

Concentrazione

- Nello studio della distribuzione della ricchezza, è di fondamentale importanza l’aspetto della ***concentrazione***. Intuitivamente, la concentrazione è elevata quando poche unità della popolazione possiedono gran parte della ricchezza. La concentrazione è minima (equidistribuzione) quando tutte le unità hanno la stessa ricchezza.
- Un *carattere quantitativo trasferibile*, le cui modalità ordinate sono y_1, \dots, y_N , si dice ***equidistribuito*** se ognuna delle N unità possiede una quota dell’ammontare del carattere pari a

$$\frac{1}{N}A, \quad \text{dove} \quad A = \sum_{i=1}^N y_i,$$

che coincide con la media aritmetica μ .

- Se non c’è equidistribuzione allora si ha ***concentrazione***.
- Si ha ***massima concentrazione*** quando una sola unità del collettivo possiede tutto l’ammontare del carattere e tutte le altre nulla, cioè

$$y_1 = \dots = y_{N-1} = 0 \quad \text{e} \quad y_N = A.$$

- ***Esempio***: si hanno 100 soggetti e l’ammontare complessivo del reddito mensile è $A = 50.000\text{€}$. Se c’è *equidistribuzione* ogni soggetto ha reddito pari a 500€ mentre nel caso di *massima concentrazione* un solo soggetto ha reddito pari a 50.000€ e gli altri soggetti non hanno reddito.

Misurazione della concentrazione ***(distribuzioni unitarie)***

- ***Ammontare del carattere*** posseduto dalla i unità “più povere”: dopo aver ordinato i termini della distribuzione ($y_1 \leq y_2 \leq \dots \leq y_N$)

$$A_i = y_1 + \dots + y_i = \sum_{j=1}^i y_j$$

- ***Ammontare relativo del carattere*** posseduto dalla i unità “più povere”:

$$Q_i = \frac{A_i}{A} = \frac{\sum_{j=1}^i y_j}{\sum_{j=1}^N y_j}$$

- ***Ammontare relativo del carattere*** posseduto dalla i unità “più povere” nel caso (*ipotetico*) di *equidistribuzione*:

$$P_i = \frac{i}{N}$$

- Per *qualsiasi distribuzione* si ha: $P_i \geq Q_i, \forall i$, e $P_N = Q_N = 1$

- All’aumentare della concentrazione aumentano le differenze: $P_i - Q_i$

- Nel caso di *massima concentrazione* si ha: $Q_1 = \dots = Q_{N-1} = 0$

- Per avere un indice sintetico si usa il **rapporto di concentrazione di Gini** che si ottiene come rapporto tra $\sum_{i=1}^{N-1} (P_i - Q_i)$ e il suo valore massimo:

$$G = \frac{\sum_{i=1}^{N-1} (P_i - Q_i)}{\sum_{i=1}^{N-1} P_i} = 1 - \frac{\sum_{i=1}^{N-1} Q_i}{\sum_{i=1}^{N-1} P_i}$$

- L’indice di Gini cresce al crescere del livello di concentrazione ed è sempre compreso tra 0 (nel caso di *equidistribuzione*) e 1 (nel caso di *massima concentrazione*).
- Un altro strumento che permette di valutare il grado di concentrazione è la **curva di Lorenz**. Si tratta di un grafico ottenuto unendo con dei segmenti i punti di coordinate (P_i, Q_i) , per $i = 1, \dots, N$. Maggiore è l’area tra la curva di Lorenz e la bisettrice, maggiore è la concentrazione.
- Dal grafico della curva di Lorenz si può ricavare una ulteriore misura di concentrazione, denominata **area di concentrazione**, strettamente legata al rapporto di concentrazione di Gini. Questa è data dall’area compresa tra la curva di concentrazione e la retta di equidistribuzione:

$$A_c = \frac{1}{2} - \frac{1}{2} \left[\sum_{i=1}^N (P_i - P_{i-1})(Q_i + Q_{i-1}) \right] \quad , \quad F_0 = Q_0 = 0$$

$$\text{Area di concentrazione massima} : \frac{1}{2} - \left(\frac{1}{N} \cdot 1 \cdot \frac{1}{2} \right) = \frac{N-1}{2N}$$

$$G = \frac{N}{N-1} 2A_c$$

Esempio

- Per un gruppo di 5 soggetti si ha la seguente distribuzione del reddito mensile

Unità (i)	Reddito (x_i)
1	250
2	650
3	3000
4	750
5	350
Totale	5000

- Non c’è equidistribuzione e quindi c’è concentrazione. Per quantificarne il livello si ordinano prima le modalità ottenendo

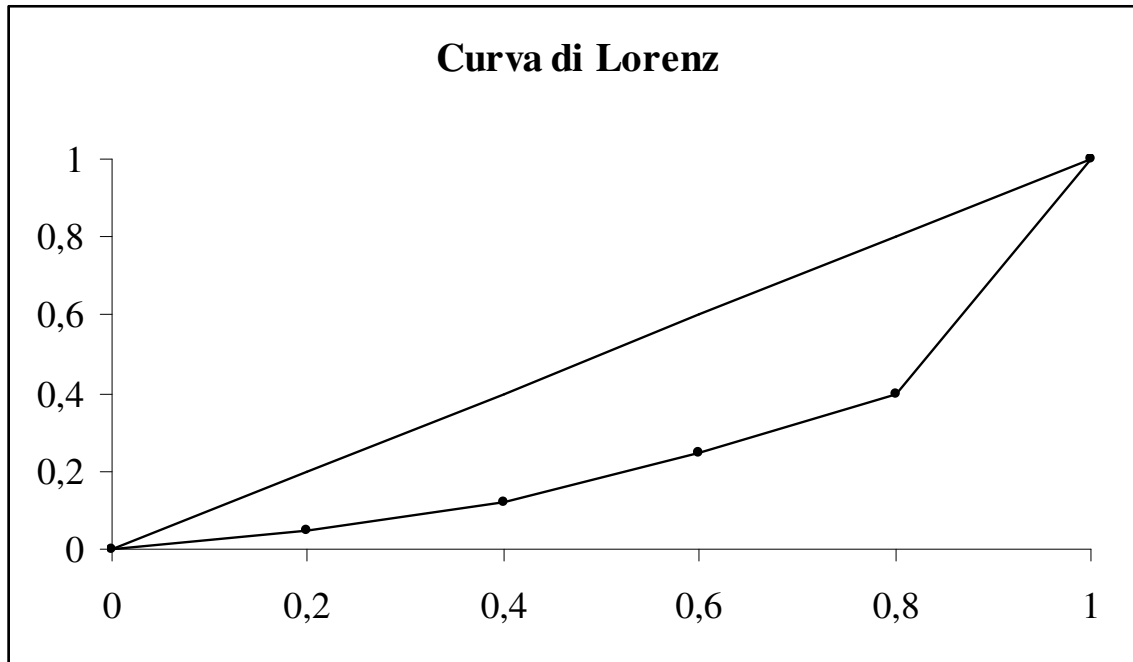
Reddito (y_i)	A_i	Q_i	P_i	$P_i - Q_i$
250	250	*	0,2	0,15
350	600	0,12	0,4	0,28
650	1250	0,25	0,6	0,35
750	2000	0,40	0,8	0,40
3000	5000	1,00	1,0	0,00
5000	–	–	2	1,18

da cui l’indice di Gini è pari a $G = 1,18/2 = 0,59$

mentre l’area di concentrazione $A_C = 0,236$

Dalla quale è possibile ricavare l’indice di Gini attraverso la formula sopra

citata $G = \frac{N}{N-1} 2A_c$ che dà appunto come risultato 0,59.



Misurazione della concentrazione (distribuzioni di frequenza)

- ***Ammontare del carattere*** posseduto dalla i modalità “più povere”: dopo aver ordinato i caratteri della distribuzione ($x_1 \leq x_2 \leq \dots \leq x_K$), dove K è il numero di modalità assunte dalla variabile statistica

$$A^*_i = \sum_{j=1}^i x_j f_j$$

- ***Ammontare relativo del carattere*** posseduto dalla i modalità “più povere”:

$$Q^*_i = \frac{A^*_i}{A^*} = \frac{\sum_{j=1}^i x_j f_j}{\sum_{j=1}^K x_j f_j} = \frac{\sum_{j=1}^i x_j f_j}{\mu}$$

- ***Ammontare relativo del carattere*** posseduto dalla i modalità “più povere” nel caso (*ipotetico*) di *equidistribuzione*:

$$P^*_i = \sum_{j=1}^i f_j$$

- Per *qualsiasi distribuzione* si ha: $P^*_i \geq Q^*_i, \forall i$, e $P^*_K = Q^*_K = 1$
- All’aumentare della concentrazione aumentano le differenze: $P^*_i - Q^*_i$
- Nel caso di *massima concentrazione* si ha: $Q^*_1 = \dots = Q^*_{K-1} = 0$

- Per avere un indice sintetico si usa il **rapporto di concentrazione di Gini** che si ottiene come rapporto tra $\sum_{i=1}^{K-1} (P_i^* - Q_i^*)$ e il suo valore massimo:

$$G^* = \frac{\sum_{i=1}^{K-1} (P_i^* - Q_i^*)}{\sum_{i=1}^{K-1} P_i^*} = 1 - \frac{\sum_{i=1}^{K-1} Q_i^*}{\sum_{i=1}^{K-1} P_i^*}$$

- L’indice di Gini cresce al crescere del livello di concentrazione ed è sempre compreso tra 0 (nel caso di *equidistribuzione*) e 1 (nel caso di *massima concentrazione*).
- Un altro strumento che permette di valutare il grado di concentrazione è la **curva di Lorenz**. Si tratta di un grafico ottenuto unendo con dei segmenti i punti di coordinate (P_i^*, Q_i^*) , per $i=1, \dots, K$. Maggiore è l’area tra la curva di Lorenz e la bisettrice, maggiore è la concentrazione. Sulla base della curva di Lorenz è possibile calcolare l’area di concentrazione che, nel caso di distribuzioni di frequenza, vale:

$$A_c^* = \frac{1}{2} - \frac{1}{2} \left[\sum_{i=1}^k (P_i^* - P_{i-1}^*) (Q_i^* + Q_{i-1}^*) \right], \quad P_0^* = Q_0^* = 0$$

Esempio

- Per un gruppo di 100 soggetti si ha la seguente distribuzione del reddito mensile

Modalità (x_i)	Frequenze (f_i)
200	30%

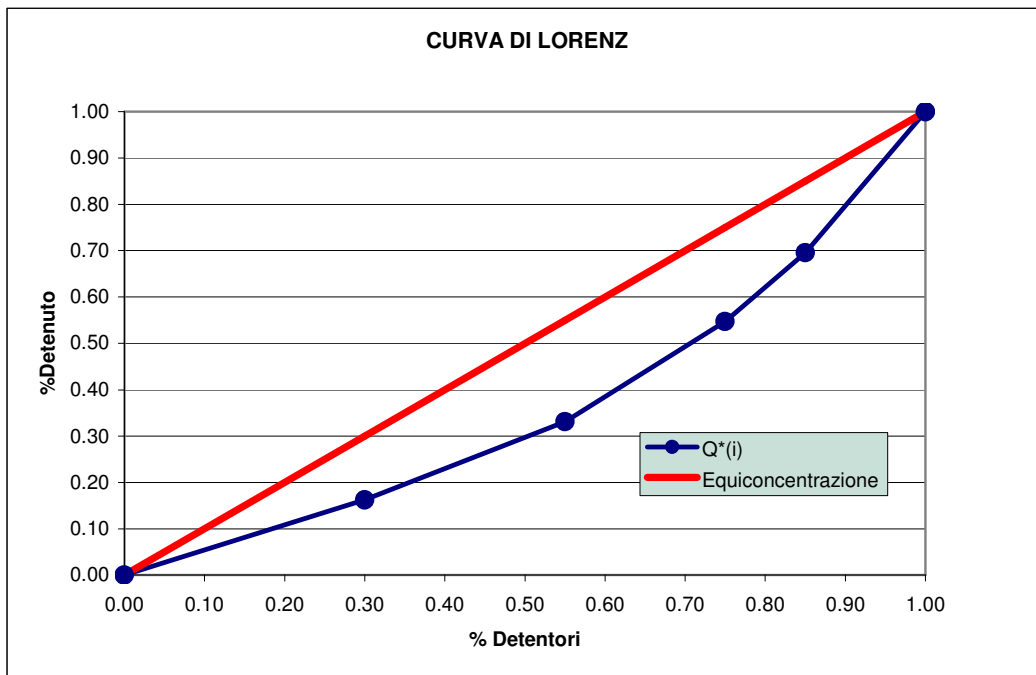
Lez. 6 – Analisi dei Dati – A. Mira – Università dell’Insubria

250	25%
400	20%
550	10%
750	15%

- Non c’è perfetta equidistribuzione e quindi c’è concentrazione. Si ricorda che le sommatorie sulle ultime due colonne coinvolgono le sole prime (k-1) modalità:

Reddito (x_i)	A_i^*	Q_i^*	P_i^*	$P_i^* - Q_i^*$
200	60	0,16	0,30	0,14
250	122,5	0,33	0,55	0,22
400	202,5	0,55	0,75	0,20
550	257,5	0,70	0,85	0,15
750	370	1,00	1,00	0,00
			2,45	0,71

da cui l’indice di Gini è pari a $G^* = 0,71/2,45 = 0,29$

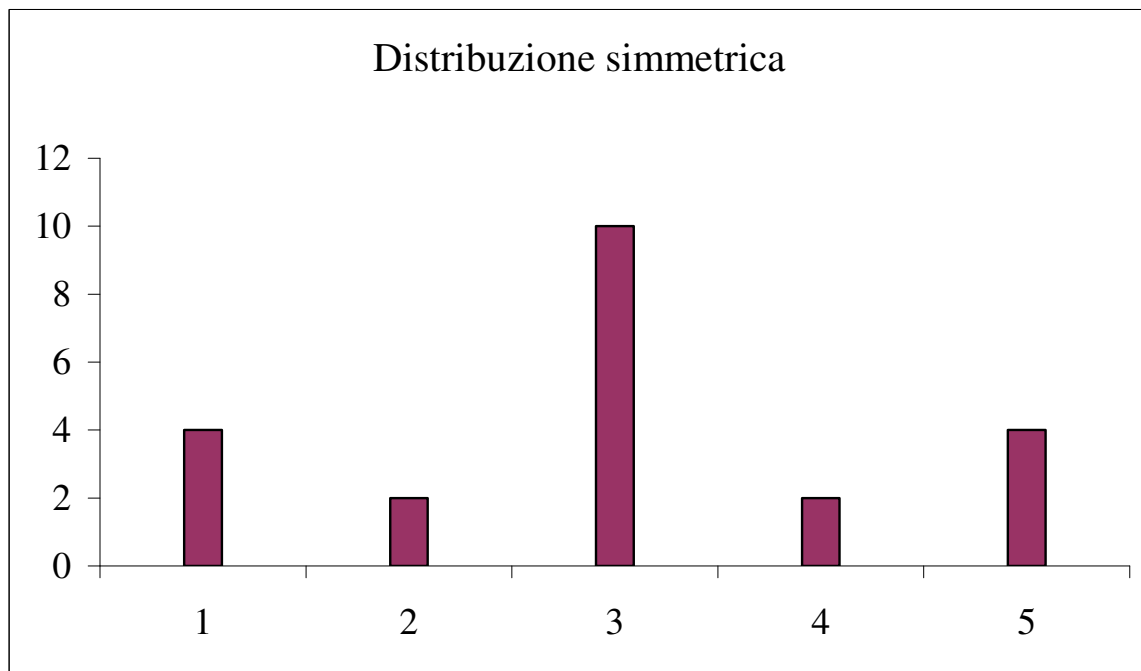


Asimmetria

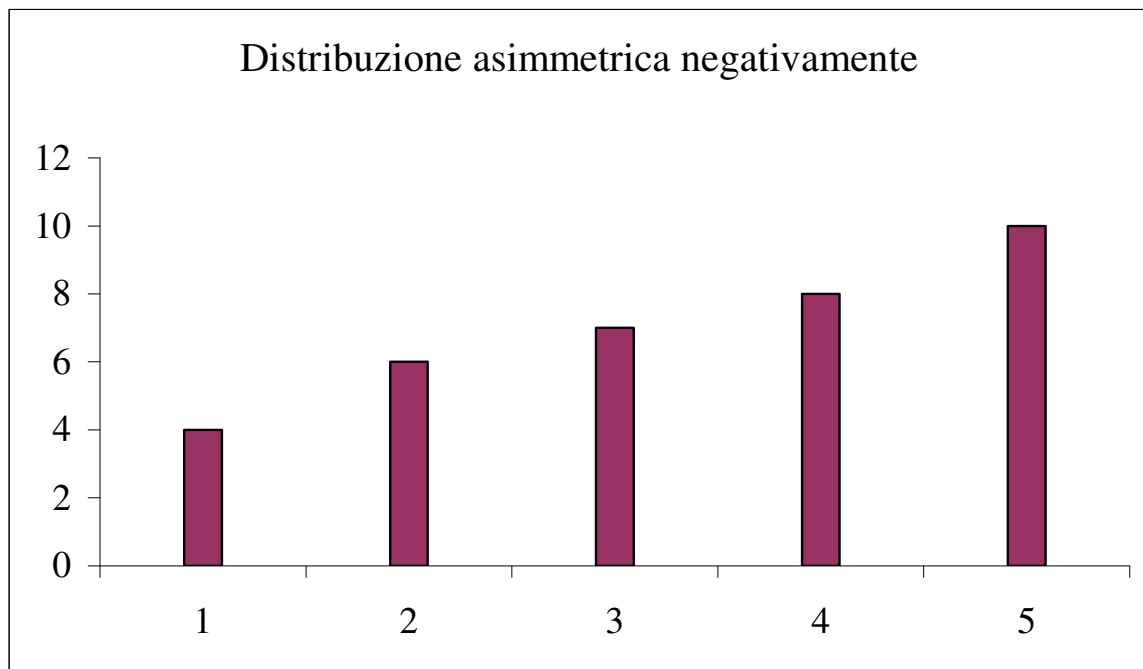
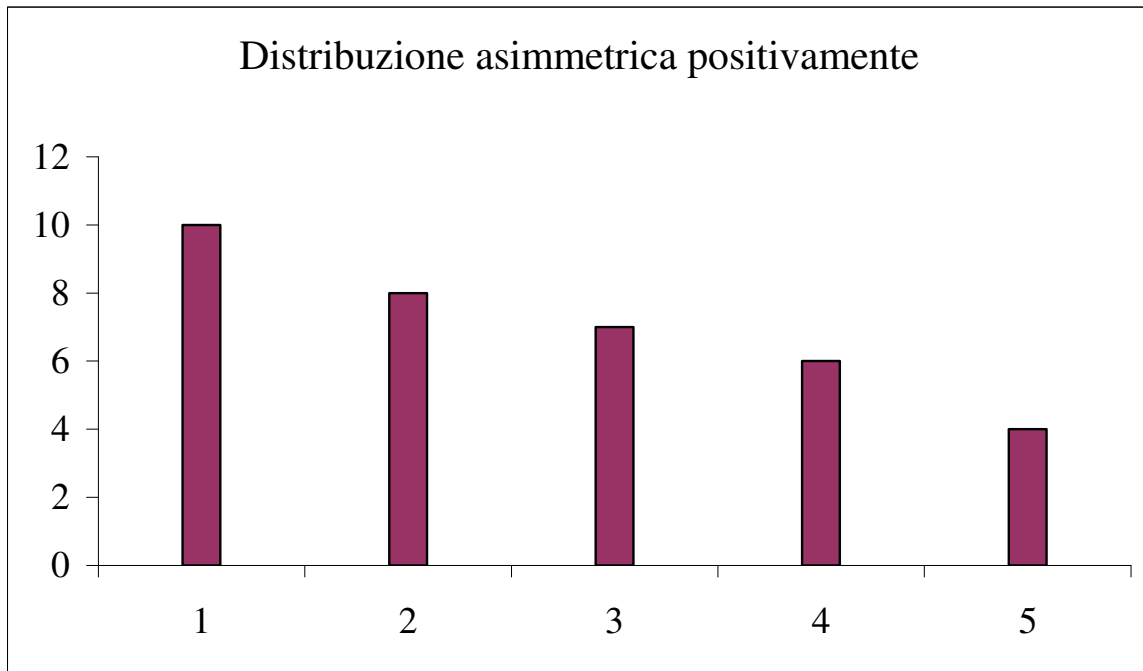
- Se Media=Moda=Mediana
- Una distribuzione di frequenza con *mediana* m è *simmetrica* se:

➤ $|x_1 - m| = |x_k - m|, |x_2 - m| = |x_{k-1} - m|, |x_3 - m| = |x_{k-2} - m|, \dots$

➤ $n_1 = n_k, n_2 = n_{k-1}, n_3 = n_{k-2}, \dots$



- Una distribuzione si dice asimmetrica se le condizioni precedenti non sono rispettate. In particolare si può avere:
 - *asimmetria positiva*: se sono più frequenti le modalità piccole, in generale (ma non sempre) risulta che $\text{moda} < \text{mediana} < \text{media}$
 - *asimmetria negativa*: se sono più frequenti le modalità più grandi, in generale (ma non sempre) risulta che $\text{moda} > \text{mediana} > \text{media}$



Misurazione dell’asimmetria

- L’*indice di asimmetria* più utilizzato è quello di *Fisher*:

$$\alpha_1 = \frac{1}{\sigma^3} \left[\frac{1}{N} \sum_{i=1}^k (x_i - \mu)^3 n_i \right]$$

- Si noti che se la distribuzione di frequenza è simmetrica si ha:

$$m - q_1 = q_3 - m$$

e quindi si può costruire un *indice di asimmetria* basato su statistiche d’ordine:

$$\alpha_2 = \frac{(q_3 - m) - (m - q_1)}{q_3 - q_1}$$

- Usualmente, quando gli indici sono maggiori di 0 si ha *asimmetria positiva* e quando sono negativi si ha *asimmetria negativa*.

Esempi

- Per la prima distribuzione del numero di figli, che ha media $\mu = 2,04$ e deviazione standard $\sigma = 0,871$, si ha:

N. figli (x_i)	Frequenze (n_i)	$(x_i - \mu)^3 n_i$
0	1	-8,4897
1	4	-4,4995
2	15	-0,0010
3	3	2,6542
4	2	15,0591
Totale	25	4,7232

da cui

$$\alpha_1 = \frac{1}{0,871^3} \left(\frac{4,723}{25} \right) = 0,286$$

che indica una leggera asimmetria positiva.

- Per la distribuzione dell’altezza per un collettivo di 50 soggetti in cui la mediana è 174, il primo quartile è 170,43 e il terzo quartile è 177,57, si ha:

$$\alpha_2 = \frac{(177,57 - 174) - (174 - 170,43)}{177,57 - 170,43} = 0$$

che indica la presenza di simmetria.

Serie storiche

- Si definisce *serie storica* una sequenza di osservazioni, relative a un certo fenomeno, effettuate in T tempi (mesi, anni, etc.):

t	y_t
1	y_1
2	y_2
\vdots	\vdots
T	y_T

Esempio

- Serie storica delle *richieste di cittadinanza* in Italia da parte di cittadini stranieri

Anno (t)	Richieste (y_t)
1996	8.931
1997	11.633
1998	10.780
1999	13.648
2000	11.566

Analisi di serie storiche

- Un’analisi preliminare può essere basata su un **grafico** che consiste nel rappresentare su un piano cartesiano punti di coordinate (t, y_t) che poi vengono congiunti con dei segmenti. In questo modo è possibile intuire l’*andamento* del fenomeno.

Esempio

- Per la serie storica delle *richieste di cittadinanza* in Italia

