Esercitazioni di statistica

Gli indici di variabilità

Stefania Spina

Universitá di Napoli Federico II

stefania.spina@unina.it

8 Ottobre 2014

Introduzione

I **valori medi** sono indici importanti per la descrizione sintetica di un fenomeno statistico.
Hanno però il **limite** di **non** darci alcuna **informazione sulla distribuzione dei dati**.

La **sintesi** mediante una media è **rappresentativa** soltanto se le **unità statistiche** presentano **modalità prossime a questa**.
Molto spesso, distribuzioni caratterizzate dall'uguaglianza nei valori degli indici di posizione possono riflettere situazioni molto diverse tra di loro.

Esempio

Si consideri la statura di tre gruppi di studenti (X, Y e Z).
I gruppi presentano le seguenti distribuzioni della statura:

$$X = \{160, 161, 164, 190, 195\}$$

$$Y = \{167, 166, 174, 178, 185\}$$

$$Z = \{174, 174, 174, 174, 174\}$$

La statura dei tre gruppi di studenti è sempre di $\bar{x} = 174cm$.

In tutte e tre le prove la media è 174 cm ma i dati sono chiaramente distribuiti in modo diverso.

La variabilità

La sola indicazione di un valore di sintesi, quindi, non consente di descrivere la distribuzione di una carattere.
E' buona norma accompagnare la media con una **misura di variabilità**.

La variabilità esprime la tendenza delle unità di un collettivo ad assumere diverse modalità del carattere.

Gli indici di variabilità

E' possibile distinguere tre categorie di indici di variabilità:

- **indici di dispersione** rispetto ad una media;
- indici di disuguaglianza a coppie (**mutua variabilità** o variabilità reciproca);
- **indici di mutabilità**, che misurano l'omogeneità/eterogeneità tra le modalità di una distribuzione di frequenza.

Nelle tre categorie sopra citate, è possibile operare un'ulteriore distinzione tra gli indici.

a. **Assoluti**: utilizzano la stessa unità di misura della modalità della distribuzione, ma non consentono di fare confronti fra distribuzioni statistiche espresse in unità di misure diverse;
b. **Relativi**: depurano la distribuzione dall'unità di misura, per questo motivo sono particolarmente adatti per operare confronti tra distribuzioni. Si ottengono rapportando un indice assoluto al suo massimo o ad una media.

Variabilità rispetto ad un centro

Gli indici di variabilità più utilizzati che tengono conto della distribuzione di tutti i dati sono:

- Lo **scarto quadratico medio (o deviazione standard**) che si indica con σ (sigma)
- la **varianza della popolazione** che si indica con σ^2.

$$\sigma^2 = \frac{\sum_{i=1}^{n} (x_i - \mu)^2}{n} \qquad \sigma = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \mu)^2}{n}}$$

Ovviamente, se si conoscono i dati mediante una distribuzione di frequenza, le precedenti formule si modificano per tenere conto dei raggruppamenti delle modalità nel modo seguente:

$$\sigma^2 = \frac{\sum_{i=1}^{k} (x_i - \mu)^2 n_i}{\sum_{i=1}^{k} n_i} = \frac{1}{N} \sum_{i=1}^{k} (x_i - \mu)^2 n_i$$

Esercizio

Riprendendo l'esempio delle stature dei tre gruppi, è possibile calcolare la varianza.
Schema di calcolo della varianza per il **gruppo X**

	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$
160	- 14	196
161	- 13	169
164	- 10	100
190	16	256
195	21	441
Totale		1162

Schema di calcolo della varianza per il **gruppo Y**

	$(y_i - \bar{y})$	$(y_i - \bar{y})^2$
167	- 7	49
166	- 8	64
174	0	0
178	4	16
185	11	121
Totale		250

Schema di calcolo della varianza per il **gruppo Z**

	$(z_i - \bar{z})$	$(z_i - \bar{z})^2$
174	0	0
174	0	0
174	0	0
174	0	0
174	0	0
Totale	0	0

Esercizio

$$\sigma_x^2 = \frac{1}{5}\sum_{i=1}^{5}(x_i - \bar{x})^2 = \frac{1162}{5} = 232,4cm^2$$

$$\sigma_y^2 = \frac{1}{5}\sum_{i=1}^{5}(y_i - \bar{y})^2 = \frac{250}{5} = 50cm^2$$

$$\sigma_z^2 = \frac{1}{5}\sum_{i=1}^{5}(z_i - \bar{z})^2 = \frac{0}{5} = 0$$

Esercizio

Di seguito c'è una tabella di frequenza per l'età studenti del Corso di Statistica.
Calcoliamo la varianza

	Frequenze assolute	$x_j n_j$
18	2	36
19	44	836
20	66	1320
21	32	672
22	18	396
23	13	299
24	9	216
25	6	150
Totale	190	3925

Esercizio

$$\bar{x} = \mu = \frac{1}{n}\sum_{j=1}^{k} x_j n_j = \frac{1}{190} * 3925 = 20,6579$$

Calcoliamo la varianza:

$$\sigma^2 = \frac{\sum_{i=1}^{k}(x_i - \mu)^2 n_i}{\sum_{i=1}^{k} n_i} = \frac{1}{N}\sum_{i=1}^{k}(x_i - \mu)^2 n_i$$

	Frequenze assolute (n_i)	$(x_i - \mu)$	$(x_i - \mu)^2$	$(x_i - \mu)^2 n_i$
18	2	-2.66	7.06	14.13
19	44	-1.66	2.75	120.94
20	66	-0.66	0.43	28.57
21	32	0.34	0.12	3.74
22	18	1.34	1.80	32.42
23	13	2.34	5.48	71.31
24	9	3.34	11.17	100.53
25	6	4.34	18.85	113.12
Totale	190			484.76

$$Var(X) = \sigma^2 = \frac{484.76}{190} = 2.55$$

$$\sigma = \sqrt{\sigma^2} = \sqrt{2.55} = 1.60$$

Esercizio

Dato il seguente insieme di numeri:

$$4.5, 6.2, 7.8, 10.4, 15.9$$

Determinare:
a) La devianza;
b) La varianza;
c) Lo scarto quadratico medio;
d) Il coefficiente di variazione.

Esercizio

Determinare la devianza:

$$\mu = (4.5 + 6.2 + 7.8 + 10.4 + 15.9)/5 = 8.96$$

$$DEV(x) = \sum_i (x_i - \mu)^2$$

	$(x_i - \mu)$	$(x_i - \mu)^2$
4.50	-4.46	19.89
6.20	-2.76	7.62
7.80	-1.16	1.35
10.40	1.44	2.07
15.90	6.94	48.16
Totale	0.00	79.09

Esercizio

Determinare la devianza, utilizzando anche la formula semplificata:

$$DEV(x) = \sum_i (x_i)^2 - n\mu^2$$

	$(x_i)^2$
4.50	20.25
6.20	38.44
7.80	60.84
10.40	108.16
15.90	252.81
Totale	480.50

Essendo

$$n\mu^2 = 5 * (8.96)^2 = 401.408$$

$$\Downarrow$$

$$Dev(x) = 480.5 - 401.408 = 79.092$$

Esercizio

Calcolo della varianza

$$\sigma^2 = \frac{\sum_{i=1}^{k}(x_i - \mu)^2}{n} = \frac{Dev(x)}{n} = \frac{79.09}{5} = 15.81$$

Calcolo dello scarto quadratico medio

$$\sigma = \sqrt{Var(x)} = \sqrt{15.81} = 3.97$$

Coefficiente di variazione

Tra gli indici relativi o normalizzati il più utilizzato è il **coefficiente di variazione CV**. Tale coefficiente è:

- **indipendente dall'unità di misura**, cioè è un numero puro che misura la variazione media del fenomeno in rapporto alla sua media aritmetica;
- **è sempre non negativo** per come è costruito;
- è utile per **confrontare la variabilità relativa di un fenomeno in circostanze differenti**

Si può usare nel caso di:
- la variabilità della distribuzione per età tra le varie regioni;
- la distribuzione dei redditi per nazioni e per anno;
- la variabilità del peso rispetto al sesso, ecc...

Si calcola

$$CV = \frac{\sigma}{\mu} * 100$$

in genere, è espresso come percentuale.

Esercizio precedente

Calcolo del coefficiente di variazione

$$CV = \frac{\sigma}{\mu} = \frac{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2}}{\mu} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\frac{(x_i - \mu)^2}{\mu}}$$

$$CV = \frac{\sigma}{\mu} * 100 = \frac{3.97}{8.96} * 100 = 44.39$$

Coefficiente di variazione

La tabella sottostante riporta i risultati di una gara scolastica di cinque allievi per una corsa di resistenza sui 1.000 m (misurata in minuti) e di salto con l'asta (misurata in metri).
Vogliamo conoscere quale delle due serie di risultati presenta maggiore variabilità.

TABELLA
Risultati di una gara scolastica di atletica leggera

Corsa 1000 m (in minuti) (x)	3,00	3,50	4,00	4,50	6,00
Salto in alto (in metri) (y)	1,85	1,60	1,65	1,00	0,80

Coefficiente di variazione

Calcoliamo la media e lo scarto quadratico medio delle due distribuzioni.

$$\mu_x = 4.20 \ \textit{minuti}$$
$$\sigma_x = 1.03 \ \textit{minuti}$$
$$\mu_y = 1.38 \ \textit{metri}$$
$$\sigma_y = 0.41 \ \textit{metri}$$

Poichè le due serie di dati sono espresse in unità di misura diverse, per confrontare la loro variabilità si ricorre al coefficiente di variazione:

$$CV_x = \frac{\sigma_x}{\mu_x} * 100 = \frac{1.03}{4.20} * 100 = 24.5\%$$

$$CV_y = \frac{\sigma_y}{\mu_y} * 100 = \frac{0.41}{1.38} = 29.4\%$$

I risultati del salto in alto presentano maggiore variabilità relativa.

Esercizio

Data la seguente distribuzione di frequenza, determinare:
a) La devianza;
b) La varianza;
c) Lo scarto quadratico medio.

	Frequenze assolute
170 ⊣ 175	14
175 ⊣ 180	18
180 ⊣ 185	28
185 ⊣ 190	33
190 ⊣ 195	17
195 ⊣ 200	15
Totale	125

Esercizio

Calcolo della devianza

$$\mu = 23142.5/125 = 185.14$$

$$c_i = \frac{x_i + x_{i+1}}{2}$$

	n_i	valori centrali (c_i)	$c_i * n_i$	$(c_i - \mu)$	$(c_i - \mu)^2$	$(c_i - \mu)^2 * n_i$
170 ⊣ 175	14	172.5	2415	-12.64	159.76	2236.77
175 ⊣ 180	18	177.5	3195	-7.64	58.36	1050.65
180 ⊣ 185	28	182.5	5110	-2.64	6.96	195.14
185 ⊣ 190	33	187.5	6187.5	2.36	5.56	183.79
190 ⊣ 195	17	192.5	3272.5	7.36	54.16	920.88
195 ⊣ 200	15	197.5	2962.5	12.36	152.76	2291.54
Totale	125		23142.5			6878.8

$$Dev(X) = 6878.8$$

Esercizio

Calcolo della varianza

$$Var(x) = \frac{Dev(x)}{n} = \frac{6878.8}{125} = 55.03$$

Calcolo dello scarto quadratico medio

$$\sigma = \sqrt{Var(x)} = \sqrt{55.03} = 7.41$$

Campo di variazione

Il **campo di variazione = Range** definito come:

$$Range(X) = max(X) - min(X) = x_{(n)} - x_{(i)}$$

Determinare il campo di variazione del seguente insieme di numeri:

$$35, 4, 28, 76, 12, 5, 7$$

1 step
Per ottenere il campo di variazione bisogna **ordinare in modo crescente la sequenza** di numeri

$$4, 5, 7, 12, 28, 35, 76$$

2 step
Fare la **differenza tra il numero più grande e il numero più piccolo**

$$w = 76 - 4 = 72$$

Campo di variazione interquartile

Poichè il campo di variazione è influenzato anche da un solo valore atipico, risulta essere **molto vulnerabile ad errori e situazioni eccezionali**, è preferibile il **campo di variazione interquartile (IQR)**:

$$IQR = Q_3 - Q_1$$

Questo indice è espresso nella **stessa unità di misura del fenomeno ed è robusto** rispetto all'esistenza di valori eccezionali.

Esercizio

Determinare il campo di variazione interquartile della distribuzione riportata in tabella

	Frequenze assolute
170 ⊣ 175	14
175 ⊣ 180	18
180 ⊣ 185	28
185 ⊣ 190	33
190 ⊣ 195	17
195 ⊣ 200	15
Totale	125

Esercizio

Calcolare le frequenze cumulate

	Frequenze assolute	**Frequenze cumulate**
170 ⊣ 175	14	14
175 ⊣ 180	18	32
180 ⊣ 185	28	60
185 ⊣ 190	33	93
190 ⊣ 195	17	110
195 ⊣ 200	15	125
Totale	125	

Esercizio

Determinare il primo e il terzo quartile

	Frequenze assolute	Frequenze cumulate
170 ⊣ 175	14	14
175 ⊣ 180	18	32
180 ⊣ 185	28	60
185 ⊣ 190	33	93
190 ⊣ 195	17	110
195 ⊣ 200	15	125
Totale	125	

$$Q_1 = L_{\inf} + \frac{N/4 - N_{Cl_{(prec)}}}{n_{Cl_{(Q)}}} * c$$

$$Pos\ Q_1 = \frac{N}{4} = \frac{125}{4} = 31.25 \rightarrow 175 \dashv 180$$

$$Pos\ Q_3 = \left(\frac{N}{4} * 3\right) = \frac{125}{4} * 3 = 93.75 \rightarrow 190 \dashv 195$$

$$\boldsymbol{Q_1} = 175 + (31.25 - 14)/18 * 5 = 179.8$$

$$\boldsymbol{Q_3} = 190 + (93.75 - 93)/17 * 5 = 190.22$$

$$IQR = Q_3 - Q_1 = 190.22 - 179.8 = 10.42$$

Introduzione

Un importante aspetto di una rilevazione statistica, e quindi della connessa distribuzione di frequenza, con carattere quantitativo è quello della concentrazione.

Un fenomeno è tanto più concentrato quanto più una piccola frazione delle unità di rilevazione della popolazione possiede una elevata quantità del carattere.

La concentrazione può variare fra due casi estremi:

- assenza di concentrazione
- massima concentrazione

Sono due estremi che sono utilizzati come termine di paragone per stabilire se un caso concreto si avvicina all'uno o all'altro estremo.

La concentrazione nulla

Si ha **concentrazione nulla** quando tutte le unità di rilevazione della popolazione posseggono lo stesso ammontare del carattere.

In questo caso si parla anche di **equiripartizione** del carattere dato che tutti gli N soggetti lo posseggono con la stessa intensità.
Ad esempio, si ha equiripartizione:

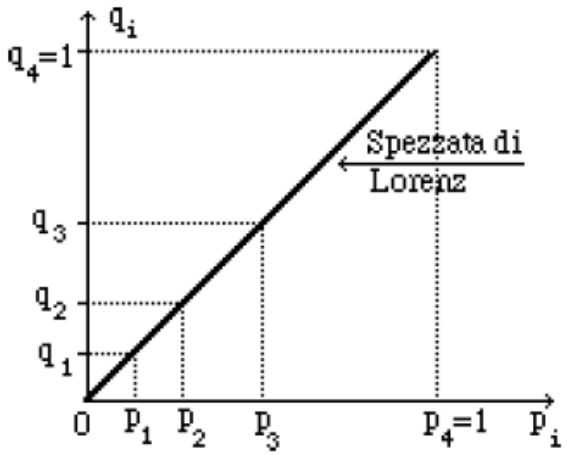
- del reddito in una data popolazione se tutti i soggetti hanno lo stesso ammontare di reddito;
- del possesso di terra in una data popolazione se tutti i componenti di quella popolazione posseggono la stessa estensione di terreno.

La concentrazione nulla

Nel caso di equiripartizione la distribuzione di frequenza associata alla rilevazione diviene semplicemente:

x_i	n_i
μ	N
	N

Dal punto di vista grafico, la concentrazione nulla viene rappresentata attraverso **la spezzata di Lorenz** che coincide con la diagonale del quadrato di lato unitario come evidenziato nella figura che segue;



La concentrazione massima

Si ha **massima concentrazione** quando una sola unità di rilevazione della popolazione possiede tutto l'ammontare del carattere e le rimanenti unità non ne posseggono:

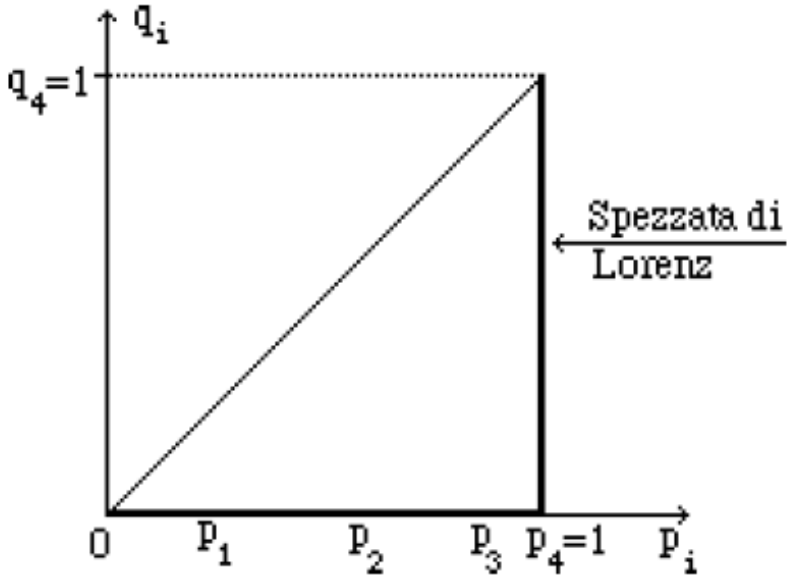
$$x_1 = \ldots = x_{N-1} = 0, x_N = N\mu$$

Se è μ la media della popolazione, nel caso di concentrazione massima la distribuzione di frequenza derivata da quella rilevazione statistica diviene

x_i	n_i
0	N-1
Nμ	1
	N

La concentrazione massima

Dal punto di vista grafico, la concentrazione massima viene rappresentata attraverso la spezzata di Lorenz che coincide, sostanzialmente, con i cateti del triangolo rettangolo definito al di sotto della diagonale del quadrato di lato unitario come evidenziato nella figura che segue



La concentrazione

Volendo elaborare degli indici che permettano di misurare il grado di concentrazione esistente in una generica rilevazione e che siano relativi di modo che possano essere facilmente confrontabili con quelli derivati da rilevazioni espresse con diversa unità di misura e differente numerosità.
Per rendere gli indici indipendenti dalla numerosità N della popolazione si considerano al posto delle frequenze assolute le frequenze relative cumulate che, per motivi storici, nell'ambito della concentrazione vengono indicate con

$$p_i = \frac{i}{n}$$

che è **la percentuale cumulata dei primi i possessori del carattere e risulta indipendente dalla numerosità N della popolazione**.

Per rendere gli **indici indipendenti dall'unità di misura del fenomeno**, al posto delle $x_{(i)}$ consideriamo le **percentuali cumulate del carattere dei primi i possessori** che si indicano di solito con

$$q_i = \frac{1}{N\mu} \sum_{j=1}^{i} x_j$$

La concentrazione

Da un punto di vista operativo, al posto della rilevazione di partenza con modalità xi si ha una nuova rilevazione composta dalle coppie q_i *e* p_i

x_i	$x_{(i)}$	p_i	q_i
x_1	$x_{(1)}$	$\frac{1}{N}$	$\frac{x_{(1)}}{N\mu}$
x_2	$x_{(2)}$	$\frac{2}{N}$	$\frac{x_{(1)}+x_{(2)}}{N\mu}$
x_3	$x_{(3)}$	$\frac{3}{N}$	$\frac{x_{(1)}+x_{(2)}+x_{(3)}}{N\mu}$
...	...	...	...
x_N	$x_{(N)}$	1	1

In generale, fra le p_i e le q_i esiste la seguente relazione $p_i \geq q_i$ che è equivalente a $p_i - q_i \geq 0$, *per* $i = 1, 2, \ldots, N$.

La concentrazione

1. la somma di tutti gli N scarti dalla media è sempre nulla, cioè scarti positivi e scarti negativi si compensano;
2. le $x_{(i)}$ sono ordinate in senso non decrescente;
3. il carattere della distribuzione, perché sia trasferibile, è sempre non negativo, il che implica $\mu > 0$.

Possiamo dire che c'è **equidistribuzione** quando $\boldsymbol{q_i} = \boldsymbol{p_i}$ per ogni $i = 1, 2, 3, \ldots, n$
Nel caso di **concentrazione massima** abbiamo visto che tutte le $\boldsymbol{q_i}$ sono **nulle esclusa l'ultima che è pari ad uno**.

Esercizio 1

Un albergo della penisola Sorrentina nel 2005 ha registrato un decremento rispetto all'anno 2000. I 145 clienti dell'anno 2000 sono diventati 105 nel 2005.
Nella tabella seguente è riportata la distribuzione dei clienti per paese di provenienza nei due anni.
La direzione dell'albergo vuole investigare se l'evoluzione della clientela ha dato luogo ad una maggiore o minore concentrazione della clientela secondo i paesi di provenienza.

	2000	**2005**
Medio Oriente	8	5
Egitto	12	8
America Latina	12	21
Svezia	14	4
Finlandia	15	6
Norvegia	15	20
Sud Africa	15	6
Israele	16	12
Venezuela	18	2
Portogallo	20	21
Totale	145	105

Esercizio 1

	p_i	Dati ordinati 2000	q_i 2000	Dati ordinati 2005	q_i 2005
1	0.1000	8	0.0552	2	0.0190
2	0.2000	12	0.1379	4	0.0571
3	0.3000	12	0.2207	5	0.1048
4	0.4000	14	0.3172	6	0.1619
5	0.5000	15	0.4207	6	0.2190
6	0.6000	15	0.5241	8	0.2952
7	0.7000	15	0.6276	12	0.4095
8	0.8000	16	0.7379	20	0.6000
9	0.9000	18	0.8621	21	0.8000
10	1.0000	20	1.0000	21	1.0000
Totale		145		105	

$$p_i = \frac{i}{n}$$

$$q_i = \frac{1}{N\mu} \sum_{j=1}^{i} x_j$$

Esercizio 1

La concentrazione di un carattere si misura rispetto ad una condizione detta di equidistribuzione.
Si ha **concentrazione nulla** quando l'ammontare totale del carattere è ripartito in parti uguali tra le unità.
Si ha **concentrazione massima** quando tutto il carattere è posseduto da una sola unità, mentre (n-1) unità non possiedono nulla.

$$p_i = \frac{i}{n}$$

$$q_i = \frac{1}{N\mu} \sum_{j=1}^{i} x_j$$

Esercizio 1

Un indice che misura la concentrazione è il **Rapporto di Concentrazione (R) di Gini**.
Si tratta di un indice relativo che varia tra 0 ed 1.

$$R = \frac{\sum_{i=1}^{N-1}(p_i - q_i)}{\sum_{i=1}^{N-1} p_i} = 1 - \frac{\sum_{i=1}^{N-1} q_i}{\sum_{i=1}^{N-1} p_i}$$

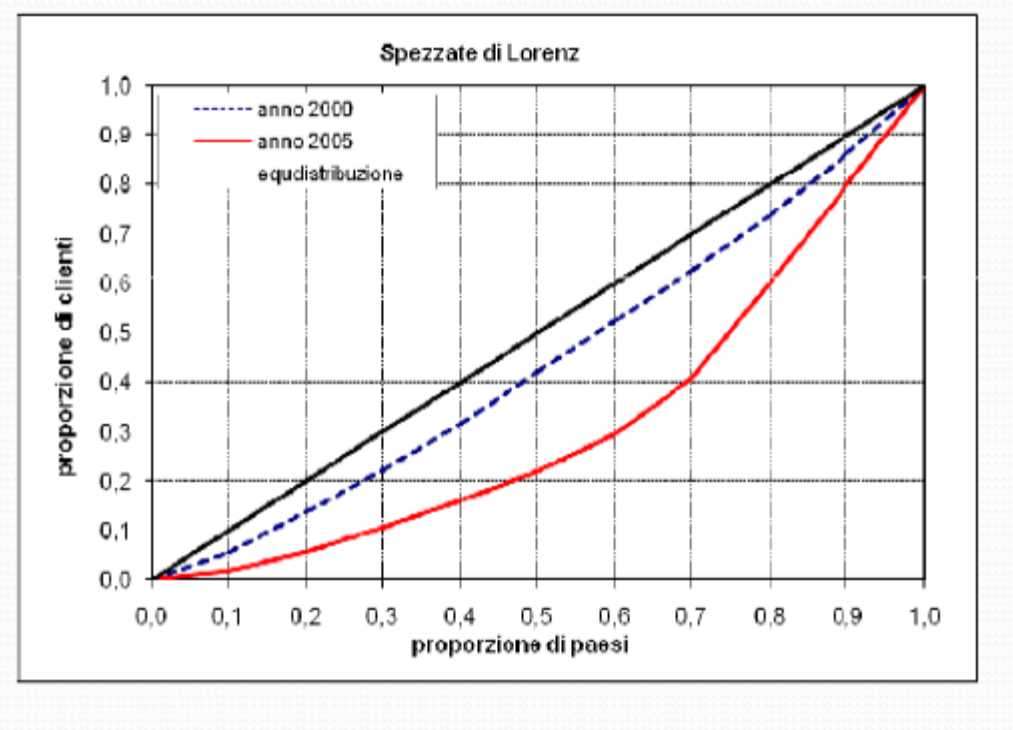
R=0 si ha concentrazione minima.
R=1 si ha concentrazione massima.

$$R_{2000} = 1 - \frac{3.903}{4.5} = 0.13$$

$$R_{2005} = 1 - \frac{2.667}{4.5} = 0.40$$

Esercizio 1

Una rappresentazione grafica della concentrazione può essere fatta attraverso la curva di Lorenz (curva di concentrazione), ovvero la spezzata che si ottiene unendo i punti di coordinate (p_i, q_i) rappresentati sul piano cartesiano.



La bisettrice rappresentata sul grafico rappresenta la situazione di equidistribuzione. L'area compresa tra la curva di concentrazione e la retta di equidistribuzione viene detta **area di concentrazione**.

Esercizio 2

La società Gamma s.p.a., dopo aver effettuato una ricerca di personale qualificato per coprire la posizione di responsabile delle relazioni con l'estero, ha ricevuto 20 curriculum vitae da cittadini sia italiani che stranieri. Tra le informazioni ritenute particolarmente rilevanti dalla società, c'è quella riguardante il livello minimo di reddito mensile desiderato, riportato nella seguente tabella:

classi	n_i
$0.9 \vert - \vert 1.525$	6
$1.525 - \vert 2.15$	8
$2.15 - \vert 2.775$	3
$2.775 - \vert 3.4$	3
Totale	*20*

1. Misurarne la concentrazione e rappresentare la corrispondente curva di Lorenz.

Esercizio 2

1. La misura della concentrazione del livello minimo di reddito tramite la distribuzione per classi di modalità richiede il calcolo del rapporto di concentrazione:

$$R = 1 - \sum_{i=1}^{k} (p_i - p_{i-1}) * (q_i + q_{i-1})$$

con

$$p_i = \frac{i}{n} \sum_{j=1}^{i} n_j$$

$$q_i = \frac{1}{N\mu} \sum_{j=1}^{i} c_j n_j$$

La media del livello minimo di reddito è pari a $\mu = 1.932$, segue che il denominatore delle q_i è $N * \mu = \sum_{j=1}^{i} c_j n_j = 38.64$

Esercizio 2

Utilizzando le formule precedenti, si passa al calcolo delle p_i e delle q_i come riportato in tabella, e dei termini della sommatoria del rapporto di concentrazione.

Livello minimo di reddito

classi	n_i	c_i	$c_i \cdot n_i$	p_i	q_i	$p_i - p_{i-1} = f_i$	$q_i + q_{i-1}$	$(q_i + q_{i-1})f_i$
$0.9 \vert - \vert 1.525$	6	1.213	7.278	0.300	0.188	0.30	0.188	0.056
$1.525 - \vert 2.15$	8	1.838	14.704	0.700	0.569	0.40	0.757	0.303
$2.15 - \vert 2.775$	3	2.463	7.389	0.850	0.760	0.15	1.329	0.199
$2.775 - \vert 3.4$	3	3.088	9.264	1	1	0.15	1.760	0.264
Totale	*20*		*38.635*					*0.822*

Segue quindi che R = 1 - 0.822 = 0.178, ovvero il fenomeno presenta bassa concentrazione.

Esercizio 2

Impiegando i dati in tabella è possibile rappresentare la curva di Lorenz che dà evidenza grafica dei risultati numerici riportati.