatica mostra invece un aumento di gran lunga superiore. Si dovrebbe usare la radice dei valori.

Forma impattante, ma scorretta

Ortogrammi a figure ripetute

Consentono l' uso delle figure e di mantenere i reali rapporti numerici

L' idea è di usare una figura stilizzata che rappresenti l'unità di conto.

La si ripete per quante volte è contenuta nel valore da rappresentare

Disegnata solo in parte rappresenterà - pro rata- il resto della divisione.

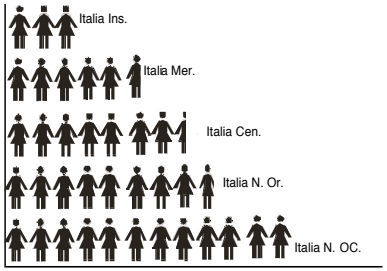
$$c = unità\ \ di\ \ conto; \quad y_i = modalità\ \ i-esima$$

$$\begin{cases} \left[\frac{y_i}{c}\right] = parte\ \ intera \Rightarrow Numero\ \ di\ \ figure \\ \frac{y_i}{c} - \left[\frac{y_i}{c}\right] = frazione \Rightarrow parte\ \ di\ \ figura \end{cases}$$

=400'000

$$\begin{cases} \left[\frac{3'158'275}{400'000}\right] = 7\ \ figure \\ \frac{3'158'275}{400'000} - \left[\frac{3'158'275}{400'000}\right] = 90\%\ \ di\ \ 1\ \ figura \end{cases}$$

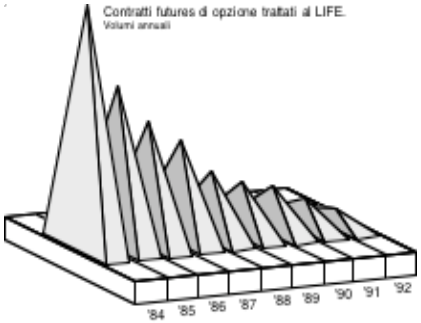
Comparto	Addetti	p.i.	p.f.
Italia N.Oc.	4'827'295	12	0.07
Italia N.Or.	3'158'275	8	0.90
Italia Cen.	2'638'275	7	0.60
Italia Mer.	1'964'644	5	0.91
Italia Ins.	875'215	2	0.19

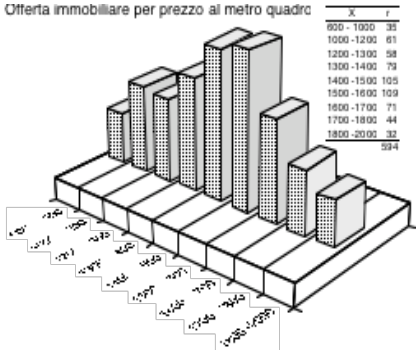


Ortogrammi stereoscopici

La redazione degli ortogrammi può essere abbellita rendendo percepibili in rilievo le colonne o le barre.

La presenza poi delle basi di appoggio impreziosirà l'illustrazione.





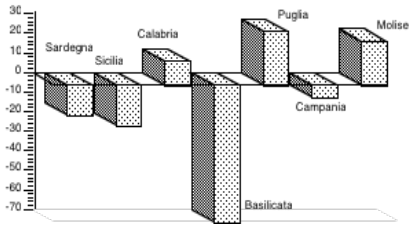
La realizzazione di questi grafici coinvolge l'uso di metodi proiettivi ed implica diverse scelte

Il software grafico solleva da molti problemi pratici, ma non tutti quelli

Effetti d'ombra

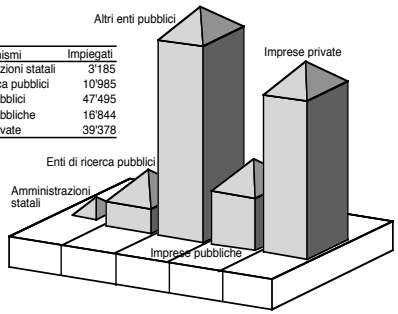
Per dare un aspetto più realistico si simulano degli effetti d'ombra (un tratteggio più scuro per la faccia che si ritiene non illuminata) che danno profondità agli oggetti, ma che possono sviarne la lettura.

Variazione percentuale degli appalti pubblici nelle regioni meridionali.



Unità di personale in organismi di ricerca

Organismi	Impiegati
Amministrazioni statali	3'185
Enti di riceca pubblici	10'985
Altri enti pubblici	47'495
Imprese pubbliche	16'844
Imprese private	39'378



Inganni visivi

L'effetto stereoscopico rende gli ortogrammi piacevoli alla vista, ma occorre essere consapevoli delle deformazioni che comportano.

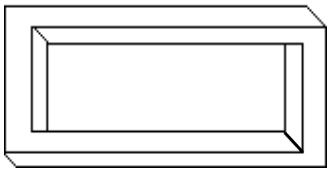
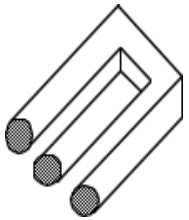
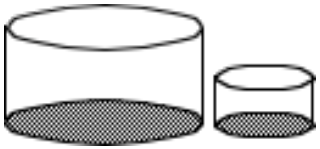


Figura 3D che esiste solo nel piano



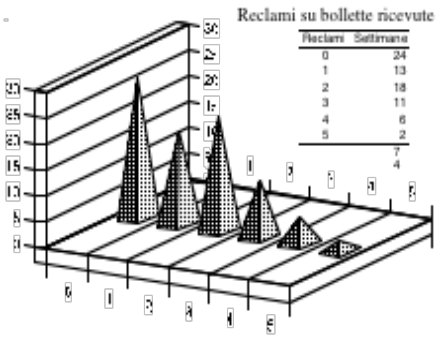
Effetto stereoscopico sbagliato indot dal disegnatore



Il più piccolo è ottenuto scalando quello più grande del 50%. Le differenze tra i due volumi appaiono comunque molto superiori.

Ortogrammi metafisici

L'effetto 3D può essere spinto oltre gli ortogrammi stereoscopici per visualizzare i valori con ortogrammi metafisici a grandissimo effetto.



Ricordano le Piazze d'Italia dei quadri del pittore "metafisico" G. De Chiri

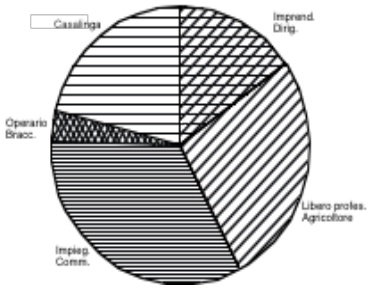
Diagrammi areali

Utili se si vuole far risaltare la relazione "parte al tutto" tra i valori.

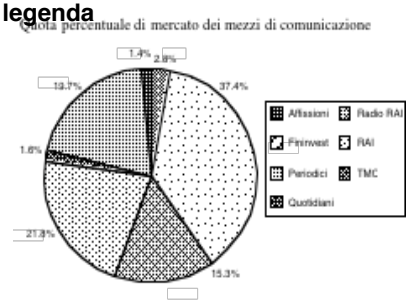
Famiglie italiane per professione del capofamiglia

X_i	n_i	f_i	g_i
Imprenditore-Dir.	2365	0.1540	55.44
Libero prof.-Agr.	4131	0.2690	96.83
Impiegato-Com.	5068	0.3300	118.80
Operaio-Bracc.	553	0.0360	12.96
Casalinga	3241	0.2110	75.97
	15358	1.0000	360.00

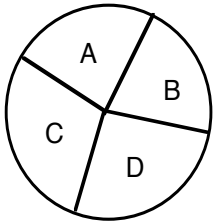
Se g_i il grado da abbinare ad X_i; si ha $g_i = 360f_i$ con f_i pari al peso della modalità o unità di pertinenza nei dati.

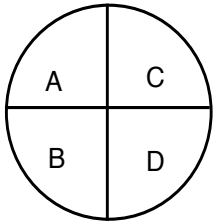


Conviene inserire a parte la legenda

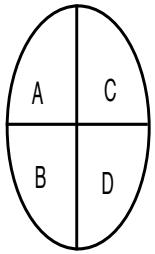


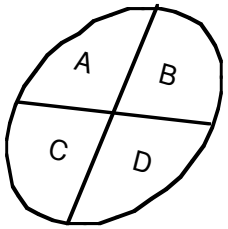
Difetti dei diagrammi a torta

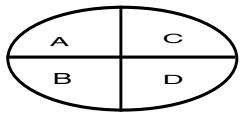




La comparazione dei settori può essere ostacolata dall'inclinazione delle linee







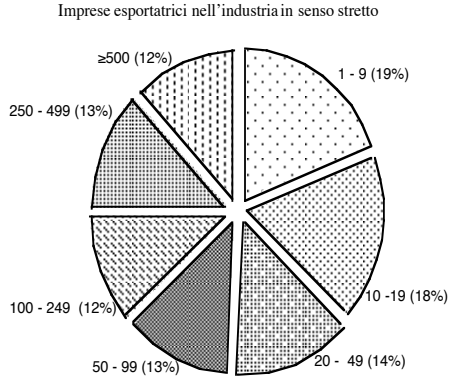
Se si cambia angolatura, si cambia la percepibilità, si cambia il messaggio

Settori separati

Si rendono più chiari i rapporti tra valori

Si evidenziano i settori più piccoli

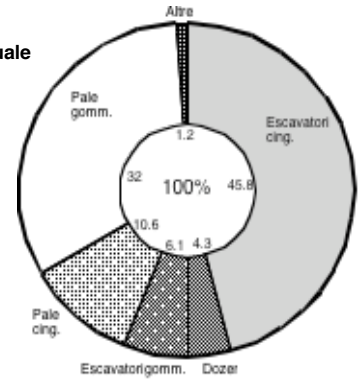
Il numero di settori può essere aumentato



Diagrammi a ciambella o anello

L'idea di proporzionalità e circolarità può essere realizzata riferendo le quote ai tranci in cui può essere suddivisa l'area racchiusa tra due cerchi concentrici

Composizione percentuale delle vendite di ruspe. Gennaio-giugno 1993.



Lo sapzio al centro è usato per esprimere il totale ripartito tra le modalità (in questo caso è 100% dato che il grafico è costruito direttamente sulle percentuali).

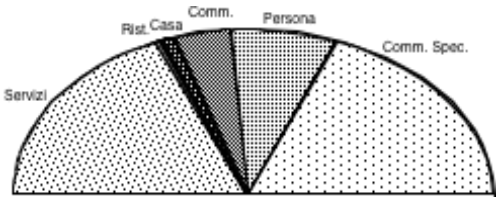
Può anche essere adoperato per inserire valori, percentuali o etichette senza ingombrare l'inquadratura esterna.

Diagrammi a ventaglio

L'idea di proporzionalità e circolarità può essere realizzata riferendo le quote ai 180° della semi circonferenza

Numero di punti vendita in franchising in Italia. Anno 1996

Settore	Franchisor	g_i
Comm. Spec.	3225	64
Comm. non Spec.	899	18
Persona	1597	32
Casa	234	5
Servizi	2968	59
Ristorazione	107	2



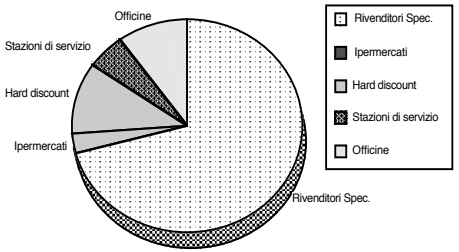
Questo grafico occupa meno spazio rispetto al diagramma a torta senza perdita rilevante di qualità.

Questo è utile su giornali e riviste o per comunicati commerciali a pagamento.

Il numero di elementi che si possono rappresentare bene diventa però più

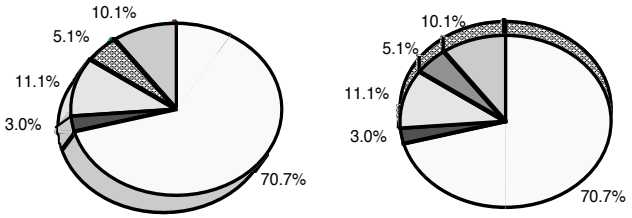
Diagrammi circolari in rilievo

Si impreziosisce la figura:



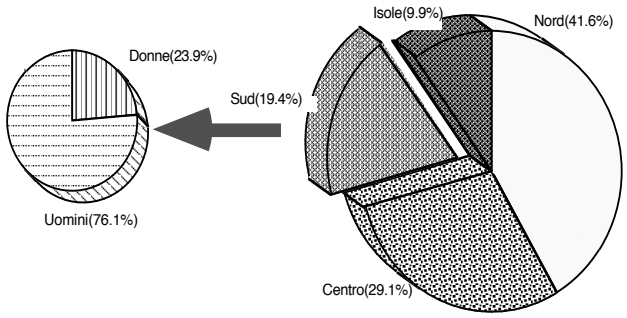
Angolature ed orientamento sono a scelta.

E' facile sbagliare.



Estrusione di un settore

Si può analizzare separatamente un settore

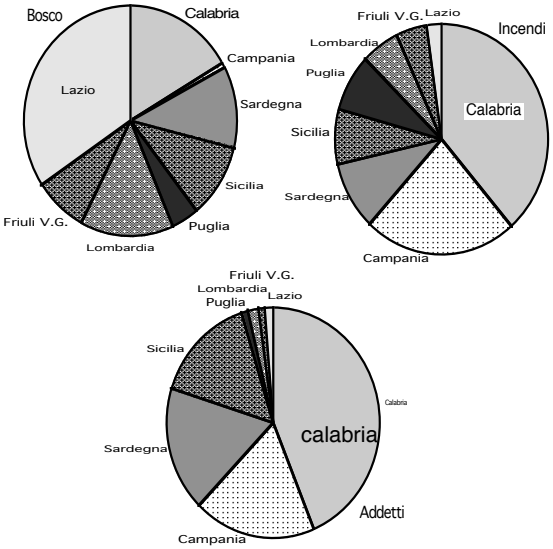


Suddivisione per zona dei consulenti del lavoro al 1997.

Diagrammi a torta multipli

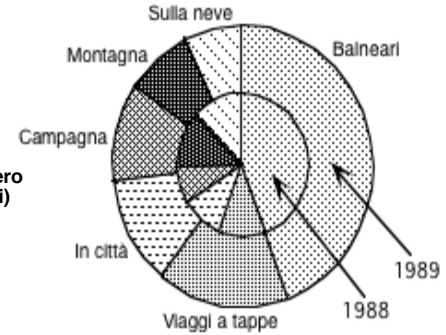
Devono essere pochi e con poche modalità

Regioni	Bosco	Addetti	Incendi
Calabria	576	15600	5.3
Campania	289	6500	3.2
Sardegna	399	6200	1.3
Sicilia	361	5500	1.1
Puglia	149	310	1.1
Lombardia	500	580	0.8
Friuli V.G.	285	222	0.6
Lazio	1194	540	0.3



Esempio

Vacanze e soggiorni degli italiani all'estero per il biennio 1988-1989 (Milioni di viaggi)



I viaggi a tappe sono cresciuti nel 1989 rispetto al 1988 laddove le vacanze sulla neve registrano una perdita di 4 punti in percentuale.

E' però difficile dire se nel complesso le vacanze e soggiorno degli italiani siano cambiate.

La tecnica dei cerchi inscritti può riguardare due occasioni (tre, inserendo un terzo cerchio, ma le modalità da comparare debbono proprio poche)

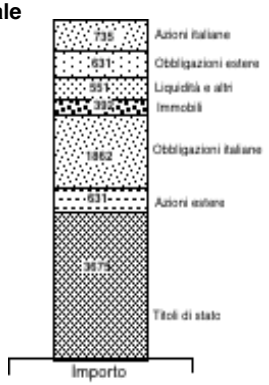
Ortogrammi frazionati

Si sfrutta la semplicità degli ortogrammi unita alla suddivisione areale dei diagrammi a torta.

Si sceglie un ortogramma di dimensione prefissata come totale per poi frazionarlo come le modalità si rapportano al totale

Composizione (in milioni) del portafoglio di una importante società finanziaria al 31.12.1995

Tipologia	Importo	Quota%
Titoli di stato	3675	43.35
Azioni estere	631	7.44
Obbligazioni italiane	1862	21.97
Immobili	392	4.62
Liquidità e altri	551	6.50
Obbligazioni estere	631	7.44
Azioni italiane	735	8.67
Totale	8477	100.00



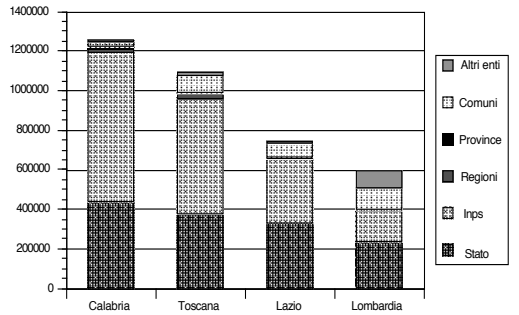
E' netta la prevalenza dei titoli di Stato e delle obbligazioni italiane.

L'impatto del grafico non è inferiore al diagramma a torta ed è più facile da realizzare; inoltre, consente di inserire un numero maggiore di categorie rispetto ad altri diagrammi areali.

Ortogrammi frazionati multipli

L'uso dell'ortogramma frazionato per confronti multipli è abbastanza diffuso

Spesa finale netta degli enti pubblici per l'assistenza nel 1994 (lire procapite)



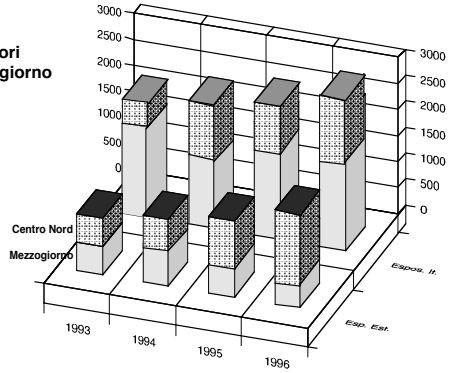
E' leggibile la categoria disposta sul livello inferiore (Stato) e quella più grande (INPS), ma è difficile individuare altri elementi di raffronto significativi.

Sarebbe forse opportuno aggregare le altre categorie in una sola voce.

Ortogrammi frazionati in 3D

Gli ortogrammi frazionati, combinati con elementi metafisici, raggiungono livelli straordinari di attrattività.

Serie storica della presenza di espositori italiani e stranieri nelle fiere del Mezzogiorno e del Centro-Nord Italia.



Gli elementi figurativi a cui è associato un dato sono solo le altezze dei segmenti che compongono i rettangoli; tutto il resto è scenografia.

La speranza è che il pubblico a cui è mostrata non si faccia beffe di ciò che

Grafici per le serie storiche

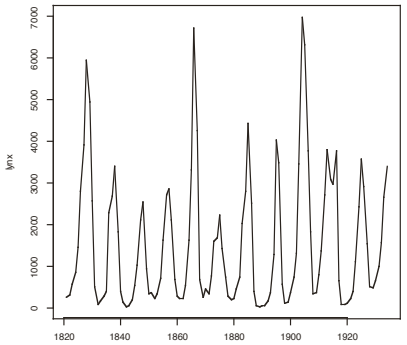
I grafici delle serie storiche sono molto diffusi perché è difficile individuare relazioni o riconoscere tendenze con la lettura di tabelle.

L'andamento del grafico di una serie storica è utile

per comprendere il suo gradiente evolutivo (Trend)

per datare la svolta in una linea di sviluppo

per circoscrivere i periodi di picco e di valle



per delimitare l'arco di variazione di un fenomeno.

L'obiettivo delle rappresentazioni grafiche è di far risaltare tali aspetti senza che per questo si debba considerare ogni singolo dato.

t	C_t
Anno	*Bovina*
1961	14.0
1962	15.6
1963	17.8
1964	17.3
1965	17.3
1966	20.4
1967	22.5
1968	22.5
1969	23.5
1970	24.8
1971	25.2
1972	24.3
1973	26.0
1974	24.5
1975	22.4
1976	23.0
1977	23.1
1978	24.1
1979	24.5
1980	25.5
1981	25.2
1982	25.1
1983	25.4
1984	25.2
1985	25.1
1986	25.5
1987	26.2
1988	26.3
1989	26.9
1990	26.2

Esempio di serie storica

Il fenomeno presenta delle regolarità. Dobbiamo scoprirle e valutarne l'entità.

Time Sequence Plot o Profilo

Questo è lo strumento più semplice ed importante per l'analisi delle serie storiche

I profili

In un sistema di assi cartesiani si associa quello delle ascisse al *continuum* temporale su cui si collocano, rispetto ad un periodo base, le circostanze di rilevazione.

Sulle ordinate si riportano le modalità (quantitative metriche).

I punti così ottenuti sono interconnessi con varie tecniche:

Con un segmento di retta (profili lineari)

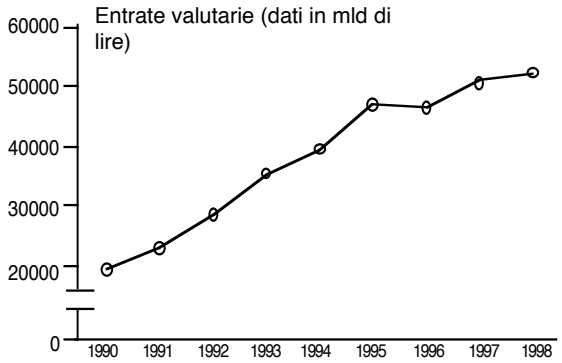
Con due segmenti fra di loro perpendicolari (a torri)

Con una linea continua detta *spline* (profili continui)

Profilo lineare

I valori sulle ordinate coprono il campo di variazione estendendosi leggermente oltre per dare respiro al disegno.

Le etichette sono riportate con arrotondamenti alle migliaia



Spesso gli zeri finali sono omessi e l'unità di valore è indicata altrove.

Si è eliminata la zona relativa ai valori (teorici) compresa tra lo zero ed il valore arrotondato -per difetto- al minimo riscontrato nella rilevazione

Esempio

La presenza della griglia, per quanto diradata e tenue, confonde e nasconde l'andamento della serie storica.

Alla cattiva qualità del grafico contribuisce anche il tono di grigio adoperato per il simbolo dei valori.

Zero e Non Zero

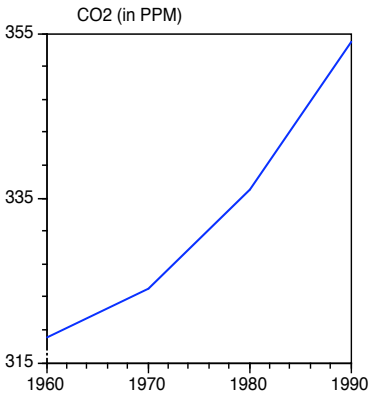
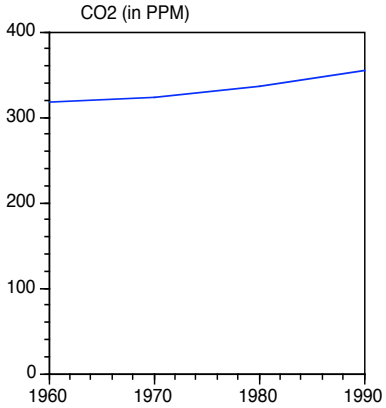


Grafico presentato da Al Gore al Senato USA per sensibilizzare la pubblica opinione sull'effetto serra.

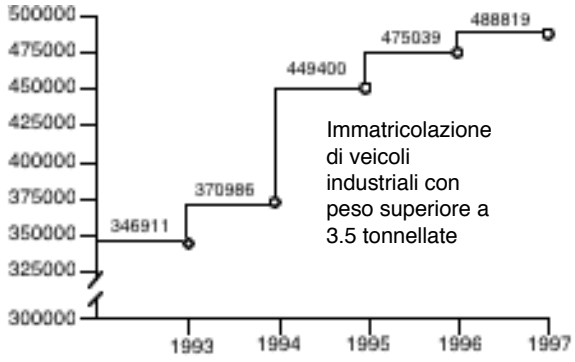
E' corretto, ma ottiene l'effetto contrario.



Questo è il grafico vero.

Profilo a torri (o a gradini)

"... Conviene dare forma grafica a tali fenomeni cumulativi in modo che l'occhio percepisca subito, la circostanza che ciascuna intensità va riferita a tutto un intero periodo, e non ad un solo istante di esso.

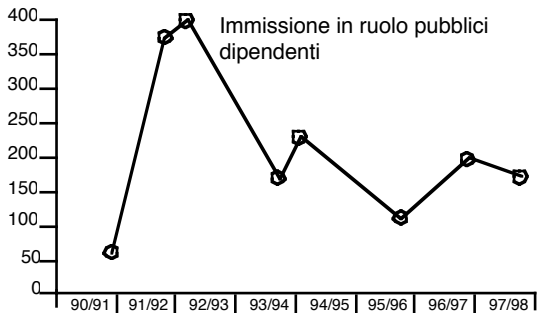


Ciò si ottiene sostituendo alla spezzata con segmenti inclinati, una spezzata con segmenti orizzontali e verticali, in modo che ciascun segmento orizzontale corrisponda ad un intero periodo di osservazione".

Esempio

Le immissioni in ruolo in un ente avvengono una volta l'anno, anche se in periodi diversi nell'anno.

Il grafico recepisce le date di assunzione disponendo i simboli in corrispondenza della posizione -nell'anno- in cui si colloca l'assunzione.

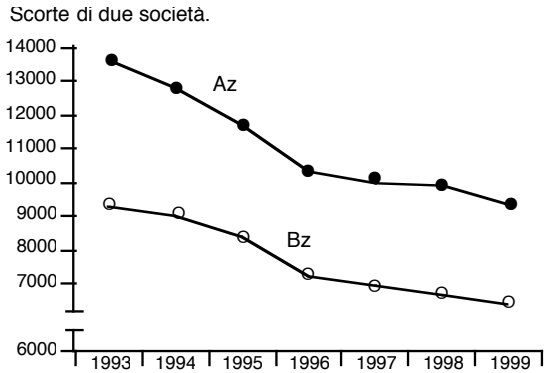


Questo è sbagliato perché l'unità di tempo è l'anno intero e non il giorno

Si deve scegliere una posizione costante per ogni unità di tempo per non creare una fonte di irregolarità non necessaria.

Un altro errore è il raccordo lineare tra i simboli che prefigura una evoluzione graduale fuori luogo per il tipo di fenomeno considerato

Esempio

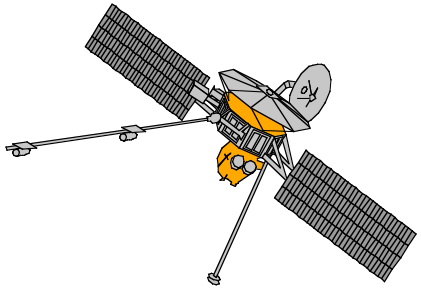


Il raccordo con segmenti di retta di due punti consecutivi mantiene il carattere unitario della rilevazione permettendo di seguire la dinamica dei fenomeni a confronto: è evidente l'andamento parallelo delle due variabili e la loro comune riduzione progressiva.

La scelta dell'inclinazione dei segmenti è fatta dal computer, ma non è neutra rispetto alla percezione del grafico. Si può sfruttare la dimensione del punto per disporre di una discreta gamma di tangenti

Fenomeni di flusso

Il valore riscontrato ad un dato istante è il livello raggiunto a partire dall'ultima rilevazione già effettuata ed include tutti i movimenti -in entrata e in uscita- che hanno interessato il fenomeno nel corso dell'unità di tempo



Poiché il dato si estende a tutto l'arco temporale è bene collocare il riferimento al centro dell'unità in ascissa.

il valore da riportare è spesso ottenuto come semisomma dei valori iniziali e finali per cui la collocazione naturale è proprio in mezzo.

Profilo continuo (o spline)

Si realizza con il computer determinando per ogni "m" punti il polinomio cubico (o di altro grado minore di "m") che passi più vicino ai punti.

E'consigliabile se i dati derivano da fenomeni di flusso osservati con appositi strumenti, ma di cui si dispone o si vogliano dare solo alcuni valori.

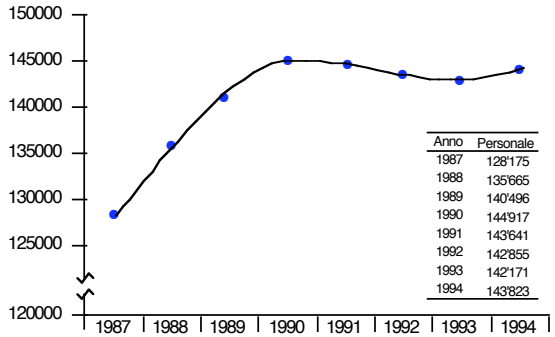
La tecnica delle *splin*e ha un elevato grado di soggettività tanto nello scegliere il grado del polinomio che il numero dei punti cui applicarlo.

E' per questo che si preferisce un raccordo con spezzate di retta, meno sinuose e piacevoli, ma più composte e stabili.

Esempio

Il personale è una variabile di flusso dato che l'organico può essere rilevato in ogni istante dell'anno aggiungendo all'ultima registrazione, il saldo tra coloro che hanno lasciato il lavoro e gli assunti nell'anno.

Personale addetto alla ricerca e sviluppo in Italia



Anno	Personale
1987	128'175
1988	135'665
1989	140'496
1990	144'917
1991	143'641
1992	142'855
1993	142'171
1994	143'823

La continuità del fenomeno risalta con un raccordo continuo che forma un oggetto unico: il profilo della serie storica.

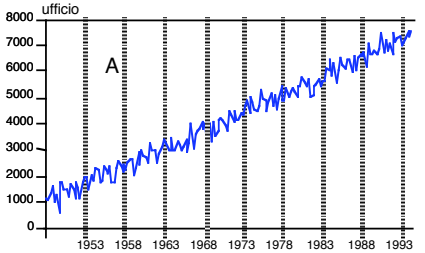
Da sottolineare l'inserimento della tabella all'interno dell'area del grafico che non desta troppo fastidio visto che sfrutta una zona periferica.

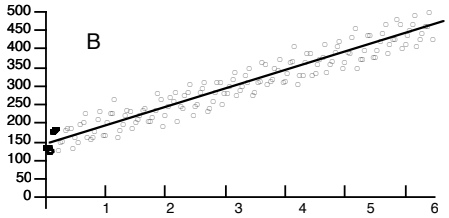
Serie storiche lunghe

Se la serie è lunga oppure le rilevazioni avvengono con frequenza si può ridurre l'affollamento sulle ascisse riportando le etichette per unità più ampie: trimestri o anni invece di mesi, quinquenni o decenni invece di anni.

Anche i raccordi tra i punti possono essere soppressi se questo serve all'economia del grafico

serie trimestrale della spesa di consumabili per ufficio

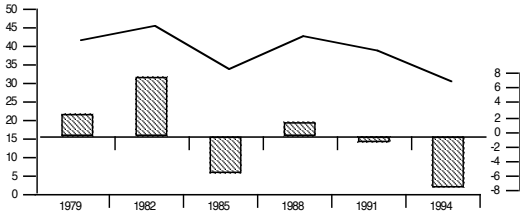




Accorgimenti migliorativi

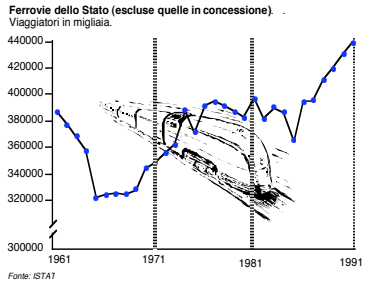
Aiuti erogati dall'AIMA (valori in miliardi di lire).

Al profilo è affiancato un ortogramma per una migliore interpretazione o per renderlo più attraente.



Le immagini hanno la duplice funzione di intrigare chi guarda e ricordare quale sia il soggetto dei dati

Se il grafico è molto prolungato non si riesce a cogliere il suo andamento complessivo con un solo colpo d'occhio. Sono utili le fincature

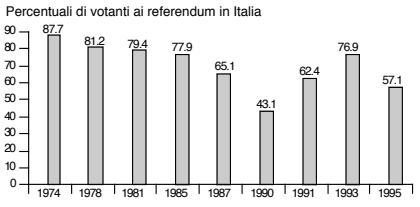


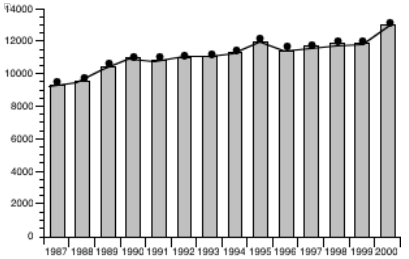
Serie storiche ed ortogrammi

Se le modalità sono ripetizioni di un fenomeno osservato in condizioni e/o tempi ristretti la serie storica è rappresentabile anche con ortogrammi.

raccolti agricoli concentrati in pochi giorni dell'anno
i diplomati delle scuole superiori a luglio
il pagamento dell'imposta personale sui redditi.

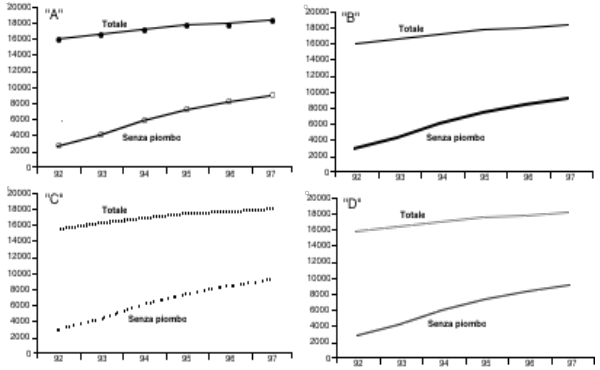
Percentuali di votanti ai referendum in Italia





Talvolta i vertici degli ortogrammi sono collegati con dei segmenti di retta, ma è una pratica incoerente e poco utile

Simboli e tracciati



In "A" c'è ridondanza. La denominazione rende inutile fornire l'ulteriore dettaglio incapsulato nel simbolo.
In "B" le linee sono poco leggibili negli incroci.
In "C" ci sono poche scelte di tratteggio distinguibili
In "D" le tonalità di grigio sono dispendiosi per il toner e precludono la possibilità di una griglia pure realizzata in grigio distintamente

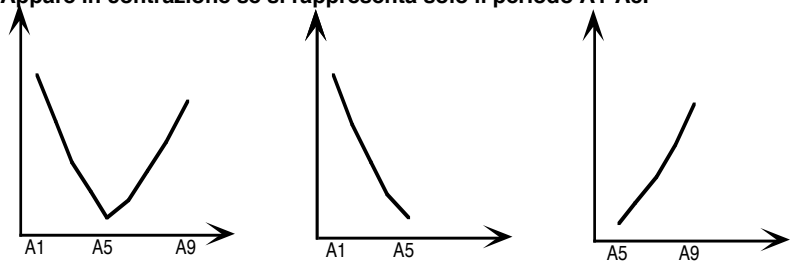
Integrità del grafico/1

A seconda dei periodi che si riportano, il lettore riceve un messaggio diverso:

Se si usano tutti sembra di essere di fronte ad un fenomeno che dopo un periodo di caduta si sia ripreso

Se si usa il solo periodo A5-A9 il fenomeno sembra in espansione

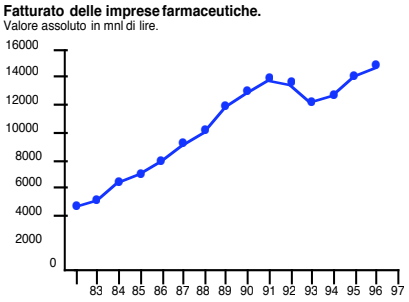
Appare in contrazione se si rappresenta solo il periodo A1-A5.

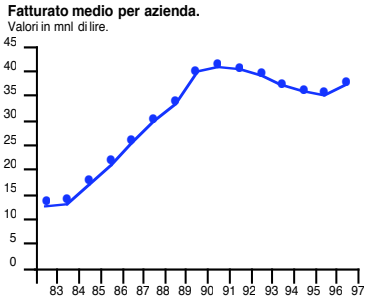


Il grafico è importante non solo per quello che mostra, ma anche per quello che nasconde

Integrità del grafico/2

I valori riportati debbono agganciarsi correttamente al fenomeno





Se invece del fatturato complessivo delle imprese, si rappresenta il fatturato medio si riceve un'immagine molto diversa del trend in atto.

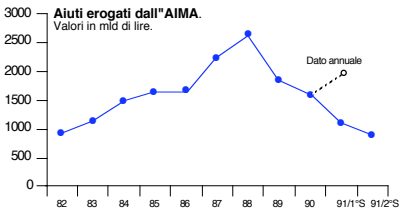
E' evidente che l'aumento del primo è solo dovuto ad un maggior numero di imprese attive e non ad un aumento reale dell'attività

Integrità del grafico/3

Chi osserva può trovarsi fuori strada a causa di una cattiva scelta del campo dell'unità temporale di rilevazione o mal riportata sulle ascisse

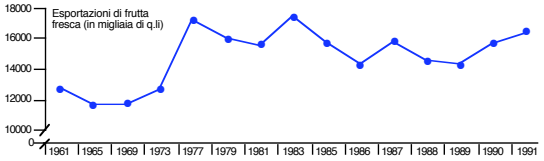
Gli ultimi due valori sono semestrali e la loro presentazione separata oscura l'avvio di quella che potrebbe essere una ripresa.

In generale, è bene dare coerenza al grafico riportando i valori per unità a composizione uniforme.



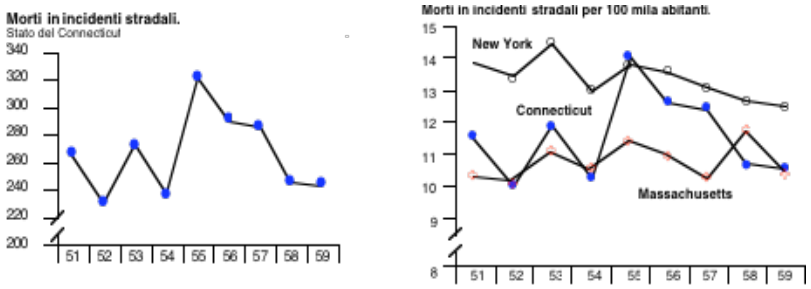
Spazi uguali sulle ascisse non corrispondono ad uguali intervalli temporali.

Invece di portare un chiarimento, la riduzione dei punti ha complicato il messaggio.



Integrità del grafico/4

Chi osserva può trovarsi fuori strada a causa di una cattiva scelta del campo dell'unità temporale di rilevazione o mal riportata sulle ascisse

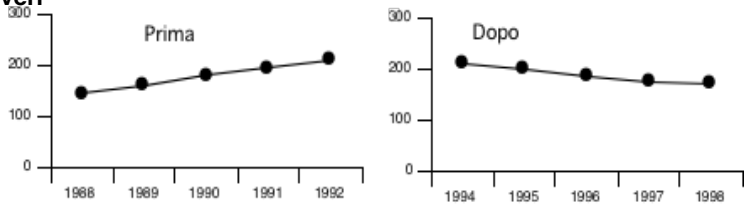


La serie storica segnala un deciso declino degli incidenti dopo il 1955 che risulta un anno disastroso.

La notizia perde molto del suo portato informativo quando la serie è messa a confronto con analoghe serie di altri Stati.

Il grafico non basta

Morti in incidenti stradali prima e dopo l'introduzione di provvedimenti severi



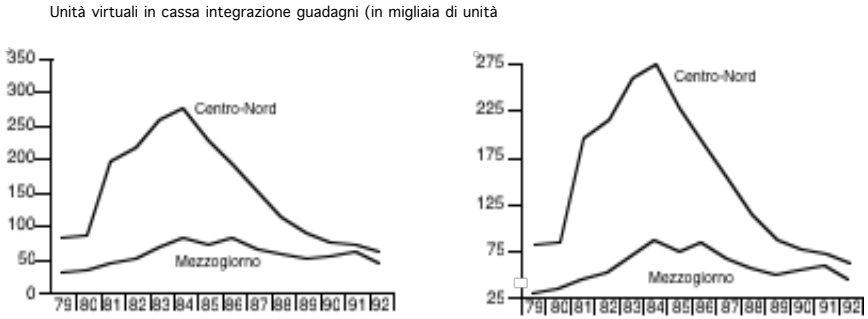
L'anno il 1993, non è stato inserito in quanto l'effetto annuncio porta ad una riduzione naturale e le sanzioni richiedono tempo prima che siano applicate.

Il cambio nel trend sembra confermare l'efficacia del provvedimento.

Qual'è l'andamento dei ritiri di patente e delle relative condanne?
Come sono variate le violazioni del limite di velocità?
Ci sono state innovazioni nei sistemi di sicurezza o nella legislazione o negli incentivi a poliziotti e magistrati?

Scelta della scala

Lo sfruttamento dell'area del grafico impone di collocare l'inizio dell'asse quantitativo sulla modalità minima e la fine sul valore massimo.



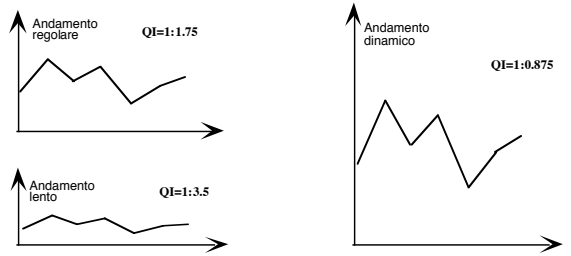
Il grafico a destra impegna più efficacemente l'area del disegno che nel primo rimane in parte inutilizzata (ha cioè un peggiore rapporto DT).

Appare quindi più chiara la marcata differenza di utilizzo della cassa integrazione tra il C-N ed il Sud.

Unità di misura degli assi

In ogni grafico si deve scegliere la scala cioè stabilire il rapporto di tra unità di misura del fenomeno e unità di lunghezza degli assi.

Q.I.=b/h



Se si modifica l'altezza o la base o entrambe si modifica l'orientamento dei segmenti del profilo cioè l'angolo che essi formano con l'asse temporale.

Se il quoziente immagine tende a zero l'inclinazione di tutti i segmenti tende a zero con un appiattimento sull'asse delle ascisse; se tende ad infinito tutti i segmenti tenderanno a disporsi verticalmente.

In entrambi i casi sarà molto difficile cogliere delle differenze di inclinazioni e quindi percepire le tendenza evolutiva della serie storica.

La scala logaritmica

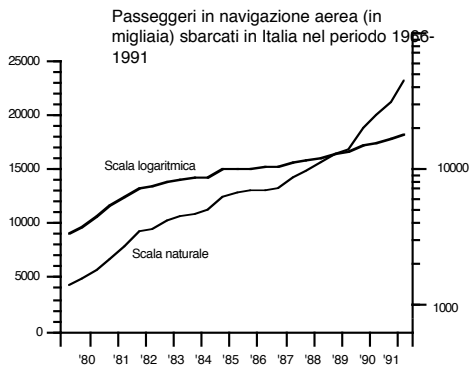
Si usa se una serie ha sia valori molto piccoli che molto grandi.

Evidenzia una forte crescita (poi esurita) nei primi anni '80 che non è ben percepibile in scala naturale.

L'asse delle ordinate -sul lato destro- riporta i logaritmi decimali dei valori.

Tale asse non può essere allineato allo zero in scala naturale (il logaritmo di zero non esiste).

per valori inferiori all'unità, i logaritmi sono negativi.



Il logaritmo ha un effetto telescopico: ingrandisce le differenze piccole e rimpiccolisce quelle grandi.

Scala logaritmica/2

Si realizza sostituendo ai valori (purché positivi) i loro logaritmi: $Y_i = Log(X_i)$ laddove sull'asse delle ascisse si riportano in scala naturale i periodi di rilevazione (diagramma semilogaritmico).

In scala naturale tra 6000 e 6001 c'è la stessa differenza che c'è tra 12000 e 12001;

In scala logaritmica, ad uno stesso segmento sulle ordinate, corrisponde uguaglianza di rapporti:

$$(1.5\text{-}1.4)=(2.8\text{-}2.6) \dashrightarrow (2^{1.5}/2^{1.4})=(2^{2.8}/2^{2.7})$$

Quindi l'uguaglianza dei due rapporti.

$$Log(X_1) - Log(X_2) = Log(X_3) - Log(X_4)$$

$$Log\left(\frac{X_1}{X_2}\right) = Log\left(\frac{X_3}{X_4}\right)$$

Esempio

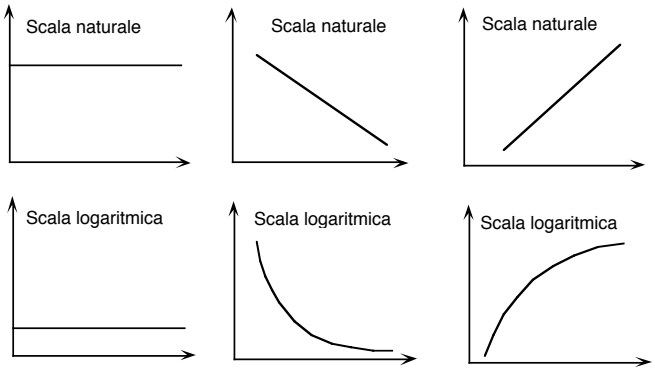
Fatturato del settore della comunicazione d'impresa e della Data Marketing Italia, s.r.l.

Nel diagramma in scala logaritmica risalta l'andamento della D.M.I. che sembra crescere ad un ritmo regolare: gli scarti tra ordinate sono pressoché costanti da un anno all'altro.

Anche il settore aggregato ha un ritmo di crescita regolare, ma nettamente superiore a quello della D.M.I.

Effetto di smussamento

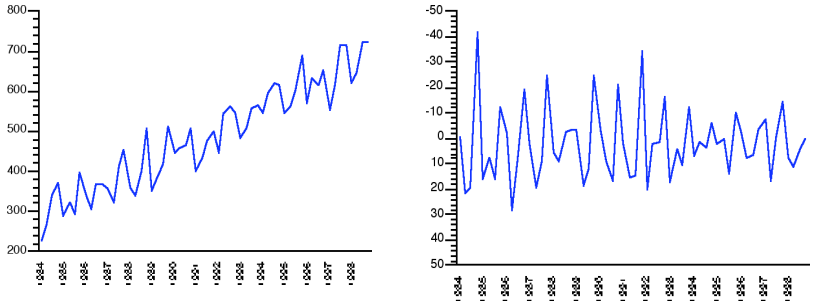
La trasformazione logaritmica modifica l'andamento della serie storica facendo apparire lineare un andamento curvilineo e viceversa.



Questo effetto può essere sfruttato per presentare un andamento tormentato da picchi e valli con un profilo più dolce o con oscillazioni più smorzate portando in primo piano il gradiente naturale del fenomeno.

Esempio

Andamento trimestrale del saldo di cassa di un'impresa. La serie mostra una crescita persistente e fluttuazioni regolari



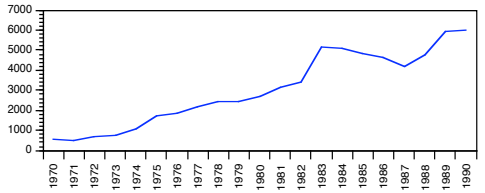
Nel grafico a destra, tutti i valori tranne l'ultimo sono stati sostituiti da: 100*[Log(Y_t/Y_{t-1})]. La nuova rappresentazione elimina entrambi gli effetti anche se compaiono imprevedibili oscillazioni smorzate.

La differenza in scala logaritmica ha di solito effetti stabilizzanti sull'andamento della serie storica anche se, di tanto in tanto, inserisce effetti artificiali il cui impatto non è sempre positivo ai fini dell'analisi.

Scelta della base

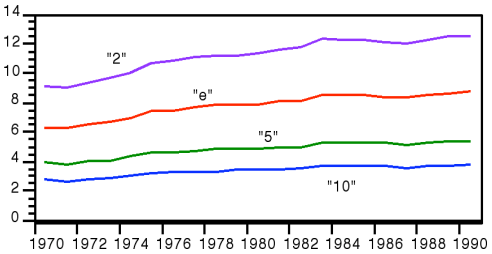
$$Log_a(x) = \frac{Log_b(x)}{Log_b(a)}$$

i logaritmi possono essere: decimali, binari, naturali.



Valore	Log_{10}	Log_5	Log_e	Log_2
1000	3	4.98	6.91	9.97
100	2	3.32	4.61	6.64
10	1	1.66	2.30	3.32
1	0	0.00	0.00	0.00
0.1	-1	-1.66	-2.30	-3.32
0.01	-2	-3.32	-4.61	-6.64
0.001	-3	-4.98	-6.91	-9.97

Anno	Int. Str.	Anno	Int. Str.
1970	571.9	1981	3153
1971	520.6	1982	3387
1972	701.9	1983	5159
1973	782.5	1984	5094
1974	1085.3	1985	4827
1975	1704.5	1986	4634
1976	1834.8	1987	4176
1977	2198.2	1988	4787
1978	2429.7	1989	5914
1979	2463.9	1990	6035
1980	2695.4		

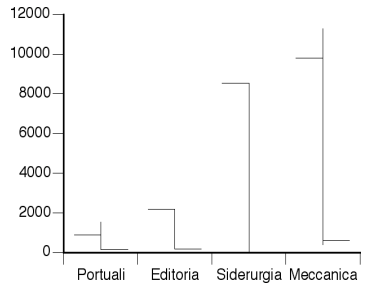


Quelli binari funzionano meglio

Diagrammi start/stop

Si scelgono dei valori rappresentativi di ogni serie storica: primo, ultimo, massimo e minimo che sono i soli ad essere riportati in grafico

Settore	Inizio	Massimo	Minimo	Fine
Portuali	877	1'521	125	134
Editoria	2'184	2'184	184	184
Siderurgia	8'537	8'537	16	16
Meccanica	9'768	11'249	374	595



Ogni serie diventa una linea verticale che si prolunga dal valore più piccolo a quello più grande.

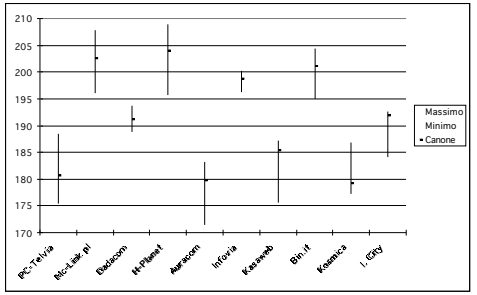
La linea è poi tagliata da due tratti orizzontali di ampiezza fissa i cui punti di attacco sono il valore iniziale ed il valore finale.

Tale diagramma può operare bene per dati borsistici (è presene come tale in Excel)

Grafico azionario

Talvolta, si desidera presentare un dato proiettandolo nel campo di variazione.

Fornitore	Massimo	Minimo	Canone
PC-Telvia	188.5	175.4	180.6
Mc-Link pl	196.1	207.9	202.5
Dadacom	193.7	188.9	191.2
H-Planet	208.9	195.7	204.0
Auracom	183.2	171.5	179.8
Infovia	200.2	196.3	198.7
Kasaweb	187.2	175.6	185.4
Bin.it	204.3	195.0	201.1
Kosmica	186.8	177.2	179.3
I. City	192.6	184.1	191.9



Questo è uno dei grafici azionari presente nell'Excel, facilmente estendibile alle temperature, ai tassi bancari, ai saldi di conto corrente, etc.

Diagramma a candela

i valori: iniziale, alto, basso e finale sono espressi da un rettangolo *(box)* di altezza pari al campo di variazione e di base fissa collocato a intervalli regolari.

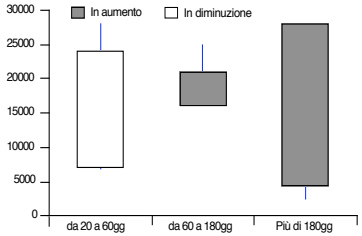
Se il valore iniziale è più piccolo di quello finale la serie è considerata in aumento ed è presentata con una campitura diversa da quelle in diminuzione;

La serie è ritenuta in diminuzione se il valore di chiusura è inferiore a quello di apertura.

Lo stoppino della candela sporge in alto -per un tratto di lunghezza fissa- se il valore finale è inferiore a quello massimo o in basso se il valore iniziale è superiore a quello minimo;

Se entrambe le configurazioni sono presenti, la candela avrà due stoppini.

Esempio:
Giustizia lenta? Tempi intercorrenti tra la data di emissione del decreto che dispone il giudizio e la data dell'udienza. Numero di decreti; periodo dal 1990 al 1997.



E' sconfortante che siano in diminuzione i tempi ridotti ed in aumento quelli lunghi.