Corso di Statistica
Facoltà di Economia

a.a. 2010-2011

La concentrazione

Lezione n° 7

Sommario

- Quando studiarla?
- Obiettivo
- Diagramma di Lorenz
- Rapporto di concentrazione
- Area di concentrazione
- Esempi

Quando studiarla?

X: carattere quantitativo *Trasferibile*

tra le unità statistiche

Es: il reddito è un carattere trasferibile (tra gli individui), come lo sono il patrimonio, il numero di azioni di una certa azienda (tra gli azionisti) e i finanziamenti ricevuti dalle regioni italiane (tra le regioni stesse).

Non sono invece trasferibili caratteri come l'età o l'intensità delle precipitazioni da una zona all'altra.

Obiettivo

- Interessa misurare se e quanto il carattere risulta *equamente suddiviso* tra le unità del collettivo o invece risulta *concentrato* in poche unità statistiche.

•Chiaramente ogni situazione reale sarà intermedia tra due situazioni estreme che chiameremo di equiripartizione e di concentrazione massima.

Casi estremi

–***Equidistribuzione***: *tutte le unità statistiche possiedono la stessa quantità del carattere* X

Unità statistiche: *1, 2, ……, n-1, n*
Quantità posseduta: $\mu, \mu, \ldots\ldots, \mu, \mu$

–***Massima concentrazione***: Il carattere è posseduto nella sua totalità da una sola unità statistica:

Unità statistiche: *1, 2, ……, n-1, n*
Quantità posseduta: 0, 0, ……, 0, $\sum_{i=1}^{N} x_i = N\mu$

Tre modi per valutare la concentrazione

- **Curva di concentrazione (o di Lorenz-Gini)**
- **Rapporto di concentrazione**
- **Area di concentrazione**

Curva di concentrazione

1. Ordiniamo le quantità del carattere (intensità) possedute dalle unità statistiche in senso non decrescente:

$$0 \le x_1 \le x_2 \le \ldots \le x_N$$

2. Definiamo:

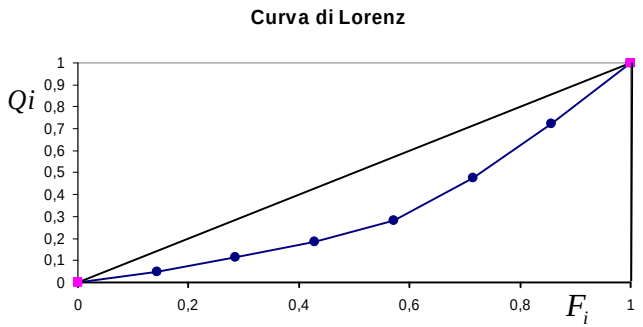
$F_i = \frac{i}{N}$ (frequenze cumulate delle i unità più povere)

$Q_i = \frac{\sum_{j=1}^{i} X_j}{\sum_{j=1}^{N} X_j} = \frac{\sum_{j=1}^{i} X_j}{N\mu}$ (quantità relative di carattere delle i unità più povere)

Curva di concentrazione

Rappresentiamo le le coppie (F_i, Q_i) in un grafico cartesiano e congiungiamole. La spezzata così ottenuta è la curva di concentrazione.

Esempio: N=7



Curva di concentrazione

Alcune osservazioni:

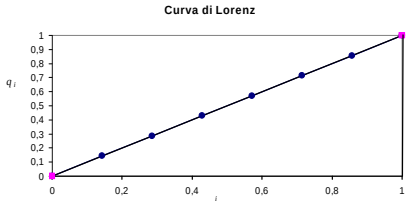
- La spezzata parte da $(F_0, Q_0) = (0,0)$ e termina in $(F_N, Q_N) = (1,1)$
- Essendo $Q_i \le F_i$ la spezzata è sempre al di sotto della bisettrice. La bisettrice è tale per cui $F_i = Q_i \ \forall i$ e rappresenta quindi il caso di equiripartizione: quanto più la curva si discosta dalla bisettrice tanto maggiore è la concentrazione del carattere.

•Il generico punto (Fi, Qi) della spezzata si può interpretare nel seguente modo: l' (Fi *100)% *più povero* di carattere possiede il (Qi*100)% del carattere totale.

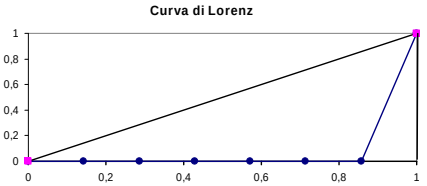
• Ad esempio se il grafico precedente fosse riferito al reddito avremmo che il 60% più povero della popolazione possiede circa il 30% del reddito totale.

Il rapporto di concentrazione

- Intuitivamente la distanza verticale tra la bisettrice e la spezzata di concentrazione, pari a $(F_i - Q_i)$, aumenta all'aumentare della concentrazione del carattere…



1) Equiripartizione (N=7)



2) Massima concentrazione (N=7)

Il rapporto di concentrazione

Concentriamoci quindi su tale distanza:

Abbiamo visto che in generale vale $F_i - Q_i \ge 0$ i = 1, 2, ..., N

Mentre:

- In caso di equidistribuzione:

  $F_i - Q_i = 0$ i = 1,2,...,N

- In caso di massima concentrazione:

  $F_i - Q_i = F_i$ i = 1,2,...,N-1 e $F_N - Q_N = 0$.

- Un modo naturale per misurare la concentrazione è quindi quello di costruire un indice basato sulla somma di tali differenze...

Il rapporto di concentrazione

- la somma delle differenze (Fi-Qi), divisa per il valore massimo che tali differenze possono assumere è nota come rapporto di concentrazione di Gini:

$$R = \frac{\sum_{1}^{N-1}(F_i - Q_i)}{\sum_{i=1}^{N-1} F_i} = \frac{2}{N-1}\sum_{1}^{N-1}(F_i - Q_i)$$

•Si ha $0 \le R \le 1$ in particolare vale 0 nel caso di equiripartizione ed 1 nel caso di massima concentrazione.

Osservazione: esiste un legame intuitivo tra concentrazione e la dispersione di un carattere. Infatti più un carattere è concentrato maggiore è la sua dispersione. Al contrario nel caso di equiripartizione il carattere presenta una sola modalità: dispersione nulla. In effetti si può mostrare che vale la seguente relazione:

$$R = \frac{\Delta}{2\mu} = \frac{N}{N-1}\frac{\Delta_R}{2\mu}$$

Quindi R può essere visto anche come un indice di variabilità relativa, in quanto corrisponde alla differenza semplice media divisa per il suo massimo (2μ).

Esempio

Data la successione di modalità di X:

80	90	21	23	32	16	62

Pe calcolare R anzitutto si ordinano le x_i in senso non decrescente (ad esempio X_1=16 perché l'unità più povera possiede 16) e poi si trovano i corrispondenti valori di F_i e Q_i (ad esempio F_1=1/7=0.1429 e Q_1=16/324=0,0494 ovvero: l'unità più povera rappresenta il 14% del totale e possiede circa il 5% del carattere totale).

i	x_i	F_i	Q_i	(F_i-Q_i)
1	16	0,1429	0,0494	0,0935
2	21	0,2857	0,1142	0,1715
3	23	0,4286	0,1852	0,2434
4	32	0,5714	0,2840	0,2875
5	62	0,7143	0,4753	0,2390
6	80	0,8571	0,7222	0,1349
7	90			
Totale	324	3		1,1698

$$R=\frac{\sum_{i=1}^{N-1}F_i-Q_i}{\sum_{i=1}^{N-1}F_i}=\frac{1{,}1698}{3}=0{,}3899$$

Esempio (continua)

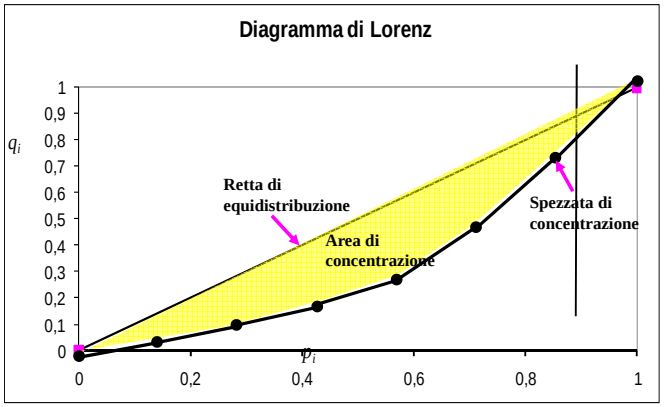
Si noti che: $R=\frac{\sum_{1}^{N-1}(F_i-Q_i)}{\sum_{i=1}^{N-1}F_i}=\frac{2}{N-1}\sum_{1}^{N-1}(F_i-Q_i)=1-\frac{2}{N-1}\sum_{1}^{N-1}Q_i$

e quindi si può calcolare il valore di R a partire dalle sole intensità cumulate

i	x_i	Q_i
1	16	0,0494
2	21	0,1142
3	23	0,1852
4	32	0,2840
5	62	0,4753
6	80	0,7222
7	90	1,0000
	324	**1,8302**

$$R=1-\frac{2}{N-1}\sum_{i=1}^{N-1}Q_i=1-\frac{2}{7-1}\cdot 1.8302=0.3899$$

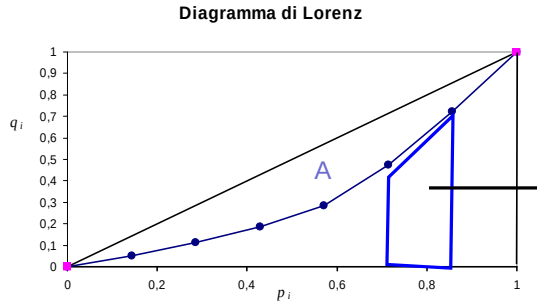
Area di concentrazione



- L'area tra la curva di concentrazione e la retta di equidistribuzione prende il nome di area di concentrazione
- Nel caso di equiripartizione l'area è di concentrazione è pari a zero; al crescere della concentrazione l'area cresce senza mai superare il valore ½.

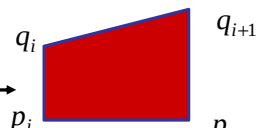
Area di concentrazione

Una misura dell'area A di concentrazione può essere ottenuta sottraendo all'area del triangolo la somma delle aree degli *n* trapezi delimitati dai punti (F_i,Q_i) per $i=0,1,\ldots,n$.



Area del trapezio:

½ x Somma delle basi x altezza



$$\frac{1}{2}\underbrace{(F_i-F_{i-1})}_{ALTEZZA}\ \underbrace{(Q_i+Q_{i-1})}_{B.MAGG+B.MINORE}$$

Area di concentrazione

- Un indice di concentrazione si può ottenere dividendo l'area di concentrazione per il valore assunto da tale area nel caso di concentrazione massima:

$$I = \frac{A}{A_{\max}} = \frac{1/2 - 1/2 \sum_{i=1}^{N-1} (F_i - F_{i-1})(Q_i + Q_{i-1})}{A_{\max}}$$

- Poiché si dimostra che:

$$A = \frac{\Delta_R}{4\mu} \qquad A_{\max} = \frac{N-1}{2N}$$

Abbiamo $I = R$.

Ovvero ritroviamo il rapporto di concentrazione di Gini che si può quindi calcolare anche dividendo l'area di concentrazione per il suo valore massimo.

Area di concentrazione

Un altro indice di concentrazione si può ottenere osservando che l'area di concentrazione massima è approssimativamente uguale ad ½ per N grande. Possiamo quindi rapportare l'area di concentrazione a questo valore:

$$\tilde{R} = \frac{1/2 - 1/2 \sum_{i=1}^{n-1} (F_i - F_{i-1})(Q_i + Q_{i-1})}{1/2} = 1 - \sum_{i=1}^{n-1} (F_i - F_{i-1})(Q_i + Q_{i-1})$$

In generale $\tilde{R} < R$ con i due indici che tendono a coincidere per N grande.

Dati raggruppati

- Finora abbiamo considerato dati non raggruppati, supponiamo ora di avere una distribuzione di frequenza, per cui abbiamo k valori distinti $x_1, ..., x_k$ ordinati in modo non decrescente con $n_1, ..., n_k$ numerosità.

- Definiamo: $F_i^* = \frac{n_1 + ... + n_i}{N} = p_1 + p_2 + ... + p_i$

$$Q_i^* = \frac{\sum_{j=1}^{i} n_j x_j}{\sum_{j=1}^{k} n_j x_j} = \frac{\sum_{j=1}^{i} n_j x_j}{N\mu} = \frac{p_1 x_1 + p_2 x_2 + ... + p_i x_i}{\mu}$$

•Si tratta chiaramente delle versioni ponderate di F_i e Q_i viste in precedenza con la differenza che ora l'indice i va riferito alle classi. Le ultime uguaglianze danno la formula per frequenze relative.

Dati raggruppati

- **Curva di concentrazione**: unendo le coppie (F_i^*, Q_i^*) si ottiene la *stessa* curva di concentrazione che si otterrebbe lavorando con i dati unitari. L'unica differenza è che ora per ottenere la spezzata si devono congiungere k punti e non più N punti. Inoltre usando la formulazione con le frequenze relative è possibile costruire la curva anche nel caso in cui si conoscono solo le frequenze relative p_i.

- **Area di concentrazione:** la curva di concentrazione è la stessa quindi il valore di $\tilde{R}$ e di R …

Dati raggruppati

- **Rapporto di concentrazione di Gini**: Calcolando il rapporto di concentrazione con le coppie (F_i^*, Q_i^*) si ottiene l'indice:

$$R^* = \frac{\sum_{i=1}^{k-1}(F_i^* - Q_i^*)}{\sum_{i=1}^{k-1} F_i^*}$$

Che è, in generale, diverso da R. Questo perché è diversa la situazione di massima concentrazione nei due casi: usando le frequenze relative la massima concentrazione si ha quando tutto il carattere è contenuto nella k-esima classe, e quindi da n_k unità statistiche (e non da una sola).
Ovviamente se si conoscono solo le frequenze relative R* è l'unico indice calcolabile.

Dati raggruppati in intervalli

• Si pensi ad un carattere continuo come il reddito: se N è grande è poco pratico costruire la curva di concentrazione (o il rapporto o l'area) a partire dai redditi individuali: si procede pertanto ad accorpare gli individui in classi di reddito. A questo punto però sorge il problema di valutare la concentrazione per una variabile continua (per intervalli).

•Il modo più semplice di procedere consiste nel "discretizzare" la variabile continua concentrando tutta la massa sui punti medi degli intervalli per poi procedere come nei casi precedenti.

• Nel caso in cui si disponga dell'ammontare totale di carattere in un intervallo si può sfruttare questa informazione per discretizzare gli intervalli in modo "coerente" con tale ammontare totale.

Es: nell'intervallo [5,15) di numerosità N_i=10 si sa che il carattere totale è pari a 60. E' preferibile discretizzare l'intervallo su X_i=6. Si noti che devono essere note le numerosità degli intervalli per poter procedere in questo modo.

Esempio:Concentrazione per distribuzioni di frequenze

Si utilizza la formula: $\tilde{R} = 1 - \sum_{i=1}^{k-1}(F_i^* - F_{i-1}^*)(Q_i^* + Q_{i-1}^*)$

Reddito $[x_i, x_{i+1})$	Numero di dipendenti N_i	Reddito Discr. X_i^D	F_i^*	$X_i^D N_i$	Q_i^*	$F_i^*-F_{i-1}^*$	$Q_i^*+Q_{i-1}^*$	$(F_i^*-F_{i-1}^*)(Q_i^*+Q_{i-1}^*)$
0	0	0	0	0	0	0.114	0.103	0.012
250-260	8	255	0.114	2040	0.103	0.143	0.339	0.048
260-270	10	265	0.257	2650	0.236	0.229	0.694	0.159
270-280	16	275	0.486	4400	0.458	0.214	1.132	0.242
280-290	15	285	0.700	4275	0.673	0.143	1.496	0.214
290-300	10	295	0.843	2950	0.822	0.114	1.769	0.202
300-320	8	310	0.957	2480	0.947	0.043	1.947	0.083
320-380	3	350	1	1050	1			
Totali	$\sum_{i=1}^{7} N_i = 70$			$\sum_{i=1}^{7} X_i^D N_i = 19845$				**0.961**

$\tilde{R}$=1-0.961=0.039

Esempio: Concentrazione per distribuzioni di frequenze

Si utilizza la formula: $\tilde{R} = 1 - \sum_{i=1}^{k-1}(F_i^* - F_{i-1}^*)(Q_i^* + Q_{i-1}^*)$

In questo caso si sfrutta un'informazione aggiuntiva:
la conoscenza dell'ammontare complessivo del carattere per ogni classe (X_i);

Fatturato	Aziende	Ammontare fatturato	N_i	F_i	X_i'	Q_i	$F_i - F_{i-1}$	$Q_{i+1}+Q_i$	$(F_{i+1}-F_i)(Q_{i+1}+Q_i)$
0	0	0	0	0.000	0	0.000	0.250	0.079	0.020
300-800	50	22500	50	0.250	22500	0.079	0.400	0.413	0.165
800-1500	80	72000	130	0.650	94500	0.333	0.200	1.005	0.201
1500-3000	40	96000	170	0.850	190500	0.672	0.150	1.672	0.251
3000-5000	30	93000	200	1.000	283500	1.000	.	.	.
Totale	200	283500							0.636772

$\tilde{R}$=1-0.636=0.364

Confronto Tra Distribuzioni

Si riportano di seguito le distribuzioni dei finanziamenti concessi da un istituto bancario per l'acquisto della prima casa a giovani coppie residenti in Campania ed in Sardegna:

Sardegna	
Finanziamenti (migliaia di Euro)	n_i
0-50	57
50-100	3
100-150	7
150-200	8
200-250	25
Totale	100

Campania	
Finanziamenti (migliaia di Euro)	ni
0-50	10
50-100	25
100-150	37
150-200	22
200-250	6
Totale	100

a) In quale regione risulta più elevata la concentrazione dei finanziamenti concessi?

b) Confrontare graficamente i diversi livelli di concentrazione dei finanziamenti nelle due regioni

Concentrazione dei finanziamenti in Sardegna

Sardegna										
Finanziamenti (migliaia di Euro)	n_i	N_i	p_i	x_i	xi n_i	X_i	q_i	$p_{i+1}-p_i$	$q_{i+1}+q_i$	$(p_{i+1}-p_i)(q_{i+1}+q_i)$
0	0	0	0.000	0	0	0	0.000	0.570	0.149	0.085
0-50	57	57	0.570	25	1425	1425	0.149	0.030	0.322	0.010
50-100	3	60	0.600	75	225	1650	0.173	0.070	0.437	0.031
100-150	7	67	0.670	125	875	2525	0.264	0.080	0.675	0.054
150-200	8	75	0.750	175	1400	3925	0.411	0.250	1.411	0.353
200-250	25	100	1.000	225	5625	9550	1.000	-	-	-
Totale	100				9550					0.532

$$\tilde{R} = 1 - \sum_{i=0}^{k-1} (p_{i+1} - p_i)(q_{i+1} + q_i) = 1 - 0.532 = 0.468$$

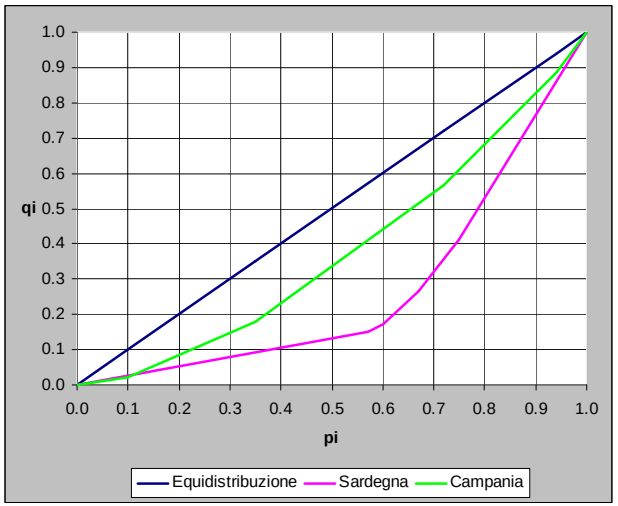
Concentrazione dei finanziamenti in Campania

Campania										
Finanziamenti (migliaia di Euro)	n_i	N_i	p_i	x_i	xi n_i	X_i	q_i	$p_{i+1}-p_i$	$q_{i+1}+q_i$	$(p_{i+1}-p_i)(q_{i+1}+q_i)$
0	0	0	0.000	0	0	0	0.000	0.100	0.021	0.002
0-50	10	10	0.100	25	250	250	0.021	0.250	0.199	0.050
50-100	25	35	0.350	75	1875	2125	0.178	0.370	0.743	0.275
100-150	37	72	0.720	125	4625	6750	0.565	0.220	1.452	0.319
150-200	22	94	0.940	175	3850	10600	0.887	0.060	1.887	0.113
200-250	6	100	1.000	225	1350	11950	1.000	-	-	-
Totale	100				11950					0.759

$$\tilde{R} = 1 - \sum_{i=0}^{k-1} (p_{i+1} - p_i)(q_{i+1} + q_i) = 1 - 0.759 = 0.241$$

$\tilde{R}_{sar} = 0.468$ $\tilde{R}_{cam} = 0.241$

La concentrazione è più elevata in Sardegna



La curva di Lorenz conferma questo risultato

Riepilogo

La concentrazione può essere calcolata solo su un carattere quantitativo trasferibile

Le due situazioni limite sono:

***Equidistribuzione*: tutte le unità statistiche possiedono la stessa quantità del carattere X**

***Massima concentrazione*: Il carattere è posseduto nella sua totalità da una sola unità statistica**

Le due quantità essenziali per il calcolo della concentrazione sono:

- F_i : *frazione sul totale delle prime i unità più povere*
- Q_i : *frazione del carattere posseduto dalle prime i* unità.

RIEPILOGO DELLE FORMULE PER IL CALCOLO DELLA CONCENTRAZIONE

$$R = \frac{\sum_{i=1}^{n-1}(F_i - Q_i)}{\sum_{i=1}^{n-1} F_i}$$

$$R = 1 - \frac{2}{n-1} \cdot \sum_{i=1}^{n-1} q_i$$

$$R = 1 - \frac{2}{(n-1) \cdot n\mu} \cdot \sum_{i=1}^{n-1} X_i$$

$$\tilde{R} = 1 - \sum_{i=0}^{n-1} (p_{i+1} - p_i)(q_{i+1} + q_i)$$

Solo per distribuzioni unitarie

RIEPILOGO DELLE FORMULE PER IL CALCOLO DELLA CONCENTRAZIONE

$$\tilde{R} = 1 - \sum_{i=0}^{n-1} (p_{i+1} - p_i)(q_{i+1} + q_i)$$ **Per distribuzioni di frequenze assolute**

DOVE $pi = \frac{N_i}{n} = F_i$ $\qquad Q_i = \frac{\sum_{j=1}^{i} x_{(j)} \cdot n_{(j)}}{\sum_{j=1}^{k} x_{(j)} \cdot n_{(j)}} = \frac{X_i}{X_k}$

$$R = \frac{\Delta}{2\mu}$$ **Per distribuzioni unitarie e di frequenze assolute**

Formule di transizione

$$\tilde{R} = \frac{n-1}{n} R \qquad R = \frac{n}{n-1} \tilde{R}$$