CORSO DI STATISTICA (parte 1) - ESERCITAZIONE 4

Dott.ssa Antonella Costanzo

a.costanzo@unicas.it

Esercizio 1. Differenze semplici medie, confronti in termini di mutua variabilità

La distribuzione del prezzo (in lire) del pane al chilogrammo nei capoluoghi di 27 province nel 1970 e nel 1989 è riportata nella seguente tabella:

Prezzo al Kg 1970	**700**	**800**	**900**	**950**	**1000**	**1200**	**Totale N**
frequenze	1	4	2	3	7	10	27
Prezzo al Kg 1982	**2100**	**2500**	**2600**	**2950**	**3000**	**3600**	**Totale N**
frequenze	2	3	2	4	6	10	27

a) Determinare la differenza semplice media con e senza ripetizione del prezzo del pane nel 1970
b) Si può dire che dal 1970 al 1989 ci sia stato un aumento della variabilità del fenomeno?
c) Calcolare il rapporto di concentrazione dei prezzi per il 1970

Sol.

In presenza di caratteri trasferibili (reddito, prezzi, consumo di beni) è di maggior interesse lo studio della variabilità tra le singole unità statistiche piuttosto che la variabilità rispetto ad un centro.

Differenza semplice media: tale indice rappresenta la media dei valori assoluti delle differenze calcolate rispetto a tutte le possibili coppie di modalità. Esso corrisponde a:

$$\Delta = \frac{\sum_{i \neq j=1}^{N} |x_i - x_j| n_i n_j}{N(N-1)}$$

La quantità a denominatore N(N-1) rappresenta il numero di possibili coppie di N osservazioni (senza ripetizione).

$\Delta = 0$ (valore minimo), quando tutte le modalità coincidono per cui le differenze semplici sono nulle

$\Delta = 2\mu$ (valore massimo), quando tutte le modalità tranne una sono nulle.

La versione normalizzata dell'indice è quindi:

$$R = \frac{\Delta}{2\mu}$$

R viene denominato rapporto di concentrazione di Gini. Il rapporto di concentrazione di Gini è un indice relativo che varia tra 0 e 1. Se $R = 0$ si ha concentrazione minima mentre se $R = 1$ si ha la concentrazione massima

a)

Per agevolare i calcoli della differenza semplice media (con o senza ripetizione) si possono costruire le seguenti tabelle nelle quali vengono calcolati rispettivamente le differenze $|x_i - x_j|$ e i prodotti $n_i * n_j$:

Matrice delle differenze (in valore assoluto)

$\|x_i - x_j\|$	700	800	900	950	1000	1200
700	0	100	200	250	300	500
800	100	0	100	150	200	400
900	200	100	0	50	100	300
950	250	150	50	0	50	250
1000	300	200	100	50	0	200
1200	500	400	300	250	200	0

Matrice dei prodotti delle frequenze

$n_i * n_j$	1	4	2	3	7	10
1	1	4	2	3	7	10
4	4	16	8	12	28	40
2	2	8	4	6	14	20
3	3	12	6	9	21	30
7	7	28	14	21	49	70
10	10	40	20	30	70	100

Differenza semplice media(senza ripetizione) con formula rapida:

$$\Delta^{1970} = \frac{2\sum_{i=i}^{k-1}\sum_{j=(i+1)}^{k}|x_i - x_j|n_i n_j}{N(N-1)}$$

Il numeratore di questa espressione può essere determinato nel modo seguente:

$\lvert x_i - x_j \rvert$	$n_i n_j$
100	4
200	2
250	3
300	7
500	10
100	8
150	12
200	28
400	40
50	6
100	14
300	20
50	21
250	30
200	70

$$\Delta^{1970} = \frac{2[(100 \times 4) + (200 \times 2) + (250 \times 3) + (300 \times 7) + (500 \times 10) + \cdots}{27 * 26} = \frac{2(63100)}{702} = 179.77$$

Tale valore indica che i prezzi del pane nei 27 capoluoghi nel 1970 differiscono mediamente tra loro di 179.77 lire.

La differenza media con ripetizione è data da:

$$\Delta_R{}^{1970} = \frac{126200}{N^2} = \frac{126200}{27^2} = 173.11$$

Tale valore indica che i prezzi del pane nei 27 capoluoghi nel 1970 differiscono mediamente tra loro (e con loro stessi) di 173.11 lire.

b)

Osservando i valori del prezzo del pane nei due anni presi in esame, ci rendiamo conto che l'ordine di grandezza è differente, ragion per cui per confrontare le variabilità dei prezzi del pane nei due anni (1970 e 1989) è necessario ricorrere a indici relativi di variabilità. In tal caso la scelta corretta è quella di confrontare la variabilità dei prezzi del 1970 e del 1989 con gli indici relativi di variabilità:

$$\Delta_{rel} = \frac{\Delta}{\mu}$$

$$\Delta_{R(rel)} = \frac{\Delta_R}{\mu}$$

Calcoliamo quindi la differenza semplice media per la distribuzione dei prezzi per l'anno 1989

Matrice delle differenze (in valore assoluto)

$\lvert x_i - x_j \rvert$	2100	2500	2600	2950	3000	3600
2100	0	400	500	850	900	1500
2500	400	0	100	450	500	1100
2600	500	100	0	350	400	1000
2950	850	450	350	0	50	650
3000	900	500	400	50	0	600
3600	1500	1100	1000	650	600	0

Matrice dei prodotti delle frequenze

$n_i * n_j$	2	3	2	4	6	10
2	4	6	4	8	12	20
3	6	9	6	12	18	30
2	4	6	4	8	12	20
4	8	12	8	16	24	40
6	12	18	12	24	36	60
10	20	30	20	40	60	100

$\lvert x_i - x_j \rvert$	$n_i n_j$
400	6
500	4
850	8
900	12
1500	20
100	6
450	12
500	18
1100	30
350	8
400	12
1000	20
50	24
650	40
600	60

$$\Delta^{1989} = \frac{2[6(400) + 4(500) + 8(850) + 12(900) + 20(1500) + 6(500) + \cdots}{N(N-1)} = \frac{381600}{27(26)} = 543.59$$

La differenza media con ripetizione è data da:

$$\Delta_R{}^{1989} = \frac{381600}{N^2} = \frac{381600}{27^2} = 523.45$$

Per il confronto riportiamo quindi gli indici relativi per ciascuno delle due distribuzioni:

$\mu_{1970} = 1020.37$

$\mu_{1989} = 3062.96$

Differenza semplice media (senza ripetizione) relativa	
Δ_{rel}^{1970}	$\frac{179.77}{1020.37} = 0.1762$
Δ_{rel}^{1989}	$\frac{543.59}{3062.96} = 0.1774$

Il valore 0.1762 indica che la differenza media semplice del prezzo del pane nel 1970 è pari al 17.62% della media aritmetica. Il valore 0.1774 indica che la differenza media semplice del prezzo del pane nel 1989 è pari al 17.74% della media aritmetica

c)

Rapporto di concentrazione per i prezzi dell'anno 1970

$$R_{1970} = \frac{\Delta}{2\mu} = \frac{179.77}{2(1020.37)} = \frac{179.77}{2040.74} = 0.088$$

Esercizio 2. Mutua variabilità e concentrazione

Sia X= reddito mensile in migliaia di euro, rilevata su un collettivo di famiglie come segue:

Reddito (X_i)	**Numero di famiglie n_i**
1	1
2	0
3	5
4	4

Quesiti:

1) Trovare lo scarto quadratico medio del reddito

2) Trovare lo scarto quadratico medio del reddito nell'ipotesi che ad ogni famiglia venga dato un aumento di stipendio di 500 euro

3) Trovare la differenza semplice media

3) Calcolare il rapporto di concentrazione per il reddito

Sol.

1)

La media del reddito è data da: $\mu = \frac{1*1+2*0+3*5+4*4}{10} = 3.2$

quindi: $\sigma = \sqrt{\frac{(1-3.2)^2*1+(2-3.2)^2*0+(3-3.2)^2*5+(4-3.2)^2*4}{10}}$=0.87

2)

Ricordando che dati a e b costanti vale la seguente proprietà per la varianza: $\sigma^2_{a+bX} = b^2\sigma^2_X$ si avrà che:

$\sigma_{a+bX} = |b|\sigma_X$ =$\sqrt{500}$*0.87

3)

La differenza semplice media (senza ripetizione) è data da $\Delta = \frac{\sum_{i=1}^{k}\sum_{j=1}^{k}|x_i-x_j|n_i*n_j}{N(N-1)}$

$|X_i - X_j|$

	X_j			
X_i	**1**	**2**	**3**	**4**
1	0	1	2	3
2	1	0	1	2
3	2	1	0	1
4	3	2	1	0

$n_i * n_j$

	n_j			
n_i	**1**	**0**	**5**	**4**
1	1	0	5	4
0	0	0	0	0
5	5	0	25	20
4	4	0	20	16

A questo punto il numeratore della differenza media semplice si ottiene moltiplicando elemento per elemento le due tabelle precedenti e sommando i prodotti ottenuti (vedi metodo rapido come nell'es.1)

$$\Delta = \frac{2[(0 \times 1) + (1 \times 0) + \cdots}{10(10-1)} = 0.993$$

4) Il rapporto di concentrazione per il reddito è dato da $R = \frac{\Delta}{2*\mu}$ ovvero come rapporto fra la differenza semplice media e il valore che tale indice di variabilità assume nel caso di massima concentrazione (il doppio della media aritmetica). Nel nostro caso, il rapporto di concentrazione è dato da:

$$R = \frac{0.993}{2 * 3.2} = 0.146$$

Esercizio 3. La concentrazione per distribuzioni di frequenza

Promemoria:

La concentrazione rappresenta un aspetto particolare della variabilità di un carattere statistico. In generale lo studio della concentrazione consente di valutare il modo in cui un carattere si distribuisce tra le unità statistiche del collettivo (es. ricchezza di un paese distribuita su n abitanti).

Lo studio della concentrazione è chiaramente interessante nel caso di caratteri trasferibili, al fine di valutare il grado di vicinanza/lontananza rispetto alla condizione di equidistribuzione, quando cioè tutte le unità detengono lo stesso ammontare del carattere.

La concentrazione è pari a 0 se tutte le unità detengono lo stesso ammontare del carattere (equidistribuzione); è massima invece se l'intero ammontare del carattere è detenuto da una sola osservazione.

Sia X="reddito annuo percepito" (migliaia di euro) un carattere quantitativo osservato su 12 impiegati di un'azienda. Sia:

X=reddito	n_i
[50,70)	1
[70,130)	5
[130,170)	4
[170,200)	2
	12

X=reddito	c_i	n_i	$c_i n_i$	$\sum c_i n_i$	f_i	p_i	q_i	$q_i + q_{i-1}$	$f_i(q_i + q_{i-1})$
[50,70)	60	1	60	60	0.083	0.083	60/1530=0.039	0.039	0.003
[70,130)	100	5	500	560	0.416	0.499	560/1530=0.366	0.405	0.168
[130,170)	150	4	600	1160	0.333	0.832	1160/1530=0.758	1.124	0.374
[170,200)	185	2	370	1530	0.166	1	1	1.758	0.292
		12	1530		1				0.837

$p_i = \frac{\sum_{j=1}^{i} n_j}{n}$ frequenza relativa cumulata ossia le frazioni dell'ammontare complessivo che dovrebbero essere possedute dalle prime i unità statistiche, in situazione di equidistribuzione (frazioni teoriche)

$q_i = \frac{\sum_{j=1}^{i} c_j n_j}{\sum_{j=1}^{k} c_j n_j}$ frazione dell'ammontare complessivo possedute dalle prime i unità statistiche (frazioni empiriche).

In caso di equidistribuzione si avrebbe $p_i = q_i$

Indice di concentrazione di Gini (relativo):

$$G = 1 - \sum_{i=0}^{k-1} (p_i - p_{i-1})(q_i + q_{i-1}) = 1 - \sum_{i=0}^{k-1} f_i(q_i + q_{i-1})$$

infatti: $p_i - p_{i-1} = f_i$

per cui nel nostro caso:

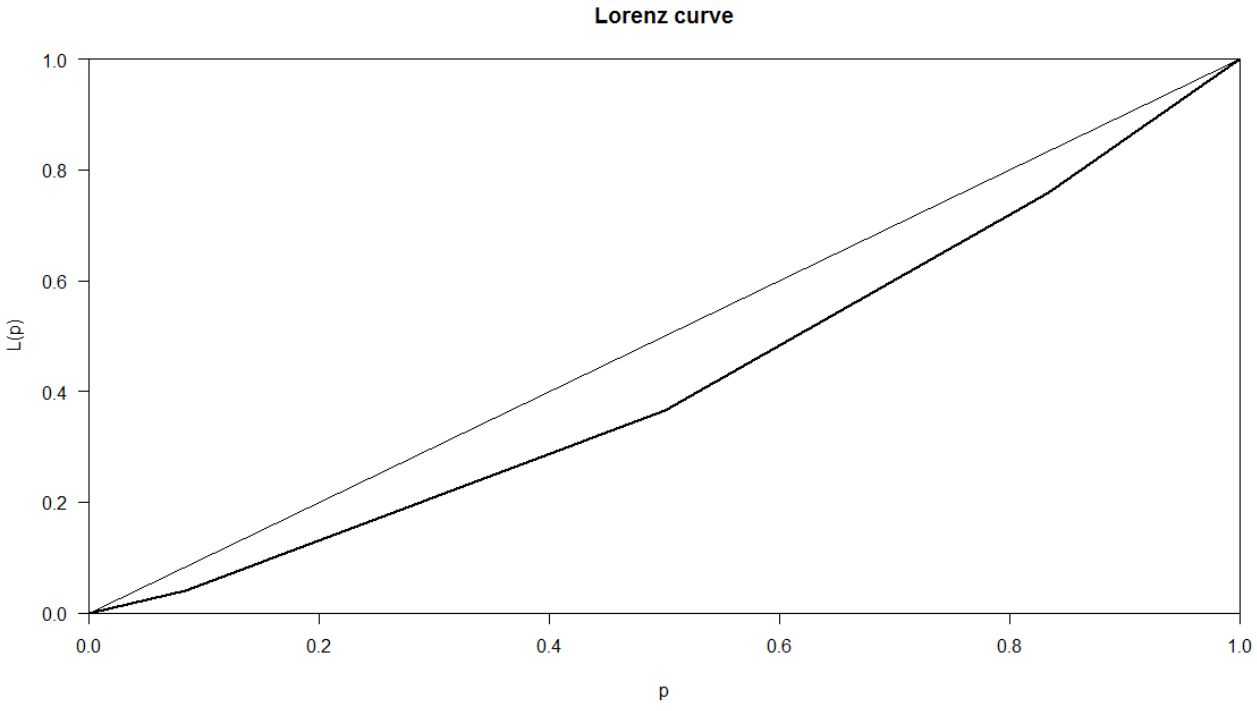
$$G = 1 - \sum_{i=0}^{k-1} f_i(q_i + q_{i-1}) = 1 - 0.837 = 0.163$$

L'indice varia tra 0 e 1 per cui tale valore è indicativo di una situazione molto prossima alla equidistribuzione dei redditi annui percepiti dalle unità statistiche del collettivo.

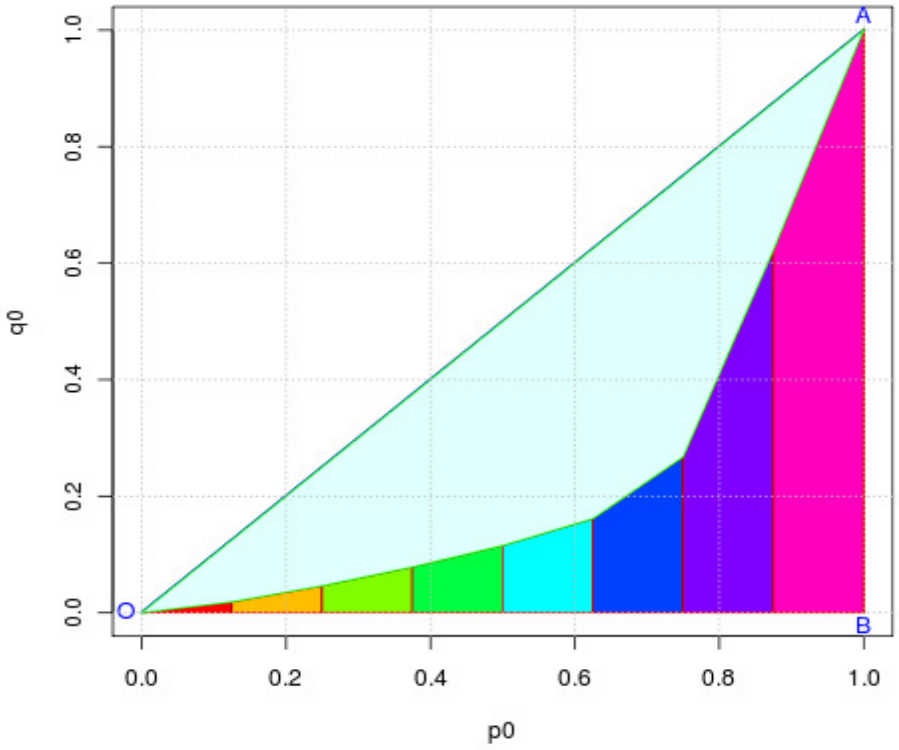
Metodo grafico per determinare il rapporto di concentrazione

La curva di Lorenz è la spezzata passante per i punti di coordinate (p_i, q_i). L'area compresa tra la retta di equidistribuzione e la spezzata di Lorenz è detta area di concentrazione.

Per i nostri dati:



Appendice: derivazione dell'indice di Gini con metodo grafico



$$G = \frac{area\ di\ concentrazione}{area\ del\ triangolo\ OAB} = \frac{area\ del\ triangolo\ OAB - somma\ dei\ trapezi}{area\ del\ triangono\ OAB}$$

- Area del triangolo OAB: $\frac{OB \times BH}{2} = \frac{1 \times 1}{2} = \frac{1}{2}$
- Area di un generico trapezio: $\frac{(B+b) \times h}{2}$
  B=base maggiore=q_i
  b=base minore=q_{i-1}
  h=altezza=$p_i - p_{i-1} = f_i$
- Somma delle aree dei trapezi: $\frac{\sum_{i=0}^{k-1}(q_i + q_{i-1}) \times (p_i - p_{i-1})}{2}$

Sostituendo nella formula di partenza si ha:

$$G = \frac{area\ del\ triangolo\ OAB - somma\ dei\ trapezi}{area\ del\ triangono\ OAB} = \frac{\frac{1}{2} - \frac{\sum_{i=0}^{k-1}(q_i + q_{i-1}) \times (p_i - p_{i-1})}{2}}{\frac{1}{2}} =$$

$$= \frac{\frac{1}{2}(1 - \sum_{i=0}^{k-1}(q_i + q_{i-1}) \times (p_i - p_{i-1}))}{\frac{1}{2}} = 1 - \sum_{i=0}^{k-1}(q_i + q_{i-1}) \times (p_i - p_{i-1})$$